Solutions to Casella & Berger's Statistical Inference

Aayush Bajaj

2026-09-21T11:00:00+10:00

Contents

1 Probability Theory **4**
1.1 Exercises 1.1–1.7 . . . 5
1.2 Exercises 1.8–1.14 . . . 8
1.3 Exercises 1.15–1.21 . . . 11
1.4 Exercises 1.22–1.28 . . . 14
1.5 Exercises 1.29–1.35 . . . 18
1.6 Exercises 1.36–1.42 . . . 22
1.7 Exercises 1.43–1.49 . . . 27
1.8 Exercises 1.50–1.55 . . . 31

2 Transformations and Expectations **35**
2.1 Exercises 2.1–2.7 . . . 36
2.2 Exercises 2.8–2.14 . . . 41
2.3 Exercises 2.15–2.21 . . . 45
2.4 Exercises 2.22–2.28 . . . 49
2.5 Exercises 2.29–2.35 . . . 53
2.6 Exercises 2.36–2.40 . . . 58

3 Common Families of Distributions **64**
3.1 Exercises 3.1–3.7 . . . 65
3.2 Exercises 3.8–3.14 . . . 68
3.3 Exercises 3.15–3.21 . . . 73
3.4 Exercises 3.22–3.28 . . . 76
3.5 Exercises 3.29–3.35 . . . 82
3.6 Exercises 3.36–3.42 . . . 88
3.7 Exercises 3.43–3.49 . . . 91
3.8 Exercises 3.50–3.50 . . . 96

4 Multiple Random Variables **98**
4.1 Exercises 4.1–4.7 . . . 99
4.2 Exercises 4.8–4.14 . . . 103
4.3 Exercises 4.15–4.21 . . . 106
4.4 Exercises 4.22–4.28 . . . 111
4.5 Exercises 4.29–4.35 . . . 116
4.6 Exercises 4.36–4.42 . . . 121
4.7 Exercises 4.43–4.49 . . . 126
4.8 Exercises 4.50–4.56 . . . 132
4.9 Exercises 4.57–4.63 . . . 136
4.10 Exercises 4.64–4.65 . . . 141

5 Properties of a Random Sample **143**
5.1 Exercises 5.1–5.7 . . . 144
5.2 Exercises 5.8–5.14 . . . 149
5.3 Exercises 5.15–5.21 . . . 153
5.4 Exercises 5.22–5.28 . . . 158
5.5 Exercises 5.29–5.35 . . . 164
5.6 Exercises 5.36–5.42 . . . 168
5.7 Exercises 5.43–5.49 . . . 173
5.8 Exercises 5.50–5.56 . . . 178
5.9 Exercises 5.57–5.63 . . . 183
5.10 Exercises 5.64–5.69 . . . 188

6 Principles of Data Reduction **195**
6.1 Exercises 6.1–6.7 . . . 196
6.2 Exercises 6.8–6.14 . . . 199
6.3 Exercises 6.15–6.21 . . . 203
6.4 Exercises 6.22–6.28 . . . 207
6.5 Exercises 6.29–6.35 . . . 213
6.6 Exercises 6.36–6.42 . . . 219
6.7 Exercises 6.43–6.43 . . . 223

7 Point Estimation **225**
7.1 Exercises 7.1–7.7 . . . 226
7.2 Exercises 7.8–7.14 . . . 229
7.3 Exercises 7.15–7.21 . . . 234
7.4 Exercises 7.22–7.28 . . . 241
7.5 Exercises 7.29–7.35 . . . 248
7.6 Exercises 7.36–7.42 . . . 256
7.7 Exercises 7.43–7.49 . . . 260
7.8 Exercises 7.50–7.56 . . . 265
7.9 Exercises 7.57–7.63 . . . 271
7.10 Exercises 7.64–7.66 . . . 275

8 Hypothesis Testing **279**
8.1 Exercises 8.1–8.7 . . . 280
8.2 Exercises 8.8–8.14 . . . 285
8.3 Exercises 8.15–8.21 . . . 290
8.4 Exercises 8.22–8.28 . . . 295
8.5 Exercises 8.29–8.35 . . . 299
8.6 Exercises 8.36–8.42 . . . 304
8.7 Exercises 8.43–8.49 . . . 312
8.8 Exercises 8.50–8.56 . . . 318
8.9 Exercises 8.57–8.58 . . . 325

9 Interval Estimation **327**
9.1 Exercises 9.1–9.7 . . . 328
9.2 Exercises 9.8–9.14 . . . 334
9.3 Exercises 9.15–9.21 . . . 339
9.4 Exercises 9.22–9.28 . . . 345
9.5 Exercises 9.29–9.35 . . . 352
9.6 Exercises 9.36–9.42 . . . 358
9.7 Exercises 9.43–9.49 . . . 362
9.8 Exercises 9.50–9.56 . . . 368
9.9 Exercises 9.57–9.58 . . . 374

10 Asymptotic Evaluations **377**
10.1 Exercises 10.1–10.7 . . . 378
10.2 Exercises 10.8–10.14 . . . 381
10.3 Exercises 10.15–10.21 . . . 390
10.4 Exercises 10.22–10.28 . . . 397
10.5 Exercises 10.29–10.35 . . . 403
10.6 Exercises 10.36–10.42 . . . 412
10.7 Exercises 10.43–10.48 . . . 418

11 Analysis of Variance and Regression **425**
11.1 Exercises 11.1–11.7 . . . 426
11.2 Exercises 11.8–11.14 . . . 430
11.3 Exercises 11.15–11.21 . . . 436
11.4 Exercises 11.22–11.28 . . . 443
11.5 Exercises 11.29–11.35 . . . 448
11.6 Exercises 11.36–11.41 . . . 454

12 Regression Models **463**
12.1 Exercises 12.1–12.7 . . . 464
12.2 Exercises 12.8–12.14 . . . 470
12.3 Exercises 12.15–12.21 . . . 477
12.4 Exercises 12.22–12.28 . . . 484
12.5 Exercises 12.29–12.31 . . . 490

Solutions to every exercise in Casella & Berger's *Statistical Inference* (2nd edition, Duxbury, 2002) — 624 exercises across chapters 1–12, from the axioms of probability to ANOVA and regression. The book is filed at Statistical Inference.

1 Probability Theory

Contents

1.1 Exercises 1.1–1.7 5
1.2 Exercises 1.8–1.14 8
1.3 Exercises 1.15–1.21 11
1.4 Exercises 1.22–1.28 14
1.5 Exercises 1.29–1.35 18
1.6 Exercises 1.36–1.42 22
1.7 Exercises 1.43–1.49 27
1.8 Exercises 1.50–1.55 31

1.1 Exercises 1.1–1.7

Problem 1.1. For each of the following experiments, describe the sample space.
(a) Toss a coin four times.
(b) Count the number of insect-damaged leaves on a plant.
(c) Measure the lifetime (in hours) of a particular brand of light bulb.
(d) Record the weights of 10-day-old rats.
(e) Observe the proportion of defectives in a shipment of electronic components.

Solution. (a) The ordered four-tuples of outcomes,

$$S = \{H, T\}^4 = \{(x_1, x_2, x_3, x_4) : x_i \in \{H, T\}\},$$

a set of $2^4 = 16$ points.
(b) $S = \{0, 1, 2, \ldots, N\}$, where N is the number of leaves on the plant. If N is not known in advance one takes $S = \{0, 1, 2, \ldots\}$.
(c) $S = [0, \infty)$; the endpoint 0 covers a bulb dead on installation, and no upper bound on a lifetime is imposed.
(d) $S = (0, \infty)$, since a weight is a positive real number.
(e) With n components in the shipment, the proportion defective is a multiple of $1/n$:

$$S = \left\{0, \tfrac{1}{n}, \tfrac{2}{n}, \ldots, \tfrac{n-1}{n}, 1\right\}.$$

If n is unknown, or the shipment is regarded as arbitrarily large, take $S = [0, 1]$.

Problem 1.2. Verify the following identities.
(a) $A \setminus B = A \setminus (A \cap B) = A \cap B^c$
(b) $B = (B \cap A) \cup (B \cap A^c)$
(c) $B \setminus A = B \cap A^c$
(d) $A \cup B = A \cup (B \cap A^c)$

Solution. Throughout, $A \setminus B := \{x \in S : x \in A \ \wedge \ x \notin B\}$.
(a) For $x \in A$ the conditions $x \notin B$ and $x \notin A \cap B$ coincide, so

$$\begin{aligned} x \in A \setminus B &\iff x \in A \ \wedge \ x \notin B \\ &\iff x \in A \ \wedge \ x \notin A \cap B \iff x \in A \setminus (A \cap B), \end{aligned}$$

and the middle line is verbatim the definition of $A \cap B^c$.
(b) $A \cup A^c = S$, so by the Distributive Law (Theorem 1.1.4c),

$$(B \cap A) \cup (B \cap A^c) = B \cap (A \cup A^c) = B \cap S = B.$$

(c) Interchange the roles of A and B in the second equality of (a).
(d) Distribute over $A \cup A^c = S$ and absorb:

$$\begin{aligned} A \cup (B \cap A^c) &= (A \cup B) \cap (A \cup A^c) \\ &= (A \cup B) \cap S = A \cup B, \end{aligned}$$

using the second Distributive Law of Theorem 1.1.4c.

Problem 1.3. Finish the proof of Theorem 1.1.4. For any events A, B, and C defined on a sample space S, show that
(a) $A \cup B = B \cup A$ and $A \cap B = B \cap A$. (commutativity)

(b) $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$. (associativity)
(c) $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$. (DeMorgan's Laws)

Solution. Every part is inherited from the corresponding property of $\vee$, $\wedge$, $\neg$ acting on the defining membership statements; the proof of the Distributive Law on pp. 3–4 is the template, and as there each chain below is a string of equivalences, so both containments hold at once.
(a) Since $\vee$ and $\wedge$ are commutative,

$$\begin{aligned} x \in A \cup B &\iff (x \in A) \vee (x \in B) \\ &\iff (x \in B) \vee (x \in A) \iff x \in B \cup A, \end{aligned}$$

and the same chain with $\vee$ replaced by $\wedge$ gives $A \cap B = B \cap A$.
(b) Since $\vee$ is associative,

$$\begin{aligned} x \in A \cup (B \cup C) &\iff (x \in A) \vee \big[(x \in B) \vee (x \in C)\big] \\ &\iff \big[(x \in A) \vee (x \in B)\big] \vee (x \in C) \\ &\iff x \in (A \cup B) \cup C. \end{aligned}$$

Replacing $\vee$ by the (likewise associative) $\wedge$ gives $A \cap (B \cap C) = (A \cap B) \cap C$.
(c) These are the two De Morgan equivalences of logic, $\neg(p \vee q) \iff (\neg p) \wedge (\neg q)$ and $\neg(p \wedge q) \iff (\neg p) \vee (\neg q)$, read off membership:

$$\begin{aligned} x \in (A \cup B)^c &\iff \neg\big[(x \in A) \vee (x \in B)\big] \\ &\iff (x \notin A) \wedge (x \notin B) \iff x \in A^c \cap B^c, \end{aligned}$$

$$\begin{aligned} x \in (A \cap B)^c &\iff \neg\big[(x \in A) \wedge (x \in B)\big] \\ &\iff (x \notin A) \vee (x \notin B) \iff x \in A^c \cup B^c. \end{aligned}$$

Problem 1.4. For events A and B, find formulas for the probabilities of the following events in terms of the quantities $P(A)$, $P(B)$, and $P(A \cap B)$.
(a) either A or B or both
(b) either A or B but not both
(c) at least one of A or B
(d) at most one of A or B

Solution. (a) The event is $A \cup B$, so by Theorem 1.2.9b

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

(b) The event is the symmetric difference $(A \cap B^c) \cup (A^c \cap B)$, a disjoint union, so Axiom 3 and Theorem 1.2.9a give

$$\begin{aligned} P(A \cap B^c) + P(A^c \cap B) &= \big[P(A) - P(A \cap B)\big] + \big[P(B) - P(A \cap B)\big] \\ &= P(A) + P(B) - 2P(A \cap B). \end{aligned}$$

(c) At least one of A, B is again the event $A \cup B$, so the answer is that of (a):

$$P(A) + P(B) - P(A \cap B).$$

(d) At most one of A, B excludes only the outcome that both occur, i.e. the event is $(A \cap B)^c$, and by Theorem 1.2.8c

$$P\big((A \cap B)^c\big) = 1 - P(A \cap B).$$

Problem 1.5. Approximately one-third of all human twins are identical (one-egg) and two-thirds are fraternal (two-egg) twins. Identical twins are necessarily the same sex, with male and female being equally likely. Among fraternal twins, approximately one-fourth are both female, one-fourth are both male, and half are one male and one female. Finally, among all U.S. births, approximately 1 in 90 is a twin birth. Define the following events:

$$A = \{\text{a U.S. birth results in twin females}\}$$
$$B = \{\text{a U.S. birth results in identical twins}\}$$
$$C = \{\text{a U.S. birth results in twins}\}$$

(a) State, in words, the event $A \cap B \cap C$.
(b) Find $P(A \cap B \cap C)$.

Solution. (a) $A \cap B \cap C$ is the event that a U.S. birth results in twins, that the twins are identical, and that both are female: a birth of identical female twins. (The condition C is redundant, since $A \subset C$ and $B \subset C$.)
(b) $P(A \cap B \cap C) = 1/540$. Chain the given rates by the multiplication rule (Definition 1.3.2 rearranged, $P(D \cap E) = P(E)P(D \mid E)$), whose three factors are the twin-birth rate $P(C) = 1/90$, the identical fraction among twins $P(B \mid C) = 1/3$, and $P(A \mid B \cap C) = 1/2$ because identical twins are same-sex with each sex equally likely:

$$\begin{aligned} P(A \cap B \cap C) &= P(C)\, P(B \mid C)\, P(A \mid B \cap C) \\ &= \frac{1}{90} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{540} \approx 0.00185. \end{aligned}$$

Problem 1.6. Two pennies, one with $P(\text{head}) = u$ and one with $P(\text{head}) = w$, are to be tossed together independently. Define

$$p_0 = P(0 \text{ heads occur}),$$
$$p_1 = P(1 \text{ head occurs}),$$
$$p_2 = P(2 \text{ heads occur}).$$

Can u and w be chosen such that $p_0 = p_1 = p_2$? Prove your answer.

Solution. No. By independence,

$$p_0 = (1-u)(1-w),$$
$$p_1 = u(1-w) + w(1-u),$$
$$p_2 = uw,$$

and the arithmetic–geometric mean inequality applied to the two nonnegative terms of p_1 gives

$$p_1 \;\geq\; 2\sqrt{u(1-w) \cdot w(1-u)} \;=\; 2\sqrt{p_0 p_2}.$$

The three events partition the sample space, so $p_0 + p_1 + p_2 = 1$; if they were all equal each would equal $1/3$, and the display would read $\frac{1}{3} \geq 2\sqrt{\frac{1}{3} \cdot \frac{1}{3}} = \frac{2}{3}$, which is false. Hence no $u, w \in [0,1]$ work.
Method (2): $p_0 = p_2$ forces $1 - u - w + uw = uw$, i.e. $w = 1 - u$. Writing $t = u(1-u)$, this makes

$$p_1 = u^2 + (1-u)^2 = 1 - 2t, \qquad p_2 = t,$$

so $p_1 = p_2$ demands $t = \frac{1}{3}$; but $t = u(1-u) \leq \frac{1}{4}$ for every real u.

Problem 1.7. Refer to the dart game of Example 1.2.7. In that example the dart board (Figure 1.2.1) is a disc of radius r divided by five equally spaced concentric circles, the distance between consecutive rings being $r/5$; the region scoring i points is the annulus with inner radius $(5-i)r/5$

and outer radius $(6-i)r/5$, for $i = 1, \ldots, 5$ (so the bullseye, scoring 5, is the disc of radius $r/5$ and the outermost band scores 1). The probability of hitting a region is proportional to its area, and in Example 1.2.7 the board is assumed to be hit with probability 1, giving

$$P(\text{scoring } i \text{ points}) = \frac{(6-i)^2 - (5-i)^2}{5^2}, \qquad i = 1, \ldots, 5.$$

Suppose now we do not assume that the probability of hitting the dart board is 1, but rather is proportional to the area of the dart board. Assume that the dart board is mounted on a wall that is hit with probability 1, and the wall has area A.
(a) Using the fact that the probability of hitting a region is proportional to area, construct a probability function for $P(\text{scoring } i \text{ points})$, $i = 0, \ldots, 5$. (No points are scored if the dart board is not hit.)
(b) Show that the conditional probability distribution $P(\text{scoring } i \text{ points} \mid \text{board is hit})$ is exactly the probability distribution of Example 1.2.7.

Solution. (a) The wall is hit with probability 1, so the proportionality constant is $1/A$ and $P(\text{region } R) = \text{area}(R)/A$. The region scoring i points, $1 \le i \le 5$, is the annulus of radii $(5-i)r/5 < (6-i)r/5$, of area $\pi r^2\big[(6-i)^2 - (5-i)^2\big]/25$, whence

$$P(\text{scoring } i \text{ points}) = \frac{\pi r^2}{A} \cdot \frac{(6-i)^2 - (5-i)^2}{5^2}, \quad i = 1, \ldots, 5,$$
$$P(\text{scoring } 0 \text{ points}) = 1 - \frac{\pi r^2}{A},$$

the last line being the complement of the board (Theorem 1.2.8c), since a miss scores nothing. These six numbers are nonnegative (the board lies on the wall, so $\pi r^2 \le A$) and sum to 1, because the bracketed terms telescope:

$$\sum_{i=1}^{5} \big[(6-i)^2 - (5-i)^2\big] = 5^2 - 0^2 = 25.$$

Assigning these masses to the six disjoint outcomes defines a probability function by Theorem 1.2.6.
(b) Let $H = \{\text{board is hit}\} = \bigcup_{i=1}^{5} \{\text{score } i\}$, a disjoint union, so by Axiom 3 and the telescoping sum above $P(H) = \pi r^2/A > 0$. Since $\{\text{score } i\} \subset H$ for $i \ge 1$, Definition 1.3.2 gives

$$\begin{aligned} P(\text{score } i \mid H) &= \frac{P(\{\text{score } i\} \cap H)}{P(H)} \\ &= \frac{(\pi r^2/A)\big[(6-i)^2 - (5-i)^2\big]/5^2}{\pi r^2/A} \\ &= \frac{(6-i)^2 - (5-i)^2}{5^2}, \qquad i = 1, \ldots, 5, \end{aligned}$$

which is Example 1.2.7's distribution; the factor $\pi r^2/A$ cancels, and $P(\text{score } 0 \mid H) = 0$.

1.2 Exercises 1.8–1.14

Problem 1.8. Again refer to the game of darts explained in Example 1.2.7. (In that example the dart board is a disc of radius r divided into five concentric regions by circles of radii $r/5, 2r/5, 3r/5, 4r/5, r$; the innermost disc scores 5 points and the outermost annulus scores 1 point, so the region scoring i points is the annulus with inner radius $(5-i)r/5$ and outer radius $(6-i)r/5$. The board is always hit, and the probability of landing in a region is proportional to its area.)
(a) Derive the general formula for the probability of scoring i points.
(b) Show that $P(\text{scoring } i \text{ points})$ is a decreasing function of i, that is, as the points increase, the probability of scoring them decreases.

(c) Show that $P(\text{scoring } i \text{ points})$ is a probability function according to the Kolmogorov Axioms.

Solution. (a) $P(\text{scoring } i \text{ points}) = \dfrac{11-2i}{25}$, $i = 1, \ldots, 5$. Indeed, the region scoring i points is the annulus between radii $(5-i)r/5$ and $(6-i)r/5$, so proportionality to area gives

$$\begin{aligned} P(\text{scoring } i) &= \frac{\pi\big((6-i)r/5\big)^2 - \pi\big((5-i)r/5\big)^2}{\pi r^2} \\ &= \frac{(6-i)^2 - (5-i)^2}{5^2} = \frac{11-2i}{25}, \end{aligned}$$

independent of π and r.
(b) $P(\text{scoring } i+1) - P(\text{scoring } i) = -\frac{2}{25} < 0$ for $i = 1, \ldots, 4$.
(c) Apply Theorem 1.2.6 with the finite sample space $S = \{1, 2, 3, 4, 5\}$ and $p_i = (11-2i)/25$. Its two hypotheses hold: $p_i \geq \frac{1}{25} > 0$, and

$$\sum_{i=1}^{5} \frac{11-2i}{25} = \frac{55 - 2 \cdot 15}{25} = \frac{25}{25} = 1.$$

Hence $P(A) = \sum_{\{i \,:\, i \in A\}} p_i$ is a probability function on any sigma algebra of subsets of S, i.e. it satisfies the Kolmogorov Axioms of Definition 1.2.4.

Problem 1.9. Prove the general version of DeMorgan's Laws. Let $\{A_\alpha : \alpha \in \Gamma\}$ be a (possibly uncountable) collection of sets. Prove that
(a) $\big(\bigcup_\alpha A_\alpha\big)^c = \bigcap_\alpha A_\alpha^c$.
(b) $\big(\bigcap_\alpha A_\alpha\big)^c = \bigcup_\alpha A_\alpha^c$.

Solution. (a) Both sides consist of exactly the points lying in no A_α:

$$\begin{aligned} x \in \Big(\bigcup_{\alpha\in\Gamma} A_\alpha\Big)^c &\iff \neg\big(\exists\, \alpha \in \Gamma :\ x \in A_\alpha\big) \\ &\iff \forall\, \alpha \in \Gamma :\ x \notin A_\alpha \\ &\iff \forall\, \alpha \in \Gamma :\ x \in A_\alpha^c \\ &\iff x \in \bigcap_{\alpha\in\Gamma} A_\alpha^c. \end{aligned}$$

The second equivalence is the quantifier duality $\neg\exists = \forall\neg$, valid for Γ of any cardinality; countability was never used.
(b) Apply (a) to the collection $\{A_\alpha^c : \alpha \in \Gamma\}$ and use $(A^c)^c = A$:

$$\Big(\bigcup_\alpha A_\alpha^c\Big)^c = \bigcap_\alpha \big(A_\alpha^c\big)^c = \bigcap_\alpha A_\alpha.$$

Complementing both ends gives $\bigcup_\alpha A_\alpha^c = \big(\bigcap_\alpha A_\alpha\big)^c$.

Problem 1.10. Formulate and prove a version of DeMorgan's Laws that applies to a finite collection of sets $A_1, \ldots, A_n$.

Solution. Formulation: for any sets $A_1, \ldots, A_n$,

$$\Big(\bigcup_{i=1}^n A_i\Big)^c = \bigcap_{i=1}^n A_i^c, \qquad \Big(\bigcap_{i=1}^n A_i\Big)^c = \bigcup_{i=1}^n A_i^c.$$

This is Exercise 1.9 with the index set $\Gamma = \{1, 2, \ldots, n\}$, which is a legitimate (indeed finite) choice there, so both identities are already proved.

Method (2): induction from the two-set law, Theorem 1.1.4(c). The case $n = 2$ is that theorem. Assuming the first identity for $n-1$ sets,

$$\Big(\bigcup_{i=1}^{n} A_i\Big)^c = \Big(\Big(\bigcup_{i=1}^{n-1} A_i\Big) \cup A_n\Big)^c$$
$$= \Big(\bigcup_{i=1}^{n-1} A_i\Big)^c \cap A_n^c = \Big(\bigcap_{i=1}^{n-1} A_i^c\Big) \cap A_n^c = \bigcap_{i=1}^{n} A_i^c,$$

the second equality by Theorem 1.1.4(c) and the third by the induction hypothesis. The intersection identity follows by the same induction, or by substituting A_i^c for A_i and complementing as in Exercise 1.9(b).

Problem 1.11. Let S be a sample space.
(a) Show that the collection $\mathcal{B} = \{\emptyset, S\}$ is a sigma algebra.
(b) Let $\mathcal{B} = \{\text{all subsets of } S, \text{ including } S \text{ itself}\}$. Show that $\mathcal{B}$ is a sigma algebra.
(c) Show that the intersection of two sigma algebras is a sigma algebra.

Solution. Each part is a check of the three conditions of Definition 1.2.1.
(a) $\mathcal{B} = \{\emptyset, S\}$. (i) $\emptyset \in \mathcal{B}$. (ii) $\emptyset^c = S \in \mathcal{B}$ and $S^c = \emptyset \in \mathcal{B}$. (iii) If $A_1, A_2, \cdots \in \mathcal{B}$, then $\bigcup_{i=1}^{\infty} A_i = S$ if some $A_i = S$, and $= \emptyset$ otherwise; either way it is in $\mathcal{B}$.
(b) $\mathcal{B}$ is the power set of S, so every condition reduces to the observation that the set produced is again a subset of S: (i) $\emptyset \subset S$; (ii) $A^c = S \setminus A \subset S$; (iii) $\bigcup_{i=1}^{\infty} A_i \subset S$ whenever every $A_i \subset S$.
(c) Let $\mathcal{B}_1, \mathcal{B}_2$ be sigma algebras of subsets of S and put $\mathcal{B} = \mathcal{B}_1 \cap \mathcal{B}_2$, the collection of sets belonging to both. (i) $\emptyset \in \mathcal{B}_1$ and $\emptyset \in \mathcal{B}_2$, so $\emptyset \in \mathcal{B}$. (ii) If $A \in \mathcal{B}$ then $A \in \mathcal{B}_j$ for $j = 1, 2$, hence $A^c \in \mathcal{B}_j$ for $j = 1, 2$, hence $A^c \in \mathcal{B}$. (iii) If $A_1, A_2, \cdots \in \mathcal{B}$ then for each $j = 1, 2$ all the A_i lie in $\mathcal{B}_j$, so $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}_j$; therefore $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$.

Problem 1.12. It was noted in Section 1.2.1 that statisticians who follow the deFinetti school do not accept the Axiom of Countable Additivity, instead adhering to the Axiom of Finite Additivity.
(a) Show that the Axiom of Countable Additivity implies Finite Additivity.
(b) Although, by itself, the Axiom of Finite Additivity does not imply Countable Additivity, suppose we supplement it with the following. Let $A_1 \supset A_2 \supset \cdots \supset A_n \supset \cdots$ be an infinite sequence of nested sets whose limit is the empty set, which we denote by $A_n \downarrow \emptyset$. Consider the following:
Axiom of Continuity: If $A_n \downarrow \emptyset$, then $P(A_n) \to 0$.
Prove that the Axiom of Continuity and the Axiom of Finite Additivity imply Countable Additivity.

Solution. (a) Pad with empty sets. First, $P(\emptyset) = 0$: taking $A_i = \emptyset$ for every i (a pairwise disjoint sequence, since $\emptyset \cap \emptyset = \emptyset$) Axiom 3 gives

$$P(\emptyset) = \sum_{i=1}^{\infty} P(\emptyset),$$

and the right side is $+\infty$ unless $P(\emptyset) = 0$, whereas the left side is a real number. Now let $A, B \in \mathcal{B}$ be disjoint and set $A_1 = A$, $A_2 = B$, $A_i = \emptyset$ for $i \geq 3$. These are pairwise disjoint with union $A \cup B$, so

$$P(A \cup B) = \sum_{i=1}^{\infty} P(A_i)$$
$$= P(A) + P(B) + \sum_{i=3}^{\infty} 0 = P(A) + P(B).$$

(b) Let $A_1, A_2, \cdots \in \mathcal{B}$ be pairwise disjoint and put $A = \bigcup_{i=1}^{\infty} A_i$. Finite additivity extends by induction to any n pairwise disjoint sets, since $\bigcup_{i=1}^{n} A_i = \big(\bigcup_{i=1}^{n-1} A_i\big) \cup A_n$ and these two sets are disjoint.

Let $B_n = \bigcup_{i=n+1}^{\infty} A_i$, which lies in $\mathcal{B}$ by Definition 1.2.1(c). For each n the $n+1$ sets $A_1, \ldots, A_n, B_n$ are pairwise disjoint with union A, so finite additivity gives

$$P(A) = \sum_{i=1}^{n} P(A_i) + P(B_n).$$

The B_n are nested downward, $B_1 \supset B_2 \supset \cdots$, and $\bigcap_{n=1}^{\infty} B_n = \emptyset$: if $x \in B_n$ for every n, then $x \in A_{i_0}$ for some i_0, while $x \in B_{i_0}$ forces $x \in A_i$ for some $i > i_0$, contradicting disjointness. Thus $B_n \downarrow \emptyset$, so the Axiom of Continuity gives $P(B_n) \to 0$ and

$$\begin{aligned} P(A) &= \lim_{n\to\infty} \left(\sum_{i=1}^{n} P(A_i) + P(B_n) \right) \\ &= \sum_{i=1}^{\infty} P(A_i). \end{aligned}$$

Problem 1.13. If $P(A) = \frac{1}{3}$ and $P(B^c) = \frac{1}{4}$, can A and B be disjoint? Explain.

Solution. No. By Theorem 1.2.8(c), $P(B) = 1 - P(B^c) = \frac{3}{4}$. If A and B were disjoint, finite additivity (Exercise 1.12(a)) would give

$$P(A \cup B) = P(A) + P(B) = \tfrac{1}{3} + \tfrac{3}{4} = \tfrac{13}{12} > 1,$$

contradicting $P(A \cup B) \leq 1$ (Theorem 1.2.8(b)).

Problem 1.14. Suppose that a sample space S has n elements. Prove that the number of subsets that can be formed from the elements of S is 2^n.

Solution. There are 2^n: specifying a subset of $S = \{s_1, \ldots, s_n\}$ is a job of n tasks, the ith being the in-or-out decision for s_i and having 2 possible outcomes, so the Fundamental Theorem of Counting (Theorem 1.2.14) gives

$$\underbrace{2 \times 2 \times \cdots \times 2}_{n \text{ factors}} = 2^n.$$

The tasks do indeed determine the subset uniquely: $A \mapsto \big(I_A(s_1), \ldots, I_A(s_n)\big)$ and $(x_1, \ldots, x_n) \mapsto \{s_i : x_i = 1\}$ are mutually inverse (Check!), so the subsets of S are in bijection with $\{0,1\}^n$.
Method (2): count by size. There are $\binom{n}{k}$ subsets of size k, and these classes are disjoint and exhaust all subsets, so the total is

$$\sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n} \binom{n}{k} 1^k 1^{n-k} = (1+1)^n = 2^n$$

by the Binomial Theorem (Theorem 3.2.2).

1.3 Exercises 1.15–1.21

Problem 1.15. Finish the proof of Theorem 1.2.14. Use the result established for $k = 2$ as the basis of an induction argument.
(Theorem 1.2.14, the Fundamental Theorem of Counting: if a job consists of k separate tasks, the ith of which can be done in n_i ways, $i = 1, \ldots, k$, then the entire job can be done in $n_1 \times n_2 \times \cdots \times n_k$ ways. The text proves the case $k = 2$ by direct counting.)

Solution. Bundle the first k tasks into one composite task and apply the case $k = 2$.
Induct on k, the base case $k = 2$ being the counting argument in the proof of Theorem 1.2.14. Assume the theorem for some $k \geq 2$ and let a job consist of $k+1$ tasks, doable in $n_1, \ldots, n_{k+1}$ ways. Let T be the single composite task of doing tasks 1 through k; a way of doing T is exactly a choice of a way for each of those k tasks, so by the induction hypothesis T can be done in $n_1 \times \cdots \times n_k$ ways. The job is now the two tasks T and task $k+1$, so the case $k = 2$ gives

$$(n_1 \times \cdots \times n_k) \times n_{k+1} = n_1 \times \cdots \times n_{k+1}$$

ways, which is the assertion for $k+1$.

Problem 1.16. How many different sets of initials can be formed if every person has one surname and
(a) exactly two given names?
(b) either one or two given names?
(c) either one or two or three given names?
(Answers: (a) 26^3 (b) $26^3 + 26^2$ (c) $26^4 + 26^3 + 26^2$)

Solution. (a) $26^3 = 17{,}576$. A set of initials is an ordered string of three letters (two given names, then the surname), each of the 26 letters available for each position with repetition allowed: the ordered-with-replacement entry n^r of Table 1.2.1 with $n = 26$, $r = 3$.
(b) $26^2 + 26^3 = 18{,}252$, since the one-given-name strings (length 2) and the two-given-name strings (length 3) are disjoint collections.
(c) $26^2 + 26^3 + 26^4 = 475{,}228$, by the same disjoint decomposition into lengths 2, 3 and 4.

Problem 1.17. In the game of dominoes, each piece is marked with two numbers. The pieces are symmetrical so that the number pair is not ordered (so, for example, $(2, 6) = (6, 2)$). How many different pieces can be formed using the numbers $1, 2, \ldots, n$?
(Answer: $n(n+1)/2$)

Solution.

$$\binom{n+2-1}{2} = \binom{n+1}{2} = \frac{n(n+1)}{2}.$$

A piece is an unordered selection of size $r = 2$ from the n numbers, with replacement permitted since doubles (i, i) are legal pieces; that is the unordered-with-replacement entry $\binom{n+r-1}{r}$ of Table 1.2.1.
Method (2): count directly. There are $\binom{n}{2}$ pieces with two distinct numbers and n doubles, and

$$\binom{n}{2} + n = \frac{n(n-1)}{2} + n = \frac{n(n+1)}{2}.$$

Problem 1.18. If n balls are placed at random into n cells, find the probability that exactly one cell remains empty.
(Answer: $\binom{n}{2} n!/n^n$)

Solution.

$$P(\text{exactly one cell empty}) = \frac{\binom{n}{2} n!}{n^n}.$$

The balls are distinguishable and each of the n^n placements (ordered with replacement, Table 1.2.1) is equally likely. With n balls in n cells, exactly one empty cell forces the occupancy pattern to be one cell with 2 balls, one cell with 0, and the remaining $n-2$ cells with 1 each. Build such a placement by Theorem 1.2.14:
(i) choose the empty cell: n ways;
(ii) choose the doubly occupied cell: $n-1$ ways;

(iii) choose the two balls that share it: $\binom{n}{2}$ ways;
(iv) distribute the remaining $n-2$ balls one apiece among the remaining $n-2$ cells: $(n-2)!$ ways.
Hence the count of favorable placements is

$$n(n-1)\binom{n}{2}(n-2)! = \binom{n}{2} n(n-1)(n-2)! \\ = \binom{n}{2} n!,$$

and dividing by n^n gives the stated probability.

Problem 1.19. If a multivariate function has continuous partial derivatives, the order in which the derivatives are calculated does not matter. Thus, for example, the function $f(x, y)$ of two variables has equal third partials

$$\frac{\partial^3}{\partial x^2 \partial y} f(x, y) = \frac{\partial^3}{\partial y \, \partial x^2} f(x, y).$$

(a) How many fourth partial derivatives does a function of three variables have?
(b) Prove that a function of n variables has $\binom{n+r-1}{r}$ rth partial derivatives.

Solution. (a) $\binom{3+4-1}{4} = \binom{6}{4} = 15$, by (b) with $n = 3$, $r = 4$.
(b) Because the order of differentiation is immaterial, an rth partial derivative of $f(x_1, \ldots, x_n)$ is determined by nothing but how many times each variable is differentiated, i.e. by the vector of nonnegative integers $(r_1, \ldots, r_n)$ with $r_1 + \cdots + r_n = r$, the derivative being

$$\frac{\partial^r}{\partial x_1^{r_1} \cdots \partial x_n^{r_n}} f.$$

Such a vector is exactly an unordered sample of size r drawn with replacement from the n variables (x_i drawn r_i times), and by Table 1.2.1 the number of these is $\binom{n+r-1}{r}$.

Problem 1.20. My telephone rings 12 times each week, the calls being randomly distributed among the 7 days. What is the probability that I get at least one call each day?
(Answer: .2285)

Solution.

$$P = \frac{1}{7^{12}} \sum_{k=0}^{7} (-1)^k \binom{7}{k} (7-k)^{12} = 0.2285.$$

Model the week as an ordered-with-replacement assignment of each of the 12 calls to one of the 7 days, so the sample space has 7^{12} equally likely points (Table 1.2.1). Let A_i be the event that day i receives no call. If a fixed set of k days all go without a call, then all 12 calls land among the other $7-k$ days, so

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = \frac{(7-k)^{12}}{7^{12}},$$

a value depending only on k. Inclusion-exclusion over the $\binom{7}{k}$ choices of the k days gives

$$P\left(\bigcup_{i=1}^{7} A_i\right) = \sum_{k=1}^{7} (-1)^{k+1} \binom{7}{k} \frac{(7-k)^{12}}{7^{12}},$$

and the event of interest is the complement, whence the displayed alternating sum. The terms of the numerator are

k	$(-1)^k\binom{7}{k}(7-k)^{12}$
0	13,841,287,201
1	−15,237,476,352
2	5,126,953,125
3	−587,202,560
4	18,600,435
5	−86,016
6	7
7	0

summing to 3,162,075,840, and

$$\frac{3{,}162{,}075{,}840}{13{,}841{,}287{,}201} = 0.22845.$$

Problem 1.21. A closet contains n pairs of shoes. If $2r$ shoes are chosen at random ($2r < n$), what is the probability that there will be no matching pair in the sample?
(Answer: $\binom{n}{2r}2^{2r}/\binom{2n}{2r}$)

Solution.

$$P(\text{no matching pair}) = \frac{\binom{n}{2r}\,2^{2r}}{\binom{2n}{2r}}.$$

The closet holds $2n$ individual shoes, and a random sample of $2r$ of them is one of the $\binom{2n}{2r}$ unordered-without-replacement selections (Table 1.2.1), all equally likely. The sample contains no matching pair exactly when its $2r$ shoes come from $2r$ distinct pairs, one shoe from each; such samples are built by Theorem 1.2.14 in two tasks:
(i) choose which $2r$ of the n pairs are represented: $\binom{n}{2r}$ ways (possible since $2r < n$);
(ii) choose left or right from each represented pair: 2^{2r} ways.
Dividing the product by $\binom{2n}{2r}$ gives the stated probability.

1.4 Exercises 1.22–1.28

Problem 1.22. (a) In a draft lottery containing the 366 days of the year (including February 29), what is the probability that the first 180 days drawn (without replacement) are evenly distributed among the 12 months?
(b) What is the probability that the first 30 days drawn contain none from September?

Solution. (a) $.167 \times 10^{-8}$; (b) $\binom{336}{30}/\binom{366}{30} = .0687$.
The first 180 days form an unordered sample drawn without replacement, so all $\binom{366}{180}$ subsets are equally likely. Evenly distributed means $180/12 = 15$ days from each month, and such a subset is built by choosing 15 days from each month in turn, so by the Fundamental Theorem of Counting (Theorem 1.2.14) the favourable subsets number $\prod_i \binom{d_i}{15}$:

$$P_a = \frac{\prod_{i=1}^{12}\binom{d_i}{15}}{\binom{366}{180}}, \qquad (d_1,\ldots,d_{12}) = (31,29,31,30,31,30,31,31,30,31,30,31).$$

Grouping the seven 31-day months, the four 30-day months and February,

$$P_a = \frac{\binom{31}{15}^7\binom{30}{15}^4\binom{29}{15}}{\binom{366}{180}} = .167 \times 10^{-8}.$$

(b) September contributes 30 days, leaving $366 - 30 = 336$ admissible days. The first 30 draws are an equally likely unordered 30-subset of the 366, and the favourable ones are exactly the 30-subsets of those 336:

$$P_b = \frac{\binom{336}{30}}{\binom{366}{30}} = .0687.$$

Problem 1.23. Two people each toss a fair coin n times. Find the probability that they will toss the same number of heads.

Solution. $\left(\frac{1}{4}\right)^n \binom{2n}{n}$.
Let X and Y be the two head counts. The tosses are independent, and each person's 2^n outcome strings are equally likely with exactly $\binom{n}{k}$ of them carrying k heads, so $P(X = k) = P(Y = k) = \binom{n}{k}2^{-n}$. Decomposing on the common value,

$$\begin{aligned} P(X = Y) &= \sum_{k=0}^{n} P(X = k)P(Y = k) \\ &= \left(\tfrac{1}{2}\right)^{2n} \sum_{k=0}^{n} \binom{n}{k}^2. \end{aligned}$$

Since $\binom{n}{k} = \binom{n}{n-k}$, the sum is $\sum_k \binom{n}{k}\binom{n}{n-k}$, which is the coefficient of x^n in $(1+x)^n(1+x)^n = (1+x)^{2n}$, namely $\binom{2n}{n}$ (Vandermonde). Hence

$$P(X = Y) = \left(\tfrac{1}{4}\right)^n \binom{2n}{n}.$$

Problem 1.24. Two players, A and B, alternately and independently flip a coin and the first player to obtain a head wins. Assume player A flips first.
(a) If the coin is fair, what is the probability that A wins?
(b) Suppose that $P(\text{head}) = p$, not necessarily $\frac{1}{2}$. What is the probability that A wins?
(c) Show that for all p, $0 < p < 1$, $P(\text{A wins}) > \frac{1}{2}$. (Hint: Try to write $P(\text{A wins})$ in terms of the events $E_1, E_2, \ldots$, where $E_i = \{\text{head first appears on } i\text{th toss}\}$.)

Solution. (a) 2/3; (b) $\dfrac{p}{1-(1-p)^2} = \dfrac{1}{2-p}$; (c) immediate from (b).
Do (b) first. The tosses are independent, so $P(E_i) = (1-p)^{i-1}p$, and the E_i are disjoint. Player A makes the odd-numbered tosses, so $\{\text{A wins}\} = \bigcup_{i=1}^{\infty} E_{2i-1}$ and countable additivity (Definition 1.2.4) gives

$$\begin{aligned} P(\text{A wins}) &= \sum_{i=1}^{\infty} p\,(1-p)^{2i-2} \\ &= p\sum_{j=0}^{\infty} \left[(1-p)^2\right]^j \\ &= \frac{p}{1-(1-p)^2} = \frac{1}{2-p}, \end{aligned}$$

the geometric series converging because $(1-p)^2 < 1$ for $0 < p \le 1$.
(a) Put $p = \frac{1}{2}$: $P(\text{A wins}) = 1/(2 - \frac{1}{2}) = 2/3$.
(c) For $0 < p < 1$ we have $2 - p < 2$, hence

$$P(\text{A wins}) = \frac{1}{2-p} > \frac{1}{2}.$$

Problem 1.25. The Smiths have two children. At least one of them is a boy. What is the probability that both children are boys? (See Gardner 1961 for a complete discussion of this problem.)

Solution. 1/3.
Order the children by birth and take the equally likely sample space

$$S = \{BB,\ BG,\ GB,\ GG\},$$

each outcome having probability $\frac{1}{4}$ (boys and girls equally likely, sexes of the two children independent). With $A = \{BB\}$ and $B = \{\text{at least one boy}\} = \{BB, BG, GB\}$ we have $A \cap B = A$, so by Definition 1.3.2

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Problem 1.26. A fair die is cast until a 6 appears. What is the probability that it must be cast more than five times?

Solution. $(5/6)^5 = .4019$.
Let N be the number of the cast on which the first 6 appears. The event $\{N > 5\}$ is exactly the event that each of the first five casts fails to show a 6, and the casts are independent with failure probability $\frac{5}{6}$ each, so

$$P(N > 5) = \left(\frac{5}{6}\right)^5 = \frac{3125}{7776} = .4019.$$

Problem 1.27. Verify the following identities for $n \geq 2$.
(a) $\displaystyle\sum_{k=0}^{n}(-1)^k\binom{n}{k} = 0$
(b) $\displaystyle\sum_{k=1}^{n} k\binom{n}{k} = n2^{n-1}$
(c) $\displaystyle\sum_{k=1}^{n}(-1)^{k+1}k\binom{n}{k} = 0$

Solution. All three are the Binomial Theorem $\sum_{k=0}^{m}\binom{m}{k}x^k = (1+x)^m$ evaluated at $x = \pm 1$, after the absorption identity

$$k\binom{n}{k} = k \cdot \frac{n!}{k!(n-k)!} = n \cdot \frac{(n-1)!}{(k-1)!\,(n-k)!} = n\binom{n-1}{k-1}, \qquad 1 \leq k \leq n.$$

(a) Take $m = n$, $x = -1$:

$$\sum_{k=0}^{n}(-1)^k\binom{n}{k} = (1-1)^n = 0,$$

valid since $n \geq 2 \geq 1$.
(b) Absorb, reindex $j = k - 1$, then take $m = n - 1$, $x = 1$:

$$\begin{aligned}\sum_{k=1}^{n} k\binom{n}{k} &= n\sum_{k=1}^{n}\binom{n-1}{k-1} \\ &= n\sum_{j=0}^{n-1}\binom{n-1}{j} = n2^{n-1}.\end{aligned}$$

(c) The same two moves with $x = -1$; note $(-1)^{k+1} = (-1)^j$ when $j = k - 1$:

$$\begin{aligned}\sum_{k=1}^{n}(-1)^{k+1}k\binom{n}{k} &= n\sum_{j=0}^{n-1}(-1)^j\binom{n-1}{j} \\ &= n(1-1)^{n-1} = 0,\end{aligned}$$

where $n - 1 \geq 1$ is exactly what the hypothesis $n \geq 2$ supplies (for $n = 1$ the sum equals 1).

Problem 1.28. A way of approximating large factorials is through the use of Stirling's Formula:

$$n! \approx \sqrt{2\pi}\, n^{n+(1/2)} e^{-n},$$

a complete derivation of which is difficult. Instead, prove the easier fact,

$$\lim_{n\to\infty} \frac{n!}{n^{n+(1/2)}e^{-n}} = \text{a constant.}$$

(Hint: Feller 1968 proceeds by using the monotonicity of the logarithm to establish that

$$\int_{k-1}^{k} \log x\, dx < \log k < \int_{k}^{k+1} \log x\, dx, \qquad k = 1, \ldots, n,$$

and hence

$$\int_{0}^{n} \log x\, dx < \log n! < \int_{1}^{n+1} \log x\, dx.$$

Now compare $\log n!$ to the average of the two integrals. See Exercise 5.35 for another derivation.)

Solution. Put $d_n = \log\bigl(n!\,/\,n^{n+(1/2)}e^{-n}\bigr)$ and show d_n decreases to a finite limit; the required limit is then $e^{\lim d_n}$.

Explicitly $d_n = \log n! - \bigl(n+\tfrac12\bigr)\log n + n$; this is $\log n!$ measured against the average of the hint's two integrals $n\log n - n$ and $(n+1)\log(n+1) - n$, up to a bounded term.

Monotonicity. Telescoping one step and writing $t = 1/(2n+1)$, so that $(n+1)/n = (1+t)/(1-t)$ and $n+\tfrac12 = 1/(2t)$,

$$\begin{aligned} d_n - d_{n+1} &= \bigl(n+\tfrac12\bigr)\log\frac{n+1}{n} - 1 \\ &= \frac{1}{2t}\log\frac{1+t}{1-t} - 1 \\ &= \frac{1}{2t}\cdot 2\left(t + \frac{t^3}{3} + \frac{t^5}{5} + \cdots\right) - 1 \\ &= \frac{t^2}{3} + \frac{t^4}{5} + \cdots \;>\; 0, \end{aligned}$$

the series being the standard expansion of $\log\frac{1+t}{1-t}$, legitimate since $0 < t < 1$. So (d_n) is strictly decreasing.

Lower bound. Enlarging every coefficient to $\frac13$ and summing the resulting geometric series,

$$\begin{aligned} 0 < d_n - d_{n+1} &< \frac13 \cdot \frac{t^2}{1-t^2} = \frac{1}{3\bigl[(2n+1)^2 - 1\bigr]} \\ &= \frac{1}{12n(n+1)} = \frac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right). \end{aligned}$$

Hence $d_n - \frac{1}{12n} < d_{n+1} - \frac{1}{12(n+1)}$, so $d_n - \frac{1}{12n}$ is increasing and therefore

$$d_n > d_n - \frac{1}{12n} \;\geq\; d_1 - \frac{1}{12} = 1 - \frac{1}{12} \quad \text{for all } n \geq 1.$$

A decreasing sequence bounded below converges, say $d_n \downarrow c$ with $c \geq \frac{11}{12}$ finite. Since exp is continuous,

$$\frac{n!}{n^{n+(1/2)}e^{-n}} = e^{d_n} \longrightarrow e^{c},$$

a finite positive constant.

1.5 Exercises 1.29–1.35

Problem 1.29. (In Example 1.2.20 four numbers are drawn with replacement from $\{2, 4, 9, 12\}$, and all $4^4 = 256$ ordered draws are equally likely.)
(a) For the situation of Example 1.2.20, enumerate the ordered samples that make up the unordered samples $\{4, 4, 12, 12\}$ and $\{2, 9, 9, 12\}$.
(b) Suppose that we had a collection of six numbers, $\{1, 2, 7, 8, 14, 20\}$. What is the probability of drawing, with replacement, the unordered sample $\{2, 7, 7, 8, 14, 14\}$?
(c) Verify that an unordered sample of size k, from m different numbers repeated $k_1, k_2, \ldots, k_m$ times, has

$$\frac{k!}{k_1!\, k_2! \cdots k_m!}$$

ordered components, where $k_1 + k_2 + \cdots + k_m = k$.
(d) Establish that the number of multinomial coefficients, and hence the number of distinct bootstrap samples, is $\binom{k+m-1}{k}$. In other words,

$$\sum_{k_1, k_2, \ldots, k_m} I_{\{k_1 + k_2 + \cdots + k_m = k\}} = \binom{k + m - 1}{k}.$$

Solution. (a) Six and twelve ordered samples respectively, matching $4!/(2!\,2!) = 6$ and $4!/(1!\,2!\,1!) = 12$ from part (c):

$$\begin{aligned}
\{4,4,12,12\}: \quad & (4,4,12,12),\ (4,12,4,12),\ (4,12,12,4),\\
& (12,4,4,12),\ (12,4,12,4),\ (12,12,4,4);\\
\{2,9,9,12\}: \quad & (2,9,9,12),\ (2,9,12,9),\ (2,12,9,9),\\
& (9,2,9,12),\ (9,2,12,9),\ (9,9,2,12),\\
& (9,9,12,2),\ (9,12,2,9),\ (9,12,9,2),\\
& (12,2,9,9),\ (12,9,2,9),\ (12,9,9,2).
\end{aligned}$$

(b) $5/1296 \approx .0039$. The sample has size $k = 6$ with multiplicities $1, 2, 1, 2$ for $2, 7, 8, 14$, so by (c) it is made up of

$$\frac{6!}{1!\,2!\,1!\,2!} = \frac{720}{4} = 180$$

ordered samples, out of $6^6 = 46{,}656$ equally likely ordered draws:

$$P\big(\{2,7,7,8,14,14\}\big) = \frac{180}{46656} = \frac{5}{1296}.$$

(c) Treat the k drawn numbers as distinguishable; counting method 1 of Section 1.2.3 orders them in $k!$ ways. Two of those orderings give the same ordered sample exactly when one comes from the other by permuting the k_i copies of the ith number among themselves, so the map onto ordered samples is exactly $k_1!\, k_2! \cdots k_m!$-to-one. Dividing out that double counting, as for the two 4s in Example 1.2.20, leaves

$$\frac{k!}{k_1!\, k_2! \cdots k_m!}.$$

(d) The sum counts the ordered m-tuples $(k_1, \ldots, k_m)$ of nonnegative integers with $k_1 + \cdots + k_m = k$; since an unordered with-replacement sample of size k from m numbers is determined by, and determines, its multiplicity vector, this is also the number of distinct bootstrap samples. Encode such a tuple as a row of k stars and $m - 1$ bars,

$$\underbrace{\star\cdots\star}_{k_1} \mid \underbrace{\star\cdots\star}_{k_2} \mid \cdots \mid \underbrace{\star\cdots\star}_{k_m},$$

a bijection onto the arrangements of $k+m-1$ symbols of which k are stars. Choosing the star positions gives

$$\sum_{k_1, \ldots, k_m} I_{\{k_1 + \cdots + k_m = k\}} = \binom{k + m - 1}{k},$$

which is the unordered-with-replacement count $\binom{n+r-1}{r}$ of Section 1.2.3 with $n = m$, $r = k$.

Problem 1.30. For the collection of six numbers, $\{1, 2, 7, 8, 14, 20\}$, draw a histogram of the distribution of all possible sample averages calculated from samples drawn with replacement. (As in Example 1.2.20, a sample consists of $n = 6$ draws made with replacement, so that all 6^6 ordered samples are equally likely.)

Solution. Each of the $\binom{11}{6} = 462$ unordered samples (Exercise 1.29(d)) gets the multinomial weight of Exercise 1.29(c),

$$P\big(\{k_1, \ldots, k_6\}\big) = \frac{1}{6^6} \cdot \frac{6!}{k_1!\, k_2! \cdots k_6!},$$

since all $6^6 = 46{,}656$ ordered samples are equally likely; this is the multinomial weighting of Figure 1.2.2. Collecting the ordered samples by their average $\bar{x} = \frac{1}{6}\sum_{j=1}^{6} x_j$ gives 85 distinct values in $[1, 20]$, and grouping into unit-width cells gives the histogram

$\bar{x}$ in	count	probability	bar
$[1, 2)$	63	.0014	
$[2, 3)$	379	.0081	##
$[3, 4)$	1137	.0244	#####
$[4, 5)$	2397	.0514	##########
$[5, 6)$	4016	.0861	##################
$[6, 7)$	5568	.1193	########################
$[7, 8)$	6584	.1411	############################
$[8, 9)$	6792	.1456	##############################
$[9, 10)$	6162	.1321	##########################
$[10, 11)$	4966	.1064	#####################
$[11, 12)$	3586	.0769	###############
$[12, 13)$	2322	.0498	##########
$[13, 14)$	1361	.0292	######
$[14, 15)$	726	.0156	###
$[15, 16)$	351	.0075	##
$[16, 17)$	156	.0033	#
$[17, 18)$	62	.0013	
$[18, 19)$	21	.0005	
$[19, 20)$	6	.0001	
$\{20\}$	1	.00002	

The modal cell $[8, 9)$ holds the population mean $\bar{x} = 52/6 = 26/3$, whose probability $2520/46656 = 35/648$ is the largest of any single average.

Problem 1.31. For the situation of Example 1.2.20, the average of the original set of numbers $\{2, 4, 9, 12\}$ is $\frac{27}{4}$, which has the highest probability.
(a) Prove that, in general, if we sample with replacement from the set $\{x_1, x_2, \ldots, x_n\}$, the outcome with average $(x_1 + x_2 + \cdots + x_n)/n$ is the most likely, having probability $\dfrac{n!}{n^n}$.
(b) Use Stirling's Formula (Exercise 1.28) to show that $n!/n^n \approx \sqrt{2n\pi}/e^n$ (Hall 1992, Appendix I).
(c) Show that the probability that a particular x_i is missing from an outcome is $\big(1 - \frac{1}{n}\big)^n \to e^{-1}$ as $n \to \infty$.

Solution. (a) The winner is the outcome $\{x_1, x_2, \ldots, x_n\}$ that contains each population value exactly once; its average is $\bar{x} = (x_1 + \cdots + x_n)/n$ and its probability is $n!/n^n$.
Label the population items $1, \ldots, n$, so that a sample of size n drawn with replacement is one of n^n equally likely ordered n-tuples of labels. An unordered outcome is described by its multiplicity vector $(k_1, \ldots, k_n)$ with $k_i \geq 0$, $\sum_i k_i = n$, and by Exercise 1.29(c) it consists of $n!/(k_1! \cdots k_n!)$ ordered tuples,

so

$$P(k_1,\dots,k_n)=\frac{1}{n^n}\,\frac{n!}{k_1!\,k_2!\cdots k_n!}.$$

Maximizing P is maximizing the multinomial coefficient $M(k)=n!/\prod_i k_i!$. Suppose some $k_i\ge 2$. Since $\sum_i k_i=n$ over n coordinates, some $k_j=0$. Let k' move one unit from i to j; then

$$\frac{M(k')}{M(k)}=\frac{k_i!\,k_j!}{(k_i-1)!\,(k_j+1)!}=\frac{k_i}{1}=k_i\ \ge 2>1.$$

So no vector with a coordinate ≥ 2 can be maximal, and as the feasible set is finite the maximum is attained only at $k_1=\cdots=k_n=1$, i.e. at the outcome $\{x_1,\dots,x_n\}$, with

$$P(1,1,\dots,1)=\frac{1}{n^n}\,\frac{n!}{1!\cdots 1!}=\frac{n!}{n^n}.$$

(Other outcomes may share the average $\bar{x}$, as in Example 1.2.20; the claim is that none is more probable than this one.)
(b) By Exercise 1.28, $n!\approx\sqrt{2\pi}\,n^{n+(1/2)}e^{-n}$, so

$$\frac{n!}{n^n}\approx\frac{\sqrt{2\pi}\,n^{n+(1/2)}e^{-n}}{n^n}=\sqrt{2\pi n}\,e^{-n}=\frac{\sqrt{2n\pi}}{e^n}.$$

(c) The n draws are independent and each avoids x_i with probability $(n-1)/n$, so

$$P(x_i\text{ missing})=\left(\frac{n-1}{n}\right)^n=\left(1-\frac{1}{n}\right)^n,$$

and since $n\log\left(1-\frac{1}{n}\right)=n\left(-\frac{1}{n}-\frac{1}{2n^2}-\cdots\right)\to -1$, this tends to $e^{-1}\approx .368$.

Problem 1.32. An employer is about to hire one new employee from a group of N candidates, whose future potential can be rated on a scale from 1 to N. The employer proceeds according to the following rules:
(a) Each candidate is seen in succession (in random order) and a decision is made whether to hire the candidate.
(b) Having rejected $m-1$ candidates ($m>1$), the employer can hire the mth candidate only if the mth candidate is better than the previous $m-1$.
Suppose a candidate is hired on the ith trial. What is the probability that the best candidate was hired?

Solution. $P(\text{best candidate hired}\mid\text{hire occurs on trial }i)=\dfrac{i}{N}$.
By rule (b), all that "a candidate was hired on the ith trial" tells us about the ratings is

$$H_i=\{\text{candidate } i \text{ is better than candidates } 1,\dots,i-1\},$$

i.e. candidate i is the best of the first i seen. Write $B=\{\text{candidate } i \text{ is the best of all } N\}$. The interview order is a uniformly random permutation of the N candidates, so by exchangeability the best of the first i is equally likely to occupy any one of those i positions, and the best of all N is equally likely to occupy any one of the N positions:

$$P(H_i)=\frac{1}{i},\qquad P(B)=\frac{1}{N}.$$

Since $B\subset H_i$ (the overall best is in particular better than candidates $1,\dots,i-1$), Definition 1.3.2 gives

$$P(B\mid H_i)=\frac{P(B\cap H_i)}{P(H_i)}=\frac{P(B)}{P(H_i)}=\frac{1/N}{1/i}=\frac{i}{N}.$$

Problem 1.33. Suppose that 5% of men and .25% of women are color-blind. A person is chosen at random and that person is color-blind. What is the probability that the person is male? (Assume males and females to be in equal numbers.)

Solution. $P(M \mid C) = \dfrac{20}{21} \approx .952$.
With M, F the partition into males and females, $P(M) = P(F) = \frac{1}{2}$, and C the event of color-blindness, we are given $P(C \mid M) = .05$ and $P(C \mid F) = .0025$. Bayes' Rule (Theorem 1.3.5) gives

$$\begin{aligned} P(M \mid C) &= \frac{P(C \mid M)P(M)}{P(C \mid M)P(M) + P(C \mid F)P(F)} \\ &= \frac{(.05)(.5)}{(.05)(.5) + (.0025)(.5)} \\ &= \frac{.05}{.0525} = \frac{20}{21} \approx .952. \end{aligned}$$

Problem 1.34. Two litters of a particular rodent species have been born, one with two brown-haired and one gray-haired (litter 1), and the other with three brown-haired and two gray-haired (litter 2). We select a litter at random and then select an offspring at random from the selected litter.
(a) What is the probability that the animal chosen is brown-haired?
(b) Given that a brown-haired offspring was selected, what is the probability that the sampling was from litter 1?

Solution. (a) $P(B) = \dfrac{19}{30} \approx .633$. Litters L_1, L_2 partition the sample space with $P(L_1) = P(L_2) = \frac{1}{2}$, and sampling uniformly within a litter gives $P(B \mid L_1) = \frac{2}{3}$ (two brown of three) and $P(B \mid L_2) = \frac{3}{5}$ (three brown of five). By the law of total probability (the denominator of Theorem 1.3.5),

$$\begin{aligned} P(B) &= \tfrac{1}{2} \cdot \tfrac{2}{3} + \tfrac{1}{2} \cdot \tfrac{3}{5} \\ &= \tfrac{1}{3} + \tfrac{3}{10} = \frac{19}{30}. \end{aligned}$$

(b) $P(L_1 \mid B) = \dfrac{10}{19} \approx .526$:

$$P(L_1 \mid B) = \frac{P(B \mid L_1)P(L_1)}{P(B)} = \frac{\frac{1}{2} \cdot \frac{2}{3}}{19/30} = \frac{1/3}{19/30} = \frac{10}{19}.$$

Problem 1.35. Prove that if $P(\cdot)$ is a legitimate probability function and B is a set with $P(B) > 0$, then $P(\cdot \mid B)$ also satisfies Kolmogorov's Axioms.

Solution. Each of the three axioms of Definition 1.2.4 is inherited from P after division by the constant $P(B) > 0$. Throughout, $P(A \mid B) = P(A \cap B)/P(B)$ by Definition 1.3.2, which is well defined on the same sigma algebra $\mathcal{B}$ since $A \cap B \in \mathcal{B}$ whenever $A, B \in \mathcal{B}$.
(i) Nonnegativity: for any $A \in \mathcal{B}$, $P(A \cap B) \geq 0$ by Axiom 1 for P, and $P(B) > 0$, so

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \;\geq\; 0.$$

(ii) Normalization: $S \cap B = B$, so

$$P(S \mid B) = \frac{P(S \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

(iii) Countable additivity: let $A_1, A_2, \ldots \in \mathcal{B}$ be pairwise disjoint. Then the sets $A_i \cap B$ are also

pairwise disjoint, since

$$(A_i \cap B) \cap (A_j \cap B) \subset A_i \cap A_j = \emptyset \quad (i \neq j),$$

and $(\bigcup_{i=1}^{\infty} A_i) \cap B = \bigcup_{i=1}^{\infty} (A_i \cap B)$ by the distributive law of Theorem 1.1.4, extended to a countable union. Hence, applying Axiom 3 for P to the disjoint sets $A_i \cap B$,

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} A_i \,\Big|\, B\right) &= \frac{P(\bigcup_{i=1}^{\infty}(A_i \cap B))}{P(B)} \\ &= \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} \\ &= \sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)} = \sum_{i=1}^{\infty} P(A_i \mid B), \end{aligned}$$

dividing the convergent nonnegative series termwise by the constant $P(B)$. So $P(\cdot \mid B)$ is a probability function on $(S, \mathcal{B})$.

1.6 Exercises 1.36–1.42

Problem 1.36. If the probability of hitting a target is $\frac{1}{5}$, and ten shots are fired independently, what is the probability of the target being hit at least twice? What is the conditional probability that the target is hit at least twice, given that it is hit at least once?

Solution. P(at least two hits) $= 6095609/9765625 = .6242$, and the conditional probability is $6095609/8717049 = .6993$.
Let X be the number of hits. The shots are independent with hit probability $\frac{1}{5}$, so by Definition 1.3.7 the probability of any specified pattern of j hits and $10 - j$ misses is $(\frac{1}{5})^j (\frac{4}{5})^{10-j}$; there is one pattern with $j = 0$ and ten with $j = 1$. Hence

$$\begin{aligned} P(X \geq 2) &= 1 - \left(\tfrac{4}{5}\right)^{10} - 2\left(\tfrac{4}{5}\right)^{9} \\ &= \frac{5^{10} - 4^{10} - 10 \cdot 4^9}{5^{10}} = \frac{6095609}{9765625} = .6242. \end{aligned}$$

Since $\{X \geq 2\} \subset \{X \geq 1\}$, Exercise 1.38(b) gives the conditional probability as a ratio, with $P(X \geq 1) = 1 - (4/5)^{10} = 8717049/9765625$:

$$P(X \geq 2 \mid X \geq 1) = \frac{P(X \geq 2)}{P(X \geq 1)} = \frac{6095609}{8717049} = .6993.$$

Problem 1.37. Here we look at some variations of Example 1.3.4 (Three prisoners). In that example, prisoners A, B, C are on death row; the governor pardons one of the three, chosen at random so that $P(A) = P(B) = P(C) = \frac{1}{3}$, and tells the warden, who then tells A the name of one of B, C who will die. $\mathcal{W}$ denotes the event that the warden says B will die.
(a) In the warden's calculation of Example 1.3.4 it was assumed that if A were to be pardoned, then with equal probability the warden would tell A that either B or C would die. However, this need not be the case. The warden can assign probabilities γ and $1 - \gamma$ to these events, as shown here:

Prisoner pardoned	Warden tells A	
A	B dies	with probability γ
A	C dies	with probability $1 - \gamma$
B	C dies	
C	B dies	

Calculate $P(A \mid \mathcal{W})$ as a function of γ. For what values of γ is $P(A \mid \mathcal{W})$ less than, equal to, or greater than $\frac{1}{3}$?

(b) Suppose again that $\gamma = \frac{1}{2}$, as in the example. After the warden tells A that B will die, A thinks for a while and realizes that his original calculation was false. However, A then gets a bright idea. A asks the warden if he can swap fates with C. The warden, thinking that no information has been passed, agrees to this. Prove that A's reasoning is now correct and that his probability of survival has jumped to $\frac{2}{3}$.

Solution. (a) $P(A \mid \mathcal{W}) = \dfrac{\gamma}{1+\gamma}$, which is $< \frac{1}{3}$ for $\gamma < \frac{1}{2}$, $= \frac{1}{3}$ at $\gamma = \frac{1}{2}$, and $> \frac{1}{3}$ for $\gamma > \frac{1}{2}$.
With $P(A) = P(B) = P(C) = \frac{1}{3}$, the table gives the joint probabilities of "pardoned" and "warden says B dies":

$$P(A \cap \mathcal{W}) = \tfrac{1}{3}\gamma, \qquad P(B \cap \mathcal{W}) = 0,$$
$$P(C \cap \mathcal{W}) = \tfrac{1}{3} \cdot 1 = \tfrac{1}{3},$$

(if B is pardoned the warden must name C, and if C is pardoned he must name B). Since A, B, C partition the sample space,

$$P(\mathcal{W}) = \tfrac{1}{3}\gamma + 0 + \tfrac{1}{3} = \frac{1+\gamma}{3},$$

so by Definition 1.3.2,

$$P(A \mid \mathcal{W}) = \frac{\gamma/3}{(1+\gamma)/3} = \frac{\gamma}{1+\gamma}.$$

The map $\gamma \mapsto \gamma/(1+\gamma)$ is increasing on $[0,1]$, and since $1+\gamma > 0$,

$$\frac{\gamma}{1+\gamma} < \frac{1}{3} \iff 3\gamma < 1+\gamma \iff \gamma < \tfrac{1}{2},$$
$$\frac{\gamma}{1+\gamma} = \frac{1}{3} \iff \gamma = \tfrac{1}{2},$$
$$\frac{\gamma}{1+\gamma} > \frac{1}{3} \iff \gamma > \tfrac{1}{2}.$$

(b) After the swap A is pardoned exactly when C is pardoned, so A's survival probability is $P(C \mid \mathcal{W}) = \frac{2}{3}$. Indeed, with $\gamma = \frac{1}{2}$ part (a) gives $P(\mathcal{W}) = \frac{1}{2}$ and $P(C \cap \mathcal{W}) = \frac{1}{3}$, so

$$P(C \mid \mathcal{W}) = \frac{1/3}{1/2} = \frac{2}{3},$$

against $P(A \mid \mathcal{W}) = \frac{1}{3}$ from (1.3.2) of Example 1.3.4 had he not swapped.

Problem 1.38. Prove each of the following statements. (Assume that any conditioning event has positive probability.)
(a) If $P(B) = 1$, then $P(A \mid B) = P(A)$ for any A.
(b) If $A \subset B$, then $P(B \mid A) = 1$ and $P(A \mid B) = P(A)/P(B)$.
(c) If A and B are mutually exclusive, then

$$P(A \mid A \cup B) = \frac{P(A)}{P(A) + P(B)}.$$

(d) $P(A \cap B \cap C) = P(A \mid B \cap C)\, P(B \mid C)\, P(C)$.

Solution. (a) $P(B) = 1$ forces $P(B^c) = 0$ by Theorem 1.2.8c, so $P(A \cap B^c) \le P(B^c) = 0$ by Theorem 1.2.9c. Hence by Theorem 1.2.9a,

$$P(A \cap B) = P(A) - P(A \cap B^c) = P(A),$$

and $P(A \mid B) = P(A \cap B)/P(B) = P(A)/1 = P(A)$.

(b) $A \subset B$ gives $A \cap B = A$, so Definition 1.3.2 applied twice yields

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A)}{P(A)} = 1, \qquad P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)}.$$

(c) $A \subset A \cup B$, so $A \cap (A \cup B) = A$, while $A \cap B = \emptyset$ makes $P(A \cup B) = P(A) + P(B)$ by Theorem 1.2.9b (or Axiom 3). Therefore

$$P(A \mid A \cup B) = \frac{P(A \cap (A \cup B))}{P(A \cup B)} = \frac{P(A)}{P(A) + P(B)}.$$

(d) Apply (1.3.3) twice, first with conditioning event $B \cap C$ and then with conditioning event C:

$$\begin{aligned} P(A \cap B \cap C) &= P\big(A \mid B \cap C\big)\, P(B \cap C) \\ &= P\big(A \mid B \cap C\big)\, P(B \mid C)\, P(C). \end{aligned}$$

(Both steps are legitimate since $P(B \cap C) > 0$ forces $P(C) > 0$.)

Problem 1.39. A pair of events A and B cannot be simultaneously mutually exclusive and independent. Prove that if $P(A) > 0$ and $P(B) > 0$, then:
(a) If A and B are mutually exclusive, they cannot be independent.
(b) If A and B are independent, they cannot be mutually exclusive.

Solution. (a) Mutual exclusivity gives $A \cap B = \emptyset$, so $P(A \cap B) = 0$ by Theorem 1.2.8a, whereas

$$P(A)P(B) > 0$$

because both factors are strictly positive. Thus $P(A \cap B) \neq P(A)P(B)$ and (1.3.8) fails: A and B are not independent.
(b) This is the contrapositive of (a). Explicitly, if A and B are independent then $P(A \cap B) = P(A)P(B) > 0$, so $A \cap B \neq \emptyset$ and the events are not mutually exclusive.

Problem 1.40. Finish the proof of Theorem 1.3.9 by proving parts (b) and (c). Theorem 1.3.9 states that if A and B are independent events, then the following pairs are also independent: (a) A and B^c; (b) A^c and B; (c) A^c and B^c.

Solution. (b) Independence (1.3.8) is symmetric in the two events, so part (a) of the theorem, applied to the ordered pair (B, A), already says that B and A^c are independent.
(c) By (b) the pair (A^c, B) is independent, so part (a) applied to it gives independence of A^c and B^c. Method (2): direct computation. For (b), by Theorem 1.2.9a, independence of A and B, and Theorem 1.2.8c,

$$\begin{aligned} P(A^c \cap B) &= P(B) - P(A \cap B) = P(B) - P(A)P(B) \\ &= P(B)\big(1 - P(A)\big) = P(A^c)P(B). \end{aligned}$$

For (c), by DeMorgan's Law, Theorem 1.2.8c and Theorem 1.2.9b,

$$\begin{aligned} P(A^c \cap B^c) &= P\big((A \cup B)^c\big) = 1 - P(A \cup B) \\ &= 1 - P(A) - P(B) + P(A)P(B) \\ &= \big(1 - P(A)\big)\big(1 - P(B)\big) = P(A^c)P(B^c). \end{aligned}$$

Problem 1.41. As in Example 1.3.6, consider telegraph signals "dot" and "dash" sent in the proportion 3:4, where erratic transmissions cause a dot to become a dash with probability $\frac{1}{4}$ and a dash to become a dot with probability $\frac{1}{3}$.

(a) If a dash is received, what is the probability that a dash has been sent?
(b) Assuming independence between signals, if the message dot-dot was received, what is the probability distribution of the four possible messages that could have been sent?

Solution. (a) $P(\text{dash sent} \mid \text{dash received}) = 32/41$.
The proportion 3:4 gives the prior $P(\text{dot sent}) = \frac{3}{7}$, $P(\text{dash sent}) = \frac{4}{7}$, and the stated error rates give the transmission channel

$$P(\text{dash rec} \mid \text{dot sent}) = \tfrac{1}{4}, \qquad P(\text{dot rec} \mid \text{dot sent}) = \tfrac{3}{4},$$
$$P(\text{dot rec} \mid \text{dash sent}) = \tfrac{1}{3}, \quad P(\text{dash rec} \mid \text{dash sent}) = \tfrac{2}{3}.$$

Since $\{\text{dot sent}, \text{dash sent}\}$ partitions the sample space, Bayes' Rule (Theorem 1.3.5) applies:

$$P(\text{dash rec}) = \tfrac{2}{3} \cdot \tfrac{4}{7} + \tfrac{1}{4} \cdot \tfrac{3}{7} = \tfrac{8}{21} + \tfrac{3}{28} = \tfrac{41}{84},$$
$$P(\text{dash sent} \mid \text{dash rec}) = \frac{(2/3)(4/7)}{41/84} = \frac{32/84}{41/84} = \frac{32}{41}.$$

(b) The posterior on the four messages dot-dot, dot-dash, dash-dot, dash-dash is

$$\frac{729}{1849}, \ \frac{432}{1849}, \ \frac{432}{1849}, \ \frac{256}{1849}.$$

Write $a = P(\text{dot sent, dot rec}) = \frac{3}{4} \cdot \frac{3}{7} = \frac{27}{84}$ and $b = P(\text{dash sent, dot rec}) = \frac{1}{3} \cdot \frac{4}{7} = \frac{16}{84}$, so that $P(\text{dot rec}) = a + b = \frac{43}{84}$. By independence of the two symbols, the joint probability of sending (s_1, s_2) and receiving dot-dot factors, so Bayes' Rule gives, for each message (s_1, s_2),

$$P\big((s_1, s_2) \text{ sent} \mid \text{dot-dot rec}\big) = \frac{c_{s_1} c_{s_2}}{(a+b)^2}, \qquad c_{\text{dot}} = a, \ c_{\text{dash}} = b.$$

With 84^2 cancelling throughout, the numerators are $27^2, 27 \cdot 16, 16 \cdot 27, 16^2$ over $43^2 = 1849$:

message sent	posterior probability	decimal
dot-dot	729/1849	.3943
dot-dash	432/1849	.2336
dash-dot	432/1849	.2336
dash-dash	256/1849	.1385

Problem 1.42. The *inclusion-exclusion identity* of Miscellanea 1.8.1 gets its name from the fact that it is proved by the method of inclusion and exclusion (Feller 1968, Section IV.1). Here we go into the details. The probability $P(\cup_{i=1}^n A_i)$ is the sum of the probabilities of all the sample points that are contained in at least one of the A_is. The method of inclusion and exclusion is a recipe for counting these points. Recall from Miscellanea 1.8.1 that

$$P_1 = \sum_{i=1}^{n} P(A_i), \qquad P_2 = \sum_{1 \le i < j \le n} P(A_i \cap A_j),$$

$$P_3 = \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k), \quad \ldots, \quad P_n = P(A_1 \cap \cdots \cap A_n).$$

(a) Let E_k denote the set of all sample points that are contained in exactly k of the events $A_1, A_2, \ldots, A_n$. Show that $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(E_i)$.
(b) Without loss of generality, assume that E_k is contained in $A_1, A_2, \ldots, A_k$. Show that $P(E_k)$ appears k times in the sum P_1, $\binom{k}{2}$ times in the sum P_2, $\binom{k}{3}$ times in the sum P_3, etc.
(c) Show that

$$k - \binom{k}{2} + \binom{k}{3} - \cdots \pm \binom{k}{k} = 1.$$

(See Exercise 1.27.)
(d) Show that parts (a)-(c) imply $\sum_{i=1}^{n} P(E_i) = P_1 - P_2 + \cdots \pm P_n$, establishing the inclusion-exclusion identity.

Solution. (a) A sample point lies in $\cup_{i=1}^{n} A_i$ if and only if it lies in exactly k of the A_i for some k with $1 \le k \le n$, so

$$\bigcup_{i=1}^{n} A_i = \bigcup_{k=1}^{n} E_k,$$

and the E_k are pairwise disjoint (a point lies in exactly one of the counts $1, \ldots, n$). Axiom 3 of Definition 1.2.4 gives $P(\cup_{i=1}^{n} A_i) = \sum_{k=1}^{n} P(E_k)$.
(b) Fix the k-subset and count. Under the stated normalization the points of E_k lie in $A_1, \ldots, A_k$ and in no other A_i, so for any index set $J \subset \{1, \ldots, n\}$,

$$E_k \cap \bigcap_{i \in J} A_i = \begin{cases} E_k, & J \subset \{1, \ldots, k\}, \\ \emptyset, & \text{otherwise.} \end{cases}$$

The sum P_j runs over the $\binom{n}{j}$ index sets J of size j, and E_k contributes $P(E_k)$ to the term for J exactly when $J \subset \{1, \ldots, k\}$, which happens for $\binom{k}{j}$ of them. Hence $P(E_k)$ appears $\binom{k}{1} = k$ times in P_1, $\binom{k}{2}$ times in P_2, $\binom{k}{3}$ times in P_3, and $\binom{k}{j}$ times in P_j (so not at all once $j > k$).
(This is the force of the without-loss-of-generality: in general E_k splits into disjoint pieces, one for each k-subset S of the indices, the display above applies to each piece with $\{1, \ldots, k\}$ replaced by S, and the counts $\binom{k}{j}$ depend only on $|S| = k$.)
(c) Since $\binom{k}{0} = 1$ and $\binom{k}{1} = k$, Exercise 1.27(a) with $n = k$ (legitimate for $k \ge 2$),

$$\sum_{j=0}^{k} (-1)^j \binom{k}{j} = 0,$$

rearranges after moving the $j = 0$ term across as

$$\sum_{j=1}^{k} (-1)^{j+1} \binom{k}{j} = k - \binom{k}{2} + \binom{k}{3} - \cdots \pm \binom{k}{k} = 1.$$

For $k = 1$ the left side is the single term $\binom{1}{1} = 1$.
(d) By (b), each $P_j = \sum_{k=j}^{n} \binom{k}{j} P(E_k)$, so interchanging the two finite sums and applying (c),

$$\begin{aligned} \sum_{j=1}^{n} (-1)^{j+1} P_j &= \sum_{j=1}^{n} (-1)^{j+1} \sum_{k=1}^{n} \binom{k}{j} P(E_k) \\ &= \sum_{k=1}^{n} P(E_k) \sum_{j=1}^{k} (-1)^{j+1} \binom{k}{j} \\ &= \sum_{k=1}^{n} P(E_k) \cdot 1, \end{aligned}$$

where $\binom{k}{j} = 0$ for $j > k$ removes the extra terms. Combining with (a),

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{k=1}^{n} P(E_k) = P_1 - P_2 + P_3 - \cdots \pm P_n.$$

1.7 Exercises 1.43–1.49

Problem 1.43. For the *inclusion-exclusion identity* of Miscellanea 1.8.1: for sets $A_1, A_2, \ldots, A_n$ set

$$P_1 = \sum_{i=1}^{n} P(A_i), \qquad P_2 = \sum_{1\le i<j\le n} P(A_i \cap A_j),$$
$$P_3 = \sum_{1\le i<j<k\le n} P(A_i \cap A_j \cap A_k), \ \ldots,$$
$$P_n = P(A_1 \cap A_2 \cap \cdots \cap A_n),$$

so that $P(A_1 \cup A_2 \cup \cdots \cup A_n) = P_1 - P_2 + P_3 - P_4 + \cdots \pm P_n$, with the asserted sequence of upper and lower bounds

$$P_1 \ge P(\cup_{i=1}^n A_i) \ge P_1 - P_2,$$
$$P_1 - P_2 + P_3 \ge P(\cup_{i=1}^n A_i) \ge P_1 - P_2 + P_3 - P_4, \ \ldots$$

(a) Derive both Boole's and Bonferroni's Inequality from the inclusion-exclusion identity.
(b) Show that the P_i satisfy $P_i \ge P_j$ if $i \le j$ and that the sequence of bounds in Miscellanea 1.8.1 improves as the number of terms increases.
(c) Typically as the number of terms in the bound increases, the bound becomes more useful. However, Schwager (1984) cautions that there are some cases where there is not much improvement, in particular if the A_is are highly correlated. Examine what happens to the sequence of bounds in the extreme case when $A_i = A$ for every i. (See Worsley 1982 and the correspondence of Worsley 1985 and Schwager 1985.)

Solution. (a) Boole's Inequality (Theorem 1.2.11b) *is* the first bound of the display, $P(\cup_{i=1}^n A_i) \le P_1 = \sum_{i=1}^n P(A_i)$, which is the case $m = 1$ of the error formula proved in (b). Bonferroni's Inequality follows by applying Boole to the complements: since $\cup_i A_i^c = (\cap_i A_i)^c$,

$$1 - P\Big(\bigcap_{i=1}^n A_i\Big) = P\Big(\bigcup_{i=1}^n A_i^c\Big) \le \sum_{i=1}^n \big(1 - P(A_i)\big),$$

that is, $P(\cap_{i=1}^n A_i) \ge \sum_{i=1}^n P(A_i) - (n-1)$, which is (1.2.10); the case $n = 2$ is (1.2.9).
(b) Let E_k be the set of sample points lying in exactly k of $A_1, \ldots, A_n$. Exercise 1.42 supplies the two facts $P(\cup_{i=1}^n A_i) = \sum_{k\ge1} P(E_k)$ and $P_j = \sum_{k\ge j} \binom{k}{j} P(E_k)$. Write $S_m = \sum_{j=1}^m (-1)^{j+1} P_j$ for the m-term bound. Interchanging the (finite) sums and using $\sum_{j=0}^m (-1)^j \binom{k}{j} = (-1)^m \binom{k-1}{m}$ (Exercise 1.27),

$$S_m = \sum_{k\ge1} P(E_k) \sum_{j=1}^{m} (-1)^{j+1} \binom{k}{j} = \sum_{k\ge1} P(E_k) \Big[1 - (-1)^m \binom{k-1}{m}\Big],$$

whence the exact error of the m-term bound,

$$S_m - P\Big(\bigcup_{i=1}^n A_i\Big) = (-1)^{m+1} \sum_{k>m} \binom{k-1}{m} P(E_k).$$

Odd m therefore give upper bounds and even m lower bounds (this is the displayed sequence), with error magnitude $D_m = \sum_{k>m} \binom{k-1}{m} P(E_k)$. Since $\binom{k-1}{m+1} \le \binom{k-1}{m}$ exactly when $k \le 2m+2$, and $k \le n$ always, $D_{m+1} \le D_m$ for every $m \ge (n-2)/2$; and $D_n = 0$, so the bounds do improve and close to equality, the identity itself being the case $m = n$.

The printed ordering $P_i \geq P_j$ for $i \leq j$ is, however, false as stated: with $n = 4$ and $A_1 = A_2 = A_3 = A_4 = A$, $P(A) = 1$, one has $P_1 = 4 < 6 = P_2$. Since $\binom{k}{j+1} \leq \binom{k}{j}$ exactly when $k \leq 2j + 1$, what is true is that $P_j \geq P_{j+1}$ whenever $P(E_k) = 0$ for all $k > 2j + 1$; and that ordering is precisely what makes the bounds improve at every step, because

$$S_{m+2} - S_m = (-1)^m \big(P_{m+1} - P_{m+2}\big).$$

Part (c) is the extreme failure of the criterion.
(c) With $A_i = A$ for every i we have $P_j = \binom{n}{j} P(A)$, so by the computation in (b) with $P(E_n) = P(A)$,

$$S_m = P(A) \sum_{j=1}^{m} (-1)^{j+1} \binom{n}{j} = P(A) \Big[1 - (-1)^m \binom{n-1}{m}\Big].$$

The m-term bound thus misses the true value $P(\cup_i A_i) = P(A)$ by $P(A)\binom{n-1}{m}$, an error that *increases* with m up to $m = \lfloor (n-1)/2 \rfloor$ before decreasing, and is zero only at $m = n$. For $n = 10$ and $P(A) = .1$ the successive bounds are

$$S_1 = 1.0, \quad S_2 = -3.5, \quad S_3 = 8.5, \quad S_4 = -12.5, \ \ldots, \quad S_{10} = .1,$$

every intermediate one vacuous: with perfectly correlated events no partial expansion improves on the trivial bounds, and all n terms are needed.

Problem 1.44. Standardized tests provide an interesting application of probability theory. Suppose first that a test consists of 20 multiple-choice questions, each with 4 possible answers. If the student guesses on each question, then the taking of the exam can be modeled as a sequence of 20 independent events. Find the probability that the student gets at least 10 questions correct, given that he is guessing.

Solution. $P(X \geq 10) = .0139$.
Let C_i be the event that question i is answered correctly. Guessing uniformly among the 4 answers gives $P(C_i) = \frac{1}{4}$, and the C_i are mutually independent (Definition 1.3.12), so $X = \#\{i : C_i \text{ occurs}\}$ has the binomial form: a fixed pattern with x correct and $20 - x$ incorrect has probability $(\frac{1}{4})^x (\frac{3}{4})^{20-x}$ by independence, and there are $\binom{20}{x}$ such patterns. Hence

$$P(X \geq 10) = \sum_{x=10}^{20} \binom{20}{x} \Big(\frac{1}{4}\Big)^x \Big(\frac{3}{4}\Big)^{20-x} = \frac{1}{4^{20}} \sum_{x=10}^{20} \binom{20}{x} 3^{20-x}.$$

The sum is 15,244,087,642 and $4^{20} = 1{,}099{,}511{,}627{,}776$, so

$$P(X \geq 10) = \frac{15{,}244{,}087{,}642}{1{,}099{,}511{,}627{,}776} = .01386.$$

Problem 1.45. Show that the induced probability function defined in (1.4.1) defines a legitimate probability function in that it satisfies the Kolmogorov Axioms. (Recall (1.4.1): for a sample space S with probability function P and a random variable X with range $\mathcal{X}$, $P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\})$, and more generally (1.4.2), $P_X(X \in A) = P(\{s \in S : X(s) \in A\})$ for $A \subset \mathcal{X}$.)

Solution. Write $X^{-1}(A) = \{s \in S : X(s) \in A\}$, so that (1.4.2) reads $P_X(A) = P\big(X^{-1}(A)\big)$; each of the three axioms of Definition 1.2.4 for P_X is the corresponding axiom for P transported through X^{-1}.
(1) $P_X(A) = P\big(X^{-1}(A)\big) \geq 0$, since P satisfies the first axiom.
(2) $X(s) \in \mathcal{X}$ for every $s \in S$, so $X^{-1}(\mathcal{X}) = S$ and $P_X(\mathcal{X}) = P(S) = 1$.
(3) Let $A_1, A_2, \cdots \subset \mathcal{X}$ be pairwise disjoint. Their preimages are pairwise disjoint, since $s \in X^{-1}(A_i) \cap$

$X^{-1}(A_j)$ would force $X(s) \in A_i \cap A_j = \emptyset$, and preimages respect unions:

$$X^{-1}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \bigcup_{i=1}^{\infty} X^{-1}(A_i),$$

both sides being $\{s : X(s) \in A_i \text{ for some } i\}$. Countable additivity of P therefore gives

$$P_X\Big(\bigcup_{i=1}^{\infty} A_i\Big) = P\Big(\bigcup_{i=1}^{\infty} X^{-1}(A_i)\Big) = \sum_{i=1}^{\infty} P\big(X^{-1}(A_i)\big) = \sum_{i=1}^{\infty} P_X(A_i).$$

Problem 1.46. Seven balls are distributed randomly into seven cells. Let $X_i =$ the number of cells containing exactly i balls. What is the probability distribution of X_3? (That is, find $P(X_3 = x)$ for every possible x.)

Solution.

x	0	1	2
$P(X_3 = x)$	$\frac{10627}{16807}$	$\frac{120}{343}$	$\frac{300}{16807}$
decimal	.6323	.3499	.0178

Here *randomly* means the $7^7 = 823{,}543$ assignments of the seven (distinguishable) balls to the seven cells are equally likely, so every probability below is a count divided by 7^7. Only $x = 0, 1, 2$ are possible, since three cells with three balls each would need nine balls.
$x = 2$: pick the two triply occupied cells, $\binom{7}{2} = 21$ ways; pick their balls, $\binom{7}{3}\binom{4}{3} = 35 \cdot 4 = 140$ ways; drop the last ball in any of the remaining five cells, 5 ways. Thus

$$P(X_3 = 2) = \frac{21 \cdot 140 \cdot 5}{7^7} = \frac{14{,}700}{823{,}543} = \frac{300}{16807}.$$

$x = 1$: pick the triple cell, 7 ways, and its three balls, $\binom{7}{3} = 35$ ways; the remaining four balls go into the other six cells with no cell receiving exactly three. Of the $6^4 = 1296$ placements, those with a triple are counted as $6 \cdot \binom{4}{3} \cdot 5 = 120$ (cell, its three balls, the stray ball elsewhere), and no placement of four balls has two triples, so $1296 - 120 = 1176$ are admissible. Thus

$$P(X_3 = 1) = \frac{7 \cdot 35 \cdot 1176}{7^7} = \frac{288{,}120}{823{,}543} = \frac{120}{343}.$$

$x = 0$: by complementation,

$$\begin{aligned} P(X_3 = 0) &= 1 - \frac{288{,}120 + 14{,}700}{823{,}543} \\ &= \frac{520{,}723}{823{,}543} = \frac{10627}{16807}. \end{aligned}$$

Problem 1.47. Prove that the following functions are cdfs.
(a) $\frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x)$, $x \in (-\infty, \infty)$
(b) $(1 + e^{-x})^{-1}$, $x \in (-\infty, \infty)$
(c) $e^{-e^{-x}}$, $x \in (-\infty, \infty)$
(d) $1 - e^{-x}$, $x \in (0, \infty)$
(e) the function defined in (1.5.6), namely, for some ϵ with $0 < \epsilon < 1$,

$$F_Y(y) = \begin{cases} \dfrac{1-\epsilon}{1+e^{-y}} & \text{if } y < 0, \\ \epsilon + \dfrac{1-\epsilon}{1+e^{-y}} & \text{if } y \geq 0. \end{cases}$$

Solution. In (a)-(d) the function is differentiable with $F' > 0$ on the interior of its domain, which delivers conditions (b) and (c) of Theorem 1.5.3 at once, so only the limits of (a) are at issue; (e) is not continuous and is treated separately.
(a) $\lim_{x\to-\infty} \tan^{-1} x = -\pi/2$ and $\lim_{x\to\infty} \tan^{-1} x = \pi/2$, so F runs from $\frac{1}{2} - \frac{1}{2} = 0$ to $\frac{1}{2} + \frac{1}{2} = 1$, and

$$F'(x) = \frac{1}{\pi(1+x^2)} > 0.$$

(b) As $x \to -\infty$, $e^{-x} \to \infty$ and $F(x) \to 0$; as $x \to \infty$, $e^{-x} \to 0$ and $F(x) \to 1$. Moreover

$$F'(x) = \frac{e^{-x}}{(1+e^{-x})^2} > 0.$$

(This is the logistic cdf (1.5.5) of Example 1.5.5.)
(c) As $x \to -\infty$, $e^{-x} \to \infty$, so $e^{-e^{-x}} \to 0$; as $x \to \infty$, $e^{-x} \to 0$, so $e^{-e^{-x}} \to e^0 = 1$. And

$$F'(x) = e^{-e^{-x}} e^{-x} > 0.$$

(d) Extend F to the whole line by $F(x) = 0$ for $x \le 0$, the only extension that is a cdf, since $1 - e^{-x} \to 0$ as $x \downarrow 0$ makes the extension continuous there. Then $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} (1 - e^{-x}) = 1$, and $F'(x) = e^{-x} > 0$ on $(0, \infty)$ while F is constant on $(-\infty, 0]$; a continuous function that is nondecreasing on each of two abutting intervals is nondecreasing.
(e) Both pieces are $(1 - \epsilon)$ times the increasing function of part (b), shifted by a constant, so F_Y is increasing on $(-\infty, 0)$ and on $[0, \infty)$. Its limits are

$$\lim_{y\to-\infty} F_Y(y) = 0, \qquad \lim_{y\to\infty} F_Y(y) = \epsilon + (1-\epsilon) = 1.$$

At the only possible break, $y = 0$,

$$\lim_{y\uparrow 0} F_Y(y) = \frac{1-\epsilon}{2} \;<\; \epsilon + \frac{1-\epsilon}{2} = F_Y(0) = \lim_{y\downarrow 0} F_Y(y),$$

since F_Y is given by its second branch on all of $[0, \infty)$. So F_Y is nondecreasing across 0 as well, and right-continuous everywhere, with a jump of height ϵ at the origin.

Problem 1.48. Prove the necessity part of Theorem 1.5.3. (Theorem 1.5.3: the function $F(x)$ is a cdf if and only if (a) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$; (b) $F(x)$ is a nondecreasing function of x; (c) $F(x)$ is right-continuous, that is, for every number x_0, $\lim_{x\downarrow x_0} F(x) = F(x_0)$.)

Solution. Let $F(x) = F_X(x) = P(X \le x)$ for a random variable X on (S, P) (Definition 1.5.1); all three properties are properties of P along monotone sequences of events. Those we need first: if $A_1 \subset A_2 \subset \cdots$ with $A = \cup_n A_n$, then writing A as the disjoint union $A_1 \cup \bigcup_{n\ge 2}(A_n \cap A_{n-1}^c)$ and applying the third Kolmogorov Axiom gives $P(A_n) \to P(A)$; taking complements gives the decreasing version, $A_1 \supset A_2 \supset \cdots$ with $A = \cap_n A_n$ implies $P(A_n) \to P(A)$.
(b) If $x < y$ then $\{X \le x\} \subset \{X \le y\}$, so $F(x) \le F(y)$ by Theorem 1.2.9c. Monotonicity is proved first because it is what turns each sequential limit below into a genuine limit.
(a) The events $\{X \le -n\}$ decrease, and their intersection is empty: $X(s)$ is a real number for every $s \in S$, so $X(s) > -n$ for all large n. Hence $F(-n) \to P(\emptyset) = 0$. Likewise $\{X \le n\}$ increase to S, so $F(n) \to P(S) = 1$. Since F is monotone and bounded, the limits along these sequences are the limits as $x \to -\infty$ and $x \to \infty$.
(c) Fix x_0. The events $\{X \le x_0 + \frac{1}{n}\}$ decrease with

$$\bigcap_{n=1}^{\infty} \left\{X \le x_0 + \tfrac{1}{n}\right\} = \{X \le x_0\},$$

because $X(s) \le x_0 + \frac{1}{n}$ for all n forces $X(s) \le x_0$. Therefore $F(x_0 + \frac{1}{n}) \to F(x_0)$, and monotonicity of F upgrades this to $\lim_{x \downarrow x_0} F(x) = F(x_0)$.

Problem 1.49. A cdf F_X is *stochastically greater* than a cdf F_Y if $F_X(t) \le F_Y(t)$ for all t and $F_X(t) < F_Y(t)$ for some t. Prove that if $X \sim F_X$ and $Y \sim F_Y$, then

$$P(X > t) \ge P(Y > t) \quad \text{for every } t$$

and

$$P(X > t) > P(Y > t) \quad \text{for some } t;$$

that is, X tends to be bigger than Y.

Solution. Complementation turns each hypothesis directly into its conclusion. Since $\{X > t\} = \{X \le t\}^c$, Theorem 1.2.8c gives $P(X > t) = 1 - F_X(t)$ and likewise $P(Y > t) = 1 - F_Y(t)$. Hence for every t, the hypothesis $F_X(t) \le F_Y(t)$ yields

$$P(X > t) = 1 - F_X(t) \;\ge\; 1 - F_Y(t) = P(Y > t),$$

and at a t with $F_X(t) < F_Y(t)$ – one exists by the definition of stochastically greater – the same line with strict inequalities gives $P(X > t) > P(Y > t)$.

1.8 Exercises 1.50–1.55

Problem 1.50. Verify formula (1.5.4), the formula for the partial sum of the geometric series:

$$\sum_{k=1}^{n} t^{k-1} = \frac{1 - t^n}{1 - t}, \qquad t \ne 1.$$

Solution. Induction on n; the entire content is the one-line step

$$\frac{1 - t^n}{1 - t} + t^n = \frac{1 - t^{n+1}}{1 - t}.$$

For $n = 1$ both sides equal 1. Assuming the formula for n,

$$\begin{aligned} \sum_{k=1}^{n+1} t^{k-1} &= \frac{1 - t^n}{1 - t} + t^n \\ &= \frac{1 - t^n + t^n(1 - t)}{1 - t} \\ &= \frac{1 - t^{n+1}}{1 - t}, \end{aligned}$$

which is the formula for $n + 1$.
Method (2): telescoping, with no induction. For $t \ne 1$,

$$(1 - t)\sum_{k=1}^{n} t^{k-1} = \sum_{k=1}^{n} t^{k-1} - \sum_{k=1}^{n} t^k = 1 - t^n,$$

and divide by $1 - t$.

Problem 1.51. An appliance store receives a shipment of 30 microwave ovens, 5 of which are (unknown to the manager) defective. The store manager selects 4 ovens at random, without replacement,

and tests to see if they are defective. Let X = number of defectives found. Calculate the pmf and cdf of X and plot the cdf.

Solution. X is hypergeometric:

$$f_X(x) = P(X = x) = \frac{\binom{5}{x}\binom{25}{4-x}}{\binom{30}{4}}, \qquad x = 0, 1, 2, 3, 4,$$

since the $\binom{30}{4} = 27{,}405$ equally likely samples of size 4 split according to how many of the 5 defectives they contain. (The five counts below sum to 27,405, so $\sum_x f_X(x) = 1$.)

x	$\binom{5}{x}\binom{25}{4-x}$	$f_X(x)$ exact	$f_X(x)$	$F_X(x)$
0	12,650	2530/5481	0.4616	0.4616
1	11,500	2300/5481	0.4196	0.8812
2	3,000	200/1827	0.1095	0.9907
3	250	50/5481	0.0091	0.9998
4	5	1/5481	0.0002	1.0000

The cdf is the step function

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 0.4616 & 0 \le x < 1 \\ 0.8812 & 1 \le x < 2 \\ 0.9907 & 2 \le x < 3 \\ 0.9998 & 3 \le x < 4 \\ 1 & x \ge 4. \end{cases}$$

The plot is a right-continuous staircase, flat between integers, with jumps at $x = 0, 1, 2, 3, 4$ of heights 0.4616, 0.4196, 0.1095, 0.0091, 0.0002.

Problem 1.52. Let X be a continuous random variable with pdf $f(x)$ and cdf $F(x)$. For a fixed number x_0, define the function

$$g(x) = \begin{cases} f(x)/[1 - F(x_0)] & x \ge x_0 \\ 0 & x < x_0. \end{cases}$$

Prove that $g(x)$ is a pdf. (Assume that $F(x_0) < 1$.)

Solution. g satisfies the two conditions of Theorem 1.6.5: it is nonnegative, and it integrates to 1.
Nonnegativity: $f(x) \ge 0$ and $1 - F(x_0) > 0$ by the standing assumption $F(x_0) < 1$, so $g(x) \ge 0$ for every x.
Total mass: by (1.6.3), $\int_{x_0}^{\infty} f(x)\,dx = 1 - F(x_0)$, hence

$$\int_{-\infty}^{\infty} g(x)\,dx = \frac{1}{1 - F(x_0)} \int_{x_0}^{\infty} f(x)\,dx = \frac{1 - F(x_0)}{1 - F(x_0)} = 1.$$

Problem 1.53. A certain river floods every year. Suppose that the low-water mark is set at 1 and the high-water mark Y has distribution function

$$F_Y(y) = P(Y \le y) = 1 - \frac{1}{y^2}, \qquad 1 \le y < \infty.$$

(a) Verify that $F_Y(y)$ is a cdf.
(b) Find $f_Y(y)$, the pdf of Y.
(c) If the low-water mark is reset at 0 and we use a unit of measurement that is $\frac{1}{10}$ of that given

previously, the high-water mark becomes $Z = 10(Y-1)$. Find $F_Z(z)$.

Solution. (a) The three conditions of Theorem 1.5.3 hold for

$$F_Y(y) = \begin{cases} 0 & y < 1 \\ 1 - y^{-2} & y \geq 1, \end{cases}$$

namely: $\lim_{y\to-\infty} F_Y(y) = 0$ and $\lim_{y\to\infty} F_Y(y) = 1 - \lim_{y\to\infty} y^{-2} = 1$; F_Y is nondecreasing, being constant on $(-\infty, 1)$ and having $\frac{d}{dy}(1-y^{-2}) = 2y^{-3} > 0$ on $(1,\infty)$; and F_Y is continuous everywhere (at the join, $F_Y(1) = 1-1 = 0$), hence in particular right-continuous.
(b) Differentiating, by (1.6.3),

$$f_Y(y) = \frac{d}{dy}\left(1 - \frac{1}{y^2}\right) = \frac{2}{y^3}, \qquad 1 < y < \infty,$$

and $f_Y(y) = 0$ otherwise.
(c) $Z = 10(Y-1)$ is increasing in Y, so for $z \geq 0$,

$$\begin{aligned} F_Z(z) &= P\big(10(Y-1) \leq z\big) \\ &= P\big(Y \leq 1 + \tfrac{z}{10}\big) \\ &= 1 - \frac{1}{\big(1+\frac{z}{10}\big)^2} = 1 - \frac{100}{(z+10)^2}, \end{aligned}$$

and $F_Z(z) = 0$ for $z < 0$ (since $Y \geq 1$ gives $Z \geq 0$).

Problem 1.54. For each of the following, determine the value of c that makes $f(x)$ a pdf.
(a) $f(x) = c\sin x$, $0 < x < \pi/2$
(b) $f(x) = ce^{-|x|}$, $-\infty < x < \infty$

Solution. (a) $c = 1$; (b) $c = \frac{1}{2}$. In each case the integrand is already nonnegative for $c > 0$, so condition (a) of Theorem 1.6.5 is automatic and only $\int f = 1$ is at issue.
(a)

$$\int_0^{\pi/2} c\sin x\, dx = c\big[-\cos x\big]_0^{\pi/2} = c(0+1) = c,$$

so $c = 1$.
(b) The integrand is even, so

$$\int_{-\infty}^{\infty} ce^{-|x|}\, dx = 2c\int_0^{\infty} e^{-x}\, dx = 2c,$$

so $c = \frac{1}{2}$.

Problem 1.55. An electronic device has lifetime denoted by T. The device has value $V = 5$ if it fails before time $t = 3$; otherwise, it has value $V = 2T$. Find the cdf of V, if T has pdf

$$f_T(t) = \frac{1}{1.5}\, e^{-t/(1.5)}, \qquad t > 0.$$

Solution.

$$F_V(v) = \begin{cases} 0 & v < 5 \\ 1 - e^{-2} & 5 \leq v < 6 \\ 1 - e^{-v/3} & v \geq 6. \end{cases}$$

V is mixed: it puts an atom at 5 and is otherwise supported on $[6, \infty)$, since $V = 2T \geq 6$ on the event $\{T \geq 3\}$. From f_T, $P(T > t) = e^{-t/1.5}$ for $t > 0$, so

$$P(V = 5) = P(T < 3) = 1 - e^{-3/1.5} = 1 - e^{-2}$$

and nothing is contributed between 5 and 6, giving the middle line. For $v \geq 6$,

$$\begin{aligned} F_V(v) &= P(T < 3) + P\big(3 \leq T \leq \tfrac{v}{2}\big) \\ &= \big(1 - e^{-2}\big) + \big(e^{-2} - e^{-v/3}\big) \\ &= 1 - e^{-v/3}, \end{aligned}$$

using $e^{-(v/2)/1.5} = e^{-v/3}$.

2 Transformations and Expectations

Contents

2.1 Exercises 2.1–2.7 36
2.2 Exercises 2.8–2.14 41
2.3 Exercises 2.15–2.21 45
2.4 Exercises 2.22–2.28 49
2.5 Exercises 2.29–2.35 53
2.6 Exercises 2.36–2.40 58

2.1 Exercises 2.1–2.7

Problem 2.1. In each of the following find the pdf of Y. Show that the pdf integrates to 1.
(a) $Y = X^3$ and $f_X(x) = 42x^5(1-x)$, $0 < x < 1$
(b) $Y = 4X + 3$ and $f_X(x) = 7e^{-7x}$, $0 < x < \infty$
(c) $Y = X^2$ and $f_X(x) = 30x^2(1-x)^2$, $0 < x < 1$
(See Example A.0.2 in Appendix A.)

Solution. Every g here is monotone on the support, so Theorem 2.1.5 applies directly:

$$f_Y(y) = f_X\big(g^{-1}(y)\big)\left|\frac{d}{dy}g^{-1}(y)\right|, \qquad y \in \mathcal{Y}.$$

(a) $g(x) = x^3$ is increasing on $\mathcal{X} = (0,1)$, so $\mathcal{Y} = (0,1)$, $g^{-1}(y) = y^{1/3}$ and $\frac{d}{dy}g^{-1}(y) = \frac{1}{3}y^{-2/3}$:

$$\begin{aligned} f_Y(y) &= 42\, y^{5/3}\big(1 - y^{1/3}\big) \cdot \tfrac{1}{3}y^{-2/3} \\ &= 14\big(y - y^{4/3}\big), \qquad 0 < y < 1. \end{aligned}$$

Then $\int_0^1 14(y - y^{4/3})\,dy = 14\big(\frac{1}{2} - \frac{3}{7}\big) = 14 \cdot \frac{1}{14} = 1$.
(b) $g(x) = 4x + 3$ is increasing on $\mathcal{X} = (0,\infty)$, so $\mathcal{Y} = (3,\infty)$, $g^{-1}(y) = (y-3)/4$ and $\frac{d}{dy}g^{-1}(y) = \frac{1}{4}$:

$$f_Y(y) = \tfrac{7}{4}\, e^{-7(y-3)/4}, \qquad 3 < y < \infty,$$

an exponential shifted to start at 3; $\int_3^\infty \frac{7}{4}e^{-7(y-3)/4}\,dy = \frac{7}{4} \cdot \frac{4}{7} = 1$.
(c) $g(x) = x^2$ is increasing on $\mathcal{X} = (0,1)$ (the negative branch is not in the support, so the two-to-one difficulty of Example 2.1.7 does not arise), so $\mathcal{Y} = (0,1)$, $g^{-1}(y) = \sqrt{y}$ and $\frac{d}{dy}g^{-1}(y) = 1/(2\sqrt{y})$:

$$\begin{aligned} f_Y(y) &= 30\, y\big(1 - \sqrt{y}\big)^2 \cdot \frac{1}{2\sqrt{y}} \\ &= 15\sqrt{y}\,\big(1 - \sqrt{y}\big)^2, \qquad 0 < y < 1. \end{aligned}$$

Substituting $u = \sqrt{y}$, $dy = 2u\,du$,

$$\int_0^1 15\sqrt{y}\,(1-\sqrt{y})^2\,dy = 30\int_0^1 u^2(1-u)^2\,du = 30 \cdot \frac{2!\,2!}{5!} = \frac{30 \cdot 4}{120} = 1,$$

the beta integral of Section 3.3 with $\alpha = \beta = 3$.

Problem 2.2. In each of the following find the pdf of Y.
(a) $Y = X^2$ and $f_X(x) = 1$, $0 < x < 1$
(b) $Y = -\log X$ and X has pdf

$$f_X(x) = \frac{(n+m+1)!}{n!\,m!}\, x^n(1-x)^m, \qquad 0 < x < 1,$$

with m, n positive integers
(c) $Y = e^X$ and X has pdf

$$f_X(x) = \frac{1}{\sigma^2}\, x\, e^{-(x/\sigma)^2/2}, \qquad 0 < x < \infty,$$

with σ^2 a positive constant

Solution. Each g is monotone on the support, so Theorem 2.1.5 gives the answer at once.

(a) $g^{-1}(y) = \sqrt{y}$ on $\mathcal{Y} = (0,1)$, so

$$f_Y(y) = 1 \cdot \frac{1}{2\sqrt{y}} = \frac{1}{2\sqrt{y}}, \qquad 0 < y < 1.$$

(b) $g(x) = -\log x$ is decreasing on $(0,1)$ with $\mathcal{Y} = (0,\infty)$, $g^{-1}(y) = e^{-y}$ and $|\frac{d}{dy} g^{-1}(y)| = e^{-y}$ (as in Example 2.1.4):

$$\begin{aligned} f_Y(y) &= \frac{(n+m+1)!}{n!\,m!}\, e^{-ny}\big(1-e^{-y}\big)^m e^{-y} \\ &= \frac{(n+m+1)!}{n!\,m!}\, e^{-(n+1)y}\big(1-e^{-y}\big)^m, \qquad 0 < y < \infty. \end{aligned}$$

(c) $g(x) = e^x$ is increasing on $(0,\infty)$ with $\mathcal{Y} = (1,\infty)$, $g^{-1}(y) = \log y$ and $\frac{d}{dy} g^{-1}(y) = 1/y$:

$$f_Y(y) = \frac{\log y}{\sigma^2 y}\, \exp\!\left(-\frac{(\log y)^2}{2\sigma^2}\right), \qquad 1 < y < \infty.$$

Problem 2.3. Suppose X has the geometric pmf $f_X(x) = \frac{1}{3}\left(\frac{2}{3}\right)^x$, $x = 0, 1, 2, \ldots$. Determine the probability distribution of $Y = X/(X+1)$. Note that here both X and Y are discrete random variables. To specify the probability distribution of Y, specify its pmf.

Solution.

$$f_Y(y) = \frac{1}{3}\left(\frac{2}{3}\right)^{y/(1-y)}, \qquad y = 0,\ \tfrac{1}{2},\ \tfrac{2}{3},\ \tfrac{3}{4},\ \ldots,\ \tfrac{x}{x+1},\ \ldots$$

Indeed $g(x) = x/(x+1)$ is strictly increasing on $\mathcal{X} = \{0, 1, 2, \ldots\}$, hence one-to-one, and

$$y = \frac{x}{x+1} \iff x = \frac{y}{1-y},$$

so $g^{-1}(y) = y/(1-y)$ is single-valued and $\mathcal{Y} = \{x/(x+1) : x = 0, 1, 2, \ldots\}$, a subset of $[0,1)$ increasing to 1. By the discrete transformation formula following (2.1.2), $f_Y(y) = \sum_{x \in g^{-1}(y)} f_X(x) = f_X\left(\frac{y}{1-y}\right)$, which is the display above.

Problem 2.4. Let λ be a fixed positive constant, and define the function $f(x)$ by $f(x) = \frac{1}{2}\lambda e^{-\lambda x}$ if $x \geq 0$ and $f(x) = \frac{1}{2}\lambda e^{\lambda x}$ if $x < 0$.
(a) Verify that $f(x)$ is a pdf.
(b) If X is a random variable with pdf given by $f(x)$, find $P(X < t)$ for all t. Evaluate all integrals.
(c) Find $P(|X| < t)$ for all t. Evaluate all integrals.

Solution. The two branches are the single formula $f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}$ on $-\infty < x < \infty$, the double exponential pdf.
(a) $f \geq 0$ everywhere since $\lambda > 0$, and by symmetry of $e^{-\lambda|x|}$,

$$\int_{-\infty}^{\infty} f(x)\,dx = 2\int_0^{\infty} \tfrac{1}{2}\lambda e^{-\lambda x}\,dx = \Big[-e^{-\lambda x}\Big]_0^{\infty} = 1,$$

so the two conditions of Theorem 1.6.5 hold and f is a pdf.
(b) For $t < 0$ only the left branch is involved:

$$P(X < t) = \int_{-\infty}^{t} \tfrac{1}{2}\lambda e^{\lambda x}\,dx = \Big[\tfrac{1}{2}e^{\lambda x}\Big]_{-\infty}^{t} = \tfrac{1}{2}e^{\lambda t}.$$

For $t \geq 0$ split at 0, where the display just obtained gives $P(X < 0) = \frac{1}{2}$:

$$\begin{aligned} P(X < t) &= \tfrac{1}{2} + \int_0^t \tfrac{1}{2}\lambda e^{-\lambda x}\, dx \\ &= \tfrac{1}{2} + \tfrac{1}{2}\big(1 - e^{-\lambda t}\big) = 1 - \tfrac{1}{2}e^{-\lambda t}. \end{aligned}$$

(c) $P(|X| < t) = 0$ for $t \leq 0$. For $t > 0$, symmetry gives

$$P(|X| < t) = P(-t < X < t) = 2\int_0^t \tfrac{1}{2}\lambda e^{-\lambda x}\, dx = 1 - e^{-\lambda t}.$$

Problem 2.5. Use Theorem 2.1.8 to find the pdf of Y in Example 2.1.2. Show that the same answer is obtained by differentiating the cdf given in (2.1.6).
For reference, Example 2.1.2 (Uniform transformation) takes X uniform on $(0, 2\pi)$, that is

$$f_X(x) = \begin{cases} 1/(2\pi) & 0 < x < 2\pi \\ 0 & \text{otherwise,} \end{cases}$$

and sets $Y = \sin^2(X)$; equation (2.1.6) reads

$$P(Y \leq y) = 2P(X \leq x_1) + 2P(x_2 \leq X \leq \pi),$$

where x_1 and x_2 are the two solutions to $\sin^2(x) = y$ with $0 < x < \pi$. Theorem 2.1.8 states that if $A_0, A_1, \ldots, A_k$ partition $\mathcal{X}$ with $P(X \in A_0) = 0$, and g_i is monotone on A_i with $g_i = g$ there, with common image $\mathcal{Y} = g_i(A_i)$ and g_i^{-1} continuously differentiable on $\mathcal{Y}$, then

$$f_Y(y) = \sum_{i=1}^{k} f_X\big(g_i^{-1}(y)\big) \left| \frac{d}{dy} g_i^{-1}(y) \right|, \qquad y \in \mathcal{Y}.$$

Solution.

$$f_Y(y) = \frac{1}{\pi\sqrt{y(1-y)}}, \qquad 0 < y < 1,$$

the beta$(\frac{1}{2}, \frac{1}{2})$ (arcsine) pdf.
Method (1), Theorem 2.1.8. The function $g(x) = \sin^2 x$ is monotone on each quarter-period, so take

$$\begin{aligned} A_0 &= \{\pi/2,\ \pi,\ 3\pi/2\}, & P(X \in A_0) &= 0, \\ A_1 &= (0, \pi/2), & g_1^{-1}(y) &= \arcsin\sqrt{y}, \\ A_2 &= (\pi/2, \pi), & g_2^{-1}(y) &= \pi - \arcsin\sqrt{y}, \\ A_3 &= (\pi, 3\pi/2), & g_3^{-1}(y) &= \pi + \arcsin\sqrt{y}, \\ A_4 &= (3\pi/2, 2\pi), & g_4^{-1}(y) &= 2\pi - \arcsin\sqrt{y}. \end{aligned}$$

These four sets and A_0 partition $\mathcal{X} = (0, 2\pi)$, each g_i carries A_i onto the common $\mathcal{Y} = (0, 1)$, so hypothesis (iii) holds, and by the chain rule

$$\frac{d}{dy}\arcsin\sqrt{y} = \frac{1}{\sqrt{1-y}} \cdot \frac{1}{2\sqrt{y}} = \frac{1}{2\sqrt{y(1-y)}},$$

which is continuous on $(0, 1)$, giving (iv). Hence $|\frac{d}{dy} g_i^{-1}(y)| = 1/(2\sqrt{y(1-y)})$ for all four i, and since $f_X \equiv 1/(2\pi)$ on each A_i,

$$f_Y(y) = 4 \cdot \frac{1}{2\pi} \cdot \frac{1}{2\sqrt{y(1-y)}} = \frac{1}{\pi\sqrt{y(1-y)}}, \qquad 0 < y < 1.$$

Method (2), differentiating (2.1.6). The two solutions of $\sin^2 x = y$ in $(0,\pi)$ are $x_1 = \arcsin\sqrt{y}$ and $x_2 = \pi - \arcsin\sqrt{y}$, so by uniformity

$$P(X \le x_1) = \frac{\arcsin\sqrt{y}}{2\pi}, \qquad P(x_2 \le X \le \pi) = \frac{\pi - x_2}{2\pi} = \frac{\arcsin\sqrt{y}}{2\pi},$$

and (2.1.6) becomes

$$F_Y(y) = \frac{2}{\pi}\arcsin\sqrt{y}, \qquad 0 < y < 1.$$

Differentiating with the derivative computed above,

$$f_Y(y) = \frac{2}{\pi}\cdot\frac{1}{2\sqrt{y(1-y)}} = \frac{1}{\pi\sqrt{y(1-y)}},$$

the same answer.

Problem 2.6. In each of the following find the pdf of Y and show that the pdf integrates to 1.
(a) $f_X(x) = \frac{1}{2}e^{-|x|}$, $-\infty < x < \infty$; $Y = |X|^3$
(b) $f_X(x) = \frac{3}{8}(x+1)^2$, $-1 < x < 1$; $Y = 1 - X^2$
(c) $f_X(x) = \frac{3}{8}(x+1)^2$, $-1 < x < 1$; $Y = 1 - X^2$ if $X \le 0$ and $Y = 1 - X$ if $X > 0$

Solution. Each g is two-piece, so Theorem 2.1.8 applies with $A_0 = \{0\}$ and A_1, A_2 the negative and positive parts of $\mathcal{X}$.
(a) Here $g_1(x) = -x^3$ on $A_1 = (-\infty, 0)$ and $g_2(x) = x^3$ on $A_2 = (0,\infty)$, both with common image $\mathcal{Y} = (0,\infty)$; the inverses are $g_1^{-1}(y) = -y^{1/3}$, $g_2^{-1}(y) = y^{1/3}$, each with $|\frac{d}{dy}g_i^{-1}(y)| = \frac{1}{3}y^{-2/3}$, continuous on $(0,\infty)$. Since $f_X(\pm y^{1/3}) = \frac{1}{2}e^{-y^{1/3}}$,

$$f_Y(y) = 2\cdot\tfrac{1}{2}e^{-y^{1/3}}\cdot\tfrac{1}{3}y^{-2/3} = \tfrac{1}{3}y^{-2/3}e^{-y^{1/3}}, \qquad 0 < y < \infty.$$

With $u = y^{1/3}$, $du = \frac{1}{3}y^{-2/3}dy$, we get $\int_0^\infty f_Y(y)\,dy = \int_0^\infty e^{-u}\,du = 1$.
(b) $g(x) = 1 - x^2$ is increasing on $A_1 = (-1, 0)$ and decreasing on $A_2 = (0, 1)$, with common image $\mathcal{Y} = (0,1)$, inverses $g_1^{-1}(y) = -\sqrt{1-y}$ and $g_2^{-1}(y) = \sqrt{1-y}$, and $|\frac{d}{dy}g_i^{-1}(y)| = 1/(2\sqrt{1-y})$. Writing $a = \sqrt{1-y}$ and using $(1-a)^2 + (1+a)^2 = 2 + 2a^2 = 4 - 2y$,

$$\begin{aligned} f_Y(y) &= \frac{1}{2\sqrt{1-y}}\Big[\tfrac{3}{8}(1-a)^2 + \tfrac{3}{8}(1+a)^2\Big] \\ &= \frac{3}{8}\cdot\frac{4-2y}{2\sqrt{1-y}} = \frac{3(2-y)}{8\sqrt{1-y}}, \qquad 0 < y < 1. \end{aligned}$$

Substituting $u = 1 - y$,

$$\int_0^1 \frac{3(2-y)}{8\sqrt{1-y}}\,dy = \frac{3}{8}\int_0^1 \frac{1+u}{\sqrt{u}}\,du = \frac{3}{8}\Big(2 + \tfrac{2}{3}\Big) = 1.$$

(c) Now $g_1(x) = 1 - x^2$ on $A_1 = (-1, 0)$ and $g_2(x) = 1 - x$ on $A_2 = (0, 1)$, each monotone with common image $\mathcal{Y} = (0,1)$; the inverses are $g_1^{-1}(y) = -\sqrt{1-y}$ with $|\frac{d}{dy}g_1^{-1}(y)| = 1/(2\sqrt{1-y})$, and $g_2^{-1}(y) = 1 - y$ with $|\frac{d}{dy}g_2^{-1}(y)| = 1$. Hence

$$\begin{aligned} f_Y(y) &= \tfrac{3}{8}\big(1 - \sqrt{1-y}\big)^2\cdot\frac{1}{2\sqrt{1-y}} + \tfrac{3}{8}(2-y)^2 \\ &= \frac{3\big(1-\sqrt{1-y}\big)^2}{16\sqrt{1-y}} + \frac{3(2-y)^2}{8}, \qquad 0 < y < 1. \end{aligned}$$

For the total mass, the second term gives $\frac{3}{8}\int_0^1 (2-y)^2 dy = \frac{3}{8}\cdot\frac{7}{3} = \frac{7}{8}$, while with $u = \sqrt{1-y}$, $dy = -2u\,du$, the first gives

$$\frac{3}{16}\int_0^1 \frac{(1-u)^2}{u}\,2u\,du = \frac{3}{8}\int_0^1 (1-u)^2\,du = \frac{3}{8}\cdot\frac{1}{3} = \frac{1}{8},$$

and $\frac{7}{8} + \frac{1}{8} = 1$.

Problem 2.7. Let X have pdf $f_X(x) = \frac{2}{9}(x+1)$, $-1 \le x \le 2$.
(a) Find the pdf of $Y = X^2$. Note that Theorem 2.1.8 is not directly applicable in this problem.
(b) Show that Theorem 2.1.8 remains valid if the sets $A_0, A_1, \ldots, A_k$ contain $\mathcal{X}$, and apply the extension to solve part (a) using $A_0 = \emptyset$, $A_1 = (-1, 1)$, and $A_2 = (1, 2)$.

Solution.

$$f_Y(y) = \begin{cases} \dfrac{2}{9\sqrt{y}} & 0 < y < 1, \\ \dfrac{\sqrt{y}+1}{9\sqrt{y}} & 1 < y < 4. \end{cases}$$

(a) Theorem 2.1.8 fails at hypothesis (iii): the sets on which $g(x) = x^2$ is monotone are $(-1, 0)$, $(0, 1)$ and $(1, 2)$, whose images $(0, 1)$, $(0, 1)$ and $(1, 4)$ are not all the same $\mathcal{Y}$. So work from the cdf, as in Example 2.1.7. Here

$$F_X(x) = \int_{-1}^{x} \tfrac{2}{9}(t+1)\,dt = \frac{(x+1)^2}{9}, \qquad -1 \le x \le 2,$$

and $\mathcal{Y} = [0, 4)$, with two cases.
(i) $0 < y < 1$: both roots $\pm\sqrt{y}$ lie in $\mathcal{X}$, so

$$\begin{aligned} F_Y(y) &= P\big(-\sqrt{y} \le X \le \sqrt{y}\big) \\ &= \frac{(\sqrt{y}+1)^2 - (1-\sqrt{y})^2}{9} = \frac{4\sqrt{y}}{9}. \end{aligned}$$

(ii) $1 \le y < 4$: only $+\sqrt{y}$ lies in $\mathcal{X}$, so

$$F_Y(y) = P\big(X \le \sqrt{y}\big) = \frac{(\sqrt{y}+1)^2}{9}.$$

Differentiating each piece gives the displayed f_Y.
(b) Extension. Only containment is needed, not equality: let $A_0, \ldots, A_k$ be disjoint sets with $P(X \notin \bigcup_{i \ge 0} A_i) = 0$ and $P(X \in A_0) = 0$, let $g_i = g$ be monotone on $A_i \cap \mathcal{X}$ with g_i^{-1} continuously differentiable on $\mathcal{Y}_i = g_i(A_i \cap \mathcal{X})$, and drop hypothesis (iii), asking only $\bigcup_i \mathcal{Y}_i = \mathcal{Y}$. For $B \subset \mathcal{Y}$, disjointness and the two null sets give

$$\begin{aligned} P(Y \in B) &= \sum_{i=1}^{k} P\big(X \in A_i \cap \mathcal{X},\ g_i(X) \in B\big) \\ &= \sum_{i=1}^{k} \int_{A_i \cap \mathcal{X} \cap g_i^{-1}(B)} f_X(x)\,dx \\ &= \sum_{i=1}^{k} \int_{B \cap \mathcal{Y}_i} f_X\big(g_i^{-1}(y)\big) \left| \frac{d}{dy} g_i^{-1}(y) \right| dy, \end{aligned}$$

the last step by the change of variable $x = g_i^{-1}(y)$ on each piece, legitimate because g_i is monotone (hence one-to-one) there with continuously differentiable inverse. As B is arbitrary,

$$f_Y(y) = \sum_{i\,:\,y \in \mathcal{Y}_i} f_X\big(g_i^{-1}(y)\big) \left| \frac{d}{dy} g_i^{-1}(y) \right|,$$

the formula of Theorem 2.1.8 read with the convention $f_X(g_i^{-1}(y)) = 0$ whenever $g_i^{-1}(y) \notin \mathcal{X}$.
Now take $A_0 = \emptyset$, $A_1 = (-1, 1)$, $A_2 = (1, 2)$: the omitted points $-1, 1, 2$ carry no probability, so A_0 may indeed be empty, and $\mathcal{Y}_1 = [0, 1)$, $\mathcal{Y}_2 = (1, 4)$ now need not agree. Since x^2 is two-to-one on

A_1, split it at the null point 0 into $A_{1a} = (-1, 0)$ and $A_{1b} = (0, 1)$ (formula (2.1.11) of Example 2.1.7 applied inside A_1), with

$$g_{1a}^{-1}(y) = -\sqrt{y}, \quad g_{1b}^{-1}(y) = \sqrt{y}, \quad g_2^{-1}(y) = \sqrt{y},$$

and $|\frac{d}{dy} g_i^{-1}(y)| = 1/(2\sqrt{y})$ in all three cases. For $0 < y < 1$ only the two A_1 terms are present:

$$f_Y(y) = \frac{1}{2\sqrt{y}} \left[\tfrac{2}{9}(\sqrt{y} + 1) + \tfrac{2}{9}(-\sqrt{y} + 1) \right] = \frac{1}{2\sqrt{y}} \cdot \frac{4}{9} = \frac{2}{9\sqrt{y}}.$$

For $1 < y < 4$ only the A_2 term is present:

$$f_Y(y) = \frac{1}{2\sqrt{y}} \cdot \tfrac{2}{9}(\sqrt{y} + 1) = \frac{\sqrt{y} + 1}{9\sqrt{y}},$$

agreeing with part (a).

2.2 Exercises 2.8–2.14

Problem 2.8. In each of the following show that the given function is a cdf and find $F_X^{-1}(y)$.
(a)

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-x} & \text{if } x \geq 0 \end{cases}$$

(b)

$$F_X(x) = \begin{cases} e^x/2 & \text{if } x < 0 \\ 1/2 & \text{if } 0 \leq x < 1 \\ 1 - \left(e^{1-x}/2\right) & \text{if } 1 \leq x \end{cases}$$

(c)

$$F_X(x) = \begin{cases} e^x/4 & \text{if } x < 0 \\ 1 - \left(e^{-x}/4\right) & \text{if } x \geq 0 \end{cases}$$

Note that, in part (c), $F_X(x)$ is discontinuous but (2.1.13) is still the appropriate definition of $F_X^{-1}(y)$.

Solution. Each function satisfies the three conditions of Theorem 1.5.3 (limits 0 and 1, nondecreasing, right-continuous), and (2.1.13) reads $F_X^{-1}(y) = \inf\{x : F_X(x) \geq y\}$.
(a) F_X is 0 on $(-\infty, 0)$ and rises continuously from 0 to 1 on $[0, \infty)$ with $F_X'(x) = e^{-x} > 0$ there; it is continuous at 0 since $1 - e^0 = 0$. So it is a continuous, strictly increasing cdf on its range, and (2.1.13) reduces to (2.1.12): solving $1 - e^{-x} = y$,

$$F_X^{-1}(y) = -\log(1 - y), \qquad 0 < y < 1.$$

(b) On $(-\infty, 0)$, $e^x/2$ increases from 0 to 1/2; the value at $x = 0$ is 1/2, so F_X is continuous there; on $[1, \infty)$, $1 - e^{1-x}/2$ starts at $1 - e^0/2 = 1/2$ and increases to 1. Hence F_X is continuous and nondecreasing with the correct limits, so it is a cdf — but it is flat at height 1/2 on $[0, 1]$, which is exactly where (2.1.13) is needed. For $0 < y < 1/2$, $e^x/2 = y$ gives $x = \log(2y)$; for $y = 1/2$ the set $\{x : F_X(x) \geq 1/2\} = [0, \infty)$ has infimum $0 = \log(2 \cdot \frac{1}{2})$; for $1/2 < y < 1$, $1 - e^{1-x}/2 = y$ gives $x = 1 - \log(2(1 - y))$. Thus

$$F_X^{-1}(y) = \begin{cases} \log(2y) & 0 < y \leq 1/2 \\ 1 - \log\big(2(1 - y)\big) & 1/2 < y < 1. \end{cases}$$

(c) Each piece is increasing, the limits are 0 and 1, and at $x = 0$ the left limit is 1/4 while $F_X(0) = 3/4$: F_X is right-continuous with a jump of size 1/2 at 0, so it is a cdf of a variable with an atom $P(X = 0) = 1/2$. Now apply (2.1.13) directly. For $0 < y \leq 1/4$ the equation $e^x/4 = y$ is solvable with

$x = \log(4y) \le 0$. For $1/4 < y \le 3/4$ every $x < 0$ has $F_X(x) < 1/4 < y$ while $F_X(0) = 3/4 \ge y$, so the infimum is 0. For $3/4 < y < 1$, $1 - e^{-x}/4 = y$ gives $x = -\log(4(1-y))$. Hence

$$F_X^{-1}(y) = \begin{cases} \log(4y) & 0 < y \le 1/4 \\ 0 & 1/4 < y \le 3/4 \\ -\log\big(4(1-y)\big) & 3/4 < y < 1. \end{cases}$$

Problem 2.9. If the random variable X has pdf

$$f(x) = \begin{cases} \dfrac{x-1}{2} & 1 < x < 3 \\ 0 & \text{otherwise,} \end{cases}$$

find a monotone function $u(x)$ such that the random variable $Y = u(X)$ has a uniform$(0,1)$ distribution.

Solution. Take $u(x) = F_X(x) = (x-1)^2/4$ on $1 < x < 3$, since

$$F_X(x) = \int_1^x \frac{t-1}{2}\, dt = \frac{(x-1)^2}{4}, \qquad 1 < x < 3,$$

which is continuous and strictly increasing from 0 to 1; the probability integral transform (Theorem 2.1.10) applies, its hypothesis of a continuous cdf being met, and gives $Y = F_X(X) \sim$ uniform$(0,1)$.

Problem 2.10. In Theorem 2.1.10 the probability integral transform was proved, relating the uniform cdf to any continuous cdf. In this exercise we investigate the relationship between discrete random variables and uniform random variables. Let X be a discrete random variable with cdf $F_X(x)$ and define the random variable Y as $Y = F_X(X)$.
(a) Prove that Y is stochastically greater than a uniform$(0,1)$; that is, if $U \sim$ uniform$(0,1)$, then

$$P(Y > y) \ge P(U > y) = 1 - y, \quad \text{for all } y,\ 0 < y < 1,$$
$$P(Y > y) > P(U > y) = 1 - y, \quad \text{for some } y,\ 0 < y < 1.$$

(Recall that *stochastically greater* was defined in Exercise 1.49.)
(b) Equivalently, show that the cdf of Y satisfies $F_Y(y) \le y$ for all $0 < y < 1$ and $F_Y(y) < y$ for some $0 < y < 1$. (Hint: Let x_0 be a jump point of F_X, and define $y_0 = F_X(x_0)$. Show that $P(Y \le y_0) = y_0$. Now establish the inequality by considering $y = y_0 + \epsilon$. Pictures of the cdfs will help.)

Solution. (a) is (b) complemented: $P(Y > y) = 1 - F_Y(y)$ and $P(U > y) = 1 - y$, so $F_Y(y) \le y$ is $P(Y > y) \ge P(U > y)$ and $F_Y(y) < y$ is the strict version. Prove (b).
Let $\mathcal{X}$ be the countable support of X. Along $\mathcal{X}$ the cdf is *strictly* increasing: for $x < x'$ both in $\mathcal{X}$,

$$F_X(x') \ge F_X(x) + P(X = x') > F_X(x),$$

so F_X is one-to-one on $\mathcal{X}$ and $F_X(x) \le F_X(x')$ forces $x \le x'$.
Fix $0 < y < 1$ and put $S_y = \{x \in \mathcal{X} : F_X(x) \le y\}$, so that $F_Y(y) = P(X \in S_y)$. If $S_y = \emptyset$ then $F_Y(y) = 0 \le y$. Otherwise let $M = \sup_{x \in S_y} F_X(x) \le y$ and choose $x_n \in S_y$ with $F_X(x_n) \uparrow M$. Every $x \in S_y$ has $F_X(x) \le F_X(x_n)$ for some n, hence $x \le x_n$; so $\{X \in S_y\} = \bigcup_n \{X \le x_n\}$, an increasing union, and

$$F_Y(y) = \lim_{n \to \infty} F_X(x_n) = M \le y.$$

For the strict inequality take any $x_0 \in \mathcal{X}$ — a jump point of F_X — and set $y_0 = F_X(x_0)$ and $j = P(X = x_0) > 0$. By the strict increase, $F_X(x) \le y_0$ holds for exactly the support points $x \le x_0$, so

$$P(Y \le y_0) = P(X \le x_0) = F_X(x_0) = y_0,$$

the hint's equality, while $P(Y = y_0) = P(X = x_0) = j$ by injectivity on $\mathcal{X}$. Hence $F_Y(y_0^-) = y_0 - j$. Take $y = y_0 - j/2$, which lies in $(0, 1)$ since $y_0 \geq j$ and $j > 0$; then

$$F_Y(y) \leq F_Y(y_0^-) = y_0 - j < y_0 - \tfrac{j}{2} = y.$$

(The hint's $y = y_0 + \epsilon$ needs a next support point above x_0, which a discrete X need not have; taking y just below y_0 avoids that.)

Problem 2.11. Let X have the standard normal pdf, $f_X(x) = (1/\sqrt{2\pi})e^{-x^2/2}$.
(a) Find $\mathrm{E}X^2$ directly, and then by using the pdf of $Y = X^2$ from Example 2.1.7 and calculating $\mathrm{E}Y$.
(b) Find the pdf of $Y = |X|$, and find its mean and variance.

Solution. (a) $\mathrm{E}X^2 = 1$. Directly, integrate by parts with $u = x$ and $dv = xe^{-x^2/2}\,dx$:

$$\begin{aligned}\mathrm{E}X^2 &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x \cdot xe^{-x^2/2}\,dx \\ &= \frac{1}{\sqrt{2\pi}}\Big[-xe^{-x^2/2}\Big]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\,dx \\ &= 0 + 1 = 1.\end{aligned}$$

Via the square transformation (2.1.11) of Example 2.1.7, worked out for the normal in Example 2.1.9,

$$f_Y(y) = \frac{1}{2\sqrt{y}}\big(f_X(\sqrt{y}) + f_X(-\sqrt{y})\big) = \frac{1}{\sqrt{2\pi}}\,\frac{1}{\sqrt{y}}\,e^{-y/2}, \qquad y > 0,$$

the chi squared pdf with 1 degree of freedom. Substituting $y = 2t$,

$$\begin{aligned}\mathrm{E}Y &= \frac{1}{\sqrt{2\pi}}\int_0^{\infty} y^{1/2}e^{-y/2}\,dy = \frac{2^{3/2}}{\sqrt{2\pi}}\int_0^{\infty} t^{1/2}e^{-t}\,dt \\ &= \frac{2\sqrt{2}}{\sqrt{2\pi}}\,\Gamma(3/2) = \frac{2\sqrt{2}}{\sqrt{2\pi}}\cdot\frac{\sqrt{\pi}}{2} = 1,\end{aligned}$$

agreeing with the direct computation.
(b) $g(x) = |x|$ is monotone on $A_1 = (-\infty, 0)$ and $A_2 = (0, \infty)$, both mapped onto $\mathcal{Y} = (0, \infty)$ with $|dg_i^{-1}/dy| = 1$, and $P(X \in A_0) = P(X = 0) = 0$; Theorem 2.1.8 (its hypotheses (i)–(iv) hold, the inverses being $y \mapsto -y$ and $y \mapsto y$) gives the half-normal pdf

$$f_Y(y) = f_X(y) + f_X(-y) = \sqrt{\frac{2}{\pi}}\,e^{-y^2/2}, \qquad 0 < y < \infty.$$

Then, substituting $u = y^2/2$,

$$\mathrm{E}Y = \sqrt{\frac{2}{\pi}}\int_0^{\infty} y\,e^{-y^2/2}\,dy = \sqrt{\frac{2}{\pi}} \approx 0.7979,$$

while $\mathrm{E}Y^2 = \mathrm{E}X^2 = 1$ because $|X|^2 = X^2$. Hence

$$\operatorname{Var} Y = 1 - \frac{2}{\pi} \approx 0.3634.$$

Problem 2.12. A random right triangle can be constructed in the following manner. Let X be a random angle whose distribution is uniform on $(0, \pi/2)$. For each X, construct a triangle as pictured below. Here, Y = height of the random triangle. For a fixed constant d, find the distribution of Y

and $\mathrm{E}Y$.
(The figure shows a right triangle with vertices at the origin, at $(d, 0)$, and at (d, y): the horizontal leg runs along the axis from the origin to $(d, 0)$ and has fixed length d; the vertical leg runs from $(d, 0)$ up to (d, y) and has length Y, so the right angle sits at $(d, 0)$; the hypotenuse runs from the origin to (d, y) and makes the angle X with the horizontal leg at the origin.)

Solution. $Y = d \tan X$, a half-Cauchy variable with scale d, and $\mathrm{E}Y = \infty$ (the mean does not exist). Assume $d > 0$. Since tan is strictly increasing on $(0, \pi/2)$ onto $(0, \infty)$, for $y > 0$

$$F_Y(y) = P(d \tan X \le y) = P\big(X \le \arctan(y/d)\big) = \frac{\arctan(y/d)}{\pi/2},$$

using the uniform$(0, \pi/2)$ cdf. Differentiating,

$$f_Y(y) = \frac{2}{\pi} \cdot \frac{1/d}{1 + (y/d)^2} = \frac{2d}{\pi\,(d^2 + y^2)}, \qquad 0 < y < \infty,$$

which integrates to 1 (Check!). For the mean, the integrand behaves like $2d/(\pi y)$ at infinity:

$$\mathrm{E}Y = \int_0^\infty \frac{2d\,y}{\pi(d^2 + y^2)}\,dy = \frac{d}{\pi}\,\log\big(d^2 + y^2\big)\Big|_0^\infty = \infty.$$

Problem 2.13. Consider a sequence of independent coin flips, each of which has probability p of being heads. Define a random variable X as the length of the run (of either heads or tails) started by the first trial. (For example, $X = 3$ if either TTTH or HHHT is observed.) Find the distribution of X, and find $\mathrm{E}X$.

Solution. With $q = 1 - p$ and $0 < p < 1$,

$$P(X = x) = p^x q + q^x p, \qquad x = 1, 2, \ldots, \qquad \mathrm{E}X = \frac{p}{q} + \frac{q}{p} = \frac{p^2 + q^2}{pq}.$$

The event $\{X = x\}$ is the disjoint union of $\{x$ heads then a tail$\}$ and $\{x$ tails then a head$\}$, whose probabilities are $p^x q$ and $q^x p$ by independence, giving the pmf above. It sums to 1:

$$\sum_{x=1}^{\infty} \big(p^x q + q^x p\big) = q \cdot \frac{p}{1 - p} + p \cdot \frac{q}{1 - q} = p + q = 1.$$

For the mean, use $\sum_{x \ge 1} x r^x = r/(1 - r)^2$ for $|r| < 1$ twice:

$$\begin{aligned} \mathrm{E}X &= q \sum_{x=1}^{\infty} x\, p^x + p \sum_{x=1}^{\infty} x\, q^x \\ &= q \cdot \frac{p}{q^2} + p \cdot \frac{q}{p^2} = \frac{p}{q} + \frac{q}{p}. \end{aligned}$$

Problem 2.14. (a) Let X be a continuous, nonnegative random variable $[f(x) = 0$ for $x < 0]$. Show that

$$\mathrm{E}X = \int_0^\infty \big[1 - F_X(x)\big]\,dx,$$

where $F_X(x)$ is the cdf of X.

(b) Let X be a discrete random variable whose range is the nonnegative integers. Show that

$$\mathrm{E}X = \sum_{k=0}^{\infty} \big(1 - F_X(k)\big),$$

where $F_X(k) = P(X \le k)$. Compare this with part (a).

Solution. Both identities are the same interchange of order of integration on the region $\{(x,t) : 0 \le x < t\}$, legitimate because the integrand is nonnegative (Fubini–Tonelli).
(a) Since $1 - F_X(x) = P(X > x) = \int_x^\infty f(t)\,dt$ for $x \ge 0$,

$$\begin{aligned}\int_0^\infty \big[1 - F_X(x)\big]\,dx &= \int_0^\infty \int_x^\infty f(t)\,dt\,dx \\ &= \int_0^\infty \int_0^t dx\, f(t)\,dt \\ &= \int_0^\infty t\, f(t)\,dt = \mathrm{E}X,\end{aligned}$$

the last equality because f vanishes on $(-\infty, 0)$.
(b) Here $1 - F_X(k) = P(X > k) = \sum_{j=k+1}^\infty P(X = j)$, so summing over $k \ge 0$ and counting each j once for every $k \in \{0, 1, \ldots, j-1\}$,

$$\begin{aligned}\sum_{k=0}^{\infty} \big(1 - F_X(k)\big) &= \sum_{k=0}^{\infty} \sum_{j=k+1}^{\infty} P(X = j) \\ &= \sum_{j=1}^{\infty} \sum_{k=0}^{j-1} P(X = j) \\ &= \sum_{j=1}^{\infty} j\, P(X = j) = \mathrm{E}X.\end{aligned}$$

Part (a) is the integral version of exactly this count: the area under the horizontal slices of the survival function $P(X > x)$ is the expectation, with dx replacing the unit steps $k \to k+1$.

2.3 Exercises 2.15–2.21

Problem 2.15. Betteley (1977) provides an interesting addition law for expectations. Let X and Y be any two random variables and define

$$X \wedge Y = \min(X, Y) \qquad \text{and} \qquad X \vee Y = \max(X, Y).$$

Analogous to the probability law $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, show that

$$\mathrm{E}(X \vee Y) = \mathrm{E}\,X + \mathrm{E}\,Y - \mathrm{E}(X \wedge Y).$$

(Hint: Establish that $X + Y = (X \vee Y) + (X \wedge Y)$.)

Solution. The whole exercise is the pointwise identity

$$X + Y = (X \vee Y) + (X \wedge Y),$$

which holds at every sample point s because $\{X(s), Y(s)\}$ and $\{(X \wedge Y)(s), (X \vee Y)(s)\}$ are the same pair of numbers: if $X(s) \le Y(s)$ the right side is $Y(s) + X(s)$, and if $X(s) > Y(s)$ it is $X(s) + Y(s)$. Both $X \vee Y$ and $X \wedge Y$ have expectations whenever X and Y do, since

$$|X \vee Y| \le |X| + |Y| \qquad \text{and} \qquad |X \wedge Y| \le |X| + |Y|.$$

So apply E and use linearity – Theorem 2.2.5a, which Section 4.1 notes persists with X replaced by (X, Y), here with $g_1(x, y) = \max(x, y)$, $g_2(x, y) = \min(x, y)$:

$$\begin{aligned} \mathrm{E}\, X + \mathrm{E}\, Y &= \mathrm{E}\big[(X \vee Y) + (X \wedge Y)\big] \\ &= \mathrm{E}(X \vee Y) + \mathrm{E}(X \wedge Y), \end{aligned}$$

and move $\mathrm{E}(X \wedge Y)$ to the other side.

Problem 2.16. Use the result of Exercise 2.14 to find the mean duration of certain telephone calls, where we assume that the duration, T, of a particular call can be described probabilistically by $P(T > t) = ae^{-\lambda t} + (1 - a)e^{-\mu t}$, where a, λ, and μ are constants, $0 < a < 1$, $\lambda > 0$, $\mu > 0$.

Solution.

$$\mathrm{E}\, T = \frac{a}{\lambda} + \frac{1 - a}{\mu}.$$

T is continuous and nonnegative, so Exercise 2.14(a) applies, and $1 - F_T(t) = P(T > t)$:

$$\begin{aligned} \mathrm{E}\, T &= \int_0^\infty \big[1 - F_T(t)\big]\, dt \\ &= \int_0^\infty \Big(ae^{-\lambda t} + (1 - a)e^{-\mu t}\Big)\, dt \\ &= a\,\frac{1}{\lambda} + (1 - a)\,\frac{1}{\mu}, \end{aligned}$$

the mixture of the two exponential means of Example 2.2.2.

Problem 2.17. A *median* of a distribution is a value m such that $P(X \le m) \ge \frac{1}{2}$ and $P(X \ge m) \ge \frac{1}{2}$. (If X is continuous, m satisfies $\int_{-\infty}^{m} f(x)\, dx = \int_m^\infty f(x)\, dx = \frac{1}{2}$.) Find the median of the following distributions.
(a) $f(x) = 3x^2, \quad 0 < x < 1$
(b) $f(x) = \dfrac{1}{\pi(1 + x^2)}, \quad -\infty < x < \infty$

Solution. (a) $m = 2^{-1/3} \approx 0.7937$. The cdf is $F(x) = \int_0^x 3t^2\, dt = x^3$ on $(0, 1)$, so

$$\tfrac{1}{2} = F(m) = m^3 \qquad \Longrightarrow \qquad m = \big(\tfrac{1}{2}\big)^{1/3}.$$

(b) $m = 0$. The Cauchy density of Example 2.2.4 is symmetric about 0, $f(-x) = f(x)$, so

$$\int_{-\infty}^{0} f(x)\, dx = \int_0^\infty f(x)\, dx = \tfrac{1}{2}.$$

Explicitly, $F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x$, and $F(0) = \frac{1}{2}$.

Problem 2.18. Show that if X is a continuous random variable, then

$$\min_a \mathrm{E}\,|X - a| = \mathrm{E}\,|X - m|,$$

where m is the median of X (see Exercise 2.17).

Solution. For $a > m$ the integrand difference $|x - a| - |x - m|$ is at least $a - m$ on $\{x \le m\}$ and at least $-(a - m)$ off it, and the median splits the mass evenly, so the two halves cancel. (Assume $\mathrm{E}|X| < \infty$; otherwise $\mathrm{E}|X - a| = \infty$ for every a.) Leaf by leaf:

(i) $x \le m$: $|x-a| - |x-m| = (a-x) - (m-x) = a - m$.
(ii) $m < x < a$: $|x-a| - |x-m| = (a-x) - (x-m) = a + m - 2x > -(a-m)$.
(iii) $x \ge a$: $|x-a| - |x-m| = (x-a) - (x-m) = -(a-m)$.
Comparing with the step function $(a-m)\big[I(x \le m) - I(x > m)\big]$, Theorem 2.2.5c gives

$$\begin{aligned} \mathrm{E}|X-a| - \mathrm{E}|X-m| &\ge (a-m)P(X \le m) - (a-m)P(X > m) \\ &= (a-m)\big[\tfrac{1}{2} - \tfrac{1}{2}\big] \;=\; 0, \end{aligned}$$

using $P(X \le m) = P(X > m) = \frac{1}{2}$, the median's defining property for continuous X (Exercise 2.17). The case $a < m$ is the mirror image, with $\{x \ge m\}$ and $\{x < m\}$ and the constant $m - a$ (Check!). So $\mathrm{E}|X-a| \ge \mathrm{E}|X-m|$ for every a, with equality at $a = m$.

Problem 2.19. Prove that

$$\frac{d}{da}\,\mathrm{E}(X-a)^2 = 0 \iff \mathrm{E}\,X = a$$

by differentiating the integral. Verify, using calculus, that $a = \mathrm{E}\,X$ is indeed a minimum. List the assumptions about F_X and f_X that are needed.

Solution. Differentiating under the integral sign,

$$\begin{aligned} \frac{d}{da}\,\mathrm{E}(X-a)^2 &= \frac{d}{da}\int_{-\infty}^{\infty} (x-a)^2 f_X(x)\,dx \\ &= \int_{-\infty}^{\infty} \frac{\partial}{\partial a}(x-a)^2 f_X(x)\,dx \\ &= -2\int_{-\infty}^{\infty} (x-a) f_X(x)\,dx \\ &= -2\,(\mathrm{E}\,X - a), \end{aligned}$$

which vanishes if and only if $a = \mathrm{E}\,X$. The same interchange applied once more gives

$$\frac{d^2}{da^2}\,\mathrm{E}(X-a)^2 = 2\int_{-\infty}^{\infty} f_X(x)\,dx = 2 > 0$$

for every a, so $\mathrm{E}(X-a)^2$ is strictly convex in a and its unique stationary point $a = \mathrm{E}\,X$ is the global minimum – the calculus proof of (2.2.3) promised in Example 2.2.6.
Assumptions.
(i) F_X is absolutely continuous with $F_X' = f_X$, so that $\mathrm{E}\,g(X) = \int g(x) f_X(x)\,dx$ for the g used above.
(ii) $\mathrm{E}\,X^2 < \infty$, so that $\mathrm{E}(X-a)^2$ is finite for all a (and then $\mathrm{E}|X| < \infty$ too).
(iii) The hypotheses of Corollary 2.4.4 for the interchange in the second line: for a ranging in a neighborhood $(a_0 - \delta, a_0 + \delta)$,

$$\left|\frac{\partial}{\partial a}(x-a)^2 f_X(x)\right| \le 2\big(|x| + |a_0| + \delta\big) f_X(x),$$

and the dominating function on the right is integrable by (ii); likewise $2f_X$ dominates itself for the second derivative. No continuity or boundedness of f_X is required.

Problem 2.20. A couple decides to continue to have children until a daughter is born. What is the expected number of children of this couple? (Hint: See Example 1.5.4.)

Solution. Two children, on the assumption that births are independent with $P(\text{daughter}) = p = \frac{1}{2}$.
Let X be the number of children. "Keep going until the first success" is exactly the setup of Example

1.5.4 with a daughter playing the role of a head, so X has the geometric pmf (1.5.2) and cdf

$$P(X = x) = (1-p)^{x-1}p, \qquad F_X(x) = 1 - (1-p)^x, \quad x = 1, 2, \ldots,$$

with $F_X(0) = 0$. Since X takes values in the nonnegative integers, Exercise 2.14(b) turns the mean into a geometric series with no algebra at all:

$$\begin{aligned} \mathrm{E}\, X &= \sum_{k=0}^{\infty} \bigl[1 - F_X(k)\bigr] \\ &= \sum_{k=0}^{\infty} (1-p)^k \\ &= \frac{1}{p}, \end{aligned}$$

the $k = 0$ term being $1 - F_X(0) = 1 = (1-p)^0$. With $p = \frac{1}{2}$, $\mathrm{E}\, X = 2$.

Problem 2.21. Prove the "two-way" rule for expectations, equation (2.2.5), which says $\mathrm{E}\, g(X) = \mathrm{E}\, Y$, where $Y = g(X)$. Assume that $g(x)$ is a monotone function.

Solution. Change of variable $y = g(x)$ in $\mathrm{E}\, Y = \int_{\mathcal{Y}} y f_Y(y)\, dy$, with f_Y supplied by Theorem 2.1.5.
Continuous case. Let $\mathcal{X} = (c, d)$ be the support of X (possibly infinite) and $\mathcal{Y} = g(\mathcal{X})$. Monotonicity makes g^{-1} single-valued; assume as Theorem 2.1.5 requires that f_X is continuous on $\mathcal{X}$ and g^{-1} continuously differentiable on $\mathcal{Y}$. Then

$$f_Y(y) = f_X\bigl(g^{-1}(y)\bigr) \left| \frac{d}{dy} g^{-1}(y) \right|, \qquad y \in \mathcal{Y}.$$

If g is increasing, $\mathcal{Y} = (g(c), g(d))$ and $\frac{d}{dy} g^{-1}(y) > 0$; substituting $x = g^{-1}(y)$, so that $dx = \frac{d}{dy} g^{-1}(y)\, dy$ and the limits map $g(c) \mapsto c$, $g(d) \mapsto d$,

$$\begin{aligned} \mathrm{E}\, Y &= \int_{g(c)}^{g(d)} y\, f_X\bigl(g^{-1}(y)\bigr)\, \frac{d}{dy} g^{-1}(y)\, dy \\ &= \int_c^d g(x)\, f_X(x)\, dx \;=\; \mathrm{E}\, g(X). \end{aligned}$$

If g is decreasing, $\mathcal{Y} = (g(d), g(c))$ and the absolute value contributes $-\frac{d}{dy} g^{-1}(y)$, while the substitution now sends $g(d) \mapsto d$, $g(c) \mapsto c$; the two sign reversals cancel:

$$\begin{aligned} \mathrm{E}\, Y &= \int_{g(d)}^{g(c)} y\, f_X\bigl(g^{-1}(y)\bigr) \left(-\frac{d}{dy} g^{-1}(y) \right) dy \\ &= -\int_d^c g(x)\, f_X(x)\, dx \\ &= \int_c^d g(x)\, f_X(x)\, dx \;=\; \mathrm{E}\, g(X). \end{aligned}$$

Discrete case. Monotone means one-to-one, so the Section 2.1 formula $f_Y(y) = \sum_{x \in g^{-1}(y)} f_X(x)$ collapses to a single term,

$$f_Y(y) = P\bigl(X = g^{-1}(y)\bigr) = f_X\bigl(g^{-1}(y)\bigr), \qquad y \in \mathcal{Y},$$

and $x \mapsto g(x)$ is a bijection of $\mathcal{X}$ onto $\mathcal{Y}$. Reindexing the sum by $y = g(x)$,

$$\mathrm{E}\, Y = \sum_{y \in \mathcal{Y}} y\, f_Y(y) = \sum_{x \in \mathcal{X}} g(x)\, f_X(x) = \mathrm{E}\, g(X),$$

the reindexing being unconditional since $\mathrm{E}|g(X)| < \infty$ is exactly Definition 2.2.1's requirement for either side to exist.

2.4 Exercises 2.22–2.28

Problem 2.22. Let X have the pdf

$$f(x) = \frac{4}{\beta^3\sqrt{\pi}}\, x^2 e^{-x^2/\beta^2}, \qquad 0 < x < \infty, \quad \beta > 0.$$

(a) Verify that $f(x)$ is a pdf.
(b) Find $\mathrm{E}\, X$ and $\operatorname{Var} X$.

Solution. Everything follows from the single gamma evaluation

$$\int_0^\infty u^n e^{-u^2}\, du = \tfrac{1}{2}\,\Gamma\!\left(\tfrac{n+1}{2}\right), \qquad n > -1,$$

(substitute $t = u^2$), which gives $\sqrt{\pi}/4$, $1/2$ and $3\sqrt{\pi}/8$ at $n = 2, 3, 4$.
(a) $f \geq 0$ on $(0, \infty)$, and with $x = \beta u$,

$$\begin{aligned}\int_0^\infty f(x)\, dx &= \frac{4}{\beta^3\sqrt{\pi}} \int_0^\infty \beta^3 u^2 e^{-u^2}\, du \\ &= \frac{4}{\sqrt{\pi}} \cdot \frac{\sqrt{\pi}}{4} = 1.\end{aligned}$$

(b) The same substitution turns each moment into the displayed integral:

$$\begin{aligned}\mathrm{E}\, X &= \frac{4}{\beta^3\sqrt{\pi}}\, \beta^4 \int_0^\infty u^3 e^{-u^2}\, du = \frac{4\beta}{\sqrt{\pi}} \cdot \frac{1}{2} = \frac{2\beta}{\sqrt{\pi}}, \\ \mathrm{E}\, X^2 &= \frac{4}{\beta^3\sqrt{\pi}}\, \beta^5 \int_0^\infty u^4 e^{-u^2}\, du = \frac{4\beta^2}{\sqrt{\pi}} \cdot \frac{3\sqrt{\pi}}{8} = \frac{3\beta^2}{2}.\end{aligned}$$

Hence, by Definition 2.3.2 and (2.3.1),

$$\operatorname{Var} X = \frac{3\beta^2}{2} - \frac{4\beta^2}{\pi} = \beta^2\, \frac{3\pi - 8}{2\pi}.$$

Problem 2.23. Let X have the pdf

$$f(x) = \tfrac{1}{2}(1 + x), \qquad -1 < x < 1.$$

(a) Find the pdf of $Y = X^2$.
(b) Find $\mathrm{E}\, Y$ and $\operatorname{Var} Y$.

Solution. (a) $f_Y(y) = \dfrac{1}{2\sqrt{y}}$, $0 < y < 1$: the square transformation formula (2.1.11) of Example 2.1.7 gives, for $0 < y < 1$,

$$\begin{aligned}f_Y(y) &= \frac{1}{2\sqrt{y}}\Big(f_X(\sqrt{y}) + f_X(-\sqrt{y})\Big) \\ &= \frac{1}{2\sqrt{y}}\Big(\tfrac{1}{2}(1 + \sqrt{y}) + \tfrac{1}{2}(1 - \sqrt{y})\Big) = \frac{1}{2\sqrt{y}}.\end{aligned}$$

(b) Directly from f_Y,

$$\begin{aligned}\mathrm{E}\, Y &= \frac{1}{2}\int_0^1 y^{1/2}\, dy = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}, \\ \mathrm{E}\, Y^2 &= \frac{1}{2}\int_0^1 y^{3/2}\, dy = \frac{1}{2} \cdot \frac{2}{5} = \frac{1}{5},\end{aligned}$$

so $\operatorname{Var} Y = \dfrac{1}{5} - \dfrac{1}{9} = \dfrac{4}{45}$.

Problem 2.24. Compute $\mathrm{E}\, X$ and $\operatorname{Var} X$ for each of the following probability distributions.
(a) $f_X(x) = ax^{a-1}$, $0 < x < 1$, $a > 0$.
(b) $f_X(x) = \dfrac{1}{n}$, $x = 1, 2, \ldots, n$, $n > 0$ an integer.
(c) $f_X(x) = \frac{3}{2}(x-1)^2$, $0 < x < 2$.

Solution. (a) $\mathrm{E}\, X = \dfrac{a}{a+1}$ and $\operatorname{Var} X = \dfrac{a}{(a+2)(a+1)^2}$. Indeed $\mathrm{E}\, X^k = a\int_0^1 x^{a+k-1}dx = \dfrac{a}{a+k}$, so

$$\begin{aligned}\operatorname{Var} X &= \frac{a}{a+2} - \frac{a^2}{(a+1)^2} = \frac{a\big[(a+1)^2 - a(a+2)\big]}{(a+2)(a+1)^2} \\ &= \frac{a}{(a+2)(a+1)^2}.\end{aligned}$$

(b) $\mathrm{E}\, X = \dfrac{n+1}{2}$ and $\operatorname{Var} X = \dfrac{n^2-1}{12}$, from $\sum_{x=1}^n x = \frac{n(n+1)}{2}$ and $\sum_{x=1}^n x^2 = \frac{n(n+1)(2n+1)}{6}$:

$$\begin{aligned}\operatorname{Var} X &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\ &= \frac{(n+1)\big[2(2n+1) - 3(n+1)\big]}{12} = \frac{(n+1)(n-1)}{12}.\end{aligned}$$

(c) $\mathrm{E}\, X = 1$ and $\operatorname{Var} X = \dfrac{3}{5}$. The pdf is symmetric about 1, so $\mathrm{E}\, X = 1$ by Exercise 2.26(c), and

$$\operatorname{Var} X = \mathrm{E}(X-1)^2 = \frac{3}{2}\int_0^2 (x-1)^4\, dx = \frac{3}{2}\cdot\frac{2}{5} = \frac{3}{5}.$$

Problem 2.25. Suppose the pdf $f_X(x)$ of a random variable X is an even function. ($f_X(x)$ is an even function if $f_X(x) = f_X(-x)$ for every x.) Show that
(a) X and $-X$ are identically distributed.
(b) $M_X(t)$ is symmetric about 0.

Solution. (a) The two cdfs agree: substituting $u = -t$ and using $f_X(-u) = f_X(u)$,

$$\begin{aligned}F_{-X}(x) = P(-X \le x) = P(X \ge -x) &= \int_{-x}^{\infty} f_X(t)\, dt \\ &= \int_{-\infty}^{x} f_X(-u)\, du = \int_{-\infty}^{x} f_X(u)\, du = F_X(x).\end{aligned}$$

(b) $M_X(t) = M_X(-t)$ for every t, the two sides finite together. By (a) the variables X and $-X$ have one distribution, so they have the same expectation of $g(x) = e^{tx}$ (Definition 2.2.1):

$$M_X(t) = \mathrm{E}\, e^{tX} = \mathrm{E}\, e^{t(-X)} = \mathrm{E}\, e^{(-t)X} = M_X(-t).$$

Problem 2.26. Let $f(x)$ be a pdf and let a be a number such that, for all $\epsilon > 0$, $f(a+\epsilon) = f(a-\epsilon)$. Such a pdf is said to be *symmetric* about the point a.
(a) Give three examples of symmetric pdfs.
(b) Show that if $X \sim f(x)$, symmetric, then the median of X (see Exercise 2.17) is the number a.
(c) Show that if $X \sim f(x)$, symmetric, and $\mathrm{E}\, X$ exists, then $\mathrm{E}\, X = a$.
(d) Show that $f(x) = e^{-x}$, $x \ge 0$, is not a symmetric pdf.

(e) Show that for the pdf in part (d), the median is less than the mean.

Solution. (a) $\mathrm{n}(a, \sigma^2)$; uniform$(a-1, a+1)$; the double exponential $f(x) = \frac{1}{2}e^{-|x-a|}$, $-\infty < x < \infty$. (Check!)
(b) Split the total mass at a and reflect the left half onto the right half with $x = a - \epsilon$, then $x = a + \epsilon$:

$$\begin{aligned} P(X \le a) &= \int_{-\infty}^{a} f(x)\,dx = \int_{0}^{\infty} f(a-\epsilon)\,d\epsilon \\ &= \int_{0}^{\infty} f(a+\epsilon)\,d\epsilon = \int_{a}^{\infty} f(x)\,dx = P(X \ge a). \end{aligned}$$

The two ends sum to 1, so each equals $\frac{1}{2}$ and a satisfies the defining requirement of Exercise 2.17.
(c) The same reflection kills the first central term. Since $\mathrm{E}\,X$ exists, $\mathrm{E}(X-a)$ exists and splits at a:

$$\begin{aligned} \mathrm{E}(X-a) &= \int_{a}^{\infty} (x-a)f(x)\,dx + \int_{-\infty}^{a} (x-a)f(x)\,dx \\ &= \int_{0}^{\infty} \epsilon f(a+\epsilon)\,d\epsilon - \int_{0}^{\infty} \epsilon f(a-\epsilon)\,d\epsilon = 0, \end{aligned}$$

the two integrals being equal by symmetry. Hence $\mathrm{E}\,X = a$.
(d) Here $f(x) = 0$ for $x < 0$, so no centre can work. Given any candidate a, take $\epsilon = |a| + 1 > 0$. Then $a + \epsilon \ge 1$ and $a - \epsilon \le -1$, so

$$f(a+\epsilon) = e^{-(a+\epsilon)} > 0 = f(a-\epsilon),$$

contradicting symmetry about a.
(e) The median m solves $\int_0^m e^{-x}dx = \frac{1}{2}$, i.e. $1 - e^{-m} = \frac{1}{2}$, so $m = \log 2 \approx 0.693$, while $\mathrm{E}\,X = \int_0^\infty xe^{-x}dx = 1$. Thus $m = \log 2 < 1 = \mathrm{E}\,X$.

Problem 2.27. Let $f(x)$ be a pdf, and let a be a number such that if $a \ge x \ge y$, then $f(a) \ge f(x) \ge f(y)$, and if $a \le x \le y$, then $f(a) \ge f(x) \ge f(y)$. Such a pdf is called *unimodal* with a *mode* equal to a.
(a) Give an example of a unimodal pdf for which the mode is unique.
(b) Give an example of a unimodal pdf for which the mode is not unique.
(c) Show that if $f(x)$ is both symmetric (see Exercise 2.26) and unimodal, then the point of symmetry is a mode.
(d) Consider the pdf $f(x) = e^{-x}$, $x \ge 0$. Show that this pdf is unimodal. What is its mode?

Solution. The definition says exactly that f is nondecreasing on $(-\infty, a]$ and nonincreasing on $[a, \infty)$.
(a) $\mathrm{n}(0,1)$: $f(x) = (2\pi)^{-1/2}e^{-x^2/2}$ is strictly increasing on $(-\infty, 0]$ and strictly decreasing on $[0, \infty)$, so 0 is a mode and strictness makes it the only one.
(b) uniform$(0,1)$: for every $a \in (0,1)$, f is nondecreasing on $(-\infty, a]$ (values 0 then 1) and nonincreasing on $[a, \infty)$ (values 1 then 0). Every point of $(0,1)$ is a mode.
(c) Let b be a mode of f, symmetric about a. Symmetry reads $f(x) = f(2a - x)$, and $x \mapsto 2a - x$ reverses order while preserving the value of f, so $b' := 2a - b$ is a mode too. Since $a = \frac{1}{2}(b + b')$, relabelling if necessary gives $b \le a \le b'$, and then

$$\begin{aligned} [a, \infty) &\subseteq [b, \infty), \quad \text{on which } f \text{ is nonincreasing,} \\ (-\infty, a] &\subseteq (-\infty, b'], \quad \text{on which } f \text{ is nondecreasing,} \end{aligned}$$

which is exactly the assertion that a is a mode.
(d) The mode is 0, and it is unique. Here $f(x) = 0$ for $x < 0$ and $f(x) = e^{-x}$ for $x \ge 0$. For $0 \le x \le y$,

$$f(0) = 1 \ge e^{-x} \ge e^{-y},$$

and for $0 \ge x \ge y$ we have $f(0) = 1 \ge f(x) \ge f(y)$ since f is 0 at every negative argument. Uniqueness: $f(x) < 1 = f(0)$ for every $x \ne 0$, so no other point can dominate f.

Problem 2.28. Let μ_n denote the nth central moment of a random variable X. Two quantities of interest, in addition to the mean and variance, are

$$\alpha_3 = \frac{\mu_3}{(\mu_2)^{3/2}} \qquad \text{and} \qquad \alpha_4 = \frac{\mu_4}{\mu_2^2}.$$

The value α_3 is called the *skewness* and α_4 is called the *kurtosis*. The skewness measures the lack of symmetry in the pdf (see Exercise 2.26). The kurtosis, although harder to interpret, measures the peakedness or flatness of the pdf.
(a) Show that if a pdf is symmetric about a point a, then $\alpha_3 = 0$.
(b) Calculate α_3 for $f(x) = e^{-x}$, $x \geq 0$, a pdf that is skewed to the right.
(c) Calculate α_4 for each of the following pdfs and comment on the peakedness of each.

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, && -\infty < x < \infty, \\ f(x) &= \tfrac{1}{2}, && -1 < x < 1, \\ f(x) &= \tfrac{1}{2} e^{-|x|}, && -\infty < x < \infty. \end{aligned}$$

Solution. (a) $\mu_3 = 0$, hence $\alpha_3 = 0$. By Exercise 2.26(c) the mean is the point of symmetry a, so $\mu_3 = \mathrm{E}(X-a)^3$; splitting at a and reflecting each half as in Exercise 2.26(c),

$$\mu_3 = \int_0^\infty \epsilon^3 f(a+\epsilon)\, d\epsilon - \int_0^\infty \epsilon^3 f(a-\epsilon)\, d\epsilon = 0,$$

the two integrals being equal because $f(a+\epsilon) = f(a-\epsilon)$.
(b) $\alpha_3 = 2$. For $X \sim$ exponential(1), $\mathrm{E}\, X^k = \int_0^\infty x^k e^{-x} dx = k!$, so with $\mathrm{E}\, X = 1$,

$$\begin{aligned} \mu_2 &= \mathrm{E}\, X^2 - 1 = 1, \\ \mu_3 &= \mathrm{E}\, X^3 - 3\,\mathrm{E}\, X^2 + 3\,\mathrm{E}\, X - 1 = 6 - 6 + 3 - 1 = 2, \end{aligned}$$

and $\alpha_3 = 2/1^{3/2} = 2$.
(c) Each pdf is symmetric about 0, so the central moments are the raw moments.
(i) n$(0,1)$: $\mu_2 = 1$ and $\mu_4 = 3$ (the standard normal moments, Section 3.3), so

$$\alpha_4 = \frac{3}{1^2} = 3.$$

(ii) uniform$(-1,1)$: $\mu_2 = \int_{-1}^1 \tfrac{1}{2} x^2 dx = \tfrac{1}{3}$ and $\mu_4 = \int_{-1}^1 \tfrac{1}{2} x^4 dx = \tfrac{1}{5}$, so

$$\alpha_4 = \frac{1/5}{(1/3)^2} = \frac{9}{5} = 1.8.$$

(iii) Double exponential: by evenness $\mu_k = \int_0^\infty x^k e^{-x} dx = k!$ for even k, so $\mu_2 = 2$, $\mu_4 = 24$ and

$$\alpha_4 = \frac{24}{2^2} = 6.$$

The ordering $1.8 < 3 < 6$ is the comment asked for: the uniform is flattest, the normal intermediate, and the double exponential the most peaked, having a spike at 0 and heavy exponential tails.

2.5 Exercises 2.29–2.35

Problem 2.29. To calculate moments of discrete distributions, it is often easier to work with the *factorial moments* (see Miscellanea 2.6.2).
(a) Calculate the factorial moment $\mathrm{E}[X(X-1)]$ for the binomial and Poisson distributions.
(b) Use the results of part (a) to calculate the variances of the binomial and Poisson distributions.
(c) A particularly nasty discrete distribution is the *beta-binomial*, with pmf

$$P(Y=y) = a\left(\frac{1}{y+a}\right)\frac{\binom{n}{y}\binom{a+b-1}{a}}{\binom{n+a+b-1}{y+a}},$$

where n, a, and b are integers, and $y = 0, 1, 2, \ldots, n$. Use factorial moments to calculate the variance of the beta-binomial. (See Exercise 4.34 for another approach to this calculation.)

Solution. (a) $\mathrm{E}[X(X-1)] = n(n-1)p^2$ for the binomial(n,p) and λ^2 for the Poisson(λ). In both cases the factor $x(x-1)$ cancels the first two factors of $x!$ and leaves the same pmf with its index shifted by two. For the binomial,

$$\begin{aligned}\mathrm{E}[X(X-1)] &= \sum_{x=2}^{n}\frac{n!}{(x-2)!\,(n-x)!}\,p^x(1-p)^{n-x}\\ &= n(n-1)p^2\sum_{x=2}^{n}\binom{n-2}{x-2}p^{x-2}(1-p)^{(n-2)-(x-2)}\\ &= n(n-1)p^2,\end{aligned}$$

the last sum being the total mass of a binomial$(n-2,p)$ pmf. For the Poisson,

$$\begin{aligned}\mathrm{E}[X(X-1)] &= \sum_{x=2}^{\infty}x(x-1)\,\frac{e^{-\lambda}\lambda^x}{x!} = \lambda^2\sum_{x=2}^{\infty}\frac{e^{-\lambda}\lambda^{x-2}}{(x-2)!}\\ &= \lambda^2.\end{aligned}$$

(b) Since $X^2 = X(X-1)+X$, Theorem 2.2.5 gives

$$\operatorname{Var}X = \mathrm{E}[X(X-1)] + \mathrm{E}X - (\mathrm{E}X)^2.$$

With $\mathrm{E}X = np$ for the binomial,

$$\operatorname{Var}X = n(n-1)p^2 + np - n^2p^2 = np(1-p),$$

and with $\mathrm{E}X = \lambda$ for the Poisson,

$$\operatorname{Var}X = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

(c) $\operatorname{Var}Y = \dfrac{nab(n+a+b)}{(a+b)^2(a+b+1)}$.
First put the pmf in beta form: write out the three binomial coefficients and use $a/a! = 1/(a-1)!$ together with $(y+a)!/(y+a) = (y+a-1)!$, which is where the two stray factors a and $1/(y+a)$ go. This gives

$$\begin{aligned}P(Y=y) &= \binom{n}{y}\frac{(a+b-1)!\,(y+a-1)!\,(n-y+b-1)!}{(a-1)!\,(b-1)!\,(n+a+b-1)!}\\ &= \binom{n}{y}\frac{B(y+a,\;n-y+b)}{B(a,b)},\end{aligned}$$

where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. That these probabilities sum to 1 is the identity

$$\sum_{j=0}^{m}\binom{m}{j}B(j+\alpha,\,m-j+\beta) = B(\alpha,\beta), \qquad m \ge 0,$$

for positive integers α, β; it is the only tool needed below.
For the first factorial moment, drop the $y = 0$ term and set $j = y - 1$, $m = n - 1$:

$$\begin{aligned} \mathrm{E}Y &= \sum_{y=1}^{n} \frac{n!}{(y-1)!\,(n-y)!} \, \frac{B(y+a,\, n-y+b)}{B(a,b)} \\ &= \frac{n}{B(a,b)} \sum_{j=0}^{m} \binom{m}{j} B\big(j+(a+1),\, m-j+b\big) \\ &= n\,\frac{B(a+1,b)}{B(a,b)} \;=\; \frac{na}{a+b}. \end{aligned}$$

For the second, the factor $y(y-1)$ again shifts the index by two; with $j = y - 2$, $m = n - 2$,

$$\begin{aligned} \mathrm{E}[Y(Y-1)] &= \sum_{y=2}^{n} \frac{n!}{(y-2)!\,(n-y)!} \, \frac{B(y+a,\, n-y+b)}{B(a,b)} \\ &= \frac{n(n-1)}{B(a,b)} \sum_{j=0}^{m} \binom{m}{j} B\big(j+(a+2),\, m-j+b\big) \\ &= n(n-1)\,\frac{B(a+2,b)}{B(a,b)} \;=\; \frac{n(n-1)a(a+1)}{(a+b)(a+b+1)}. \end{aligned}$$

(The two ratios of beta functions collapse by $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.) Now the identity of part (b):

$$\begin{aligned} \operatorname{Var} Y &= \frac{n(n-1)a(a+1)}{(a+b)(a+b+1)} + \frac{na}{a+b} - \frac{n^2a^2}{(a+b)^2} \\ &= \frac{nab(n+a+b)}{(a+b)^2(a+b+1)}. \end{aligned}$$

(Check! Put everything over $(a+b)^2(a+b+1)$; the numerator is $na\big[(n-1)(a+1)(a+b) + (a+b)(a+b+1) - na(a+b+1)\big]$, whose bracket simplifies to $b(n+a+b)$.)

Problem 2.30. Find the moment generating function corresponding to
(a) $f(x) = \dfrac{1}{c},\ 0 < x < c.$
(b) $f(x) = \dfrac{2x}{c^2},\ 0 < x < c.$
(c) $f(x) = \dfrac{1}{2\beta} e^{-|x-\alpha|/\beta},\ -\infty < x < \infty,\ -\infty < \alpha < \infty,\ \beta > 0.$
(d) $P(X = x) = \dbinom{r+x-1}{x} p^r (1-p)^x,\ x = 0, 1, \ldots,\ 0 < p < 1,\ r > 0$ an integer.

Solution. (a) $M_X(t) = \dfrac{e^{ct} - 1}{ct}$ for $t \neq 0$, and $M_X(0) = 1$:

$$M_X(t) = \frac{1}{c} \int_0^c e^{tx}\, dx = \frac{1}{c} \cdot \frac{e^{ct} - 1}{t}.$$

It is finite for every t, and $(e^{ct} - 1)/(ct) \to 1$ as $t \to 0$, so M_X is continuous at 0 with the required value $M_X(0) = 1$.
(b) $M_X(t) = \dfrac{2\big(cte^{ct} - e^{ct} + 1\big)}{c^2 t^2}$ for $t \neq 0$, and $M_X(0) = 1$. Integrating by parts,

$$\begin{aligned} M_X(t) &= \frac{2}{c^2} \int_0^c xe^{tx}\, dx = \frac{2}{c^2} \left[\frac{xe^{tx}}{t}\right]_0^c - \frac{2}{c^2 t} \int_0^c e^{tx}\, dx \\ &= \frac{2}{c^2} \left(\frac{ce^{ct}}{t} - \frac{e^{ct} - 1}{t^2}\right). \end{aligned}$$

(c) $M_X(t) = \dfrac{e^{\alpha t}}{1-\beta^2 t^2}$, $|t| < 1/\beta$. Substituting $y = x - \alpha$ and splitting the absolute value at 0,

$$\begin{aligned} M_X(t) &= \frac{e^{\alpha t}}{2\beta} \int_{-\infty}^{\infty} e^{ty} e^{-|y|/\beta}\, dy \\ &= \frac{e^{\alpha t}}{2\beta} \left(\int_0^{\infty} e^{-(1/\beta - t)y}\, dy + \int_0^{\infty} e^{-(1/\beta + t)y}\, dy \right) \\ &= \frac{e^{\alpha t}}{2\beta} \left(\frac{1}{1/\beta - t} + \frac{1}{1/\beta + t} \right) = \frac{e^{\alpha t}}{1-\beta^2 t^2}. \end{aligned}$$

Both integrals converge exactly when $|t| < 1/\beta$, which is the stated neighborhood of 0.
(d) $M_X(t) = \left(\dfrac{p}{1-(1-p)e^t} \right)^r$, $t < -\log(1-p)$. Write $q = 1-p$ and use the negative binomial series $\sum_{x=0}^{\infty} \binom{r+x-1}{x} u^x = (1-u)^{-r}$, valid for $|u| < 1$:

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{\infty} \binom{r+x-1}{x} p^r \left(qe^t\right)^x \\ &= \frac{p^r}{(1-qe^t)^r}. \end{aligned}$$

The series converges precisely when $qe^t < 1$, that is $t < \log(1/q)$, an interval containing 0.

Problem 2.31. Does a distribution exist for which $M_X(t) = t/(1-t)$, $|t| < 1$? If yes, find it. If no, prove it.

Solution. No. Every moment generating function satisfies

$$M_X(0) = \mathrm{E}\, e^{0 \cdot X} = \mathrm{E}\, 1 = 1,$$

whereas the proposed function takes the value $0/(1-0) = 0$ at $t = 0$. So $t/(1-t)$ is not the mgf of any random variable.
Method (2): the mgf generates the moments by $M_X^{(n)}(0) = \mathrm{E}X^n$ (Definition 2.3.6 and Theorem 2.3.7 in the form stated there). Expanding in a geometric series for $|t| < 1$,

$$\frac{t}{1-t} = \sum_{n=1}^{\infty} t^n = \sum_{n=0}^{\infty} \frac{\mathrm{E}X^n}{n!} t^n \implies \mathrm{E}X^n = n!,\ n \geq 1,$$

but the constant term forces $\mathrm{E}X^0 = 0$, which is impossible since $X^0 \equiv 1$.

Problem 2.32. Let $M_X(t)$ be the moment generating function of X, and define $S(t) = \log(M_X(t))$. Show that

$$\left.\frac{d}{dt} S(t)\right|_{t=0} = \mathrm{E}X \qquad \text{and} \qquad \left.\frac{d^2}{dt^2} S(t)\right|_{t=0} = \mathrm{Var}\, X.$$

Solution. Differentiate the logarithm and evaluate at 0, where $M_X(0) = 1$ and $M_X^{(n)}(0) = \mathrm{E}X^n$ by Theorem 2.3.7 (legitimate because the mgf is assumed to exist in a neighborhood of 0, where it is infinitely differentiable). First,

$$\left.\frac{d}{dt} S(t)\right|_{t=0} = \left.\frac{M_X'(t)}{M_X(t)}\right|_{t=0} = \frac{\mathrm{E}X}{1} = \mathrm{E}X.$$

Second, by the quotient rule,

$$\frac{d^2}{dt^2}S(t) = \frac{M_X''(t)M_X(t) - \big(M_X'(t)\big)^2}{\big(M_X(t)\big)^2},$$
$$\left.\frac{d^2}{dt^2}S(t)\right|_{t=0} = \frac{\mathrm{E}X^2 \cdot 1 - (\mathrm{E}X)^2}{1^2} = \operatorname{Var} X,$$

the last equality being Definition 2.3.2 expanded by the linearity of Theorem 2.2.5(a).

Problem 2.33. In each of the following cases verify the expression given for the moment generating function, and in each case use the mgf to calculate $\mathrm{E}X$ and $\operatorname{Var} X$.

(a) $P(X = x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$, $\quad M_X(t) = e^{\lambda(e^t-1)}$, $\quad x = 0, 1, \ldots;\ \lambda > 0.$

(b) $P(X = x) = p(1-p)^x$, $\quad M_X(t) = \dfrac{p}{1-(1-p)e^t}$, $\quad x = 0, 1, \ldots;\ 0 < p < 1.$

(c) $f_X(x) = \dfrac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi}\,\sigma}$, $\quad M_X(t) = e^{\mu t + \sigma^2 t^2/2}$, $\quad -\infty < x < \infty;\ -\infty < \mu < \infty,\ \sigma > 0.$

Solution. Throughout, $\mathrm{E}X = S'(0)$ and $\operatorname{Var} X = S''(0)$ for $S(t) = \log M_X(t)$, by Exercise 2.32.
(a) Summing the Poisson series,

$$M_X(t) = \sum_{x=0}^{\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)},$$

finite for every real t. Here $S(t) = \lambda(e^t - 1)$, so $S'(t) = S''(t) = \lambda e^t$ and

$$\mathrm{E}X = \lambda, \qquad \operatorname{Var} X = \lambda.$$

(b) With $q = 1 - p$, the geometric series gives

$$M_X(t) = p\sum_{x=0}^{\infty}\big(qe^t\big)^x = \frac{p}{1-qe^t},$$

convergent for $t < \log(1/q)$, a neighborhood of 0 since $q < 1$. Here $S(t) = \log p - \log(1 - qe^t)$, so

$$S'(t) = \frac{qe^t}{1-qe^t},$$
$$S''(t) = \frac{qe^t}{1-qe^t} + \frac{q^2e^{2t}}{(1-qe^t)^2},$$

and evaluating at $t = 0$, where $1 - q = p$,

$$\mathrm{E}X = \frac{q}{p} = \frac{1-p}{p}, \qquad \operatorname{Var} X = \frac{q}{p} + \frac{q^2}{p^2} = \frac{q(p+q)}{p^2} = \frac{1-p}{p^2}.$$

(c) Complete the square in the exponent: with $A = x - \mu$,

$$tx - \frac{(x-\mu)^2}{2\sigma^2} = \mu t + \frac{\sigma^2 t^2}{2} - \frac{\big(A - \sigma^2 t\big)^2}{2\sigma^2}.$$

Hence

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\left(x-(\mu+\sigma^2 t)\right)^2/(2\sigma^2)}\,dx$$
$$= e^{\mu t + \sigma^2 t^2/2},$$

the remaining integral being that of an $\mathrm{n}(\mu + \sigma^2 t, \sigma^2)$ pdf, hence 1, for every t. Now $S(t) = \mu t + \sigma^2 t^2/2$ is a quadratic, so $S'(t) = \mu + \sigma^2 t$ and $S''(t) = \sigma^2$, giving

$$\mathrm{E}X = \mu, \qquad \operatorname{Var} X = \sigma^2.$$

Problem 2.34. A distribution cannot be uniquely determined by a finite collection of moments, as this example from Romano and Siegel (1986) shows. Let X have the normal distribution, that is, X has pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad -\infty < x < \infty.$$

Define a discrete random variable Y by

$$P\big(Y = \sqrt{3}\big) = P\big(Y = -\sqrt{3}\big) = \frac{1}{6}, \qquad P(Y = 0) = \frac{2}{3}.$$

Show that

$$\mathrm{E}X^r = \mathrm{E}Y^r \qquad \text{for } r = 1, 2, 3, 4, 5.$$

(Romano and Siegel point out that for any finite n there exists a discrete, and hence nonnormal, random variable whose first n moments are equal to those of X.)

Solution. Both distributions are symmetric about 0, which kills $r = 1, 3, 5$ on both sides at once; the two even moments then match by arithmetic. Precisely: $f_X(-x) = f_X(x)$ and $P(Y = -y) = P(Y = y)$, so for odd r the integrand $x^r f_X(x)$ and the summand $y^r P(Y = y)$ are odd functions and

$$\mathrm{E}X^r = 0 = \mathrm{E}Y^r, \qquad r = 1, 3, 5$$

(the normal integrals converge absolutely, so the cancellation is legitimate).
For the even moments of X, take $\mu = 0$, $\sigma = 1$ in Exercise 2.33(c): $M_X(t) = e^{t^2/2}$. Expanding both this and the defining series of Definition 2.3.6,

$$\sum_{k=0}^{\infty} \frac{t^{2k}}{2^k k!} = e^{t^2/2} = M_X(t) = \sum_{n=0}^{\infty} \frac{\mathrm{E}X^n}{n!}\, t^n,$$

and matching coefficients of t^{2k} gives $\mathrm{E}X^{2k} = (2k)!/(2^k k!)$. Hence

$$\mathrm{E}X^2 = \frac{2!}{2 \cdot 1} = 1, \qquad \mathrm{E}X^4 = \frac{4!}{2^2 \cdot 2!} = 3.$$

For Y, direct summation over the three atoms:

$$\mathrm{E}Y^2 = 3 \cdot \tfrac{1}{6} + 0 \cdot \tfrac{2}{3} + 3 \cdot \tfrac{1}{6} = 1,$$
$$\mathrm{E}Y^4 = 9 \cdot \tfrac{1}{6} + 0 \cdot \tfrac{2}{3} + 9 \cdot \tfrac{1}{6} = 3.$$

So $\mathrm{E}X^r = \mathrm{E}Y^r$ for $r = 1, 2, 3, 4, 5$.

Problem 2.35. Fill in the gaps in Example 2.3.10. There X_1 has the lognormal pdf

$$f_1(x) = \frac{1}{\sqrt{2\pi}\, x}\, e^{-(\log x)^2/2}, \qquad 0 \le x < \infty,$$

and X_2 has the pdf $f_2(x) = f_1(x)\big[1 + \sin(2\pi \log x)\big]$, $0 \le x < \infty$; these are the two distinct pdfs with identical moments pictured in Figure 2.3.2.
(a) Show that if $X_1 \sim f_1(x)$, then

$$\mathrm{E}X_1^r = e^{r^2/2}, \qquad r = 0, 1, \ldots.$$

So $f_1(x)$ has all of its moments, and all of the moments are finite.
(b) Now show that

$$\int_0^{\infty} x^r f_1(x) \sin(2\pi \log x)\, dx = 0$$

for all positive integers r, so $\mathrm{E}X_1^r = \mathrm{E}X_2^r$ for all r. (Romano and Siegel 1986 discuss an extreme version of this example, where an entire class of distinct pdfs have the same moments. Also, Berg 1988 has shown that this moment behavior can arise with simpler transforms of the normal distribution such as X^3.)

Solution. (a) The substitution $y = \log x$, $dy = dx/x$, turns the rth moment into the standard normal mgf evaluated at r:

$$\begin{aligned}\mathrm{E}X_1^r &= \int_0^\infty x^r \, \frac{1}{\sqrt{2\pi}\, x} e^{-(\log x)^2/2}\, dx \\ &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{ry} e^{-y^2/2}\, dy \\ &= M_Z(r) = e^{r^2/2},\end{aligned}$$

where $Z \sim \mathrm{n}(0,1)$ and $M_Z(t) = e^{t^2/2}$ by Exercise 2.33(c) with $\mu = 0, \sigma = 1$. Since the normal mgf is finite for every real argument, every moment of X_1 is finite.
(b) The same substitution, followed by the shift $u = y - r$, makes the integrand odd. With $y = \log x$,

$$\begin{aligned}I_r &:= \int_0^\infty x^r f_1(x) \sin(2\pi \log x)\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{ry - y^2/2} \sin(2\pi y)\, dy \\ &= \frac{e^{r^2/2}}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-(y-r)^2/2} \sin(2\pi y)\, dy\end{aligned}$$

by the completion of the square $ry - y^2/2 = r^2/2 - (y-r)^2/2$. Now put $u = y - r$ and use $\sin(2\pi u + 2\pi r) = \sin(2\pi u)$, which holds because r is an integer:

$$I_r = \frac{e^{r^2/2}}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-u^2/2} \sin(2\pi u)\, du = 0,$$

the integrand being an odd function of u and absolutely integrable (it is dominated by $e^{-u^2/2}$). Consequently, for every $r = 0, 1, 2, \ldots$,

$$\begin{aligned}\mathrm{E}X_2^r &= \int_0^\infty x^r f_1(x) \big[1 + \sin(2\pi \log x)\big]\, dx \\ &= \mathrm{E}X_1^r + I_r = \mathrm{E}X_1^r.\end{aligned}$$

2.6 Exercises 2.36–2.40

Problem 2.36. The *lognormal distribution*, on which Example 2.3.10 is based, has an interesting property. If we have the pdf

$$f(x) = \frac{1}{\sqrt{2\pi}\, x}\, e^{-(\log x)^2/2}, \qquad 0 \le x < \infty,$$

then Exercise 2.35 shows that all moments exist and are finite. However, this distribution does not have a moment generating function, that is,

$$M_X(t) = \int_0^\infty \frac{e^{tx}}{\sqrt{2\pi}\, x}\, e^{-(\log x)^2/2}\, dx$$

does not exist. Prove this.

Solution. The integral is $+\infty$ for every $t > 0$, so M_X is finite on no neighbourhood of 0.
Fix $t > 0$. Every term of the exponential series is a lower bound for e^{tx} on $x > 0$, so for each $n = 0, 1, 2, \ldots$,

$$M_X(t) = \mathrm{E}\, e^{tX} \;\geq\; \mathrm{E}\left[\frac{(tX)^n}{n!}\right] = \frac{t^n}{n!}\,\mathrm{E}\, X^n \;=\; \frac{t^n e^{n^2/2}}{n!},$$

the last step by Exercise 2.35(a), which gives $\mathrm{E}\, X^n = e^{n^2/2}$ (and in particular guarantees every moment is finite, so the bound is not vacuous). Since $n! \leq n^n$,

$$\frac{t^n e^{n^2/2}}{n!} \;\geq\; \exp\bigl\{\tfrac{1}{2}n^2 + n\log t - n\log n\bigr\} \;\xrightarrow[n\to\infty]{}\; \infty,$$

because the $n^2/2$ term dominates both $n\log t$ and $n\log n$. The left side $M_X(t)$ does not depend on n, so $M_X(t) = \infty$.
For $t \leq 0$ we have $e^{tx} \leq 1$ on $x > 0$ and hence $M_X(t) \leq 1 < \infty$. Thus $\{t : M_X(t) < \infty\} = (-\infty, 0]$, which contains no open interval about 0, and the moment generating function does not exist.
Method (2): substituting $y = \log x$ turns the integral into

$$M_X(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, \exp\bigl\{te^y - \tfrac{1}{2}y^2\bigr\}\, dy,$$

and for $t > 0$ the exponent $te^y - \frac{1}{2}y^2 \to +\infty$ as $y \to \infty$ (since $e^y > y^3/6$), so the integrand exceeds 1 for all large y and the integral diverges.

Problem 2.37. Referring to the situation described in Miscellanea 2.6.3: McCullagh (1994) considers the pair of densities

$$f_1 = n(0,1) \qquad \text{and} \qquad f_2(x) = f_1(x)\left[1 + \tfrac{1}{2}\sin(2\pi x)\right],$$

that is, $f_1(x) = (2\pi)^{-1/2}e^{-x^2/2}$ on $-\infty < x < \infty$, with cumulant generating functions

$$K_1(t) = \frac{t^2}{2}, \qquad K_2(t) = K_1(t) + \log\Bigl[1 + \tfrac{1}{2}e^{-2\pi^2}\sin(2\pi t)\Bigr].$$

He notes that although the densities are visibly dissimilar, the cgfs are virtually identical, with maximum difference less than 1.34×10^{-9} over the entire range (less than the size of one pixel).
(a) Plot the pdfs f_1 and f_2 to illustrate their difference.
(b) Plot the cumulant generating functions K_1 and K_2 to illustrate their similarity.
(c) Calculate the moment generating functions of the pdfs f_1 and f_2. Are they similar or different?
(d) How do the pdfs f_1 and f_2 relate to the pdfs described in Example 2.3.10?

Solution. (c) is the computation the rest of the exercise dramatizes:

$$M_1(t) = e^{t^2/2}, \qquad M_2(t) = e^{t^2/2}\left[1 + \tfrac{1}{2}e^{-2\pi^2}\sin(2\pi t)\right].$$

(a) f_2 is f_1 modulated by the period-1 wave $1 + \frac{1}{2}\sin(2\pi x)$: the curves meet at every $x \in \frac{1}{2}\mathbb{Z}$, with $f_2 = \frac{3}{2}f_1$ at $x = k + \frac{1}{4}$ and $f_2 = \frac{1}{2}f_1$ at $x = k - \frac{1}{4}$. Plotted together on $[-3, 3]$, f_1 is the single-humped bell while f_2 ripples, rising to 0.580 at $x = \frac{1}{4}$ and falling to 0.193 at $x = -\frac{1}{4}$, against a bell whose own maximum is 0.399:

x	$f_1(x)$	$f_2(x)$
-1.0	0.2420	0.2420
-0.75	0.3011	0.4517
-0.5	0.3521	0.3521
-0.25	0.3867	0.1933
0.0	0.3989	0.3989
0.25	0.3867	0.5800
0.5	0.3521	0.3521
0.75	0.3011	0.1506
1.0	0.2420	0.2420

(b) Plotted on the same axes, K_1 and K_2 are one parabola: the vertical gap is

$$K_2(t) - K_1(t) = \log\Big[1 + \tfrac{1}{2}e^{-2\pi^2}\sin(2\pi t)\Big],$$

whose modulus is at most $\log(1 + \frac{1}{2}e^{-2\pi^2}) = 1.3376 \times 10^{-9}$, uniformly in t.

t	$K_1(t)$	$K_2(t)$	difference
-1.00	0.5000000000	0.5000000000	0
-0.25	0.0312500000	0.0312499987	-1.34×10^{-9}
0.00	0.0000000000	0.0000000000	0
0.25	0.0312500000	0.0312500013	$+1.34 \times 10^{-9}$
1.00	0.5000000000	0.5000000000	0

(c) $M_1(t) = e^{t^2/2}$ is Exercise 2.33(c) with $\mu = 0$, $\sigma = 1$. For M_2, complete the square, $e^{tx}f_1(x) = e^{t^2/2}f_1(x-t)$, and substitute $u = x - t$:

$$\begin{aligned}\int_{-\infty}^{\infty} e^{tx} f_1(x)\sin(2\pi x)\,dx &= e^{t^2/2}\int_{-\infty}^{\infty} f_1(u)\sin\big(2\pi(u+t)\big)\,du \\ &= e^{t^2/2}\Big[\sin(2\pi t)\,\mathrm{E}\cos(2\pi Z) \\ &\qquad + \cos(2\pi t)\,\mathrm{E}\sin(2\pi Z)\Big] \\ &= e^{t^2/2}e^{-2\pi^2}\sin(2\pi t),\end{aligned}$$

where $Z \sim n(0,1)$: the second expectation vanishes because f_1 is even and sin odd, and $\mathrm{E}\cos(sZ) = e^{-s^2/2}$ is the (real) characteristic function of $n(0,1)$, evaluated at $s = 2\pi$. Hence

$$M_2(t) = e^{t^2/2} + \tfrac{1}{2}e^{t^2/2}e^{-2\pi^2}\sin(2\pi t),$$

which is the stated formula, and $K_2 = \log M_2$ is the cgf quoted above. Both mgfs are finite for every t. They are mathematically **different**, as Theorem 2.3.11(b) demands: the mgfs being finite on a neighbourhood of 0, equal mgfs would force $f_1 = f_2$. Already the first moments separate them,

$$\mathrm{E}_2\, X = M_2'(0) = \pi e^{-2\pi^2} \neq 0 = \mathrm{E}_1\, X.$$

But every difference carries the factor $\frac{1}{2}e^{-2\pi^2} = 1.34 \times 10^{-9}$, so numerically the two mgfs are the same function: McCullagh's answer is yes for mathematical purposes but a resounding no for numerical purposes.

(d) They are the pair of Example 2.3.10 pulled back through the logarithm. There $g_1(y) = (\sqrt{2\pi}\,y)^{-1}e^{-(\log y)^2/2}$ is the lognormal and $g_2(y) = g_1(y)[1+\sin(2\pi\log y)]$. If X has pdf f_1 then $Y = e^X$ has pdf g_1; if X has pdf f_2 then, by Theorem 2.1.5 with $y = e^x$,

$$f_Y(y) = f_2(\log y)\frac{1}{y} = g_1(y)\left[1 + \tfrac{1}{2}\sin(2\pi\log y)\right],$$

which is g_2 with the oscillation amplitude halved. So f_1, f_2 are the Example 2.3.10 pair transported to the log scale (up to that factor $\frac{1}{2}$), and the transport is what restores the mgf: g_1 and g_2 have none at all (Exercise 2.36) and identical moments, whereas M_1, M_2 exist everywhere and do differ.

Problem 2.38. Let X have the negative binomial distribution with pmf

$$f(x) = \binom{r+x-1}{x} p^r (1-p)^x, \qquad x = 0, 1, 2, \ldots,$$

where $0 < p < 1$ and $r > 0$ is an integer.
(a) Calculate the mgf of X.
(b) Define a new random variable by $Y = 2pX$. Show that, as $p \to 0$, the mgf of Y converges to that of a chi squared random variable with $2r$ degrees of freedom by showing that

$$\lim_{p \to 0} M_Y(t) = \left(\frac{1}{1-2t} \right)^r, \qquad |t| < \tfrac{1}{2}.$$

Solution. (a)

$$M_X(t) = \left(\frac{p}{1-(1-p)e^t} \right)^r, \qquad t < -\log(1-p).$$

Indeed, with $u = (1-p)e^t$ the negative binomial series $\sum_{x \geq 0} \binom{r+x-1}{x} u^x = (1-u)^{-r}$ gives

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{\infty} e^{tx} \binom{r+x-1}{x} p^r (1-p)^x \\ &= p^r \sum_{x=0}^{\infty} \binom{r+x-1}{x} \left[(1-p)e^t \right]^x = \frac{p^r}{\left(1-(1-p)e^t\right)^r}, \end{aligned}$$

the series converging exactly when $u < 1$, i.e. $t < -\log(1-p)$.
(b) Fix $|t| < \frac{1}{2}$. Since $-\log(1-p) > p > 2pt$ for all $0 < p < 1$, the point $2pt$ lies in the region of convergence of M_X, so

$$M_Y(t) = \mathrm{E}\, e^{t(2pX)} = M_X(2pt) = \left(\frac{p}{1-(1-p)e^{2pt}} \right)^r = \left(\frac{1}{h(p)} \right)^r,$$

where $h(p) = \left[1-(1-p)e^{2pt}\right]/p$. Both numerator and denominator of h vanish at $p = 0$, and by l'Hopital's rule in p,

$$\begin{aligned} \lim_{p \to 0} h(p) &= \lim_{p \to 0} \frac{\partial}{\partial p} \left[1-(1-p)e^{2pt} \right] \\ &= \lim_{p \to 0} \left[e^{2pt} - 2t(1-p)e^{2pt} \right] \\ &= \lim_{p \to 0} e^{2pt} \left(1 - 2t + 2pt \right) \;=\; 1 - 2t. \end{aligned}$$

Because $1 - 2t > 0$, the map $v \mapsto v^{-r}$ is continuous there and

$$\lim_{p \to 0} M_Y(t) = \left(\frac{1}{1-2t} \right)^r, \qquad |t| < \tfrac{1}{2},$$

which is the mgf of a chi squared random variable with $2r$ degrees of freedom (the χ^2_ν mgf is $(1-2t)^{-\nu/2}$ for $t < \frac{1}{2}$; here $\nu = 2r$).

Problem 2.39. In each of the following cases calculate the indicated derivatives, justifying all operations.
(a) $\dfrac{d}{dx} \displaystyle\int_0^x e^{-\lambda t}\, dt$
(b) $\dfrac{d}{d\lambda} \displaystyle\int_0^\infty e^{-\lambda t}\, dt$

(c) $\dfrac{d}{dt}\displaystyle\int_t^1 \frac{1}{x^2}\,dx$
(d) $\dfrac{d}{dt}\displaystyle\int_1^\infty \frac{1}{(x-t)^2}\,dx$

Solution. (a) $e^{-\lambda x}$. Leibnitz's Rule (Theorem 2.4.1) with $a(x)=0$, $b(x)=x$ and an integrand free of the differentiation variable leaves only the boundary term,

$$\frac{d}{dx}\int_0^x e^{-\lambda t}\,dt = e^{-\lambda b(x)}\,b'(x) = e^{-\lambda x};$$

no interchange is involved, and $t \mapsto e^{-\lambda t}$ is continuous, so this is just the Fundamental Theorem of Calculus.
(b) $-1/\lambda^2$, for $\lambda > 0$. Here the range is infinite, so the interchange needs Corollary 2.4.4. Fix $\lambda > 0$ and pick δ_0 with $0 < \delta_0 < \lambda$. For all λ' with $|\lambda' - \lambda| \le \delta_0$ and all $t > 0$,

$$\left|\frac{\partial}{\partial\lambda}e^{-\lambda t}\right|_{\lambda=\lambda'} = te^{-\lambda' t} \le te^{-(\lambda-\delta_0)t} =: g(t,\lambda),$$

and $\int_0^\infty g(t,\lambda)\,dt = (\lambda-\delta_0)^{-2} < \infty$ because $\lambda - \delta_0 > 0$. So (2.4.3) holds and

$$\frac{d}{d\lambda}\int_0^\infty e^{-\lambda t}\,dt = -\int_0^\infty te^{-\lambda t}\,dt = -\frac{1}{\lambda^2},$$

agreeing with the direct evaluation $\frac{d}{d\lambda}(1/\lambda) = -1/\lambda^2$.
(c) $-1/t^2$, for $t > 0$. The integrand has no t in it, so Leibnitz's Rule contributes only the lower-limit term, with $a(t) = t$, $b = 1$:

$$\frac{d}{dt}\int_t^1 \frac{dx}{x^2} = -\frac{1}{a(t)^2}\,a'(t) = -\frac{1}{t^2}.$$

(Directly, $\int_t^1 x^{-2}dx = t^{-1} - 1$. Check!) For $t \le 0$ the integral diverges and there is nothing to differentiate.
(d) $1/(1-t)^2$, for $t < 1$ (the integral diverges otherwise). Evaluating first,

$$\int_1^\infty \frac{dx}{(x-t)^2} = \left[\frac{-1}{x-t}\right]_1^\infty = \frac{1}{1-t},$$

whose derivative is $1/(1-t)^2$. Differentiating under the integral sign gives the same number, and that interchange is what needs justifying. Fix $t_0 < 1$ and choose $\delta_0 > 0$ with $t_0 + \delta_0 < 1$. For $x \ge 1$ and $|t' - t_0| \le \delta_0$,

$$\left|\frac{\partial}{\partial t}\frac{1}{(x-t)^2}\right|_{t=t'} = \frac{2}{(x-t')^3} \le \frac{2}{(x-t_0-\delta_0)^3} =: g(x,t_0),$$

and, since $1 - t_0 - \delta_0 > 0$,

$$\int_1^\infty g(x,t_0)\,dx = \frac{1}{(1-t_0-\delta_0)^2} < \infty.$$

Corollary 2.4.4 therefore applies and

$$\frac{d}{dt}\int_1^\infty \frac{dx}{(x-t)^2} = \int_1^\infty \frac{2\,dx}{(x-t)^3} = \frac{1}{(1-t)^2},$$

which is indeed $\frac{d}{dt}\left[(1-t)^{-1}\right]$.

Problem 2.40. Prove

$$\sum_{k=0}^{x}\binom{n}{k}p^k(1-p)^{n-k}=(n-x)\binom{n}{x}\int_0^{1-p}t^{\,n-x-1}(1-t)^x\,dt.$$

(Hint: Integrate by parts or differentiate both sides with respect to p.)

Solution. Both sides vanish at $p=1$ and have the same derivative on $(0,1)$. (Throughout $0\le x\le n-1$: at $x=n$ the right side reads 0 times a divergent integral, so the identity is stated only for $x<n$.)
Write $L(p)$ and $R(p)$ for the two sides. For R, Leibnitz's Rule (Theorem 2.4.1) with $a=0$, $b(p)=1-p$ and an integrand free of p (continuous on $[0,1-p]$, since $n-x-1\ge 0$) leaves only the boundary term:

$$\begin{aligned}R'(p)&=(n-x)\binom{n}{x}\,b(p)^{\,n-x-1}\big(1-b(p)\big)^x\,b'(p)\\&=-\,(n-x)\binom{n}{x}\,p^x(1-p)^{\,n-x-1}.\end{aligned}$$

For L, differentiate termwise and use $k\binom{n}{k}=n\binom{n-1}{k-1}$ and $(n-k)\binom{n}{k}=n\binom{n-1}{k}$. Putting

$$a_j=\binom{n-1}{j}p^j(1-p)^{\,n-1-j},\qquad a_{-1}:=0,$$

the kth derivative is

$$\begin{aligned}\frac{d}{dp}\binom{n}{k}p^k(1-p)^{n-k}&=\binom{n}{k}\Big[kp^{k-1}(1-p)^{n-k}\\&\qquad-(n-k)p^k(1-p)^{\,n-k-1}\Big]\\&=n\big(a_{k-1}-a_k\big),\end{aligned}$$

so the sum telescopes:

$$L'(p)=n\sum_{k=0}^{x}\big(a_{k-1}-a_k\big)=-\,n\,a_x=-\,n\binom{n-1}{x}p^x(1-p)^{\,n-1-x}.$$

Since $n\binom{n-1}{x}=(n-x)\binom{n}{x}$, we get $L'\equiv R'$ on $(0,1)$.
At $p=1$ every term of L carries a factor $(1-p)^{n-k}$ with $n-k\ge n-x\ge 1$, so $L(1)=0$; and $R(1)=0$ because the integral is over $[0,0]$. Both functions are continuous on $[0,1]$, so $L\equiv R$ there.
Method (2) (integration by parts): with $u=(1-t)^x$ and $dv=(n-x)t^{\,n-x-1}dt$,

$$\begin{aligned}R_x(p)&=\binom{n}{x}\Big[t^{\,n-x}(1-t)^x\Big]_0^{1-p}\\&\qquad+x\binom{n}{x}\int_0^{1-p}t^{\,n-x}(1-t)^{x-1}\,dt\\&=\binom{n}{x}p^x(1-p)^{\,n-x}+R_{x-1}(p),\end{aligned}$$

the last step because $x\binom{n}{x}=(n-x+1)\binom{n}{x-1}$. Since $R_0(p)=n\int_0^{1-p}t^{\,n-1}dt=(1-p)^n$, induction on x gives $R_x=L$.

3 Common Families of Distributions

Contents

3.1 Exercises 3.1–3.7 65
3.2 Exercises 3.8–3.14 68
3.3 Exercises 3.15–3.21 73
3.4 Exercises 3.22–3.28 76
3.5 Exercises 3.29–3.35 82
3.6 Exercises 3.36–3.42 88
3.7 Exercises 3.43–3.49 91
3.8 Exercises 3.50–3.50 96

3.1 Exercises 3.1–3.7

Problem 3.1. Find expressions for EX and $\operatorname{Var} X$ if X is a random variable with the general discrete uniform(N_0, N_1) distribution that puts equal probability on each of the values $N_0, N_0+1, \ldots, N_1$. Here $N_0 \le N_1$ and both are integers.

Solution.

$$EX = \frac{N_0 + N_1}{2}, \qquad \operatorname{Var} X = \frac{(N_1 - N_0)(N_1 - N_0 + 2)}{12}.$$

Put $n = N_1 - N_0 + 1$ (the number of mass points) and $Y = X - N_0 + 1$. Since $P(X = x) = 1/(N_1 - N_0 + 1)$ on $\{N_0, \ldots, N_1\}$ (p. 86), Y has the discrete uniform$(1, n)$ pmf (3.2.1), for which the book computes $EY = (n+1)/2$ and $\operatorname{Var} Y = (n+1)(n-1)/12$. Location shifts move the mean and leave the variance alone, so

$$\begin{aligned} EX &= EY + N_0 - 1 = \frac{n+1}{2} + N_0 - 1 \\ &= \frac{N_1 - N_0 + 2}{2} + N_0 - 1 = \frac{N_0 + N_1}{2}, \\ \operatorname{Var} X &= \operatorname{Var} Y = \frac{(n+1)(n-1)}{12} \\ &= \frac{(N_1 - N_0 + 2)(N_1 - N_0)}{12}. \end{aligned}$$

Problem 3.2. A manufacturer receives a lot of 100 parts from a vendor. The lot will be unacceptable if more than five of the parts are defective. The manufacturer is going to select randomly K parts from the lot for inspection and the lot will be accepted if no defective parts are found in the sample.
(a) How large does K have to be to ensure that the probability that the manufacturer accepts an unacceptable lot is less than .10?
(b) Suppose the manufacturer decides to accept the lot if there is at most one defective in the sample. How large does K have to be to ensure that the probability that the manufacturer accepts an unacceptable lot is less than .10?

Solution. (a) $K = 32$. (b) $K = 51$.
This is Example 3.2.1 (acceptance sampling) with $N = 100$. If the lot holds M defectives, the number X of defectives in a sample of K drawn without replacement is hypergeometric$(100, M, K)$, pmf (3.2.2), and an unacceptable lot is one with $M \ge 6$. Both acceptance rules have the form $\{X \le c\}$, and for such a rule the worst unacceptable lot is $M = 6$: couple a lot of M defectives with one of $M+1$ by nesting the defective sets, so the sampled defective count can only increase and $P(X \le c \mid M)$ is nonincreasing in M.
(a) Here $c = 0$, so the requirement is

$$P(X = 0 \mid M = 6) = \frac{\binom{94}{K}}{\binom{100}{K}} = \frac{(100-K)(99-K)\cdots(95-K)}{100 \cdot 99 \cdots 95} < .10,$$

whose left side is decreasing in K, and

$$\begin{aligned} K = 31: &\quad .1006, \\ K = 32: &\quad .0918. \end{aligned}$$

Hence $K = 32$ is the smallest sample size that works.
(b) Now $c = 1$, so at $M = 6$ the requirement is

$$P(X \le 1 \mid M = 6) = \frac{\binom{94}{K} + 6\binom{94}{K-1}}{\binom{100}{K}} < .10,$$

and

$$K = 50: \quad .1022,$$
$$K = 51: \quad .0933,$$

so $K = 51$.

Problem 3.3. The flow of traffic at certain street corners can sometimes be modeled as a sequence of Bernoulli trials by assuming that the probability of a car passing during any given second is a constant p and that there is no interaction between the passing of cars at different seconds. If we treat seconds as indivisible time units (trials), the Bernoulli model applies. Suppose a pedestrian can cross the street only if no car is to pass during the next 3 seconds. Find the probability that the pedestrian has to wait for exactly 4 seconds before starting to cross.

Solution.

$$P(W = 4) = p(1-p)^3\big[1 - (1-p)^3\big].$$

Number the seconds after the pedestrian's arrival $1, 2, 3, \ldots$ and let C_i be the event that a car passes in second i; the C_i are independent with $P(C_i) = p$. He starts to cross at the first time $t \geq 0$ at which seconds $t+1, t+2, t+3$ are all car-free, and $W = t$ is his wait. Thus $\{W = 4\}$ is the event that seconds $5, 6, 7$ are car-free while $t = 0, 1, 2, 3$ each fail:
(i) $t = 3$ fails $\iff$ a car passes in one of $4, 5, 6$; given $5, 6$ free this is C_4.
(ii) $t = 2$ and $t = 1$ then fail automatically, since second 4 lies in $\{3, 4, 5\}$ and in $\{2, 3, 4\}$.
(iii) $t = 0$ fails $\iff$ a car passes in one of $1, 2, 3$.
Hence

$$\{W = 4\} = (C_1 \cup C_2 \cup C_3) \cap C_4 \cap C_5^c \cap C_6^c \cap C_7^c,$$

an intersection of events determined by the disjoint blocks of seconds $\{1, 2, 3\}$, $\{4\}$, $\{5, 6, 7\}$, so the three probabilities multiply:

$$\begin{aligned} P(W = 4) &= \big[1 - (1-p)^3\big] \cdot p \cdot (1-p)^3 \\ &= p(1-p)^3 - p(1-p)^6. \end{aligned}$$

Problem 3.4. A man with n keys wants to open his door and tries the keys at random. Exactly one key will open the door. Find the mean number of trials if
(a) unsuccessful keys are not eliminated from further selections.
(b) unsuccessful keys are eliminated.

Solution. (a) n. (b) $(n+1)/2$.
(a) Each draw is an independent Bernoulli trial with success probability $1/n$, and T is the trial of the first success, so T is geometric$(1/n)$ (p. 97) and

$$ET = \frac{1}{1/n} = n.$$

(b) Sampling without replacement, for $k = 1, \ldots, n$,

$$\begin{aligned} P(T = k) &= \frac{n-1}{n} \cdot \frac{n-2}{n-1} \cdots \frac{n-k+1}{n-k+2} \cdot \frac{1}{n-k+1} \\ &= \frac{1}{n}, \end{aligned}$$

the first $k-1$ factors being the conditional probabilities of successive failures and the last that of success on trial k; the product telescopes. So T is discrete uniform$(1, n)$ and, by (3.2.1) (or Exercise 3.1 with $N_0 = 1$, $N_1 = n$),

$$ET = \frac{n+1}{2}.$$

Problem 3.5. A standard drug is known to be effective in 80% of the cases in which it is used. A new drug is tested on 100 patients and found to be effective in 85 cases. Is the new drug superior? (Hint: Evaluate the probability of observing 85 or more successes assuming that the new and old drugs are equally effective.)

Solution. No: under the hypothesis of equal effectiveness, 85 or more successes has probability .128, which is not at all unusual.
Let X be the number of cases in which the new drug is effective. If the two drugs are equally effective then X is binomial(100, .8), and

$$P(X \geq 85) = \sum_{x=85}^{100} \binom{100}{x} (.8)^x (.2)^{100-x} = .1285.$$

Equivalently, by the normal approximation with the continuity correction (Example 3.3.2), with $EX = 100(.8) = 80$ and $\operatorname{Var} X = 100(.8)(.2) = 16$,

$$\begin{aligned} P(X \geq 85) &\approx P\left(Z \geq \frac{84.5 - 80}{4}\right) \\ &= P(Z \geq 1.125) = .130. \end{aligned}$$

Problem 3.6. A large number of insects are expected to be attracted to a certain variety of rose plant. A commercial insecticide is advertised as being 99% effective. Suppose 2,000 insects infest a rose garden where the insecticide has been applied, and let X = number of surviving insects.
(a) What probability distribution might provide a reasonable model for this experiment?
(b) Write down, but do not evaluate, an expression for the probability that fewer than 100 insects survive, using the model in part (a).
(c) Evaluate an approximation to the probability in part (b).

Solution. (a) $X \sim$ binomial(2000, .01).
Treat the insects as independent trials, each "success" being survival of the insecticide, which has probability $1 - .99 = .01$; with $n = 2000$ fixed trials this is exactly the binomial model.
(b)

$$P(X < 100) = \sum_{x=0}^{99} \binom{2000}{x} (.01)^x (.99)^{2000-x}.$$

(c) $P(X < 100) \approx 1$.
Here n is large and p small, so the Poisson approximation of Example 3.2.5 applies with $\lambda = np = 2000(.01) = 20$:

$$P(X < 100) \approx \sum_{x=0}^{99} \frac{e^{-20} 20^x}{x!}.$$

Since 99 sits about 18 standard deviations ($\sigma = \sqrt{20} \approx 4.47$) above the mean, the omitted tail is

$$\sum_{x=100}^{\infty} \frac{e^{-20} 20^x}{x!} \approx 3.5 \times 10^{-37},$$

so the probability is 1 to any practical accuracy.

Problem 3.7. Let the number of chocolate chips in a certain type of cookie have a Poisson distribution. We want the probability that a randomly chosen cookie has at least two chocolate chips to be greater than .99. Find the smallest value of the mean of the distribution that ensures this

probability.

Solution. $\lambda = 6.6384$ chips per cookie.
Let $X \sim \text{Poisson}(\lambda)$, so $EX = \lambda$. As in Example 3.2.4,

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - e^{-\lambda} - \lambda e^{-\lambda}, \end{aligned}$$

so the requirement $P(X \geq 2) > .99$ is

$$g(\lambda) := (1 + \lambda)e^{-\lambda} < .01.$$

Since $g'(\lambda) = -\lambda e^{-\lambda} < 0$ for $\lambda > 0$, g decreases strictly from $g(0) = 1$ to 0, so the inequality holds exactly for $\lambda > \lambda^*$, where $g(\lambda^*) = .01$. Numerically,

$$\begin{aligned} g(6.63) = .010073, \qquad g(6.64) = .0099857, \\ \lambda^* = 6.63835. \end{aligned}$$

The admissible set is open, so λ^* is its infimum rather than a minimum, and the mean must exceed 6.6384.

3.2 Exercises 3.8–3.14

Problem 3.8. Two movie theaters compete for the business of 1,000 customers. Assume that each customer chooses between the movie theaters independently and with "indifference." Let N denote the number of seats in each theater.
(a) Using a binomial model, find an expression for N that will guarantee that the probability of turning away a customer (because of a full house) is less than 1%.
(b) Use the normal approximation to get a numerical value for N.

Solution. $N = 541$ seats in each theater.
(a) Let X be the number of the 1,000 customers who pick theater A; independence and indifference make $X \sim \text{binomial}(1000, \frac{1}{2})$, and theater B draws the other $1000 - X$. Somebody is turned away exactly when $X > N$ or $1000 - X > N$. For $N \geq 500$ these two events are disjoint, and the binomial$(1000, \frac{1}{2})$ pmf is symmetric about 500, so they have the same probability:

$$\begin{aligned} P(\text{turn away}) &= P(X > N) + P(X < 1000 - N) \\ &= 2 \sum_{x=N+1}^{1000} \binom{1000}{x} \left(\tfrac{1}{2}\right)^{1000}. \end{aligned}$$

Take N to be the smallest integer with

$$\sum_{x=N+1}^{1000} \binom{1000}{x} \left(\tfrac{1}{2}\right)^{1000} < .005.$$

(b) Here $\text{E}X = 1000 \cdot \frac{1}{2} = 500$ and $\text{Var}\, X = 1000 \cdot \frac{1}{2} \cdot \frac{1}{2} = 250$, so $\sigma = \sqrt{250} = 15.81$. With the continuity correction,

$$P(X > N) \approx P\Big(Z > \frac{N + .5 - 500}{15.81}\Big) < .005$$

iff $(N + .5 - 500)/15.81 > z_{.005} = 2.5758$, i.e.

$$N > 499.5 + 2.5758(15.81) = 540.2,$$

so $N = 541$.

Problem 3.9. Often, news stories that are reported as startling "one-in-a-million" coincidences are actually, upon closer examination, not rare events and can even be expected to occur. A few years ago an elementary school in New York state reported that its incoming kindergarten class contained five sets of twins. This, of course, was reported throughout the state, with a quote from the principal that this was a "statistical impossibility". Was it? Or was it an instance of what Diaconis and Mosteller (1989) call the "law of truly large numbers"? Let us do some calculations.
(a) The probability of a twin birth is approximately 1/90, and we can assume that an elementary school will have approximately 60 children entering kindergarten (three classes of 20 each). Explain how our "statistically impossible" event can be thought of as the probability of 5 or more successes from a binomial$(60, 1/90)$. Is this even rare enough to be newsworthy?
(b) Even if the probability in part (a) is rare enough to be newsworthy, consider that this could have happened in any school in the county, and in any county in the state, and it still would have been reported exactly the same. (The "law of truly large numbers" is starting to come into play.) New York state has 62 counties, and it is reasonable to assume that each county has five elementary schools. Does the event still qualify as a "statistical impossibility", or is it becoming something that could be expected to occur?
(c) If the probability in part (b) still seems small, consider further that this event could have happened in any one of the 50 states, during any of the last 10 years, and still would have received the same news coverage.
In addition to Diaconis and Mosteller (1989), see Hanley (1992) for more on coincidences.

Solution. Not impossible at all: over ten years of American kindergartens such a class is essentially certain to appear.
(a) Index the 60 entering children by their births and let each birth be a "success" if it was a twin birth, with probability 1/90; treating the 60 births as independent Bernoulli trials makes X, the number of twin births represented in the class, a binomial$(60, 1/90)$ variable, and "five sets of twins" is $\{X \geq 5\}$. Then

$$\begin{aligned} P(X \geq 5) &= \sum_{x=5}^{60} \binom{60}{x} \left(\tfrac{1}{90}\right)^x \left(\tfrac{89}{90}\right)^{60-x} \\ &= .000557, \end{aligned}$$

the sum being dominated by its first term .000500. About one chance in 1,800 – rare enough to print, but nowhere near impossible.
(b) With $62 \times 5 = 310$ schools in the state, each an independent trial with success probability $p = .000557$, the number of schools reporting the event is binomial$(310, p)$ and

$$1 - (1-p)^{310} = .159,$$

roughly one chance in six that at least one New York school has such a class this year. Already unremarkable.
(c) Over 50 states and 10 years there are $310 \times 50 \times 10 = 155{,}000$ school-years, so the expected number of such classes is $155{,}000\, p = 86.3$ and

$$1 - (1-p)^{155{,}000} = 1 - e^{-86.3} \approx 1.$$

Not a statistical impossibility but a near certainty.

Problem 3.10. Shuster (1991) describes a number of probability calculations that he did for a court case involving the sale of cocaine. A Florida police department seized 496 *suspected* packets of cocaine, of which four were randomly selected and tested and found to actually be cocaine. The police then chose two more packets at random and, posing as drug dealers, sold the packets to the defendant. These last two packets were lost before they could be tested to verify that they were, indeed, cocaine.

(a) If the original 496 packets were composed of N packets of cocaine and $M = 496 - N$ noncocaine, show that the probability of selecting 4 cocaine packets and then 2 noncocaine packets, which is the probability that the defendant is innocent of buying cocaine, is

$$\frac{\binom{N}{4}\binom{M}{2}}{\binom{N+M}{4}\binom{N+M-4}{2}}.$$

(b) Maximizing (in M and N) the probability in part (a) maximizes the defendant's "innocence probability". Show that this probability is .022, attained at $M = 165$ and $N = 331$.

Solution. (a) The four tested packets are a simple random sample of size 4 from the $N + M$ packets, so the probability that all four are cocaine is the hypergeometric probability

$$P(4 \text{ cocaine}) = \frac{\binom{N}{4}\binom{M}{0}}{\binom{N+M}{4}} = \frac{\binom{N}{4}}{\binom{N+M}{4}}.$$

Given that event, the pool left for the second draw has $N - 4$ cocaine and M noncocaine packets, a total of $N + M - 4$, and the two sold packets are a simple random sample of size 2 from it, so

$$P(2 \text{ noncocaine} \mid 4 \text{ cocaine}) = \frac{\binom{M}{2}}{\binom{N+M-4}{2}}.$$

Multiplying the two gives the stated expression.
(b) With $N+M = 496$ fixed the denominator $\binom{496}{4}\binom{492}{2}$ is constant, so we maximize $g(N) = \binom{N}{4}\binom{496-N}{2}$ over $4 \le N \le 494$. The ratio of consecutive terms is

$$\frac{g(N+1)}{g(N)} = \frac{N+1}{N-3} \cdot \frac{494-N}{496-N},$$

and $g(N+1) \ge g(N)$ iff $(N+1)(494-N) \ge (N-3)(496-N)$, i.e. iff

$$493N + 494 \ge 499N - 1488 \iff N \le \tfrac{1982}{6} = 330.33.$$

So g increases up through $N = 331$ and decreases thereafter: the maximum is at $N = 331$, $M = 165$, where

$$\frac{\binom{331}{4}\binom{165}{2}}{\binom{496}{4}\binom{492}{2}} = .0221.$$

Problem 3.11. The hypergeometric distribution can be approximated by either the binomial or the Poisson distribution. (Of course, it can be approximated by other distributions, but in this exercise we will concentrate on only these two.) Let X have the hypergeometric distribution

$$P(X = x \mid N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, \qquad x = 0, 1, \ldots, K.$$

(a) Show that as $N \to \infty$, $M \to \infty$, and $M/N \to p$,

$$P(X = x \mid N, M, K) \to \binom{K}{x} p^x (1-p)^{K-x}, \qquad x = 0, 1, \ldots, K.$$

(Stirling's Formula from Exercise 1.23 may be helpful.)
(b) Use the fact that the binomial can be approximated by the Poisson to show that if $N \to \infty$, $M \to \infty$, $K \to \infty$, $M/N \to 0$, and $KM/N \to \lambda$, then

$$P(X = x \mid N, M, K) \to \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, \ldots.$$

(c) Verify the approximation in part (b) directly, without using the Poisson approximation to the binomial. (Lemma 2.3.14 is helpful.)

Solution. Write $a^{(b)} = a(a-1)\cdots(a-b+1)$ for the falling factorial, so that $\binom{a}{b} = a^{(b)}/b!$. Cancelling the three factorial denominators against $K!/(x!(K-x)!)$ puts the pmf in the form used throughout:

$$P(X = x \mid N, M, K) = \binom{K}{x} \frac{M^{(x)}\,(N-M)^{(K-x)}}{N^{(K)}}.$$

(a) Here K and x are fixed, so split the ratio into three products of finitely many factors:

$$\frac{M^{(x)}}{N^x} = \prod_{j=0}^{x-1}\Big(\frac{M}{N} - \frac{j}{N}\Big) \to p^x,$$

$$\frac{(N-M)^{(K-x)}}{N^{K-x}} = \prod_{j=0}^{K-x-1}\Big(1 - \frac{M}{N} - \frac{j}{N}\Big) \to (1-p)^{K-x},$$

$$\frac{N^K}{N^{(K)}} = \prod_{j=0}^{K-1}\Big(1 - \frac{j}{N}\Big)^{-1} \to 1,$$

using $M/N \to p$ and $N \to \infty$ in each. Their product is exactly $M^{(x)}(N-M)^{(K-x)}/N^{(K)}$, so

$$P(X = x \mid N, M, K) \to \binom{K}{x} p^x (1-p)^{K-x}.$$

(No Stirling needed: the falling factorials are products of boundedly many factors.)
(b) By (a), for each fixed K the hypergeometric pmf is the binomial(K, p) pmf with $p = M/N$ in the limit of large N, M; here $p = M/N \to 0$ while $Kp = KM/N \to \lambda$, which is precisely the regime of the Poisson approximation to the binomial, (2.3.9), established in Example 2.3.13 (its hypotheses are exactly $p \to 0$ with $Kp \to \lambda$, and Theorem 2.3.12 converts the mgf convergence proved there into convergence of the pmfs). Hence

$$\binom{K}{x}\Big(\frac{M}{N}\Big)^x\Big(1 - \frac{M}{N}\Big)^{K-x} \to \frac{e^{-\lambda}\lambda^x}{x!},$$

and passing the two limits in this order gives the assertion.
(c) Fix x and factor the displayed pmf as

$$P(X = x \mid N, M, K) = \frac{1}{x!}\,A\,B\,C, \qquad \text{where}$$

$$A = \frac{K^{(x)}M^{(x)}}{N^x},$$

$$B = \frac{(N-M)^{(K-x)}}{N^{(K-x)}},$$

$$C = \frac{N^{(K-x)}N^x}{N^{(K)}}.$$

Note $K/N = (KM/N)/M \to 0$ since $M \to \infty$.
$A = \prod_{j=0}^{x-1} \frac{KM}{N}\big(1 - \frac{j}{K}\big)\big(1 - \frac{j}{M}\big) \to \lambda^x$, as x is fixed and $K, M \to \infty$.
$C = \prod_{j=K-x}^{K-1}\big(1 - \frac{j}{N}\big)^{-1} \to 1$, since each of these x factors lies between 1 and $(1 - K/N)^{-1} \to 1$.
$B = \prod_{j=0}^{K-x-1}\big(1 - \frac{M}{N-j}\big)$ has factors decreasing in j and positive for large N (as $(M+K)/N \to 0$), so

$$\Big(1 - \frac{M}{N-K}\Big)^{K-x} \le B \le \Big(1 - \frac{M}{N}\Big)^{K-x}.$$

Put $\lambda_K = KM/N \to \lambda$ and $\mu_K = KM/(N-K) = \lambda_K/(1-K/N) \to \lambda$. Both bounds are of the form $(1 - c_K/K)^K(1 - c_K/K)^{-x}$ with $c_K \to \lambda$, and $(1 - c_K/K)^{-x} \to 1$ because $c_K/K \to 0$; Lemma 2.3.14 (with $a_K = -c_K \to -\lambda$) gives $(1 - c_K/K)^K \to e^{-\lambda}$. So both bounds tend to $e^{-\lambda}$ and $B \to e^{-\lambda}$.
Therefore $P(X = x \mid N, M, K) \to \lambda^x e^{-\lambda}/x!$.

Problem 3.12. Suppose X has a binomial(n,p) distribution and let Y have a negative binomial(r,p) distribution. Show that $F_X(r-1) = 1 - F_Y(n-r)$.

Solution. Both sides count the same event in one sequence of trials. Run independent Bernoulli(p) trials $Z_1, Z_2, \ldots$ and set

$$X = \sum_{i=1}^{n} Z_i, \qquad Y = (\text{number of failures before the } r\text{th success}),$$

so $X \sim$ binomial(n,p) and $Y \sim$ negative binomial(r,p) in the pmf (3.2.10) convention of Section 3.2. The rth success occurs at trial $Y + r$, so the first n trials contain at least r successes if and only if $Y + r \le n$; that is,

$$\{X \le r-1\} = \{Y + r > n\} = \{Y > n - r\}.$$

Taking probabilities,

$$F_X(r-1) = P(X \le r-1) = P(Y > n-r) = 1 - F_Y(n-r).$$

Problem 3.13. A *truncated* discrete distribution is one in which a particular class cannot be observed and is eliminated from the sample space. In particular, if X has range $0, 1, 2, \ldots$ and the 0 class cannot be observed (as is usually the case), the *0-truncated* random variable X_T has pmf

$$P(X_T = x) = \frac{P(X = x)}{P(X > 0)}, \qquad x = 1, 2, \ldots.$$

Find the pmf, mean, and variance of the 0-truncated random variable starting from
(a) $X \sim$ Poisson(λ).
(b) $X \sim$ negative binomial(r,p), as in (3.2.10).

Solution. X_T is X conditioned on $\{X > 0\}$, so $\mathrm{E}\, g(X_T) = \mathrm{E}\, g(X)/P(X > 0)$ for every g with $g(0) = 0$; since x and x^2 both vanish at 0, the untruncated first two moments carry over after division by $P(X > 0)$.
(a) $P(X > 0) = 1 - e^{-\lambda}$, so

$$P(X_T = x) = \frac{e^{-\lambda}\lambda^x}{x!\,(1 - e^{-\lambda})}, \qquad x = 1, 2, \ldots.$$

With $\mathrm{E}X = \lambda$ and $\mathrm{E}X^2 = \lambda + \lambda^2$,

$$\mathrm{E}X_T = \frac{\lambda}{1 - e^{-\lambda}}, \qquad \mathrm{E}X_T^2 = \frac{\lambda + \lambda^2}{1 - e^{-\lambda}},$$

$$\begin{aligned} \operatorname{Var} X_T &= \frac{\lambda + \lambda^2}{1 - e^{-\lambda}} - \frac{\lambda^2}{(1 - e^{-\lambda})^2} \\ &= \frac{\lambda\big[(1+\lambda)(1 - e^{-\lambda}) - \lambda\big]}{(1 - e^{-\lambda})^2}. \end{aligned}$$

(b) $P(X > 0) = 1 - P(X = 0) = 1 - p^r$, so

$$P(X_T = x) = \frac{\binom{r+x-1}{x} p^r (1-p)^x}{1 - p^r}, \qquad x = 1, 2, \ldots.$$

From Section 3.2, $\mathrm{E}X = r(1-p)/p$ and $\operatorname{Var} X = r(1-p)/p^2$, whence

$$\mathrm{E}X^2 = \frac{r(1-p)}{p^2} + \frac{r^2(1-p)^2}{p^2} = \frac{r(1-p)\big[1 + r(1-p)\big]}{p^2}.$$

Therefore

$$\mathrm{E}X_T = \frac{r(1-p)}{p\,(1-p^r)}, \qquad \mathrm{E}X_T^2 = \frac{r(1-p)\big[1+r(1-p)\big]}{p^2(1-p^r)},$$

$$\begin{aligned}\operatorname{Var} X_T &= \frac{r(1-p)\big[1+r(1-p)\big]}{p^2(1-p^r)} - \frac{r^2(1-p)^2}{p^2(1-p^r)^2} \\ &= \frac{r(1-p)\big[(1+r(1-p))(1-p^r) - r(1-p)\big]}{p^2(1-p^r)^2}.\end{aligned}$$

Problem 3.14. Starting from the 0-truncated negative binomial (refer to Exercise 3.13), if we let $r \to 0$, we get an interesting distribution, the *logarithmic series distribution*. A random variable X has a logarithmic series distribution with parameter p if

$$P(X = x) = \frac{-(1-p)^x}{x \log p}, \qquad x = 1, 2, \ldots, \quad 0 < p < 1.$$

(a) Verify that this defines a legitimate probability function.
(b) Find the mean and variance of X. (The logarithmic series distribution has proved useful in modeling species abundance. See Stuart and Ord 1987 for a more detailed discussion of this distribution.)

Solution. Everything follows from the Taylor series $-\log(1-q) = \sum_{x=1}^{\infty} q^x/x$, valid for $|q| < 1$, taken at $q = 1-p$.
(a) Each term is positive: $(1-p)^x/x > 0$ and $\log p < 0$ for $0 < p < 1$, so the leading minus sign makes $P(X = x) > 0$. And

$$\sum_{x=1}^{\infty} \frac{-(1-p)^x}{x \log p} = \frac{-1}{\log p} \sum_{x=1}^{\infty} \frac{(1-p)^x}{x} = \frac{-1}{\log p}\big(-\log p\big) = 1.$$

(b) The x in the denominator cancels, leaving geometric series:

$$\mathrm{E}X = \frac{-1}{\log p} \sum_{x=1}^{\infty} (1-p)^x = \frac{-1}{\log p} \cdot \frac{1-p}{p} = \frac{-(1-p)}{p \log p},$$

$$\mathrm{E}X^2 = \frac{-1}{\log p} \sum_{x=1}^{\infty} x(1-p)^x = \frac{-1}{\log p} \cdot \frac{1-p}{p^2} = \frac{-(1-p)}{p^2 \log p},$$

using $\sum_{x\geq 1} q^x = q/(1-q)$ and $\sum_{x\geq 1} xq^x = q/(1-q)^2$ at $q = 1-p$. Hence

$$\begin{aligned}\operatorname{Var} X &= \frac{-(1-p)}{p^2 \log p} - \frac{(1-p)^2}{p^2(\log p)^2} \\ &= \frac{-(1-p)\big[\log p + (1-p)\big]}{p^2(\log p)^2}.\end{aligned}$$

3.3 Exercises 3.15–3.21

Problem 3.15. In Section 3.2 it was claimed that the Poisson(λ) distribution is the limit of the negative binomial(r, p) distribution as $r \to \infty$, $p \to 1$, and $r(1-p) \to \lambda$. Show that under these conditions the mgf of the negative binomial converges to that of the Poisson.

Solution. With $\lambda_r = r(1-p_r) \to \lambda$, so that $1 - p_r = \lambda_r/r$, the negative binomial(r, p) mgf is

$$M_r(t) \;=\; \left(\frac{p_r}{1-(1-p_r)e^t}\right)^r \;=\; \frac{\big(1-\frac{\lambda_r}{r}\big)^r}{\big(1-\frac{\lambda_r e^t}{r}\big)^r},$$

valid for all t with $(\lambda_r/r)e^t < 1$, which holds for every fixed t once r is large. Both factors are of the form $(1 + a_r/r)^r$ with a convergent sequence a_r, so Lemma 2.3.14 applies to each:

$$\begin{aligned}\left(1 - \tfrac{\lambda_r}{r}\right)^r &\longrightarrow e^{-\lambda},\\ \left(1 - \tfrac{\lambda_r e^t}{r}\right)^r &\longrightarrow e^{-\lambda e^t} \neq 0.\end{aligned}$$

Hence for every $t \in (-\infty, \infty)$,

$$M_r(t) \;\longrightarrow\; \frac{e^{-\lambda}}{e^{-\lambda e^t}} \;=\; e^{\lambda(e^t-1)},$$

the mgf of a Poisson(λ) random variable. Since the limit mgf is finite in a neighborhood of 0, Theorem 2.3.12 converts this into convergence of the distributions.

Problem 3.16. Verify these two identities regarding the gamma function that were given in the text:
(a) $\Gamma(\alpha + 1) = \alpha\,\Gamma(\alpha)$
(b) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

Solution. (a) Integrate $\Gamma(\alpha+1) = \int_0^\infty t^\alpha e^{-t}\,dt$ by parts with $u = t^\alpha$, $dv = e^{-t}dt$:

$$\Gamma(\alpha+1) = \Big[-t^\alpha e^{-t}\Big]_0^\infty + \alpha\int_0^\infty t^{\alpha-1}e^{-t}\,dt = \alpha\,\Gamma(\alpha),$$

the boundary term vanishing at 0 because $\alpha > 0$ and at ∞ because e^{-t} beats t^α.
(b) Substitute $t = z^2/2$, so $dt = z\,dz$ and $t^{-1/2} = \sqrt{2}/z$ on $z > 0$:

$$\begin{aligned}\Gamma\left(\tfrac{1}{2}\right) &= \int_0^\infty t^{-1/2}e^{-t}\,dt = \int_0^\infty \frac{\sqrt{2}}{z}\,e^{-z^2/2}\,z\,dz\\ &= \sqrt{2}\int_0^\infty e^{-z^2/2}\,dz = \sqrt{2}\cdot\sqrt{\tfrac{\pi}{2}} = \sqrt{\pi},\end{aligned}$$

the value of the last integral being (3.3.14), which the text obtains by squaring it and converting to polar coordinates.

Problem 3.17. Establish a formula similar to (3.3.18) for the gamma distribution. If $X \sim$ gamma(α, β), then for any positive constant ν,

$$\mathrm{E}X^\nu \;=\; \frac{\beta^\nu\,\Gamma(\nu+\alpha)}{\Gamma(\alpha)}.$$

(Here (3.3.18) is the beta moment formula: for $X \sim$ beta(α, β), $\mathrm{E}X^n = B(\alpha+n, \beta)/B(\alpha, \beta) = \Gamma(\alpha+n)\Gamma(\alpha+\beta)/[\Gamma(\alpha+\beta+n)\Gamma(\alpha)]$.)

Solution. Recognize the integrand as an unnormalized gamma($\nu+\alpha, \beta$) pdf, exactly as (3.3.18) recognizes a beta kernel:

$$\begin{aligned}\mathrm{E}X^\nu &= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^\nu\,x^{\alpha-1}e^{-x/\beta}\,dx\\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{(\nu+\alpha)-1}e^{-x/\beta}\,dx\\ &= \frac{\Gamma(\nu+\alpha)\,\beta^{\nu+\alpha}}{\Gamma(\alpha)\beta^\alpha} = \frac{\beta^\nu\,\Gamma(\nu+\alpha)}{\Gamma(\alpha)}.\end{aligned}$$

The third equality is (3.3.8) with α there replaced by $\nu+\alpha$, legitimate because $\nu > 0$ makes $\nu+\alpha > 0$, so gamma($\nu+\alpha, \beta$) is a genuine pdf.

Problem 3.18. There is an interesting relationship between negative binomial and gamma random variables, which may sometimes provide a useful approximation. Let Y be a negative binomial random variable with parameters r and p, where p is the success probability. Show that as $p \to 0$, the mgf of the random variable pY converges to that of a gamma distribution with parameters r and 1.

Solution. By Theorem 2.3.15 and the negative binomial(r, p) mgf (Table of Common Distributions; here Y counts failures, pmf (3.2.10)),

$$M_{pY}(t) \;=\; M_Y(pt) \;=\; \left(\frac{p}{1-(1-p)e^{pt}}\right)^r.$$

Fix $t < 1$ and set $h(p) = 1-(1-p)e^{pt}$, a smooth function of p with $h(0) = 0$ and

$$h'(p) = e^{pt} - (1-p)t\,e^{pt}, \qquad h'(0) = 1 - t.$$

Since $h'(0) = 1 - t \neq 0$, we get $p/h(p) = p/(h(p) - h(0)) \to 1/h'(0) = 1/(1-t)$ as $p \to 0$; in particular $h(p) > 0$ for all small $p > 0$, which is exactly the condition $pt < -\log(1-p)$ under which $M_Y(pt)$ exists. Therefore, with r held fixed,

$$M_{pY}(t) \;\longrightarrow\; \left(\frac{1}{1-t}\right)^r, \qquad t < 1,$$

which is the gamma(α, β) mgf $(1-\beta t)^{-\alpha}$ of Example 2.3.8 at $\alpha = r$, $\beta = 1$. As this limit is an mgf finite on a neighborhood of 0, Theorem 2.3.12 gives $pY \to$ gamma$(r, 1)$ in distribution.

Problem 3.19. Show that

$$\int_x^\infty \frac{1}{\Gamma(\alpha)} z^{\alpha-1}e^{-z}\,dz \;=\; \sum_{y=0}^{\alpha-1} \frac{x^y e^{-x}}{y!}, \qquad \alpha = 1, 2, 3, \ldots$$

(Hint: Use integration by parts.) Express this formula as a probabilistic relationship between Poisson and gamma random variables.

Solution. Induct on α, writing $I_\alpha(x) = \frac{1}{\Gamma(\alpha)}\int_x^\infty z^{\alpha-1}e^{-z}\,dz$.
Base case $\alpha = 1$: $I_1(x) = \int_x^\infty e^{-z}dz = e^{-x}$, which is the $y = 0$ term alone.
Induction step: for $\alpha \geq 2$, integrate by parts with $u = z^{\alpha-1}$, $dv = e^{-z}dz$,

$$\begin{aligned}\int_x^\infty z^{\alpha-1}e^{-z}\,dz &= \Big[-z^{\alpha-1}e^{-z}\Big]_x^\infty + (\alpha-1)\int_x^\infty z^{\alpha-2}e^{-z}\,dz \\ &= x^{\alpha-1}e^{-x} + (\alpha-1)\int_x^\infty z^{\alpha-2}e^{-z}\,dz.\end{aligned}$$

Dividing by $\Gamma(\alpha) = (\alpha-1)!$ and using $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ (Exercise 3.16(a)),

$$I_\alpha(x) \;=\; \frac{x^{\alpha-1}e^{-x}}{(\alpha-1)!} \;+\; I_{\alpha-1}(x),$$

which with the inductive hypothesis gives $I_\alpha(x) = \sum_{y=0}^{\alpha-1}\frac{x^y e^{-x}}{y!}$.
Probabilistically: the left side is $P(X > x)$ for $X \sim$ gamma$(\alpha, 1)$ and the right side is $P(Y \leq \alpha-1)$ for $Y \sim$ Poisson(x), so

$$P(X > x) \;=\; P(Y \leq \alpha-1), \qquad \text{equivalently} \qquad P(X \leq x) = P(Y \geq \alpha).$$

Rescaling X by β (so $X \sim$ gamma(α, β) and $Y \sim$ Poisson(x/β)) recovers (3.3.9) of Example 3.3.1.

Problem 3.20. Let the random variable X have the pdf

$$f(x) \;=\; \frac{2}{\sqrt{2\pi}}\, e^{-x^2/2}, \qquad 0 < x < \infty.$$

(a) Find the mean and variance of X. (This distribution is sometimes called a *folded normal*.)
(b) If X has the folded normal distribution, find the transformation $g(X) = Y$ and values of α and β so that $Y \sim \text{gamma}(\alpha, \beta)$.

Solution. (a) $\mathrm{E}X = \sqrt{2/\pi}$ and $\operatorname{Var} X = 1 - \dfrac{2}{\pi}$. For the mean, $\frac{d}{dx}\bigl(-e^{-x^2/2}\bigr) = xe^{-x^2/2}$, so

$$\mathrm{E}X = \frac{2}{\sqrt{2\pi}} \int_0^\infty x\, e^{-x^2/2}\, dx = \frac{2}{\sqrt{2\pi}} \Bigl[-e^{-x^2/2}\Bigr]_0^\infty = \frac{2}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}}.$$

For the second moment, f is exactly twice the $\mathrm{n}(0,1)$ pdf restricted to $x > 0$, and x^2 is even, so

$$\mathrm{E}X^2 = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1,$$

the variance of a $\mathrm{n}(0,1)$ variable. Hence $\operatorname{Var} X = 1 - 2/\pi \approx 0.3634$.
(b) Take $Y = g(X) = X^2$, with $\alpha = \frac{1}{2}$ and $\beta = 2$; that is, $Y \sim \chi^2_1$. The map $x \mapsto x^2$ is monotone on the support $(0, \infty)$ with inverse $y^{1/2}$ and $\frac{d}{dy} y^{1/2} = \frac{1}{2} y^{-1/2}$, so Theorem 2.1.5 gives

$$f_Y(y) = \frac{2}{\sqrt{2\pi}} e^{-y/2} \cdot \frac{1}{2} y^{-1/2} = \frac{1}{\sqrt{2\pi}}\, y^{-1/2} e^{-y/2}, \qquad y > 0.$$

Since $\Gamma(\frac{1}{2}) 2^{1/2} = \sqrt{\pi}\sqrt{2} = \sqrt{2\pi}$ by Exercise 3.16(b), this is exactly the $\text{gamma}(\frac{1}{2}, 2)$ pdf (3.3.6).

Problem 3.21. Write the integral that would define the mgf of the pdf

$$f(x) \;=\; \frac{1}{\pi}\, \frac{1}{1 + x^2}.$$

Is the integral finite? (Do you expect it to be?)

Solution. The integral is

$$M_X(t) \;=\; \int_{-\infty}^{\infty} e^{tx}\, \frac{1}{\pi}\, \frac{1}{1 + x^2}\, dx,$$

and it is infinite for every $t \neq 0$. For $t > 0$, $e^{tx}/(1 + x^2) \to \infty$ as $x \to \infty$, so the integrand does not even tend to 0 and the integral diverges; for $t < 0$ the same happens as $x \to -\infty$. Only $t = 0$ gives a finite value, namely 1. So the mgf fails to exist on any neighborhood of 0.
Yes, this is expected: f is the Cauchy pdf (3.3.19) with $\theta = 0$, and the Cauchy has no moments at all — $\mathrm{E}|X| = \infty$ by (3.3.20). A finite mgf near 0 would force all moments to be finite (Theorem 2.3.7 generates them by differentiation), so no mgf can exist here.

3.4 Exercises 3.22–3.28

Problem 3.22. For each of the following distributions, verify the formulas for $\mathrm{E}X$ and $\operatorname{Var} X$ given in the text.
(a) Verify $\operatorname{Var} X$ if X has a Poisson(λ) distribution. (*Hint:* Compute $\mathrm{E}X(X-1) = \mathrm{E}X^2 - \mathrm{E}X$.)
(b) Verify $\operatorname{Var} X$ if X has a negative binomial(r, p) distribution.
(c) Verify $\operatorname{Var} X$ if X has a gamma(α, β) distribution.
(d) Verify $\mathrm{E}X$ and $\operatorname{Var} X$ if X has a beta(α, β) distribution.

(e) Verify $\mathrm{E}X$ and $\operatorname{Var} X$ if X has a double exponential(μ, σ) distribution.

Solution. (a) The factorial moment collapses the series:

$$\mathrm{E}X(X-1) = \sum_{x=2}^{\infty} x(x-1)\,\frac{e^{-\lambda}\lambda^x}{x!} = \lambda^2 \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^{x-2}}{(x-2)!} = \lambda^2.$$

Hence $\mathrm{E}X^2 = \lambda^2 + \mathrm{E}X = \lambda^2 + \lambda$ and $\operatorname{Var} X = \lambda^2 + \lambda - \lambda^2 = \lambda$.
(b) With $q = 1 - p$ and $P(X = x) = \binom{r+x-1}{x} p^r q^x$, $x = 0, 1, \ldots$, the two identities

$$x\binom{r+x-1}{x} = r\binom{r+x-1}{x-1}, \qquad x(x-1)\binom{r+x-1}{x} = r(r+1)\binom{r+x-1}{x-2}$$

(both immediate from $\binom{r+x-1}{x} = \frac{(r+x-1)!}{x!\,(r-1)!}$) reindex each sum into a negative binomial pmf with r raised. Putting $y = x - 1$, then $y = x - 2$,

$$\mathrm{E}X = \frac{rq}{p} \sum_{y=0}^{\infty} \binom{(r+1)+y-1}{y} p^{r+1} q^y = \frac{rq}{p},$$

$$\mathrm{E}X(X-1) = \frac{r(r+1)q^2}{p^2} \sum_{y=0}^{\infty} \binom{(r+2)+y-1}{y} p^{r+2} q^y = \frac{r(r+1)q^2}{p^2}.$$

Therefore

$$\begin{aligned} \operatorname{Var} X &= \frac{r(r+1)q^2}{p^2} + \frac{rq}{p} - \frac{r^2q^2}{p^2} \\ &= \frac{rq^2 + rqp}{p^2} = \frac{r(1-p)}{p^2}. \end{aligned}$$

(c) For any n with $\alpha + n > 0$, the gamma kernel integrates to $\Gamma(\alpha+n)\beta^{\alpha+n}$, so

$$\mathrm{E}X^n = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha+n-1} e^{-x/\beta}\, dx = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} \beta^n.$$

Thus $\mathrm{E}X = \alpha\beta$, $\mathrm{E}X^2 = \alpha(\alpha+1)\beta^2$, and $\operatorname{Var} X = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2$.
(d) The same device with the beta function: $\mathrm{E}X^n = B(\alpha+n, \beta)/B(\alpha, \beta)$, and $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ gives

$$\mathrm{E}X = \frac{\alpha}{\alpha+\beta}, \qquad \mathrm{E}X^2 = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}.$$

Subtracting,

$$\begin{aligned} \operatorname{Var} X &= \frac{\alpha\big[(\alpha+1)(\alpha+\beta) - \alpha(\alpha+\beta+1)\big]}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned}$$

(e) Here $f(x) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}$ is symmetric about μ, so $\mathrm{E}X = \mu$ (the integral $\int z e^{-|z|/\sigma} dz$ converges and is odd). For the variance substitute $z = x - \mu$ and use $\int_0^\infty z^2 e^{-z/\sigma} dz = 2\sigma^3$:

$$\operatorname{Var} X = \frac{1}{2\sigma} \int_{-\infty}^{\infty} z^2 e^{-|z|/\sigma}\, dz = \frac{1}{\sigma} \int_0^\infty z^2 e^{-z/\sigma}\, dz = 2\sigma^2.$$

Problem 3.23. The *Pareto distribution*, with parameters α and β, has pdf

$$f(x) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \qquad \alpha < x < \infty, \quad \alpha > 0, \quad \beta > 0.$$

(a) Verify that $f(x)$ is a pdf.
(b) Derive the mean and variance of this distribution.
(c) Prove that the variance does not exist if $\beta \le 2$.

Solution. (a) $f \ge 0$ on (α, ∞) and

$$\int_\alpha^\infty \frac{\beta\alpha^\beta}{x^{\beta+1}}\,dx = \beta\alpha^\beta\Big[\frac{x^{-\beta}}{-\beta}\Big]_\alpha^\infty = \alpha^\beta \cdot \alpha^{-\beta} = 1.$$

(b) For $\beta > n$ the integral $\int_\alpha^\infty x^{n-\beta-1}dx$ converges and

$$\mathrm{E}X^n = \beta\alpha^\beta \int_\alpha^\infty x^{n-\beta-1}\,dx = \beta\alpha^\beta \cdot \frac{\alpha^{n-\beta}}{\beta-n} = \frac{\beta\alpha^n}{\beta-n}.$$

So $\mathrm{E}X = \beta\alpha/(\beta-1)$ for $\beta > 1$, and for $\beta > 2$,

$$\begin{aligned}\operatorname{Var} X &= \frac{\beta\alpha^2}{\beta-2} - \frac{\beta^2\alpha^2}{(\beta-1)^2} \\ &= \frac{\beta\alpha^2\big[(\beta-1)^2 - \beta(\beta-2)\big]}{(\beta-2)(\beta-1)^2} = \frac{\beta\alpha^2}{(\beta-2)(\beta-1)^2},\end{aligned}$$

the bracket being $(\beta^2 - 2\beta + 1) - (\beta^2 - 2\beta) = 1$.
(c) The second moment is

$$\mathrm{E}X^2 = \beta\alpha^\beta \int_\alpha^\infty x^{1-\beta}\,dx,$$

and $\int_\alpha^\infty x^{1-\beta}dx = \infty$ exactly when $1 - \beta \ge -1$, i.e. $\beta \le 2$ (for $\beta = 2$ the integrand is x^{-1}, whose integral diverges logarithmically). Since the integrand is nonnegative, $\mathrm{E}X^2 = \infty$, so $\operatorname{Var} X$ does not exist.

Problem 3.24. Many "named" distributions are special cases of the more common distributions already discussed. For each of the following named distributions derive the form of the pdf, verify that it is a pdf, and calculate the mean and variance.
(a) If $X \sim$ exponential(β), then $Y = X^{1/\gamma}$ has the *Weibull*(γ, β) distribution, where $\gamma > 0$ is a constant.
(b) If $X \sim$ exponential(β), then $Y = (2X/\beta)^{1/2}$ has the *Rayleigh* distribution.
(c) If $X \sim$ gamma(a, b), then $Y = 1/X$ has the *inverted gamma* IG(a, b) distribution. (This distribution is useful in Bayesian estimation of variances; see Exercise 7.23.)
(d) If $X \sim$ gamma$(\frac{3}{2}, \beta)$, then $Y = (X/\beta)^{1/2}$ has the *Maxwell* distribution.
(e) If $X \sim$ exponential(1), then $Y = \alpha - \gamma \log X$ has the *Gumbel*(α, γ) distribution, where $-\infty < \alpha < \infty$ and $\gamma > 0$. (The Gumbel distribution is also known as the *extreme value distribution*.)

Solution. Every g below is a differentiable, strictly monotone bijection of the support (increasing except in (e)), so Theorem 2.1.5 gives $f_Y(y) = f_X(g^{-1}(y))\,|\frac{d}{dy}g^{-1}(y)|$; reversing each substitution returns $\int f_X = 1$, and moments of Y are moments of X by Definition 2.2.1, using $\mathrm{E}X^n = \Gamma(\alpha+n)\beta^n/\Gamma(\alpha)$ for $n > -\alpha$ (Exercise 3.22(c)).
(a) $g^{-1}(y) = y^\gamma$, $g^{-1\prime}(y) = \gamma y^{\gamma-1}$, so for $y > 0$

$$f_Y(y) = \frac{1}{\beta}e^{-y^\gamma/\beta}\,\gamma y^{\gamma-1} = \frac{\gamma}{\beta}y^{\gamma-1}e^{-y^\gamma/\beta},$$

the Weibull(γ, β) pdf; $u = y^\gamma/\beta$ returns $\int_0^\infty e^{-u}du = 1$. With $\alpha = 1$ in the gamma moment formula,

$$\begin{aligned}\mathrm{E}Y &= \mathrm{E}X^{1/\gamma} = \beta^{1/\gamma}\Gamma(1+\tfrac{1}{\gamma}), \\ \operatorname{Var} Y &= \beta^{2/\gamma}\Big[\Gamma\big(1+\tfrac{2}{\gamma}\big) - \Gamma^2\big(1+\tfrac{1}{\gamma}\big)\Big].\end{aligned}$$

(b) $g^{-1}(y) = \beta y^2/2$, $g^{-1\prime}(y) = \beta y$, so for $y > 0$

$$f_Y(y) = \frac{1}{\beta} e^{-y^2/2}\, \beta y = y\, e^{-y^2/2},$$

free of β, and $\int_0^\infty y e^{-y^2/2} dy = [-e^{-y^2/2}]_0^\infty = 1$. The integrand below is even, and $\int_{-\infty}^\infty y^2 e^{-y^2/2} dy = \sqrt{2\pi}$ is the variance of an $\mathrm{n}(0,1)$ times $\sqrt{2\pi}$, so

$$\mathrm{E}Y = \int_0^\infty y^2 e^{-y^2/2} dy = \tfrac{1}{2} \int_{-\infty}^\infty y^2 e^{-y^2/2} dy = \sqrt{\tfrac{\pi}{2}},$$

while $\mathrm{E}Y^2 = \mathrm{E}(2X/\beta) = 2$ directly. Hence $\operatorname{Var} Y = 2 - \pi/2$.
(c) $g^{-1}(y) = 1/y$, $|g^{-1\prime}(y)| = 1/y^2$, so for $y > 0$

$$f_Y(y) = \frac{1}{\Gamma(a) b^a} \Big(\frac{1}{y}\Big)^{a-1} e^{-1/(by)} \cdot \frac{1}{y^2} = \frac{1}{\Gamma(a) b^a} \Big(\frac{1}{y}\Big)^{a+1} e^{-1/(by)},$$

the $\mathrm{IG}(a,b)$ pdf; $u = 1/y$ carries the integral back to the gamma(a,b) pdf, so it is 1. Negative moments of the gamma exist for $n < a$:

$$\mathrm{E}Y = \mathrm{E}X^{-1} = \frac{\Gamma(a-1) b^{-1}}{\Gamma(a)} = \frac{1}{b(a-1)} \quad (a > 1),$$
$$\mathrm{E}Y^2 = \mathrm{E}X^{-2} = \frac{\Gamma(a-2) b^{-2}}{\Gamma(a)} = \frac{1}{b^2(a-1)(a-2)} \quad (a > 2),$$

whence

$$\operatorname{Var} Y = \frac{1}{b^2(a-1)} \Big[\frac{1}{a-2} - \frac{1}{a-1}\Big] = \frac{1}{b^2(a-1)^2(a-2)}, \qquad a > 2.$$

(d) $g^{-1}(y) = \beta y^2$, $g^{-1\prime}(y) = 2\beta y$, and $\Gamma(\frac{3}{2}) = \sqrt{\pi}/2$, so for $y > 0$

$$f_Y(y) = \frac{(\beta y^2)^{1/2} e^{-y^2}}{\Gamma(\frac{3}{2}) \beta^{3/2}} \cdot 2\beta y = \frac{2 y^2 e^{-y^2}}{\Gamma(\frac{3}{2})} = \frac{4}{\sqrt{\pi}}\, y^2 e^{-y^2},$$

the Maxwell pdf, again free of β (and it integrates to 1, reversing the substitution). From the gamma moment formula with $\alpha = \frac{3}{2}$,

$$\mathrm{E}Y = \beta^{-1/2} \mathrm{E}X^{1/2} = \beta^{-1/2} \frac{\Gamma(2) \beta^{1/2}}{\Gamma(\frac{3}{2})} = \frac{2}{\sqrt{\pi}}, \qquad \mathrm{E}Y^2 = \frac{\mathrm{E}X}{\beta} = \frac{3}{2},$$

so $\operatorname{Var} Y = \frac{3}{2} - 4/\pi$.
(e) $g^{-1}(y) = e^{(\alpha - y)/\gamma}$, $|g^{-1\prime}(y)| = \frac{1}{\gamma} e^{(\alpha-y)/\gamma}$, so for $-\infty < y < \infty$

$$f_Y(y) = \frac{1}{\gamma}\, e^{(\alpha-y)/\gamma} \exp\big(-e^{(\alpha-y)/\gamma}\big),$$

the Gumbel(α,γ) pdf, with antiderivative $\exp(-e^{(\alpha-y)/\gamma})$ running from 0 to 1. Differentiating $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$ under the integral at $t = 1$ gives $\mathrm{E}\log X = \Gamma'(1) = -\gamma_E$ and $\mathrm{E}(\log X)^2 = \Gamma''(1) = \gamma_E^2 + \pi^2/6$, where $\gamma_E \approx 0.5772$ is Euler's constant; hence $\operatorname{Var}(\log X) = \pi^2/6$ and

$$\mathrm{E}Y = \alpha + \gamma\gamma_E,$$
$$\operatorname{Var} Y = \gamma^2 \operatorname{Var}(\log X) = \frac{\gamma^2 \pi^2}{6}.$$

Problem 3.25. Suppose the random variable T is the length of life of an object (possibly the lifetime of an electrical component or of a subject given a particular treatment). The *hazard function* $h_T(t)$ associated with the random variable T is defined by

$$h_T(t) = \lim_{\delta \to 0} \frac{P(t \le T < t+\delta \mid T \ge t)}{\delta}.$$

Thus, we can interpret $h_T(t)$ as the rate of change of the probability that the object survives a little past time t, given that the object survives to time t. Show that if T is a continuous random variable, then

$$h_T(t) = \frac{f_T(t)}{1 - F_T(t)} = -\frac{d}{dt} \log\big(1 - F_T(t)\big).$$

Solution. Since T is continuous, $P(T \ge t) = 1 - F_T(t)$ and $P(t \le T < t+\delta) = F_T(t+\delta) - F_T(t)$, so for any t with $F_T(t) < 1$

$$\begin{aligned} h_T(t) &= \lim_{\delta \to 0} \frac{1}{\delta} \cdot \frac{F_T(t+\delta) - F_T(t)}{1 - F_T(t)} \\ &= \frac{1}{1 - F_T(t)} \lim_{\delta \to 0} \frac{F_T(t+\delta) - F_T(t)}{\delta} \\ &= \frac{F_T'(t)}{1 - F_T(t)} = \frac{f_T(t)}{1 - F_T(t)}, \end{aligned}$$

the last step by (1.6.2), whose hypothesis holds at every continuity point of f_T. The second equality is the chain rule applied to $\log(1 - F_T(t))$:

$$-\frac{d}{dt} \log\big(1 - F_T(t)\big) = -\frac{-f_T(t)}{1 - F_T(t)} = \frac{f_T(t)}{1 - F_T(t)}.$$

Problem 3.26. Verify that the following pdfs have the indicated hazard functions (see Exercise 3.25).
(a) If $T \sim$ exponential(β), then $h_T(t) = 1/\beta$.
(b) If $T \sim$ Weibull(γ, β), then $h_T(t) = (\gamma/\beta)t^{\gamma-1}$.
(c) If $T \sim$ logistic(μ, β), that is,

$$F_T(t) = \frac{1}{1 + e^{-(t-\mu)/\beta}},$$

then $h_T(t) = (1/\beta)F_T(t)$.

Solution. In each case substitute into $h_T = f_T/(1 - F_T)$ from Exercise 3.25.
(a) $1 - F_T(t) = e^{-t/\beta}$ and $f_T(t) = \frac{1}{\beta} e^{-t/\beta}$ for $t > 0$, so

$$h_T(t) = \frac{\frac{1}{\beta} e^{-t/\beta}}{e^{-t/\beta}} = \frac{1}{\beta}.$$

(b) From Exercise 3.24(a), $f_T(t) = \frac{\gamma}{\beta} t^{\gamma-1} e^{-t^\gamma/\beta}$ for $t > 0$, and $\int_t^\infty f_T = e^{-t^\gamma/\beta}$ gives $1 - F_T(t) = e^{-t^\gamma/\beta}$. Hence

$$h_T(t) = \frac{(\gamma/\beta) t^{\gamma-1} e^{-t^\gamma/\beta}}{e^{-t^\gamma/\beta}} = \frac{\gamma}{\beta} t^{\gamma-1}.$$

(c) Write $u = e^{-(t-\mu)/\beta}$, so $F_T(t) = (1+u)^{-1}$ and $du/dt = -u/\beta$. Then

$$\begin{aligned} f_T(t) &= \frac{d}{dt}(1+u)^{-1} = -\frac{1}{(1+u)^2} \cdot \frac{-u}{\beta} = \frac{1}{\beta} \cdot \frac{u}{(1+u)^2}, \\ 1 - F_T(t) &= 1 - \frac{1}{1+u} = \frac{u}{1+u}, \end{aligned}$$

and therefore

$$h_T(t) = \frac{u/\big[\beta(1+u)^2\big]}{u/(1+u)} = \frac{1}{\beta}\cdot\frac{1}{1+u} = \frac{1}{\beta}F_T(t).$$

Problem 3.27. For each of the following families, show whether all the pdfs in the family are unimodal (see Exercise 2.27).
(a) uniform(a, b)
(b) gamma(α, β)
(c) n(μ, σ^2)
(d) beta(α, β)

Solution. Yes for (a) and (c); no for (b) and (d). Throughout, "unimodal with mode m" is Exercise 2.27's condition: f is nondecreasing on $(-\infty, m]$ and nonincreasing on $[m, \infty)$. In particular a unimodal pdf is bounded, by $f(m)$ — that is what kills (b) and (d).
(a) Every uniform(a, b) pdf is unimodal, with every point of $[a, b]$ a mode. Take $m = (a + b)/2$: for $y < x \le m$ the values $f(y), f(x) \in \{0, (b-a)^{-1}\}$ satisfy $f(y) \le f(x)$, since f takes the value 0 only to the left of a; symmetrically on $[m, \infty)$.
(b) Not all. Differentiating $f(x) = x^{\alpha-1}e^{-x/\beta}/(\Gamma(\alpha)\beta^\alpha)$ on $x > 0$,

$$f'(x) = \frac{x^{\alpha-2}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}\Big[(\alpha-1) - \frac{x}{\beta}\Big],$$

whose bracket is positive for $x < \beta(\alpha-1)$ and negative for $x > \beta(\alpha-1)$.

- (i) $\alpha > 1$: unimodal with mode $m = \beta(\alpha-1)$.
- (ii) $\alpha = 1$: $f(x) = \beta^{-1}e^{-x/\beta}$ is nonincreasing on $[0, \infty)$ and vanishes to the left, so it is unimodal with $m = 0$.
- (iii) $\alpha < 1$: $f(x) \to \infty$ as $x \to 0^+$, so f is unbounded and no m can satisfy $f(m) \ge f(x)$ for all x. Not unimodal.

(c) All are unimodal, with mode $m = \mu$: since

$$f'(x) = -\frac{x-\mu}{\sigma^2}\,f(x),$$

$f' > 0$ on $(-\infty, \mu)$ and $f' < 0$ on (μ, ∞).
(d) Not all; the beta(α, β) pdf is unimodal exactly when $\alpha \ge 1$ and $\beta \ge 1$. On $0 < x < 1$,

$$f'(x) = \frac{x^{\alpha-2}(1-x)^{\beta-2}}{B(\alpha,\beta)}\big[(\alpha-1) - (\alpha+\beta-2)x\big].$$

- (i) $\alpha > 1,\ \beta > 1$: $\alpha + \beta - 2 > 0$, so the bracket changes sign once, from + to −; unimodal with $m = (\alpha-1)/(\alpha+\beta-2)$.
- (ii) $\alpha = 1,\ \beta \ge 1$ (or $\alpha \ge 1,\ \beta = 1$): f is monotone on $(0, 1)$ and bounded, hence unimodal with $m = 0$ (resp. $m = 1$).
- (iii) $\alpha < 1$: $f(x) \to \infty$ as $x \to 0^+$, so f is unbounded and not unimodal; symmetrically for $\beta < 1$ as $x \to 1^-$.

Problem 3.28. Show that each of the following families is an exponential family.
(a) normal family with either parameter μ or σ known
(b) gamma family with either parameter α or β known or both unknown
(c) beta family with either parameter α or β known or both unknown
(d) Poisson family
(e) negative binomial family with r known, $0 < p < 1$

Solution. In each case pull the parameter-free factors into h, the x-free factors into c, and write the remaining x-and-parameter cross terms as $\exp(\sum_i w_i(\theta)t_i(x))$ — Definition 3.4.1. The algebraic rewrites are:

$$\text{(a) } \sigma \text{ known: } \frac{e^{-x^2/(2\sigma^2)}}{\sqrt{2\pi}\,\sigma} \cdot e^{-\mu^2/(2\sigma^2)} \exp\Big(\frac{\mu}{\sigma^2}\,x\Big),$$

$$\text{(a) } \mu \text{ known: } \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2\Big),$$

$$\text{(b) } \frac{1}{\Gamma(\alpha)\beta^\alpha} \exp\Big((\alpha-1)\log x - \frac{1}{\beta}x\Big), \qquad x>0,$$

$$\text{(c) } \frac{1}{B(\alpha,\beta)} \exp\big((\alpha-1)\log x + (\beta-1)\log(1-x)\big), \quad 0<x<1,$$

$$\text{(d) } \frac{1}{x!} \cdot e^{-\lambda} \exp\big(x\log\lambda\big), \quad x=0,1,2,\dots,$$

$$\text{(e) } \binom{r+x-1}{x} \cdot p^r \exp\big(x\log(1-p)\big), \qquad x=0,1,2,\dots$$

Reading off h, c, w_i, t_i (a factor already free of the unknown parameter moves into h when the other parameter is known) gives the following, with $k=1$ except in the two "both unknown" rows, where $k=2$:

case	$h(x)$	$c(\theta)$	$w_1(\theta)$	$t_1(x)$	$w_2(\theta)$	$t_2(x)$
(a) normal, σ known	$\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}$	$e^{-\mu^2/(2\sigma^2)}$	μ/σ^2	x	—	—
(a) normal, μ known	$1/\sqrt{2\pi}$	$1/\sigma$	$-1/(2\sigma^2)$	$(x-\mu)^2$	—	—
(b) gamma, α known	$x^{\alpha-1}I_{(0,\infty)}(x)$	$1/(\Gamma(\alpha)\beta^\alpha)$	$-1/\beta$	x	—	—
(b) gamma, β known	$e^{-x/\beta}I_{(0,\infty)}(x)$	$1/(\Gamma(\alpha)\beta^\alpha)$	$\alpha-1$	$\log x$	—	—
(b) gamma, both unknown	$I_{(0,\infty)}(x)$	$1/(\Gamma(\alpha)\beta^\alpha)$	$\alpha-1$	$\log x$	$-1/\beta$	x
(c) beta, β known	$(1-x)^{\beta-1}I_{(0,1)}(x)$	$1/B(\alpha,\beta)$	$\alpha-1$	$\log x$	—	—
(c) beta, α known	$x^{\alpha-1}I_{(0,1)}(x)$	$1/B(\alpha,\beta)$	$\beta-1$	$\log(1-x)$	—	—
(c) beta, both unknown	$I_{(0,1)}(x)$	$1/B(\alpha,\beta)$	$\alpha-1$	$\log x$	$\beta-1$	$\log(1-x)$
(d) Poisson(λ)	$1/x!$	$e^{-\lambda}$	$\log\lambda$	x	—	—
(e) negative binomial, r known	$\binom{r+x-1}{x}$	p^r	$\log(1-p)$	x	—	—

Definition 3.4.1's hypotheses hold in every row: $h \ge 0$ and $c \ge 0$; each w_i depends only on the parameter and each t_i only on x; and the support — $\mathbb{R}$, $(0,\infty)$, $(0,1)$, or $\{0,1,2,\dots\}$ — is the same for all parameter values, being carried by h alone.

3.5 Exercises 3.29–3.35

Problem 3.29. For each family in Exercise 3.28, describe the natural parameter space. (Exercise 3.28 asks one to show that each of the following families is an exponential family.)
(a) normal family with either μ or σ known
(b) gamma family with either α or β known or both unknown
(c) beta family with either α or β known or both unknown
(d) Poisson family
(e) negative binomial family with r known, $0<p<1$

Solution. Only one integral has to be tested in each case, since by the definition following (3.4.7),

$$\mathcal{H} = \Big\{\boldsymbol{\eta} : \int h(x)\exp\Big(\sum_{i=1}^{k}\eta_i t_i(x)\Big)dx < \infty\Big\},$$

with the integral replaced by a sum over $\{x : h(x)>0\}$ in the discrete cases.
(a) Normal. With σ known, $h(x)=e^{-x^2/(2\sigma^2)}$, $t_1(x)=x$ and $\eta_1=\mu/\sigma^2$; the Gaussian factor dominates $e^{\eta_1 x}$ for every real η_1, so

$$\mathcal{H} = (-\infty,\infty).$$

With μ known, $h(x) = 1$, $t_1(x) = (x-\mu)^2$ and $\eta_1 = -1/(2\sigma^2)$; $\int e^{\eta_1(x-\mu)^2}dx < \infty$ if and only if $\eta_1 < 0$, so $\mathcal{H} = (-\infty, 0)$. (With both unknown this is Example 3.4.6: $\mathcal{H} = \{(\eta_1, \eta_2) : \eta_1 > 0,\ -\infty < \eta_2 < \infty\}$.)
(b) Gamma. Write $f(x \mid \alpha, \beta) = h(x)c(\alpha, \beta)\exp\{(\alpha-1)\log x - x/\beta\}$ with $h(x) = I_{(0,\infty)}(x)$, $t_1(x) = \log x$, $t_2(x) = x$, $\eta_1 = \alpha - 1$, $\eta_2 = -1/\beta$. Both unknown:

$$\int_0^\infty x^{\eta_1} e^{\eta_2 x}\, dx < \infty \iff \eta_1 > -1 \text{ and } \eta_2 < 0,$$

the first condition for integrability at 0, the second at ∞; so $\mathcal{H} = (-1, \infty) \times (-\infty, 0)$. With β known the factor $e^{-x/\beta}$ joins h, leaving $\mathcal{H} = \{\eta_1 : \eta_1 > -1\}$; with α known, $x^{\alpha-1}$ joins h and $\mathcal{H} = \{\eta_2 : \eta_2 < 0\}$.
(c) Beta. Here $h(x) = I_{(0,1)}(x)$, $t_1(x) = \log x$, $t_2(x) = \log(1-x)$, $\eta_1 = \alpha - 1$, $\eta_2 = \beta - 1$, and

$$\int_0^1 x^{\eta_1}(1-x)^{\eta_2} dx < \infty \iff \eta_1 > -1,\ \eta_2 > -1,$$

so with both unknown $\mathcal{H} = (-1, \infty) \times (-1, \infty)$. With β known, $(1-x)^{\beta-1}$ is absorbed into h and $\mathcal{H} = (-1, \infty)$; symmetrically for α known.
(d) Poisson. $h(x) = 1/x!$ on $x = 0, 1, 2, \ldots$, $t_1(x) = x$, $\eta_1 = \log\lambda$, and $\sum_{x\ge 0} e^{\eta_1 x}/x! = e^{e^{\eta_1}} < \infty$ for every real η_1. Hence $\mathcal{H} = (-\infty, \infty)$.
(e) Negative binomial, r known. $h(x) = \binom{r+x-1}{x}$ on $x = 0, 1, 2, \ldots$, $t_1(x) = x$, $\eta_1 = \log(1-p)$, and

$$\sum_{x\ge 0} \binom{r+x-1}{x} e^{\eta_1 x} = (1 - e^{\eta_1})^{-r} < \infty \iff e^{\eta_1} < 1,$$

so $\mathcal{H} = (-\infty, 0)$.

Problem 3.30. Use the identities of Theorem 3.4.2 to
(a) calculate the variance of a binomial random variable.
(b) calculate the mean and variance of a Poisson(λ) random variable.

Solution. (a) $\operatorname{Var} X = np(1-p)$. From Example 3.4.1, $w_1(p) = \log\frac{p}{1-p}$, $t_1(x) = x$, $c(p) = (1-p)^n$, and

$$\frac{d}{dp}w_1(p) = \frac{1}{p(1-p)}, \qquad \frac{d^2}{dp^2}w_1(p) = \frac{2p-1}{p^2(1-p)^2},$$
$$\frac{d}{dp}\log c(p) = \frac{-n}{1-p}, \qquad \frac{d^2}{dp^2}\log c(p) = \frac{-n}{(1-p)^2}.$$

Example 3.4.3 has already delivered $\mathrm{E}X = np$ from (3.4.4), so (3.4.5) reads

$$\operatorname{Var}\Big(\frac{X}{p(1-p)}\Big) = \frac{n}{(1-p)^2} - \frac{2p-1}{p^2(1-p)^2}\, np.$$

Multiplying through by $p^2(1-p)^2$,

$$\operatorname{Var} X = np^2 - np(2p-1) = np(1-p).$$

(b) $\mathrm{E}X = \operatorname{Var} X = \lambda$. Writing $f(x \mid \lambda) = \frac{1}{x!}e^{-\lambda}\exp(x\log\lambda)$ identifies $h(x) = 1/x!$, $c(\lambda) = e^{-\lambda}$, $w_1(\lambda) = \log\lambda$, $t_1(x) = x$, whence $w_1' = 1/\lambda$, $w_1'' = -1/\lambda^2$, $\log c(\lambda) = -\lambda$, $(\log c)' = -1$, $(\log c)'' = 0$. Then (3.4.4) gives

$$\mathrm{E}\Big(\frac{X}{\lambda}\Big) = 1 \quad\Longrightarrow\quad \mathrm{E}X = \lambda,$$

and (3.4.5) gives

$$\operatorname{Var}\Big(\frac{X}{\lambda}\Big) = -0 - \mathrm{E}\Big(-\frac{X}{\lambda^2}\Big) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda},$$

so $\operatorname{Var} X = \lambda^2 \cdot \lambda^{-1} = \lambda$.

Problem 3.31. In this exercise we will prove Theorem 3.4.2.
(a) Start from the equality

$$\int h(x)c(\boldsymbol{\theta})\exp\Big(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\Big)dx = 1,$$

differentiate both sides, and then rearrange terms to establish (3.4.4). (The fact that $\frac{d}{dx}\log g(x) = g'(x)/g(x)$ will be helpful.)
(b) Differentiate the above equality a second time; then rearrange to establish (3.4.5). (The fact that $\frac{d^2}{dx^2}\log g(x) = (g''(x)/g(x)) - (g'(x)/g(x))^2$ will be helpful.)

Solution. Fix x and j and abbreviate the $\boldsymbol{\theta}$-dependent factor of the density by

$$g(\boldsymbol{\theta}) = c(\boldsymbol{\theta})\exp\Big(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\Big), \qquad f(x \mid \boldsymbol{\theta}) = h(x)g(\boldsymbol{\theta}),$$

so that, by the hint,

$$\begin{aligned}\frac{\partial g/\partial\theta_j}{g} &= \frac{\partial}{\partial\theta_j}\log g \\ &= \underbrace{\frac{\partial}{\partial\theta_j}\log c(\boldsymbol{\theta})}_{=:L} + \underbrace{\sum_{i=1}^{k}\frac{\partial w_i(\boldsymbol{\theta})}{\partial\theta_j}t_i(x)}_{=:A(x)}.\end{aligned}$$

Throughout, the derivative passes inside the integral: on the interior of the natural parameter space the integral in (3.4.7) converges uniformly on compact $\boldsymbol{\eta}$-neighbourhoods, which is exactly the hypothesis needed in Section 2.4 to interchange $\partial/\partial\theta_j$ and $\int$.
(a) Differentiating $\int h(x)g(\boldsymbol{\theta})dx = 1$ with respect to θ_j and writing $\partial g/\partial\theta_j = g\cdot(L + A(x))$,

$$0 = \int h(x)\frac{\partial g}{\partial\theta_j}dx = \int f(x \mid \boldsymbol{\theta})\big(L + A(x)\big)dx = L + \mathrm{E}\,A(X),$$

since $\int f = 1$ and L does not depend on x. That is,

$$\mathrm{E}\Big(\sum_{i=1}^{k}\frac{\partial w_i(\boldsymbol{\theta})}{\partial\theta_j}t_i(X)\Big) = -\frac{\partial}{\partial\theta_j}\log c(\boldsymbol{\theta}),$$

which is (3.4.4).
(b) The second hint, in the form $g''/g = (\log g)'' + ((\log g)')^2$, gives

$$\frac{\partial^2 g/\partial\theta_j^2}{g} = \frac{\partial^2}{\partial\theta_j^2}\log c(\boldsymbol{\theta}) + \sum_{i=1}^{k}\frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial\theta_j^2}t_i(x) + \big(L + A(x)\big)^2.$$

Differentiating $\int h(x)g(\boldsymbol{\theta})dx = 1$ twice and multiplying the integrand by g/g therefore yields

$$\begin{aligned}0 &= \int h(x)\frac{\partial^2 g}{\partial\theta_j^2}dx \\ &= \frac{\partial^2}{\partial\theta_j^2}\log c(\boldsymbol{\theta}) + \mathrm{E}\Big(\sum_{i=1}^{k}\frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial\theta_j^2}t_i(X)\Big) + \mathrm{E}\big(L + A(X)\big)^2.\end{aligned}$$

By part (a) $\mathrm{E}(L + A(X)) = 0$, so the last term is $\mathrm{Var}(L + A(X)) = \mathrm{Var}\,A(X)$, L being constant in x. Rearranging,

$$\mathrm{Var}\Big(\sum_{i=1}^{k}\frac{\partial w_i(\boldsymbol{\theta})}{\partial\theta_j}t_i(X)\Big) = -\frac{\partial^2}{\partial\theta_j^2}\log c(\boldsymbol{\theta}) - \mathrm{E}\Big(\sum_{i=1}^{k}\frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial\theta_j^2}t_i(X)\Big),$$

which is (3.4.5).

Problem 3.32. (a) If an exponential family can be written in the form (3.4.7), show that the identities of Theorem 3.4.2 simplify to

$$\mathrm{E}(t_j(X)) = -\frac{\partial}{\partial \eta_j} \log c^*(\boldsymbol{\eta}),$$
$$\mathrm{Var}(t_j(X)) = -\frac{\partial^2}{\partial \eta_j^2} \log c^*(\boldsymbol{\eta}).$$

(b) Use this identity to calculate the mean and variance of a gamma(a, b) random variable.

Solution. (a) The natural parameterization (3.4.7) is (3.4.1) with $\boldsymbol{\theta} = \boldsymbol{\eta}$, $c = c^*$ and $w_i(\boldsymbol{\eta}) = \eta_i$, so the derivatives appearing in Theorem 3.4.2 collapse:

$$\frac{\partial w_i(\boldsymbol{\eta})}{\partial \eta_j} = \delta_{ij}, \qquad \frac{\partial^2 w_i(\boldsymbol{\eta})}{\partial \eta_j^2} = 0.$$

The sum in (3.4.4) is then the single term $t_j(X)$, giving $\mathrm{E}(t_j(X)) = -\frac{\partial}{\partial \eta_j} \log c^*(\boldsymbol{\eta})$, and in (3.4.5) the expectation term vanishes, leaving $\mathrm{Var}(t_j(X)) = -\frac{\partial^2}{\partial \eta_j^2} \log c^*(\boldsymbol{\eta})$.
(b) $\mathrm{E}X = ab$ and $\mathrm{Var}\, X = ab^2$. The gamma pdf in the form (3.4.7) has

$$h(x) = I_{(0,\infty)}(x), \quad t_1(x) = \log x, \quad t_2(x) = x, \quad \eta_1 = a - 1, \quad \eta_2 = -\frac{1}{b},$$

and since $\int_0^\infty x^{\eta_1} e^{\eta_2 x} dx = \Gamma(\eta_1 + 1)(-\eta_2)^{-(\eta_1+1)}$ on the natural parameter space $\eta_1 > -1$, $\eta_2 < 0$ (Exercise 3.29(b)),

$$\log c^*(\boldsymbol{\eta}) = (\eta_1 + 1)\log(-\eta_2) - \log\Gamma(\eta_1 + 1).$$

Using $\partial \log(-\eta_2)/\partial \eta_2 = 1/\eta_2$ and $t_2(x) = x$,

$$\mathrm{E}X = -\frac{\partial}{\partial \eta_2} \log c^* = -\frac{\eta_1 + 1}{\eta_2} = -\frac{a}{-1/b} = ab,$$
$$\mathrm{Var}\, X = -\frac{\partial^2}{\partial \eta_2^2} \log c^* = \frac{\eta_1 + 1}{\eta_2^2} = \frac{a}{1/b^2} = ab^2.$$

Problem 3.33. For each of the following families:
(i) Verify that it is an exponential family.
(ii) Describe the curve on which the $\boldsymbol{\theta}$ parameter vector lies.
(iii) Sketch a graph of the curved parameter space.
(a) n(θ, θ)
(b) n$(\theta, a\theta^2)$, a known
(c) gamma$(\alpha, 1/\alpha)$
(d) $f(x \mid \theta) = C \exp(-(x - \theta)^4)$, C a normalizing constant

Solution. All four are curved exponential families in the sense of Definition 3.4.7: $d = 1$ while $k = 2, 2, 2, 3$.
(a) Completing the square in the exponent,

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\Big(-\frac{x^2}{2\theta} + x - \frac{\theta}{2}\Big), \qquad \theta > 0,$$

which is (3.4.1) with $h(x) = e^x$, $c(\theta) = (2\pi\theta)^{-1/2} e^{-\theta/2}$, $w_1(\theta) = 1/\theta$, $t_1(x) = -x^2/2$ (and, in the Example 3.4.4 parameterization, $w_2 = \mu/\sigma^2 = 1$, $t_2(x) = x$). Taking $\boldsymbol{\theta} = (\mu, \sigma^2)$ as in Example 3.4.8, the parameter vector is (θ, θ): the curve is the line

$$\sigma^2 = \mu, \qquad \mu > 0,$$

i.e. the open 45° ray out of the origin in the first quadrant of the (μ, σ^2) plane, a one-dimensional subset of the full space $\mathcal{R} \times (0, \infty)$.
(b) Here $a > 0$ and $\theta \neq 0$ (else $\sigma^2 = 0$), and the verification is Example 3.4.4 with $\mu = \theta$, $\sigma^2 = a\theta^2$:

$$w_1 = \frac{1}{a\theta^2}, \quad t_1(x) = -\frac{x^2}{2}, \qquad w_2 = \frac{1}{a\theta}, \quad t_2(x) = x.$$

The parameter vector $(\mu, \sigma^2) = (\theta, a\theta^2)$ traces the parabola

$$\sigma^2 = a\mu^2, \qquad \mu \neq 0,$$

a parabola through the origin with the vertex deleted, opening upward and narrowing as a increases; $a = 1$ is Example 3.4.8.
(c) With $\beta = 1/\alpha$,

$$f(x \mid \alpha) = \frac{\alpha^\alpha}{\Gamma(\alpha)} \exp\big((\alpha - 1)\log x - \alpha x\big) I_{(0,\infty)}(x),$$

so $h(x) = I_{(0,\infty)}(x)$, $c(\alpha) = \alpha^\alpha/\Gamma(\alpha)$, $w_1(\alpha) = \alpha - 1$, $t_1(x) = \log x$, $w_2(\alpha) = -\alpha$, $t_2(x) = x$. The vector (α, β) lies on

$$\alpha\beta = 1, \qquad \alpha > 0,$$

the branch of the rectangular hyperbola $\beta = 1/\alpha$ in the open first quadrant, i.e. exactly the gamma densities with mean 1.
(d) Expanding $(x - \theta)^4 = x^4 - 4\theta x^3 + 6\theta^2 x^2 - 4\theta^3 x + \theta^4$,

$$f(x \mid \theta) = e^{-x^4} \cdot Ce^{-\theta^4} \exp\big(4\theta x^3 - 6\theta^2 x^2 + 4\theta^3 x\big),$$

an exponential family with $k = 3$: $h(x) = e^{-x^4}$, $c(\theta) = Ce^{-\theta^4}$ (with $C = \big(\int e^{-u^4} du\big)^{-1} = 2/\Gamma(1/4) \approx 0.5516$), and

$$\begin{aligned} w_1(\theta) &= 4\theta, & t_1(x) &= x^3, \\ w_2(\theta) &= -6\theta^2, & t_2(x) &= x^2, \\ w_3(\theta) &= 4\theta^3, & t_3(x) &= x. \end{aligned}$$

The parameter vector $(w_1, w_2, w_3) = (4\theta, -6\theta^2, 4\theta^3)$, $-\infty < \theta < \infty$, is a twisted cubic in $\mathcal{R}^3$: eliminating $\theta = w_1/4$ gives

$$w_2 = -\tfrac{3}{8} w_1^2, \qquad w_3 = \tfrac{1}{16} w_1^3.$$

To sketch it: the curve lies in the downward parabolic cylinder $w_2 = -3w_1^2/8$ and runs along it with w_3 increasing cubically, through the origin at $\theta = 0$.

Problem 3.34. In Example 3.4.9 we saw that normal approximations can result in curved exponential families. For each of the following normal approximations:
(i) Describe the curve on which the $\boldsymbol{\theta}$ parameter vector lies.
(ii) Sketch a graph of the curved parameter space.
(a) Poisson approximation: $\bar{X} \sim \mathrm{n}(\lambda, \lambda/n)$
(b) binomial approximation: $\bar{X} \sim \mathrm{n}(p, p(1-p)/n)$
(c) negative binomial approximation: $\bar{X} \sim \mathrm{n}\big(r(1-p)/p,\ r(1-p)/np^2\big)$

Solution. In each case the underlying family is the two-parameter normal family with $\boldsymbol{\theta} = (\mu, \sigma^2)$, so a single free parameter confines $\boldsymbol{\theta}$ to a curve in the half plane $\mathcal{R} \times (0, \infty)$; n is fixed throughout.
(a) $(\mu, \sigma^2) = (\lambda, \lambda/n)$, $\lambda > 0$, so the curve is the straight ray

$$\sigma^2 = \frac{\mu}{n}, \qquad \mu > 0,$$

a line of slope $1/n$ out of the (deleted) origin: the whole ray is covered as λ runs over $(0, \infty)$, and increasing n flattens it toward the μ axis. (For $n = 1$ this is Exercise 3.33(a).)

(b) $(\mu, \sigma^2) = (p, p(1-p)/n)$, $0 < p < 1$, so

$$\sigma^2 = \frac{\mu(1-\mu)}{n}, \qquad 0 < \mu < 1,$$

the arch of a downward parabola with (deleted) zeros at $\mu = 0$ and $\mu = 1$ and maximum height $1/(4n)$ at $\mu = 1/2$.
(c) Solve the mean for p: $\mu = r(1-p)/p$ gives $p = r/(\mu + r)$, and μ runs over all of $(0, \infty)$ as p runs over $(0, 1)$. Substituting into the variance,

$$\sigma^2 = \frac{r(1-p)}{np^2} = \frac{1}{np} \cdot \frac{r(1-p)}{p} = \frac{\mu}{np} = \frac{\mu(\mu + r)}{nr},$$

so the curve is the arc of the upward parabola

$$\sigma^2 = \frac{\mu(\mu + r)}{nr}, \qquad \mu > 0,$$

leaving the (deleted) origin with slope $1/n$ and bending upward thereafter.

Problem 3.35. (a) The normal family that approximates a Poisson can also be parameterized as $\mathrm{n}(e^\theta, e^\theta)$, where $-\infty < \theta < \infty$. Sketch a graph of the parameter space, and compare with the approximation in Exercise 3.34(a).
(b) Suppose that $X \sim \text{gamma}(\alpha, \beta)$ and we assume that $\mathrm{E}X = \mu$. Sketch a graph of the parameter space.
(c) Suppose that $X_i \sim \text{gamma}(\alpha_i, \beta_i)$, $i = 1, 2, \ldots, n$, and we assume that $\mathrm{E}X_i = \mu$. Describe the parameter space $(\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n)$.

Solution. (a) The same ray as in Exercise 3.34(a) with $n = 1$: since $(\mu, \sigma^2) = (e^\theta, e^\theta)$,

$$\sigma^2 = \mu, \qquad \mu > 0,$$

the open 45° ray in the first quadrant. The map $\theta \mapsto e^\theta$ is a bijection of $(-\infty, \infty)$ onto $(0, \infty)$, so this parameterization sweeps out exactly the set that $\lambda \mapsto (\lambda, \lambda)$ does; only the coordinate used to march along the ray differs, and in both cases $d = 1 < k = 2$, a curved family. (For general n, Exercise 3.34(a) gives the ray $\sigma^2 = \mu/n$, the same ray with slope $1/n$ instead of 1.)
(b) With μ fixed, $\mathrm{E}X = \alpha\beta = \mu$ confines the parameter vector to

$$\{(\alpha, \beta) : \alpha\beta = \mu,\ \alpha > 0,\ \beta > 0\} = \Big\{\Big(\alpha, \frac{\mu}{\alpha}\Big) : \alpha > 0\Big\},$$

the branch of the rectangular hyperbola $\beta = \mu/\alpha$ lying in the open first quadrant, asymptotic to both axes and crossing the line $\beta = \alpha$ at $(\sqrt{\mu}, \sqrt{\mu})$; raising μ pushes the branch out along the diagonal. Exercise 3.33(c) is the case $\mu = 1$.
(c) The constraint acts on each coordinate pair separately, so the parameter space is the n-fold product of the hyperbolas of part (b),

$$\Big\{(\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n) \in (0, \infty)^{2n} :\ \alpha_i\beta_i = \mu,\ i = 1, \ldots, n\Big\},$$

an n-dimensional surface in $\mathcal{R}^{2n}$ smoothly parameterized by $(\alpha_1, \ldots, \alpha_n) \in (0, \infty)^n$ through $\beta_i = \mu/\alpha_i$. The joint density of independent $X_1, \ldots, X_n$ is an exponential family with $k = 2n$ terms $\log x_i$ and x_i, so $d = n < k = 2n$: a curved family.

3.6 Exercises 3.36–3.42

Problem 3.36. Consider the pdf $f(x) = \frac{63}{4}\left(x^6 - x^8\right)$, $-1 < x < 1$. Graph $(1/\sigma)f\big((x-\mu)/\sigma\big)$ for each of the following on the same axes.
(a) $\mu = 0$, $\sigma = 1$
(b) $\mu = 3$, $\sigma = 1$
(c) $\mu = 3$, $\sigma = 2$

Solution. All three curves are the same symmetric bimodal "camel" shape, shifted to be centred at μ and stretched by σ; the standard one is

$$f(x) = \frac{63}{4}\,x^6\left(1 - x^2\right), \qquad -1 < x < 1,$$

which vanishes at $x = 0, \pm 1$ and has its two modes where $\frac{d}{dx}(x^6 - x^8) = 2x^5(3 - 4x^2) = 0$, i.e. at $x = \pm\frac{\sqrt{3}}{2} \approx \pm 0.866$, with common height $f\left(\pm\frac{\sqrt{3}}{2}\right) = \frac{1701}{1024} \approx 1.661$. (Check that $\int_{-1}^{1} f = \frac{63}{4}\cdot 2\left(\frac{1}{7} - \frac{1}{9}\right) = 1$.) The three members are

$$\begin{aligned} &\text{(a)} \quad f(x), && -1 < x < 1, \\ &\text{(b)} \quad f(x-3), && 2 < x < 4, \\ &\text{(c)} \quad \tfrac{1}{2} f\left(\tfrac{x-3}{2}\right), && -1 < x < 5, \end{aligned}$$

with landmarks

curve	support	zero at	modes at	modal height
(a)	$(-1, 1)$	$0, \pm 1$	± 0.866	1.661
(b)	$(2, 4)$	3, 2, 4	2.134, 3.866	1.661
(c)	$(-1, 5)$	3, -1, 5	1.268, 4.732	0.831

On one set of axes: (b) is (a) slid three units to the right, shape and height unchanged; (c) sits at the same centre as (b) but is twice as wide and exactly half as tall.

Problem 3.37. Show that if $f(x)$ is a pdf, symmetric about 0, then μ is the median of the location-scale pdf $(1/\sigma)f\big((x-\mu)/\sigma\big)$, $-\infty < x < \infty$.

Solution. By Theorem 3.5.6 we may write $X = \sigma Z + \mu$, where Z has pdf f; since $\sigma > 0$,

$$P(X \ge \mu) = P(\sigma Z + \mu \ge \mu) = P(Z \ge 0) = \int_0^\infty f(z)\,dz.$$

Substituting $z = -u$ and using the symmetry $f(-u) = f(u)$,

$$\int_0^\infty f(z)\,dz = \int_{-\infty}^0 f(-u)\,du = \int_{-\infty}^0 f(u)\,du = P(Z \le 0).$$

These two integrals are equal and sum to $\int_{-\infty}^{\infty} f = 1$, so each equals $\frac{1}{2}$; that is,

$$P(X \ge \mu) = P(X \le \mu) = \tfrac{1}{2},$$

which is precisely the statement that μ is a median of X.

Problem 3.38. Let Z be a random variable with pdf $f(z)$. Define z_α to be a number that satisfies this relationship:

$$\alpha = P(Z > z_\alpha) = \int_{z_\alpha}^\infty f(z)\,dz.$$

Show that if X is a random variable with pdf $(1/\sigma)f\big((x-\mu)/\sigma\big)$ and $x_\alpha = \sigma z_\alpha + \mu$, then $P(X >$

$x_\alpha) = \alpha$. (Thus if a table of z_α values were available, then values of x_α could be easily computed for any member of the location-scale family.)

Solution. By Theorem 3.5.6, X has pdf $(1/\sigma)f\big((x-\mu)/\sigma\big)$ if and only if $X = \sigma Z + \mu$ for a random variable Z with pdf $f(z)$. Since $\sigma > 0$, the map $t \mapsto \sigma t + \mu$ is strictly increasing, so the events $\{X > x_\alpha\}$ and $\{Z > z_\alpha\}$ coincide:

$$\begin{aligned} P(X > x_\alpha) &= P\left(\sigma Z + \mu > \sigma z_\alpha + \mu\right) \\ &= P\left(Z > z_\alpha\right) = \alpha. \end{aligned}$$

Method (2): substitute $z = (x-\mu)/\sigma$ in the integral directly, with $dx = \sigma\, dz$ and the lower limit $\sigma z_\alpha + \mu$ going to z_α:

$$\begin{aligned} P(X > x_\alpha) &= \int_{\sigma z_\alpha + \mu}^{\infty} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx \\ &= \int_{z_\alpha}^{\infty} f(z)\, dz = \alpha. \end{aligned}$$

Problem 3.39. Consider the Cauchy family defined in Section 3.3. This family can be extended to a location-scale family yielding pdfs of the form

$$f(x|\mu,\sigma) = \frac{1}{\sigma\pi\left(1 + \left(\frac{x-\mu}{\sigma}\right)^2\right)}, \qquad -\infty < x < \infty.$$

The mean and variance do not exist for the Cauchy distribution. So the parameters μ and σ^2 are not the mean and variance. But they do have important meaning. Show that if X is a random variable with a Cauchy distribution with parameters μ and σ, then:
(a) μ is the median of the distribution of X, that is, $P(X \geq \mu) = P(X \leq \mu) = \frac{1}{2}$.
(b) $\mu + \sigma$ and $\mu - \sigma$ are the quartiles of the distribution of X, that is, $P(X \geq \mu + \sigma) = P(X \leq \mu - \sigma) = \frac{1}{4}$. (Hint: Prove this first for $\mu = 0$ and $\sigma = 1$ and then use Exercise 3.38.)

Solution. Write the family in standard form: with $f(z) = \dfrac{1}{\pi\left(1+z^2\right)}$, the standard Cauchy pdf (3.3.19) with $\theta = 0$, we have $f(x|\mu,\sigma) = (1/\sigma)f\big((x-\mu)/\sigma\big)$, so by Theorem 3.5.6 $X = \sigma Z + \mu$ with Z standard Cauchy.
(a) f is symmetric about 0, since $f(-z) = f(z)$; hence μ is the median by Exercise 3.37.
(b) For the standard variable, using $\frac{d}{dz}\arctan z = (1+z^2)^{-1}$ and $\arctan(\pm\infty) = \pm\pi/2$,

$$\begin{aligned} P(Z \geq 1) &= \frac{1}{\pi}\int_1^{\infty} \frac{dz}{1+z^2} = \frac{1}{\pi}\Big[\arctan z\Big]_1^{\infty} \\ &= \frac{1}{\pi}\left(\frac{\pi}{2} - \frac{\pi}{4}\right) = \frac{1}{4}, \end{aligned}$$

so $z_{1/4} = 1$ in the notation of Exercise 3.38. That exercise, applied with $\alpha = \frac{1}{4}$, gives $x_{1/4} = \sigma \cdot 1 + \mu = \mu + \sigma$ and

$$P(X \geq \mu + \sigma) = P(Z \geq 1) = \tfrac{1}{4}.$$

By the symmetry of f about 0, $P(Z \leq -1) = P(Z \geq 1) = \frac{1}{4}$, and therefore

$$P(X \leq \mu - \sigma) = P(\sigma Z + \mu \leq \mu - \sigma) = P(Z \leq -1) = \tfrac{1}{4}.$$

Problem 3.40. Let $f(x)$ be any pdf with mean μ and variance σ^2. Show how to create a location-scale family based on $f(x)$ such that the standard pdf of the family, say $f^*(x)$, has mean 0 and variance 1.

Solution. Take

$$f^*(x) = \sigma f(\sigma x + \mu), \qquad -\infty < x < \infty,$$

and let the family be $\big\{\,(1/\tau)f^*\big((x-\nu)/\tau\big) : -\infty < \nu < \infty,\ \tau > 0\,\big\}$.
First, f^* is a pdf and generates the same family as f: writing it as

$$f^*(x) = \frac{1}{(1/\sigma)}\, f\left(\frac{x - (-\mu/\sigma)}{1/\sigma}\right),$$

it is the member of the location-scale family built from f with location $-\mu/\sigma$ and scale $1/\sigma$, hence a pdf by Theorem 3.5.1. The two families coincide, since composing affine maps gives

$$\frac{1}{\tau} f^*\left(\frac{x-\nu}{\tau}\right) = \frac{1}{\tau/\sigma}\, f\left(\frac{x - (\nu - \mu\tau/\sigma)}{\tau/\sigma}\right),$$

and $(\nu, \tau) \mapsto (\nu - \mu\tau/\sigma,\ \tau/\sigma)$ maps $\mathbb{R} \times (0,\infty)$ onto itself; f itself is the member with $\nu = \mu$, $\tau = \sigma$.
Second, f^* has the required moments. If X has pdf f, then by Theorem 3.5.6 a random variable with pdf f^* is $Z = (X - \mu)/\sigma$, so

$$\mathrm{E}Z = \frac{\mathrm{E}X - \mu}{\sigma} = 0, \qquad \mathrm{Var}\, Z = \frac{\mathrm{Var}\, X}{\sigma^2} = 1,$$

both existing because $\mathrm{E}X = \mu$ and $\mathrm{Var}\, X = \sigma^2$ were assumed finite. By Theorem 3.5.7 the general member $(1/\tau)f^*\big((x-\nu)/\tau\big)$ then has mean ν and variance τ^2.

Problem 3.41. A family of cdfs $\{F(x|\theta), \theta \in \Theta\}$ is *stochastically increasing in* θ if $\theta_1 > \theta_2 \Rightarrow F(x|\theta_1)$ is stochastically greater than $F(x|\theta_2)$. (See Exercise 1.49 for the definition of stochastically greater: F_X is stochastically greater than F_Y if $F_X(t) \le F_Y(t)$ for all t, with strict inequality for some t.)
(a) Show that the $\mathrm{n}(\mu, \sigma^2)$ family is stochastically increasing in μ for fixed σ^2.
(b) Show that the gamma(α, β) family of (3.3.6) is stochastically increasing in β (scale parameter) for fixed α (shape parameter).

Solution. In both parts the cdf is a strictly decreasing function of the parameter at every x where the density has mass, which is more than the definition requires.
(a) Let $\mu_1 > \mu_2$ and let Φ be the standard normal cdf. The $\mathrm{n}(\mu, \sigma^2)$ family is the location-scale family with standard pdf $\mathrm{n}(0,1)$, so by the standardization $P(X \le x) = P\big(Z \le (x-\mu)/\sigma\big)$ closing Section 3.5,

$$F(x|\mu_i) = \Phi\left(\frac{x - \mu_i}{\sigma}\right), \qquad i = 1, 2.$$

Since $\sigma > 0$ we have $\frac{x-\mu_1}{\sigma} < \frac{x-\mu_2}{\sigma}$ for every x, and Φ is strictly increasing on $\mathbb{R}$ (its density ϕ is everywhere positive), so

$$F(x|\mu_1) < F(x|\mu_2) \quad \text{for every } x \in \mathbb{R}.$$

Hence $F(x|\mu_1)$ is stochastically greater than $F(x|\mu_2)$.
(b) Let $\beta_1 > \beta_2 > 0$ and fix $\alpha > 0$. As in the derivation of (3.3.6) from (3.3.5), $X \sim \text{gamma}(\alpha, \beta)$ is the same as $X = \beta T$ with $T \sim \text{gamma}(\alpha, 1)$, so for every x

$$F(x|\beta) = P(\beta T \le x) = F_T\left(\frac{x}{\beta}\right).$$

For $x \le 0$ both sides are 0, since the gamma sample space is $(0, \infty)$. For $x > 0$ we have $x/\beta_1 < x/\beta_2$, and F_T is strictly increasing on $(0, \infty)$ because the gamma$(\alpha, 1)$ density is positive there; hence

$$F(x|\beta_1) = F_T\left(\frac{x}{\beta_1}\right) < F_T\left(\frac{x}{\beta_2}\right) = F(x|\beta_2).$$

So $F(x|\beta_1) \le F(x|\beta_2)$ for all x with strict inequality on $(0, \infty)$, which is the required stochastic ordering.

Problem 3.42. Refer to Exercise 3.41 for the definition of a stochastically increasing family.
(a) Show that a location family is stochastically increasing in its location parameter.
(b) Show that a scale family is stochastically increasing in its scale parameter if the sample space is $[0, \infty)$.

Solution. Both parts are the monotonicity of the standard cdf F applied to a shifted or rescaled argument; the only work is producing the one point of strict inequality that Exercise 1.49 demands.
(a) Let the standard pdf be f with cdf F, so that the location family of Definition 3.5.2 has cdfs $F(x|\mu) = F(x - \mu)$. Take $\mu_1 > \mu_2$ and put $d = \mu_1 - \mu_2 > 0$. Since $x - \mu_1 < x - \mu_2$ and F is nondecreasing,

$$F(x|\mu_1) = F(x - \mu_1) \le F(x - \mu_2) = F(x|\mu_2) \qquad \text{for all } x.$$

For strictness at some point, suppose instead that $F(t) = F(t + d)$ for every $t \in \mathbb{R}$. Iterating gives $F(t) = F(t + nd)$ for every $n \ge 1$, and letting $n \to \infty$,

$$F(t) = \lim_{n\to\infty} F(t + nd) = 1 \qquad \text{for every } t,$$

contradicting $\lim_{t\to-\infty} F(t) = 0$. So there is a t with $F(t) < F(t + d)$, and at $x = t + \mu_1$,

$$F(x|\mu_1) = F(t) < F(t + d) = F(x|\mu_2).$$

Hence $F(x|\mu_1)$ is stochastically greater than $F(x|\mu_2)$.
(b) Now the standard pdf f vanishes off $[0, \infty)$, the scale family (Definition 3.5.4) has pdfs $(1/\sigma)f(x/\sigma)$, and the corresponding cdfs are $F(x|\sigma) = F(x/\sigma)$, with $F(x) = 0$ for $x \le 0$. Take $\sigma_1 > \sigma_2 > 0$ and put $r = \sigma_1/\sigma_2 > 1$. For $x \le 0$, $F(x|\sigma_1) = 0 = F(x|\sigma_2)$ (this is where the sample space $[0, \infty)$ is used: for $x < 0$ the inequality $x/\sigma_1 > x/\sigma_2$ runs the wrong way, and only the vanishing of F there saves it). For $x > 0$,

$$\frac{x}{\sigma_1} < \frac{x}{\sigma_2} \quad \Longrightarrow \quad F(x|\sigma_1) = F\left(\frac{x}{\sigma_1}\right) \le F\left(\frac{x}{\sigma_2}\right) = F(x|\sigma_2).$$

For strictness, suppose $F(t) = F(rt)$ for every $t > 0$. Iterating, $F(t) = F(r^n t)$ for every $n \ge 1$, and $r^n t \to \infty$, so $F(t) = 1$ for every $t > 0$; but F is continuous with $F(0) = 0$, so $F(t) \to 0$ as $t \downarrow 0$, a contradiction. Hence some $t > 0$ has $F(t) < F(rt)$, and at $x = \sigma_1 t > 0$,

$$F(x|\sigma_1) = F(t) < F(rt) = F\left(\frac{\sigma_1 t}{\sigma_2}\right) = F(x|\sigma_2),$$

which completes the stochastic ordering.

3.7 Exercises 3.43–3.49

Problem 3.43. A family of cdfs $\{F(x|\theta),\ \theta \in \Theta\}$ is *stochastically decreasing in* θ if $\theta_1 > \theta_2 \ \Rightarrow$ $F(x|\theta_2)$ is stochastically greater than $F(x|\theta_1)$. (See Exercises 3.41 and 3.42.)
(a) Prove that if $X \sim F_X(x|\theta)$, where the sample space of X is $(0, \infty)$ and $F_X(x|\theta)$ is stochastically increasing in θ, then $F_Y(y|\theta)$ is stochastically decreasing in θ, where $Y = 1/X$.
(b) Prove that if $X \sim F_X(x|\theta)$, where $F_X(x|\theta)$ is stochastically increasing in θ and $\theta > 0$, then

$F_X\big(x\big|\frac{1}{\theta}\big)$ is stochastically decreasing in θ.
(Recall Exercise 1.49: a cdf F_X is *stochastically greater* than a cdf F_Y if $F_X(t) \le F_Y(t)$ for all t and $F_X(t) < F_Y(t)$ for some t.)

Solution. Both statements are the definition read backwards; the only content is that $x \mapsto 1/x$ and $\theta \mapsto 1/\theta$ each reverse order.
(a) Since the sample space is $(0,\infty)$, for $y > 0$ the event $\{1/X \le y\}$ is the event $\{X \ge 1/y\}$, so

$$F_Y(y|\theta) = P\Big(X \ge \tfrac{1}{y}\,\Big|\,\theta\Big) = 1 - F_X\Big(\tfrac{1}{y}\,\Big|\,\theta\Big),$$

taking $F_X(\cdot|\theta)$ continuous, as in Exercises 3.41–3.42. Fix $\theta_1 > \theta_2$. Stochastic increase gives

$$F_X(x|\theta_1) \le F_X(x|\theta_2) \quad \text{for all } x,$$

with strict inequality at some $x_0 > 0$. Applying this at $x = 1/y$ and negating,

$$\begin{aligned} F_Y(y|\theta_2) &= 1 - F_X\Big(\tfrac{1}{y}\big|\theta_2\Big) \\ &\le 1 - F_X\Big(\tfrac{1}{y}\big|\theta_1\Big) = F_Y(y|\theta_1) \end{aligned}$$

for every $y > 0$, and strictly at $y_0 = 1/x_0$. Thus $F_Y(\cdot|\theta_2)$ is stochastically greater than $F_Y(\cdot|\theta_1)$ whenever $\theta_1 > \theta_2$, which is exactly stochastic decrease in θ.
(b) Write $G(x|\theta) = F_X\big(x\big|\frac{1}{\theta}\big)$ for $\theta > 0$. If $\theta_1 > \theta_2 > 0$ then $\frac{1}{\theta_2} > \frac{1}{\theta_1}$, so stochastic increase of F_X in its parameter, applied to the pair $\frac{1}{\theta_2} > \frac{1}{\theta_1}$, gives

$$G(x|\theta_2) = F_X\Big(x\big|\tfrac{1}{\theta_2}\Big) \le F_X\Big(x\big|\tfrac{1}{\theta_1}\Big) = G(x|\theta_1)$$

for all x, with strict inequality for some x. Hence $G(\cdot|\theta_2)$ is stochastically greater than $G(\cdot|\theta_1)$, i.e. G is stochastically decreasing in θ.

Problem 3.44. For any random variable X for which $\mathrm{E}X^2$ and $\mathrm{E}|X|$ exist, show that $P(|X| \ge b)$ does not exceed either $\mathrm{E}X^2/b^2$ or $\mathrm{E}|X|/b$, where b is a positive constant. If $f(x) = e^{-x}$ for $x > 0$, show that one bound is better when $b = 3$ and the other when $b = \sqrt{2}$. (Notice Markov's Inequality in Miscellanea 3.8.2.)

Solution. Both bounds are Chebychev's Inequality (Theorem 3.6.1), $P(g(X) \ge r) \le \mathrm{E}g(X)/r$, applied to two different nonnegative g.
Take $g(x) = x^2$ and $r = b^2 > 0$. Since $\{|X| \ge b\} = \{X^2 \ge b^2\}$,

$$P(|X| \ge b) = P(X^2 \ge b^2) \le \frac{\mathrm{E}X^2}{b^2}.$$

Take instead $g(x) = |x|$ and $r = b$ (this is Markov's Inequality):

$$P(|X| \ge b) \le \frac{\mathrm{E}|X|}{b}.$$

Both g are nonnegative and both expectations are assumed finite, so the hypotheses of Theorem 3.6.1 hold in each case.
For $f(x) = e^{-x}$, $x > 0$ (exponential with mean 1), $X > 0$ so $|X| = X$, and $\mathrm{E}X = 1$, $\mathrm{E}X^2 = 2$. The two bounds are $2/b^2$ and $1/b$, and

$$\frac{2}{b^2} < \frac{1}{b} \iff b > 2.$$

Hence:

(i) $b = 3$: $\mathrm{E}X^2/b^2 = \frac{2}{9} \approx .222$ versus $\mathrm{E}|X|/b = \frac{1}{3} \approx .333$, so the second-moment bound is better (true value $e^{-3} \approx .0498$).
(ii) $b = \sqrt{2}$: $\mathrm{E}X^2/b^2 = 1$ (vacuous) versus $\mathrm{E}|X|/b = 1/\sqrt{2} \approx .707$, so the Markov bound is better (true value $e^{-\sqrt{2}} \approx .243$).

Problem 3.45. Let X be a random variable with moment generating function $M_X(t)$, $-h < t < h$.
(a) Prove that $P(X \geq a) \leq e^{-at} M_X(t)$, $0 < t < h$. (A proof similar to that used for Chebychev's Inequality will work.)
(b) Similarly, prove that $P(X \leq a) \leq e^{-at} M_X(t)$, $-h < t < 0$.
(c) A special case of part (a) is that $P(X \geq 0) \leq \mathrm{E}e^{tX}$ for all $t \geq 0$ for which the mgf is defined. What are general conditions on a function $h(t, x)$ such that $P(X \geq 0) \leq \mathrm{E}h(t, X)$ for all $t \geq 0$ for which $\mathrm{E}h(t, X)$ exists? (In part (a), $h(t, x) = e^{tx}$.)

Solution. (a) For $0 < t < h$ the map $u \mapsto e^{tu}$ is strictly increasing, so $\{X \geq a\} = \{e^{tX} \geq e^{at}\}$. Apply Chebychev's Inequality (Theorem 3.6.1) with the nonnegative function $g(x) = e^{tx}$ and $r = e^{at} > 0$; $\mathrm{E}g(X) = M_X(t)$ is finite because $|t| < h$:

$$P(X \geq a) = P\big(e^{tX} \geq e^{at}\big) \leq \frac{\mathrm{E}e^{tX}}{e^{at}} = e^{-at} M_X(t).$$

(b) For $-h < t < 0$ the map $u \mapsto e^{tu}$ is strictly decreasing, so now $\{X \leq a\} = \{e^{tX} \geq e^{at}\}$, and the identical application of Theorem 3.6.1 gives

$$P(X \leq a) = P\big(e^{tX} \geq e^{at}\big) \leq e^{-at} M_X(t).$$

(c) It suffices that, for each $t \geq 0$,

$$h(t, x) \geq 0 \ \text{ for all } x, \qquad h(t, x) \geq 1 \ \text{ for all } x \geq 0.$$

Indeed, with I the indicator function (Definition 3.4.5), these two conditions give the pointwise bound $h(t, x) \geq I_{[0,\infty)}(x)$ for every x, whence

$$\mathrm{E}h(t, X) \ \geq \ \mathrm{E}\, I_{[0,\infty)}(X) \ = \ P(X \geq 0).$$

In part (a) with $a = 0$ and $h(t, x) = e^{tx}$: $e^{tx} > 0$ always, and $e^{tx} \geq 1$ for $x \geq 0$ when $t \geq 0$.

Problem 3.46. Calculate $P(|X - \mu_X| \geq k\sigma_X)$ for $X \sim \text{uniform}(0, 1)$ and $X \sim \text{exponential}(\lambda)$, and compare your answers to the bound from Chebychev's Inequality.

Solution.

$$P(|X - \mu_X| \geq k\sigma_X) = \begin{cases} \left(1 - \dfrac{k}{\sqrt{3}}\right)^+ & \text{uniform}(0, 1), \\ e^{-(1+k)} + \big(1 - e^{-(1-k)}\big)^+ & \text{exponential}(\lambda), \end{cases}$$

against Chebychev's bound $1/k^2$ (Example 3.6.2). Here $u^+ = \max(u, 0)$.
Uniform$(0, 1)$: $\mu = \frac{1}{2}$, $\sigma^2 = \frac{1}{12}$, so $k\sigma = k/(2\sqrt{3})$. The event $|X - \frac{1}{2}| < k\sigma$ is the interval $\left(\frac{1}{2} - \frac{k}{2\sqrt{3}},\ \frac{1}{2} + \frac{k}{2\sqrt{3}}\right)$, whose intersection with $(0, 1)$ has length $\min\big(k/\sqrt{3},\, 1\big)$. Hence

$$P\big(|X - \tfrac{1}{2}| \geq k\sigma\big) = 1 - \frac{k}{\sqrt{3}} \quad (0 \leq k < \sqrt{3}),$$

and it is 0 for $k \geq \sqrt{3} \approx 1.732$.
Exponential(λ) (pdf $\lambda^{-1} e^{-x/\lambda}$ on $(0, \infty)$): $\mu = \lambda$, $\sigma = \lambda$, so the deviation event is $\{X \geq \lambda(1+k)\} \cup \{X \leq$

$\lambda(1-k)\}$ and

$$P(X \geq \lambda(1+k)) = e^{-(1+k)},$$
$$P(X \leq \lambda(1-k)) = 1 - e^{-(1-k)} \quad (k < 1),$$

the second event being empty for $k \geq 1$; λ cancels, being a scale parameter (Definition 3.5.4). Numerically:

k	uniform	exponential	Chebychev $1/k^2$
1	.4226	.1353	1.0000
2	0	.0498	.2500
3	0	.0183	.1111

Both exact probabilities sit far below $1/k^2$, as Section 3.6.1 warns: the uniform is already 0 past $k = \sqrt{3}$, and the exponential tail decays like e^{-k} against the bound's k^{-2}.

Problem 3.47. If Z is a standard normal random variable, prove this companion to the inequality in Example 3.6.3:

$$P(|Z| \geq t) \;\geq\; \sqrt{\frac{2}{\pi}}\,\frac{t}{1+t^2}\,e^{-t^2/2}.$$

(Example 3.6.3 established the upper bound $P(|Z| \geq t) \leq \sqrt{2/\pi}\,e^{-t^2/2}/t$ for all $t > 0$.)

Solution. The whole proof is the antiderivative

$$\frac{d}{dx}\left(-\frac{e^{-x^2/2}}{x}\right) = \left(1+\frac{1}{x^2}\right)e^{-x^2/2}, \qquad x > 0,$$

which integrates over (t, ∞) to

$$\int_t^\infty \left(1+\frac{1}{x^2}\right)e^{-x^2/2}\,dx = \frac{e^{-t^2/2}}{t}.$$

On that range $x > t$, so $1 + x^{-2} \leq 1 + t^{-2}$ and therefore

$$\begin{aligned}\frac{e^{-t^2/2}}{t} &= \int_t^\infty \left(1+\frac{1}{x^2}\right)e^{-x^2/2}\,dx \\ &\leq \left(1+\frac{1}{t^2}\right)\int_t^\infty e^{-x^2/2}\,dx.\end{aligned}$$

Dividing by $1 + t^{-2} = (1+t^2)/t^2$,

$$\int_t^\infty e^{-x^2/2}\,dx \;\geq\; \frac{t^2}{1+t^2}\cdot\frac{e^{-t^2/2}}{t} = \frac{t}{1+t^2}e^{-t^2/2}.$$

Multiplying by $1/\sqrt{2\pi}$ gives $P(Z \geq t) \geq \frac{1}{\sqrt{2\pi}}\frac{t}{1+t^2}e^{-t^2/2}$, and by symmetry $P(|Z| \geq t) = 2P(Z \geq t)$, so

$$P(|Z| \geq t) \;\geq\; \frac{2}{\sqrt{2\pi}}\,\frac{t}{1+t^2}e^{-t^2/2} = \sqrt{\frac{2}{\pi}}\,\frac{t}{1+t^2}e^{-t^2/2}.$$

Problem 3.48. Derive recursion relations, similar to the one given in (3.6.2), for the binomial, negative binomial, and hypergeometric distributions. (Equation (3.6.2) is the Poisson relation $P(X = x+1) = \frac{\lambda}{x+1}P(X = x)$, which allows Poisson probabilities to be computed recursively starting from $P(X = 0) = e^{-\lambda}$.)

Solution. In each case form the ratio $P(X=x+1)/P(X=x)$ and cancel factorials.
Binomial (n,p), $x=0,1,\dots,n-1$:

$$\frac{P(X=x+1)}{P(X=x)} = \frac{\binom{n}{x+1}p^{x+1}(1-p)^{n-x-1}}{\binom{n}{x}p^{x}(1-p)^{n-x}} = \frac{n-x}{x+1}\cdot\frac{p}{1-p},$$

using $\binom{n}{x+1}/\binom{n}{x} = (n-x)/(x+1)$. Hence

$$P(X=x+1) = \frac{(n-x)\,p}{(x+1)(1-p)}\,P(X=x), \qquad P(X=0)=(1-p)^n.$$

Negative binomial (r,p) in the form (3.2.10), $P(X=x)=\binom{r+x-1}{x}p^r(1-p)^x$, $x=0,1,\dots$:

$$\frac{\binom{r+x}{x+1}}{\binom{r+x-1}{x}} = \frac{(r+x)!\,x!}{(x+1)!\,(r+x-1)!} = \frac{r+x}{x+1},$$

so

$$P(X=x+1) = \frac{(r+x)(1-p)}{x+1}\,P(X=x), \qquad P(X=0)=p^r.$$

Hypergeometric (N,M,K), $P(X=x)=\binom{M}{x}\binom{N-M}{K-x}/\binom{N}{K}$:

$$\frac{P(X=x+1)}{P(X=x)} = \frac{\binom{M}{x+1}}{\binom{M}{x}}\cdot\frac{\binom{N-M}{K-x-1}}{\binom{N-M}{K-x}} = \frac{M-x}{x+1}\cdot\frac{K-x}{N-M-K+x+1},$$

the second factor from $\binom{m}{k-1}/\binom{m}{k} = k/(m-k+1)$ with $m=N-M$, $k=K-x$. Hence

$$P(X=x+1) = \frac{(M-x)(K-x)}{(x+1)(N-M-K+x+1)}\,P(X=x),$$

for $\max(0,K-N+M)\le x<\min(M,K)$, started at $P(X=0)=\binom{N-M}{K}/\binom{N}{K}$.

Problem 3.49. Prove the following analogs to Stein's Lemma, assuming appropriate conditions on the function g.
(a) If $X\sim\text{gamma}(\alpha,\beta)$, then

$$\mathrm{E}\big(g(X)(X-\alpha\beta)\big) = \beta\,\mathrm{E}\big(Xg'(X)\big).$$

(b) If $X\sim\text{beta}(\alpha,\beta)$, then

$$\mathrm{E}\left[g(X)\left(\beta-(\alpha-1)\frac{1-X}{X}\right)\right] = \mathrm{E}\big((1-X)g'(X)\big).$$

Solution. Each identity is integration by parts against a pdf, exactly as in the proof of Stein's Lemma (Lemma 3.6.5); the only new ingredient is the right exact derivative.
(a) With $f(x|\alpha,\beta)=\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}$ on $(0,\infty)$,

$$\frac{d}{dx}\big[x\,f(x|\alpha,\beta)\big] = \frac{\alpha x^{\alpha-1}-\beta^{-1}x^\alpha}{\Gamma(\alpha)\beta^\alpha}e^{-x/\beta} = -\frac{1}{\beta}(x-\alpha\beta)f(x|\alpha,\beta),$$

i.e. $(x-\alpha\beta)f(x) = -\beta\,\frac{d}{dx}[xf(x)]$. Therefore

$$\begin{aligned}\mathrm{E}\big(g(X)(X-\alpha\beta)\big) &= -\beta\int_0^\infty g(x)\,\frac{d}{dx}\big[xf(x)\big]\,dx\\ &= -\beta\Big[g(x)\,xf(x)\Big]_0^\infty + \beta\int_0^\infty g'(x)\,xf(x)\,dx\\ &= \beta\,\mathrm{E}\big(Xg'(X)\big),\end{aligned}$$

the boundary term vanishing under the appropriate condition on g: $g(x)\,x\,f(x|\alpha,\beta)\to 0$ at 0 and at ∞, with $\mathrm{E}|Xg'(X)|<\infty$ (the analog of $\mathrm{E}|g'(X)|<\infty$ in Lemma 3.6.5).
(b) With $f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ on $(0,1)$, put

$$h(x) = (1-x)f(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta}}{B(\alpha,\beta)}.$$

Differentiating,

$$\begin{aligned}h'(x) &= \frac{(\alpha-1)x^{\alpha-2}(1-x)^{\beta} - \beta x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\\ &= -f(x|\alpha,\beta)\left[\beta-(\alpha-1)\frac{1-x}{x}\right],\end{aligned}$$

since $x^{\alpha-2}(1-x)^{\beta} = x^{\alpha-1}(1-x)^{\beta-1}\cdot\frac{1-x}{x}$. Hence

$$\begin{aligned}\mathrm{E}\left[g(X)\left(\beta-(\alpha-1)\tfrac{1-X}{X}\right)\right] &= -\int_0^1 g(x)h'(x)\,dx\\ &= -\Big[g(x)h(x)\Big]_0^1 + \int_0^1 g'(x)h(x)\,dx\\ &= \mathrm{E}\big((1-X)g'(X)\big),\end{aligned}$$

the boundary term vanishing under the condition $g(x)(1-x)f(x|\alpha,\beta)\to 0$ at 0 and at 1, with $\mathrm{E}|(1-X)g'(X)|<\infty$.

3.8 Exercises 3.50–3.50

Problem 3.50. Prove the identity for the negative binomial distribution given in Theorem 3.6.8, part (b).
For reference, Theorem 3.6.8 (Hwang) states: let $g(x)$ be a function with $-\infty < \mathrm{E}\,g(X) < \infty$ and $-\infty < g(-1) < \infty$. Then:
(a) If $X\sim$ Poisson(λ),

$$\mathrm{E}\big(\lambda g(X)\big) = \mathrm{E}\big(Xg(X-1)\big).$$

(b) If $X\sim$ negative binomial(r,p),

$$\mathrm{E}\big((1-p)g(X)\big) = \mathrm{E}\left(\frac{X}{r+X-1}\,g(X-1)\right).$$

Here the negative binomial pmf is the one in (3.2.10), namely

$$P(X=x) = \binom{r+x-1}{x}p^r(1-p)^x, \qquad x=0,1,2,\ldots$$

Solution. The weight $x/(r+x-1)$ is absorbed by the binomial coefficient, which lowers both of its

arguments by one:

$$\frac{x}{r+x-1}\binom{r+x-1}{x} = \frac{x\,(r+x-2)!}{x!\,(r-1)!} = \binom{r+x-2}{x-1}, \qquad x \geq 1.$$

Hence, using the pmf (3.2.10) and substituting $y = x-1$,

$$\begin{aligned}
\mathrm{E}\left(\frac{X}{r+X-1}g(X-1)\right) &= \sum_{x=1}^{\infty}\binom{r+x-2}{x-1}g(x-1)\,p^r(1-p)^x \\
&= (1-p)\sum_{y=0}^{\infty}\binom{r+y-1}{y}g(y)\,p^r(1-p)^y \\
&= \mathrm{E}\big((1-p)g(X)\big).
\end{aligned}$$

The omitted $x = 0$ term contributes nothing: its weight $x/(r+x-1)$ is 0 when $r > 1$, and when $r = 1$ it reads $0/0$ and is taken to be 0, which is meaningful only because $g(-1)$ is finite by hypothesis. The other hypothesis, $-\infty < \mathrm{E}\,g(X) < \infty$, makes the last sum finite, and the chain above is a term-by-term re-indexing rather than a rearrangement.

4 Multiple Random Variables

Contents

4.1 Exercises 4.1–4.7 99
4.2 Exercises 4.8–4.14 103
4.3 Exercises 4.15–4.21 106
4.4 Exercises 4.22–4.28 111
4.5 Exercises 4.29–4.35 116
4.6 Exercises 4.36–4.42 121
4.7 Exercises 4.43–4.49 126
4.8 Exercises 4.50–4.56 132
4.9 Exercises 4.57–4.63 136
4.10 Exercises 4.64–4.65 141

4.1 Exercises 4.1–4.7

Problem 4.1. A random point (X, Y) is distributed uniformly on the square with vertices $(1, 1)$, $(1, -1)$, $(-1, 1)$, and $(-1, -1)$. That is, the joint pdf is $f(x, y) = \frac{1}{4}$ on the square. Determine the probabilities of the following events.
(a) $X^2 + Y^2 < 1$
(b) $2X - Y > 0$
(c) $|X + Y| < 2$

Solution. $\pi/4$, $1/2$, and 1. The square $S = (-1, 1) \times (-1, 1)$ has area 4, so for any $A \subset S$ the uniform density $f = \frac{1}{4}$ gives $P((X, Y) \in A) = \frac{1}{4}\,\text{area}(A)$.
(a) The disk $\{x^2 + y^2 < 1\}$ is contained in S and has area π, so

$$P(X^2 + Y^2 < 1) \;=\; \tfrac{1}{4}\pi \;=\; \frac{\pi}{4} \approx 0.785.$$

(b) The map $(x, y) \mapsto (-x, -y)$ carries S onto itself and preserves area, while it carries $\{2x - y > 0\}$ onto $\{2x - y < 0\}$. Hence those two events have equal probability, and since the line $2x = y$ has area 0,

$$P(2X - Y > 0) \;=\; \tfrac{1}{2}\big(1 - P(2X = Y)\big) \;=\; \frac{1}{2}.$$

(c) On S we have $|x + y| \le |x| + |y| < 2$ at every interior point, so the event contains all of S except a null set:

$$P(|X + Y| < 2) \;=\; 1.$$

Problem 4.2. Prove the following properties of bivariate expectations (the bivariate analog to Theorem 2.2.5). For random variables X and Y, functions $g_1(x, y)$ and $g_2(x, y)$, and constants a, b, and c:
(a) $\mathrm{E}(ag_1(X, Y) + bg_2(X, Y) + c) = a\mathrm{E}(g_1(X, Y)) + b\mathrm{E}(g_2(X, Y)) + c$.
(b) If $g_1(x, y) \ge 0$, then $\mathrm{E}(g_1(X, Y)) \ge 0$.
(c) If $g_1(x, y) \ge g_2(x, y)$, then $\mathrm{E}(g_1(X, Y)) \ge \mathrm{E}(g_2(X, Y))$.
(d) If $a \le g_1(x, y) \le b$, then $a \le \mathrm{E}(g_1(X, Y)) \le b$.

Solution. All four follow from linearity and positivity of the integral, exactly as in the univariate Theorem 2.2.5, whose hypothesis that $\mathrm{E}(g_1(X, Y))$ and $\mathrm{E}(g_2(X, Y))$ exist is what licenses the splitting in (a). Only the continuous case is written; the discrete case replaces $\iint \cdot\, f(x, y)\, dx\, dy$ by $\sum_{(x,y)} \cdot\, f(x, y)$.
(a) Since the integrals of $|g_1|f$ and $|g_2|f$ are finite, the integral of the sum splits:

$$\begin{aligned}
\mathrm{E}(ag_1 + bg_2 + c) &= \iint \big(ag_1(x, y) + bg_2(x, y) + c\big) f(x, y)\, dx\, dy \\
&= a \iint g_1 f\, dx\, dy + b \iint g_2 f\, dx\, dy + c \iint f\, dx\, dy \\
&= a\mathrm{E}(g_1(X, Y)) + b\mathrm{E}(g_2(X, Y)) + c,
\end{aligned}$$

the last step using $\iint f(x, y)\, dx\, dy = 1$.
(b) If $g_1 \ge 0$ then the integrand $g_1(x, y)f(x, y) \ge 0$ everywhere, since $f \ge 0$; an integral of a nonnegative function is nonnegative, so $\mathrm{E}(g_1(X, Y)) \ge 0$.
(c) Apply (b) to $g_1 - g_2 \ge 0$ and then (a) with $a = 1$, $b = -1$, $c = 0$:

$$\begin{aligned}
0 \;&\le\; \mathrm{E}(g_1(X, Y) - g_2(X, Y)) \\
&=\; \mathrm{E}(g_1(X, Y)) - \mathrm{E}(g_2(X, Y)).
\end{aligned}$$

(d) A constant function has expectation equal to that constant, by (a) with $a = b = 0$. So apply (c) to the pair (g_1, a) for the lower bound and to the pair (b, g_1) for the upper:

$$a = \mathrm{E}(a) \;\le\; \mathrm{E}(g_1(X, Y)) \;\le\; \mathrm{E}(b) = b.$$

Problem 4.3. Using Definition 4.1.1, show that the random vector (X, Y) defined at the end of Example 4.1.5 has the pmf given in that example.
(Definition 4.1.1: an n-dimensional random vector is a function from a sample space S into $\Re^n$. Example 4.1.5 first defines a pmf by $f(0,0) = f(0,1) = \frac{1}{6}$, $f(1,0) = f(1,1) = \frac{1}{3}$, and $f(x,y) = 0$ otherwise, and then exhibits a sample space and random vector realizing it: let S be the 36-point sample space for the experiment of tossing two fair dice; let $X = 0$ if the first die shows at most 2 and $X = 1$ if the first die shows more than 2, and let $Y = 0$ if the second die shows an odd number and $Y = 1$ if the second die shows an even number.)

Solution. Count sample points. Here $S = \{(i,j) : 1 \le i, j \le 6\}$ with all 36 points equally likely, and Definition 4.1.1 makes (X, Y) the function $S \to \Re^2$ given by

$$X(i,j) = \begin{cases} 0 & i \in \{1,2\}, \\ 1 & i \in \{3,4,5,6\}, \end{cases}$$

$$Y(i,j) = \begin{cases} 0 & j \in \{1,3,5\}, \\ 1 & j \in \{2,4,6\}. \end{cases}$$

Since X depends only on i and Y only on j, the event $\{X = x, Y = y\}$ is the product set $\{i : X = x\} \times \{j : Y = y\}$, whose cardinality is the product of the two cardinalities. The first factor has 2 elements when $x = 0$ and 4 when $x = 1$; the second has 3 elements for either value of y. By Definition 4.1.3, $f(x,y) = P(X = x, Y = y) = \#\{X = x, Y = y\}/36$:

$$f(0,0) = f(0,1) = \frac{2 \cdot 3}{36} = \frac{1}{6},$$
$$f(1,0) = f(1,1) = \frac{4 \cdot 3}{36} = \frac{1}{3}.$$

No sample point maps to any other (x, y), so $f(x,y) = 0$ elsewhere; these are exactly the values listed in Example 4.1.5.

Problem 4.4. A pdf is defined by

$$f(x,y) = \begin{cases} C(x+2y) & \text{if } 0 < y < 1 \text{ and } 0 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the value of C.
(b) Find the marginal distribution of X.
(c) Find the joint cdf of X and Y.
(d) Find the pdf of the random variable $Z = 9/(X+1)^2$.

Solution. $C = 1/4$, $f_X(x) = (x+1)/4$ on $(0,2)$, and $f_Z(z) = 9/(8z^2)$ on $(1,9)$.
(a) Integrate over the rectangle:

$$1 = C \int_0^2 \int_0^1 (x + 2y)\, dy\, dx = C \int_0^2 \left[xy + y^2\right]_{y=0}^1 dx$$
$$= C \int_0^2 (x+1)\, dx = C\left[\tfrac{x^2}{2} + x\right]_0^2 = 4C,$$

so $C = \frac{1}{4}$.
(b) The inner integral just computed is the marginal:

$$f_X(x) = \int_0^1 \tfrac{1}{4}(x + 2y)\, dy = \frac{x+1}{4}, \qquad 0 < x < 2,$$

and 0 otherwise. Integrating, $F_X(x) = \frac{1}{8}x(x+2)$ for $0 < x < 2$.
(c) For $0 < x < 2$, $0 < y < 1$,

$$F(x,y) = \int_0^x \int_0^y \tfrac{1}{4}(s+2t)\,dt\,ds = \tfrac{1}{4}\int_0^x \left(sy + y^2\right) ds$$
$$= \tfrac{1}{4}\left(\tfrac{x^2 y}{2} + xy^2\right) = \frac{xy(x+2y)}{8}.$$

Off that rectangle the cdf is obtained by capping each argument:

$$F(x,y) = \begin{cases} 0 & x \le 0 \text{ or } y \le 0, \\ \dfrac{xy(x+2y)}{8} & 0 < x < 2,\ 0 < y < 1, \\ \dfrac{y(1+y)}{2} & x \ge 2,\ 0 < y < 1, \\ \dfrac{x(x+2)}{8} & 0 < x < 2,\ y \ge 1, \\ 1 & x \ge 2,\ y \ge 1. \end{cases}$$

(d) On $0 < x < 2$ the map $x \mapsto 9/(x+1)^2$ is strictly decreasing from 9 to 1, so Z takes values in $(1, 9)$ and, for such z,

$$Z \le z \iff (X+1)^2 \ge \frac{9}{z} \iff X \ge \frac{3}{\sqrt{z}} - 1.$$

Hence, with $u = 3/\sqrt{z} - 1 \in (0, 2)$,

$$F_Z(z) = 1 - F_X(u) = 1 - \tfrac{1}{8}u(u+2) = 1 - \tfrac{1}{8}\left(\tfrac{3}{\sqrt{z}} - 1\right)\left(\tfrac{3}{\sqrt{z}} + 1\right)$$
$$= 1 - \tfrac{1}{8}\left(\tfrac{9}{z} - 1\right) = \frac{9}{8}\left(1 - \frac{1}{z}\right),$$

which is 0 at $z = 1$ and 1 at $z = 9$. (Check!) Differentiating,

$$f_Z(z) = \frac{9}{8z^2}, \qquad 1 < z < 9,$$

and 0 otherwise.

Problem 4.5. (a) Find $P(X > \sqrt{Y})$ if X and Y are jointly distributed with pdf

$$f(x,y) = x + y, \qquad 0 \le x \le 1,\ 0 \le y \le 1.$$

(b) Find $P(X^2 < Y < X)$ if X and Y are jointly distributed with pdf

$$f(x,y) = 2x, \qquad 0 \le x \le 1,\ 0 \le y \le 1.$$

Solution. 7/20 and 1/6.
(a) On the unit square, $x > \sqrt{y} \iff y < x^2$, so integrate y from 0 to x^2:

$$P(X > \sqrt{Y}) = \int_0^1 \int_0^{x^2} (x+y)\,dy\,dx = \int_0^1 \left[xy + \tfrac{y^2}{2}\right]_{y=0}^{x^2} dx$$
$$= \int_0^1 \left(x^3 + \tfrac{x^4}{2}\right) dx = \frac{1}{4} + \frac{1}{10} = \frac{7}{20}.$$

(b) For $0 \le x \le 1$ we have $x^2 \le x$, and both lie in $[0,1]$, so the inner limits are x^2 and x:

$$P(X^2 < Y < X) = \int_0^1 \int_{x^2}^{x} 2x \, dy \, dx = \int_0^1 2x\,(x - x^2)\, dx$$
$$= \int_0^1 (2x^2 - 2x^3)\, dx = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}.$$

Problem 4.6. A and B agree to meet at a certain place between 1 PM and 2 PM. Suppose they arrive at the meeting place independently and randomly during the hour. Find the distribution of the length of time that A waits for B. (If B arrives before A, define A's waiting time as 0.)

Solution. A mixed distribution: an atom of size $\frac{1}{2}$ at 0, plus density $1 - w$ on $(0,1)$. Measure time in hours after 1 PM, let X and Y be A's and B's arrival times, independent uniform$(0,1)$, and let $W = \max(Y - X,\, 0)$ be A's waiting time.
The joint pdf is $f(x,y) = 1$ on the unit square, so probabilities are areas. First the atom:

$$P(W = 0) = P(Y \le X) = \tfrac{1}{2},$$

by symmetry of the square about the diagonal $y = x$ (the diagonal itself has area 0).
For $0 \le w < 1$ the event $\{W > w\} = \{Y > X + w\}$ is the triangle with vertices $(0, w)$, $(0, 1)$, $(1 - w, 1)$, whose legs both have length $1 - w$:

$$P(W > w) = \frac{(1-w)^2}{2},$$

so the cdf is

$$F_W(w) = \begin{cases} 0 & w < 0, \\ 1 - \dfrac{(1-w)^2}{2} & 0 \le w < 1, \\ 1 & w \ge 1. \end{cases}$$

This jumps from 0 to $\frac{1}{2}$ at $w = 0$ and is absolutely continuous on $(0,1)$ with

$$f_W(w) = \frac{d}{dw}\Big(1 - \tfrac{1}{2}(1-w)^2\Big) = 1 - w, \qquad 0 < w < 1,$$

and 0 elsewhere.

Problem 4.7. A woman leaves for work between 8 AM and 8:30 AM and takes between 40 and 50 minutes to get there. Let the random variable X denote her time of departure, and the random variable Y the travel time. Assuming that these variables are independent and uniformly distributed, find the probability that the woman arrives at work before 9 AM.

Solution. 1/2. Measure X in minutes after 8 AM, so $X \sim$ uniform$(0, 30)$ and $Y \sim$ uniform$(40, 50)$; by independence the joint pdf is the product

$$f(x,y) = \frac{1}{30} \cdot \frac{1}{10} = \frac{1}{300}, \qquad 0 < x < 30,\ 40 < y < 50.$$

She arrives before 9 AM exactly when $X + Y < 60$. For each $y \in (40, 50)$ the cut $60 - y$ lies in $(10, 20)$, hence strictly inside $(0, 30)$, so the x-section has length $60 - y$:

$$P(X + Y < 60) = \frac{1}{300} \int_{40}^{50} \int_0^{60-y} dx \, dy = \frac{1}{300} \int_{40}^{50} (60 - y)\, dy$$
$$= \frac{1}{300} \Big[60y - \tfrac{y^2}{2}\Big]_{40}^{50} = \frac{1750 - 1600}{300} = \frac{150}{300} = \frac{1}{2}.$$

4.2 Exercises 4.8–4.14

Problem 4.8. Referring to Miscellanea 4.9.1 (the Exchange Paradox). A swami puts \$m in one envelope and \$2m in another. You and your opponent each pick one of the two envelopes at random. You open yours and find x dollars, and the swami then asks whether you want to trade envelopes. You reason that a switch yields either $x/2$ or $2x$ dollars, each with probability $1/2$, for an expected value of $\frac{1}{2}(x/2) + \frac{1}{2}(2x) = 5x/4 > x$, so you offer to trade — and so does your opponent, by the same calculation. The resolutions offered in the Miscellanea are:
(i) (Christensen and Utts 1992) Let $M \sim \pi(m)$ be the distribution of the amount placed in the first envelope, let X be the amount in your envelope and Y the amount in your opponent's. Then $P(X = m \mid M = m) = P(X = 2m \mid M = m) = 1/2$, and hence

$$P(M = x \mid X = x) = \frac{\pi(x)}{\pi(x) + \pi(x/2)}, \qquad P(M = x/2 \mid X = x) = \frac{\pi(x/2)}{\pi(x) + \pi(x/2)}.$$

It then follows that the expected winning from a trade is

$$\frac{\pi(x)}{\pi(x) + \pi(x/2)}\, 2x \;+\; \frac{\pi(x/2)}{\pi(x) + \pi(x/2)}\, \frac{x}{2}.$$

(ii) A more classical approach assumes no pdf on the amount in the first envelope, and holds that the correct conditional distributions are $P(Y = 2x \mid X = m) = 1$ and $P(Y = x/2 \mid X = 2m) = 1$.
(a) Show that $P(X = m \mid M = m) = P(X = 2m \mid M = m) = 1/2$, and verify the expressions for $P(M = x \mid X = x)$ and $P(M = x/2 \mid X = x)$.
(b) Verify that one should trade only if $\pi(x/2) < 2\pi(x)$, and if π is the exponential(λ) density, show that it is optimal to trade if $x < 2\log 2/\lambda$.
(c) For the classical approach, show that $P(Y = 2x \mid X = m) = 1$ and $P(Y = x/2 \mid X = 2m) = 1$ and that your expected winning if you trade or keep your envelope is $E(Y) = 3m/2$.

Solution. (a) The envelope you open is chosen by a fair coin flip that is made independently of the swami's choice of m, so with I the indicator of "you hold the first envelope", $P(I = 1) = 1/2$ independently of M, and

$$P(X = m \mid M = m) = P(I = 1) = \tfrac{1}{2}, \qquad P(X = 2m \mid M = m) = P(I = 0) = \tfrac{1}{2}.$$

Bayes' Rule (Theorem 1.3.5) applied to the two exhaustive hypotheses $\{M = x\}$ (you hold the small envelope) and $\{M = x/2\}$ (you hold the large one) gives

$$\begin{aligned} P(M = x \mid X = x) &= \frac{P(X = x \mid M = x)\,\pi(x)}{P(X = x \mid M = x)\,\pi(x) + P(X = x \mid M = x/2)\,\pi(x/2)} \\ &= \frac{\frac{1}{2}\pi(x)}{\frac{1}{2}\pi(x) + \frac{1}{2}\pi(x/2)} = \frac{\pi(x)}{\pi(x) + \pi(x/2)}, \end{aligned}$$

and $P(M = x/2 \mid X = x) = 1 - P(M = x \mid X = x) = \pi(x/2)/(\pi(x) + \pi(x/2))$. (The Miscellanea calls π a pdf, so these are posterior densities at the two points x and $x/2$; the arithmetic is the display above either way.)
(b) Conditional on $X = x$, a trade yields $2x$ on $\{M = x\}$ and $x/2$ on $\{M = x/2\}$, so the expected winning from a trade is the quantity displayed in the statement, and it beats the x in hand exactly when

$$\begin{aligned} & 2x\,\pi(x) + \tfrac{x}{2}\pi(x/2) > x\big(\pi(x) + \pi(x/2)\big) \\ \iff \quad & \pi(x) > \tfrac{1}{2}\pi(x/2) \quad \iff \quad \pi(x/2) < 2\pi(x). \end{aligned}$$

For $\pi(t) = \lambda e^{-\lambda t}$, $t > 0$, this reads

$$\lambda e^{-\lambda x/2} < 2\lambda e^{-\lambda x} \iff e^{\lambda x/2} < 2 \iff x < \frac{2\log 2}{\lambda}.$$

(c) Under the classical reading the envelopes hold m and $2m$ with m fixed, and X and Y are these two amounts as assigned to you and your opponent by the fair flip. If $X = m$ your opponent necessarily holds the other envelope, so $Y = 2m = 2x$ with probability 1; if $X = 2m$ then $Y = m = x/2$ with probability 1. Since $Y = 2m$ or $Y = m$ according as $I = 1$ or $I = 0$,

$$E(Y) = \tfrac{1}{2}(2m) + \tfrac{1}{2}(m) = \frac{3m}{2} = E(X).$$

Problem 4.9. Prove that if the joint cdf of X and Y satisfies

$$F_{X,Y}(x, y) = F_X(x)F_Y(y),$$

then for any pair of intervals (a, b) and (c, d),

$$P(a \le X \le b,\ c \le Y \le d) = P(a \le X \le b)\, P(c \le Y \le d).$$

Solution. Inclusion-exclusion on four lower quadrants, then the assumed factorization: for any $s < b$ and $t < d$,

$$\begin{aligned} P(s < X \le b,\, t < Y \le d) &= F_{X,Y}(b, d) - F_{X,Y}(s, d) - F_{X,Y}(b, t) + F_{X,Y}(s, t) \\ &= F_X(b)F_Y(d) - F_X(s)F_Y(d) - F_X(b)F_Y(t) + F_X(s)F_Y(t) \\ &= \big(F_X(b) - F_X(s)\big)\big(F_Y(d) - F_Y(t)\big). \end{aligned}$$

Now close the rectangle. Take $s_n \uparrow a$ and $t_n \uparrow c$ strictly; since $\bigcap_n \{X > s_n\} = \{X \ge a\}$, the events $\{s_n < X \le b,\ t_n < Y \le d\}$ decrease to $\{a \le X \le b,\ c \le Y \le d\}$, so continuity of P from above (Axiom 3 of Definition 1.2.4, countable additivity) gives

$$\begin{aligned} P(a \le X \le b,\, c \le Y \le d) &= \big(F_X(b) - F_X(a^-)\big)\big(F_Y(d) - F_Y(c^-)\big) \\ &= P(a \le X \le b)\, P(c \le Y \le d), \end{aligned}$$

the last line by that same limit taken one coordinate at a time.

Problem 4.10. The random pair (X, Y) has the distribution given by the joint pmf below (rows indexed by $Y \in \{2, 3, 4\}$, columns by $X \in \{1, 2, 3\}$).

$Y \backslash X$	1	2	3
2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
3	$\frac{1}{6}$	0	$\frac{1}{6}$
4	0	$\frac{1}{3}$	0

(a) Show that X and Y are dependent.
(b) Give a probability table for random variables U and V that have the same marginals as X and Y but are independent.

Solution. (a) $P(X = 2, Y = 3) = 0$ but $P(X = 2)\, P(Y = 3) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} \neq 0$, so the joint pmf does not factor and Definition 4.2.5 fails.
Summing down columns and across rows, the marginals are

$$\begin{aligned} f_X(1) &= \tfrac{1}{12} + \tfrac{1}{6} + 0 = \tfrac{1}{4}, & f_X(2) &= \tfrac{1}{6} + 0 + \tfrac{1}{3} = \tfrac{1}{2}, \\ f_X(3) &= \tfrac{1}{12} + \tfrac{1}{6} + 0 = \tfrac{1}{4}, & f_Y(2) &= \tfrac{1}{12} + \tfrac{1}{6} + \tfrac{1}{12} = \tfrac{1}{3}, \\ f_Y(3) &= \tfrac{1}{6} + 0 + \tfrac{1}{6} = \tfrac{1}{3}, & f_Y(4) &= 0 + \tfrac{1}{3} + 0 = \tfrac{1}{3}. \end{aligned}$$

(b) Take $f_{U,V}(u, v) = f_X(u)f_Y(v)$, which is a pmf with the required marginals and is independent by construction (Lemma 4.2.7):

$V\backslash U$	1	2	3
2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
3	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
4	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$

The columns sum to $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ and the rows to $\frac{1}{3}$ each. (Check!)

Problem 4.11. Let U = the number of trials needed to get the first head and V = the number of trials needed to get two heads in repeated tosses of a fair coin. Are U and V independent random variables?

Solution. No. The second head cannot arrive before the first, so $V \geq U + 1$ with probability 1 and in particular

$$P(U = 2,\, V = 2) = 0,$$

while each marginal event has positive probability:

$$P(U = 2) = P(\mathrm{TH}) = \tfrac{1}{4}, \qquad P(V = 2) = P(\mathrm{HH}) = \tfrac{1}{4},$$

so $P(U = 2)P(V = 2) = \frac{1}{16} \neq 0$. The joint pmf therefore does not factor, and Definition 4.2.5 fails.

Problem 4.12. If a stick is broken at random into three pieces, what is the probability that the pieces can be put together in a triangle? (See Gardner 1961 for a complete discussion of this problem.)

Solution. 1/4. Model the stick as $[0, 1]$ broken at two independent uniform$(0, 1)$ points X and Y, so the pieces are $\min(X, Y)$, $|X - Y|$, $1 - \max(X, Y)$. Three nonnegative lengths summing to 1 satisfy all three triangle inequalities exactly when no piece exceeds the sum of the other two, i.e. exactly when every piece is less than $1/2$.
On the half $\{X < Y\}$ the pieces are X, $Y - X$, $1 - Y$, and the three constraints read

$$X < \tfrac{1}{2}, \qquad Y > \tfrac{1}{2}, \qquad Y - X < \tfrac{1}{2},$$

which cut out of the unit square the triangle with vertices $(0, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, 1)$ — a right triangle with legs $\frac{1}{2}$, hence area $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$. Since (X, Y) is uniform on the unit square and the region for $\{X > Y\}$ is its reflection in the diagonal,

$$P(\text{triangle}) = \tfrac{1}{8} + \tfrac{1}{8} = \tfrac{1}{4}.$$

Problem 4.13. Let X and Y be random variables with finite means.
(a) Show that

$$\min_{g(x)} E\big(Y - g(X)\big)^2 = E\big(Y - E(Y \mid X)\big)^2,$$

where $g(x)$ ranges over all functions. ($E(Y \mid X)$ is sometimes called the *regression of* Y *on* X, the "best" predictor of Y conditional on X.)
(b) Show that equation (2.2.3), $\min_b E(X - b)^2 = E(X - EX)^2$, can be derived as a special case of part (a).

Solution. (a) Add and subtract $E(Y \mid X)$ and expand:

$$\begin{aligned} E\big(Y - g(X)\big)^2 &= E\Big[\big(Y - E(Y \mid X)\big) + \big(E(Y \mid X) - g(X)\big)\Big]^2 \\ &= E\big(Y - E(Y \mid X)\big)^2 + E\big(E(Y \mid X) - g(X)\big)^2 \\ &\quad + 2E\Big[\big(Y - E(Y \mid X)\big)\big(E(Y \mid X) - g(X)\big)\Big]. \end{aligned}$$

The cross term vanishes. Condition on X using the tower property (Theorem 4.4.3, whose only hypothesis is that the expectations exist — assume the second moments are finite, else both sides are $+\infty$);

the factor $E(Y \mid X) - g(X)$ is a function of X, so it comes out of the inner conditional expectation:

$$E\Big[\big(Y - E(Y \mid X)\big)\big(E(Y \mid X) - g(X)\big)\Big]$$
$$= E\Big\{\big(E(Y \mid X) - g(X)\big)\, E\big[\,Y - E(Y \mid X) \;\big|\; X\,\big]\Big\} = E\{(\,\cdot\,)\cdot 0\} = 0,$$

because $E[Y \mid X] - E[E(Y \mid X) \mid X] = E(Y \mid X) - E(Y \mid X) = 0$. Hence

$$E\big(Y - g(X)\big)^2 = E\big(Y - E(Y \mid X)\big)^2 + E\big(E(Y \mid X) - g(X)\big)^2 \;\geq\; E\big(Y - E(Y \mid X)\big)^2,$$

with equality attained by $g(x) = E(Y \mid X = x)$, which is the asserted minimum.
(b) Take X degenerate, say $P(X = 1) = 1$. Every function $g(X)$ is then a constant b, and the conditional distribution of Y given $X = 1$ is the marginal (Definition 4.2.1 with $f_X(1) = 1$), so $E(Y \mid X) = EY$ and part (a) reads

$$\min_b E(Y - b)^2 = E(Y - EY)^2,$$

which is (2.2.3) with Y in the role of the book's X.

Problem 4.14. Suppose X and Y are independent $\mathrm{n}(0,1)$ random variables.
(a) Find $P(X^2 + Y^2 < 1)$.
(b) Find $P(X^2 < 1)$, after verifying that X^2 is distributed χ^2_1.

Solution. (a) $P(X^2 + Y^2 < 1) = 1 - e^{-1/2} \approx 0.3935$. By independence the joint pdf is the circularly symmetric $f(x,y) = (2\pi)^{-1}e^{-(x^2+y^2)/2}$, so passing to polar coordinates $x = r\cos\theta$, $y = r\sin\theta$ (Jacobian r),

$$P(X^2 + Y^2 < 1) = \int_0^{2\pi}\!\!\int_0^1 \frac{1}{2\pi} e^{-r^2/2}\, r\, dr\, d\theta$$
$$= \int_0^1 r\, e^{-r^2/2}\, dr = \Big[-e^{-r^2/2}\Big]_0^1 = 1 - e^{-1/2}.$$

(b) $X^2 \sim \chi^2_1$ is Example 2.1.9: $g(x) = x^2$ is monotone on $A_1 = (-\infty, 0)$ and $A_2 = (0, \infty)$ with $g_1^{-1}(t) = -\sqrt{t}$, $g_2^{-1}(t) = \sqrt{t}$, and $A_0 = \{0\}$ has probability 0, so Theorem 2.1.8 gives, for $t > 0$,

$$f_{X^2}(t) = \frac{1}{\sqrt{2\pi}} e^{-t/2}\Big|\frac{-1}{2\sqrt{t}}\Big| + \frac{1}{\sqrt{2\pi}} e^{-t/2}\Big|\frac{1}{2\sqrt{t}}\Big| = \frac{1}{\sqrt{2\pi}}\, t^{-1/2} e^{-t/2},$$

which is the gamma$(\frac{1}{2}, 2)$ pdf $\big(\Gamma(\frac{1}{2}) = \sqrt{\pi}\big)$, that is χ^2_1. Hence

$$P(X^2 < 1) = P(-1 < X < 1) = 2\Phi(1) - 1 \approx 0.6827.$$

4.3 Exercises 4.15–4.21

Problem 4.15. Let $X \sim \text{Poisson}(\theta)$, $Y \sim \text{Poisson}(\lambda)$, independent. It was shown in Theorem 4.3.2 that the distribution of $X + Y$ is Poisson$(\theta + \lambda)$. Show that the distribution of $X \mid X + Y$ is binomial with success probability $\theta/(\theta + \lambda)$. What is the distribution of $Y \mid X + Y$?

Solution. $X \mid X + Y = n \sim \text{binomial}(n, \theta/(\theta + \lambda))$. For $x = 0, 1, \ldots, n$, independence gives $P(X = x, X + Y = n) = P(X = x)P(Y = n - x)$, so by Definition 4.2.1 and Theorem 4.3.2,

$$P(X=x \mid X+Y=n) = \frac{\dfrac{e^{-\theta}\theta^x}{x!}\cdot\dfrac{e^{-\lambda}\lambda^{n-x}}{(n-x)!}}{\dfrac{e^{-(\theta+\lambda)}(\theta+\lambda)^n}{n!}}$$
$$= \binom{n}{x}\frac{\theta^x\lambda^{n-x}}{(\theta+\lambda)^n}$$
$$= \binom{n}{x}\left(\frac{\theta}{\theta+\lambda}\right)^x\left(1-\frac{\theta}{\theta+\lambda}\right)^{n-x},$$

the binomial$(n, \theta/(\theta+\lambda))$ pmf.
Since $Y = (X+Y) - X$, the same computation with the roles of θ and λ interchanged gives

$$Y \mid X+Y=n \;\sim\; \text{binomial}\left(n,\ \frac{\lambda}{\theta+\lambda}\right).$$

Problem 4.16. Let X and Y be independent random variables with the same geometric distribution.
(a) Show that U and V are independent, where U and V are defined by

$$U = \min(X,Y) \qquad \text{and} \qquad V = X - Y.$$

(b) Find the distribution of $Z = X/(X+Y)$, where we define $Z = 0$ if $X+Y=0$.
(c) Find the joint pmf of X and $X+Y$.

Solution. The convention in (b) forces the version of the geometric supported on $\{0,1,2,\dots\}$ (with the Section 3.2 support $\{1,2,\dots\}$ the event $\{X+Y=0\}$ is empty), so throughout $P(X=x) = pq^x$, $x = 0,1,2,\dots$, with $q = 1-p$, and likewise for Y.
(a) The joint pmf of (U,V) factors. If $v \ge 0$ then $\{U=u,\, V=v\} = \{X = u+v,\ Y=u\}$, and if $v<0$ then $\{U=u,\, V=v\} = \{X=u,\ Y=u-v\}$; either way a single pair (x,y) is involved and

$$P(U=u, V=v) \;=\; p^2q^{\,2u+|v|} \;=\; \underbrace{\left(p^2q^{2u}\right)}_{g(u)}\ \underbrace{q^{|v|}}_{h(v)},$$

for $u = 0,1,2,\dots$ and $v \in \mathbb{Z}$. The support is the product set $\{0,1,2,\dots\}\times\mathbb{Z}$, so by Lemma 4.2.7 U and V are independent. Summing out gives the marginals

$$P(U=u) = p^2q^{2u}\sum_{v\in\mathbb{Z}} q^{|v|} = p^2q^{2u}\,\frac{1+q}{1-q} = (1-q^2)(q^2)^u,$$
$$P(V=v) = q^{|v|}\,p^2\sum_{u\ge 0} q^{2u} = \frac{p}{1+q}\,q^{|v|},$$

so U is geometric with success probability $1-q^2$ and V has a two-sided geometric law. (Check that each sums to 1.)
(b) Write $z = a/b$ in lowest terms. For $0 < z \le 1$ the event $\{Z=z\}$ is the disjoint union

$$\{Z=z\} = \bigcup_{m\ge 1}\{X = ma,\ Y = m(b-a)\},$$

and each pair there has probability $p^2q^{ma}q^{m(b-a)} = p^2q^{mb}$, whence

$$P(Z=z) = \sum_{m\ge1} p^2q^{mb} = \frac{p^2q^b}{1-q^b}, \qquad z = \frac{a}{b} \in (0,1].$$

The value $z=0$ picks up the convention, since $\{Z=0\} = \{X=0\}$ (this already contains $\{X=Y=0\}$):

$$P(Z=0) = P(X=0) = p = \frac{p^2q}{1-q} + p^2,$$

the first term being what the displayed formula gives at $b = 1$.
(c) With $T = X + Y$, for $0 \le x \le t$,

$$P(X = x,\ T = t) = P(X = x)P(Y = t - x) = p^2 q^x q^{\,t-x} = p^2 q^t,$$

and 0 otherwise. Being free of x on each line $T = t$, this says T is negative binomial with $P(T = t) = (t + 1)p^2 q^t$ and $X \mid T = t$ is uniform on $\{0, 1, \ldots, t\}$.

Problem 4.17. Let X be an exponential(1) random variable, and define Y to be the integer part of $X + 1$, that is

$$Y = i + 1 \quad \text{if and only if} \quad i \le X < i + 1, \qquad i = 0, 1, 2, \ldots.$$

(a) Find the distribution of Y. What well-known distribution does Y have?
(b) Find the conditional distribution of $X - 4$ given $Y \ge 5$.

Solution. (a) Y is geometric with success probability $p = 1 - e^{-1}$. Indeed, for $i = 0, 1, 2, \ldots$,

$$P(Y = i + 1) = P(i \le X < i + 1) = e^{-i} - e^{-(i+1)} = \left(1 - e^{-1}\right)\left(e^{-1}\right)^i,$$

so writing $y = i + 1$,

$$P(Y = y) = p(1 - p)^{\,y-1}, \qquad y = 1, 2, \ldots, \quad p = 1 - e^{-1} \approx 0.6321,$$

which is the geometric(p) pmf of Section 3.2.
(b) $X - 4 \mid Y \ge 5 \ \sim$ exponential(1). The conditioning event is $\{Y \ge 5\} = \{X \ge 4\}$, and for $t > 0$

$$P(X - 4 > t \mid X \ge 4) = \frac{P(X > 4 + t)}{P(X \ge 4)} = \frac{e^{-(4+t)}}{e^{-4}} = e^{-t},$$

the exponential survival function; this is the memoryless property of the exponential (Section 3.3).

Problem 4.18. Given that $g(x) \ge 0$ has the property that

$$\int_0^\infty g(x)\,dx = 1,$$

show that

$$f(x, y) = \frac{2\,g\left(\sqrt{x^2 + y^2}\right)}{\pi\sqrt{x^2 + y^2}}, \qquad x, y > 0,$$

is a pdf.

Solution. Pass to polar coordinates. $f \ge 0$ on the open first quadrant $A = \{(x, y) : x, y > 0\}$ because $g \ge 0$, so only the total integral is at issue. The map

$$x = r\cos\theta, \qquad y = r\sin\theta$$

is one-to-one from $B = (0, \infty) \times (0, \pi/2)$ onto A, with Jacobian

$$J = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r \ > \ 0.$$

Hence, by the change of variables (4.3.2),

$$\begin{aligned}\int_0^\infty\int_0^\infty f(x,y)\,dx\,dy &= \int_0^{\pi/2}\int_0^\infty \frac{2g(r)}{\pi r}\,r\,dr\,d\theta\\ &= \frac{2}{\pi}\cdot\frac{\pi}{2}\int_0^\infty g(r)\,dr\\ &= 1.\end{aligned}$$

So f is a pdf on A.

Problem 4.19. (a) Let X_1 and X_2 be independent $\mathrm{n}(0,1)$ random variables. Find the pdf of $(X_1-X_2)^2/2$.
(b) If $X_i,\ i=1,2$, are independent $\mathrm{gamma}(\alpha_i,1)$ random variables, find the marginal distributions of $X_1/(X_1+X_2)$ and $X_2/(X_1+X_2)$.

Solution. (a) $(X_1-X_2)^2/2\sim\chi^2_1$, with pdf

$$f_U(u)=\frac{1}{\sqrt{2\pi}}\,u^{-1/2}e^{-u/2},\qquad u>0.$$

By Theorem 4.2.14, $X_1-X_2\sim\mathrm{n}(0,2)$, so $Z=(X_1-X_2)/\sqrt{2}\sim\mathrm{n}(0,1)$, and $U=Z^2$ is χ^2_1 by Example 2.1.9.
(b) $X_1/(X_1+X_2)\sim\mathrm{beta}(\alpha_1,\alpha_2)$ and $X_2/(X_1+X_2)\sim\mathrm{beta}(\alpha_2,\alpha_1)$. Put

$$U=X_1+X_2,\qquad V=\frac{X_1}{X_1+X_2},$$

a one-to-one map of $A=(0,\infty)^2$ onto $B=(0,\infty)\times(0,1)$ with inverse $x_1=uv$, $x_2=u(1-v)$ and

$$J=\begin{vmatrix} v & u\\ 1-v & -u\end{vmatrix}=-u,\qquad |J|=u.$$

Since $f_{X_1,X_2}(x_1,x_2)=x_1^{\alpha_1-1}x_2^{\alpha_2-1}e^{-(x_1+x_2)}/(\Gamma(\alpha_1)\Gamma(\alpha_2))$ on A, formula (4.3.2) gives

$$\begin{aligned}f_{U,V}(u,v) &= \frac{(uv)^{\alpha_1-1}\big(u(1-v)\big)^{\alpha_2-1}e^{-u}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\,u\\ &= \underbrace{\frac{u^{\alpha_1+\alpha_2-1}e^{-u}}{\Gamma(\alpha_1+\alpha_2)}}_{\mathrm{gamma}(\alpha_1+\alpha_2,1)\ \mathrm{pdf}}\\ &\qquad\times\underbrace{\frac{\Gamma(\alpha_1+\alpha_2)\,v^{\alpha_1-1}(1-v)^{\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}}_{\mathrm{beta}(\alpha_1,\alpha_2)\ \mathrm{pdf}}\end{aligned}$$

on the product set B. By Lemma 4.2.7 the two factors are the marginals (and U,V are independent), so $V=X_1/(X_1+X_2)\sim\mathrm{beta}(\alpha_1,\alpha_2)$. Interchanging the indices, $X_2/(X_1+X_2)=1-V\sim\mathrm{beta}(\alpha_2,\alpha_1)$.

Problem 4.20. X_1 and X_2 are independent $\mathrm{n}(0,\sigma^2)$ random variables.
(a) Find the joint distribution of Y_1 and Y_2, where

$$Y_1=X_1^2+X_2^2\qquad\text{and}\qquad Y_2=\frac{X_1}{\sqrt{Y_1}}.$$

(b) Show that Y_1 and Y_2 are independent, and interpret this result geometrically.

Solution. (a) The joint pdf is

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{2\sigma^2}e^{-y_1/(2\sigma^2)} \cdot \frac{1}{\pi\sqrt{1-y_2^2}}, \qquad y_1 > 0,\ -1 < y_2 < 1.$$

The transformation is two-to-one, so use (4.3.6) with $A_1 = \{x_2 > 0\}$ and $A_2 = \{x_2 < 0\}$ (the line $x_2 = 0$ has probability 0). On A_ε ($\varepsilon = \pm 1$) the inverse is

$$x_1 = y_2\sqrt{y_1}, \qquad x_2 = \varepsilon\sqrt{y_1}\sqrt{1-y_2^2},$$

mapping $B = (0,\infty)\times(-1,1)$ onto A_ε, with

$$J_\varepsilon = \begin{vmatrix} \dfrac{y_2}{2\sqrt{y_1}} & \sqrt{y_1} \\ \dfrac{\varepsilon\sqrt{1-y_2^2}}{2\sqrt{y_1}} & \dfrac{-\varepsilon y_2\sqrt{y_1}}{\sqrt{1-y_2^2}} \end{vmatrix}$$

$$= -\frac{\varepsilon}{2}\cdot\frac{y_2^2+(1-y_2^2)}{\sqrt{1-y_2^2}} = \frac{-\varepsilon}{2\sqrt{1-y_2^2}}.$$

Both inverses carry $x_1^2 + x_2^2 = y_1$ and both have $|J_\varepsilon| = \left(2\sqrt{1-y_2^2}\right)^{-1}$, so

$$f_{Y_1,Y_2}(y_1,y_2) = \sum_{\varepsilon=\pm1} \frac{1}{2\pi\sigma^2}e^{-y_1/(2\sigma^2)}\,\frac{1}{2\sqrt{1-y_2^2}}$$

$$= \frac{1}{2\pi\sigma^2}\,\frac{e^{-y_1/(2\sigma^2)}}{\sqrt{1-y_2^2}}$$

on B, which is the displayed factorization.
(b) The support B is a product set and the pdf factors as $g(y_1)h(y_2)$ above, so Y_1 and Y_2 are independent by Lemma 4.2.7; moreover each factor is already a pdf, so

$$Y_1 \sim \text{exponential}(2\sigma^2) = \sigma^2\chi_2^2, \qquad f_{Y_2}(y_2) = \frac{1}{\pi\sqrt{1-y_2^2}}, \quad |y_2| < 1.$$

Geometrically, in polar form $R = \sqrt{Y_1}$ and $Y_2 = \cos\Theta$ with Θ the angle to the x_1-axis, so squared distance from the origin carries no information about direction. That is the rotational symmetry of the spherical normal: its pdf depends on (x_1,x_2) only through $x_1^2+x_2^2$, forcing $\Theta \sim \text{uniform}(0,2\pi)$ whatever R is, and $\cos\Theta$ then has the arcsine pdf above.

Problem 4.21. A point is generated at random in the plane according to the following polar scheme. A radius R is chosen, where the distribution of R^2 is χ^2 with 2 degrees of freedom. Independently, an angle θ is chosen, where $\theta \sim \text{uniform}(0,2\pi)$. Find the joint distribution of $X = R\cos\theta$ and $Y = R\sin\theta$.

Solution. X and Y are independent n$(0,1)$:

$$f_{X,Y}(x,y) = \frac{1}{2\pi}\,e^{-(x^2+y^2)/2}, \qquad (x,y)\in\mathbb{R}^2.$$

Since $\chi_2^2 = \text{exponential}(2)$, $f_{R^2}(w) = \frac{1}{2}e^{-w/2}$ for $w > 0$; the map $r \mapsto r^2$ is increasing on $(0,\infty)$, so by Theorem 2.1.5

$$f_R(r) = \tfrac{1}{2}e^{-r^2/2}\cdot 2r = re^{-r^2/2}, \qquad r > 0.$$

By independence, $f_{R,\theta}(r,\theta) = re^{-r^2/2}/(2\pi)$ on $A = (0,\infty)\times(0,2\pi)$. The polar map $(r,\theta)\mapsto(r\cos\theta,\, r\sin\theta)$ is one-to-one from A onto the plane less the nonnegative x-axis (a set of measure zero), with inverse $r=\sqrt{x^2+y^2}$, and

$$J = \frac{\partial(r,\theta)}{\partial(x,y)} = \frac{1}{r} = \frac{1}{\sqrt{x^2+y^2}}.$$

Hence by (4.3.2),

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{\sqrt{x^2+y^2}\, e^{-(x^2+y^2)/2}}{2\pi}\cdot\frac{1}{\sqrt{x^2+y^2}} \\ &= \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\cdot\frac{1}{\sqrt{2\pi}}e^{-y^2/2}, \end{aligned}$$

a product of two $\mathrm{n}(0,1)$ pdfs on the product set $\mathbb{R}^2$, so by Lemma 4.2.7 X and Y are independent standard normals.

4.4 Exercises 4.22–4.28

Problem 4.22. Let (X,Y) be a bivariate random vector with joint pdf $f(x,y)$. Let $U=aX+b$ and $V=cY+d$, where a, b, c, and d are fixed constants with $a>0$ and $c>0$. Show that the joint pdf of (U,V) is

$$f_{U,V}(u,v) \;=\; \frac{1}{ac}\, f\!\left(\frac{u-b}{a},\, \frac{v-d}{c}\right).$$

Solution. Apply (4.3.2). Since $a>0$ and $c>0$, the map $(x,y)\mapsto(ax+b,\, cy+d)$ is one-to-one from $\mathcal{A}=\{(x,y) : f(x,y)>0\}$ onto its image $\mathcal{B}$, with unique inverse $x=h_1(u,v)=(u-b)/a$, $y=h_2(u,v)=(v-d)/c$ and Jacobian

$$J \;=\; \begin{vmatrix} 1/a & 0 \\ 0 & 1/c \end{vmatrix} \;=\; \frac{1}{ac} \;\neq\; 0,$$

so $|J|=1/(ac)$. Hence

$$f_{U,V}(u,v) \;=\; f\!\left(\frac{u-b}{a},\, \frac{v-d}{c}\right)\frac{1}{ac}$$

on $\mathcal{B}$; off $\mathcal{B}$ both sides vanish, since there $f\big((u-b)/a, (v-d)/c\big)=0$.

Problem 4.23. For X and Y as in Example 4.3.3, find the distribution of XY by making the transformations given in (a) and (b) and integrating out V.
(a) $U=XY$, $V=Y$
(b) $U=XY$, $V=X/Y$
(In Example 4.3.3, $X\sim\mathrm{beta}(\alpha,\beta)$ and $Y\sim\mathrm{beta}(\alpha+\beta,\gamma)$ are independent, so that

$$f_{X,Y}(x,y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}\frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha+\beta)\Gamma(\gamma)}y^{\alpha+\beta-1}(1-y)^{\gamma-1}$$

for $0<x<1$, $0<y<1$; the example uses $U=XY$, $V=X$ to show $XY\sim\mathrm{beta}(\alpha,\beta+\gamma)$.)

Solution. Both transformations return $XY\sim\mathrm{beta}(\alpha,\beta+\gamma)$, the answer of Example 4.3.3. Throughout write

$$C \;=\; \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)},$$

the constant left after the $\Gamma(\alpha+\beta)$ factors cancel in $f_{X,Y}$.
(a) $U=XY$, $V=Y$. On $\mathcal{A}=(0,1)^2$ the inverse is $x=u/v$, $y=v$, unique and one-to-one onto

$\mathcal{B} = \{(u,v) : 0 < u < v < 1\}$ (for fixed $v = y$, $u = xy$ ranges over $(0, v)$). The Jacobian is

$$J = \begin{vmatrix} 1/v & -u/v^2 \\ 0 & 1 \end{vmatrix} = \frac{1}{v}.$$

By (4.3.2),

$$\begin{aligned} f_{U,V}(u,v) &= C\left(\frac{u}{v}\right)^{\alpha-1}\Big(1-\frac{u}{v}\Big)^{\beta-1} v^{\alpha+\beta-1}(1-v)^{\gamma-1}\frac{1}{v} \\ &= C\,u^{\alpha-1}(v-u)^{\beta-1}(1-v)^{\gamma-1}, \end{aligned}$$

on $0 < u < v < 1$, every power of v cancelling. (Check!) Integrate out v with $v = u + (1-u)t$, $dv = (1-u)\,dt$:

$$\begin{aligned} \int_u^1 (v-u)^{\beta-1}(1-v)^{\gamma-1}\,dv &= (1-u)^{\beta+\gamma-1}\int_0^1 t^{\beta-1}(1-t)^{\gamma-1}dt \\ &= (1-u)^{\beta+\gamma-1}\frac{\Gamma(\beta)\Gamma(\gamma)}{\Gamma(\beta+\gamma)}. \end{aligned}$$

Hence

$$f_U(u) = \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta+\gamma)}u^{\alpha-1}(1-u)^{\beta+\gamma-1}, \qquad 0 < u < 1,$$

i.e. $XY \sim \text{beta}(\alpha, \beta+\gamma)$.
(b) $U = XY$, $V = X/Y$. Then $UV = X^2$ and $U/V = Y^2$, so with $X, Y > 0$ the inverse is

$$x = \sqrt{uv}, \qquad y = \sqrt{u/v},$$

one-to-one from $\mathcal{B} = \{(u,v) : 0 < u < 1,\ u < v < 1/u\}$ onto $\mathcal{A}$ (the constraints $x < 1$, $y < 1$ read $uv < 1$, $u < v$). The Jacobian is

$$J = \begin{vmatrix} \frac{1}{2}v^{1/2}u^{-1/2} & \frac{1}{2}u^{1/2}v^{-1/2} \\ \frac{1}{2}v^{-1/2}u^{-1/2} & -\frac{1}{2}u^{1/2}v^{-3/2} \end{vmatrix} = -\frac{1}{4v} - \frac{1}{4v} = -\frac{1}{2v},$$

so $|J| = 1/(2v)$ and, by (4.3.2),

$$\begin{aligned} f_{U,V}(u,v) = C\,&(uv)^{\frac{\alpha-1}{2}}\big(1-\sqrt{uv}\big)^{\beta-1} \\ &\times (u/v)^{\frac{\alpha+\beta-1}{2}}\big(1-\sqrt{u/v}\big)^{\gamma-1}\frac{1}{2v} \end{aligned}$$

on $\mathcal{B}$. Collecting the pure powers, $(uv)^{\frac{\alpha-1}{2}}(u/v)^{\frac{\alpha+\beta-1}{2}}\frac{1}{2v} = \frac{1}{2}u^{\alpha+\frac{\beta}{2}-1}v^{-\frac{\beta}{2}-1}$. Now integrate out v by $t = \sqrt{u/v}$, i.e. $v = u/t^2$, $|dv| = 2u\,t^{-3}dt$, which maps $v \in (u, 1/u)$ onto $t \in (u, 1)$ and gives $\sqrt{uv} = u/t$. The power factors become

$$\tfrac{1}{2}u^{\alpha+\frac{\beta}{2}-1}\big(u/t^2\big)^{-\frac{\beta}{2}-1}\,2u\,t^{-3} = u^{\alpha-1}t^{\beta-1},$$

so that

$$f_U(u) = C\,u^{\alpha-1}\int_u^1 t^{\beta-1}\Big(1-\frac{u}{t}\Big)^{\beta-1}(1-t)^{\gamma-1}dt = C\,u^{\alpha-1}\int_u^1 (t-u)^{\beta-1}(1-t)^{\gamma-1}dt.$$

This is the integral already evaluated in (a), so again

$$f_U(u) = \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta+\gamma)}u^{\alpha-1}(1-u)^{\beta+\gamma-1}, \qquad 0 < u < 1.$$

Problem 4.24. Let X and Y be independent random variables with $X \sim \text{gamma}(r,1)$ and $Y \sim \text{gamma}(s,1)$. Show that $Z_1 = X+Y$ and $Z_2 = X/(X+Y)$ are independent, and find the distribution

of each. (Z_1 is gamma and Z_2 is beta.)

Solution. $Z_1 \sim \text{gamma}(r+s, 1)$ and $Z_2 \sim \text{beta}(r, s)$, independent, because the transformed joint pdf factors.
The inverse of $z_1 = x+y$, $z_2 = x/(x+y)$ is $x = z_1 z_2$, $y = z_1(1-z_2)$, a one-to-one map of $\mathcal{B} = \{(z_1, z_2) : z_1 > 0,\ 0 < z_2 < 1\}$ onto $\mathcal{A} = \{(x, y) : x > 0,\ y > 0\}$, with Jacobian

$$J = \begin{vmatrix} z_2 & z_1 \\ 1 - z_2 & -z_1 \end{vmatrix} = -z_1, \qquad |J| = z_1.$$

By independence $f_{X,Y}(x, y) = x^{r-1}e^{-x}y^{s-1}e^{-y}/\big(\Gamma(r)\Gamma(s)\big)$ on $\mathcal{A}$, so (4.3.2) gives, for $(z_1, z_2) \in \mathcal{B}$,

$$\begin{aligned} f_{Z_1,Z_2}(z_1, z_2) &= \frac{(z_1 z_2)^{r-1}\big(z_1(1-z_2)\big)^{s-1}e^{-z_1}}{\Gamma(r)\Gamma(s)}\, z_1 \\ &= \underbrace{\frac{z_1^{r+s-1}e^{-z_1}}{\Gamma(r+s)}}_{\text{gamma}(r+s,1)} \cdot \underbrace{\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} z_2^{r-1}(1-z_2)^{s-1}}_{\text{beta}(r,s)}. \end{aligned}$$

The set $\mathcal{B}$ is a rectangle (a cross product), and the pdf is a function of z_1 times a function of z_2 on it, so Z_1 and Z_2 are independent by Lemma 4.2.7. Each factor is already a pdf, so they are the marginals: $Z_1 \sim \text{gamma}(r+s, 1)$, $Z_2 \sim \text{beta}(r, s)$.

Problem 4.25. Use the techniques of Section 4.3 to derive the joint distribution of (X, Y) from the joint distribution of (X, Z) in Examples 4.5.8 and 4.5.9.
In Example 4.5.8, $X \sim \text{uniform}(0, 1)$ and $Z \sim \text{uniform}(0, \frac{1}{10})$ are independent and $Y = X + Z$; the stated joint pdf is $f(x, y) = 10$ on $0 < x < 1$, $x < y < x + \frac{1}{10}$.
In Example 4.5.9, $X \sim \text{uniform}(-1, 1)$ and $Z \sim \text{uniform}(0, \frac{1}{10})$ are independent and $Y = X^2 + Z$; the stated joint pdf is $f(x, y) = 5$ on $-1 < x < 1$, $x^2 < y < x^2 + \frac{1}{10}$.

Solution. In both examples the transformation $(X, Z) \mapsto (X, Y)$ is a shear, so $|J| = 1$ and the joint pdf of (X, Y) is the joint pdf of (X, Z) with z replaced by its expression in (x, y).
For Example 4.5.8, independence gives $f_{X,Z}(x, z) = 1 \cdot 10 = 10$ on $\mathcal{A} = \{0 < x < 1,\ 0 < z < \frac{1}{10}\}$. Take $U = X$, $Y = X + Z$; the inverse $x = u$, $z = y - u$ is one-to-one onto

$$\mathcal{B} = \Big\{(u, y) : 0 < u < 1,\ u < y < u + \tfrac{1}{10}\Big\},$$

since $0 < z < \frac{1}{10}$ reads $u < y < u + \frac{1}{10}$. Its Jacobian is

$$J = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1.$$

So by (4.3.2), $f_{X,Y}(x, y) = f_{X,Z}(x,\, y - x) \cdot 1 = 10$ on $0 < x < 1$, $x < y < x + \frac{1}{10}$, as claimed.
For Example 4.5.9, now $f_{X,Z}(x, z) = \frac{1}{2} \cdot 10 = 5$ on $\mathcal{A} = \{-1 < x < 1,\ 0 < z < \frac{1}{10}\}$. Take $U = X$, $Y = X^2 + Z$; the inverse $x = u$, $z = y - u^2$ is one-to-one onto

$$\mathcal{B} = \Big\{(u, y) : -1 < u < 1,\ u^2 < y < u^2 + \tfrac{1}{10}\Big\},$$

with

$$J = \begin{vmatrix} 1 & 0 \\ -2u & 1 \end{vmatrix} = 1.$$

Hence $f_{X,Y}(x, y) = f_{X,Z}(x,\, y - x^2) = 5$ on $-1 < x < 1$, $x^2 < y < x^2 + \frac{1}{10}$, the region of Figure 4.5.2.

Problem 4.26. X and Y are independent random variables with $X \sim$ exponential(λ) and $Y \sim$ exponential(μ). It is impossible to obtain direct observations of X and Y. Instead, we observe the random variables Z and W, where

$$Z = \min\{X, Y\} \qquad \text{and} \qquad W = \begin{cases} 1 & \text{if } Z = X, \\ 0 & \text{if } Z = Y. \end{cases}$$

(This is a situation that arises, in particular, in medical experiments. The X and Y variables are *censored*.)
(a) Find the joint distribution of Z and W.
(b) Prove that Z and W are independent. (Hint: Show that $P(Z \le z \mid W = i) = P(Z \le z)$ for $i = 0$ or 1.)

Solution. (a) Z and W have the mixed joint distribution

$$f_{Z,W}(z,1) = \frac{1}{\lambda} e^{-\tau z}, \qquad f_{Z,W}(z,0) = \frac{1}{\mu} e^{-\tau z}, \qquad z > 0,$$

where $\tau = \dfrac{1}{\lambda} + \dfrac{1}{\mu} = \dfrac{\lambda + \mu}{\lambda\mu}$ and exponential(β) carries the Section 3.3 pdf $\beta^{-1} e^{-x/\beta}$, $x > 0$.
Indeed $\{Z > z,\ W = 1\} = \{z < X < Y\}$ (ties have probability 0), so by independence

$$\begin{aligned} P(Z > z,\, W = 1) &= \int_z^\infty \frac{1}{\lambda} e^{-x/\lambda}\, P(Y > x)\, dx \\ &= \int_z^\infty \frac{1}{\lambda} e^{-x/\lambda} e^{-x/\mu}\, dx = \frac{1}{\lambda\tau} e^{-\tau z}, \end{aligned}$$

and differentiating in $-z$ gives the stated $f_{Z,W}(z,1)$; interchanging the roles of X and Y gives $f_{Z,W}(z,0)$.
Summing over w recovers the marginals: with $z = 0$ above,

$$P(W = 1) = \frac{1}{\lambda\tau} = \frac{\mu}{\lambda + \mu}, \qquad P(W = 0) = \frac{\lambda}{\lambda + \mu},$$

and

$$P(Z > z) = P(X > z) P(Y > z) = e^{-z/\lambda} e^{-z/\mu} = e^{-\tau z},$$

so $Z \sim$ exponential$(1/\tau) =$ exponential$\big(\lambda\mu/(\lambda+\mu)\big)$ and $W \sim$ Bernoulli$\big(\mu/(\lambda+\mu)\big)$.
(b) The joint distribution found in (a) already factors:

$$f_{Z,W}(z,1) = \frac{1}{\lambda} e^{-\tau z} = \underbrace{\frac{\mu}{\lambda + \mu}}_{P(W=1)} \cdot \underbrace{\tau e^{-\tau z}}_{f_Z(z)},$$

and symmetrically $f_{Z,W}(z,0) = \frac{\lambda}{\lambda+\mu} \cdot \tau e^{-\tau z}$. Equivalently, in the form of the hint, for $i = 0, 1$,

$$P(Z \le z \mid W = i) = \frac{P(Z \le z,\, W = i)}{P(W = i)} = 1 - e^{-\tau z} = P(Z \le z),$$

the conditional cdf of Z not depending on i. Hence Z and W are independent by Definition 4.2.5.

Problem 4.27. Let $X \sim \mathrm{n}(\mu, \sigma^2)$ and let $Y \sim \mathrm{n}(\gamma, \sigma^2)$. Suppose X and Y are independent. Define $U = X + Y$ and $V = X - Y$. Show that U and V are independent normal random variables. Find the distribution of each of them.

Solution. $U \sim \mathrm{n}(\mu+\gamma,\, 2\sigma^2)$ and $V \sim \mathrm{n}(\mu-\gamma,\, 2\sigma^2)$, and they are independent.
Standardize: $X^* = (X-\mu)/\sigma$ and $Y^* = (Y-\gamma)/\sigma$ are independent (Theorem 4.3.5, applied to the functions $g(x) = (x-\mu)/\sigma$ and $h(y) = (y-\gamma)/\sigma$ of the independent X and Y) standard normal variables. Example 4.3.4 shows that

$$S = X^* + Y^* \quad \text{and} \quad D = X^* - Y^*$$

are independent, with $S \sim \mathrm{n}(0,2)$ and $D \sim \mathrm{n}(0,2)$. Since $X = \sigma X^* + \mu$ and $Y = \sigma Y^* + \gamma$,

$$U = \sigma S + (\mu+\gamma), \qquad V = \sigma D + (\mu-\gamma),$$

so U is a function of S alone and V a function of D alone; U and V are independent by Theorem 4.3.5 again. Finally a location-scale change of a normal is normal (Corollary 4.6.10 with $n=1$), giving

$$U \sim \mathrm{n}\big(\mu+\gamma,\, 2\sigma^2\big), \qquad V \sim \mathrm{n}\big(\mu-\gamma,\, 2\sigma^2\big).$$

Method (2): directly, $u = x+y$, $v = x-y$ inverts to $x = (u+v)/2$, $y = (u-v)/2$ on all of $\mathbb{R}^2$ with $J = -\frac{1}{2}$, so by (4.3.2)

$$\begin{aligned} f_{U,V}(u,v) &= \frac{1}{2\pi\sigma^2}\exp\left(-\frac{\big(\frac{u+v}{2}-\mu\big)^2 + \big(\frac{u-v}{2}-\gamma\big)^2}{2\sigma^2}\right)\frac{1}{2} \\ &= \frac{1}{\sqrt{4\pi\sigma^2}}e^{-\frac{(u-\mu-\gamma)^2}{4\sigma^2}} \cdot \frac{1}{\sqrt{4\pi\sigma^2}}e^{-\frac{(v-\mu+\gamma)^2}{4\sigma^2}}, \end{aligned}$$

the cross terms in uv cancelling exactly because $\operatorname{Var} X = \operatorname{Var} Y$. Lemma 4.2.7 then gives independence and identifies both marginals.

Problem 4.28. Let X and Y be independent standard normal random variables.
(a) Show that $X/(X+Y)$ has a Cauchy distribution.
(b) Find the distribution of $X/|Y|$.
(c) Is the answer to part (b) surprising? Can you formulate a general theorem?

Solution. (a) $X/(X+Y)$ is Cauchy with location $\frac{1}{2}$ and scale $\frac{1}{2}$, i.e. it has pdf

$$f(t) = \frac{1}{\pi}\cdot\frac{1/2}{(1/2)^2 + (t-1/2)^2} = \frac{2}{\pi\big(1+(2t-1)^2\big)}, \qquad -\infty < t < \infty.$$

Put $S = X+Y$ and $D = X-Y$. By Example 4.3.4, S and D are independent $\mathrm{n}(0,2)$ variables, and $X = (S+D)/2$, so

$$\frac{X}{X+Y} = \frac{(S+D)/2}{S} = \frac{1}{2} + \frac{1}{2}\cdot\frac{D}{S}.$$

Now $D/\sqrt{2}$ and $S/\sqrt{2}$ are independent $\mathrm{n}(0,1)$, so D/S is a ratio of independent standard normals and is standard Cauchy by Example 4.3.6. Applying Theorem 2.1.5 to the increasing map $c \mapsto \frac{1}{2} + \frac{1}{2}c$ shifts and scales the standard Cauchy pdf $1/\big(\pi(1+c^2)\big)$ into the display above.
(b) $X/|Y|$ is also standard Cauchy. Let $U = X/|Y|$ and $V = |Y|$. As in Example 4.3.6 the map is two-to-one: with $A_1 = \{y>0\}$, $A_2 = \{y<0\}$ and $A_0 = \{y=0\}$ (which has probability 0), the two inverses are

$$(x,y) = (uv,\ v) \quad \text{and} \quad (x,y) = (uv,\ -v),$$

each with $|J_i| = v$, and $\mathcal{B} = \{(u,v) : -\infty < u < \infty,\ v > 0\}$. Then (4.3.6) gives

$$f_{U,V}(u,v) = 2\cdot\frac{1}{2\pi}e^{-(uv)^2/2}e^{-v^2/2}\,v = \frac{v}{\pi}e^{-(u^2+1)v^2/2},$$

which is exactly the joint pdf obtained in Example 4.3.6. Hence, by the integration carried out there,

$$f_U(u) = \frac{1}{\pi(u^2+1)}, \qquad -\infty < u < \infty.$$

(c) Not surprising. Write $X/Y = (X/|Y|)\,\operatorname{sign}(Y)$: conditionally on $Y = y$, the variable X/y equals $X/|y|$ when $y > 0$ and $-X/|y|$ when $y < 0$, and $-X$ has the same distribution as X. So the conditional distribution of X/Y given $Y = y$ equals that of $X/|Y|$ given $Y = y$ for every $y \neq 0$, and the two ratios have the same distribution. The general theorem is:
If X and Y are independent and the distribution of X is symmetric about 0 (that is, $-X$ and X have the same distribution), and $P(Y = 0) = 0$, then X/Y and $X/|Y|$ have the same distribution.

4.5 Exercises 4.29–4.35

Problem 4.29. Jones (1999) looked at the distribution of functions of X and Y when $X = R\cos\theta$ and $Y = R\sin\theta$, where $\theta \sim \mathrm{U}(0, 2\pi)$ and R is a positive random variable. Here are two of the many situations that he considered.
(a) Show that X/Y has a Cauchy distribution.
(b) Show that the distribution of $(2XY)/\sqrt{X^2 + Y^2}$ is the same as the distribution of X. Specialize this result to one about $\mathrm{n}(0, \sigma^2)$ random variables.

Solution. (a) $X/Y = \cot\theta$, because $R > 0$ cancels. The function cot has period π and is a decreasing bijection from $(0, \pi)$ onto $\mathbb{R}$ (and likewise from $(\pi, 2\pi)$), so each of the two branches contributes equally and

$$f_{X/Y}(u) = 2 \cdot \frac{1}{2\pi}\left|\frac{d}{du}\operatorname{arccot} u\right| = \frac{1}{\pi(1+u^2)}, \qquad -\infty < u < \infty,$$

the Cauchy(0,1) pdf (3.3.19).
(b) (Part (b) needs the implicit hypothesis that R and θ are independent; part (a) does not.) Since $X^2 + Y^2 = R^2$ with $R > 0$,

$$\frac{2XY}{\sqrt{X^2+Y^2}} = \frac{2R^2\cos\theta\sin\theta}{R} = R\sin 2\theta.$$

Put $\psi = (2\theta - \pi/2) \bmod 2\pi$. Reduction mod 2π of any fixed affine image $a\theta + b$ with integer $a \neq 0$ sends $\mathrm{U}(0, 2\pi)$ to $\mathrm{U}(0, 2\pi)$, and ψ is a function of θ alone, so $\psi \sim \mathrm{U}(0, 2\pi)$ independent of R. Since $\sin 2\theta = \cos\psi$,

$$\frac{2XY}{\sqrt{X^2+Y^2}} = R\cos\psi \overset{d}{=} R\cos\theta = X.$$

Specialization: if X, Y are independent $\mathrm{n}(0, \sigma^2)$, their joint pdf $(2\pi\sigma^2)^{-1}e^{-(x^2+y^2)/(2\sigma^2)}$ is rotationally invariant, so in polar form $R = \sqrt{X^2+Y^2}$ and θ are independent with $\theta \sim \mathrm{U}(0, 2\pi)$. Hence

$$\frac{2XY}{\sqrt{X^2+Y^2}} \sim \mathrm{n}(0, \sigma^2).$$

Problem 4.30. Suppose the distribution of Y, conditional on $X = x$, is $\mathrm{n}(x, x^2)$ and that the marginal distribution of X is uniform$(0, 1)$.
(a) Find $\mathrm{E}Y$, $\operatorname{Var} Y$, and $\operatorname{Cov}(X, Y)$.
(b) Prove that Y/X and X are independent.

Solution. (a) $\mathrm{E}Y = 1/2$, $\operatorname{Var} Y = 5/12$, $\operatorname{Cov}(X, Y) = 1/12$. By Theorem 4.4.3 and Theorem 4.4.7, using $\mathrm{E}(Y|X) = X$ and $\operatorname{Var}(Y|X) = X^2$,

$$\begin{aligned}
\mathrm{E}Y &= \mathrm{E}\big(\mathrm{E}(Y|X)\big) = \mathrm{E}X = \tfrac{1}{2},\\
\operatorname{Var} Y &= \mathrm{E}\big(\operatorname{Var}(Y|X)\big) + \operatorname{Var}\big(\mathrm{E}(Y|X)\big)\\
&= \mathrm{E}X^2 + \operatorname{Var} X = \tfrac{1}{3} + \tfrac{1}{12} = \tfrac{5}{12},\\
\mathrm{E}(XY) &= \mathrm{E}\big(X\,\mathrm{E}(Y|X)\big) = \mathrm{E}X^2 = \tfrac{1}{3},
\end{aligned}$$

so $\operatorname{Cov}(X, Y) = \frac{1}{3} - \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{12}$.

(b) On $0 < x < 1$ we have $x > 0$, so conditionally on $X = x$ the variable $Y/X = Y/x$ is a positive rescaling of an $\mathrm{n}(x, x^2)$ variable:

$$\frac{Y}{X}\,\Big|\, X = x \;\sim\; \mathrm{n}\left(\frac{x}{x}, \frac{x^2}{x^2}\right) = \mathrm{n}(1,1).$$

The conditional distribution of Y/X given $X = x$ does not depend on x, hence Y/X is independent of X (and $Y/X \sim \mathrm{n}(1,1)$).

Problem 4.31. Suppose that the random variable Y has a binomial distribution with n trials and success probability X, where n is a given constant and X is a uniform$(0,1)$ random variable.
(a) Find $\mathrm{E}Y$ and $\operatorname{Var} Y$.
(b) Find the joint distribution of X and Y.
(c) Find the marginal distribution of Y.

Solution. (a) $\mathrm{E}Y = n/2$ and $\operatorname{Var} Y = n/6 + n^2/12$. By Theorems 4.4.3 and 4.4.7 with $\mathrm{E}(Y|X) = nX$, $\operatorname{Var}(Y|X) = nX(1-X)$,

$$\begin{aligned}\mathrm{E}Y &= \mathrm{E}(nX) = \frac{n}{2},\\ \operatorname{Var} Y &= \mathrm{E}\big(nX(1-X)\big) + \operatorname{Var}(nX)\\ &= n\left(\tfrac{1}{2} - \tfrac{1}{3}\right) + \frac{n^2}{12} = \frac{n}{6} + \frac{n^2}{12}.\end{aligned}$$

(b) The joint distribution is the mixed pdf-pmf obtained from $f(x,y) = P(Y = y|X = x) f_X(x)$:

$$f(x,y) = \binom{n}{y} x^y (1-x)^{n-y}, \qquad 0 < x < 1,\ y = 0, 1, \ldots, n.$$

(c) Y is discrete uniform on $\{0, 1, \ldots, n\}$: integrating out x against the beta kernel,

$$P(Y = y) = \binom{n}{y} \int_0^1 x^y (1-x)^{n-y} dx = \binom{n}{y} \frac{y!\,(n-y)!}{(n+1)!} = \frac{1}{n+1}$$

for $y = 0, 1, \ldots, n$.

Problem 4.32. (a) For the hierarchical model

$$Y|\Lambda \sim \text{Poisson}(\Lambda) \qquad \text{and} \qquad \Lambda \sim \text{gamma}(\alpha, \beta)$$

find the marginal distribution, mean, and variance of Y. Show that the marginal distribution of Y is a negative binomial if α is an integer.
(b) Show that the three-stage model

$$Y|N \sim \text{binomial}(N, p), \quad N|\Lambda \sim \text{Poisson}(\Lambda), \quad \Lambda \sim \text{gamma}(\alpha, \beta)$$

leads to the same marginal (unconditional) distribution of Y.

Solution. (a) $Y \sim$ negative binomial$\big(\alpha,\, p = 1/(1+\beta)\big)$, with $\mathrm{E}Y = \alpha\beta$ and $\operatorname{Var} Y = \alpha\beta(1+\beta)$.

Integrating out λ against a gamma kernel,

$$\begin{aligned} P(Y=y) &= \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!}\,\frac{\lambda^{\alpha-1}e^{-\lambda/\beta}}{\Gamma(\alpha)\beta^\alpha}\,d\lambda \\ &= \frac{1}{y!\,\Gamma(\alpha)\beta^\alpha}\,\Gamma(y+\alpha)\left(\frac{\beta}{1+\beta}\right)^{y+\alpha} \\ &= \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\,y!}\left(\frac{1}{1+\beta}\right)^{\alpha}\left(\frac{\beta}{1+\beta}\right)^{y}, \end{aligned}$$

for $y = 0, 1, 2, \ldots$. With $p = 1/(1+\beta)$ this is $\frac{\Gamma(y+\alpha)}{\Gamma(\alpha)y!}p^\alpha(1-p)^y$, and for integer α the ratio of gamma functions is $\binom{\alpha+y-1}{y}$, giving exactly the negative binomial pmf (3.2.10) with $r = \alpha$. By Theorems 4.4.3 and 4.4.7,

$$\begin{aligned} \mathrm{E}Y &= \mathrm{E}\Lambda = \alpha\beta, \\ \mathrm{Var}\,Y &= \mathrm{E}\Lambda + \mathrm{Var}\,\Lambda = \alpha\beta + \alpha\beta^2 = \alpha\beta(1+\beta). \end{aligned}$$

(b) Collapse the top two stages by the binomial-Poisson computation of Example 4.4.2, which gives $Y|\Lambda \sim \text{Poisson}(p\Lambda)$. Since $\Lambda \sim \text{gamma}(\alpha, \beta)$ and $p > 0$ is a scale factor, $p\Lambda \sim \text{gamma}(\alpha, p\beta)$. Thus the three-stage model is the two-stage model

$$Y|\Lambda' \sim \text{Poisson}(\Lambda'), \qquad \Lambda' \sim \text{gamma}(\alpha, p\beta),$$

and part (a) gives the same marginal family,

$$P(Y=y) = \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\,y!}\left(\frac{1}{1+p\beta}\right)^{\alpha}\left(\frac{p\beta}{1+p\beta}\right)^{y},$$

i.e. negative binomial with $r = \alpha$ and success probability $1/(1+p\beta)$ — identical to (a) after the reparametrization $\beta \mapsto p\beta$.

Problem 4.33. (Alternative derivation of the negative binomial distribution) Solomon (1983) details the following biological model. Suppose that each of a random number, N, of insects lays X_i eggs, where the X_is are independent, identically distributed random variables. The total number of eggs laid is $H = X_1 + \cdots + X_N$. What is the distribution of H? It is common to assume that N is Poisson(λ). Furthermore, if we assume that each X_i has the logarithmic series distribution (see Exercise 3.14) with success probability p, we have the hierarchical model

$$H|N = X_1 + \cdots + X_N, \qquad P(X_i = t) = \frac{-1}{\log(p)}\,\frac{(1-p)^t}{t},$$

for $t = 1, 2, \ldots$, and

$$N \sim \text{Poisson}(\lambda).$$

Show that the marginal distribution of H is negative binomial(r, p), where $r = -\lambda/\log(p)$. (It is easiest to calculate and identify the mgf of H using Theorems 4.4.3 and 4.6.7. Stuart and Ord 1987, Section 5.21, also mention this derivation of the logarithmic series distribution. They refer to H as a randomly stopped sum.)

Solution. The mgf of H is $\big(p/(1-(1-p)e^t)\big)^r$ with $r = -\lambda/\log p$, which is the negative binomial(r, p) mgf.
First, summing the series $\sum_{x\ge1} z^x/x = -\log(1-z)$ at $z = (1-p)e^t$,

$$M_X(t) = \frac{-1}{\log p}\sum_{x=1}^{\infty}\frac{\big((1-p)e^t\big)^x}{x} = \frac{\log\big(1-(1-p)e^t\big)}{\log p},$$

finite for $t < -\log(1-p)$, so all the mgfs below exist in a neighbourhood of 0.

The X_i are independent of each other and of N, so Theorem 4.6.7 gives $\mathrm{E}(e^{tH}|N=n) = \big(M_X(t)\big)^n$ for every $n \geq 0$, and Theorem 4.4.3 then gives, with the Poisson(λ) mgf $\mathrm{E}s^N = e^{\lambda(s-1)}$ at $s = M_X(t) > 0$,

$$\begin{aligned} M_H(t) &= \mathrm{E}\Big(\big(M_X(t)\big)^N\Big) = \exp\big\{\lambda\big(M_X(t)-1\big)\big\} \\ &= \exp\left\{\frac{\lambda}{\log p}\log\big(1-(1-p)e^t\big) - \lambda\right\} \\ &= \exp\big\{-r\log\big(1-(1-p)e^t\big) + r\log p\big\} \\ &= \left(\frac{p}{1-(1-p)e^t}\right)^r, \end{aligned}$$

using $\lambda = -r\log p$. This is the mgf of the negative binomial pmf (3.2.10) with parameters r and p, so by Theorem 2.3.11 $H \sim$ negative binomial(r,p).

Problem 4.34. (a) For the hierarchy in Example 4.4.6, that is,

$$X|P \sim \text{binomial}(n,P), \qquad P \sim \text{beta}(\alpha,\beta),$$

show that the marginal distribution of X is given by the beta-binomial distribution,

$$P(X=x) = \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}.$$

(b) A variation on the hierarchical model in part (a) is

$$X|P \sim \text{negative binomial}(r,P) \qquad \text{and} \qquad P \sim \text{beta}(\alpha,\beta).$$

Find the marginal pmf of X and its mean and variance. (This distribution is the beta-Pascal.)

Solution. (a) Integrate out p; the integrand is a beta($x+\alpha$, $n-x+\beta$) kernel:

$$\begin{aligned} P(X=x) &= \int_0^1 \binom{n}{x} p^x(1-p)^{n-x}\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,dp \\ &= \binom{n}{x}\frac{B(x+\alpha,\, n-x+\beta)}{B(\alpha,\beta)} \\ &= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}, \end{aligned}$$

for $x = 0, 1, \ldots, n$.
(b) With the pmf (3.2.10), $P(X=x|p) = \binom{r+x-1}{x}p^r(1-p)^x$ for $x = 0,1,2,\ldots$, the same integration gives the beta-Pascal pmf

$$\begin{aligned} P(X=x) &= \binom{r+x-1}{x}\frac{B(r+\alpha,\, x+\beta)}{B(\alpha,\beta)} \\ &= \frac{\Gamma(r+x)}{\Gamma(r)\,x!}\,\frac{\Gamma(\alpha+\beta)\Gamma(r+\alpha)\Gamma(x+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(r+\alpha+x+\beta)}. \end{aligned}$$

For the moments, use $\mathrm{E}(X|P) = r(1-P)/P$ and $\mathrm{Var}(X|P) = r(1-P)/P^2$, together with the beta moment formula (valid whenever $\alpha - k > 0$)

$$\mathrm{E}\left[\frac{(1-P)^j}{P^k}\right] = \frac{B(\alpha-k,\beta+j)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha-k)\Gamma(\beta+j)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+j-k)}.$$

The three instances needed are

$$\begin{aligned}(j,k)=(1,1): &\quad \frac{\beta}{\alpha-1},\\ (j,k)=(1,2): &\quad \frac{\beta(\alpha+\beta-1)}{(\alpha-1)(\alpha-2)},\\ (j,k)=(2,2): &\quad \frac{\beta(\beta+1)}{(\alpha-1)(\alpha-2)}.\end{aligned}$$

(Check!) Hence by Theorem 4.4.3, for $\alpha > 1$,

$$\mathrm{E}X = r\,\mathrm{E}\left[\frac{1-P}{P}\right] = \frac{r\beta}{\alpha-1},$$

and by Theorem 4.4.7, for $\alpha > 2$,

$$\begin{aligned}\operatorname{Var} X &= r\,\mathrm{E}\left[\frac{1-P}{P^2}\right] + r^2\left(\mathrm{E}\left[\frac{(1-P)^2}{P^2}\right] - \frac{\beta^2}{(\alpha-1)^2}\right)\\ &= \frac{r\beta(\alpha+\beta-1)}{(\alpha-1)(\alpha-2)} + r^2\,\frac{\beta(\alpha+\beta-1)}{(\alpha-1)^2(\alpha-2)}\\ &= \frac{r\beta(\alpha+\beta-1)(\alpha-1+r)}{(\alpha-1)^2(\alpha-2)},\end{aligned}$$

where the middle line used $\beta(\beta+1)(\alpha-1) - \beta^2(\alpha-2) = \beta(\alpha+\beta-1)$.

Problem 4.35. (a) For the hierarchy in Example 4.4.6, that is, $X|P \sim \text{binomial}(n,P)$ and $P \sim \text{beta}(\alpha,\beta)$, show that the variance of X can be written

$$\operatorname{Var} X = n\,\mathrm{E}P(1-\mathrm{E}P) + n(n-1)\operatorname{Var} P.$$

(The first term reflects binomial variation with success probability $\mathrm{E}P$, and the second term is often called "extra-binomial" variation, showing how the hierarchical model has a variance that is larger than the binomial alone.)
(b) For the hierarchy in Exercise 4.32, that is, $Y|\Lambda \sim \text{Poisson}(\Lambda)$ and $\Lambda \sim \text{gamma}(\alpha,\beta)$, show that the variance of Y can be written

$$\operatorname{Var} Y = \mathrm{E}\Lambda + \operatorname{Var}\Lambda = \mu + \frac{1}{\alpha}\mu^2,$$

where $\mu = \mathrm{E}\Lambda$. Identify the "extra-Poisson" variation induced by the hierarchy.

Solution. (a) Apply Theorem 4.4.7 with $\mathrm{E}(X|P) = nP$ and $\operatorname{Var}(X|P) = nP(1-P)$, then substitute $\mathrm{E}P^2 = \operatorname{Var} P + (\mathrm{E}P)^2$:

$$\begin{aligned}\operatorname{Var} X &= \mathrm{E}\big(nP(1-P)\big) + \operatorname{Var}(nP)\\ &= n\mathrm{E}P - n\big(\operatorname{Var} P + (\mathrm{E}P)^2\big) + n^2\operatorname{Var} P\\ &= n\,\mathrm{E}P(1-\mathrm{E}P) + n(n-1)\operatorname{Var} P.\end{aligned}$$

Here $\mathrm{E}P = \alpha/(\alpha+\beta)$ and $\operatorname{Var} P = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$.
(b) Theorem 4.4.7 with $\mathrm{E}(Y|\Lambda) = \operatorname{Var}(Y|\Lambda) = \Lambda$ gives

$$\operatorname{Var} Y = \mathrm{E}\Lambda + \operatorname{Var}\Lambda = \alpha\beta + \alpha\beta^2,$$

and with $\mu = \mathrm{E}\Lambda = \alpha\beta$ we have $\alpha\beta^2 = (\alpha\beta)^2/\alpha = \mu^2/\alpha$, so

$$\operatorname{Var} Y = \mu + \frac{1}{\alpha}\mu^2.$$

The extra-Poisson variation is $\operatorname{Var}\Lambda = \mu^2/\alpha$, the amount by which the marginal variance exceeds that of a Poisson(μ) variable.

4.6 Exercises 4.36–4.42

Problem 4.36. One generalization of the Bernoulli trials hierarchy in Example 4.4.6 is to allow the success probability to vary from trial to trial, keeping the trials independent. A standard model for this situation is

$$X_i \mid P_i \sim \text{Bernoulli}(P_i), \quad i = 1, \dots, n,$$
$$P_i \sim \text{beta}(\alpha, \beta).$$

This model might be appropriate, for example, if we are measuring the success of a drug on n patients and, because the patients are different, we are reluctant to assume that the success probabilities are constant. (This can be thought of as an *empirical Bayes model*; see Miscellanea 7.5.6.)
A random variable of interest is $Y = \sum_{i=1}^n X_i$, the total number of successes.
(a) Show that $\mathrm{E}Y = n\alpha/(\alpha+\beta)$.
(b) Show that $\operatorname{Var} Y = n\alpha\beta/(\alpha+\beta)^2$, and hence Y has the same mean and variance as a binomial$\Big(n, \frac{\alpha}{\alpha+\beta}\Big)$ random variable. What is the distribution of Y?
(c) Suppose now that the model is

$$X_i \mid P_i \sim \text{binomial}(n_i, P_i), \quad i = 1, \dots, k,$$
$$P_i \sim \text{beta}(\alpha, \beta).$$

Show that for $Y = \sum_{i=1}^k X_i$, $\mathrm{E}Y = \frac{\alpha}{\alpha+\beta}\sum_{i=1}^k n_i$ and $\operatorname{Var} Y = \sum_{i=1}^k \operatorname{Var} X_i$, where

$$\operatorname{Var} X_i = n_i \frac{\alpha\beta(\alpha+\beta+n_i)}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Solution. $Y \sim$ binomial$\Big(n, \frac{\alpha}{\alpha+\beta}\Big)$ exactly: the pairs (X_i, P_i) are independent across i, and each X_i is a marginally Bernoulli variable with

$$P(X_i = 1) = \mathrm{E}\big(P(X_i = 1 \mid P_i)\big) = \mathrm{E}P_i = \frac{\alpha}{\alpha+\beta},$$

by Theorem 4.4.3 and the beta mean of Section 3.3.
(a) By Theorem 4.4.3,

$$\mathrm{E}Y = \sum_{i=1}^n \mathrm{E}\big(\mathrm{E}(X_i \mid P_i)\big) = \sum_{i=1}^n \mathrm{E}P_i = \frac{n\alpha}{\alpha+\beta}.$$

(b) Write $\mu = \alpha/(\alpha+\beta)$ and recall $\operatorname{Var} P_i = \alpha\beta/\big[(\alpha+\beta)^2(\alpha+\beta+1)\big]$. Since $\mathrm{E}P_i - \mathrm{E}P_i^2 = \mu - \mu^2 - \operatorname{Var} P_i$ and $\mu - \mu^2 = \alpha\beta/(\alpha+\beta)^2$, the conditional variance identity (Theorem 4.4.7) gives

$$\begin{aligned}\operatorname{Var} X_i &= \mathrm{E}\big(P_i(1-P_i)\big) + \operatorname{Var}(P_i) \\ &= \frac{\alpha\beta}{(\alpha+\beta)^2} - \operatorname{Var} P_i + \operatorname{Var} P_i \\ &= \frac{\alpha\beta}{(\alpha+\beta)^2} = \mu(1-\mu).\end{aligned}$$

The X_i are independent (they are functions of the independent pairs (X_i, P_i)), so $\operatorname{Var} Y = n\alpha\beta/(\alpha+\beta)^2 = n\mu(1-\mu)$, the binomial$(n, \mu)$ variance.
(c) Now $\mathrm{E}(X_i \mid P_i) = n_i P_i$ and $\operatorname{Var}(X_i \mid P_i) = n_i P_i(1-P_i)$, so by Theorem 4.4.3,

$$\mathrm{E}Y = \sum_{i=1}^k n_i \,\mathrm{E}P_i = \frac{\alpha}{\alpha+\beta}\sum_{i=1}^k n_i,$$

and by Theorem 4.4.7, using $\mathrm{E}\big(P_i(1-P_i)\big) = \alpha\beta/\big[(\alpha+\beta)(\alpha+\beta+1)\big]$ (the calculation in (b)),

$$\begin{aligned}\operatorname{Var} X_i &= n_i \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} + n_i^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{n_i\,\alpha\beta\,\big[(\alpha+\beta)+n_i\big]}{(\alpha+\beta)^2(\alpha+\beta+1)}.\end{aligned}$$

The X_i are again independent, so $\operatorname{Var} Y = \sum_{i=1}^{k} \operatorname{Var} X_i$.

Problem 4.37. A generalization of the hierarchy in Exercise 4.34 is described by D. G. Morrison (1978), who gives a model for *forced binary choices*. A forced binary choice occurs when a person is forced to choose between two alternatives, as in a taste test. It may be that a person cannot actually discriminate between the two choices (can you tell Coke from Pepsi?), but the setup of the experiment is such that a choice must be made. Therefore, there is a confounding between discriminating correctly and guessing correctly. Morrison modeled this by defining the following parameters:

$$\begin{aligned}p &= \text{probability that a person can actually discriminate,} \\ c &= \text{probability that a person discriminates correctly.}\end{aligned}$$

Then

$$c = p + \tfrac{1}{2}(1-p) = \tfrac{1}{2}(1+p), \qquad \tfrac{1}{2} < c < 1,$$

where $\frac{1}{2}(1-p)$ is the probability that a person guesses correctly. We now run the experiment and observe $X_1, \ldots, X_n \sim \text{Bernoulli}(c)$, so

$$P\Big(\sum X_i = k \,\Big|\, c\Big) = \tbinom{n}{k} c^k (1-c)^{n-k}.$$

However, it is probably the case that p is not constant from person to person, so p is allowed to vary according to a beta distribution,

$$P \sim \text{beta}(a, b).$$

(a) Show that the distribution of $\sum X_i$ is a mixture of beta-binomials.
(b) Find the mean and variance of $\sum X_i$.

Solution. (a) $S = \sum_{i=1}^{n} X_i$ is a binomial$(n, \frac{1}{2})$ mixture of shifted beta-binomials. Split each trial into its two mechanisms: let $U_i \sim \text{Bernoulli}(P)$ record that the person discriminates and $V_i \sim \text{Bernoulli}(1/2)$ record that a guess would be correct, all independent given P, and set $X_i = U_i + (1-U_i)V_i$, so that $P(X_i = 1 \mid P) = P + \frac{1}{2}(1-P) = \frac{1}{2}(1+P) = c$. The number $M = \sum_{i=1}^{n} V_i \sim \text{binomial}(n, \frac{1}{2})$ is independent of P; the M trials with $V_i = 1$ are correct outright and the remaining $n - M$ contribute $\sum U_i$, so

$$S \mid M = m,\ P = p \ \sim\ m + \text{binomial}(n-m,\, p),$$

and averaging over $P \sim \text{beta}(a, b)$ (Example 4.4.6),

$$S \mid M = m \ \sim\ m + \text{beta-binomial}(n-m,\, a,\, b).$$

Explicitly,

$$P(S = k) = \sum_{m=0}^{k} \binom{n}{m} \frac{1}{2^n} \binom{n-m}{k-m} \frac{B(a+k-m,\, b+n-k)}{B(a,b)}.$$

That this is the stated model, and not merely a construction matching it in one coordinate, follows

from $\binom{n}{m}\binom{n-m}{k-m} = \binom{n}{k}\binom{k}{m}$:

$$\sum_{m=0}^{k} \binom{n}{m}\binom{n-m}{k-m} p^{k-m}(1-p)^{n-k} = \binom{n}{k}(1-p)^{n-k} \sum_{m=0}^{k} \binom{k}{m} p^{k-m}$$
$$= \binom{n}{k}(1+p)^k(1-p)^{n-k},$$

which upon division by 2^n is exactly $\binom{n}{k}c^k(1-c)^{n-k}$ with $c = (1+p)/2$.
(b) Write $\mu = \mathrm{E}P = a/(a+b)$ and $\sigma^2 = \mathrm{Var}\,P = ab/\big[(a+b)^2(a+b+1)\big]$. Since $\mathrm{E}(S \mid P) = n(1+P)/2$, Theorem 4.4.3 gives

$$\mathrm{E}S = \frac{n}{2}(1+\mu) = \frac{n(2a+b)}{2(a+b)}.$$

With $\mathrm{Var}(S \mid P) = n\frac{1+P}{2}\cdot\frac{1-P}{2} = \frac{n}{4}(1-P^2)$, Theorem 4.4.7 gives

$$\mathrm{Var}\,S = \frac{n}{4}\big(1-\mu^2-\sigma^2\big) + \frac{n^2}{4}\sigma^2$$
$$= \frac{n\,b(2a+b)}{4(a+b)^2} + \frac{n(n-1)\,ab}{4(a+b)^2(a+b+1)},$$

using $1-\mu^2 = b(2a+b)/(a+b)^2$.

Problem 4.38. (*The gamma as a mixture of exponentials*) Gleser (1989) shows that, in certain cases, the gamma distribution can be written as a scale mixture of exponentials, an identity suggested by different analyses of the same data. Let $f(x)$ be a gamma(r,λ) pdf.
(a) Show that if $r \le 1$, then $f(x)$ can be written

$$f(x) = \int_0^\lambda \frac{1}{\nu} e^{-x/\nu}\, p_\lambda(\nu)\, d\nu,$$

where

$$p_\lambda(\nu) = \frac{1}{\Gamma(r)\Gamma(1-r)}\,\frac{\nu^{\,r-1}}{(\lambda-\nu)^r}, \qquad 0 < \nu < \lambda.$$

(*Hint:* Make a change of variable from ν to u, where $u = x/\nu - x/\lambda$.)
(b) Show that $p_\lambda(\nu)$ is a pdf, for $r \le 1$, by showing that

$$\int_0^\lambda p_\lambda(\nu)\, d\nu = 1.$$

(c) Show that the restriction $r \le 1$ is necessary for the representation in part (a) to be valid; that is, there is no such representation if $r > 1$. (*Hint:* Suppose $f(x)$ can be written $f(x) = \int_0^\infty \big(e^{-x/\nu}/\nu\big) q_\lambda(\nu)\, d\nu$ for some pdf $q_\lambda(\nu)$. Show that $\frac{\partial}{\partial x}\log\big(f(x)\big) > 0$ but $\frac{\partial}{\partial x}\log\big(\int_0^\infty (e^{-x/\nu}/\nu) q_\lambda(\nu)\, d\nu\big) < 0$, a contradiction.)

Solution. (a) The hinted substitution collapses the integral to $\Gamma(1-r)$. Here $f(x) = x^{r-1}e^{-x/\lambda}/\big(\Gamma(r)\lambda^r\big)$, $x > 0$, is the gamma(r,λ) pdf of (3.3.6), and $0 < r < 1$ throughout (a) and (b); at $r = 1$ the gamma is itself exponential and p_λ degenerates to the point mass at $\nu = \lambda$. The map $u = x/\nu - x/\lambda$ carries $\nu \in (0,\lambda)$ decreasingly onto $u \in (0,\infty)$, with

$$du = -\frac{x}{\nu^2}\, d\nu, \qquad \lambda - \nu = \frac{u\,\nu\lambda}{x}, \qquad e^{-x/\nu} = e^{-u}e^{-x/\lambda}.$$

Substituting these into the integrand and reversing the orientation,

$$\begin{aligned}\frac{1}{\nu}e^{-x/\nu}\frac{\nu^{r-1}}{(\lambda-\nu)^r}\,d\nu &= e^{-u}e^{-x/\lambda}\,\nu^{r-2}\Big(\frac{u\nu\lambda}{x}\Big)^{-r}\frac{\nu^2}{x}\,du\\ &= e^{-x/\lambda}\,x^{r-1}\lambda^{-r}\,u^{-r}e^{-u}\,du,\end{aligned}$$

every power of ν having cancelled. Hence

$$\begin{aligned}\int_0^\lambda \frac{1}{\nu}e^{-x/\nu}p_\lambda(\nu)\,d\nu &= \frac{x^{r-1}e^{-x/\lambda}\lambda^{-r}}{\Gamma(r)\Gamma(1-r)}\int_0^\infty u^{-r}e^{-u}\,du\\ &= \frac{x^{r-1}e^{-x/\lambda}}{\Gamma(r)\lambda^r} = f(x),\end{aligned}$$

since $\int_0^\infty u^{-r}e^{-u}du = \Gamma(1-r)$, which is finite precisely because $r<1$.
(b) $p_\lambda \ge 0$, and with $\nu=\lambda t$,

$$\int_0^\lambda p_\lambda(\nu)\,d\nu = \frac{1}{\Gamma(r)\Gamma(1-r)}\int_0^1 t^{r-1}(1-t)^{-r}\,dt = \frac{B(r,1-r)}{\Gamma(r)\Gamma(1-r)} = 1,$$

because $B(r,1-r)=\Gamma(r)\Gamma(1-r)/\Gamma(1)$, the integral converging since both exponents $r-1$ and $-r$ exceed -1 exactly when $0<r<1$.
(c) A scale mixture of exponentials is strictly decreasing, while a gamma(r,λ) pdf with $r>1$ increases near the origin. Precisely, suppose $f(x)=g(x):=\int_0^\infty \frac{1}{\nu}e^{-x/\nu}q_\lambda(\nu)\,d\nu$ for some pdf q_λ. On the one hand,

$$\frac{\partial}{\partial x}\log f(x) = \frac{r-1}{x}-\frac{1}{\lambda} > 0 \qquad \text{for } 0<x<\lambda(r-1),$$

an interval that is nonempty exactly when $r>1$. On the other hand, differentiation under the integral sign is legitimate on $x\ge x_0>0$ (there $\nu^{-2}e^{-x/\nu}\le \big(2/(ex_0)\big)^2$, a constant, hence q_λ-integrable), so

$$g'(x) = -\int_0^\infty \frac{1}{\nu^2}e^{-x/\nu}q_\lambda(\nu)\,d\nu < 0,$$

giving $\frac{\partial}{\partial x}\log g(x)<0$ for every $x>0$. The two statements contradict each other at any $x\in\big(0,\lambda(r-1)\big)$, so no such q_λ exists when $r>1$.

Problem 4.39. Let $(X_1,\dots,X_n)$ have a multinomial distribution with m trials and cell probabilities $p_1,\dots,p_n$ (see Definition 4.6.2). Show that, for every i and j,

$$X_i \mid X_j = x_j \;\sim\; \text{binomial}\left(m-x_j,\ \frac{p_i}{1-p_j}\right),$$

$$X_j \;\sim\; \text{binomial}(m,p_j)$$

and that $\text{Cov}(X_i,X_j) = -mp_ip_j$.

Solution. Both distributional claims are the computations carried out on pages 182–183 for the last coordinate, and the labelling of cells is arbitrary. Collapsing all cells but j into one makes X_j the number of successes in m independent trials of success probability p_j, so $X_j\sim\text{binomial}(m,p_j)$ (the marginalizing sum is the one evaluated there by Theorem 4.6.4). Given $X_j=x_j$, the remaining coordinates are multinomial with $m-x_j$ trials and cell probabilities $p_\ell/(1-p_j)$, $\ell\ne j$; taking the i-th marginal of that multinomial by the same argument,

$$X_i \mid X_j = x_j \;\sim\; \text{binomial}\left(m-x_j,\ \frac{p_i}{1-p_j}\right).$$

For the covariance, condition on X_j (Theorem 4.4.3) and use $\mathrm{E}X_j = mp_j$, $\mathrm{E}X_j^2 = mp_j(1-p_j) + m^2p_j^2$:

$$\begin{aligned}\mathrm{E}(X_iX_j) &= \mathrm{E}\big(X_j\,\mathrm{E}(X_i \mid X_j)\big) = \frac{p_i}{1-p_j}\,\mathrm{E}\big(X_j(m-X_j)\big)\\ &= \frac{p_i}{1-p_j}\Big(m^2p_j - mp_j(1-p_j) - m^2p_j^2\Big)\\ &= \frac{p_i}{1-p_j}\,m(m-1)p_j(1-p_j) = m(m-1)p_ip_j.\end{aligned}$$

Hence, by Theorem 4.5.3,

$$\mathrm{Cov}(X_i, X_j) = m(m-1)p_ip_j - (mp_i)(mp_j) = -mp_ip_j.$$

Problem 4.40. A generalization of the beta distribution is the *Dirichlet* distribution. In its bivariate version, (X, Y) have pdf

$$f(x,y) = Cx^{a-1}y^{b-1}(1-x-y)^{c-1},$$
$$0<x<1, \quad 0<y<1, \quad 0<y<1-x<1,$$

where $a>0$, $b>0$, and $c>0$ are constants.
(a) Show that $C = \dfrac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)}$.
(b) Show that, marginally, both X and Y are beta.
(c) Find the conditional distribution of $Y \mid X = x$, and show that $Y/(1-x)$ is beta(b,c).
(d) Show that $\mathrm{E}(XY) = \dfrac{ab}{(a+b+c+1)(a+b+c)}$, and find their covariance.

Solution. (a) Integrating y first over $0<y<1-x$, with $y=(1-x)t$ and $s=a+b+c$ throughout:

$$\begin{aligned}\int_0^{1-x} y^{b-1}(1-x-y)^{c-1}dy &= (1-x)^{b+c-1}\int_0^1 t^{b-1}(1-t)^{c-1}dt\\ &= (1-x)^{b+c-1}B(b,c),\end{aligned}$$

so that $\iint f = C\,B(b,c)B(a,b+c)$, and

$$C^{-1} = \frac{\Gamma(b)\Gamma(c)}{\Gamma(b+c)}\cdot\frac{\Gamma(a)\Gamma(b+c)}{\Gamma(s)} = \frac{\Gamma(a)\Gamma(b)\Gamma(c)}{\Gamma(a+b+c)}.$$

(b) The inner integral just computed is the marginal:

$$f_X(x) = C\,B(b,c)\,x^{a-1}(1-x)^{b+c-1} = \frac{\Gamma(s)}{\Gamma(a)\Gamma(b+c)}x^{a-1}(1-x)^{b+c-1},$$

so $X \sim \mathrm{beta}(a, b+c)$, and by the symmetry of f in $(a,x) \leftrightarrow (b,y)$, $Y \sim \mathrm{beta}(b, a+c)$.
(c) Dividing,

$$f(y \mid x) = \frac{y^{b-1}(1-x-y)^{c-1}}{B(b,c)\,(1-x)^{b+c-1}}, \qquad 0<y<1-x.$$

Put $T = Y/(1-x)$, a linear map with Jacobian $1-x$, giving

$$f_T(t) = \frac{(1-x)^{b-1}t^{b-1}(1-x)^{c-1}(1-t)^{c-1}(1-x)}{B(b,c)(1-x)^{b+c-1}} = \frac{t^{b-1}(1-t)^{c-1}}{B(b,c)},$$

for $0<t<1$: that is, $Y/(1-x) \sim \mathrm{beta}(b,c)$, free of x.
(d) By (c), $\mathrm{E}(Y \mid X) = \dfrac{b}{b+c}(1-X)$, so by Theorem 4.4.3 and the beta$(a, b+c)$ moments $\mathrm{E}X = a/s$, $\mathrm{E}X^2 = a(a+1)/\big[s(s+1)\big]$,

$$\begin{aligned}\mathrm{E}(XY) &= \frac{b}{b+c}\,\mathrm{E}\big(X - X^2\big) = \frac{b}{b+c}\cdot\frac{a(b+c)}{s(s+1)}\\ &= \frac{ab}{(a+b+c)(a+b+c+1)},\end{aligned}$$

using $\mathrm{E}X - \mathrm{E}X^2 = a\big[(s+1)-(a+1)\big]/\big[s(s+1)\big]$. Since $\mathrm{E}X\,\mathrm{E}Y = ab/s^2$, Theorem 4.5.3 gives

$$\mathrm{Cov}(X,Y) = ab\left(\frac{1}{s(s+1)} - \frac{1}{s^2}\right) = \frac{-ab}{(a+b+c)^2(a+b+c+1)}.$$

Problem 4.41. Show that any random variable is uncorrelated with a constant.

Solution. $\mathrm{Cov}(X,c) = 0$. Taking $Y \equiv c$ gives $\mathrm{E}Y = c$, so by Definition 4.5.1,

$$\mathrm{Cov}(X,c) = \mathrm{E}\big[(X-\mathrm{E}X)(c-c)\big] = \mathrm{E}[0] = 0,$$

a vanishing covariance being exactly what uncorrelatedness asserts; the ratio ρ of Definition 4.5.2 is not itself defined here, since $\mathrm{Var}\,c = 0$ sits in its denominator.

Problem 4.42. Let X and Y be independent random variables with means μ_X, μ_Y and variances σ_X^2, σ_Y^2. Find an expression for the correlation of XY and Y in terms of these means and variances.

Solution.

$$\rho_{XY,\,Y} = \frac{\mu_X \sigma_Y}{\sqrt{\sigma_X^2\sigma_Y^2 + \sigma_X^2\mu_Y^2 + \mu_X^2\sigma_Y^2}}.$$

Independence lets every expectation factor (Theorem 4.2.10). For the covariance, by Theorem 4.5.3,

$$\begin{aligned}\mathrm{Cov}(XY,Y) &= \mathrm{E}(XY^2) - \mathrm{E}(XY)\,\mathrm{E}Y = \mu_X\,\mathrm{E}Y^2 - \mu_X\mu_Y^2 \\ &= \mu_X\big(\sigma_Y^2 + \mu_Y^2\big) - \mu_X\mu_Y^2 = \mu_X\sigma_Y^2,\end{aligned}$$

and, since X^2 and Y^2 are also independent,

$$\begin{aligned}\mathrm{Var}(XY) &= \mathrm{E}X^2\,\mathrm{E}Y^2 - \mu_X^2\mu_Y^2 \\ &= \big(\sigma_X^2 + \mu_X^2\big)\big(\sigma_Y^2 + \mu_Y^2\big) - \mu_X^2\mu_Y^2 \\ &= \sigma_X^2\sigma_Y^2 + \sigma_X^2\mu_Y^2 + \mu_X^2\sigma_Y^2.\end{aligned}$$

Dividing by $\sqrt{\mathrm{Var}(XY)\,\mathrm{Var}(Y)}$ cancels one factor of σ_Y and gives the displayed expression (Definition 4.5.2), valid whenever $0 < \sigma_X^2, \sigma_Y^2 < \infty$ – by independence these second moments are all that $\mathrm{Var}(XY)$ requires.

4.7 Exercises 4.43–4.49

Problem 4.43. Let X_1, X_2, and X_3 be uncorrelated random variables, each with mean μ and variance σ^2. Find, in terms of μ and σ^2, $\mathrm{Cov}(X_1+X_2, X_2+X_3)$ and $\mathrm{Cov}(X_1+X_2, X_1-X_2)$.

Solution. $\mathrm{Cov}(X_1+X_2, X_2+X_3) = \sigma^2$ and $\mathrm{Cov}(X_1+X_2, X_1-X_2) = 0$; neither depends on μ. Covariance is bilinear (expand $\mathrm{E}[(\cdot - \mathrm{E}\cdot)(\cdot - \mathrm{E}\cdot)]$ termwise), and $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$ by hypothesis, while $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}\,X_i = \sigma^2$. Hence

$$\begin{aligned}\mathrm{Cov}(X_1+X_2,\ X_2+X_3) &= \mathrm{Cov}(X_1,X_2) + \mathrm{Cov}(X_1,X_3) \\ &\quad + \mathrm{Cov}(X_2,X_2) + \mathrm{Cov}(X_2,X_3) \\ &= 0 + 0 + \sigma^2 + 0 = \sigma^2,\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(X_1+X_2,\ X_1-X_2) &= \mathrm{Var}\,X_1 - \mathrm{Cov}(X_1,X_2) \\ &\quad + \mathrm{Cov}(X_2,X_1) - \mathrm{Var}\,X_2 \\ &= \sigma^2 - 0 + 0 - \sigma^2 = 0.\end{aligned}$$

Problem 4.44. Prove the following generalization of Theorem 4.5.6: For any random vector $(X_1, \dots, X_n)$,

$$\operatorname{Var}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \operatorname{Var} X_i + 2 \sum_{1 \le i < j \le n} \operatorname{Cov}(X_i, X_j).$$

Solution. Write $\mu_i = \mathrm{E}X_i$ and expand the square. With $S = \sum_{i=1}^{n} X_i$ we have $\mathrm{E}S = \sum_i \mu_i$, so

$$\begin{aligned} \operatorname{Var} S &= \mathrm{E}\Big(\sum_{i=1}^{n}(X_i - \mu_i)\Big)^2 = \mathrm{E}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - \mu_i)(X_j - \mu_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j), \end{aligned}$$

the interchange of E with the finite double sum being linearity of expectation (Theorem 4.2.2), and the last step Definition 4.5.1. Splitting the double sum into its diagonal and off-diagonal parts and using the symmetry $\operatorname{Cov}(X_i, X_j) = \operatorname{Cov}(X_j, X_i)$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j) = \sum_{i=1}^{n} \operatorname{Var} X_i + 2 \sum_{1 \le i < j \le n} \operatorname{Cov}(X_i, X_j).$$

Problem 4.45. Show that if $(X, Y) \sim$ bivariate normal$(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, then the following are true.
(a) The marginal distribution of X is $\mathrm{n}(\mu_X, \sigma_X^2)$ and the marginal distribution of Y is $\mathrm{n}(\mu_Y, \sigma_Y^2)$.
(b) The conditional distribution of Y given $X = x$ is

$$\mathrm{n}\big(\mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X),\ \sigma_Y^2(1 - \rho^2)\big)\,.$$

(c) For any constants a and b, the distribution of $aX + bY$ is

$$\mathrm{n}\big(a\mu_X + b\mu_Y,\ a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y\big)\,.$$

Solution. Everything falls out of one completion of the square in the exponent of Definition 4.5.10. Put

$$u = \frac{x - \mu_X}{\sigma_X}, \qquad v = \frac{y - \mu_Y}{\sigma_Y},$$

so that the exponent of $f(x, y)$ is $-\frac{1}{2(1-\rho^2)}(u^2 - 2\rho uv + v^2)$. Since

$$u^2 - 2\rho uv + v^2 = (v - \rho u)^2 + (1 - \rho^2)u^2,$$

and since $\sigma_Y(v - \rho u) = y - \big[\mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X)\big]$, the pdf factors as

$$f(x, y) = g(x)\, h(y \mid x),$$

where, with $m(x) = \mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X)$,

$$\begin{aligned} g(x) &= \frac{1}{\sqrt{2\pi}\,\sigma_X}\, e^{-u^2/2}, \\ h(y \mid x) &= \frac{1}{\sqrt{2\pi}\,\sigma_Y\sqrt{1 - \rho^2}} \exp\!\left(-\frac{(y - m(x))^2}{2\sigma_Y^2(1 - \rho^2)}\right). \end{aligned}$$

(a) For each fixed x, $h(\cdot \mid x)$ is the $\mathrm{n}(m(x), \sigma_Y^2(1 - \rho^2))$ pdf, hence integrates to 1 in y. So $f_X(x) = \int f(x, y)\, dy = g(x)$, the $\mathrm{n}(\mu_X, \sigma_X^2)$ pdf. The exponent is symmetric under interchanging the roles of (x, μ_X, σ_X) and (y, μ_Y, σ_Y), so the same factorization with the roles reversed gives f_Y the $\mathrm{n}(\mu_Y, \sigma_Y^2)$ pdf.

(b) By Definition 4.2.3 and part (a), $f(y \mid x) = f(x,y)/f_X(x) = h(y \mid x)$, which is exactly the claimed normal pdf. (Note $1-\rho^2 > 0$ since $-1 < \rho < 1$.)
(c) Compute the mgf (for $a = b = 0$ read $\mathrm{n}(0,0)$ as the point mass at 0), using the tower property $\mathrm{E}(\mathrm{E}(\cdot \mid X)) = \mathrm{E}(\cdot)$ of Theorem 4.4.3. For $t \in \mathbb{R}$, by (b),

$$\mathrm{E}\Big(e^{tbY} \mid X\Big) = \exp\Big(tb\,m(X) + \tfrac{1}{2}t^2b^2\sigma_Y^2(1-\rho^2)\Big),$$

by the $\mathrm{n}(m(X), \sigma_Y^2(1-\rho^2))$ mgf. Writing $c = a + b\rho\sigma_Y/\sigma_X$, so that $taX + tb\,m(X) = tcX + tb\mu_Y - tb\rho(\sigma_Y/\sigma_X)\mu_X$,

$$\begin{aligned} M_{aX+bY}(t) &= \mathrm{E}\Big[e^{taX}\,\mathrm{E}(e^{tbY} \mid X)\Big] \\ &= e^{tb\mu_Y - tb\rho(\sigma_Y/\sigma_X)\mu_X + \frac{1}{2}t^2b^2\sigma_Y^2(1-\rho^2)}\,\mathrm{E}\,e^{tcX} \\ &= e^{tb\mu_Y - tb\rho(\sigma_Y/\sigma_X)\mu_X + \frac{1}{2}t^2b^2\sigma_Y^2(1-\rho^2)} e^{tc\mu_X + \frac{1}{2}t^2c^2\sigma_X^2}, \end{aligned}$$

the last step by (a) and the normal mgf. The coefficient of t is

$$b\mu_Y - b\rho\frac{\sigma_Y}{\sigma_X}\mu_X + \Big(a + b\rho\frac{\sigma_Y}{\sigma_X}\Big)\mu_X = a\mu_X + b\mu_Y,$$

and the coefficient of $t^2/2$ is

$$\begin{aligned} b^2\sigma_Y^2(1-\rho^2) + c^2\sigma_X^2 &= b^2\sigma_Y^2 - b^2\rho^2\sigma_Y^2 \\ &\quad + a^2\sigma_X^2 + 2ab\rho\sigma_X\sigma_Y + b^2\rho^2\sigma_Y^2 \\ &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y. \end{aligned}$$

Thus M_{aX+bY} is the mgf of the asserted normal distribution, and since mgfs exist in a neighbourhood of 0 here, Theorem 2.3.11 identifies the distribution.

Problem 4.46. (A derivation of the bivariate normal distribution) Let Z_1 and Z_2 be independent $\mathrm{n}(0,1)$ random variables, and define new random variables X and Y by

$$X = a_X Z_1 + b_X Z_2 + c_X \quad \text{and} \quad Y = a_Y Z_1 + b_Y Z_2 + c_Y,$$

where a_X, b_X, c_X, a_Y, b_Y, and c_Y are constants.
(a) Show that

$$\begin{aligned} \mathrm{E}X = c_X, \quad \mathrm{Var}\,X = a_X^2 + b_X^2, \\ \mathrm{E}Y = c_Y, \quad \mathrm{Var}\,Y = a_Y^2 + b_Y^2, \\ \mathrm{Cov}(X,Y) = a_Xa_Y + b_Xb_Y. \end{aligned}$$

(b) If we define the constants a_X, b_X, c_X, a_Y, b_Y, and c_Y by

$$\begin{aligned} a_X &= \sqrt{\tfrac{1+\rho}{2}}\,\sigma_X, \quad b_X = \sqrt{\tfrac{1-\rho}{2}}\,\sigma_X, \quad c_X = \mu_X, \\ a_Y &= \sqrt{\tfrac{1+\rho}{2}}\,\sigma_Y, \quad b_Y = -\sqrt{\tfrac{1-\rho}{2}}\,\sigma_Y, \quad c_Y = \mu_Y, \end{aligned}$$

where $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and ρ are constants, $-1 \le \rho \le 1$, then show that

$$\begin{aligned} \mathrm{E}X = \mu_X, \quad \mathrm{Var}\,X = \sigma_X^2, \\ \mathrm{E}Y = \mu_Y, \quad \mathrm{Var}\,Y = \sigma_Y^2, \end{aligned} \qquad \rho_{XY} = \rho.$$

(c) Show that (X,Y) has the bivariate normal pdf with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ and ρ.
(d) If we start with bivariate normal parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and ρ, we can define constants a_X, b_X, c_X, a_Y, b_Y, and c_Y as the solutions to the equations

$$\begin{aligned} \mu_X = c_X, \quad \sigma_X^2 = a_X^2 + b_X^2, \\ \mu_Y = c_Y, \quad \sigma_Y^2 = a_Y^2 + b_Y^2, \end{aligned}$$

$$\rho\sigma_X\sigma_Y = a_X a_Y + b_X b_Y.$$

Show that the solution given in part (b) is not unique by exhibiting another solution to these equations. How many solutions are there?

Solution. (a) Immediate from linearity and $\mathrm{E}Z_i = 0$, $\operatorname{Var} Z_i = 1$, $\operatorname{Cov}(Z_1, Z_2) = 0$ (Theorem 4.5.5, since $Z_1 \perp Z_2$):

$$\begin{aligned}\mathrm{E}X &= a_X \mathrm{E}Z_1 + b_X \mathrm{E}Z_2 + c_X = c_X,\\ \operatorname{Var} X &= a_X^2 \operatorname{Var} Z_1 + b_X^2 \operatorname{Var} Z_2 = a_X^2 + b_X^2,\end{aligned}$$

by Theorem 4.5.6 (independent case), and likewise for Y. Bilinearity of covariance, plus the vanishing of the cross terms, gives

$$\begin{aligned}\operatorname{Cov}(X,Y) &= a_X a_Y \operatorname{Var} Z_1 + b_X b_Y \operatorname{Var} Z_2\\ &\quad + (a_X b_Y + b_X a_Y)\operatorname{Cov}(Z_1, Z_2)\\ &= a_X a_Y + b_X b_Y.\end{aligned}$$

(b) Substitute into (a). Writing $p = \sqrt{(1+\rho)/2}$ and $q = \sqrt{(1-\rho)/2}$, so $p^2 + q^2 = 1$ and $p^2 - q^2 = \rho$,

$$\begin{aligned}\operatorname{Var} X = \sigma_X^2(p^2+q^2) = \sigma_X^2, \qquad \operatorname{Var} Y = \sigma_Y^2(p^2+q^2) = \sigma_Y^2,\\ \operatorname{Cov}(X,Y) = \sigma_X\sigma_Y\, p^2 + (\sigma_X q)(-\sigma_Y q) = \sigma_X\sigma_Y(p^2 - q^2) = \rho\sigma_X\sigma_Y,\end{aligned}$$

and $\mathrm{E}X = c_X = \mu_X$, $\mathrm{E}Y = c_Y = \mu_Y$. Hence $\rho_{XY} = \rho\sigma_X\sigma_Y/(\sigma_X\sigma_Y) = \rho$.
(c) Take $|\rho| < 1$ (Definition 4.5.10 requires it; at $\rho = \pm 1$ the map below is singular and (X, Y) is degenerate). With p, q as above the transformation

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \sigma_X p & \sigma_X q \\ \sigma_Y p & -\sigma_Y q \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} + \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

is one-to-one onto $\mathbb{R}^2$ with determinant $-2pq\,\sigma_X\sigma_Y$, and $2pq = \sqrt{(1+\rho)(1-\rho)} = \sqrt{1-\rho^2}$. Inverting: with $u = (x-\mu_X)/\sigma_X$ and $v = (y-\mu_Y)/\sigma_Y$ we get $u = pz_1 + qz_2$, $v = pz_1 - qz_2$, hence $z_1 = (u+v)/(2p)$, $z_2 = (u-v)/(2q)$ and

$$\begin{aligned}z_1^2 + z_2^2 &= \frac{(u+v)^2}{2(1+\rho)} + \frac{(u-v)^2}{2(1-\rho)}\\ &= \frac{(1-\rho)(u+v)^2 + (1+\rho)(u-v)^2}{2(1-\rho^2)}\\ &= \frac{2u^2 + 2v^2 - 4\rho uv}{2(1-\rho^2)} = \frac{u^2 - 2\rho uv + v^2}{1-\rho^2}.\end{aligned}$$

Since Z_1 and Z_2 are independent, $f_{Z_1,Z_2}(z_1, z_2) = (2\pi)^{-1} e^{-(z_1^2+z_2^2)/2}$ (Definition 4.2.5), and the bivariate change-of-variables formula (4.3.2), whose Jacobian here is the constant $J = -(2pq\,\sigma_X\sigma_Y)^{-1}$, gives

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\!\left(-\frac{u^2 - 2\rho uv + v^2}{2(1-\rho^2)}\right),$$

which is exactly Definition 4.5.10 with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$.
(d) Another solution: $c_X = \mu_X$, $c_Y = \mu_Y$ and

$$a_X = \sigma_X, \quad b_X = 0, \quad a_Y = \rho\sigma_Y, \quad b_Y = \sigma_Y\sqrt{1-\rho^2},$$

for which $a_X^2 + b_X^2 = \sigma_X^2$, $a_Y^2 + b_Y^2 = \sigma_Y^2(\rho^2 + 1 - \rho^2) = \sigma_Y^2$, and $a_X a_Y + b_X b_Y = \rho\sigma_X\sigma_Y$. (Check!)
There are infinitely many. The last three equations say precisely that the plane vectors $\mathbf{a} = (a_X, b_X)$ and $\mathbf{b} = (a_Y, b_Y)$ have lengths σ_X, σ_Y and inner product $\rho\sigma_X\sigma_Y$, i.e. $\|\mathbf{a}\| = \sigma_X$, $\|\mathbf{b}\| = \sigma_Y$ and the angle between them has cosine ρ. Writing $\mathbf{a} = \sigma_X(\cos\alpha, \sin\alpha)$ and $\mathbf{b} = \sigma_Y(\cos\beta, \sin\beta)$, the constraints reduce to $\cos(\alpha - \beta) = \rho$, so $\alpha \in [0, 2\pi)$ is free and $\beta = \alpha \mp \arccos\rho$: a one-parameter family, doubled by the sign (the two branches coinciding only when $\rho = \pm 1$).

Problem 4.47. (Marginal normality does not imply bivariate normality.) Let X and Y be independent $\mathrm{n}(0,1)$ random variables, and define a new random variable Z by

$$Z = \begin{cases} X & \text{if } XY > 0, \\ -X & \text{if } XY < 0. \end{cases}$$

(a) Show that Z has a normal distribution.
(b) Show that the joint distribution of Z and Y is not bivariate normal. (Hint: Show that Z and Y always have the same sign.)

Solution. $Z = \operatorname{sgn}(Y)\,|X|$, which settles both parts.
Indeed $P(XY = 0) = 0$, so with probability one: if $Y > 0$ then $XY > 0$ forces $X > 0$ and $Z = X = |X|$, while $XY < 0$ forces $X < 0$ and $Z = -X = |X|$; if $Y < 0$ the two cases give $Z = X = -|X|$ and $Z = -X = -|X|$ respectively.
(a) $Z \sim \mathrm{n}(0,1)$. Since X and Y are independent, so are $|X|$ and $\operatorname{sgn}(Y)$, and $P(Y > 0) = P(Y < 0) = \frac{1}{2}$. For $z > 0$,

$$\begin{aligned} P(Z \le z) &= P(Y < 0) + P(Y > 0)\,P(|X| \le z) \\ &= \tfrac{1}{2} + \tfrac{1}{2}\big(2\Phi(z) - 1\big) = \Phi(z), \end{aligned}$$

and for $z < 0$,

$$P(Z \le z) = P(Y < 0)\,P(|X| \ge -z) = \tfrac{1}{2} \cdot 2\big(1 - \Phi(-z)\big) = \Phi(z),$$

using the symmetry $P(|X| \le t) = 2\Phi(t) - 1$ for $t \ge 0$. So Z has the standard normal cdf.
(b) $ZY = \operatorname{sgn}(Y)|X|\,Y = |X|\,|Y| \ge 0$ with probability one, so

$$P(Z > 0,\ Y < 0) = 0.$$

But every bivariate normal pdf of Definition 4.5.10 is strictly positive on all of $\mathbb{R}^2$ (its exponent is finite everywhere, and $-1 < \rho < 1$ by definition), so it assigns strictly positive probability to the open quadrant $\{z > 0,\ y < 0\}$. Hence (Z, Y) cannot be bivariate normal, even though both marginals are $\mathrm{n}(0,1)$.

Problem 4.48. Gelman and Meng (1991) give an example of a bivariate family of distributions that are not bivariate normal but have normal conditionals. Define the joint pdf of (X, Y) as

$$f(x,y) \propto \exp\left\{-\tfrac{1}{2}\left[Ax^2y^2 + x^2 + y^2 - 2Bxy - 2Cx - 2Dy\right]\right\},$$

where A, B, C, D are constants.
(a) Show that the distribution of $X \mid Y = y$ is normal with mean $\dfrac{By + C}{Ay^2 + 1}$ and variance $\dfrac{1}{Ay^2 + 1}$. Derive a corresponding result for the distribution of $Y \mid X = x$.
(b) A most interesting configuration is $A = 1$, $B = 0$, $C = D = 8$. Show that this joint distribution is bimodal.

Solution. (a) Fix y and read the exponent as a quadratic in x. Collecting terms,

$$\begin{aligned} Ax^2y^2 + x^2 - 2Bxy - 2Cx &= (Ay^2 + 1)x^2 - 2(By + C)x \\ &= (Ay^2 + 1)\left(x - \frac{By + C}{Ay^2 + 1}\right)^2 + k(y), \end{aligned}$$

where $k(y)$ does not involve x. Hence, for fixed y,

$$f(x \mid y) = \frac{f(x,y)}{f_Y(y)} \propto \exp\left\{-\frac{1}{2}\,\frac{\left(x - \frac{By+C}{Ay^2+1}\right)^2}{1/(Ay^2+1)}\right\}$$

as a function of x, which (the proportionality constant being whatever makes it integrate to 1) is the

$$\mathrm{n}\left(\frac{By+C}{Ay^2+1},\ \frac{1}{Ay^2+1}\right)$$

pdf, provided $Ay^2+1>0$. The exponent is symmetric under $(x,C)\leftrightarrow(y,D)$, so the same computation gives

$$Y\mid X=x\ \sim\ \mathrm{n}\left(\frac{Bx+D}{Ax^2+1},\ \frac{1}{Ax^2+1}\right).$$

(b) With $A=1,B=0,C=D=8$ the modes of f are the minimizers of

$$h(x,y)=x^2y^2+x^2+y^2-16x-16y,$$

and there are exactly two of them: $(4-\sqrt{15},\,4+\sqrt{15})$ and $(4+\sqrt{15},\,4-\sqrt{15})$.
Setting the partials to zero,

$$\frac{\partial h}{\partial x}=2x(y^2+1)-16=0,\qquad \frac{\partial h}{\partial y}=2y(x^2+1)-16=0,$$

i.e. $x(y^2+1)=8$ and $y(x^2+1)=8$. Subtracting,

$$xy(y-x)-(y-x)=0\implies(y-x)(xy-1)=0.$$

(i) $xy=1$. Then $8=x(y^2+1)=xy\cdot y+x=x+y$, so $x+y=8$, $xy=1$, and x,y are the roots $4\pm\sqrt{15}$ of $t^2-8t+1=0$. At such a point

$$h=(xy)^2+(x+y)^2-2xy-16(x+y)=1+64-2-128=-65.$$

The Hessian there has $h_{xx}=2y^2+2>0$ and

$$\det=4(x^2+1)(y^2+1)-16(xy)^2=4\cdot64-16=240>0,$$

using $x^2y^2+x^2+y^2+1=1+62+1=64$; so both points are strict local minima of h.
(ii) $y=x$. Then $x(x^2+1)=8$, i.e. $x^3+x-8=0$, which has the single real root $x_0\approx1.8338$ (the cubic is strictly increasing). There $h=x_0^4+2x_0^2-32x_0\approx-40.65$ and

$$\det=4(x_0^2+1)^2-16x_0^4\approx-104.8<0,$$

so this critical point is a saddle, not a mode.
Finally $h(x,y)\geq x^2+y^2-16x-16y=(x-8)^2+(y-8)^2-128\to\infty$ as $\|(x,y)\|\to\infty$, so h attains a global minimum, which must be one of the three critical points; it is -65, attained at the two points in (i). Thus $f\propto e^{-h/2}$ has exactly two modes, and the distribution is bimodal.

Problem 4.49. Behboodian (1990) illustrates how to construct bivariate random variables that are uncorrelated but dependent. Suppose that f_1,f_2,g_1,g_2 are univariate densities with means μ_1,μ_2,ξ_1,ξ_2, respectively, and the bivariate random variable (X,Y) has density

$$(X,Y)\sim af_1(x)g_1(y)+(1-a)f_2(x)g_2(y),$$

where $0<a<1$ is known.
(a) Show that the marginal distributions are given by $f_X(x)=af_1(x)+(1-a)f_2(x)$ and $f_Y(y)=ag_1(y)+(1-a)g_2(y)$.
(b) Show that X and Y are independent if and only if $[f_1(x)-f_2(x)][g_1(y)-g_2(y)]=0$.
(c) Show that $\mathrm{Cov}(X,Y)=a(1-a)[\mu_1-\mu_2][\xi_1-\xi_2]$, and thus explain how to construct dependent uncorrelated random variables.
(d) Letting f_1,f_2,g_1,g_2 be binomial pmfs, give examples of combinations of parameters that lead to independent (X,Y) pairs, correlated (X,Y) pairs, and uncorrelated but dependent (X,Y) pairs.

Solution. (a) Integrate out, using that g_1, g_2 are densities:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = af_1(x)\int g_1 + (1-a)f_2(x)\int g_2,$$

which is $af_1(x) + (1-a)f_2(x)$; symmetrically $f_Y(y) = ag_1(y) + (1-a)g_2(y)$.
(b) One algebraic identity does it. Using (a),

$$\begin{aligned} f(x,y) - f_X(x)f_Y(y) &= af_1g_1 + (1-a)f_2g_2 \\ &\quad - \big[af_1 + (1-a)f_2\big]\big[ag_1 + (1-a)g_2\big] \\ &= a(1-a)f_1g_1 + a(1-a)f_2g_2 \\ &\quad - a(1-a)f_1g_2 - a(1-a)f_2g_1 \\ &= a(1-a)\,[f_1(x) - f_2(x)]\,[g_1(y) - g_2(y)], \end{aligned}$$

where all functions are evaluated at x or y as appropriate. By Definition 4.2.5, X and Y are independent if and only if $f(x,y) = f_X(x)f_Y(y)$ for all (x,y); since $a(1-a) \neq 0$, this holds exactly when $[f_1(x) - f_2(x)][g_1(y) - g_2(y)] = 0$ for all (x,y).
(c) From (a), $\mathrm{E}X = a\mu_1 + (1-a)\mu_2$ and $\mathrm{E}Y = a\xi_1 + (1-a)\xi_2$, while

$$\mathrm{E}XY = \iint xy\,f(x,y)\,dx\,dy = a\mu_1\xi_1 + (1-a)\mu_2\xi_2.$$

Hence, by the identity of (b) with (f_i, g_i) replaced by (μ_i, ξ_i),

$$\begin{aligned} \mathrm{Cov}(X,Y) &= a\mu_1\xi_1 + (1-a)\mu_2\xi_2 \\ &\quad - \big[a\mu_1 + (1-a)\mu_2\big]\big[a\xi_1 + (1-a)\xi_2\big] \\ &= a(1-a)(\mu_1 - \mu_2)(\xi_1 - \xi_2). \end{aligned}$$

So: choose $f_1 \neq f_2$ with $\mu_1 = \mu_2$ (equal means, different densities) and any $g_1 \neq g_2$. Then the covariance vanishes while $[f_1 - f_2][g_1 - g_2] \not\equiv 0$, so by (b) X and Y are dependent.
(d) Take $a = \frac{1}{2}$ throughout, $g_1 = \mathrm{binomial}(10, 0.2)$ (mean $\xi_1 = 2$) and $g_2 = \mathrm{binomial}(10, 0.8)$ (mean $\xi_2 = 8$).
(i) Independent: $f_1 = f_2 = \mathrm{binomial}(10, 0.5)$. Then $f_1 - f_2 \equiv 0$, so (b) gives independence (indeed $f(x,y) = f_1(x)f_Y(y)$).
(ii) Correlated: $f_1 = \mathrm{binomial}(10, 0.2)$, $f_2 = \mathrm{binomial}(10, 0.8)$, so $\mu_1 = 2$, $\mu_2 = 8$ and

$$\mathrm{Cov}(X,Y) = \tfrac{1}{2} \cdot \tfrac{1}{2}(2-8)(2-8) = 9 \neq 0.$$

(iii) Uncorrelated but dependent: $f_1 = \mathrm{binomial}(4, 1/2)$ and $f_2 = \mathrm{binomial}(8, 1/4)$, which have the same mean $\mu_1 = \mu_2 = 2$ but are different pmfs (e.g. $f_1(0) = 1/16$, $f_2(0) = (3/4)^8 = 6561/65536$). Then

$$\mathrm{Cov}(X,Y) = \tfrac{1}{4}(2-2)(2-8) = 0,$$

while $[f_1(x) - f_2(x)][g_1(y) - g_2(y)] \neq 0$ at, say, $(x,y) = (0,0)$, so X and Y are dependent by (b).

4.8 Exercises 4.50–4.56

Problem 4.50. If (X,Y) has the bivariate normal pdf

$$f(x,y) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left(\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right),$$

show that $\mathrm{Corr}(X,Y) = \rho$ and $\mathrm{Corr}(X^2, Y^2) = \rho^2$. (Conditional expectations will simplify calculations.)

Solution. Here $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$ in Definition 4.5.10, so by its property (a) and the conditional distribution derived in Section 4.5,

$$X \sim \mathrm{n}(0,1), \qquad Y \mid X = x \sim \mathrm{n}\big(\rho x,\ 1-\rho^2\big),$$

and symmetrically $Y \sim \mathrm{n}(0,1)$. Hence $\operatorname{Var} X = \operatorname{Var} Y = 1$, and by the tower property (Theorem 4.4.3),

$$\mathrm{E}\, XY = \mathrm{E}\big[X\, \mathrm{E}(Y \mid X)\big] = \mathrm{E}\big[\rho X^2\big] = \rho.$$

Since $\mathrm{E}\, X = \mathrm{E}\, Y = 0$, $\operatorname{Cov}(X,Y) = \rho$ and $\operatorname{Corr}(X,Y) = \rho/(1\cdot 1) = \rho$.
For the squares, variance plus mean squared gives $\mathrm{E}(Y^2 \mid X = x) = (1-\rho^2) + \rho^2 x^2$, and the standard normal moments are $\mathrm{E}\, X^2 = 1$, $\mathrm{E}\, X^4 = 3$:

$$\begin{aligned} \mathrm{E}\, X^2Y^2 &= \mathrm{E}\big[X^2\, \mathrm{E}(Y^2 \mid X)\big] \\ &= \mathrm{E}\big[(1-\rho^2)X^2 + \rho^2 X^4\big] \\ &= (1-\rho^2) + 3\rho^2 = 1 + 2\rho^2. \end{aligned}$$

Therefore $\operatorname{Cov}(X^2, Y^2) = (1+2\rho^2) - (1)(1) = 2\rho^2$, while

$$\operatorname{Var} X^2 = \mathrm{E}\, X^4 - (\mathrm{E}\, X^2)^2 = 3 - 1 = 2 = \operatorname{Var} Y^2,$$

so

$$\operatorname{Corr}(X^2, Y^2) = \frac{2\rho^2}{\sqrt{2}\sqrt{2}} = \rho^2.$$

Problem 4.51. Let X, Y, and Z be independent uniform$(0,1)$ random variables.
(a) Find $P(X/Y \le t)$ and $P(XY \le t)$. (Pictures will help.)
(b) Find $P(XY/Z \le t)$.

Solution. (a) For $t > 0$,

$$P(X/Y \le t) = \begin{cases} t/2, & 0 < t \le 1, \\ 1 - \dfrac{1}{2t}, & t > 1. \end{cases}$$

Indeed $\{X/Y \le t\} = \{X \le tY\}$ is the part of the unit square above the line $x = ty$: for $t \le 1$ that line exits at $(t,1)$, cutting off the triangle $(0,0),(t,1),(0,1)$ of area $t/2$; for $t > 1$ it exits at $(1,1/t)$ and the complement is the triangle $(0,0),(1,0),(1,1/t)$ of area $1/(2t)$.
For XY the answer is 1 once $t \ge 1$, and for $0 < t < 1$ the complementary region $\{xy > t\}$ lies between the hyperbola $y = t/x$ and $y = 1$ for $x \in (t,1)$:

$$P(XY > t) = \int_t^1 \Big(1 - \frac{t}{x}\Big)\,dx = 1 - t + t\log t,$$

so $P(XY \le t) = t - t\log t$ for $0 < t < 1$ (and $= 1$ for $t \ge 1$).
(b) Condition on $Z = z$ and use part (a): $P(XY \le tz) = tz - tz\log(tz)$ when $tz \le 1$, and $= 1$ otherwise. The two integrals needed are $\int_0^1 z\,dz = \frac{1}{2}$ and $\int_0^1 z\log z\,dz = -\frac{1}{4}$.
(i) $0 < t \le 1$. Then $tz \le 1$ for every $z \in (0,1)$, so

$$\begin{aligned} P(XY/Z \le t) &= \int_0^1 \big[tz - tz\log t - tz\log z\big]\,dz \\ &= \frac{t}{2} - \frac{t\log t}{2} + \frac{t}{4} = \frac{3t}{4} - \frac{t}{2}\log t. \end{aligned}$$

(ii) $t > 1$. Split at $z = 1/t$; on $z > 1/t$ the conditional probability is 1. Substituting $u = tz$ in the first

piece,

$$P(XY/Z \le t) = \frac{1}{t}\int_0^1 (u - u\log u)\,du + \Big(1 - \frac{1}{t}\Big)$$
$$= \frac{1}{t}\Big(\frac{1}{2} + \frac{1}{4}\Big) + 1 - \frac{1}{t} = 1 - \frac{1}{4t}.$$

Problem 4.52. Bullets are fired at the origin of an (x, y) coordinate system, and the point hit, say (X, Y), is a random variable. The variables X and Y are taken to be independent $\mathrm{n}(0, 1)$ random variables. If two bullets are fired independently, what is the distribution of the distance between them?

Solution. The distance D has the Rayleigh density

$$f_D(d) = \frac{d}{2}\,e^{-d^2/4}, \qquad d > 0,$$

equivalently $D^2 \sim$ exponential(4) (mean 4 in the C&B scale parametrization).
Indeed, with the two impacts (X_1, Y_1) and (X_2, Y_2) built from four independent $\mathrm{n}(0, 1)$ variables, Corollary 4.6.10 (linear combinations of independent normals are normal) gives

$$X_1 - X_2 \sim \mathrm{n}(0, 2), \qquad Y_1 - Y_2 \sim \mathrm{n}(0, 2),$$

and these are independent, being functions of disjoint sets of independent variables (Theorem 4.6.12). Writing $Z_1 = (X_1 - X_2)/\sqrt{2}$ and $Z_2 = (Y_1 - Y_2)/\sqrt{2}$, which are independent $\mathrm{n}(0, 1)$,

$$D^2 = (X_1 - X_2)^2 + (Y_1 - Y_2)^2$$
$$= 2\big(Z_1^2 + Z_2^2\big) \sim 2\chi^2_2,$$

by Lemma 5.3.2, parts (a) and (b). Since $\chi^2_2 = \mathrm{gamma}(1, 2) = \mathrm{exponential}(2)$, multiplying by the scale factor 2 gives $D^2 \sim$ exponential(4). Hence for $d > 0$

$$P(D > d) = P(D^2 > d^2) = e^{-d^2/4},$$

and differentiating $F_D(d) = 1 - e^{-d^2/4}$ yields the stated pdf.

Problem 4.53. Let A, B, and C be independent random variables, uniformly distributed on $(0, 1)$. What is the probability that $Ax^2 + Bx + C$ has real roots? (Hint: If $X \sim$ uniform$(0, 1)$, then $-\log X \sim$ exponential. The sum of two independent exponentials is gamma.)

Solution.

$$P(B^2 \ge 4AC) = \frac{5}{36} + \frac{\log 2}{6} \approx 0.2544.$$

The roots are real exactly when the discriminant $B^2 - 4AC$ is nonnegative. If $X \sim$ uniform$(0, 1)$ then $P(-\log X > s) = P(X < e^{-s}) = e^{-s}$, so $-\log A$ and $-\log C$ are independent exponential(1); their sum is gamma (Example 4.6.8),

$$T = -\log(AC) \sim \mathrm{gamma}(2, 1), \qquad P(T \ge s) = e^{-s}(1 + s),\ s > 0.$$

Condition on $B = b$, $0 < b < 1$, and note $AC \le b^2/4 \iff T \ge \log(4/b^2) =: s_b > 0$. Since $e^{-s_b} = b^2/4$,

$$P\big(4AC \le b^2\big) = \frac{b^2}{4}\Big(1 + 2\log 2 - 2\log b\Big).$$

Integrating over b, with $\int_0^1 b^2 db = \frac{1}{3}$ and $\int_0^1 b^2 \log b \, db = -\frac{1}{9}$,

$$\begin{aligned} P(B^2 \geq 4AC) &= \frac{1}{4}\int_0^1 \big[(1+2\log 2)b^2 - 2b^2 \log b\big] db \\ &= \frac{1}{4}\Big[\frac{1+2\log 2}{3} + \frac{2}{9}\Big] \\ &= \frac{1}{4}\Big[\frac{5}{9} + \frac{2\log 2}{3}\Big] = \frac{5}{36} + \frac{\log 2}{6}. \end{aligned}$$

Problem 4.54. Find the pdf of $\prod_{i=1}^n X_i$, where the X_is are independent uniform$(0,1)$ random variables. (Hint: Try to calculate the cdf, and remember the relationship between uniforms and exponentials.)

Solution.

$$f_U(u) = \frac{(-\log u)^{n-1}}{(n-1)!}, \qquad 0 < u < 1,$$

where $U = \prod_{i=1}^n X_i$.
Each $-\log X_i$ is exponential(1), since $P(-\log X_i > s) = P(X_i < e^{-s}) = e^{-s}$, and these are independent (Theorem 4.6.12), so by Example 4.6.8

$$T = -\log U = \sum_{i=1}^n (-\log X_i) \sim \text{gamma}(n,1), \qquad f_T(t) = \frac{t^{n-1}e^{-t}}{\Gamma(n)}.$$

The map $u \mapsto t = -\log u$ is a decreasing one-to-one transformation of $(0,1)$ onto $(0,\infty)$ with $|dt/du| = 1/u$, so Theorem 2.1.5 gives, for $0 < u < 1$,

$$f_U(u) = f_T(-\log u) \cdot \frac{1}{u} = \frac{(-\log u)^{n-1} e^{\log u}}{\Gamma(n)} \cdot \frac{1}{u} = \frac{(-\log u)^{n-1}}{(n-1)!}.$$

Problem 4.55. A *parallel system* is one that functions as long as at least one component of it functions. A particular parallel system is composed of three independent components, each of which has a lifelength with an exponential(λ) distribution. The lifetime of the system is the maximum of the individual lifelengths. What is the distribution of the lifetime of the system?

Solution.

$$f_T(t) = \frac{3}{\lambda}\big(1 - e^{-t/\lambda}\big)^2 e^{-t/\lambda}, \qquad t > 0,$$

where $T = \max\{X_1, X_2, X_3\}$ and exponential(λ) carries C&B's scale parametrization, $f_X(x) = \lambda^{-1}e^{-x/\lambda}$, $F_X(x) = 1 - e^{-x/\lambda}$.
The system functions past t unless all three components have failed, so independence gives directly

$$F_T(t) = P(X_1 \leq t,\, X_2 \leq t,\, X_3 \leq t) = \big[F_X(t)\big]^3 = \big(1 - e^{-t/\lambda}\big)^3,$$

the cdf of the sample maximum; differentiating yields the pdf above.

Problem 4.56. A large number, $N = mk$, of people are subject to a blood test. This can be administered in two ways.
(i) Each person can be tested separately. In this case N tests are required.
(ii) The blood samples of k people can be pooled and analyzed together. If the test is negative, this one test suffices for k people. If the test is positive, each of the k persons must be tested separately, and, in all, $k+1$ tests are required for the k people.
Assume that the probability, p, that the test is positive is the same for all people and that the test

results for different people are statistically independent.
(a) What is the probability that the test for a pooled sample of k people will be positive?
(b) Let X = number of blood tests necessary under plan (ii). Find $\mathrm{E}\,X$.
(c) In terms of minimizing the expected number of blood tests to be performed on the N people, which plan [(i) or (ii)] would be preferred if it is known that p is close to 0? Justify your answer using the expression derived in part (b).

Solution. (a) The pooled test is negative exactly when all k individuals are negative, so by independence

$$P(\text{pool positive}) = 1 - (1-p)^k.$$

(b) Write $X = \sum_{j=1}^{m} X_j$, where X_j is the number of tests spent on group j. Each X_j takes the value 1 with probability $(1-p)^k$ and $k+1$ with probability $1-(1-p)^k$, so

$$\begin{aligned}\mathrm{E}\,X_j &= (1-p)^k + (k+1)\big[1-(1-p)^k\big] \\ &= 1 + k\big[1-(1-p)^k\big].\end{aligned}$$

By linearity of expectation (Theorem 2.2.5, extended to random vectors in Section 4.2), and since $m = N/k$,

$$\mathrm{E}\,X = m\Big(1 + k\big[1-(1-p)^k\big]\Big) = N\Big(\frac{1}{k} + 1 - (1-p)^k\Big).$$

(c) Plan (ii). Plan (i) costs N tests, so (ii) is preferred precisely when

$$\frac{1}{k} + 1 - (1-p)^k < 1 \quad\Longleftrightarrow\quad (1-p)^k > \frac{1}{k} \quad\Longleftrightarrow\quad p < 1 - k^{-1/k},$$

and for any fixed $k \geq 2$ the right-hand threshold is strictly positive, so the inequality holds for all p near 0. Quantitatively, $(1-p)^k = 1 - kp + O(p^2)$ gives

$$\mathrm{E}\,X = N\Big(\frac{1}{k} + kp\Big) + O(Np^2),$$

which tends to N/k as $p \to 0$: pooling cuts the expected workload by roughly the factor k.

4.9 Exercises 4.57–4.63

Problem 4.57. Refer to Miscellanea 4.9.2, where for $x_i \geq 0$ the *power means* are defined by

$$A_r = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^r\right]^{1/r}.$$

(a) Show that A_1 is the arithmetic mean, A_{-1} is the harmonic mean, and $A_0 = \lim_{r\to 0} A_r$ is the geometric mean.
(b) The arithmetic-geometric-harmonic mean inequality will follow if it can be established that A_r is a nondecreasing function of r over the range $-\infty < r < \infty$.
(i) Verify that if $\log A_r$ is nondecreasing in r, then it will follow that A_r is nondecreasing in r.
(ii) Show that

$$\frac{d}{dr}\log A_r = \frac{1}{r^2}\left\{\frac{r\sum_i x_i^r \log x_i}{\sum_i x_i^r} - \log\left(\frac{1}{n}\sum_i x_i^r\right)\right\}.$$

(iii) Define $a_i = x_i^r / \sum_j x_j^r$ and write the quantity in braces as

$$\log(n) - \sum_i a_i \log(1/a_i),$$

where $\sum_i a_i = 1$. Now prove that this quantity is nonnegative, establishing the monotonicity of A_r and the arithmetic-geometric-harmonic mean inequality as a special case.
The quantity $\sum_i a_i \log(1/a_i)$ is called *entropy*, sometimes considered an absolute measure of uncertainty (see Bernardo and Smith 1994, Section 2.7). The result of part (iii) states that the maximum entropy is attained when all probabilities are the same (randomness).
(*Hint:* To prove the inequality note that the a_i are a probability distribution, and we can write

$$\mathrm{E} \log\left(\frac{1}{a}\right) = \sum_i a_i \log\left(\frac{1}{a_i}\right),$$

and Jensen's Inequality shows that $\mathrm{E} \log\left(\frac{1}{a}\right) \le \log\left(\mathrm{E}\frac{1}{a}\right)$.)

Solution. (a) Directly from the definition, $A_1 = \frac{1}{n}\sum_i x_i$ and

$$A_{-1} = \left[\frac{1}{n}\sum_i x_i^{-1}\right]^{-1} = \frac{n}{\sum_i (1/x_i)},$$

the arithmetic and harmonic means. For A_0, apply l'Hopital's rule to $\log A_r = r^{-1} \log\left(\frac{1}{n}\sum_i x_i^r\right)$, whose numerator vanishes at $r = 0$:

$$\begin{aligned}\lim_{r\to 0} \log A_r &= \lim_{r\to 0} \frac{\sum_i x_i^r \log x_i}{\sum_i x_i^r} \\ &= \frac{1}{n}\sum_i \log x_i,\end{aligned}$$

so $A_0 = \left(\prod_i x_i\right)^{1/n}$, the geometric mean.
(b)(i) $A_r = \exp(\log A_r)$ and exp is increasing, so A_r inherits the monotonicity of $\log A_r$.
(ii) For $r \ne 0$, differentiating $\log A_r = r^{-1} \log\left(\frac{1}{n}\sum_i x_i^r\right)$ and using $\frac{d}{dr} x_i^r = x_i^r \log x_i$,

$$\begin{aligned}\frac{d}{dr} \log A_r &= -\frac{1}{r^2} \log\left(\frac{1}{n}\sum_i x_i^r\right) + \frac{1}{r} \cdot \frac{\sum_i x_i^r \log x_i}{\sum_i x_i^r} \\ &= \frac{1}{r^2}\left\{\frac{r\sum_i x_i^r \log x_i}{\sum_i x_i^r} - \log\left(\frac{1}{n}\sum_i x_i^r\right)\right\}.\end{aligned}$$

(iii) With $a_i = x_i^r / \sum_j x_j^r$ we have $\log a_i = r \log x_i - \log \sum_j x_j^r$, hence $r\sum_i a_i \log x_i = \sum_i a_i \log a_i + \log \sum_j x_j^r$, and the braces become

$$\begin{aligned}&\sum_i a_i \log a_i + \log \sum_j x_j^r - \log \sum_i x_i^r + \log n \\ &\quad = \log(n) - \sum_i a_i \log(1/a_i).\end{aligned}$$

Since $a_i > 0$ and $\sum_i a_i = 1$, let a be the random variable taking the value a_i with probability a_i. Jensen's Inequality (Theorem 4.7.7) applied to the concave function log gives

$$\sum_i a_i \log\left(\frac{1}{a_i}\right) = \mathrm{E} \log\left(\frac{1}{a}\right) \le \log\, \mathrm{E}\left(\frac{1}{a}\right) = \log \sum_i a_i \frac{1}{a_i} = \log n,$$

so the braces are nonnegative and $\frac{d}{dr} \log A_r \ge 0$ for every $r \ne 0$. Since A_r is continuous at $r = 0$ by part (a), A_r is nondecreasing on $-\infty < r < \infty$, and $A_{-1} \le A_0 \le A_1$ reads

$$\frac{n}{\sum_i (1/x_i)} \le \left(\prod_i x_i\right)^{1/n} \le \frac{1}{n}\sum_i x_i.$$

Problem 4.58. For any two random variables X and Y with finite variances, prove that
(a) $\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, \mathrm{E}(Y|X))$.
(b) X and $Y - \mathrm{E}(Y|X)$ are uncorrelated.
(c) $\mathrm{Var}(Y - \mathrm{E}(Y|X)) = \mathrm{E}(\mathrm{Var}(Y|X))$.

Solution. (a) Condition on X and use Theorem 4.4.3, noting that X is a constant given X:

$$\begin{aligned}\mathrm{Cov}(X, \mathrm{E}(Y|X)) &= \mathrm{E}\big[X\,\mathrm{E}(Y|X)\big] - \mathrm{E}X\,\mathrm{E}\big[\mathrm{E}(Y|X)\big] \\ &= \mathrm{E}\big[\mathrm{E}(XY|X)\big] - \mathrm{E}X\,\mathrm{E}Y \\ &= \mathrm{E}(XY) - \mathrm{E}X\,\mathrm{E}Y = \mathrm{Cov}(X, Y).\end{aligned}$$

(Finite variances give $\mathrm{E}|XY| < \infty$ by the Cauchy-Schwarz Inequality, Theorem 4.7.3, so the expectations exist as Theorem 4.4.3 requires.)
(b) By bilinearity of covariance and part (a),

$$\mathrm{Cov}\big(X, Y - \mathrm{E}(Y|X)\big) = \mathrm{Cov}(X, Y) - \mathrm{Cov}(X, \mathrm{E}(Y|X)) = 0.$$

(c) $\mathrm{E}[Y - \mathrm{E}(Y|X)] = \mathrm{E}Y - \mathrm{E}Y = 0$, so the variance is the raw second moment, and conditioning on X turns it into a conditional variance:

$$\begin{aligned}\mathrm{Var}\big(Y - \mathrm{E}(Y|X)\big) &= \mathrm{E}\big[(Y - \mathrm{E}(Y|X))^2\big] \\ &= \mathrm{E}\Big[\mathrm{E}\big((Y - \mathrm{E}(Y|X))^2 \,\big|\, X\big)\Big] \\ &= \mathrm{E}\big[\mathrm{Var}(Y|X)\big].\end{aligned}$$

Problem 4.59. For any three random variables X, Y, and Z with finite variances, prove (in the spirit of Theorem 4.4.7) the covariance identity

$$\mathrm{Cov}(X, Y) = \mathrm{E}\big(\mathrm{Cov}(X, Y|Z)\big) + \mathrm{Cov}\big(\mathrm{E}(X|Z), \mathrm{E}(Y|Z)\big),$$

where $\mathrm{Cov}(X, Y|Z)$ is the covariance of X and Y under the pdf $f(x, y|z)$.

Solution. Expand each term with the definition $\mathrm{Cov}(X, Y|Z) = \mathrm{E}(XY|Z) - \mathrm{E}(X|Z)\mathrm{E}(Y|Z)$ and apply Theorem 4.4.3, whose hypothesis holds because finite variances give $\mathrm{E}|XY| < \infty$ (Cauchy-Schwarz, Theorem 4.7.3):

$$\begin{aligned}\mathrm{E}\big(\mathrm{Cov}(X, Y|Z)\big) &= \mathrm{E}\big[\mathrm{E}(XY|Z)\big] - \mathrm{E}\big[\mathrm{E}(X|Z)\mathrm{E}(Y|Z)\big] \\ &= \mathrm{E}(XY) - \mathrm{E}\big[\mathrm{E}(X|Z)\mathrm{E}(Y|Z)\big],\end{aligned}$$

while, since $\mathrm{E}[\mathrm{E}(X|Z)] = \mathrm{E}X$ and likewise for Y,

$$\mathrm{Cov}\big(\mathrm{E}(X|Z), \mathrm{E}(Y|Z)\big) = \mathrm{E}\big[\mathrm{E}(X|Z)\mathrm{E}(Y|Z)\big] - \mathrm{E}X\,\mathrm{E}Y.$$

Adding, the cross terms cancel and

$$\begin{aligned}&\mathrm{E}\big(\mathrm{Cov}(X, Y|Z)\big) + \mathrm{Cov}\big(\mathrm{E}(X|Z), \mathrm{E}(Y|Z)\big) \\ &\quad = \mathrm{E}(XY) - \mathrm{E}X\,\mathrm{E}Y,\end{aligned}$$

which is $\mathrm{Cov}(X, Y)$.

Problem 4.60. Referring to Miscellanea 4.9.3, find the conditional distribution of Y given that $Y = X$ for each of the three interpretations given for the condition $Y = X$.
(In Miscellanea 4.9.3, X and Y are independent standard normal random variables, and the condition $Y = X$ is interpreted in the following three ways:

(1) $Z_1 = 0$, where $Z_1 = Y - X$;
(2) $Z_2 = 1$, where $Z_2 = Y/X$;
(3) $Z_3 = 1$, where $Z_3 = I(Y = X)$.
Each condition is a correct interpretation of the condition $Y = X$, and each leads to a different conditional distribution – this is the Borel Paradox.)

Solution. The three interpretations give $\mathrm{n}(0, \frac{1}{2})$, the density $|y|e^{-y^2}$, and no determinate answer at all.
(1) (Y, Z_1) is a linear map of the independent pair (X, Y), hence bivariate normal (Exercise 4.46), with means 0, $\operatorname{Var} Y = 1$, $\operatorname{Var} Z_1 = 2$, and $\operatorname{Cov}(Y, Z_1) = \operatorname{Var} Y - \operatorname{Cov}(Y, X) = 1$, so $\rho = 1/\sqrt{2}$. The conditional moments following Definition 4.5.10 give

$$\begin{aligned} \mathrm{E}(Y \mid Z_1 = 0) &= 0 + \tfrac{1}{2}(0 - 0) = 0, \\ \operatorname{Var}(Y \mid Z_1 = 0) &= 1 \cdot \left(1 - \tfrac{1}{2}\right) = \tfrac{1}{2}, \end{aligned}$$

so $Y \mid Z_1 = 0 \sim \mathrm{n}(0, \frac{1}{2})$.
(2) Change variables from (x, y) to (y, z) by $z = y/x$, that is $x = y/z$, with $|\partial x/\partial z| = |y|/z^2$ for fixed y:

$$f_{Y,Z_2}(y, z) = \frac{1}{2\pi} \exp\left\{-\frac{y^2}{2}\left(\frac{1}{z^2} + 1\right)\right\} \frac{|y|}{z^2}.$$

At $z = 1$ this is $\frac{1}{2\pi}|y|e^{-y^2}$, and $\int_{-\infty}^{\infty} |y|e^{-y^2}\, dy = 1$, so $f_{Z_2}(1) = 1/(2\pi)$ and

$$f(y \mid Z_2 = 1) = |y|e^{-y^2}, \qquad -\infty < y < \infty,$$

equivalently $Y^2 \sim$ exponential(1) with a symmetric random sign; in particular $\operatorname{Var}(Y|Z_2 = 1) = 1 \neq \frac{1}{2}$.
(3) Z_3 is discrete with $P(Z_3 = 1) = P(Y = X) = 0$, so there is no conditional density to normalize and the answer is not determined. In the sub sigma-algebra formulation of Miscellanea 4.9.3, $\mathcal{G} = \sigma(Z_3) = \{\emptyset, \Omega, \{Y = X\}, \{Y \neq X\}\}$ contains only sets of probability 0 or 1, so the defining relation

$$\mathrm{E}\big[Y I_G\big] = \mathrm{E}\big[\mathrm{E}(Y|\mathcal{G})\, I_G\big], \qquad G \in \mathcal{G},$$

holds whatever value $\mathrm{E}(Y|\mathcal{G})$ is assigned on $\{Y = X\}$; every distribution is a version. The default version, inherited from the almost-trivial $\mathcal{G}$, is the marginal $\mathrm{n}(0, 1)$ – different again from (1) and (2), which is the paradox.

Problem 4.61. DeGroot (1986) gives the following example of the Borel Paradox (Miscellanea 4.9.3): Suppose that X_1 and X_2 are iid exponential(1) random variables, and define $Z = (X_2 - 1)/X_1$. The probability-zero sets $\{Z = 0\}$ and $\{X_2 = 1\}$ seem to be giving us the same information but lead to different conditional distributions.
(a) Find the distribution of $X_1|Z = 0$, and compare it to the distribution of $X_1|X_2 = 1$.
(b) For small $\varepsilon > 0$ and $x_1 > 0$, $x_2 > 0$, consider the sets

$$B_1 = \left\{(x_1, x_2) : -\varepsilon < \frac{x_2 - 1}{x_1} < \varepsilon\right\} \quad \text{and} \quad B_2 = \{(x_1, x_2) : 1 - \varepsilon < x_2 < 1 + \varepsilon\}.$$

Draw these sets and support the argument that B_1 is informative about X_1 but B_2 is not.
(c) Calculate $P(X_1 \leq x|B_1)$ and $P(X_1 \leq x|B_2)$, and show that their limits (as $\varepsilon \to 0$) agree with part (a).
(*Communicated by L. Mark Berliner, Ohio State University.*)

Solution. (a) $X_1|Z = 0 \sim$ gamma$(2, 1)$, while $X_1|X_2 = 1 \sim$ exponential(1).
The second is immediate from independence. For the first, transform $(x_1, x_2) \mapsto (x_1, z)$ with $z = (x_2 - 1)/x_1$, so $x_2 = 1 + zx_1$ and $|\partial x_2/\partial z| = x_1$:

$$f_{X_1,Z}(x_1, z) = e^{-x_1} e^{-(1+zx_1)} x_1, \qquad x_1 > 0,\ 1 + zx_1 > 0.$$

At $z = 0$ this is $e^{-1}x_1e^{-x_1}$, and $\int_0^\infty x_1e^{-x_1}dx_1 = 1$, so

$$f(x_1|Z = 0) = x_1e^{-x_1}, \qquad x_1 > 0,$$

the gamma$(2, 1)$ pdf. Conditioning on $Z = 0$ tilts the exponential toward larger X_1; conditioning on $X_2 = 1$ leaves it alone.
(b) Since $x_1 > 0$, $B_1 = \{1 - \varepsilon x_1 < x_2 < 1 + \varepsilon x_1\}$ is the wedge opening rightward from the vertex $(0, 1)$ between the lines $x_2 = 1 \pm \varepsilon x_1$; its vertical cross-section at x_1 has length $2\varepsilon x_1$, growing with x_1. B_2 is the horizontal strip of constant height 2ε about $x_2 = 1$. The wedge admits a wider band of x_2 values the larger x_1 is, so B_1 is easier for large X_1 to satisfy and is informative about X_1; B_2 involves X_2 alone and X_1 is independent of X_2, so it says nothing about X_1.
(c) B_2 first: by independence, for every ε,

$$P(X_1 \le x \,|\, B_2) = P(X_1 \le x) = 1 - e^{-x},$$

the exponential(1) cdf, as in part (a). For B_1, take $\varepsilon < 1/x$ so that $1 - \varepsilon t > 0$ on $0 < t \le x$:

$$\begin{aligned} P\big(B_1 \cap \{X_1 \le x\}\big) &= \int_0^x e^{-t}\left[e^{-(1-\varepsilon t)} - e^{-(1+\varepsilon t)}\right] dt \\ &= 2e^{-1}\int_0^x e^{-t}\sinh(\varepsilon t)\, dt, \end{aligned}$$

and, splitting at $t = 1/\varepsilon$ where the wedge's lower edge hits $x_2 = 0$,

$$\begin{aligned} P(B_1) = 2e^{-1}&\int_0^{1/\varepsilon} e^{-t}\sinh(\varepsilon t)\, dt \\ &+ \int_{1/\varepsilon}^\infty e^{-t}\left[1 - e^{-(1+\varepsilon t)}\right] dt. \end{aligned}$$

Divide numerator and denominator by ε and let $\varepsilon \to 0$. The tail term is at most $e^{-1/\varepsilon} = o(\varepsilon)$; on $0 < t < 1/\varepsilon$ we have $\sinh(\varepsilon t)/\varepsilon \le t\cosh(\varepsilon t) \le t\cosh(1)$, so the Dominated Convergence Theorem applies with dominating function $\cosh(1)\, te^{-t}$, and $\sinh(\varepsilon t)/\varepsilon \to t$. Hence

$$\begin{aligned} \lim_{\varepsilon\to 0} P(X_1 \le x \,|\, B_1) &= \frac{\int_0^x te^{-t}dt}{\int_0^\infty te^{-t}dt} \\ &= 1 - (1 + x)e^{-x}, \end{aligned}$$

the gamma$(2, 1)$ cdf, again as in part (a).

Problem 4.62. Finish the proof of the equality in Jensen's Inequality (Theorem 4.7.7). Let $g(x)$ be a convex function. Suppose $a + bx$ is a line tangent to $g(x)$ at $x = \mathrm{E}X$, and $g(x) > a + bx$ except at $x = \mathrm{E}X$. Then $\mathrm{E}g(X) > g(\mathrm{E}X)$ unless $P(X = \mathrm{E}X) = 1$.

Solution. Put $\mu = \mathrm{E}X$ and $h(x) = g(x) - (a + bx)$, so that $h \ge 0$ with $h(x) = 0$ only at $x = \mu$, and, since the line is tangent at μ,

$$\mathrm{E}h(X) = \mathrm{E}g(X) - (a + b\mu) = \mathrm{E}g(X) - g(\mathrm{E}X).$$

Since $h \ge 0$ we always have $\mathrm{E}g(X) \ge g(\mathrm{E}X)$, so it suffices to show that a nonnegative $W = h(X)$ with $\mathrm{E}W = 0$ has $P(W = 0) = 1$. Chebychev's Inequality (Theorem 3.6.1, with the nonnegative $g(w) = w$ and $r = 1/n$) gives $P(W \ge \frac{1}{n}) \le n\,\mathrm{E}W = 0$ for every n, and

$$P(W > 0) = P\left(\bigcup_{n=1}^\infty \{W \ge \tfrac{1}{n}\}\right) \le \sum_{n=1}^\infty P\left(W \ge \tfrac{1}{n}\right) = 0.$$

So $\mathrm{E}g(X) = g(\mathrm{E}X)$ forces $P(h(X) = 0) = 1$, that is $P(X = \mu) = 1$; contrapositively, $P(X = \mathrm{E}X) < 1$ gives $\mathrm{E}g(X) > g(\mathrm{E}X)$.

Problem 4.63. A random variable X is defined by $Z = \log X$, where $\mathrm{E}Z = 0$. Is $\mathrm{E}X$ greater than, less than, or equal to 1?

Solution. Greater than 1 (unless $X \equiv 1$, in which case it equals 1).
Write $X = e^Z$. The function $g(z) = e^z$ is convex, with $g''(z) = e^z > 0$, so its tangent line at $z = \mathrm{E}Z$ lies strictly below g everywhere else, and Jensen's Inequality (Theorem 4.7.7) together with Exercise 4.62 gives

$$\mathrm{E}X = \mathrm{E}e^Z \;\geq\; e^{\mathrm{E}Z} = e^0 = 1,$$

with strict inequality unless $P(Z = \mathrm{E}Z) = 1$, that is unless $P(X = 1) = 1$.

4.10 Exercises 4.64–4.65

Problem 4.64. This exercise involves a well-known inequality known as the *triangle inequality* (a special case of Minkowski's Inequality).
(a) Prove (without using Minkowski's Inequality) that for any numbers a and b

$$|a + b| \leq |a| + |b|.$$

(b) Use part (a) to establish that for any random variables X and Y with finite expectations,

$$\mathrm{E}|X + Y| \leq \mathrm{E}|X| + \mathrm{E}|Y|.$$

Solution. (a) Square both sides and use $ab \leq |ab| = |a|\,|b|$:

$$\begin{aligned}(a+b)^2 &= a^2 + 2ab + b^2 \\ &\leq |a|^2 + 2|a|\,|b| + |b|^2 \\ &= \big(|a| + |b|\big)^2.\end{aligned}$$

Both sides are nonnegative, so taking square roots (increasing on $[0, \infty)$) gives $|a + b| \leq |a| + |b|$.
(b) Part (a) with $a = X(\omega)$, $b = Y(\omega)$ gives $|X+Y| \leq |X|+|Y|$ at every sample point, so monotonicity of bivariate expectation (Exercise 4.2(c)) and then its linearity (Exercise 4.2(a)) give

$$\begin{aligned}\mathrm{E}|X + Y| &\leq \mathrm{E}\big(|X| + |Y|\big) \\ &= \mathrm{E}|X| + \mathrm{E}|Y|,\end{aligned}$$

the right side being finite by hypothesis.

Problem 4.65. Prove the Covariance Inequality-II by generalizing the argument given in the text immediately preceding the inequality.
The inequality to be proved is Theorem 4.7.9 (Covariance Inequality-II): let X be any random variable and $g(x)$ and $h(x)$ any functions such that $\mathrm{E}g(X)$, $\mathrm{E}h(X)$, and $\mathrm{E}(g(X)h(X))$ exist.
(a) If $g(x)$ is a nondecreasing function and $h(x)$ is a nonincreasing function, then

$$\mathrm{E}(g(X)h(X)) \leq (\mathrm{E}g(X))(\mathrm{E}h(X)).$$

(b) If $g(x)$ and $h(x)$ are either both nondecreasing or both nonincreasing, then

$$\mathrm{E}(g(X)h(X)) \geq (\mathrm{E}g(X))(\mathrm{E}h(X)).$$

The argument immediately preceding the theorem is the following. If X has finite mean μ and $g(x)$

is nondecreasing, then $\mathrm{E}(g(X)(X-\mu)) \geq 0$, since

$$\begin{aligned}
&\mathrm{E}\big(g(X)(X-\mu)\big) \\
&\quad = \mathrm{E}\big(g(X)(X-\mu)I_{(-\infty,0)}(X-\mu)\big) \\
&\qquad + \mathrm{E}\big(g(X)(X-\mu)I_{[0,\infty)}(X-\mu)\big) \\
&\quad \geq \mathrm{E}\big(g(\mu)(X-\mu)I_{(-\infty,0)}(X-\mu)\big) \\
&\qquad + \mathrm{E}\big(g(\mu)(X-\mu)I_{[0,\infty)}(X-\mu)\big) \\
&\quad = g(\mu)\,\mathrm{E}(X-\mu) = 0,
\end{aligned}$$

the middle step using that g is nondecreasing.

Solution. Split where $h-\nu$ changes sign, $\nu = \mathrm{E}h(X)$, instead of where $X - \mathrm{E}X$ does. Since all three expectations exist, linearity (Theorem 2.2.5) gives, for every constant c,

$$\mathrm{E}\big[(g(X)-c)(h(X)-\nu)\big] = \mathrm{E}(g(X)h(X)) - \nu\,\mathrm{E}g(X),$$

the terms carrying c summing to $-c\,\mathrm{E}(h(X)-\nu) = 0$. So it suffices to produce a $c = g(\mu)$ making $(g(X)-g(\mu))(h(X)-\nu)$ of one fixed sign pointwise.
(a) g nondecreasing, h nonincreasing. Put

$$A = \{x : h(x) > \nu\}, \qquad \mu = \sup A \quad (\sup\emptyset = -\infty).$$

(i) $\mu = +\infty$: every x lies below some point of A, so $h > \nu$ everywhere and $\mathrm{E}(h(X)-\nu) > 0$, contradicting the definition of ν.
(ii) $\mu = -\infty$: then $A = \emptyset$, so $h(X) \leq \nu$ with $\mathrm{E}h(X) = \nu$, forcing $h(X) = \nu$ a.s. and (a) holds with equality.
(iii) μ finite, $c = g(\mu)$: for $x < \mu$ some $x' \in A$ exceeds x, so $h(x) \geq h(x') > \nu$ while $g(x) \leq g(\mu)$; for $x > \mu$, $x \notin A$ gives $h(x) \leq \nu$ while $g(x) \geq g(\mu)$; at $x = \mu$ the first factor vanishes. In every case

$$\big(g(x)-g(\mu)\big)\big(h(x)-\nu\big) \leq 0,$$

so taking expectations in the display above,

$$\mathrm{E}(g(X)h(X)) \leq (\mathrm{E}g(X))(\mathrm{E}h(X)).$$

(b) g and h both nondecreasing. Now put

$$A = \{x : h(x) > \nu\}, \qquad \mu = \inf A \quad (\inf\emptyset = +\infty),$$

the degenerate cases going exactly as in (i) and (ii): $\mu = -\infty$ forces $h > \nu$ everywhere, and $\mu = +\infty$ gives $h(X) = \nu$ a.s. and equality. For finite μ and $c = g(\mu)$: if $x < \mu$ then $x \notin A$, so $h(x) \leq \nu$ and $g(x) \leq g(\mu)$, both factors ≤ 0; if $x > \mu$ some $x' \in A$ is below x, so $h(x) \geq h(x') > \nu$ and $g(x) \geq g(\mu)$, both factors ≥ 0. Hence

$$\big(g(x)-g(\mu)\big)\big(h(x)-\nu\big) \geq 0$$

for all x, and taking expectations gives

$$\mathrm{E}(g(X)h(X)) \geq (\mathrm{E}g(X))(\mathrm{E}h(X)).$$

If g and h are both nonincreasing, apply this to $-g$ and $-h$, both nondecreasing: $(-g)(-h) = gh$ and $(\mathrm{E}(-g(X)))(\mathrm{E}(-h(X))) = (\mathrm{E}g(X))(\mathrm{E}h(X))$.

5 Properties of a Random Sample

Contents

5.1 Exercises 5.1–5.7 144
5.2 Exercises 5.8–5.14 149
5.3 Exercises 5.15–5.21 153
5.4 Exercises 5.22–5.28 158
5.5 Exercises 5.29–5.35 164
5.6 Exercises 5.36–5.42 168
5.7 Exercises 5.43–5.49 173
5.8 Exercises 5.50–5.56 178
5.9 Exercises 5.57–5.63 183
5.10 Exercises 5.64–5.69 188

5.1 Exercises 5.1–5.7

Problem 5.1. Color blindness appears in 1% of the people in a certain population. How large must a sample be if the probability of its containing a color-blind person is to be .95 or more? (Assume that the population is large enough to be considered infinite, so that sampling can be considered to be with replacement.)

Solution. $n = 299$.
Sampling with replacement makes the n indicators iid Bernoulli(.01), so the number X of color-blind people in the sample is binomial$(n, .01)$ and

$$P(X \geq 1) = 1 - (.99)^n \geq .95 \iff (.99)^n \leq .05.$$

Taking logarithms (and dividing by $\log(.99) < 0$, which reverses the inequality),

$$n \;\geq\; \frac{\log(.05)}{\log(.99)} \;=\; 298.073\ldots,$$

so the smallest admissible integer is $n = 299$ (at $n = 298$ the probability is only .94996).

Problem 5.2. Suppose $X_1, X_2, \ldots$ are jointly continuous and independent, each distributed with marginal pdf $f(x)$, where each X_i represents annual rainfall at a given location.
(a) Find the distribution of the number of years until the first year's rainfall, X_1, is exceeded for the first time.
(b) Show that the mean number of years until X_1 is exceeded for the first time is infinite.

Solution. (a) $P(N = n) = \dfrac{1}{n(n-1)}$ for $n = 2, 3, \ldots$, where $N = \min\{j \geq 2 : X_j > X_1\}$.
The event $\{N > n\}$ says that none of $X_2, \ldots, X_n$ exceeds X_1, i.e. that X_1 is the largest of $X_1, \ldots, X_n$. The X_i are iid and continuous, so ties have probability 0 and each of the n variables is equally likely to be the maximum; hence

$$P(N > n) = P\big(X_1 = \max\{X_1, \ldots, X_n\}\big) = \frac{1}{n}, \qquad n \geq 1.$$

Therefore, for $n \geq 2$,

$$P(N = n) = P(N > n-1) - P(N > n) = \frac{1}{n-1} - \frac{1}{n} = \frac{1}{n(n-1)}.$$

Method (2): condition on $X_1 = x$ and substitute $u = F(x)$ (Theorem 2.1.10),

$$P(N = n) = \int_{-\infty}^{\infty} F(x)^{n-2}\big(1 - F(x)\big) f(x)\, dx = \int_0^1 u^{n-2}(1-u)\, du = \frac{1}{n(n-1)}.$$

(b) Since $N \geq 1$ is integer valued, $\mathrm{E}N = \sum_{n=1}^{\infty} P(N \geq n) = \sum_{n=1}^{\infty} P(N > n-1)$, and by (a)

$$\mathrm{E}N = 1 + \sum_{n=2}^{\infty} \frac{1}{n-1} = 1 + \sum_{m=1}^{\infty} \frac{1}{m} = \infty,$$

the harmonic series diverging.

Problem 5.3. Let $X_1, \ldots, X_n$ be iid random variables with continuous cdf F_X, and suppose $\mathrm{E}X_i = \mu$. Define the random variables $Y_1, \ldots, Y_n$ by

$$Y_i = \begin{cases} 1 & \text{if } X_i > \mu, \\ 0 & \text{if } X_i \le \mu. \end{cases}$$

Find the distribution of $\sum_{i=1}^n Y_i$.

Solution. $\sum_{i=1}^n Y_i \sim \text{binomial}\big(n,\ 1 - F_X(\mu)\big)$.
Each Y_i takes only the values 0 and 1, with

$$P(Y_i = 1) = P(X_i > \mu) = 1 - F_X(\mu) \equiv p,$$

so $Y_i \sim \text{Bernoulli}(p)$, identically for each i. (Continuity of F_X gives $P(X_i = \mu) = 0$, so the endpoint convention in the definition of Y_i is immaterial.) Each $Y_i = I_{(\mu,\infty)}(X_i)$ is a function of X_i alone, so the Y_i are mutually independent by Theorem 4.6.12, and a sum of n iid Bernoulli(p) variables is binomial(n, p) (Example 5.2.12).

Problem 5.4. A generalization of iid random variables is *exchangeable* random variables, an idea due to deFinetti (1972). A discussion of exchangeability can also be found in Feller (1971). The random variables $X_1, \ldots, X_n$ are *exchangeable* if any permutation of any subset of them of size k ($k \le n$) has the same distribution. In this exercise we will see an example of random variables that are exchangeable but not iid. Let $X_i \mid P \sim$ iid Bernoulli(P), $i = 1, \ldots, n$, and let $P \sim \text{uniform}(0, 1)$.
(a) Show that the marginal distribution of any k of the Xs is the same as

$$P(X_1 = x_1, \ldots, X_k = x_k) = \int_0^1 p^t (1-p)^{k-t}\, dp = \frac{t!\,(k-t)!}{(k+1)!},$$

where $t = \sum_{i=1}^k x_i$. Hence, the Xs are exchangeable.
(b) Show that, marginally,

$$P(X_1 = x_1, \ldots, X_n = x_n) \neq \prod_{i=1}^n P(X_i = x_i),$$

so the distribution of the Xs is exchangeable but not iid.
(deFinetti proved an elegant characterization theorem for an infinite sequence of exchangeable random variables. He proved that any such sequence of exchangeable random variables is a mixture of iid random variables.)

Solution. (a) Fix any k indices $i_1 < \cdots < i_k$ and any $x_1, \ldots, x_k \in \{0, 1\}$, and put $t = \sum_{j=1}^k x_j$. Conditionally on $P = p$ the Xs are iid Bernoulli(p), so averaging the conditional pmf over the uniform(0, 1) marginal of P,

$$\begin{aligned} P(X_{i_1} = x_1, \ldots, X_{i_k} = x_k) &= \mathrm{E}\Big[\textstyle\prod_{j=1}^k P^{x_j}(1-P)^{1-x_j}\Big] \\ &= \int_0^1 p^t (1-p)^{k-t}\, dp. \end{aligned}$$

The integrand is the kernel of a beta($t+1, k-t+1$) pdf, so by the beta integral (Section 3.3)

$$\int_0^1 p^t (1-p)^{k-t}\, dp = \frac{\Gamma(t+1)\Gamma(k-t+1)}{\Gamma(k+2)} = \frac{t!\,(k-t)!}{(k+1)!}.$$

This value does not depend on which k indices were chosen, and depends on $(x_1, \ldots, x_k)$ only through the sum t, which is permutation invariant. Hence every size-k subset, in every order, has the same joint distribution: the Xs are exchangeable.
(b) Take $k = 1$ in (a): $P(X_i = 1) = \int_0^1 p\, dp = \frac{1}{2}$, so $\prod_{i=1}^n P(X_i = x_i) = 2^{-n}$ for every $(x_1, \ldots, x_n)$. But the joint pmf from (a) with $k = n$ is $t!\,(n-t)!/(n+1)!$, which does vary with t. Taking $x_1 = \cdots = x_n = 1$ (so $t = n$),

$$P(X_1 = 1, \ldots, X_n = 1) = \frac{n!\,0!}{(n+1)!} = \frac{1}{n+1} \;\neq\; \frac{1}{2^n} = \prod_{i=1}^{n} P(X_i = 1)$$

for every $n \geq 2$, since $2^n > n+1$ there. So the Xs are exchangeable but not independent, hence not iid.

Problem 5.5. Let $X_1, \ldots, X_n$ be iid with pdf $f_X(x)$, and let $\bar{X}$ denote the sample mean. Show that

$$f_{\bar{X}}(x) = n f_{X_1+\cdots+X_n}(nx),$$

even if the mgf of X does not exist.

Solution. With $T = X_1 + \cdots + X_n$, the map $\bar{X} = T/n$ is strictly increasing with inverse $g^{-1}(x) = nx$ and $|\frac{d}{dx} g^{-1}(x)| = n$, so Theorem 2.1.5 gives the claim directly:

$$F_{\bar{X}}(x) = P\left(\frac{T}{n} \leq x\right) = P(T \leq nx) = F_T(nx),$$

$$f_{\bar{X}}(x) = \frac{d}{dx} F_T(nx) = n\, f_T(nx) = n\, f_{X_1+\cdots+X_n}(nx).$$

(Theorem 2.1.5 needs only continuity of f_T and of $\frac{d}{dx} g^{-1}$, both immediate here.) No mgf enters, unlike the route through Theorem 5.2.7, which computes $M_{\bar{X}}(t) = [M_X(t/n)]^n$ and so presumes M_X exists.

Problem 5.6. If X has pdf $f_X(x)$ and Y, independent of X, has pdf $f_Y(y)$, establish formulas, similar to (5.2.3), for the random variable Z in each of the following situations.
(a) $Z = X - Y$
(b) $Z = XY$
(c) $Z = X/Y$
(Here (5.2.3) is the convolution formula of Theorem 5.2.9: for $Z = X+Y$, $f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z-w)\, dw$.)

Solution. Apply (4.3.2) to $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ with an auxiliary W, then integrate w out; throughout J is the determinant of the matrix of partials of (x,y) with respect to (z,w).
(a) $Z = X - Y$, $W = X$. The inverse is $x = w$, $y = w - z$, with

$$J = \begin{vmatrix} 0 & 1 \\ -1 & 1 \end{vmatrix} = 1,$$

so $f_{Z,W}(z,w) = f_X(w) f_Y(w-z)$ and

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w)\, f_Y(w-z)\, dw.$$

(b) $Z = XY$, $W = X$. The inverse is $x = w$, $y = z/w$ (defined for $w \neq 0$, and $P(X=0) = 0$ for a continuous X), with

$$J = \begin{vmatrix} 0 & 1 \\ 1/w & -z/w^2 \end{vmatrix} = -\frac{1}{w}, \qquad |J| = \frac{1}{|w|},$$

so $f_{Z,W}(z,w) = f_X(w) f_Y(z/w)/|w|$ and

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w)\, f_Y\Big(\frac{z}{w}\Big)\, \frac{1}{|w|}\, dw.$$

(c) $Z = X/Y$, $W = Y$. The inverse is $x = zw$, $y = w$, with

$$J = \begin{vmatrix} w & z \\ 0 & 1 \end{vmatrix} = w, \qquad |J| = |w|,$$

so $f_{Z,W}(z,w) = f_X(zw) f_Y(w)\,|w|$ and

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(zw)\, f_Y(w)\,|w|\, dw.$$

Problem 5.7. In Example 5.2.10, a partial fraction decomposition is needed to derive the distribution of the sum of two independent Cauchy random variables. This exercise provides the details that are skipped in that example.
(a) Find the constants A, B, C, and D that satisfy

$$\frac{1}{1+(w/\sigma)^2}\,\frac{1}{1+((z-w)/\tau)^2} = \frac{Aw}{1+(w/\sigma)^2} + \frac{B}{1+(w/\sigma)^2} - \frac{Cw}{1+((z-w)/\tau)^2} - \frac{D}{1+((z-w)/\tau)^2},$$

where A, B, C, and D may depend on z but not on w.
(b) Using the facts that

$$\int \frac{t}{1+t^2}\,dt = \tfrac{1}{2}\log(1+t^2) + \text{constant} \quad \text{and} \quad \int \frac{1}{1+t^2}\,dt = \arctan(t) + \text{constant},$$

evaluate (5.2.4) and hence verify (5.2.5).
(Note that the integration in part (b) is quite delicate. Since the mean of a Cauchy does not exist, the integrals $\int_{-\infty}^{\infty} \frac{Aw}{1+(w/\sigma)^2}\,dw$ and $\int_{-\infty}^{\infty} \frac{Cw}{1+((z-w)/\tau)^2}\,dw$ do not exist. However, the integral of the difference *does* exist, which is all that is needed.)
Here, with $U \sim \text{Cauchy}(0,\sigma)$ and $V \sim \text{Cauchy}(0,\tau)$ independent and $Z = U + V$, Example 5.2.10 writes

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{\pi\sigma}\,\frac{1}{1+(w/\sigma)^2}\,\frac{1}{\pi\tau}\,\frac{1}{1+((z-w)/\tau)^2}\,dw, \tag{5.2.4}$$

$$f_Z(z) = \frac{1}{\pi(\sigma+\tau)}\,\frac{1}{1+(z/(\sigma+\tau))^2}, \qquad -\infty < z < \infty. \tag{5.2.5}$$

Solution. (a) With $\Delta = \big(z^2+(\sigma+\tau)^2\big)\big(z^2+(\sigma-\tau)^2\big)$,

$$A = \frac{2\tau^2 z}{\Delta}, \qquad B = \frac{\tau^2\,(z^2+\tau^2-\sigma^2)}{\Delta},$$
$$C = \frac{2\sigma^2 z}{\Delta}, \qquad D = -\,\frac{\sigma^2\,(3z^2+\sigma^2-\tau^2)}{\Delta}.$$

To get them, abbreviate $u(w) = 1+(w/\sigma)^2$ and $v(w) = 1+((z-w)/\tau)^2$. The asserted identity is $1/(uv) = (Aw+B)/u - (Cw+D)/v$, i.e., after multiplying by uv, the polynomial identity in w

$$(Aw+B)\Big(1+\tfrac{w^2-2zw+z^2}{\tau^2}\Big) - (Cw+D)\Big(1+\tfrac{w^2}{\sigma^2}\Big) = 1.$$

Matching coefficients of w^3, w^2, w^1, w^0 gives

$$w^3: \ \tfrac{A}{\tau^2} - \tfrac{C}{\sigma^2} = 0, \qquad w^2: \ \tfrac{B-2Az}{\tau^2} - \tfrac{D}{\sigma^2} = 0,$$
$$w^1: \ A\tfrac{\tau^2+z^2}{\tau^2} - \tfrac{2Bz}{\tau^2} - C = 0, \qquad w^0: \ B\tfrac{\tau^2+z^2}{\tau^2} - D = 1.$$

The first gives $C = A\sigma^2/\tau^2$; substituting into the third gives $AS = 2Bz$ with $S = \tau^2+z^2-\sigma^2$. The fourth gives $D = B(\tau^2+z^2)/\tau^2 - 1$, and feeding both into the second leaves

$$B\,\frac{S^2+4\sigma^2z^2}{S}=\tau^2,\qquad\text{where}\quad S^2+4\sigma^2z^2=\Delta,$$

the last equality because $(\tau^2+z^2-\sigma^2)^2+4\sigma^2z^2=z^4+2z^2(\sigma^2+\tau^2)+(\sigma^2-\tau^2)^2$, which is Δ. (Check!) Hence $B=\tau^2S/\Delta$, $A=2Bz/S=2\tau^2z/\Delta$, $C=2\sigma^2z/\Delta$, and

$$D=\frac{(\tau^2+z^2)S-\Delta}{\Delta}=-\frac{\sigma^2(3z^2+\sigma^2-\tau^2)}{\Delta},$$

as claimed. Note for later that

$$A\sigma^2=C\tau^2=\frac{2\sigma^2\tau^2z}{\Delta}\equiv K.$$

(b) Rewrite the two pieces of (a) with their denominators cleared, and split the second numerator about $w=z$: put

$$D'=Cz+D=-\frac{\sigma^2(z^2+\sigma^2-\tau^2)}{\Delta}.$$

Then the integrand of (5.2.4), apart from the factor $1/(\pi^2\sigma\tau)$, equals

$$\frac{\sigma^2(Aw+B)}{\sigma^2+w^2}-\frac{\tau^2\big(C(w-z)+D'\big)}{\tau^2+(w-z)^2},$$

whose antiderivative, by the two stated integration facts, is

$$\begin{aligned}G(w)=\frac{K}{2}\log\frac{\sigma^2+w^2}{\tau^2+(w-z)^2}+B\sigma\arctan\frac{w}{\sigma}\\ -D'\tau\arctan\frac{w-z}{\tau},\end{aligned}$$

the two logarithms having been combined using $A\sigma^2=C\tau^2=K$ – the delicate point flagged in the problem, since each diverges separately as $w\to\pm\infty$ while their difference tends to $\frac{K}{2}\log 1=0$. With $\arctan(\pm\infty)=\pm\pi/2$,

$$\int_{-\infty}^{\infty}\frac{dw}{u(w)v(w)}=\lim_{w\to\infty}G(w)-\lim_{w\to-\infty}G(w)=\pi\big(B\sigma-D'\tau\big).$$

Finally

$$\begin{aligned}B\sigma-D'\tau&=\frac{\sigma\tau}{\Delta}\Big[\tau(z^2+\tau^2-\sigma^2)+\sigma(z^2+\sigma^2-\tau^2)\Big]\\&=\frac{\sigma\tau(\sigma+\tau)\big(z^2+(\sigma-\tau)^2\big)}{\Delta}=\frac{\sigma\tau(\sigma+\tau)}{z^2+(\sigma+\tau)^2},\end{aligned}$$

the middle step being the factorization

$$\begin{aligned}z^2(\sigma+\tau)+(\sigma^3+\tau^3)-\sigma\tau(\sigma+\tau)&=(\sigma+\tau)\big[z^2+\sigma^2-\sigma\tau+\tau^2-\sigma\tau\big]\\&=(\sigma+\tau)\big(z^2+(\sigma-\tau)^2\big),\end{aligned}$$

after which the factor $z^2+(\sigma-\tau)^2$ cancels against Δ. Therefore

$$f_Z(z)=\frac{1}{\pi^2\sigma\tau}\cdot\frac{\pi\,\sigma\tau(\sigma+\tau)}{z^2+(\sigma+\tau)^2}=\frac{\sigma+\tau}{\pi\big(z^2+(\sigma+\tau)^2\big)},$$

which is exactly

$$f_Z(z)=\frac{1}{\pi(\sigma+\tau)}\,\frac{1}{1+\big(z/(\sigma+\tau)\big)^2},$$

i.e. $Z\sim\text{Cauchy}(0,\sigma+\tau)$, verifying (5.2.5).

5.2 Exercises 5.8–5.14

Problem 5.8. Let $X_1, \dots, X_n$ be a random sample, where $\bar{X}$ and S^2 are calculated in the usual way.
(a) Show that

$$S^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^2.$$

Assume now that the X_is have a finite fourth moment, and denote $\theta_1 = \mathrm{E}X_i$, $\theta_j = \mathrm{E}(X_i - \theta_1)^j$, $j = 2, 3, 4$.
(b) Show that $\operatorname{Var} S^2 = \frac{1}{n}\left(\theta_4 - \frac{n-3}{n-1}\theta_2^2\right)$.
(c) Find $\operatorname{Cov}(\bar{X}, S^2)$ in terms of $\theta_1, \dots, \theta_4$. Under what conditions is $\operatorname{Cov}(\bar{X}, S^2) = 0$?

Solution. (a) Insert $\pm\bar{X}$ and expand; the cross-term dies because $\sum_i (X_i - \bar{X}) = 0$:

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^2 &= \sum_{i,j} \big[(X_i - \bar{X}) - (X_j - \bar{X})\big]^2 \\
&= n \sum_i (X_i - \bar{X})^2 + n \sum_j (X_j - \bar{X})^2 \\
&\qquad - 2\Big(\sum_i (X_i - \bar{X})\Big)^2 \\
&= 2n(n-1)S^2.
\end{aligned}$$

(b) Both S^2 and the θ_j, $j \geq 2$, are unchanged by a location shift, so take $\theta_1 = 0$. Put $T = \sum_i X_i^2$ and $U = \sum_i X_i$, so $(n-1)S^2 = T - U^2/n$. Since the X_i are iid with mean 0, every product whose expectation survives must pair the indices up:

$$\begin{aligned}
\mathrm{E}T^2 &= n\theta_4 + n(n-1)\theta_2^2, \\
\mathrm{E}(TU^2) &= n\theta_4 + n(n-1)\theta_2^2, \\
\mathrm{E}U^4 &= n\theta_4 + 3n(n-1)\theta_2^2
\end{aligned}$$

(in $\mathrm{E}(TU^2)$ the terms $\mathrm{E}X_i^2 X_j X_k$ with $j \neq k$ all vanish, since some index then carries an odd, unpaired power). Hence

$$\begin{aligned}
\mathrm{E}\big[(n-1)S^2\big]^2 &= \mathrm{E}T^2 - \tfrac{2}{n}\mathrm{E}(TU^2) + \tfrac{1}{n^2}\mathrm{E}U^4 \\
&= \frac{(n-1)^2}{n}\,\theta_4 + \frac{(n-1)(n^2-2n+3)}{n}\,\theta_2^2.
\end{aligned}$$

Subtracting $\big(\mathrm{E}(n-1)S^2\big)^2 = (n-1)^2\theta_2^2$ (Theorem 5.2.6c) and dividing by $(n-1)^2$,

$$\operatorname{Var} S^2 = \frac{\theta_4}{n} + \frac{(3-n)\theta_2^2}{n(n-1)} = \frac{1}{n}\left(\theta_4 - \frac{n-3}{n-1}\theta_2^2\right).$$

(c) $\operatorname{Cov}(\bar{X}, S^2) = \theta_3/n$, zero if and only if the population third central moment vanishes. Still with $\theta_1 = 0$,

$$\begin{aligned}
\operatorname{Cov}(\bar{X}, S^2) = \mathrm{E}(\bar{X}S^2) &= \frac{1}{n(n-1)}\Big[\mathrm{E}(UT) - \tfrac{1}{n}\mathrm{E}U^3\Big] \\
&= \frac{1}{n(n-1)}\Big[n\theta_3 - \theta_3\Big] = \frac{\theta_3}{n},
\end{aligned}$$

using $\mathrm{E}(UT) = \sum_i \mathrm{E}X_i^3 = n\theta_3$ and $\mathrm{E}U^3 = n\theta_3$ (all mixed terms have an unpaired mean-zero factor). So $\operatorname{Cov}(\bar{X}, S^2) = 0$ exactly when $\theta_3 = \mathrm{E}(X_i - \theta_1)^3 = 0$.

Problem 5.9. Establish the *Lagrange Identity*, that for any numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) - \left(\sum_{i=1}^{n} a_i b_i\right)^2 = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} (a_i b_j - a_j b_i)^2.$$

Use the identity to show that the squared correlation coefficient is equal to 1 if and only if all of the sample points lie on a straight line (Wright 1992). (*Hint:* Establish the identity for $n = 2$; then induct.)

Solution. Write $A_n = \sum_{i\le n} a_i^2$, $B_n = \sum_{i\le n} b_i^2$, $C_n = \sum_{i\le n} a_i b_i$ and $L_n = \sum_{i<j\le n}(a_i b_j - a_j b_i)^2$; the claim is $A_n B_n - C_n^2 = L_n$.
For $n = 2$,

$$(a_1^2 + a_2^2)(b_1^2 + b_2^2) - (a_1 b_1 + a_2 b_2)^2 = (a_1 b_2 - a_2 b_1)^2.$$

Assume the identity at n. Then

$$\begin{aligned} A_{n+1}B_{n+1} - C_{n+1}^2 &= (A_n + a_{n+1}^2)(B_n + b_{n+1}^2) - (C_n + a_{n+1}b_{n+1})^2 \\ &= \left(A_nB_n - C_n^2\right) + a_{n+1}^2 B_n + b_{n+1}^2 A_n - 2a_{n+1}b_{n+1}C_n \\ &= L_n + \sum_{i=1}^{n}\left(a_i b_{n+1} - a_{n+1}b_i\right)^2 \;=\; L_{n+1}, \end{aligned}$$

which completes the induction.
Now apply it with $a_i = x_i - \bar{x}$, $b_i = y_i - \bar{y}$. The sample correlation is $r = C_n/\sqrt{A_nB_n}$, defined when $A_nB_n > 0$, and the identity gives

$$1 - r^2 = \frac{A_nB_n - C_n^2}{A_nB_n} = \frac{\sum_{i<j}(a_ib_j - a_jb_i)^2}{A_nB_n}.$$

Hence $r^2 = 1$ if and only if $a_ib_j = a_jb_i$ for every pair $i < j$. Since $A_n > 0$ there is a k with $a_k \neq 0$; setting $c = b_k/a_k$, the vanishing of $a_kb_i - a_ib_k$ gives $b_i = ca_i$ for every i, i.e.

$$y_i - \bar{y} = c\,(x_i - \bar{x}), \qquad i = 1, \ldots, n,$$

so all the points (x_i, y_i) lie on the line through $(\bar{x}, \bar{y})$ of slope c. Conversely, if $y_i = \alpha + \beta x_i$ for all i then $b_i = \beta a_i$, every $a_ib_j - a_jb_i$ vanishes, and $r^2 = 1$.

Problem 5.10. Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ population.
(a) Find expressions for $\theta_1, \ldots, \theta_4$, as defined in Exercise 5.8, in terms of μ and σ^2.
(b) Use the results of Exercise 5.8, together with the results of part (a), to calculate $\operatorname{Var} S^2$.
(c) Calculate $\operatorname{Var} S^2$ a completely different (and easier) way: Use the fact that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Solution. (a) $\theta_1 = \mu$, $\theta_2 = \sigma^2$, $\theta_3 = 0$, $\theta_4 = 3\sigma^4$. The odd central moment vanishes by symmetry of the normal pdf about μ, and $Z = (X_i - \mu)/\sigma \sim \mathrm{n}(0,1)$ has $\mathrm{E}Z^4 = 3\,\mathrm{E}Z^2 = 3$ by the recursion $\mathrm{E}Z^k = (k-1)\mathrm{E}Z^{k-2}$ (integrate $z^{k-1}\cdot z\varphi(z)$ by parts, using $\varphi'(z) = -z\varphi(z)$).
(b) By Exercise 5.8(b),

$$\begin{aligned} \operatorname{Var} S^2 &= \frac{1}{n}\left(3\sigma^4 - \frac{n-3}{n-1}\sigma^4\right) = \frac{\sigma^4}{n}\cdot\frac{3(n-1) - (n-3)}{n-1} \\ &= \frac{\sigma^4}{n}\cdot\frac{2n}{n-1} = \frac{2\sigma^4}{n-1}. \end{aligned}$$

(c) By Theorem 5.3.1c, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, which has variance $2(n-1)$. Therefore

$$\begin{aligned}\operatorname{Var} S^2 &= \frac{\sigma^4}{(n-1)^2}\operatorname{Var}\Big(\tfrac{(n-1)S^2}{\sigma^2}\Big)\\ &= \frac{\sigma^4 \cdot 2(n-1)}{(n-1)^2} = \frac{2\sigma^4}{n-1}.\end{aligned}$$

Problem 5.11. Suppose $\bar{X}$ and S^2 are calculated from a random sample $X_1, \ldots, X_n$ drawn from a population with finite variance σ^2. We know that $\mathrm{E}S^2 = \sigma^2$. Prove that $\mathrm{E}S \le \sigma$, and if $\sigma^2 > 0$, then $\mathrm{E}S < \sigma$.

Solution. $0 \le \operatorname{Var} S = \mathrm{E}S^2 - (\mathrm{E}S)^2 = \sigma^2 - (\mathrm{E}S)^2$ by Theorem 5.2.6c, and $S \ge 0$, so $\mathrm{E}S \le \sigma$.
Equality forces $\operatorname{Var} S = 0$, i.e. S — equivalently S^2 — is degenerate. So it suffices to show that $\sigma^2 > 0$ makes S^2 non-degenerate. Pick support points $a < b$ of the population (they exist, since $\sigma^2 > 0$ rules out a point mass), and for $\varepsilon \in (0, (b-a)/2)$ let $I_a = (a-\varepsilon, a+\varepsilon)$, $I_b = (b-\varepsilon, b+\varepsilon)$, so that $p = P(X_1 \in I_a) > 0$ and $q = P(X_1 \in I_b) > 0$. Using the representation of Exercise 5.8(a), on the event

$$A = \{X_1 \in I_a, \ldots, X_n \in I_a\}, \qquad P(A) = p^n > 0,$$

every one of the $n(n-1)$ off-diagonal terms satisfies $(X_i - X_j)^2 < 4\varepsilon^2$, whence $S^2 < 2\varepsilon^2$; while on

$$B = \{X_1 \in I_a,\ X_2 \in I_b, \ldots, X_n \in I_b\}, \qquad P(B) = pq^{n-1} > 0,$$

the $2(n-1)$ terms involving index 1 each exceed $(b-a-2\varepsilon)^2$, whence

$$S^2 > \frac{2(n-1)(b-a-2\varepsilon)^2}{2n(n-1)} = \frac{(b-a-2\varepsilon)^2}{n}.$$

As $\varepsilon \downarrow 0$ the right-hand bound tends to $(b-a)^2/n > 0$ while $2\varepsilon^2 \to 0$, so fix ε small enough that $2\varepsilon^2 < (b-a-2\varepsilon)^2/n$. Then S^2 takes values in two disjoint intervals, each with positive probability, so $\operatorname{Var} S > 0$ and $\mathrm{E}S < \sigma$.

Problem 5.12. Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(0,1)$ population. Define

$$Y_1 = \left|\frac{1}{n}\sum_{i=1}^n X_i\right|, \qquad Y_2 = \frac{1}{n}\sum_{i=1}^n |X_i|.$$

Calculate $\mathrm{E}Y_1$ and $\mathrm{E}Y_2$, and establish an inequality between them.

Solution.

$$\mathrm{E}Y_1 = \sqrt{\frac{2}{\pi n}}, \qquad \mathrm{E}Y_2 = \sqrt{\frac{2}{\pi}}, \qquad \mathrm{E}Y_1 \le \mathrm{E}Y_2.$$

For $Z \sim \mathrm{n}(0, \tau^2)$,

$$\mathrm{E}|Z| = 2\int_0^\infty \frac{z}{\sqrt{2\pi}\,\tau} e^{-z^2/(2\tau^2)}\, dz = \tau\sqrt{\frac{2}{\pi}}.$$

Apply this twice: $\bar{X} \sim \mathrm{n}(0, 1/n)$ by Theorem 5.3.1b, giving $\mathrm{E}Y_1 = \mathrm{E}|\bar{X}| = \sqrt{2/(\pi n)}$; and each $X_i \sim \mathrm{n}(0,1)$, so $\mathrm{E}Y_2 = \mathrm{E}|X_1| = \sqrt{2/\pi}$.
The inequality $\mathrm{E}Y_1 \le \mathrm{E}Y_2$ is the triangle inequality $|\sum_i X_i| \le \sum_i |X_i|$ in expectation, with equality only at $n = 1$.

Problem 5.13. Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$. Find a function of S^2, the sample variance, say $g(S^2)$, that satisfies $\mathrm{E}g(S^2) = \sigma$. (*Hint:* Try $g(S^2) = c\sqrt{S^2}$, where c is a constant.)

Solution.

$$g(S^2) = c\,S, \qquad c = \sqrt{\tfrac{n-1}{2}}\,\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.$$

By Theorem 5.3.1c, $W = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, and for $W \sim \chi^2_p$ the moment $\mathrm{E}W^{1/2}$ is obtained by recognizing a gamma kernel:

$$\begin{aligned}\mathrm{E}W^{1/2} &= \int_0^\infty \frac{w^{(p-1)/2}e^{-w/2}}{\Gamma(p/2)2^{p/2}}\,dw \\ &= \frac{\Gamma\left(\frac{p+1}{2}\right)2^{(p+1)/2}}{\Gamma(p/2)2^{p/2}} = \sqrt{2}\,\frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)}.\end{aligned}$$

With $p = n-1$ and $S = \sigma\sqrt{W}/\sqrt{n-1}$,

$$\mathrm{E}S = \sigma\sqrt{\frac{2}{n-1}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)},$$

so $c = 1/\left[\sqrt{2/(n-1)}\,\Gamma(\frac{n}{2})/\Gamma(\frac{n-1}{2})\right]$ is the constant displayed above.

Problem 5.14. Lemma 5.3.3 states: let $X_j \sim \mathrm{n}(\mu_j, \sigma_j^2)$, $j = 1, \ldots, n$, independent; for constants a_{ij} and b_{rj} ($j = 1, \ldots, n$; $i = 1, \ldots, k$; $r = 1, \ldots, m$), where $k + m \le n$, define

$$U_i = \sum_{j=1}^n a_{ij}X_j, \quad i = 1, \ldots, k; \qquad V_r = \sum_{j=1}^n b_{rj}X_j, \quad r = 1, \ldots, m.$$

Part a of the lemma: U_i and V_r are independent if and only if $\mathrm{Cov}(U_i, V_r) = 0$; furthermore $\mathrm{Cov}(U_i, V_r) = \sum_{j=1}^n a_{ij}b_{rj}\sigma_j^2$. Part b: the vectors $(U_1, \ldots, U_k)$ and $(V_1, \ldots, V_m)$ are independent if and only if U_i is independent of V_r for all pairs i, r. The exercise is:
(a) Prove that the statement of Lemma 5.3.3 follows from the special case of $\mu_i = 0$ and $\sigma_i^2 = 1$. That is, show that if $X_j = \sigma_j Z_j + \mu_j$ and $Z_j \sim \mathrm{n}(0, 1)$, $j = 1, \ldots, n$, all independent, a_{ij}, b_{rj} are constants, and

$$\mathrm{Cov}\left(\sum_{j=1}^n a_{ij}Z_j,\ \sum_{j=1}^n b_{rj}Z_j\right) = 0 \ \Rightarrow\ \sum_{j=1}^n a_{ij}Z_j,\ \sum_{j=1}^n b_{rj}Z_j \ \text{ independent},$$

then

$$\mathrm{Cov}\left(\sum_{j=1}^n a_{ij}X_j,\ \sum_{j=1}^n b_{rj}X_j\right) = 0 \ \Rightarrow\ \sum_{j=1}^n a_{ij}X_j,\ \sum_{j=1}^n b_{rj}X_j \ \text{ independent}.$$

(b) Verify the expression for $\mathrm{Cov}\left(\sum_{j=1}^n a_{ij}X_j,\ \sum_{j=1}^n b_{rj}X_j\right)$ in Lemma 5.3.3.

Solution. (a) Absorb the scale into the coefficients: with $\tilde{a}_{ij} = a_{ij}\sigma_j$ and $\tilde{b}_{rj} = b_{rj}\sigma_j$,

$$U_i = \sum_j a_{ij}X_j = \sum_j \tilde{a}_{ij}Z_j + \alpha_i, \qquad V_r = \sum_j b_{rj}X_j = \sum_j \tilde{b}_{rj}Z_j + \beta_r,$$

where $\alpha_i = \sum_j a_{ij}\mu_j$ and $\beta_r = \sum_j b_{rj}\mu_j$ are constants. Covariance is unaffected by additive constants

(Definition 4.5.1 centers each argument), so

$$\operatorname{Cov}(U_i, V_r) = \operatorname{Cov}\Big(\sum_j \tilde{a}_{ij} Z_j,\ \sum_j \tilde{b}_{rj} Z_j\Big).$$

Hence if the left side is 0, the hypothesised standard-normal case applies to the coefficient arrays $\tilde{a}, \tilde{b}$ and yields that $\sum_j \tilde{a}_{ij} Z_j$ and $\sum_j \tilde{b}_{rj} Z_j$ are independent. Independence is preserved by the shifts $u \mapsto u + \alpha_i$ and $v \mapsto v + \beta_r$ (Theorem 4.6.12, applied to the two independent variables), so U_i and V_r are independent. The converse direction of the lemma needs no reduction: independent random variables with finite variances have zero covariance (Theorem 4.5.5). Part (b) of the lemma transfers by the same substitution, applied to the two vectors.
(b) Bilinearity of covariance and independence of the X_j (so $\operatorname{Cov}(X_j, X_l) = 0$ for $j \neq l$, by Theorem 4.5.5) give

$$\begin{aligned}\operatorname{Cov}\Big(\sum_j a_{ij} X_j,\ \sum_l b_{rl} X_l\Big) &= \sum_{j=1}^{n}\sum_{l=1}^{n} a_{ij} b_{rl} \operatorname{Cov}(X_j, X_l) \\ &= \sum_{j=1}^{n} a_{ij} b_{rj} \operatorname{Var} X_j = \sum_{j=1}^{n} a_{ij} b_{rj} \sigma_j^2.\end{aligned}$$

5.3 Exercises 5.15–5.21

Problem 5.15. Establish the following recursion relations for means and variances. Let $\bar{X}_n$ and S_n^2 be the mean and variance, respectively, of $X_1, \ldots, X_n$. Then suppose another observation, X_{n+1}, becomes available. Show that
(a) $\bar{X}_{n+1} = \dfrac{X_{n+1} + n\bar{X}_n}{n+1}$.
(b) $nS_{n+1}^2 = (n-1)S_n^2 + \left(\dfrac{n}{n+1}\right)(X_{n+1} - \bar{X}_n)^2$.

Solution. (a) Split off the last term of the sum:

$$\bar{X}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} X_i = \frac{X_{n+1} + \sum_{i=1}^{n} X_i}{n+1} = \frac{X_{n+1} + n\bar{X}_n}{n+1}.$$

(b) Use $\sum_{i=1}^{m}(X_i - \bar{X}_m)^2 = \sum_{i=1}^{m} X_i^2 - m\bar{X}_m^2$ for $m = n+1$ and $m = n$ and subtract:

$$nS_{n+1}^2 - (n-1)S_n^2 = X_{n+1}^2 + n\bar{X}_n^2 - (n+1)\bar{X}_{n+1}^2.$$

By part (a), $(n+1)\bar{X}_{n+1}^2 = (X_{n+1} + n\bar{X}_n)^2/(n+1)$, so the right side equals

$$\begin{aligned}&\frac{(n+1)X_{n+1}^2 + n(n+1)\bar{X}_n^2 - (X_{n+1} + n\bar{X}_n)^2}{n+1} \\ &\quad = \frac{nX_{n+1}^2 - 2nX_{n+1}\bar{X}_n + n\bar{X}_n^2}{n+1} \\ &\quad = \frac{n}{n+1}(X_{n+1} - \bar{X}_n)^2.\end{aligned}$$

Problem 5.16. Let X_i, $i = 1, 2, 3$, be independent with $n(i, i^2)$ distributions. For each of the following situations, use the X_is to construct a statistic with the indicated distribution.
(a) chi squared with 3 degrees of freedom
(b) t distribution with 2 degrees of freedom
(c) F distribution with 1 and 2 degrees of freedom

Solution. Standardize once: $Z_i = (X_i - i)/i$ are iid $n(0,1)$, i.e.

$$Z_1 = X_1 - 1, \qquad Z_2 = \frac{X_2 - 2}{2}, \qquad Z_3 = \frac{X_3 - 3}{3}.$$

(a) By Lemma 5.3.2, $Z_1^2 + Z_2^2 + Z_3^2 \sim \chi_3^2$:

$$(X_1 - 1)^2 + \left(\frac{X_2 - 2}{2}\right)^2 + \left(\frac{X_3 - 3}{3}\right)^2.$$

(b) Abbreviate $Q = Z_2^2 + Z_3^2$, the χ_2^2 variable of part (a) with its first term dropped, so

$$Q = \left(\frac{X_2 - 2}{2}\right)^2 + \left(\frac{X_3 - 3}{3}\right)^2.$$

Then $Z_1 \sim n(0,1)$ is independent of $Q \sim \chi_2^2$, so by Definition 5.3.4

$$T = \frac{Z_1}{\sqrt{Q/2}} = \frac{X_1 - 1}{\sqrt{Q/2}} \sim t_2.$$

(c) Square the statistic in (b); by Theorem 5.3.8(b) it is $F_{1,2}$:

$$T^2 = \frac{(X_1 - 1)^2}{Q/2} \sim F_{1,2}.$$

Problem 5.17. Let X be a random variable with an $F_{p,q}$ distribution.
(a) Derive the pdf of X.
(b) Derive the mean and variance of X.
(c) Show that $1/X$ has an $F_{q,p}$ distribution.
(d) Show that $(p/q)X/[1 + (p/q)X]$ has a beta distribution with parameters $p/2$ and $q/2$.

Solution. Throughout write $X = (U/p)/(V/q)$ with $U \sim \chi_p^2$, $V \sim \chi_q^2$ independent, the reduction noted after Definition 5.3.6.
(a) Transform $x = qu/(pv)$, $w = v$, so $u = (p/q)xw$, $v = w$, with Jacobian $|\partial(u,v)/\partial(x,w)| = (p/q)w$. Then

$$\begin{aligned} f_X(x) &= \int_0^\infty \frac{\left(\frac{p}{q}xw\right)^{p/2-1} e^{-\frac{p}{2q}xw}\, w^{q/2-1} e^{-w/2}}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})2^{(p+q)/2}} \cdot \frac{p}{q} w\, dw \\ &= \frac{(p/q)^{p/2} x^{p/2-1}}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})2^{(p+q)/2}} \int_0^\infty w^{\frac{p+q}{2}-1} e^{-\frac{w}{2}\left(1+\frac{p}{q}x\right)} dw. \end{aligned}$$

The integrand is a gamma$\left(\frac{p+q}{2},\ \frac{2}{1+(p/q)x}\right)$ kernel, so the integral equals $\Gamma\left(\frac{p+q}{2}\right)\left[\frac{2}{1+(p/q)x}\right]^{(p+q)/2}$, giving

$$f_X(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})} \left(\frac{p}{q}\right)^{p/2} \frac{x^{p/2-1}}{[1 + (p/q)x]^{(p+q)/2}}, \qquad 0 < x < \infty,$$

which is (5.3.9).
(b) $\mathrm{E}X = q/(q-2)$ for $q > 2$, and

$$\operatorname{Var} X = \frac{2q^2(p+q-2)}{p(q-2)^2(q-4)} \quad (q > 4).$$

Indeed, negative moments of $V \sim \chi_q^2$ are

$$\mathrm{E}V^{-r} = \frac{1}{\Gamma(\frac{q}{2})2^{q/2}} \int_0^\infty v^{\frac{q}{2}-r-1} e^{-v/2} dv = \frac{\Gamma(\frac{q}{2} - r)}{2^r \Gamma(\frac{q}{2})}, \qquad q > 2r,$$

so $\mathrm{E}V^{-1} = 1/(q-2)$ and $\mathrm{E}V^{-2} = 1/[(q-2)(q-4)]$; also $\mathrm{E}U = p$ and $\mathrm{E}U^2 = p(p+2)$. By independence,

$$\mathrm{E}X = \frac{q}{p}\,\mathrm{E}U\,\mathrm{E}V^{-1} = \frac{q}{q-2},$$
$$\mathrm{E}X^2 = \frac{q^2}{p^2}\,\mathrm{E}U^2\,\mathrm{E}V^{-2} = \frac{q^2(p+2)}{p(q-2)(q-4)}.$$

Subtracting,

$$\mathrm{Var}\,X = \frac{q^2\big[(p+2)(q-2) - p(q-4)\big]}{p(q-2)^2(q-4)} = \frac{2q^2(p+q-2)}{p(q-2)^2(q-4)}.$$

(c) $1/X = (V/q)/(U/p)$ is a ratio of independent scaled chi squareds with q and p degrees of freedom, so $1/X \sim F_{q,p}$ by Definition 5.3.6. This is Theorem 5.3.8(a).
(d) Since $(p/q)X = U/V$,

$$\frac{(p/q)X}{1+(p/q)X} = \frac{U/V}{1+U/V} = \frac{U}{U+V}.$$

Here $U/2 \sim$ gamma$(p/2, 1)$ and $V/2 \sim$ gamma$(q/2, 1)$ are independent, and $U/(U+V) = (U/2)/[(U/2)+(V/2)]$, so Exercise 4.24 gives beta$(p/2, q/2)$, which is Theorem 5.3.8(c).

Problem 5.18. Let X be a random variable with a Student's t distribution with p degrees of freedom.
(a) Derive the mean and variance of X.
(b) Show that X^2 has an F distribution with 1 and p degrees of freedom.
(c) Let $f(x|p)$ denote the pdf of X. Show that

$$\lim_{p\to\infty} f(x|p) \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

at each value of x, $-\infty < x < \infty$. This correctly suggests that as $p \to \infty$, X converges in distribution to a $n(0,1)$ random variable. (Hint: Use Stirling's Formula.)
(d) Use the results of parts (a) and (b) to argue that, as $p \to \infty$, X^2 converges in distribution to a χ^2_1 random variable.
(e) What might you conjecture about the distributional limit, as $p \to \infty$, of $qF_{q,p}$?

Solution. Write $X = Z/\sqrt{V/p}$ with $Z \sim n(0,1)$, $V \sim \chi^2_p$ independent, as in (5.3.5).
(a) $\mathrm{E}X = 0$ if $p > 1$ and $\mathrm{Var}\,X = p/(p-2)$ if $p > 2$, which is (5.3.7). By independence $\mathrm{E}X = \mathrm{E}Z \cdot p^{1/2}\mathrm{E}V^{-1/2}$, and $\mathrm{E}V^{-1/2} = \Gamma(\frac{p-1}{2})/[\sqrt{2}\,\Gamma(\frac{p}{2})] < \infty$ exactly when $p > 1$ (the negative-moment formula of Exercise 5.17(b)); the factor $\mathrm{E}Z = 0$ then gives the mean. Likewise

$$\mathrm{Var}\,X = \mathrm{E}X^2 = \mathrm{E}Z^2 \cdot p\,\mathrm{E}V^{-1} = 1 \cdot \frac{p}{p-2}, \qquad p > 2.$$

(b) $Z^2 \sim \chi^2_1$ by Lemma 5.3.2(a), and it is independent of $V \sim \chi^2_p$, so

$$X^2 = \frac{Z^2/1}{V/p} \sim F_{1,p}$$

by Definition 5.3.6.
(c) From (5.3.6),

$$f(x|p) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma(\frac{p}{2})\sqrt{p}} \cdot \frac{1}{\sqrt{\pi}} \cdot \frac{1}{(1+x^2/p)^{(p+1)/2}}.$$

The last factor tends to $e^{-x^2/2}$, since $(1+x^2/p)^{(p+1)/2} = \big[(1+x^2/p)^p\big]^{1/2}(1+x^2/p)^{1/2} \to (e^{x^2})^{1/2} \cdot 1$.

For the first factor apply Stirling's Formula $\Gamma(a) \approx \sqrt{2\pi}\, a^{a-1/2}e^{-a}$ with $a = \frac{p+1}{2}$ and $a = \frac{p}{2}$:

$$\frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma(\frac{p}{2})} \approx \frac{\left(\frac{p+1}{2}\right)^{p/2} e^{-(p+1)/2}}{\left(\frac{p}{2}\right)^{(p-1)/2} e^{-p/2}}$$
$$= \left(\frac{p+1}{p}\right)^{p/2} \left(\frac{p}{2}\right)^{1/2} e^{-1/2} \;\sim\; e^{1/2}e^{-1/2}\sqrt{p/2} = \sqrt{p/2},$$

using $(1+1/p)^{p/2} \to e^{1/2}$; so $\Gamma(\frac{p+1}{2})/[\Gamma(\frac{p}{2})\sqrt{p}] \to 1/\sqrt{2}$. Multiplying the three limits,

$$f(x|p) \longrightarrow \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{\pi}} \cdot e^{-x^2/2} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

(d) $X_p^2 \to \chi_1^2$ in distribution, equivalently $F_{1,p} \to \chi_1^2$. (The argument uses (b) and (c); (a) only corroborates it, since $\mathrm{E}X_p^2 = p/(p-2) \to 1 = \mathrm{E}\chi_1^2$.) Write $X_p \sim t_p$: part (c) gives $X_p \to Z \sim n(0,1)$ in distribution, and $\pm\sqrt{t}$ are continuity points of Φ, so for every $t > 0$

$$P(X_p^2 \le t) = P(-\sqrt{t} \le X_p \le \sqrt{t}) \longrightarrow \Phi(\sqrt{t}) - \Phi(-\sqrt{t}) = P(Z^2 \le t),$$

while $P(X_p^2 \le t) = 0 = P(Z^2 \le t)$ for $t < 0$. Here $X_p^2 \sim F_{1,p}$ by (b) and $Z^2 \sim \chi_1^2$ by Lemma 5.3.2(a).
(e) $qF_{q,p} \to \chi_q^2$ in distribution as $p \to \infty$, part (d) being the case $q = 1$. Fix $W \sim \chi_q^2$ and take $V_p \sim \chi_p^2$ independent of W; then $qF_{q,p} \overset{d}{=} W/(V_p/p)$. Since V_p is a sum of p iid χ_1^2 variables of mean 1, the weak law (Theorem 5.5.2) gives $V_p/p \to 1$ in probability, hence $p/V_p \to 1$ in probability (Theorem 5.5.4, $h(u) = 1/u$ continuous at 1). Slutsky's Theorem 5.5.17 applied to the constant sequence W then gives $W \cdot (p/V_p) \to W$ in distribution.

Problem 5.19. (a) Prove that the χ^2 distribution is *stochastically increasing* in its degrees of freedom; that is, if $p > q$, then for any a, $P(\chi_p^2 > a) \ge P(\chi_q^2 > a)$, with strict inequality for some a.
(b) Use the results of part (a) to prove that for any ν, $kF_{k,\nu}$ is stochastically increasing in k.
(c) Show that for any k, ν, and α, $kF_{\alpha,k,\nu} > (k-1)F_{\alpha,k-1,\nu}$. (The notation $F_{\alpha,k-1,\nu}$ denotes a level-α *cutoff point*; see Section 8.3.1. Also see Miscellanea 8.5.1 and Exercise 11.15.)

Solution. (a) Couple: let $W \sim \chi_q^2$ and $Y \sim \chi_{p-q}^2$ be independent, so $W + Y \sim \chi_p^2$ by Lemma 5.3.2(b) ($p > q$, and $p - q \ge 1$ since degrees of freedom are integers). Since $Y > 0$ with probability 1, $W + Y > W$ pointwise, hence for every a

$$P(\chi_p^2 > a) = P(W + Y > a) \ge P(W > a) = P(\chi_q^2 > a).$$

The gap is exact: for every $a > 0$,

$$P(W + Y > a) - P(W > a) = P(W \le a,\ W + Y > a) \ge P(W \le \tfrac{a}{2})P(Y > a) > 0,$$

both factors being positive because W and Y are independent with support $(0, \infty)$. So the inequality is strict at every $a > 0$.
(b) Fix ν and $k > k'$ (integers). Write $kF_{k,\nu} = \chi_k^2/(V/\nu)$ with $V \sim \chi_\nu^2$ independent of the numerator, and use the same V for both. Conditioning on V and applying (a) pointwise, for any $a > 0$ and any $v > 0$,

$$P\left(\chi_k^2 > \frac{av}{\nu}\right) > P\left(\chi_{k'}^2 > \frac{av}{\nu}\right),$$

strictly, since $av/\nu > 0$. Taking expectations over V,

$$P(kF_{k,\nu} > a) = \mathrm{E}\, P\left(\chi_k^2 > \tfrac{aV}{\nu} \,\middle|\, V\right) > \mathrm{E}\, P\left(\chi_{k'}^2 > \tfrac{aV}{\nu} \,\middle|\, V\right) = P(k'F_{k',\nu} > a),$$

so $kF_{k,\nu}$ is stochastically increasing in k.

(c) The cutoff points are defined by $P(F_{k,\nu} > F_{\alpha,k,\nu}) = \alpha$, equivalently

$$P\big(kF_{k,\nu} > kF_{\alpha,k,\nu}\big) = \alpha,$$
$$P\big((k-1)F_{k-1,\nu} > (k-1)F_{\alpha,k-1,\nu}\big) = \alpha.$$

Put $a^* = (k-1)F_{\alpha,k-1,\nu} > 0$. By (b) with $k' = k-1$,

$$P\big(kF_{k,\nu} > a^*\big) > P\big((k-1)F_{k-1,\nu} > a^*\big) = \alpha.$$

The map $t \mapsto P(kF_{k,\nu} > t)$ is continuous and strictly decreasing on $(0,\infty)$ (the F pdf of (5.3.9) is positive there), and it takes the value α at $t = kF_{\alpha,k,\nu}$. Since it exceeds α at $t = a^*$, we must have

$$kF_{\alpha,k,\nu} > a^* = (k-1)F_{\alpha,k-1,\nu}.$$

Problem 5.20. (a) We can see that the t distribution is a mixture of normals using the following argument:

$$P(T_\nu \le t) = P\left(\frac{Z}{\sqrt{\chi^2_\nu/\nu}} \le t\right) = \int_0^\infty P\big(Z \le t\sqrt{x}/\sqrt{\nu}\big)\, P(\chi^2_\nu = x)\, dx,$$

where T_ν is a t random variable with ν degrees of freedom. Using the Fundamental Theorem of Calculus and interpreting $P(\chi^2_\nu = \nu x)$ as a pdf, we obtain

$$f_{T_\nu}(t) = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2 x/(2\nu)} \frac{\sqrt{x}}{\sqrt{\nu}} \frac{1}{\Gamma(\nu/2)2^{\nu/2}}\, x^{(\nu/2)-1} e^{-x/2}\, dx,$$

a scale mixture of normals. Verify this formula by direct integration.
(b) A similar formula holds for the F distribution; that is, it can be written as a mixture of chi squareds. If $F_{1,\nu}$ is an F random variable with 1 and ν degrees of freedom, then we can write

$$P(F_{1,\nu} \le \nu t) = \int_0^\infty P(\chi^2_1 \le ty) f_\nu(y)\, dy,$$

where $f_\nu(y)$ is a χ^2_ν pdf. Use the Fundamental Theorem of Calculus to obtain an integral expression for the pdf of $F_{1,\nu}$, and show that the integral equals the pdf.
(c) Verify that the generalization of part (b),

$$P\Big(F_{m,\nu} \le \frac{\nu}{m} t\Big) = \int_0^\infty P(\chi^2_m \le ty) f_\nu(y)\, dy,$$

is valid for all integers $m > 1$.

Solution. (a) Collect the constants and recognize a gamma kernel in x:

$$\begin{aligned} f_{T_\nu}(t) &= \frac{1}{\sqrt{2\pi\nu}\,\Gamma(\frac{\nu}{2})2^{\nu/2}} \int_0^\infty x^{\frac{\nu+1}{2}-1} e^{-\frac{x}{2}\left(1+\frac{t^2}{\nu}\right)} dx \\ &= \frac{\Gamma\big(\frac{\nu+1}{2}\big)}{\sqrt{2\pi\nu}\,\Gamma(\frac{\nu}{2})2^{\nu/2}} \left[\frac{2}{1+t^2/\nu}\right]^{(\nu+1)/2} \\ &= \frac{\Gamma\big(\frac{\nu+1}{2}\big)}{\Gamma(\frac{\nu}{2})} \frac{1}{(\nu\pi)^{1/2}} \frac{1}{(1+t^2/\nu)^{(\nu+1)/2}}, \end{aligned}$$

using $2^{(\nu+1)/2}/\big(2^{\nu/2}\sqrt{2}\big) = 1$ and $\sqrt{2\pi\nu} = \sqrt{2}\sqrt{\pi\nu}$. This is (5.3.6).
(b) Differentiate both sides in t; the integrand is dominated locally, so differentiation passes inside. The left side gives $\nu f_{F_{1,\nu}}(\nu t)$ and the right side $\int_0^\infty y\, f_{\chi^2_1}(ty) f_\nu(y)\, dy$, whence

$$f_{F_{1,\nu}}(\nu t) = \frac{1}{\nu} \int_0^\infty y\, f_{\chi^2_1}(ty)\, f_\nu(y)\, dy.$$

Evaluate the integral with $f_{\chi^2_1}(u) = u^{-1/2}e^{-u/2}/\sqrt{2\pi}$:

$$\int_0^\infty y\, f_{\chi^2_1}(ty) f_\nu(y)\, dy = \frac{t^{-1/2}}{\sqrt{2\pi}\,\Gamma(\frac{\nu}{2})2^{\nu/2}} \int_0^\infty y^{\frac{\nu+1}{2}-1} e^{-\frac{y}{2}(1+t)} dy$$
$$= \frac{\Gamma\big(\frac{\nu+1}{2}\big)}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})} \frac{t^{-1/2}}{(1+t)^{(\nu+1)/2}},$$

so

$$f_{F_{1,\nu}}(\nu t) = \frac{\Gamma\big(\frac{\nu+1}{2}\big)}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})} \cdot \frac{1}{\nu} \cdot \frac{t^{-1/2}}{(1+t)^{(\nu+1)/2}}.$$

Setting $p = 1$, $q = \nu$, $x = \nu t$ in (5.3.9) gives

$$f_{F_{1,\nu}}(\nu t) = \frac{\Gamma\big(\frac{1+\nu}{2}\big)}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})} \left(\frac{1}{\nu}\right)^{1/2} \frac{(\nu t)^{-1/2}}{(1+t)^{(1+\nu)/2}},$$

the same expression.
(c) Immediate from Definition 5.3.6 and does not need $m = 1$. Write $F_{m,\nu} = (\chi^2_m/m)/(V/\nu)$ with $V \sim \chi^2_\nu$ independent of χ^2_m. Then

$$P\Big(F_{m,\nu} \le \frac{\nu}{m} t\Big) = P\left(\frac{\chi^2_m/m}{V/\nu} \le \frac{\nu t}{m}\right) = P(\chi^2_m \le tV),$$

and conditioning on $V = y$ (independence) gives

$$P(\chi^2_m \le tV) = \int_0^\infty P(\chi^2_m \le ty)\, f_\nu(y)\, dy.$$

Differentiating in t as in (b) confirms the pdf: the right side yields

$$\int_0^\infty y\, f_{\chi^2_m}(ty) f_\nu(y)\, dy = \frac{\Gamma\big(\frac{m+\nu}{2}\big)}{\Gamma(\frac{m}{2})\Gamma(\frac{\nu}{2})} \frac{t^{m/2-1}}{(1+t)^{(m+\nu)/2}},$$

while the left side yields $\frac{\nu}{m} f_{F_{m,\nu}}\big(\frac{\nu}{m}t\big)$, which by (5.3.9) is the same.

Problem 5.21. What is the probability that the larger of two continuous iid random variables will exceed the population median? Generalize this result to samples of size n.

Solution. $3/4$; for a sample of size n it is $1 - 2^{-n}$.
Let m be the population median, so $P(X_i \le m) = \frac{1}{2}$ for a continuous population. The largest order statistic fails to exceed m exactly when every observation fails to, so by independence

$$P(X_{(n)} > m) = 1 - P(X_1 \le m, \ldots, X_n \le m) = 1 - \big(\tfrac{1}{2}\big)^n,$$

which is $1 - (1/2)^2 = 3/4$ at $n = 2$.

5.4 Exercises 5.22–5.28

Problem 5.22. Let X and Y be iid $\mathrm{n}(0,1)$ random variables, and define $Z = \min(X, Y)$. Prove that $Z^2 \sim \chi^2_1$.

Solution. The cdf of Z^2 is $2\Phi(\sqrt{z}\,) - 1$, which is the χ^2_1 cdf.
Write Φ for the $\mathrm{n}(0,1)$ cdf and fix $z > 0$, $a = \sqrt{z}$. Since $Z \le t$ fails exactly when both $X > t$ and

$Y > t$, independence gives $P(Z \le t) = 1 - [1 - \Phi(t)]^2$, and Z is continuous, so

$$\begin{aligned} P(Z^2 \le z) &= P(-a \le Z \le a) \\ &= \big\{1 - [1 - \Phi(a)]^2\big\} - \big\{1 - [1 - \Phi(-a)]^2\big\} \\ &= [\Phi(a)]^2 - [1 - \Phi(a)]^2, \end{aligned}$$

using $1 - \Phi(-a) = \Phi(a)$. Factoring the difference of squares,

$$[\Phi(a)]^2 - [1 - \Phi(a)]^2 = \big(2\Phi(a) - 1\big) \cdot 1 = P(|X| \le a) = P(X^2 \le z).$$

By Example 2.1.9 (or Lemma 5.3.2), $X^2 \sim \chi^2_1$; hence Z^2 and X^2 have the same cdf and $Z^2 \sim \chi^2_1$.

Problem 5.23. Let $U_i,\ i = 1, 2, \ldots$, be independent uniform$(0, 1)$ random variables, and let X have distribution

$$P(X = x) = \frac{c}{x!}, \qquad x = 1, 2, 3, \ldots,$$

where $c = 1/(e - 1)$. Find the distribution of

$$Z = \min\{U_1, \ldots, U_X\}.$$

(Hint: Note that the distribution of $Z \mid X = x$ is that of the first-order statistic from a sample of size x.)

Solution. Z has pdf $f_Z(z) = e^{1-z}/(e - 1)$ on $0 < z < 1$: a standard exponential truncated to $(0, 1)$. Conditionally on $X = x$, Z is the first order statistic of x iid uniform$(0, 1)$ variables, so $P(Z > z \mid X = x) = (1 - z)^x$ for $0 < z < 1$. Averaging over X,

$$\begin{aligned} P(Z > z) &= \sum_{x=1}^{\infty} (1 - z)^x \, \frac{c}{x!} \\ &= c\Big(\sum_{x=0}^{\infty} \frac{(1 - z)^x}{x!} - 1\Big) \\ &= c\big(e^{1-z} - 1\big), \qquad 0 < z < 1. \end{aligned}$$

Hence $F_Z(z) = 1 - c(e^{1-z} - 1)$ on $(0, 1)$, with $F_Z(0) = 0$ and $F_Z(1) = 1$ (Check!), and

$$f_Z(z) = \frac{d}{dz} F_Z(z) = c\, e^{1-z} = \frac{e^{1-z}}{e - 1}, \qquad 0 < z < 1.$$

Problem 5.24. Let $X_1, \ldots, X_n$ be a random sample from a population with pdf

$$f_X(x) = \begin{cases} 1/\theta & \text{if } 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics. Show that $X_{(1)}/X_{(n)}$ and $X_{(n)}$ are independent random variables.

Solution. The joint pdf of $(X_{(1)}/X_{(n)}, X_{(n)})$ factors, so they are independent.
By (5.4.7) with $i = 1$, $j = n$, using $F_X(x) = x/\theta$ on $(0, \theta)$,

$$f_{X_{(1)}, X_{(n)}}(u, v) = \frac{n(n - 1)(v - u)^{n-2}}{\theta^n}, \qquad 0 < u < v < \theta.$$

Put $T = X_{(1)}/X_{(n)}$, $W = X_{(n)}$, so $u = tw$, $v = w$; the transformation is one-to-one from $\{0 < u < v <$

$\theta\}$ onto $\{0 < t < 1,\ 0 < w < \theta\}$ with Jacobian

$$\left|\det \begin{pmatrix} w & t \\ 0 & 1 \end{pmatrix}\right| = w.$$

By the bivariate change-of-variables formula (4.3.2),

$$\begin{aligned} f_{T,W}(t,w) &= \frac{n(n-1)(w-tw)^{n-2}}{\theta^n} \cdot w \\ &= \underbrace{(n-1)(1-t)^{n-2}}_{0<t<1} \cdot \underbrace{\frac{n\,w^{n-1}}{\theta^n}}_{0<w<\theta}. \end{aligned}$$

The joint pdf is a product of a function of t alone and a function of w alone on a rectangular support, so by Lemma 4.2.7 T and W are independent, $T \sim \text{beta}(1, n-1)$ and $W = X_{(n)}$ has pdf nw^{n-1}/θ^n.

Problem 5.25. As a generalization of the previous exercise, let $X_1, \ldots, X_n$ be iid with pdf

$$f_X(x) = \begin{cases} \dfrac{a}{\theta^a} x^{a-1} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics. Show that $X_{(1)}/X_{(2)},\ X_{(2)}/X_{(3)}, \ldots, X_{(n-1)}/X_{(n)}$, and $X_{(n)}$ are mutually independent random variables. Find the distribution of each of them.

Solution. The joint pdf of the ratios factors completely: $Y_j = X_{(j)}/X_{(j+1)} \sim \text{beta}(ja, 1)$ for $j = 1, \ldots, n-1$, and $Y_n = X_{(n)}$ has pdf $na\, y^{na-1}/\theta^{na}$ on $(0, \theta)$.
By the joint pdf of all n order statistics (the display following Theorem 5.4.6),

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n!\Big(\frac{a}{\theta^a}\Big)^n \prod_{i=1}^n x_i^{a-1}, \qquad 0 < x_1 < \cdots < x_n < \theta.$$

Set $Y_j = X_{(j)}/X_{(j+1)}$ for $j < n$ and $Y_n = X_{(n)}$; the inverse is $x_i = \prod_{j \ge i} y_j$, a one-to-one map onto $\{0 < y_j < 1\ (j < n),\ 0 < y_n < \theta\}$. The matrix $(\partial x_i/\partial y_j)$ is upper triangular ($\partial x_i/\partial y_j = 0$ for $j < i$), so

$$|J| = \prod_{i=1}^n \frac{\partial x_i}{\partial y_i} = \prod_{i=1}^n \prod_{j>i} y_j = \prod_{j=1}^n y_j^{j-1},$$

since y_j occurs once for each $i < j$. Likewise y_j occurs in x_i for each $i \le j$, so

$$\prod_{i=1}^n x_i^{a-1} = \prod_{j=1}^n y_j^{j(a-1)}.$$

Therefore, by the n-variate change-of-variables formula (4.6.7),

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = n!\Big(\frac{a}{\theta^a}\Big)^n \prod_{j=1}^n y_j^{ja-1}$$

on the rectangle above. The pdf is a product of functions of the individual y_j over a rectangular support, so by Theorem 4.6.11 $Y_1, \ldots, Y_n$ are mutually independent, with

$$\begin{aligned} f_{Y_j}(y) &= ja\, y^{ja-1}, \quad 0 < y < 1, \quad j = 1, \ldots, n-1, \\ f_{Y_n}(y) &= \frac{na}{\theta^{na}}\, y^{na-1}, \quad 0 < y < \theta. \end{aligned}$$

That is, $X_{(j)}/X_{(j+1)} \sim \text{beta}(ja, 1)$ for $j = 1, \ldots, n-1$, and $X_{(n)}/\theta \sim \text{beta}(na, 1)$.

Problem 5.26. Complete the proof of Theorem 5.4.6. (Theorem 5.4.6: if $X_{(1)}, \ldots, X_{(n)}$ are the order statistics of a random sample from a continuous population with cdf F_X and pdf f_X, then for $1 \le i < j \le n$ the joint pdf of $X_{(i)}$ and $X_{(j)}$ is

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!\,(j-1-i)!\,(n-j)!} f_X(u) f_X(v) [F_X(u)]^{i-1} \times [F_X(v) - F_X(u)]^{j-1-i} [1 - F_X(v)]^{n-j}$$

for $-\infty < u < v < \infty$.)
(a) Let U be a random variable that counts the number of $X_1, \ldots, X_n$ less than or equal to u, and let V be a random variable that counts the number of $X_1, \ldots, X_n$ greater than u and less than or equal to v. Show that $(U, V, n-U-V)$ is a multinomial random vector with n trials and cell probabilities $(F_X(u),\ F_X(v) - F_X(u),\ 1 - F_X(v))$.
(b) Show that the joint cdf of $X_{(i)}$ and $X_{(j)}$ can be expressed as

$$\begin{aligned} F_{X_{(i)},X_{(j)}}(u,v) &= P(U \ge i,\ U+V \ge j) \\ &= \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} P(U=k, V=m) + P(U \ge j) \\ &= \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \frac{n!}{k!\,m!\,(n-k-m)!} [F_X(u)]^k \\ &\quad \times [F_X(v) - F_X(u)]^m [1 - F_X(v)]^{n-k-m} \\ &\quad + P(U \ge j). \end{aligned}$$

(c) Find the joint pdf by computing the mixed partial as indicated in (4.1.4). (The mixed partial of $P(U \ge j)$ is 0 since this term depends only on u, not v. For the other terms, there is much cancellation using relationships like (5.4.6).)

Solution. Fix $u < v$ throughout, and write $p = F_X(u)$, $q = F_X(v) - F_X(u)$, $r = 1 - F_X(v)$.
(a) Each X_l falls into exactly one of the three cells $(-\infty, u]$, $(u, v]$, (v, ∞), with probabilities p, q, r respectively, and the X_l are iid. Thus the n classifications are independent identical three-cell trials, and by Definition 4.6.2 the cell counts $(U, V, n-U-V)$ are multinomial with n trials and cell probabilities (p, q, r).
(b) Since exactly U of the sample lie in $(-\infty, u]$ and exactly $U+V$ lie in $(-\infty, v]$,

$$\{X_{(i)} \le u\} = \{U \ge i\}, \qquad \{X_{(j)} \le v\} = \{U+V \ge j\},$$

giving the first line. Split on the value of U: if $U = k \ge j$ then $U+V \ge j$ automatically, while for $i \le k \le j-1$ we need $V \ge j-k$, and $V \le n-k$ always. Hence

$$F_{X_{(i)},X_{(j)}}(u,v) = \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} P(U=k, V=m) + P(U \ge j),$$

and substituting the multinomial probabilities from part (a) gives the displayed triple-product form.
(c) By (4.1.4), $f_{X_{(i)},X_{(j)}}(u,v) = \partial^2 F_{X_{(i)},X_{(j)}}(u,v)/\partial u\,\partial v$. The term $P(U \ge j)$ is a function of u only, so it dies under $\partial/\partial v$.
Differentiate in v first. Note $q + r = 1 - p$ does not depend on v, and $\partial q/\partial v = -\partial r/\partial v = f_X(v)$. Writing the inner sum with $N = n-k$, $M = j-k$,

$$\frac{n!}{k!m!(n-k-m)!} p^k q^m r^{n-k-m} = \binom{n}{k} p^k \binom{N}{m} q^m r^{N-m},$$

and the upper-tail binomial derivative (the same cancellation as in (5.4.6)) gives

$$\frac{\partial}{\partial q}\sum_{m=M}^{N}\binom{N}{m}q^m(s-q)^{N-m}=N\binom{N-1}{M-1}q^{M-1}(s-q)^{N-M},$$

valid with $s=q+r$ held fixed, since all but one term of the telescoping sum cancels. With $s-q=r=1-F_X(v)$ and $N-M=n-j$,

$$\begin{aligned}\frac{\partial F}{\partial v}&=f_X(v)\,r^{\,n-j}\sum_{k=i}^{j-1}\binom{n}{k}(n-k)\binom{n-k-1}{j-k-1}p^k q^{\,j-k-1}\\&=f_X(v)\,r^{\,n-j}\sum_{k=i}^{j-1}\frac{n!}{k!\,(j-k-1)!\,(n-j)!}\,p^k q^{\,j-1-k},\end{aligned}$$

since $\binom{n}{k}(n-k)\binom{n-k-1}{j-k-1}=\dfrac{n!}{k!(j-k-1)!(n-j)!}$ (Check!).
Now differentiate in u. Here $p+q=F_X(v)$ is free of u and $\partial p/\partial u=-\partial q/\partial u=f_X(u)$. With $L=j-1$,

$$\frac{n!}{k!(L-k)!(n-j)!}=\frac{n!}{L!\,(n-j)!}\binom{L}{k},$$

so the sum above is $\dfrac{n!}{(j-1)!(n-j)!}\sum_{k=i}^{L}\binom{L}{k}p^k q^{L-k}$, and the same derivative identity (with total $p+q=F_X(v)$) yields

$$\frac{\partial}{\partial p}\sum_{k=i}^{L}\binom{L}{k}p^k q^{L-k}=L\binom{L-1}{i-1}p^{\,i-1}q^{\,L-i}.$$

Collecting, and using

$$\frac{n!}{(j-1)!(n-j)!}\,(j-1)\binom{j-2}{i-1}=\frac{n!}{(i-1)!\,(j-1-i)!\,(n-j)!},$$

we obtain

$$\begin{aligned}f_{X_{(i)},X_{(j)}}(u,v)=\frac{n!}{(i-1)!(j-1-i)!(n-j)!}&f_X(u)f_X(v)[F_X(u)]^{i-1}\\&\times[F_X(v)-F_X(u)]^{j-1-i}[1-F_X(v)]^{n-j},\end{aligned}$$

for $u<v$, which is (5.4.7).

Problem 5.27. Let $X_1,\dots,X_n$ be iid with pdf $f_X(x)$ and cdf $F_X(x)$, and let $X_{(1)}<\cdots<X_{(n)}$ be the order statistics.
(a) Find an expression for the conditional pdf of $X_{(i)}$ given $X_{(j)}$ in terms of f_X and F_X.
(b) Find the pdf of $V\mid R=r$, where V and R are defined in Example 5.4.7. (In that example $X_1,\dots,X_n$ are iid uniform$(0,a)$, $R=X_{(n)}-X_{(1)}$ is the range and $V=(X_{(1)}+X_{(n)})/2$ is the midrange; their joint pdf is $f_{R,V}(r,v)=n(n-1)r^{n-2}/a^n$ on $0<r<a$, $r/2<v<a-r/2$, with marginal $f_R(r)=n(n-1)r^{n-2}(a-r)/a^n$.)

Solution. (a) For $i<j$, divide (5.4.7) by the pdf (5.4.4) of $X_{(j)}$:

$$\begin{aligned}f_{X_{(i)}|X_{(j)}}(u\mid v)=\frac{(j-1)!}{(i-1)!\,(j-1-i)!}&\frac{f_X(u)}{F_X(v)}\\&\times\left[\frac{F_X(u)}{F_X(v)}\right]^{i-1}\left[\frac{F_X(v)-F_X(u)}{F_X(v)}\right]^{j-1-i},\end{aligned}$$

for $u<v$; the factors $f_X(v)$ and $[1-F_X(v)]^{n-j}$ cancel. Comparing with (5.4.4), this is exactly the pdf

of the ith order statistic of a sample of size $j-1$ from the population F_X truncated to $(-\infty, v)$, i.e. with cdf $F_X(x)/F_X(v)$ and pdf $f_X(x)/F_X(v)$ for $x < v$.
For $i > j$ the mirror computation (dividing (5.4.7) with indices $j < i$ by $f_{X_{(j)}}$) gives

$$f_{X_{(i)}|X_{(j)}}(u \mid v) = \frac{(n-j)!}{(i-j-1)!\,(n-i)!}\,\frac{f_X(u)}{1-F_X(v)} \times \left[\frac{F_X(u)-F_X(v)}{1-F_X(v)}\right]^{i-j-1} \times \left[\frac{1-F_X(u)}{1-F_X(v)}\right]^{n-i},$$

for $u > v$: the $(i-j)$th order statistic of a sample of size $n-j$ from F_X truncated to (v, ∞).
(b) Dividing the joint pdf of Example 5.4.7 by $f_R(r)$ in (5.4.8),

$$f_{V|R}(v \mid r) = \frac{n(n-1)r^{n-2}/a^n}{n(n-1)r^{n-2}(a-r)/a^n} = \frac{1}{a-r}, \qquad \frac{r}{2} < v < a - \frac{r}{2}.$$

So $V \mid R = r$ is uniform on the interval $(r/2,\ a - r/2)$, of length $a - r$.

Problem 5.28. Let $X_1, \ldots, X_n$ be iid with pdf $f_X(x)$ and cdf $F_X(x)$, and let $X_{(i_1)} < \cdots < X_{(i_l)}$ and $X_{(j_1)} < \cdots < X_{(j_m)}$ be any two disjoint groups of order statistics. In terms of the pdf $f_X(x)$ and the cdf $F_X(x)$, find expressions for
(a) The marginal cdf and pdf of $X_{(i_1)}, \ldots, X_{(i_l)}$.
(b) The conditional cdf and pdf of $X_{(i_1)}, \ldots, X_{(i_l)}$ given $X_{(j_1)}, \ldots, X_{(j_m)}$.

Solution. (a) The pdf. Integrate the joint pdf $n!\prod_{t=1}^n f_X(x_t)$ of all n order statistics (the display following Theorem 5.4.6) over the unobserved coordinates in each gap; with the conventions $i_0 = 0$, $i_{l+1} = n+1$, $F_X(x_{i_0}) = 0$, $F_X(x_{i_{l+1}}) = 1$,

$$f_{X_{(i_1)},\ldots,X_{(i_l)}}(x_{i_1}, \ldots, x_{i_l}) = n! \prod_{r=1}^{l} f_X(x_{i_r}) \times \prod_{r=1}^{l+1} \frac{\left[F_X(x_{i_r}) - F_X(x_{i_{r-1}})\right]^{i_r - i_{r-1} - 1}}{(i_r - i_{r-1} - 1)!},$$

for $x_{i_1} < \cdots < x_{i_l}$, and 0 otherwise: $i_r - i_{r-1} - 1$ of the remaining observations must land in the rth gap $(x_{i_{r-1}}, x_{i_r})$. For $l = 2$ this is (5.4.7) and for $l = 1$ it is (5.4.4).
The cdf. Evaluate at arbitrary arguments $u_1, \ldots, u_l$ and set $y_r = \min_{r' \ge r} u_{r'}$, so $y_1 \le \cdots \le y_l$ and $\{X_{(i_r)} \le u_r\ \forall r\} = \{X_{(i_r)} \le y_r\ \forall r\}$ because $X_{(i_1)} \le \cdots \le X_{(i_l)}$. Let N_r count the sample values $\le y_r$. Exactly as in Exercise 5.26(b), $\{X_{(i_r)} \le y_r\} = \{N_r \ge i_r\}$, and $(N_1, N_2 - N_1, \ldots, N_l - N_{l-1}, n - N_l)$ is multinomial with n trials and cell probabilities $F_X(y_1)$, $F_X(y_2) - F_X(y_1)$, $\ldots$, $1 - F_X(y_l)$. Hence, with $n_0 = 0$ and $F_X(y_0) = 0$,

$$F_{X_{(i_1)},\ldots,X_{(i_l)}}(u_1, \ldots, u_l) = \sum_{\substack{n_1 \le \cdots \le n_l \le n \\ n_r \ge i_r\ \forall r}} \frac{n!}{(n-n_l)! \prod_{r=1}^{l} (n_r - n_{r-1})!} \times \left[1 - F_X(y_l)\right]^{n-n_l} \prod_{r=1}^{l} \left[F_X(y_r) - F_X(y_{r-1})\right]^{n_r - n_{r-1}}.$$

(b) The pdf. Write $I = \{i_1 < \cdots < i_l\}$, $J = \{j_1 < \cdots < j_m\}$, and divide the part (a) pdf for the pooled group $I \cup J$ by the part (a) pdf for J. Set $j_0 = 0$, $j_{m+1} = n+1$, $x_{j_0} = -\infty$, $x_{j_{m+1}} = +\infty$, and let block t ($t = 0, \ldots, m$) consist of those indices of I lying strictly between j_t and j_{t+1}, say $a_1 < \cdots < a_{s_t}$ (put

$a_0 = j_t$, $a_{s_t+1} = j_{t+1}$). Every factor of the J-pdf cancels, and what survives is

$$f_{I|J} = \prod_{t=0}^{m}\Big\{(j_{t+1}-j_t-1)!\prod_{r=1}^{s_t} f_X(x_{a_r}) \\ \times \prod_{r=1}^{s_t+1}\frac{\big[F_X(x_{a_r})-F_X(x_{a_{r-1}})\big]^{a_r-a_{r-1}-1}}{(a_r-a_{r-1}-1)!}\Big[F_X(x_{j_{t+1}})-F_X(x_{j_t})\Big]^{-(j_{t+1}-j_t-1)}\Big\},$$

supported on $x_{j_t} < x_{a_1} < \cdots < x_{a_{s_t}} < x_{j_{t+1}}$ for every t.
Comparing with part (a), each block factor is exactly the joint pdf of the order statistics of ranks $a_1 - j_t < \cdots < a_{s_t} - j_t$ from a sample of size $N_t = j_{t+1} - j_t - 1$ drawn from F_X truncated to $(x_{j_t}, x_{j_{t+1}})$, i.e. with cdf

$$G_t(x) = \frac{F_X(x) - F_X(x_{j_t})}{F_X(x_{j_{t+1}}) - F_X(x_{j_t})}, \qquad x_{j_t} < x < x_{j_{t+1}},$$

the powers of the truncation constant matching because $s_t + \sum_{r=1}^{s_t+1}(a_r - a_{r-1} - 1) = N_t$. So, given $X_{(j_1)}, \ldots, X_{(j_m)}$, the blocks are mutually independent (the Markov property of the order statistics).
The cdf. By that independence it is the product over blocks of the part (a) cdf for the truncated population: writing u_i for the argument attached to $X_{(i)}$, $i \in I$,

$$F_{I|J}\big((u_i)_{i\in I}\big) = \prod_{t=0}^{m} F^{(N_t,G_t)}\big(u_{a_1}, \ldots, u_{a_{s_t}}\big),$$

where $F^{(N_t,G_t)}$ is the part (a) cdf formula with n replaced by N_t, F_X by G_t, and the ranks a_r by $a_r - j_t$, and where G_t is extended by 0 below x_{j_t} and by 1 above $x_{j_{t+1}}$.

5.5 Exercises 5.29–5.35

Problem 5.29. A manufacturer of booklets packages them in boxes of 100. It is known that, on the average, the booklets weigh 1 ounce, with a standard deviation of .05 ounce. The manufacturer is interested in calculating

$$P(\text{100 booklets weigh more than 100.4 ounces}),$$

a number that would help detect whether too many booklets are being put in a box. Explain how you would calculate the (approximate?) value of this probability. Mention any relevant theorems or assumptions needed.

Solution. Approximately $P(Z > .8) = .2119$.
Write $W = \sum_{i=1}^{100} X_i$ for the weight of a box, with X_i the weight of the ith booklet. Taking the X_i iid with $EX_i = 1$ and $\operatorname{Var} X_i = (.05)^2 = .0025$ gives $EW = 100$ and $\operatorname{sd}(W) = \sqrt{100(.0025)} = .5$, and Theorem 5.5.15 (whose only hypothesis, $0 < \sigma^2 < \infty$, holds here) applies to $(W - 100)/.5 = \sqrt{n}(\bar{X}_n - \mu)/\sigma$:

$$P(W > 100.4) = P\left(\frac{W - 100}{.5} > \frac{100.4 - 100}{.5}\right) \\ \approx P(Z > .8) = .2119.$$

The assumptions, which are worth stating to the manufacturer: (i) the booklets in a box are independent draws – a box filled from one mis-set machine run has positively correlated weights, $\operatorname{Var} W > .25$, and the tail is understated; (ii) 1 and .05 are population moments, not estimates from the box being tested; (iii) $n = 100$ is large enough for the limit in Theorem 5.5.15 to have taken hold at .8 standard deviations. No distributional form is assumed, which is also why the answer can only be approximate: two moments of X_i are given, not the law of X_i.

Problem 5.30. If $\bar{X}_1$ and $\bar{X}_2$ are the means of two independent samples of size n from a population with variance σ^2, find a value for n so that $P(|\bar{X}_1 - \bar{X}_2| < \sigma/5) \approx .99$. Justify your calculations.

Solution. $n = 332$.
Pair the two samples and set $D_i = X_{1i} - X_{2i}$. The D_i are iid with $ED_i = 0$ and $\operatorname{Var} D_i = 2\sigma^2 \in (0, \infty)$, and $\bar{D} = \bar{X}_1 - \bar{X}_2$, so Theorem 5.5.15 applies verbatim:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{2/n}} = \frac{\sqrt{n}\,(\bar{D} - 0)}{\sigma\sqrt{2}} \;\longrightarrow\; \mathrm{n}(0,1).$$

Hence

$$\begin{aligned} .99 &\approx P\Big(|\bar{X}_1 - \bar{X}_2| < \frac{\sigma}{5}\Big) \\ &= P\left(|Z| < \frac{\sigma/5}{\sigma\sqrt{2/n}}\right) = P\left(|Z| < \frac{\sqrt{n}}{5\sqrt{2}}\right). \end{aligned}$$

The .995 standard normal quantile is $z_{.005} = 2.5758$, so

$$\frac{\sqrt{n}}{5\sqrt{2}} = 2.5758 \quad\Longrightarrow\quad n = 50(2.5758)^2 = 331.74.$$

The probability is increasing in n, so take $n = 332$. The only distributional input is the CLT, which needs no normality of the population, only $\sigma^2 < \infty$; and σ cancels, because the requirement is stated in units of σ.

Problem 5.31. Suppose $\bar{X}$ is the mean of 100 observations from a population with mean μ and variance $\sigma^2 = 9$. Find limits between which $\bar{X} - \mu$ will lie with probability at least .90. Use both Chebychev's Inequality and the Central Limit Theorem, and comment on each.

Solution. Chebychev gives $\pm.9487$; the CLT gives $\pm.4935$.
By Theorem 5.2.6, $\operatorname{Var} \bar{X} = \sigma^2/n = 9/100 = .09$.
Via Chebychev: Theorem 3.6.1 applied to $g(x) = (x-\mu)^2$ (its hypothesis is only that $E(\bar{X} - \mu)^2$ exists, which it does) gives, for every $\epsilon > 0$,

$$P(|\bar{X} - \mu| < \epsilon) \;\geq\; 1 - \frac{\operatorname{Var} \bar{X}}{\epsilon^2} = 1 - \frac{.09}{\epsilon^2}.$$

Setting the bound equal to .90,

$$\frac{.09}{\epsilon^2} = .10 \quad\Longrightarrow\quad \epsilon = \sqrt{.9} = .9487,$$

so $P(-.9487 < \bar{X} - \mu < .9487) \geq .90$.
Via the Central Limit Theorem: by Theorem 5.5.15, $(\bar{X} - \mu)/.3$ is approximately $\mathrm{n}(0,1)$, so

$$\begin{aligned} .90 &\approx P\big(|\bar{X} - \mu| < \epsilon\big) = P\Big(|Z| < \frac{\epsilon}{.3}\Big) \\ \Longrightarrow\quad \frac{\epsilon}{.3} &= 1.645, \qquad \epsilon = .4935. \end{aligned}$$

Comment. Chebychev's interval is valid for every population of variance 9 but is only a bound, so its true coverage usually far exceeds .90 (Example 3.6.3: at $t = 2$ Chebychev gives .25 against the normal value .054). The CLT interval is half as wide because it uses the limiting shape rather than the second moment alone, at the cost of being an approximation resting on $n = 100$ being large for the unnamed population.

Problem 5.32. Let $X_1, X_2, \ldots$ be a sequence of random variables that converges in probability to a constant a. Assume that $P(X_i > 0) = 1$ for all i.
(a) Verify that the sequences defined by $Y_i = \sqrt{X_i}$ and $Y_i' = a/X_i$ converge in probability.
(b) Use the results in part (a) to prove the fact used in Example 5.5.18, that σ/S_n converges in probability to 1.

Solution. (a) Both follow from Theorem 5.5.4 (if $X_n \to X$ in probability and h is continuous, then $h(X_n) \to h(X)$ in probability); the only thing to check is that each h is continuous on a set carrying all the X_i and the limit a.
First, $a \geq 0$: if $a < 0$ then $|X_i - a| > |a|$ with probability 1 for every i, so $P(|X_i - a| \geq |a|) = 1$ does not tend to 0, contradicting Definition 5.5.1 at $\epsilon = |a|$.
(i) $h(x) = \sqrt{x}$ is continuous on $[0, \infty) \supseteq \{x : x > 0\} \cup \{a\}$, so

$$Y_i = \sqrt{X_i} \longrightarrow \sqrt{a} \quad \text{in probability.}$$

(ii) For $a > 0$, $h(x) = a/x$ is continuous on $(0, \infty)$, which carries every X_i and the point a, so

$$Y_i' = \frac{a}{X_i} \longrightarrow \frac{a}{a} = 1 \quad \text{in probability.}$$

(If $a = 0$ then $Y_i' = 0$ with probability 1 for every i, and the sequence converges in probability to 0 trivially.)
(b) Apply (a) twice. Example 5.5.3 gives $S_n^2 \to \sigma^2$ in probability under its hypothesis $\operatorname{Var} S_n^2 \to 0$, and $P(S_n^2 \geq 0) = 1$, so (a)(i) with $a = \sigma^2$ gives

$$S_n = \sqrt{S_n^2} \longrightarrow \sigma \quad \text{in probability}$$

(this is Example 5.5.5). Now (a)(ii) applied to $S_1, S_2, \ldots$ with $a = \sigma > 0$ gives

$$\frac{\sigma}{S_n} \longrightarrow 1 \quad \text{in probability,}$$

as required. Only continuity of $x \mapsto \sigma/x$ at the one point $\sigma > 0$ is used, so a discrete population with $P(S_n = 0) > 0$ is no obstacle: picking $\delta \in (0, \sigma]$ with $|s - \sigma| < \delta \Rightarrow |\sigma/s - 1| < \epsilon$ gives

$$P\left(\left|\frac{\sigma}{S_n} - 1\right| \geq \epsilon\right) \leq P(|S_n - \sigma| \geq \delta) \longrightarrow 0,$$

the event $\{S_n = 0\}$ sitting inside the right-hand one since $\sigma \geq \delta$.

Problem 5.33. Let X_n be a sequence of random variables that converges in distribution to a random variable X. Let Y_n be a sequence of random variables with the property that for any finite number c,

$$\lim_{n \to \infty} P(Y_n > c) = 1.$$

Show that for any finite number c,

$$\lim_{n \to \infty} P(X_n + Y_n > c) = 1.$$

(This is the type of result used in the discussion of the power properties of the tests described in Section 10.3.2.)

Solution. The containment

$$\{X_n + Y_n \leq c\} \subseteq \{X_n \leq -b\} \cup \{Y_n \leq c + b\}$$

holds for every real b – if both $X_n > -b$ and $Y_n > c+b$, then $X_n + Y_n > c$ – so Boole's Inequality gives

$$P(X_n + Y_n > c) \;\geq\; 1 - P(X_n \leq -b) - P(Y_n \leq c + b).$$

Fix $\epsilon > 0$. Since F_X is a cdf, $\lim_{b\to\infty} F_X(-b) = 0$, and F_X has at most countably many discontinuities, so we may choose $b > 0$ with $-b$ a continuity point of F_X and $F_X(-b) < \epsilon$. For that b: convergence in distribution gives $P(X_n \leq -b) = F_{X_n}(-b) \to F_X(-b) < \epsilon$ (the continuity-point requirement in the definition of convergence in distribution is exactly what the choice of b secures), and the hypothesis on Y_n applied at the finite number $c + b$ gives $P(Y_n \leq c + b) \to 0$. Therefore

$$\begin{aligned}\liminf_{n\to\infty} P(X_n + Y_n > c) &\geq 1 - \lim_{n\to\infty} P(X_n \leq -b) \\ &\quad - \lim_{n\to\infty} P(Y_n \leq c + b) \\ &= 1 - F_X(-b) \;>\; 1 - \epsilon.\end{aligned}$$

Since $\epsilon > 0$ was arbitrary and each probability is at most 1, $\lim_{n\to\infty} P(X_n + Y_n > c) = 1$.

Problem 5.34. Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. Show that

$$E\,\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = 0 \qquad \text{and} \qquad \operatorname{Var}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = 1.$$

Thus, the normalization of $\bar{X}_n$ in the Central Limit Theorem gives random variables that have the same mean and variance as the limiting $\mathrm{n}(0, 1)$ distribution.

Solution. Both are immediate from Theorem 5.2.6, which gives $E\bar{X}_n = \mu$ and $\operatorname{Var}\bar{X}_n = \sigma^2/n$ (its hypothesis, $\sigma^2 < \infty$, is assumed here):

$$\begin{aligned}E\,\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} &= \frac{\sqrt{n}}{\sigma}\big(E\bar{X}_n - \mu\big) = \frac{\sqrt{n}}{\sigma}(\mu - \mu) = 0, \\ \operatorname{Var}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} &= \frac{n}{\sigma^2}\operatorname{Var}(\bar{X}_n - \mu) \\ &= \frac{n}{\sigma^2}\cdot\frac{\sigma^2}{n} = 1,\end{aligned}$$

using linearity of expectation (Theorem 2.2.5) and $\operatorname{Var}(aW + b) = a^2 \operatorname{Var} W$ (Theorem 2.3.4) with $a = \sqrt{n}/\sigma$, $b = -\sqrt{n}\mu/\sigma$.

Problem 5.35. Stirling's Formula (derived in Exercise 1.28), which gives an approximation for factorials, can be easily derived using the CLT.
(a) Argue that, if $X_i \sim$ exponential(1), $i = 1, 2, \ldots$, all independent, then for every x,

$$P\left(\frac{\bar{X}_n - 1}{1/\sqrt{n}} \leq x\right) \;\to\; P(Z \leq x),$$

where Z is a standard normal random variable.
(b) Show that differentiating both sides of the approximation in part (a) suggests

$$\frac{\sqrt{n}}{\Gamma(n)}(x\sqrt{n} + n)^{n-1} e^{-(x\sqrt{n}+n)} \;\approx\; \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

and that $x = 0$ gives Stirling's Formula.

Solution. (a) This is Theorem 5.5.14 verbatim. The X_i are iid with $EX_i = 1$ and $\operatorname{Var} X_i = 1$ (Section 3.3), and the exponential(1) mgf $M_X(t) = (1 - t)^{-1}$ exists for $|t| < 1$, hence in a neighborhood of 0 –

the one hypothesis Theorem 5.5.14 adds beyond finite variance. So G_n, the cdf of

$$\frac{\sqrt{n}(\bar{X}_n - 1)}{\sigma} = \frac{\bar{X}_n - 1}{1/\sqrt{n}},$$

satisfies $G_n(x) \to \int_{-\infty}^{x} (2\pi)^{-1/2} e^{-y^2/2}\, dy = P(Z \le x)$ for every x.
(b) By Example 4.6.8 with $\alpha_i = \beta = 1$, $\sum_{i=1}^{n} X_i \sim \text{gamma}(n, 1)$, so for n large enough that $n + x\sqrt{n} > 0$,

$$\begin{aligned} P\left(\frac{\bar{X}_n - 1}{1/\sqrt{n}} \le x\right) &= P\left(\sum_{i=1}^{n} X_i \le n + x\sqrt{n}\right) \\ &= \int_0^{n+x\sqrt{n}} \frac{1}{\Gamma(n)} t^{n-1} e^{-t}\, dt. \end{aligned}$$

Part (a) says this is approximately $\int_{-\infty}^{x} (2\pi)^{-1/2} e^{-y^2/2}\, dy$. Differentiating each side in x – on the left by the Fundamental Theorem of Calculus with the chain rule, the upper limit $n + x\sqrt{n}$ contributing the factor $\sqrt{n}$ – gives

$$\frac{\sqrt{n}}{\Gamma(n)} (x\sqrt{n} + n)^{n-1} e^{-(x\sqrt{n}+n)} \;\approx\; \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

which is the stated approximation – only suggested, since convergence of cdfs does not by itself force convergence of the densities.
Setting $x = 0$:

$$\frac{\sqrt{n}}{\Gamma(n)}\, n^{n-1} e^{-n} \approx \frac{1}{\sqrt{2\pi}} \quad \Longleftrightarrow \quad \Gamma(n) \approx \sqrt{2\pi}\, n^{n-\frac{1}{2}} e^{-n}.$$

Since $n! = n\,\Gamma(n)$,

$$n! \;\approx\; \sqrt{2\pi}\, n^{n+\frac{1}{2}} e^{-n},$$

which is Stirling's Formula as stated in Exercise 1.28.

5.6 Exercises 5.36–5.42

Problem 5.36. Given that $N = n$, the conditional distribution of Y is χ^2_{2n}. The unconditional distribution of N is Poisson(θ).
(a) Calculate $\mathrm{E}Y$ and $\operatorname{Var} Y$ (unconditional moments).
(b) Show that, as $\theta \to \infty$, $(Y - \mathrm{E}Y)/\sqrt{\operatorname{Var} Y} \to \mathrm{n}(0, 1)$ in distribution.

Solution. (a) $\mathrm{E}Y = 2\theta$ and $\operatorname{Var} Y = 8\theta$.
A χ^2_{2N} variable has conditional mean $2N$ and conditional variance $4N$, so by Theorems 4.4.3 and 4.4.7,

$$\begin{aligned} \mathrm{E}Y &= \mathrm{E}\big[\mathrm{E}(Y \mid N)\big] = \mathrm{E}(2N) = 2\theta, \\ \operatorname{Var} Y &= \mathrm{E}\big[\operatorname{Var}(Y \mid N)\big] + \operatorname{Var}\big[\mathrm{E}(Y \mid N)\big] \\ &= \mathrm{E}(4N) + \operatorname{Var}(2N) = 4\theta + 4\theta = 8\theta, \end{aligned}$$

using $\mathrm{E}N = \operatorname{Var} N = \theta$.
(b) Condition, then use the Poisson mgf. For $t < 1/2$, the χ^2_{2n} mgf is $(1-2t)^{-n}$, so with $e^s = (1-2t)^{-1}$,

$$M_Y(t) = \mathrm{E}\big[(1-2t)^{-N}\big] = M_N\big(-\log(1-2t)\big) = \exp\Big\{\theta\Big(\tfrac{1}{1-2t} - 1\Big)\Big\},$$

that is, $M_Y(t) = \exp\{2\theta t/(1-2t)\}$, finite for $t < 1/2$.
Put $Z_\theta = (Y - 2\theta)/\sqrt{8\theta}$ and, for fixed t, write $s = t/\sqrt{8\theta}$ (so $s < 1/2$ once θ is large). Then

$$\begin{aligned} \log M_{Z_\theta}(t) &= -\frac{2\theta t}{\sqrt{8\theta}} + \frac{2\theta s}{1-2s} = 2\theta s\Big(\frac{1}{1-2s} - 1\Big) \\ &= \frac{4\theta s^2}{1-2s} = \frac{t^2/2}{1 - 2t/\sqrt{8\theta}} \;\longrightarrow\; \frac{t^2}{2} \end{aligned}$$

as $\theta \to \infty$, since $4\theta s^2 = 4\theta t^2/(8\theta) = t^2/2$. Thus $M_{Z_\theta}(t) \to e^{t^2/2}$ for every t in a neighborhood of 0, and by Theorem 2.3.12, $Z_\theta \to \mathrm{n}(0,1)$ in distribution.

Problem 5.37. In Example 5.5.16, a normal approximation to the negative binomial distribution was given. Just as with the normal approximation to the binomial distribution given in Example 3.3.2, the approximation might be improved with a "continuity correction."
For X_is defined as in Example 5.5.16 (a random sample from negative binomial(r,p), so that $\mathrm{E}X = r(1-p)/p$ and $\mathrm{Var}\,X = r(1-p)/p^2$), let $V_n = \sum_{i=1}^n X_i$. For $n = 10$, $p = .7$, and $r = 2$, calculate $P(V_n = v)$ for $v = 0, 1, \ldots, 10$ using each of the following three methods.
(a) exact calculations
(b) normal approximation as given in Example 5.5.16
(c) normal approximation with continuity correction

Solution. The continuity correction is closer to the exact value at every $v \geq 3$, roughly halving the error; in the extreme lower tail ($v \leq 2$) it overshoots and the uncorrected version is better.
As noted inside Example 5.5.16 (multiply the negative binomial mgfs, Theorem 4.6.7), $V_n = \sum_{i=1}^{10} X_i \sim$ negative binomial(nr,p) = negative binomial$(20,.7)$, so the exact answer is

$$P(V_n = v) = \binom{19+v}{v}(.7)^{20}(.3)^v, \qquad v = 0, 1, 2, \ldots$$

The CLT centering and scaling of Example 5.5.16 give

$$\mu = \frac{nr(1-p)}{p} = \frac{20(.3)}{.7} = \frac{60}{7} = 8.5714, \qquad \sigma^2 = \frac{nr(1-p)}{p^2} = \frac{20(.3)}{.49} = 12.2449,$$

so $\sigma = 3.4993$. The two approximations differ only in where the normal cdf is evaluated:

$$\text{(b)} \quad P(V_n = v) \approx \Phi\Big(\frac{v-\mu}{\sigma}\Big) - \Phi\Big(\frac{v-1-\mu}{\sigma}\Big),$$

$$\text{(c)} \quad P(V_n = v) \approx \Phi\Big(\frac{v+\frac{1}{2}-\mu}{\sigma}\Big) - \Phi\Big(\frac{v-\frac{1}{2}-\mu}{\sigma}\Big).$$

v	(a) exact	(b) normal	(c) normal with c.c.
0	.0008	.0040	.0058
1	.0048	.0081	.0111
2	.0151	.0150	.0197
3	.0332	.0255	.0323
4	.0572	.0400	.0487
5	.0824	.0580	.0677
6	.1030	.0775	.0869
7	.1148	.0955	.1028
8	.1162	.1085	.1121
9	.1085	.1136	.1128
10	.0944	.1097	.1046

Problem 5.38. The following extensions of the inequalities established in Exercise 3.45 are useful in establishing a SLLN (see Miscellanea 5.8.4). Let $X_1, X_2, \ldots, X_n$ be iid with mgf $M_X(t)$, $-h < t < h$, and let $S_n = \sum_{i=1}^n X_i$ and $\bar{X}_n = S_n/n$.
(a) Show that $P(S_n > a) \leq e^{-at}[M_X(t)]^n$, for $0 < t < h$, and $P(S_n \leq a) \leq e^{-at}[M_X(t)]^n$, for $-h < t \leq 0$.
(b) Use the facts that $M_X(0) = 1$ and $M_X{}'(0) = \mathrm{E}(X)$ to show that, if $\mathrm{E}(X) < 0$, then there is a $0 < c < 1$ with $P(S_n > a) \leq c^n$. Establish a similar bound for $P(S_n \leq a)$.

(c) Define $Y_i = X_i - \mu - \varepsilon$ and use the above argument, with $a = 0$, to establish that $P(\bar{X}_n - \mu > \varepsilon) \le c^n$.
(d) Now define $Y_i = -X_i + \mu - \varepsilon$, establish an inequality similar to part (c), and combine the two to get

$$P(|\bar{X}_n - \mu| > \varepsilon) \le 2c^n \quad \text{for some } 0 < c < 1.$$

Solution. (a) Apply Exercise 3.45 to S_n. Its hypothesis holds: the X_i being iid, Theorem 4.6.7 gives $M_{S_n}(t) = [M_X(t)]^n$, finite on the same interval $(-h, h)$. Part (a) of that exercise then reads

$$P(S_n > a) \le e^{-at} M_{S_n}(t) = e^{-at}[M_X(t)]^n, \qquad 0 < t < h,$$

and part (b) gives $P(S_n \le a) \le e^{-at}[M_X(t)]^n$ for $-h < t \le 0$.
(b) M_X is differentiable on $(-h, h)$ with $M_X(0) = 1$ and $M_X{}'(0) = \mathrm{E}X < 0$, so M_X is strictly decreasing at 0: there is $t_0 \in (0, h)$ with

$$c_0 := M_X(t_0) < 1.$$

Hence for every $a \ge 0$, by part (a),

$$P(S_n > a) \le e^{-at_0} c_0^n \le c_0^n,$$

since $e^{-at_0} \le 1$. Take $c = c_0$.
Symmetrically, if $\mathrm{E}X > 0$ then $M_X{}'(0) > 0$ forces $M_X(t_1) =: c_1 < 1$ for some $t_1 \in (-h, 0)$, and for $a \le 0$ part (a) gives $P(S_n \le a) \le e^{-at_1} c_1^n \le c_1^n$.
(c) Put $\mu = \mathrm{E}X_i$ and fix $\varepsilon > 0$. The $Y_i = X_i - \mu - \varepsilon$ are iid with

$$\mathrm{E}Y_i = -\varepsilon < 0, \qquad M_Y(t) = e^{-(\mu+\varepsilon)t} M_X(t)$$

finite on $(-h, h)$. Since $\sum_{i=1}^n Y_i = S_n - n\mu - n\varepsilon$,

$$P(\bar{X}_n - \mu > \varepsilon) = P\Big(\sum_{i=1}^n Y_i > 0\Big) \le c_1^n$$

by part (b) with $a = 0$, for some $0 < c_1 < 1$.
(d) The $Y_i = -X_i + \mu - \varepsilon$ are iid with $\mathrm{E}Y_i = -\varepsilon < 0$ and mgf $e^{(\mu-\varepsilon)t} M_X(-t)$, finite on $(-h, h)$; and $\sum_{i=1}^n Y_i = n\mu - S_n - n\varepsilon$, so part (b) with $a = 0$ gives

$$P(\bar{X}_n - \mu < -\varepsilon) = P\Big(\sum_{i=1}^n Y_i > 0\Big) \le c_2^n$$

for some $0 < c_2 < 1$. Adding the two one-sided bounds and setting $c = \max\{c_1, c_2\} < 1$,

$$P(|\bar{X}_n - \mu| > \varepsilon) \le c_1^n + c_2^n \le 2c^n.$$

Problem 5.39. This exercise, and the two following, will look at some of the mathematical details of convergence.
(a) Prove Theorem 5.5.4. (Theorem 5.5.4: if $X_1, X_2, \ldots$ converges in probability to a random variable X and h is a continuous function, then $h(X_1), h(X_2), \ldots$ converges in probability to $h(X)$.) (Hint: Since h is continuous, given $\varepsilon > 0$ we can find a δ such that $|h(x_n) - h(x)| < \varepsilon$ whenever $|x_n - x| < \delta$. Translate this into probability statements.)
(b) In Example 5.5.8, find a subsequence of the X_is that converges almost surely, that is, that

converges pointwise. (Example 5.5.8 takes $S = [0,1]$ with the uniform distribution, $X(s) = s$, and

$$\begin{aligned} X_1(s) &= s + I_{[0,1]}(s), \\ X_2(s) &= s + I_{[0,1/2]}(s), \quad X_3(s) = s + I_{[1/2,1]}(s), \\ X_4(s) &= s + I_{[0,1/3]}(s), \quad X_5(s) = s + I_{[1/3,2/3]}(s), \quad X_6(s) = s + I_{[2/3,1]}(s), \end{aligned}$$

and so on: the kth block of k functions adds the indicators of the k intervals $[(j-1)/k, j/k]$, $j = 1, \ldots, k$. The sequence converges to X in probability but not almost surely.)

Solution. (a) The hint's δ turns the continuity statement into the event inclusion

$$\{|X_n - X| < \delta\} \subset \{|h(X_n) - h(X)| < \varepsilon\},$$

so $P(|h(X_n) - h(X)| < \varepsilon) \geq P(|X_n - X| < \delta) \to 1$, which is exactly Definition 5.5.1 for $h(X_n) \to h(X)$. The one gap is that continuity gives a δ depending on the point x; uniformity is bought with one compact set. Fix $\varepsilon > 0$ and $\eta > 0$, and choose M with $P(|X| > M) < \eta/2$ (possible, since $P(|X| > M) \downarrow 0$ as $M \to \infty$). On the compact interval $[-M-1, M+1]$, h is uniformly continuous, so there is $\delta \in (0,1)$ with $|h(x) - h(y)| < \varepsilon$ whenever $|x| \leq M$ and $|x - y| < \delta$. Then

$$\{|h(X_n) - h(X)| \geq \varepsilon\} \subset \{|X_n - X| \geq \delta\} \cup \{|X| > M\},$$

whence

$$\limsup_n P(|h(X_n) - h(X)| \geq \varepsilon) \leq \lim_n P(|X_n - X| \geq \delta) + \tfrac{\eta}{2} = \tfrac{\eta}{2}.$$

As $\eta > 0$ was arbitrary, $P(|h(X_n) - h(X)| \geq \varepsilon) \to 0$.
(b) Take the first member of each block, $n_k = \frac{k(k-1)}{2} + 1$, so that

$$X_{n_k}(s) = s + I_{[0,\,1/k]}(s), \qquad k = 1, 2, \ldots$$

(the indices $1, 2, 4, 7, 11, \ldots$). For every $s \in (0,1]$ we have $1/k < s$ for all $k > 1/s$, so $X_{n_k}(s) = s = X(s)$ eventually; the convergence fails only at $s = 0$, and $P(\{0\}) = 0$. Hence $X_{n_k} \to X$ almost surely by Definition 5.5.6.

Problem 5.40. Prove Theorem 5.5.12 for the case where X_n and X are continuous random variables. (Theorem 5.5.12: if the sequence $X_1, X_2, \ldots$ converges in probability to a random variable X, then it converges in distribution to X.)
(a) Given t and ε, show that $P(X \leq t - \varepsilon) \leq P(X_n \leq t) + P(|X_n - X| \geq \varepsilon)$. This gives a lower bound on $P(X_n \leq t)$.
(b) Use a similar strategy to get an upper bound on $P(X_n \leq t)$.
(c) By pinching, deduce that $P(X_n \leq t) \to P(X \leq t)$.

Solution. (a) If $X \leq t - \varepsilon$ and $|X_n - X| < \varepsilon$, then $X_n < X + \varepsilon \leq t$; hence

$$\{X \leq t - \varepsilon\} \subset \{X_n \leq t\} \cup \{|X_n - X| \geq \varepsilon\},$$

and Boole's inequality gives

$$F_X(t - \varepsilon) \leq F_{X_n}(t) + P(|X_n - X| \geq \varepsilon),$$

that is, $F_{X_n}(t) \geq F_X(t - \varepsilon) - P(|X_n - X| \geq \varepsilon)$.
(b) If $X_n \leq t$ and $|X_n - X| < \varepsilon$, then $X < X_n + \varepsilon \leq t + \varepsilon$; hence

$$\{X_n \leq t\} \subset \{X \leq t + \varepsilon\} \cup \{|X_n - X| \geq \varepsilon\},$$

so $F_{X_n}(t) \leq F_X(t + \varepsilon) + P(|X_n - X| \geq \varepsilon)$.

(c) $P(|X_n - X| \geq \varepsilon) \to 0$ by Definition 5.5.1, so letting $n \to \infty$ in (a) and (b) with ε held fixed,

$$F_X(t-\varepsilon) \;\leq\; \liminf_n F_{X_n}(t) \;\leq\; \limsup_n F_{X_n}(t) \;\leq\; F_X(t+\varepsilon).$$

Now let $\varepsilon \downarrow 0$. Since X is continuous, F_X is continuous at t, so both outer terms tend to $F_X(t)$ and the two limits are pinched together:

$$\lim_{n\to\infty} P(X_n \leq t) = P(X \leq t)$$

for every t, which is Definition 5.5.10.

Problem 5.41. Prove Theorem 5.5.13; that is, show that

$$P(|X_n - \mu| > \varepsilon) \to 0 \text{ for every } \varepsilon \quad\Longleftrightarrow\quad P(X_n \leq x) \to \begin{cases} 0 & \text{if } x < \mu, \\ 1 & \text{if } x \geq \mu. \end{cases}$$

(a) Set $\varepsilon = |x - \mu|$ and show that if $x > \mu$, then $P(X_n \leq x) \geq P(|X_n - \mu| \leq \varepsilon)$, while if $x < \mu$, then $P(X_n \leq x) \leq P(|X_n - \mu| \geq \varepsilon)$. Deduce the $\Rightarrow$ implication.
(b) Use the fact that $\{x : |x - \mu| > \varepsilon\} = \{x : x - \mu < -\varepsilon\} \cup \{x : x - \mu > \varepsilon\}$ to deduce the $\Leftarrow$ implication.
(See Billingsley 1995, Section 25, for a detailed treatment of the above results.)

Solution. The printed "1 if $x \geq \mu$" is a misprint for "1 if $x > \mu$" (Theorem 5.5.13 itself has $x > \mu$; at $x = \mu$ the claim is false, as $X_n \equiv \mu + 1/n$ shows), and that is the statement proved below.
(a) ($\Rightarrow$) Two inclusions, one for each side of μ; halving $|x - \mu|$ converts the book's weak inequalities into the strict ones that the hypothesis controls.
(i) $x > \mu$. Put $\varepsilon = (x - \mu)/2 > 0$. Then $|X_n - \mu| \leq \varepsilon$ forces $X_n \leq \mu + \varepsilon < x$, so

$$P(X_n \leq x) \geq P(|X_n - \mu| \leq \varepsilon) = 1 - P(|X_n - \mu| > \varepsilon) \longrightarrow 1,$$

and since $P(X_n \leq x) \leq 1$ always, $P(X_n \leq x) \to 1$.
(ii) $x < \mu$. Put $\varepsilon = (\mu - x)/2 > 0$. Then $X_n \leq x$ forces $\mu - X_n \geq 2\varepsilon > \varepsilon$, so

$$P(X_n \leq x) \leq P(|X_n - \mu| > \varepsilon) \longrightarrow 0.$$

(b) ($\Leftarrow$) Fix $\varepsilon > 0$ and split the event at μ:

$$\begin{aligned} P(|X_n - \mu| > \varepsilon) &= P(X_n < \mu - \varepsilon) + P(X_n > \mu + \varepsilon) \\ &\leq P\big(X_n \leq \mu - \tfrac{\varepsilon}{2}\big) + \Big[1 - P(X_n \leq \mu + \varepsilon)\Big]. \end{aligned}$$

The point $\mu - \varepsilon/2$ lies strictly below μ, so the first term tends to 0; the point $\mu + \varepsilon$ lies strictly above μ, so $P(X_n \leq \mu + \varepsilon) \to 1$ and the bracket tends to 0. Hence $P(|X_n - \mu| > \varepsilon) \to 0$ for every $\varepsilon > 0$.

Problem 5.42. Similar to Example 5.5.11, let $X_1, X_2, \ldots$ be iid and $X_{(n)} = \max_{1 \leq i \leq n} X_i$.
(a) If $X_i \sim \text{beta}(1, \beta)$, find a value of ν so that $n^\nu(1 - X_{(n)})$ converges in distribution.
(b) If $X_i \sim \text{exponential}(1)$, find a sequence a_n so that $X_{(n)} - a_n$ converges in distribution.

Solution. (a) $\nu = 1/\beta$. The beta$(1, \beta)$ cdf is $F(x) = 1 - (1 - x)^\beta$ on $(0, 1)$, so for $t > 0$ and n large enough that $tn^{-\nu} < 1$,

$$\begin{aligned} P\big(n^\nu(1 - X_{(n)}) > t\big) &= P\big(X_{(n)} < 1 - tn^{-\nu}\big) = \big[F(1 - tn^{-\nu})\big]^n \\ &= \Big[1 - t^\beta n^{-\nu\beta}\Big]^n. \end{aligned}$$

A nondegenerate limit requires $\nu\beta = 1$; with $\nu = 1/\beta$, Lemma 2.3.14 gives

$$\Big[1 - \frac{t^\beta}{n}\Big]^n \longrightarrow e^{-t^\beta},$$

so $n^{1/\beta}(1 - X_{(n)})$ converges in distribution to the Weibull law with cdf $1 - e^{-t^\beta}$, $t > 0$.
(b) $a_n = \log n$. With $F(x) = 1 - e^{-x}$ on $x > 0$, for each fixed t and all $n > e^{-t}$,

$$P\big(X_{(n)} - \log n \le t\big) = \big[F(t + \log n)\big]^n = \Big[1 - \frac{e^{-t}}{n}\Big]^n \longrightarrow \exp\{-e^{-t}\},$$

again by Lemma 2.3.14. Thus $X_{(n)} - \log n$ converges in distribution to the Gumbel (extreme value) law with cdf $e^{-e^{-t}}$, $-\infty < t < \infty$.

5.7 Exercises 5.43–5.49

Problem 5.43. Fill in the details in the proof of Theorem 5.5.24 (the Delta Method).
(a) Show that if $\sqrt{n}(Y_n - \mu) \to \mathrm{n}(0, \sigma^2)$ in distribution, then $Y_n \to \mu$ in probability.
(b) Give the details for the application of Slutsky's Theorem (Theorem 5.5.17).

Solution. (a) Write $W_n = \sqrt{n}(Y_n - \mu) \to Z \sim \mathrm{n}(0, \sigma^2)$ in distribution. Then $Y_n - \mu = n^{-1/2} W_n \to 0 \cdot Z = 0$ in distribution, by Slutsky's Theorem 5.5.17(a) applied to the constant sequence $n^{-1/2} \to 0$; convergence in distribution to a constant is convergence in probability (Theorem 5.5.13).
Method (2): fix $\varepsilon > 0$ and $M > 0$; for $n \ge (M/\varepsilon)^2$,

$$P(|Y_n - \mu| > \varepsilon) = P(|W_n| > \sqrt{n}\,\varepsilon) \le P(|W_n| > M) \to P(|Z| > M),$$

the limit because the $\mathrm{n}(0, \sigma^2)$ cdf is continuous, so $\pm M$ are continuity points. Let $M \to \infty$.
(b) Taylor's expansion (5.5.11) about θ is

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + R_n,$$

which is the definition of R_n. Put

$$h(y) = \begin{cases} \dfrac{g(y) - g(\theta) - g'(\theta)(y - \theta)}{y - \theta}, & y \ne \theta, \\ 0, & y = \theta, \end{cases}$$

so that $R_n = h(Y_n)(Y_n - \theta)$ identically. That $g'(\theta)$ exists is exactly the statement $\lim_{y \to \theta} h(y) = 0 = h(\theta)$, i.e. h is continuous at θ. By part (a), $Y_n \to \theta$ in probability, so Theorem 5.5.4 (whose proof uses continuity of h only at the limit point, here the constant θ) gives

$$h(Y_n) \to 0 \quad \text{in probability.}$$

Now $\sqrt{n}\,R_n = h(Y_n) \cdot \sqrt{n}(Y_n - \theta)$, where $\sqrt{n}(Y_n - \theta) \to Z \sim \mathrm{n}(0, \sigma^2)$ by hypothesis, so $\sqrt{n} R_n \to 0 \cdot Z = 0$ in distribution by Theorem 5.5.17(a) and hence in probability by Theorem 5.5.13. Then Theorem 5.5.17(b) applied to

$$\sqrt{n}\big[g(Y_n) - g(\theta)\big] = g'(\theta)\sqrt{n}(Y_n - \theta) + \sqrt{n} R_n$$

gives convergence in distribution to $g'(\theta)Z + 0$, and $g'(\theta)Z \sim \mathrm{n}\big(0, \sigma^2[g'(\theta)]^2\big)$, which is (5.5.10).

Problem 5.44. Let $X_i, i = 1, 2, \ldots$, be independent Bernoulli(p) random variables and let $Y_n = \frac{1}{n}\sum_{i=1}^n X_i$.
(a) Show that $\sqrt{n}\,(Y_n - p) \to \mathrm{n}[0, p(1-p)]$ in distribution.

(b) Show that for $p \neq 1/2$, the estimate of variance $Y_n(1-Y_n)$ satisfies

$$\sqrt{n}\big[Y_n(1-Y_n) - p(1-p)\big] \to \mathrm{n}\big[0, (1-2p)^2 p(1-p)\big]$$

in distribution.
(c) Show that for $p = 1/2$, $n\big[Y_n(1-Y_n) - \frac{1}{4}\big] \to -\frac{1}{4}\chi_1^2$ in distribution. (If this appears strange, note that $Y_n(1-Y_n) \leq 1/4$, so the left-hand side is always negative. An equivalent form is $2n\big[\frac{1}{4} - Y_n(1-Y_n)\big] \to \chi_1^2$.)

Solution. (a) This is the Central Limit Theorem (Theorem 5.5.14) verbatim: the X_i are iid with $\mathrm{E}X_i = p$, $\mathrm{Var}\, X_i = p(1-p) > 0$, and mgf $1 - p + pe^t$ finite for all t, so the neighborhood-of-zero mgf hypothesis holds.
(b) Take $g(y) = y(1-y)$, so $g'(y) = 1 - 2y$ and $g'(p) = 1 - 2p \neq 0$ precisely when $p \neq 1/2$. With $\sigma^2 = p(1-p)$, the Delta Method (Theorem 5.5.24) applied to the sequence of part (a) gives

$$\sqrt{n}\big[g(Y_n) - g(p)\big] \to \mathrm{n}\big(0,\ p(1-p)(1-2p)^2\big).$$

(c) Here $g'(1/2) = 0$ and $g''(y) \equiv -2 \neq 0$, so the hypotheses of the second-order Delta Method (Theorem 5.5.26) hold with $\theta = 1/2$ and $\sigma^2 = \frac{1}{4}$, and (5.5.13) reads

$$n\Big[Y_n(1-Y_n) - \tfrac{1}{4}\Big] \to \sigma^2 \frac{g''(1/2)}{2}\chi_1^2 = \frac{1}{4}\cdot(-1)\,\chi_1^2 = -\frac{1}{4}\chi_1^2.$$

(The parenthetical equivalent form has a misprinted constant: multiplying by -4 gives $4n[\frac{1}{4} - Y_n(1-Y_n)] = 4n(Y_n - \frac{1}{2})^2 \to \chi_1^2$, not $2n$.)

Problem 5.45. For the situation of Example 5.6.1, calculate the probability that at least 50% of the components last 100 hours when
(a) $c = 300$, $X \sim$ gamma(a, b), $a = 4$, $b = 25$.
(b) $c = 100$, $X \sim$ gamma(a, b), $a = 20$, $b = 5$.
(c) $c = 100$, $X \sim$ gamma(a, b), $a = 20.7$, $b = 5$.
(Hint: In parts (a) and (b) it is possible to evaluate the gamma integral in closed form, although it probably isn't worth the effort in (b). There is no closed-form expression for the integral in part (c), which has to be evaluated through either numerical integration or simulation.)
In Example 5.6.1, c components have independent lifetimes distributed as X; the probability that one component lasts at least h hours is

$$p_1 = P(X \geq h),$$

and, modelling the c components as Bernoulli trials, the probability that at least t of them last h hours is

$$p_2 = \sum_{k=t}^{c} \binom{c}{k} p_1^k (1-p_1)^{c-k}.$$

Here $h = 100$ and $t = c/2$.

Solution. The three answers are $p_2 = .0119$, $.3096$, and $.7753$.
For integer a the gamma survival function is the Poisson tail (integrate by parts $a-1$ times): if $X \sim$ gamma(a, b) then

$$P(X \geq x) = \sum_{k=0}^{a-1} e^{-x/b} \frac{(x/b)^k}{k!}.$$

(a) $x/b = 100/25 = 4$, so

$$p_1 = e^{-4}\Big(1 + 4 + 8 + \tfrac{32}{3}\Big) = \tfrac{71}{3}e^{-4} = .43347,$$

and with $c = 300$, $t = 150$,

$$p_2 = \sum_{k=150}^{300} \binom{300}{k} (.43347)^k (.56653)^{300-k} = .01194.$$

(b) $x/b = 100/5 = 20$, so $p_1 = e^{-20} \sum_{k=0}^{19} 20^k/k! = .47026$ (the closed form is a sum of twenty terms, hence the hint), and with $c = 100$, $t = 50$,

$$p_2 = \sum_{k=50}^{100} \binom{100}{k} (.47026)^k (.52974)^{100-k} = .30962.$$

(c) $a = 20.7$ is not an integer, so the tail must be evaluated numerically:

$$p_1 = \frac{1}{\Gamma(20.7)} \int_{20}^{\infty} u^{19.7} e^{-u}\, du = .53270,$$

after the substitution $u = x/5$. Then, with $c = 100$ and threshold 50,

$$p_2 = \sum_{k=50}^{100} \binom{100}{k} (.53270)^k (.46730)^{100-k} = .77529.$$

Problem 5.46. Referring to Exercise 5.45, compare your answers to what is obtained from a normal approximation to the binomial (see Example 3.3.2).
In Example 3.3.2 a binomial(n, p) variable X is approximated by $Y \sim \mathrm{n}\big(np,\ np(1-p)\big)$, the approximation being recommended when $\min\{np,\ n(1-p)\} > 5$; the continuity correction replaces the cutoff x by $x \pm \frac{1}{2}$, so that the approximating normal area matches the whole binomial bar.

Solution. With the continuity correction the normal approximation is accurate to three decimals in all three parts; without it the errors are large (a relative error of 16% in part (a)).
For $X \sim$ binomial(c, p_1) put $\mu = cp_1$ and $\sigma = \sqrt{cp_1(1-p_1)}$. The uncorrected and corrected approximations to $P(X \ge t)$, $t = c/2$, are

$$P\left(Z \ge \frac{t-\mu}{\sigma}\right) \qquad \text{and} \qquad P\left(Z \ge \frac{t-\frac{1}{2}-\mu}{\sigma}\right),$$

the second because $P(X \ge t) = P(X \ge t - \frac{1}{2})$ by discreteness. In every part $\min\{\mu,\ c-\mu\} \ge 46$, comfortably past the rule of thumb.
(a) $\mu = 300(.43347) = 130.041$, $\sigma = 8.5832$:

$$\begin{aligned} \text{uncorrected} &: P(Z \ge 2.3253) = .01003, \\ \text{corrected} &: P(Z \ge 2.2671) = .01169, \\ \text{exact} &: .01194. \end{aligned}$$

(b) $\mu = 100(.47026) = 47.026$, $\sigma = 4.9911$:

$$\begin{aligned} \text{uncorrected} &: P(Z \ge .5959) = .27562, \\ \text{corrected} &: P(Z \ge .4957) = .31004, \\ \text{exact} &: .30962. \end{aligned}$$

(c) $\mu = 100(.53270) = 53.270$, $\sigma = 4.9893$:

$$\begin{aligned} \text{uncorrected} &: P(Z \ge -.6555) = .74392, \\ \text{corrected} &: P(Z \ge -.7557) = .77509, \\ \text{exact} &: .77529. \end{aligned}$$

Problem 5.47. Verify the distributions of the random variables in (5.6.5). That is, with $U_1, U_2, \ldots$ iid uniform$(0,1)$ and ν, a, b positive integers, show that

$$Y = -2\sum_{j=1}^{\nu} \log(U_j) \sim \chi^2_{2\nu},$$
$$Y = -\beta\sum_{j=1}^{a} \log(U_j) \sim \text{gamma}(a,\beta),$$
$$Y = \frac{\sum_{j=1}^{a} \log(U_j)}{\sum_{j=1}^{a+b} \log(U_j)} \sim \text{beta}(a,b).$$

Solution. Everything follows from $V_j = -\log U_j \sim$ exponential(1): for $v > 0$,

$$P(V_j > v) = P(U_j < e^{-v}) = e^{-v},$$

since U_j is uniform$(0,1)$ and $e^{-v} \in (0,1)$. The V_j are independent, being functions of the independent U_j (Theorem 4.6.12).
First display. $-2\log U_j = 2V_j \sim$ exponential(2) $=$ gamma$(1,2) = \chi^2_2$, and chi squared degrees of freedom add over independent summands (Lemma 5.3.2(b)), so

$$-2\sum_{j=1}^{\nu} \log U_j = \sum_{j=1}^{\nu} 2V_j \sim \chi^2_{2\nu}.$$

Second display. $-\beta \log U_j = \beta V_j \sim$ gamma$(1,\beta)$, with mgf $(1-\beta t)^{-1}$ for $t < 1/\beta$. By Theorem 4.6.7 the sum of a independent such variables has mgf

$$\prod_{j=1}^{a} (1-\beta t)^{-1} = (1-\beta t)^{-a}, \qquad t < 1/\beta,$$

which is the gamma(a,β) mgf; uniqueness (Theorem 2.3.11) gives $-\beta\sum_{j=1}^{a} \log U_j \sim$ gamma(a,β). (This is Example 4.6.8.)
Third display. The minus signs cancel in the ratio, so with

$$S = \sum_{j=1}^{a} V_j \sim \text{gamma}(a,1), \qquad T = \sum_{j=a+1}^{a+b} V_j \sim \text{gamma}(b,1)$$

(by the second display with $\beta = 1$), which are independent because they are functions of disjoint subsets of the independent V_j (Theorem 4.6.12),

$$Y = \frac{\sum_{j=1}^{a} V_j}{\sum_{j=1}^{a+b} V_j} = \frac{S}{S+T} \sim \text{beta}(a,b)$$

by Exercise 4.24.

Problem 5.48. Using strategies similar to (5.6.5), show how to generate an $F_{m,n}$ random variable, where both m and n are even integers.

Solution. Take $U_1, \ldots, U_{(m+n)/2}$ iid uniform$(0,1)$ and set

$$X = \frac{n\sum_{j=1}^{m/2} \log U_j}{m\sum_{j=m/2+1}^{(m+n)/2} \log U_j} \sim F_{m,n}.$$

Indeed, $m/2$ and $n/2$ are integers, so the first line of (5.6.5) applies to each block:

$$W_1 = -2\sum_{j=1}^{m/2} \log U_j \sim \chi^2_m, \qquad W_2 = -2\sum_{j=m/2+1}^{(m+n)/2} \log U_j \sim \chi^2_n,$$

and W_1, W_2 are independent, being functions of disjoint subsets of the independent U_j (Theorem 4.6.12). By the reduction on p. 225 (see Exercise 5.17), $F_{m,n}$ is the distribution of $(W_1/m)/(W_2/n)$ for exactly such a pair, and the factors of -2 cancel:

$$\frac{W_1/m}{W_2/n} = \frac{n}{m} \cdot \frac{-2\sum_{j=1}^{m/2} \log U_j}{-2\sum_{j=m/2+1}^{(m+n)/2} \log U_j} = X.$$

Method (2): by Theorem 5.3.8(c), $B = (m/n)X/[1 + (m/n)X] \sim \text{beta}(m/2, n/2)$ when $X \sim F_{m,n}$, so $X = (n/m)B/(1 - B)$. Generating B by the third line of (5.6.5) with $a = m/2$, $b = n/2$ and substituting reproduces the displayed X exactly.

Problem 5.49. Let $U \sim \text{uniform}(0, 1)$.
(a) Show that both $-\log U$ and $-\log(1 - U)$ are exponential random variables.
(b) Show that $X = \log \dfrac{U}{1-U}$ is a logistic$(0, 1)$ random variable.
(c) Show how to generate a logistic(μ, β) random variable.
Here (Exercise 3.26) the logistic(μ, β) distribution is the one with cdf

$$F(t) = \frac{1}{1 + e^{-(t-\mu)/\beta}}, \qquad -\infty < t < \infty.$$

Solution. (a) Both are exponential(1). For $v > 0$, since $e^{-v} \in (0, 1)$,

$$P(-\log U > v) = P(U < e^{-v}) = e^{-v},$$

and $1 - U \sim \text{uniform}(0, 1)$ as well (its cdf is $P(1 - U \le u) = P(U \ge 1 - u) = u$ on $(0, 1)$), so the same computation gives $P(-\log(1 - U) > v) = e^{-v}$.
(b) For $-\infty < x < \infty$, the map $u \mapsto u/(1 - u)$ is increasing on $(0, 1)$, so

$$\begin{aligned} P(X \le x) &= P\left(\frac{U}{1-U} \le e^x\right) \\ &= P\left(U \le \frac{e^x}{1+e^x}\right) \\ &= \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}, \end{aligned}$$

the third equality because $e^x/(1 + e^x) \in (0, 1)$. This is the logistic cdf with $\mu = 0$, $\beta = 1$.
(c) Set

$$Y = \mu + \beta \log \frac{U}{1-U} = \mu + \beta X.$$

Since $\beta > 0$, $P(Y \le y) = P(X \le (y-\mu)/\beta) = \left[1 + e^{-(y-\mu)/\beta}\right]^{-1}$, the logistic$(\mu, \beta)$ cdf. (This is (5.6.6), the Probability Integral Transform: $F^{-1}(u) = \mu + \beta \log[u/(1 - u)]$.)

5.8 Exercises 5.50–5.56

Problem 5.50. The Box-Muller method for generating normal *pseudo-random variables* (Example 5.6.4) is based on the transformation

$$X_1 = \cos(2\pi U_1)\sqrt{-2\log U_2}, \qquad X_2 = \sin(2\pi U_1)\sqrt{-2\log U_2},$$

where U_1 and U_2 are iid uniform$(0,1)$. Prove that X_1 and X_2 are independent n$(0,1)$ random variables.

Solution. (X_1, X_2) is the polar point $(R\cos\Theta, R\sin\Theta)$ with $\Theta = 2\pi U_1 \sim$ uniform$(0, 2\pi)$ and $R = \sqrt{-2\log U_2}$ Rayleigh, and (4.3.2) turns that pair into the standard bivariate normal density.
For $r > 0$,

$$\begin{aligned} P(R > r) &= P\big(-2\log U_2 > r^2\big) \\ &= P\Big(U_2 < e^{-r^2/2}\Big) = e^{-r^2/2}, \end{aligned}$$

so $f_R(r) = re^{-r^2/2}$ on $r > 0$ (i.e. $R^2 \sim$ exponential$(2) = \chi^2_2$). Since U_1 and U_2 are independent, so are Θ and R (Theorem 4.3.5), whence

$$f_{R,\Theta}(r,\theta) = \frac{1}{2\pi}\, re^{-r^2/2}, \qquad r > 0,\ 0 < \theta < 2\pi.$$

The map $(r,\theta) \mapsto (r\cos\theta, r\sin\theta)$ carries $(0,\infty)\times(0,2\pi)$ one-to-one onto $\mathbb{R}^2$ minus the ray $\{(x_1, 0) : x_1 \geq 0\}$, a set of measure zero. Its inverse is $r = \sqrt{x_1^2 + x_2^2}$, $\theta = \arctan(x_2/x_1)$ taken in the correct quadrant, with Jacobian

$$J = \begin{vmatrix} \dfrac{\partial r}{\partial x_1} & \dfrac{\partial r}{\partial x_2} \\[2ex] \dfrac{\partial \theta}{\partial x_1} & \dfrac{\partial \theta}{\partial x_2} \end{vmatrix} = \frac{1}{\sqrt{x_1^2 + x_2^2}}$$

(Check!). Substituting into (4.3.2),

$$\begin{aligned} f_{X_1,X_2}(x_1,x_2) &= \frac{1}{2\pi}\sqrt{x_1^2+x_2^2}\; e^{-(x_1^2+x_2^2)/2} \cdot \frac{1}{\sqrt{x_1^2+x_2^2}} \\ &= \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-x_2^2/2} \end{aligned}$$

for all $(x_1, x_2) \in \mathbb{R}^2$. The joint pdf factors into a function of x_1 alone times a function of x_2 alone, so by Lemma 4.2.7 X_1 and X_2 are independent, and each factor is the n$(0,1)$ pdf.

Problem 5.51. One of the earlier methods (not one of the better ones) of generating pseudo-random standard normal random variables from uniform random variables is to take

$$X = \sum_{i=1}^{12} U_i - 6,$$

where the U_is are iid uniform$(0,1)$.
(a) Justify the fact that X is approximately n$(0,1)$.
(b) Can you think of any obvious way in which the approximation fails?
(c) Show how good (or bad) the approximation is by comparing the first four moments. (The fourth moment is $29/10$ and is a lengthy calculation — mgfs and computer algebra would help; see Example A.0.6 in Appendix A.)

Solution. (a) X is exactly the Central Limit Theorem statistic at $n = 12$. With $U_i \sim$ uniform$(0,1)$ we

have $\mu = EU_i = 1/2$ and $\sigma^2 = \operatorname{Var} U_i = 1/12$ (Example 5.5.3), so $\sigma = 1/\sqrt{12}$ and

$$\left.\frac{\sqrt{n}\,(\bar{U}_n - \mu)}{\sigma}\right|_{n=12} = \frac{\sqrt{12}\,\big(\bar{U}_{12} - \tfrac{1}{2}\big)}{1/\sqrt{12}} = \sum_{i=1}^{12} U_i - 6 = X.$$

The uniform mgf exists everywhere, so Theorem 5.5.14 applies and this quantity tends to $\mathrm{n}(0,1)$; the choice $n = 12$ is made purely so that $\sigma\sqrt{n} = 1$ and no division is needed.
(b) The tails. $0 \le U_i \le 1$ forces $-6 \le X \le 6$ with probability one, so $P(|X| > 6) = 0$, while a genuine $\mathrm{n}(0,1)$ puts probability 2×10^{-9} out there. No simulation built this way will ever produce a normal deviate beyond 6σ.
(c) Let $V_i = U_i - \frac{1}{2} \sim \text{uniform}\big(-\frac{1}{2}, \frac{1}{2}\big)$, so $X = \sum_{i=1}^{12} V_i$ and

$$M_V(t) = \int_{-1/2}^{1/2} e^{tv}\,dv = \frac{e^{t/2} - e^{-t/2}}{t} = \frac{\sinh(t/2)}{t/2}.$$

By Theorem 4.2.12 (independence), $M_X(t) = [\sinh(t/2)/(t/2)]^{12}$, so the cumulant generating function is

$$\begin{aligned} K_X(t) &= 12 \log \frac{\sinh(t/2)}{t/2} = 12\left[\frac{(t/2)^2}{6} - \frac{(t/2)^4}{180} + \cdots\right] \\ &= \frac{t^2}{2} - \frac{t^4}{240} + \frac{t^6}{15120} - \cdots, \end{aligned}$$

against $K(t) = t^2/2$ for the $\mathrm{n}(0,1)$. Hence $\kappa_1 = 0$, $\kappa_2 = 1$, $\kappa_3 = 0$, and $\kappa_4 = 4!\,\big(-\frac{1}{240}\big) = -\frac{1}{10}$, giving

$$EX = 0, \quad EX^2 = 1, \quad EX^3 = 0, \quad EX^4 = \kappa_4 + 3\kappa_2^2 = \frac{29}{10}.$$

(The odd moments vanish because X is symmetric about 0.) So the first three moments match the $\mathrm{n}(0,1)$ exactly and the fourth is 2.9 against 3: the approximation is slightly light-tailed, a 3.3% kurtosis deficit.
Method (2): expand $(\sum_i V_i)^4$ directly, using $EV_1^2 = 1/12$, $EV_1^4 = 1/80$ and the multinomial count $\binom{4}{2,2}\binom{12}{2} = 396$:

$$EX^4 = 12\,EV_1^4 + 3 \cdot 12 \cdot 11\,(EV_1^2)^2 = \frac{12}{80} + \frac{396}{144} = \frac{29}{10}.$$

Problem 5.52. For each of the following distributions write down an algorithm for generating the indicated random variables.
(a) $Y \sim \text{binomial}\big(8, \frac{2}{3}\big)$
(b) $Y \sim$ hypergeometric $N = 10$, $M = 8$, $K = 4$
(c) $Y \sim \text{negative binomial}\big(5, \frac{1}{3}\big)$

Solution. In each case apply the discrete inverse-cdf algorithm (5.6.7): generate $U \sim \text{uniform}(0,1)$ and set $Y = y_{i+1}$ whenever $F_Y(y_i) < U \le F_Y(y_{i+1})$, with $F_Y(y_0) = 0$.
(a) $f_Y(y) = \binom{8}{y}\big(\frac{2}{3}\big)^y\big(\frac{1}{3}\big)^{8-y}$, $y = 0, \ldots, 8$, giving

y	0	1	2	3	4	5	6	7	8
$f_Y(y)$	0.0002	0.0024	0.0171	0.0683	0.1707	0.2731	0.2731	0.1561	0.0390
$F_Y(y)$	0.0002	0.0026	0.0197	0.0879	0.2586	0.5318	0.8049	0.9610	1.0000

so generate U and set $Y = \min\{y : U \le F_Y(y)\}$; e.g. $Y = 5$ if $0.2586 < U \le 0.5318$. Scanning outward from the mean $EY = 16/3$ rather than from $y = 0$ cuts the expected number of comparisons (see Exercise 5.55).
Method (2): generate $U_1, \ldots, U_8$ iid uniform$(0,1)$ and set $Y = \sum_{i=1}^{8} I\big(U_i \le \frac{2}{3}\big)$, a sum of 8 iid Bernoulli$\big(\frac{2}{3}\big)$, which is binomial$\big(8, \frac{2}{3}\big)$ by the very definition of the binomial as a count of successes in independent Bernoulli trials (Section 3.2).
(b) With $N = 10$, $M = 8$, $K = 4$ the support is $y = 2, 3, 4$ (at most 2 of the 4 draws can come from

the $N - M = 2$ non-special items), and $f_Y(y) = \binom{8}{y}\binom{2}{4-y}/\binom{10}{4}$:

y	2	3	4
$f_Y(y)$	2/15	8/15	1/3
$F_Y(y)$	0.1333	0.6667	1.0000

Generate U and set

$$Y = \begin{cases} 2 & \text{if } U \le 0.1333, \\ 3 & \text{if } 0.1333 < U \le 0.6667, \\ 4 & \text{if } 0.6667 < U \le 1. \end{cases}$$

Method (2): simulate the urn directly — draw 4 of the 10 items without replacement (at draw j pick item $\lceil (10 - j + 1)U_j \rceil$ from those remaining) and let Y count how many of the 8 special items were drawn.

(c) $f_Y(y) = \binom{y+4}{4}\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^y$, $y = 0, 1, 2, \ldots$, with $EY = r(1-p)/p = 10$:

y	0	1	2	3	4	5	6	7
$F_Y(y)$	0.0041	0.0178	0.0453	0.0879	0.1448	0.2131	0.2890	0.3685
y	8	9	10	11	12	13	14	15
$F_Y(y)$	0.4480	0.5245	0.5959	0.6609	0.7186	0.7689	0.8121	0.8485

Generate U and return $Y = \min\{y : U \le F_Y(y)\}$, computing F_Y on the fly by the recursion $f_Y(y+1) = \frac{y+5}{y+1} \cdot \frac{2}{3} f_Y(y)$, $f_Y(0) = (1/3)^5$, and starting the scan near $y = 10$.

Method (2): the support is unbounded, so the direct construction is preferable. Generate $U_1, \ldots, U_5$ iid uniform$(0, 1)$ and set

$$Y = \sum_{j=1}^{5} \left\lfloor \frac{\log U_j}{\log(2/3)} \right\rfloor.$$

Each summand is geometric on $\{0, 1, 2, \ldots\}$ with $p = \frac{1}{3}$, since

$$P\Big(\Big\lfloor \tfrac{\log U}{\log(2/3)} \Big\rfloor = g\Big) = P\Big(\big(\tfrac{2}{3}\big)^{g+1} < U \le \big(\tfrac{2}{3}\big)^{g}\Big) = \big(\tfrac{2}{3}\big)^{g}\,\tfrac{1}{3},$$

and a sum of 5 iid such geometrics is negative binomial$\big(5, \frac{1}{3}\big)$.

Problem 5.53. For each of the distributions in the previous exercise:
(a) Generate 1,000 variables from the indicated distribution.
(b) Compare the mean, variance, and histogram of the generated random variables with the theoretical values.

Solution. All three simulations reproduce their targets to within the sampling error the Weak Law (Theorem 5.5.2) allows. Running the three algorithms of Exercise 5.52 with $n = 1000$ replications each:

distribution	$\bar{y}$	EY	s^2	$\operatorname{Var} Y$
binomial$\big(8, \frac{2}{3}\big)$	5.315	5.333	1.699	1.778
hypergeometric$(10, 8, 4)$	3.188	3.200	0.443	0.427
negative binomial$\big(5, \frac{1}{3}\big)$	9.750	10	30.08	30.00

The theoretical values are $EY = np = 16/3$, $\operatorname{Var} Y = np(1-p) = 16/9$; $EY = KM/N = 3.2$, $\operatorname{Var} Y = K\frac{M}{N}\frac{N-M}{N}\frac{N-K}{N-1} = 0.4267$; and $EY = r(1-p)/p = 10$, $\operatorname{Var} Y = r(1-p)/p^2 = 30$. Each discrepancy is within one standard error: e.g. for the negative binomial, $\operatorname{sd}(\bar{Y}) = \sqrt{30/1000} = 0.173$ and 9.750 is 1.4 standard errors from 10.

The histograms (observed counts against the expected counts $1000 f_Y(y)$) agree equally well.

Binomial$\big(8, \frac{2}{3}\big)$:

y	0	1	2	3	4	5	6	7	8
observed	0	1	13	64	186	300	239	158	39
expected	0.2	2.4	17.1	68.3	170.7	273.1	273.1	156.1	39.0

Hypergeometric$(10, 8, 4)$:

y	2	3	4
observed	145	522	333
expected	133.3	533.3	333.3

Negative binomial$\left(5, \frac{1}{3}\right)$:

y	0	1	2	3	4	5	6	7	8	9	10	11	≥ 12
observed	3	16	30	40	60	76	88	81	76	81	72	55	322
expected	4.1	13.7	27.4	42.7	56.9	68.3	75.9	79.5	79.5	76.5	71.4	64.9	339.1

Pearson statistics $\sum(\text{obs}-\text{exp})^2/\text{exp}$ (pooling $y \leq 2$ in the binomial) come to 10.22 on 6 df, 1.26 on 2 df, and 6.90 on 12 df, with p-values 0.12, 0.53, 0.86: the simulated relative frequencies match the pmfs.

Problem 5.54. Refer to Example 5.6.6. Another sample of bay anchovy larvae counts yielded the data

158 143 106 57 97 80 109 109 350 224 109 214 84

(a) Use the technique of Example 5.6.6 to construct a simulated distribution of S^2 to see if the assumption of Poisson counts is tenable.

(b) A possible explanation of the failure of the Poisson assumptions (and increased variance) is the failure of the assumption that the larvae are uniformly distributed in the river. If the larvae tend to clump, the negative binomial(r, p) distribution (with mean $\mu = r\frac{1-p}{p}$ and variance $\mu + \frac{\mu^2}{r}$) is a reasonable alternative model. For $\mu = \bar{x}$, what values of r lead to simulated distributions that are consistent with the data?

Solution. (a) The Poisson assumption is untenable: no simulated sample came close. For these $n = 13$ counts, $\sum x_i = 1840$ and $\sum x_i^2 = 336{,}878$, so

$$\bar{x} = \frac{1840}{13} = 141.54, \qquad s^2 = \frac{336{,}878 - 13\,\bar{x}^2}{12} = 6370.60,$$

a variance-to-mean ratio of $s^2/\bar{x} = 45.0$ where the Poisson demands 1. Following Example 5.6.6, simulate $M = 10{,}000$ samples of size 13 from a Poisson(141.54) and compute S_i^2 for each. The resulting distribution is centred exactly where it should be, $\overline{S^2} = 140.9 \approx \lambda$, with

quantile	5%	50%	95%
S^2	61.6	132.9	244.3

and

$$P_\lambda\left(S^2 \geq 6370.60\right) \approx \frac{1}{M}\sum_{i=1}^{M} I\left(S_i^2 \geq 6370.60\right) = \frac{0}{10{,}000}.$$

The observed s^2 exceeds the largest of 10,000 simulated values by more than an order of magnitude, so Poisson counts are rejected.

(b) Matching the negative binomial variance to s^2 at $\mu = \bar{x}$ gives

$$\hat{r} = \frac{\bar{x}^2}{s^2 - \bar{x}} = \frac{(141.54)^2}{6370.60 - 141.54} = 3.22,$$

and the simulation confirms that this and its neighbours are consistent. Setting $p = r/(r + \bar{x})$ so the mean stays at $\bar{x}$, draw $M = 20{,}000$ samples of size 13 from negative binomial(r, p) and record S^2:

r	$\mu + \mu^2/r$	5% of S^2	median	95% of S^2	$P(S^2 \geq s^2)$
1	20,175	4,662	15,819	49,438	0.890
2	10,158	3,179	8,661	22,200	0.697
3	6,819	2,354	5,948	14,081	0.445
3.2	6,402	2,231	5,627	13,073	0.407
4	5,150	1,882	4,569	10,226	0.256
5	4,148	1,583	3,683	7,959	0.124
6	3,480	1,342	3,131	6,720	0.064
6.3	3,321	1,306	3,020	6,366	0.050
8	2,646	1,050	2,429	5,001	0.012
10	2,145	864	1,982	4,007	0.003

Reading off the last column, $s^2 = 6370.60$ falls in neither 5% tail of the simulated S^2 distribution exactly when

$$r \lesssim 6.3,$$

$r = 6.3$ being where $P(S^2 \geq s^2)$ drops through 0.05; $r = 8$ and beyond are rejected. Below $r = 1$ the tail probability peaks near 0.93 (at $r \approx \frac{1}{2}$) and never reaches 0.95, so S^2 alone never contradicts heavy clumping. Any r in $(0, 6.3]$ is thus consistent, and the moment estimate $\hat{r} = 3.22$ sits comfortably inside: there the simulated median of S^2 is 5627 and the central 90% interval [2231, 13073] covers 6370.60.

Problem 5.55. Suppose the method of (5.6.7) is used to generate the random variable Y, where $y_i = i$, $i = 0, 1, 2, \ldots$. Show that the expected number of comparisons is $E(Y + 1)$. (Hint: See Exercise 2.14.)

Solution. The number of comparisons is $N = Y + 1$ pointwise: algorithm (5.6.7) tests "$U \leq F_Y(0)$?", then "$U \leq F_Y(1)$?", and so on, stopping at the first success and returning that index, so returning $Y = y$ costs exactly $y + 1$ comparisons.
For the route the hint intends, note that the algorithm reaches its $(k+1)$st comparison only if the first k failed, that is only if $U > F_Y(k-1)$, which happens exactly when $Y \geq k$. Hence $P(N > k) = P(Y \geq k) = P(Y > k - 1)$ for $k \geq 1$, and $P(N > 0) = 1$. Both N and Y are nonnegative-integer valued, so Exercise 2.14(b) applies to each:

$$\begin{aligned} EN &= \sum_{k=0}^{\infty} P(N > k) = 1 + \sum_{k=1}^{\infty} P(Y > k - 1) \\ &= 1 + \sum_{j=0}^{\infty} P(Y > j) = 1 + EY = E(Y + 1), \end{aligned}$$

the first and last equalities being that exercise applied to N and to Y.

Problem 5.56. Let Y have the Cauchy distribution, $f_Y(y) = \frac{1}{1+y^2}$, $-\infty < y < \infty$.
(a) Show that $F_Y(y) = \tan^{-1}(y)$.
(b) Show how to simulate a Cauchy(a, b) random variable starting from a uniform$(0, 1)$ random variable.
(See Exercise 2.12 for a related result.)

Solution. (a) Errata: the printed pdf and cdf both drop a factor $1/\pi$ — as printed f_Y integrates to π and $\tan^{-1}$ is negative on $y < 0$, so neither is a legitimate density or cdf; the intended Cauchy$(0, 1)$ statement, and the one part (b) needs, is $f_Y(y) = \frac{1}{\pi}\frac{1}{1+y^2}$ with

$$F_Y(y) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(y),$$

which follows from $\frac{d}{dt}\tan^{-1}(t) = \frac{1}{1+t^2}$:

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^{y} \frac{1}{\pi}\frac{1}{1+t^2}\,dt = \frac{1}{\pi}\Big[\tan^{-1}(t)\Big]_{-\infty}^{y} \\ &= \frac{1}{\pi}\left(\tan^{-1}(y) + \frac{\pi}{2}\right) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(y). \end{aligned}$$

(b) $X = a + b\tan\big(\pi\big(U - \frac{1}{2}\big)\big)$ with $U \sim$ uniform$(0, 1)$.
The Cauchy(a, b) pdf is $f_X(x) = \frac{1}{\pi b}\left[1 + \left(\frac{x-a}{b}\right)^2\right]^{-1}$, so $X = a + bY$ and, by part (a),

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{x-a}{b}\right), \qquad F_X^{-1}(u) = a + b\tan\big(\pi\big(u - \tfrac{1}{2}\big)\big),$$

this inverse being genuine since F_X is continuous and strictly increasing on $\mathbb{R}$. By the Probability Integral Transform (Theorem 2.1.10), $F_X^{-1}(U) \sim F_X$, which is the displayed recipe.
Method (2): πU is uniform on $(0, \pi)$ and tan has period π, so $\tan(\pi U)$ has the same distribution as $\tan\big(\pi(U - \frac{1}{2})\big)$, and $X = a + b\tan(\pi U)$ serves equally well.

5.9 Exercises 5.57–5.63

Problem 5.57. Park et al. (1996) describe a method for generating correlated binary variables based on the following scheme. Let X_1, X_2, X_3 be independent Poisson random variables with means $\lambda_1, \lambda_2, \lambda_3$, respectively, and create the random variables

$$Y_1 = X_1 + X_3 \qquad \text{and} \qquad Y_2 = X_2 + X_3.$$

(a) Show that $\mathrm{Cov}(Y_1, Y_2) = \lambda_3$.
(b) Define $Z_i = I(Y_i = 0)$ and $p_i = e^{-(\lambda_i + \lambda_3)}$. Show that the Z_i are Bernoulli(p_i) with

$$\mathrm{Corr}(Z_1, Z_2) = \frac{p_1 p_2 \left(e^{\lambda_3} - 1\right)}{\sqrt{p_1(1-p_1)}\,\sqrt{p_2(1-p_2)}}.$$

(c) Show that the correlation of Z_1 and Z_2 is not unrestricted in the range $[-1, 1]$, but

$$\mathrm{Corr}(Z_1, Z_2) \le \min\left\{ \sqrt{\frac{p_2(1-p_1)}{p_1(1-p_2)}},\ \sqrt{\frac{p_1(1-p_2)}{p_2(1-p_1)}} \right\}.$$

Solution. (a) The shared summand carries all the covariance:

$$\begin{aligned} \mathrm{Cov}(Y_1, Y_2) &= \mathrm{Cov}(X_1 + X_3,\, X_2 + X_3) \\ &= \mathrm{Var}(X_3) = \lambda_3, \end{aligned}$$

the three cross terms vanishing by independence (Theorem 4.5.5) and $\mathrm{Var}(X_3) = \lambda_3$ for a Poisson(λ_3) variable.
(b) Z_i is an indicator, hence Bernoulli with success probability

$$P(Y_i = 0) = P(X_i = 0)P(X_3 = 0) = e^{-\lambda_i} e^{-\lambda_3} = p_i,$$

using independence and $P(X = 0) = e^{-\lambda}$ for $X \sim$ Poisson(λ). For the product, $Z_1 Z_2 = 1$ exactly when all three Poissons vanish, so

$$\begin{aligned} E(Z_1 Z_2) &= P(X_1 = 0, X_2 = 0, X_3 = 0) \\ &= e^{-(\lambda_1 + \lambda_2 + \lambda_3)} = p_1 p_2 e^{\lambda_3}, \end{aligned}$$

since $p_1 p_2 = e^{-(\lambda_1 + \lambda_2 + 2\lambda_3)}$. Therefore $\mathrm{Cov}(Z_1, Z_2) = p_1 p_2 e^{\lambda_3} - p_1 p_2 = p_1 p_2 (e^{\lambda_3} - 1)$, and dividing by $\sqrt{\mathrm{Var}\, Z_1 \,\mathrm{Var}\, Z_2} = \sqrt{p_1(1-p_1)}\sqrt{p_2(1-p_2)}$ gives the stated correlation.
(c) $Z_1 Z_2 \le Z_i$ pointwise, so $E(Z_1 Z_2) \le \min\{p_1, p_2\}$ and hence

$$\mathrm{Cov}(Z_1, Z_2) \le \min\{p_1, p_2\} - p_1 p_2.$$

(i) If $p_1 \le p_2$ the bound is $p_1(1 - p_2)$ and

$$\mathrm{Corr}(Z_1, Z_2) \le \frac{p_1(1-p_2)}{\sqrt{p_1(1-p_1)p_2(1-p_2)}} = \sqrt{\frac{p_1(1-p_2)}{p_2(1-p_1)}}.$$

(ii) If $p_2 \le p_1$ the same computation gives the reciprocal expression $\sqrt{p_2(1-p_1)/[p_1(1-p_2)]}$.
The two candidate bounds are reciprocals of one another, and $p_1 \le p_2 \iff p_1(1-p_2) \le p_2(1-p_1)$, so in each case the bound produced is the smaller of the two, i.e. the minimum. Since that minimum is ≤ 1, with equality only when $p_1 = p_2$, the attainable correlations are strictly confined inside $[-1, 1]$.

Problem 5.58. Suppose that $U_1, U_2, \ldots, U_n$ are iid uniform$(0,1)$ random variables, and let $S_n = \sum_{i=1}^n U_i$. Define the random variable N by

$$N = \min\{k : S_k > 1\}.$$

(a) Show that $P(S_k \le t) = t^k/k!$ if $0 \le t \le 1$.
(b) Show that $P(N = n) = P(S_{n-1} < 1) - P(S_n < 1)$ and, surprisingly, that $E(N) = e$, the base of the natural logarithms.
(c) Use the result of part (b) to calculate the value of e by simulation.
(d) How large should n be so that you are 95% confident that you have the first four digits of e correct?
(Russell 1991, who attributes the problem to Gnedenko 1978, describes such a simulation experiment.)

Solution. (a) Induction on k. For $k = 1$, $P(U_1 \le t) = t$. If the claim holds for k, then conditioning on U_{k+1} and using $0 \le t \le 1$,

$$\begin{aligned} P(S_{k+1} \le t) &= \int_0^1 P(S_k \le t - u)\, du \\ &= \int_0^t \frac{(t-u)^k}{k!}\, du = \frac{t^{k+1}}{(k+1)!}. \end{aligned}$$

(b) Since the U_i are positive, $\{S_n < 1\} \subset \{S_{n-1} < 1\}$, and $\{N = n\} = \{S_{n-1} < 1\} \setminus \{S_n < 1\}$, so

$$P(N = n) = P(S_{n-1} < 1) - P(S_n < 1) = \frac{1}{(n-1)!} - \frac{1}{n!},$$

by part (a) at $t = 1$ (the S_k are continuous, so $<$ and $\le$ agree). The same identity in tail form reads $P(N > n) = P(S_n < 1) = 1/n!$ for $n \ge 0$, whence by Exercise 2.14(b) applied to the nonnegative integer variable N,

$$E(N) = \sum_{n=0}^{\infty} P(N > n) = \sum_{n=0}^{\infty} \frac{1}{n!} = e.$$

(c) Repeat independently m times: draw uniform$(0,1)$ variables one at a time, accumulating the sum, and record the index N_j at which the running sum first exceeds 1. By the WLLN (Theorem 5.5.2), $\bar{N} = m^{-1} \sum_{j=1}^m N_j \to E(N) = e$ in probability, so $\bar{N}$ estimates e.
(d) About $n = 1.2 \times 10^7$ replications. From part (b),

$$\begin{aligned} E[N(N-1)] &= \sum_{n \ge 2} \left(\frac{n}{(n-2)!} - \frac{1}{(n-2)!} \right) \\ &= (e + 2e) - e = 2e, \end{aligned}$$

so $EN^2 = 2e + e = 3e$ and

$$\mathrm{Var}(N) = 3e - e^2 = 0.7658, \qquad \sigma = 0.8751.$$

Getting the first four digits right means $|\bar{N} - e| < 0.0005$, and by the CLT (Theorem 5.5.14) this holds with probability 0.95 when $1.96\,\sigma/\sqrt{n} \le 0.0005$, i.e.

$$n \ge \left(\frac{1.96 \times 0.8751}{0.0005} \right)^2 = 1.177 \times 10^7.$$

Problem 5.59. Prove that the algorithm of Example 5.6.7 generates a beta(a, b) random variable. (That algorithm, for f_Y the beta(a, b) pdf and any constant $c \geq \max_y f_Y(y)$, is: (a) generate (U, V) independent uniform$(0, 1)$; (b) if $U < \frac{1}{c} f_Y(V)$, set $Y = V$; otherwise return to step (a).)

Solution. The accepted V has cdf F_Y. A finite $c \geq \max_y f_Y(y)$ exists exactly when the beta pdf is bounded, i.e. $a \geq 1$ and $b \geq 1$; then $f_Y(v)/c \leq 1$, and $y = 1$ in (5.6.10) gives the per-trial acceptance probability

$$q = P\big(U < \tfrac{1}{c} f_Y(V)\big) = \frac{1}{c} > 0,$$

so the pass count $T = \min\{i : U_i < \frac{1}{c} f_Y(V_i)\}$ at which the algorithm stops is geometric(q), finite with probability one, and $Y = V_T$. The pairs (U_i, V_i) drawn on successive passes are iid, so $\{T = i\}$ splits into $i - 1$ rejections and acceptance on the ith pair, and

$$\begin{aligned} P(Y \leq y) &= \sum_{i=1}^{\infty} (1-q)^{i-1} P\big(V \leq y,\, U < \tfrac{1}{c} f_Y(V)\big) \\ &= \frac{1}{q} \cdot \frac{F_Y(y)}{c} = F_Y(y), \end{aligned}$$

the inner probability being $F_Y(y)/c$ by (5.6.10).

Problem 5.60. Generalize the algorithm of Example 5.6.7 to apply to any bounded pdf; that is, for an arbitrary bounded pdf $f(x)$ on $[a, b]$, define $c = \max_{a \leq x \leq b} f(x)$. Let X and Y be independent, with $X \sim$ uniform(a, b) and $Y \sim$ uniform$(0, c)$. Let d be a number greater than b, and define a new random variable

$$W = \begin{cases} X & \text{if } Y < f(X) \\ d & \text{if } Y \geq f(X). \end{cases}$$

(a) Show that $P(W \leq w) = \int_a^w f(t)\, dt / [c(b-a)]$ for $a \leq w \leq b$.
(b) Using part (a), explain how a random variable with pdf $f(x)$ can be generated. (Hint: Use a geometric argument; a picture will help.)

Solution. (a) For $a \leq w \leq b$ the event $\{W \leq w\}$ is exactly $\{X \leq w,\, Y < f(X)\}$ (since $d > b \geq w$), so conditioning on X and using $P(Y < f(x)) = f(x)/c$,

$$\begin{aligned} P(W \leq w) &= \int_a^w P\big(Y < f(x)\big)\, \frac{dx}{b-a} \\ &= \int_a^w \frac{f(x)}{c} \cdot \frac{dx}{b-a} = \frac{\int_a^w f(t)\, dt}{c(b-a)}. \end{aligned}$$

(b) Generate independent pairs (X, Y) until one satisfies $Y < f(X)$, and output that X; its pdf is f. Indeed, taking $w = b$ in part (a),

$$P(W \leq b) = \frac{1}{c(b-a)} = P\big(Y < f(X)\big),$$

so for $a \leq w \leq b$,

$$P(W \leq w \mid W \neq d) = \frac{\int_a^w f(t)\, dt / [c(b-a)]}{1/[c(b-a)]} = \int_a^w f(t)\, dt,$$

which is the cdf belonging to f; independence across pairs turns this conditional law into the law of the first accepted X, exactly as in Exercise 5.59. Geometrically, (X, Y) is uniform on the rectangle $[a, b] \times [0, c]$, whose area is $c(b-a)$; acceptance means the point lands in the region under the curve $y = f(x)$, of area 1, and a uniform point conditioned to lie in that region has horizontal coordinate with density f.

Problem 5.61. (a) Suppose it is desired to generate $Y \sim \text{beta}(a, b)$, where a and b are not integers. Show that using $V \sim \text{beta}([a], [b])$ will result in a finite value of $M = \sup_y f_Y(y)/f_V(y)$.
(b) Suppose it is desired to generate $Y \sim \text{gamma}(a, b)$, where a and b are not integers. Show that using $V \sim \text{gamma}([a], b)$ will result in a finite value of $M = \sup_y f_Y(y)/f_V(y)$.
(c) Show that, in each of parts (a) and (b), if V had parameter $[a] + 1$, then M would be infinite.
(d) In each of parts (a) and (b) find optimal values for the parameters of V in the sense of minimizing $E(N)$ (see (5.6.12)).
(Recall that $[a]$ = greatest integer $\le a$.)

Solution. Write $\alpha = [a]$, $\beta = [b]$, $r = a - \alpha \in (0, 1)$ and $s = b - \beta \in (0, 1)$; both $a, b > 1$ so that $\alpha, \beta \ge 1$ and the candidate densities exist. Two elementary suprema are used below: $\sup_{0 \le y \le 1} y^r(1-y)^s = r^r s^s/(r+s)^{r+s}$ at $y = r/(r+s)$, and $\sup_{y>0} y^r e^{-\lambda y} = (r/\lambda)^r e^{-r}$ at $y = r/\lambda$.
(a) The ratio is a bounded power function:

$$\frac{f_Y(y)}{f_V(y)} = \frac{B(\alpha, \beta)}{B(a, b)}\, y^{a-\alpha}(1-y)^{b-\beta}, \qquad 0 < y < 1,$$

and since $a - \alpha = r > 0$ and $b - \beta = s > 0$ both exponents are positive, so the ratio extends continuously to $[0, 1]$ with

$$M = \frac{B(\alpha, \beta)}{B(a, b)} \cdot \frac{r^r s^s}{(r+s)^{r+s}} < \infty.$$

(b) As printed this is false and the scale must be changed: with the common scale b the exponential factors cancel and

$$\frac{f_Y(y)}{f_V(y)} = \frac{\Gamma(\alpha)\, b^\alpha}{\Gamma(a)\, b^a}\, y^{a-\alpha} \xrightarrow[y\to\infty]{} \infty,$$

so $M = \infty$. The intended statement is $V \sim \text{gamma}([a], \beta_0)$ with a scale $\beta_0 > b$ (the candidate must have the heavier tail), and then

$$\frac{f_Y(y)}{f_V(y)} = \frac{\Gamma(\alpha)\beta_0^\alpha}{\Gamma(a) b^a}\, y^r e^{-\lambda y}, \qquad \lambda = \frac{1}{b} - \frac{1}{\beta_0} > 0,$$

whence

$$M = \frac{\Gamma(\alpha)\beta_0^\alpha}{\Gamma(a) b^a} \left(\frac{r}{\lambda}\right)^r e^{-r} < \infty.$$

(c) Replacing α by $\alpha + 1$ turns the exponent of y into $a - \alpha - 1 = r - 1 \in (-1, 0)$, and in both cases the remaining factor tends to a positive constant as $y \to 0$:

$$\frac{f_Y(y)}{f_V(y)} \sim \text{const} \cdot y^{\,r-1} \xrightarrow[y\to 0^+]{} \infty.$$

(In the beta case $\beta + 1$ fails the same way at $y \to 1^-$, the exponent of $1 - y$ becoming $s - 1 < 0$.) So $M = \infty$: the candidate must have both parameters no larger than the target's.
(d) By the remark following Theorem 5.6.8, the number N of candidate pairs needed per accepted variate is geometric$(1/M)$, so $E(N) = M$ and optimizing means minimizing M over the admissible candidates.
Gamma. Here the shape must be an integer $\alpha \le a$ (so that V is a sum of α iid exponentials, generable directly by (5.6.5)), while the scale $\beta_0 > b$ is a free real parameter. Substituting $\lambda = (\beta_0 - b)/(b\beta_0)$ into (b) and using $\alpha + r = a$,

$$M(\beta_0) = \frac{\Gamma(\alpha) r^r e^{-r} b^{\,r-a}}{\Gamma(a)} \cdot \frac{\beta_0^a}{(\beta_0 - b)^r},$$

and $\frac{d}{d\beta_0}\left[a \log \beta_0 - r \log(\beta_0 - b)\right] = 0$ gives

$$\frac{a}{\beta_0} = \frac{r}{\beta_0 - b} \implies \beta_0^* = \frac{ab}{a - r} = \frac{ab}{\alpha},$$

a minimum since the derivative changes sign from $-$ to $+$. Note $\alpha\beta_0^* = ab$: the optimal candidate has exactly the target's mean. Back-substituting,

$$M^*(\alpha) = \frac{\Gamma(\alpha)}{\Gamma(a)} \cdot \frac{a^a}{\alpha^\alpha}\, e^{-(a-\alpha)},$$

which is free of b and decreasing in α, because

$$\frac{M^*(\alpha+1)}{M^*(\alpha)} = \frac{e}{(1+1/\alpha)^{\alpha+1}} < 1$$

(the sequence $(1+1/\alpha)^{\alpha+1}$ decreases to e from above). Hence take the largest admissible shape: $V \sim$ gamma($[a]$, $ab/[a]$), with

$$E(N) = \frac{\Gamma([a])\, a^a}{\Gamma(a)\, [a]^{[a]}}\, e^{-(a-[a])}.$$

For $a = 2.7$ this is $\beta_0^* = 1.35\, b$ and $E(N) = 1.174$.
Beta. Both parameters must be integers with $\alpha \le a$, $\beta \le b$ by part (c), so the admissible set is the finite grid $1 \le \alpha \le [a]$, $1 \le \beta \le [b]$, and one minimizes the closed form of part (a),

$$M(\alpha,\beta) = \frac{B(\alpha,\beta)}{B(a,b)} \cdot \frac{r^r s^s}{(r+s)^{r+s}}, \quad r = a-\alpha,\ s = b-\beta,$$

over that grid; there is no closed-form minimizer, and $([a],[b])$ is generally not it. For $a = 2.7$, $b = 6.3$ the grid of $M(\alpha,\beta)$ values is

$\alpha\backslash\beta$	1	2	3	4	5	6
1	2.67	1.81	1.75	2.11	3.32	9.26
2	7.45	2.85	1.69	1.27	1.18	1.67

so the optimum is $V \sim$ beta$(2,5)$ with $E(N) = 1.18$ – better than the beta$(2,6)$ of Example 5.6.9 ($E(N) = 1.67$) and than the uniform = beta$(1,1)$ of Example 5.6.7 ($E(N) = c = 2.67$).

Problem 5.62. Find the values of M so that an Accept/Reject Algorithm can generate $Y \sim$ n$(0,1)$ using $U \sim$ uniform$(0,1)$ and
(a) $V \sim$ Cauchy.
(b) $V \sim$ double exponential.
(c) Compare the algorithms. Which one do you recommend?

Solution. (a) $M = \sqrt{2\pi}\, e^{-1/2} = 1.520$. With $f_V(v) = [\pi(1+v^2)]^{-1}$,

$$\frac{f_Y(y)}{f_V(y)} = \sqrt{\frac{\pi}{2}}\,(1+y^2)\, e^{-y^2/2},$$

and writing $t = y^2 \ge 0$, $\frac{d}{dt}\left[(1+t)e^{-t/2}\right] = \frac{1-t}{2}e^{-t/2}$ vanishes at $t = 1$, a maximum, with value $2e^{-1/2}$. Hence $M = 2\sqrt{\pi/2}\, e^{-1/2} = \sqrt{2\pi}\, e^{-1/2} = 1.520$.
(b) $M = \sqrt{2/\pi}\, e^{1/2} = 1.316$. With $f_V(v) = \frac{1}{2}e^{-|v|}$,

$$\frac{f_Y(y)}{f_V(y)} = \sqrt{\frac{2}{\pi}}\, e^{\,|y|-y^2/2},$$

and $|y| - y^2/2$ is maximized at $|y| = 1$ with value $1/2$, so $M = \sqrt{2/\pi}\, e^{1/2} = 1.316$.
Both suprema are finite because each candidate has heavier tails than the normal, as Theorem 5.6.8 requires.
(c) The double exponential. Its M is smaller, so by the remark after Theorem 5.6.8 it needs $E(N) = 1.32$ candidate pairs per normal variate against 1.52 for the Cauchy, a 13% saving; and a signed exponential costs only a logarithm of a uniform, while a Cauchy costs the transcendental $\tan(\pi(U-1/2))$ of Exercise 5.56.

Problem 5.63. For generating $Y \sim \mathrm{n}(0,1)$ using an Accept/Reject Algorithm, we could generate $U \sim$ uniform, $V \sim$ exponential(λ), and attach a random sign to V ($\pm$ each with equal probability). What value of λ will optimize this algorithm?

Solution. $\lambda = 1$, giving $M = \sqrt{2/\pi}\, e^{1/2} = 1.316$. The signed exponential has density $f_V(v) = \frac{1}{2\lambda} e^{-|v|/\lambda}$ on $(-\infty, \infty)$, so

$$\frac{f_Y(y)}{f_V(y)} = \frac{2\lambda}{\sqrt{2\pi}}\, \exp\left\{ \frac{|y|}{\lambda} - \frac{y^2}{2} \right\},$$

whose exponent is maximized at $|y| = 1/\lambda$ with value $1/(2\lambda^2)$. Hence

$$M(\lambda) = \lambda \sqrt{\frac{2}{\pi}}\, e^{1/(2\lambda^2)}, \qquad \frac{d}{d\lambda} \log M = \frac{1}{\lambda} - \frac{1}{\lambda^3},$$

which vanishes only at $\lambda = 1$ and is negative for $\lambda < 1$, positive for $\lambda > 1$, so $\lambda = 1$ minimizes $E(N) = M$, at $M(1) = \sqrt{2/\pi}\, e^{1/2} = 1.316$.

5.10 Exercises 5.64–5.69

Problem 5.64. A technique similar to Accept/Reject is *importance sampling*, which is quite useful for calculating features of a distribution. Suppose that $X \sim f$, but the pdf f is difficult to simulate from. Generate $Y_1, Y_2, \ldots, Y_m$, iid from g, and, for any function h, calculate

$$\frac{1}{m} \sum_{i=1}^{m} \frac{f(Y_i)}{g(Y_i)}\, h(Y_i).$$

We assume that the supports of f and g are the same and $\operatorname{Var} h(X) < \infty$.

(a) Show that $E\left(\frac{1}{m} \sum_{i=1}^{m} \frac{f(Y_i)}{g(Y_i)} h(Y_i) \right) = Eh(X)$.

(b) Show that $\frac{1}{m} \sum_{i=1}^{m} \frac{f(Y_i)}{g(Y_i)} h(Y_i) \to Eh(X)$ in probability.

(c) Although the estimator of part (a) has the correct expectation, in practice the estimator

$$\sum_{i=1}^{m} \left(\frac{f(Y_i)/g(Y_i)}{\sum_{j=1}^{m} f(Y_j)/g(Y_j)} \right) h(Y_i)$$

is preferred. Show that this estimator converges in probability to $Eh(X)$. Moreover, show that if h is constant, this estimator is superior to the one in part (a). (Casella and Robert 1996 further explore the properties of this estimator.)

Solution. Write $W_i = \frac{f(Y_i)}{g(Y_i)} h(Y_i)$ and $w_i = \frac{f(Y_i)}{g(Y_i)}$; the whole exercise is the single change-of-density identity $E\big[\phi(Y) f(Y)/g(Y)\big] = E\phi(X)$.

(a) Since f and g have the same support, $g > 0$ wherever $f > 0$ and

$$\begin{aligned} EW_1 &= \int \frac{f(y)}{g(y)} h(y)\, g(y)\, dy \\ &= \int h(y) f(y)\, dy = Eh(X), \end{aligned}$$

and the W_i are identically distributed, so the average has this same mean.

(b) The W_i are iid with finite mean $Eh(X)$ by (a), so the WLLN (Theorem 5.5.2, in the version

requiring only a finite mean, noted in the remark following that theorem) gives

$$\frac{1}{m}\sum_{i=1}^{m} W_i \longrightarrow Eh(X) \quad \text{in probability.}$$

(c) Divide numerator and denominator by m:

$$\sum_{i=1}^{m}\left(\frac{w_i}{\sum_j w_j}\right) h(Y_i) = \frac{\frac{1}{m}\sum_{i=1}^{m} w_i h(Y_i)}{\frac{1}{m}\sum_{j=1}^{m} w_j}.$$

The numerator is exactly the estimator of (b), so it converges in probability to $Eh(X)$. The denominator is an average of iid variables with

$$Ew_1 = \int \frac{f(y)}{g(y)}\, g(y)\, dy = \int f(y)\, dy = 1,$$

so it converges in probability to 1. By Slutsky's Theorem (Theorem 5.5.17) the ratio converges in probability to $Eh(X)/1 = Eh(X)$.
If $h \equiv c$, the estimator of (c) is

$$\sum_{i=1}^{m} \frac{w_i}{\sum_j w_j} \cdot c = c$$

identically: it estimates $Eh(X) = c$ with zero error. The estimator of (a) is $c\,\overline{w}$, with

$$\operatorname{Var}\left(c\,\overline{w}\right) = \frac{c^2}{m}\operatorname{Var}(w_1) > 0$$

unless $f = g$ a.e. So (c) has MSE 0 and (a) has strictly positive MSE.

Problem 5.65. A variation of the importance sampling algorithm of Exercise 5.64 can actually produce an approximate sample from f. Again let $X \sim f$ and generate $Y_1, Y_2, \ldots, Y_m$, iid from g. Calculate

$$q_i = \frac{f(Y_i)/g(Y_i)}{\sum_{j=1}^{m} f(Y_j)/g(Y_j)}.$$

Then generate random variables X^* from the discrete distribution on $Y_1, Y_2, \ldots, Y_m$, where $P(X^* = Y_k) = q_k$. Show that $X_1^*, X_2^*, \ldots, X_r^*$ is approximately a random sample from f.
(Hint: Show that $P(X^* \le x) = \sum_{i=1}^{m} q_i I(Y_i \le x)$, let $m \to \infty$, and use the WLLN in the numerator and denominator.)
This algorithm is called the Sampling/Importance Resampling (SIR) algorithm by Rubin (1988) and is referred to as the weighted bootstrap by Smith and Gelfand (1992).

Solution. Conditionally on the Ys, the cdf of X^* is the self-normalized importance-sampling estimator of Exercise 5.64(c) applied to $h(y) = I(y \le x)$, and therefore converges to $F_X(x)$.
Write $w_i = f(Y_i)/g(Y_i)$. Since X^* is supported on $\{Y_1, \ldots, Y_m\}$ with masses q_i,

$$\widehat{F}_m(x) := P\big(X^* \le x \mid Y_1, \ldots, Y_m\big) = \frac{\frac{1}{m}\sum_{i=1}^{m} w_i I(Y_i \le x)}{\frac{1}{m}\sum_{j=1}^{m} w_j}.$$

The numerator is an average of iid variables with mean

$$E\left[\frac{f(Y)}{g(Y)} I(Y \le x)\right] = \int_{-\infty}^{x} \frac{f(y)}{g(y)}\, g(y)\, dy = F_X(x),$$

and the denominator is an average of iid variables with mean $\int f = 1$; both are finite, so the WLLN (Theorem 5.5.2) and Slutsky's Theorem (Theorem 5.5.17) give, for each fixed x,

$$\widehat{F}_m(x) \longrightarrow F_X(x) \quad \text{in probability as } m \to \infty.$$

Given the Ys, $X_1^*, \dots, X_r^*$ are drawn independently from this discrete distribution, so for fixed $x_1, \dots, x_r$,

$$P\big(X_1^* \le x_1, \dots, X_r^* \le x_r\big) = E\left[\prod_{k=1}^{r} \widehat{F}_m(x_k)\right].$$

The integrand lies in $[0,1]$ and converges in probability to $\prod_k F_X(x_k)$ (Theorem 5.5.4, h = product), so its expectation converges to $\prod_{k=1}^r F_X(x_k)$ by bounded convergence. The joint cdf of $X_1^*, \dots, X_r^*$ therefore tends to that of r iid observations from f, which for finite m they are only approximately, being drawn from the common pool $Y_1, \dots, Y_m$.

Problem 5.66. If $X_1, \dots, X_n$ are iid $\mathrm{n}(\mu, \sigma^2)$, we have seen that the distribution of the sample mean $\bar{X}$ is $\mathrm{n}(\mu, \sigma^2/n)$. If we are interested in using a more robust estimator of location, such as the median (5.4.1), it becomes a more difficult task to derive its distribution. Recall that (5.4.1) defines the sample median as

$$M = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \big(X_{(n/2)} + X_{(n/2+1)}\big)/2 & \text{if } n \text{ is even.} \end{cases}$$

(a) Show that M is the median of the X_is if and only if $(M-\mu)/\sigma$ is the median of $(X_i - \mu)/\sigma$. Thus, we only need consider the distribution of the median from a $\mathrm{n}(0,1)$ sample.
(b) For a sample of size $n = 15$ from a $\mathrm{n}(0,1)$, simulate the distribution of the median M.
(c) Compare the distribution in part (b) to the asymptotic distribution of the median $\sqrt{n}(M-\mu) \sim \mathrm{n}\big[0, 1/4f^2(0)\big]$, where f is the pdf. Is $n = 15$ large enough for the asymptotics to be valid?

Solution. (a) The map $x \mapsto (x-\mu)/\sigma$ is strictly increasing, so it preserves the ordering: $Z_{(i)} = (X_{(i)} - \mu)/\sigma$ for every i. Applying it to (5.4.1),

$$\begin{aligned} n \text{ odd}: &\quad \frac{X_{((n+1)/2)} - \mu}{\sigma} = Z_{((n+1)/2)}, \\ n \text{ even}: &\quad \frac{\frac{1}{2}\big(X_{(n/2)} + X_{(n/2+1)}\big) - \mu}{\sigma} = \tfrac{1}{2}\big(Z_{(n/2)} + Z_{(n/2+1)}\big), \end{aligned}$$

which is exactly the median of the $Z_i = (X_i - \mu)/\sigma$. Since $Z_i \sim \mathrm{n}(0,1)$, the distribution of M is that of $\mu + \sigma M_Z$ with M_Z the median of a standard normal sample.
(b) With $n = 15$ (odd), $M = X_{(8)}$, so by Theorem 5.4.4 its exact pdf is

$$f_M(m) = \frac{15!}{7!\,7!}\,\Phi(m)^7\big(1-\Phi(m)\big)^7\varphi(m),$$

symmetric about 0. A simulation of 200,000 samples of size 15 gave

$$\bar{M} = 0.0012, \qquad s_M^2 = 0.1010, \qquad s_M = 0.3178,$$

matching the exact moments obtained by integrating f_M: $EM = 0$ and $\operatorname{Var} M = 0.10169$, $\mathrm{sd}(M) = 0.31890$.
(c) Here $f = \varphi$, $\varphi(0) = 1/\sqrt{2\pi}$, so the asymptotic variance is

$$\frac{1}{n} \cdot \frac{1}{4\varphi^2(0)} = \frac{1}{n} \cdot \frac{2\pi}{4} = \frac{\pi}{2n} = \frac{\pi}{30} = 0.10472,$$

i.e. $M \approx \mathrm{n}(0, 0.10472)$, sd 0.32360. The asymptotic variance overstates the exact one by only $0.10472/0.10169 = 1.030$, i.e. 3%, and the quantiles agree closely:

q	exact q-quantile of M	$\mathrm{n}(0, \pi/30)$ quantile
0.75	0.2147	0.2183
0.90	0.4083	0.4147
0.95	0.5244	0.5323
0.99	0.7430	0.7528

(exact quantiles from $M = \Phi^{-1}(B)$, $B \sim \mathrm{beta}(8,8)$, since $\Phi(X_{(8)})$ is the 8th order statistic of 15 uniforms). Every discrepancy is under 1.5%: yes, $n = 15$ is already large enough here.

Problem 5.67. In many instances the Metropolis Algorithm is the algorithm of choice because either (i) there are no obvious candidate densities that satisfy the Accept/Reject supremum condition, or (ii) the supremum condition is difficult to verify, or (iii) laziness leads us to substitute computing power for brain power.
For each of the following situations show how to implement the Metropolis Algorithm to generate a sample of size 100 from the specified distribution.
(a) $X \sim \frac{1}{\sigma} f[(x-\mu)/\sigma]$, f = Student's t with ν degrees of freedom, ν, μ, and σ known
(b) $X \sim \text{lognormal}(\mu, \sigma^2)$, μ, σ^2 known
(c) $X \sim \text{Weibull}(\alpha, \beta)$, α, β known
Here the Metropolis Algorithm (Section 5.6.3) is: let $Y \sim f_Y$ (target) and $V \sim f_V$ (candidate) have common support. Generate $V \sim f_V$ and set $Z_0 = V$. For $i = 1, 2, \ldots$: generate $U_i \sim \text{uniform}(0,1)$ and $V_i \sim f_V$, compute

$$\rho_i = \min\left\{ \frac{f_Y(V_i)}{f_V(V_i)} \cdot \frac{f_V(Z_{i-1})}{f_Y(Z_{i-1})},\, 1 \right\},$$

and set $Z_i = V_i$ if $U_i \le \rho_i$, $Z_i = Z_{i-1}$ if $U_i > \rho_i$.

Solution. In every case pick a candidate f_V with the same support as the target that is trivial to generate, and note that

$$\rho_i = \min\left\{ \frac{r(V_i)}{r(Z_{i-1})},\, 1 \right\}, \qquad r(x) := \frac{f_Y(x)}{f_V(x)},$$

depends on f_Y and f_V only up to multiplicative constants, so every normalizing constant may be dropped from r. Run the chain for $i = 1, \ldots, 1100$ and report $Z_{1001}, \ldots, Z_{1100}$: by the convergence statement at the end of Section 5.6.3, $Z_i \to Y$ in distribution, so after a burn-in of 1000 the retained 100 values behave like a sample from f_Y. (No supremum condition is needed; common support is all the algorithm requires.)
(a) Target $f_Y(x) = \frac{1}{\sigma} f\big((x-\mu)/\sigma\big)$ with f the t_ν pdf; support $\mathbb{R}$. Take the candidate $V \sim \text{n}(\mu, \sigma^2)$ (same support, and generated by Example 5.6.4). With $t(x) = (x-\mu)/\sigma$, all constants cancelling,

$$r(x) \propto \left(1 + \frac{t(x)^2}{\nu}\right)^{-(\nu+1)/2} e^{t(x)^2/2},$$

so with $t_i = (V_i - \mu)/\sigma$ and $s_{i-1} = (Z_{i-1} - \mu)/\sigma$,

$$\rho_i = \min\left\{ \left(\frac{1 + t_i^2/\nu}{1 + s_{i-1}^2/\nu} \right)^{-(\nu+1)/2} e^{(t_i^2 - s_{i-1}^2)/2},\; 1 \right\}.$$

This is exactly case (i): $\sup_x r(x) = \infty$, since the t tails are polynomial and the normal tails Gaussian, so Accept/Reject with a normal candidate is unavailable while Metropolis remains valid.
(b) Target $f_Y(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\{-(\log x - \mu)^2/(2\sigma^2)\}$ on $(0, \infty)$. Take $V \sim \text{exponential}(\beta_0)$ in the book's scale parameterization, $f_V(x) = \beta_0^{-1} e^{-x/\beta_0}$, with $\beta_0 = EX = e^{\mu + \sigma^2/2}$; it has the same support $(0, \infty)$ and Example 5.6.3 generates it as $-\beta_0 \log(1-U)$. Then

$$r(x) \propto \frac{1}{x} \exp\left\{ -\frac{(\log x - \mu)^2}{2\sigma^2} + \frac{x}{\beta_0} \right\},$$

$$\rho_i = \min\left\{ \frac{Z_{i-1}}{V_i} \exp\left\{ \frac{(\log Z_{i-1} - \mu)^2 - (\log V_i - \mu)^2}{2\sigma^2} + \frac{V_i - Z_{i-1}}{\beta_0} \right\},\, 1 \right\}.$$

(c) Target $f_Y(x) = \frac{\alpha}{\beta} x^{\alpha-1} e^{-x^\alpha/\beta}$ on $(0, \infty)$. Again take $V \sim \text{exponential}(\beta_0)$, say $\beta_0 = \beta^{1/\alpha}$, matching the scale of the target. Then

$$r(x) \propto x^{\alpha-1} \exp\left\{ -\frac{x^\alpha}{\beta} + \frac{x}{\beta_0} \right\},$$

$$\rho_i = \min\left\{ \left(\frac{V_i}{Z_{i-1}} \right)^{\alpha-1} \exp\left\{ \frac{Z_{i-1}^\alpha - V_i^\alpha}{\beta} + \frac{V_i - Z_{i-1}}{\beta_0} \right\},\, 1 \right\}.$$

Problem 5.68. If we use the Metropolis Algorithm rather than the Accept/Reject Algorithm we are freed from verifying the supremum condition of the Accept/Reject Algorithm. Of course, we give up the property of getting random variables with exactly the distribution we want and must settle for an approximation.
(a) Show how to use the Metropolis Algorithm to generate a random variable with an approximate Student's t distribution with ν degrees of freedom, starting from $\mathrm{n}(0,1)$ random variables.
(b) Show how to use the Accept/Reject Algorithm to generate a random variable with a Student's t distribution with ν degrees of freedom, starting from Cauchy random variables.
(c) Show how to use transformations to generate directly a random variable with a Student's t distribution with ν degrees of freedom.
(d) For $\nu = 2, 10, 25$ compare the methods by generating a sample of size 100. Which method do you prefer? Why?

Solution. Throughout, f_Y is the t_ν pdf (5.3.6),

$$f_Y(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \left(1 + \frac{y^2}{\nu}\right)^{-(\nu+1)/2}.$$

(a) Take $f_V = \mathrm{n}(0,1)$ (common support $\mathbb{R}$) and run the Metropolis Algorithm of Section 5.6.3 with

$$r(x) = \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} e^{x^2/2}, \qquad \rho_i = \min\left\{\frac{r(V_i)}{r(Z_{i-1})}, 1\right\},$$

the constants having cancelled from f_Y/f_V. Then $Z_i \to t_\nu$ in distribution. Accept/Reject is unavailable here because $\sup_x r(x) = \infty$.
(b) Here the supremum is finite. With $f_V(y) = 1/[\pi(1+y^2)]$ the Cauchy pdf,

$$\frac{f_Y(y)}{f_V(y)} = \frac{\sqrt{\pi}\,\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu}\,\Gamma\left(\frac{\nu}{2}\right)} h(y^2), \qquad h(u) = (1+u)\left(1 + \frac{u}{\nu}\right)^{-(\nu+1)/2}.$$

Differentiating,

$$h'(u) = \left(1 + \frac{u}{\nu}\right)^{-(\nu+3)/2} \frac{1-\nu}{2\nu}\,(u-1),$$

which for $\nu > 1$ is positive on $u < 1$ and negative on $u > 1$; hence h is maximized at $u = 1$ with $h(1) = 2\left(\frac{\nu}{\nu+1}\right)^{(\nu+1)/2}$, and

$$M = \sup_y \frac{f_Y(y)}{f_V(y)} = \frac{2\sqrt{\pi}\,\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\nu}{\nu+1}\right)^{(\nu+1)/2} < \infty$$

(at $\nu = 1$ this gives $M = 1$, as it must). Theorem 5.6.8 then applies: generate $U \sim \mathrm{uniform}(0,1)$ and $V \sim$ Cauchy (by $V = \tan(\pi(U' - \frac{1}{2}))$, Example 5.6.3); if $U < \frac{1}{M} f_Y(V)/f_V(V)$ set $Y = V$, else repeat. The output is exactly t_ν, and $EN = M$.
(c) By the representation preceding Definition 5.3.4, if $U \sim \mathrm{n}(0,1)$ and $W \sim \chi^2_\nu$ are independent then $U/\sqrt{W/\nu} \sim t_\nu$. Generate $Z_0, Z_1, \ldots, Z_\nu$ iid $\mathrm{n}(0,1)$ by Box-Muller (Example 5.6.4) and set $W = \sum_{j=1}^{\nu} Z_j^2$, which is χ^2_ν by Lemma 5.3.2 and is independent of Z_0. Then

$$T = \frac{Z_0}{\sqrt{W/\nu}} = \frac{\sqrt{\nu}\,Z_0}{\left(\sum_{j=1}^{\nu} Z_j^2\right)^{1/2}} \sim t_\nu.$$

(For even ν one may instead take $W = -2\sum_{j=1}^{\nu/2} \log U_j$, since $\chi^2_\nu = \mathrm{gamma}(\nu/2, 2)$.)
(d) The theoretical cost of (b) is $EN = M$:

ν	$M = EN$	acceptance rate $1/M$
2	1.2092	0.827
10	1.4474	0.691
25	1.4904	0.671

so Accept/Reject needs on average under 1.5 Cauchy draws per t variate, at any ν. Samples of size 100 from the Metropolis chain of (a) (burn-in 1000) had acceptance rates 0.91, 0.96, 0.99 for $\nu = 2, 10, 25$, but only 94, 99, 99 distinct values, since the chain repeats on rejection. Its real failure is in the tail, which is precisely where $\sup f_Y/f_V = \infty$: over six runs of length 300,000,

$$\begin{aligned} \nu = 2: &\quad \widehat{P}(|T| > 3) \in [0.048,\, 0.081], &&\text{true } 0.0955,\\ \nu = 10: &\quad \widehat{P}(|T| > 3) = 0.0132, &&\text{true } 0.0133,\\ \nu = 25: &\quad \widehat{P}(|T| > 3) = 0.0059, &&\text{true } 0.0060. \end{aligned}$$

At $\nu = 2$ the normal candidate is far lighter-tailed than the target, the chain visits the tail rarely and sticks when it does, and the estimate is both biased low and badly seed-dependent; by $\nu = 25$ the two densities are close and the problem disappears.
Prefer (c) at small ν and (b) at large ν. Both are exact and deliver independent variates; (c) needs no rejection loop and no supremum, but its cost of $\nu + 1$ normals (or 1 normal plus $\nu/2$ uniforms) grows with ν, while (b) stays under 1.5 Cauchy draws per variate at every ν once M has been found. Method (a) is last in all three cases: it is only approximate, its output is dependent with repeated values, and it errs exactly in the tail one usually wants a t for.

Problem 5.69. (Mengersen and Tweedie 1996 show that the convergence of the Metropolis Algorithm is much faster if the supremum condition is satisfied, that is, if $\sup f/g \le M < \infty$, where f is the target density and g is the candidate density.)
Show that the pdf $f_Y(y)$ is a *stable point* of the Metropolis Algorithm. That is, if $Z_i \sim f_Y(y)$, then $Z_{i+1} \sim f_Y(y)$.
Recall the algorithm of Section 5.6.3: f_Y (target) and f_V (candidate) have common support; given Z_i, generate $U_{i+1} \sim \text{uniform}(0,1)$ and $V_{i+1} \sim f_V$, put

$$\rho(Z_i, V_{i+1}) = \min\left\{ \frac{f_Y(V_{i+1})}{f_V(V_{i+1})} \cdot \frac{f_V(Z_i)}{f_Y(Z_i)},\, 1 \right\},$$

and set $Z_{i+1} = V_{i+1}$ if $U_{i+1} \le \rho(Z_i, V_{i+1})$ and $Z_{i+1} = Z_i$ otherwise.

Solution. The acceptance probability satisfies detailed balance, and stability is the one-line consequence. Write $w(x) = f_Y(x)/f_V(x)$, so that $\rho(z,v) = \min\{w(v)/w(z),\, 1\}$. Then

$$\begin{aligned} f_Y(z) f_V(v)\, \rho(z,v) &= f_V(z) f_V(v)\, w(z) \min\left\{ \frac{w(v)}{w(z)}, 1 \right\} \\ &= f_V(z) f_V(v) \min\{w(v),\, w(z)\}, \end{aligned}$$

which is symmetric in (z, v); hence

$$f_Y(z) f_V(v)\, \rho(z,v) = f_Y(v) f_V(z)\, \rho(v,z) \qquad \text{for all } z, v.$$

Now suppose $Z_i \sim f_Y$ and fix a set A. Splitting on accept/reject and conditioning on $Z_i = z$,

$$\begin{aligned} P(Z_{i+1} \in A) = &\iint_A f_Y(z) f_V(v) \rho(z,v)\, dv\, dz \\ &+ \int_A f_Y(z) \left(1 - \int f_V(v) \rho(z,v)\, dv \right) dz. \end{aligned}$$

Apply detailed balance to the first term and interchange the order of integration (Fubini; the integrand is nonnegative):

$$\int_A \int f_Y(v) f_V(z) \rho(v,z)\, dz\, dv = \int_A f_Y(v) \left(\int f_V(z) \rho(v,z)\, dz \right) dv.$$

Renaming z as v in the second term and adding,

$$P(Z_{i+1} \in A) = \int_A f_Y(v)\Big[c(v) + 1 - c(v)\Big]\, dv = \int_A f_Y(v)\, dv,$$

where $c(v) = \int f_V(z)\rho(v,z)\, dz \in [0,1]$. So $Z_{i+1} \sim f_Y$.

6 Principles of Data Reduction

Contents

6.1 Exercises 6.1–6.7 196
6.2 Exercises 6.8–6.14 199
6.3 Exercises 6.15–6.21 203
6.4 Exercises 6.22–6.28 207
6.5 Exercises 6.29–6.35 213
6.6 Exercises 6.36–6.42 219
6.7 Exercises 6.43–6.43 223

6.1 Exercises 6.1–6.7

Problem 6.1. Let X be one observation from a $\mathrm{n}(0, \sigma^2)$ population. Is $|X|$ a sufficient statistic?

Solution. Yes. The pdf of the single observation is

$$f(x \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-x^2/(2\sigma^2)} = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-|x|^2/(2\sigma^2)},$$

which depends on x only through $|x|$. So with

$$g\big(t \mid \sigma^2\big) = (2\pi\sigma^2)^{-1/2} e^{-t^2/(2\sigma^2)}, \qquad h(x) = 1,$$

we have $f(x \mid \sigma^2) = g(|x| \mid \sigma^2)h(x)$ for every x and every $\sigma^2 > 0$, and the Factorization Theorem (Theorem 6.2.6) makes $|X|$ sufficient for σ^2.

Problem 6.2. Let $X_1, \ldots, X_n$ be independent random variables with densities

$$f_{X_i}(x \mid \theta) = \begin{cases} e^{i\theta - x} & x \geq i\theta, \\ 0 & x < i\theta. \end{cases}$$

Prove that $T = \min_i(X_i/i)$ is a sufficient statistic for θ.

Solution. Factor the joint density. Writing $I(\cdot)$ for the indicator,

$$\begin{aligned} f(\mathbf{x} \mid \theta) &= \prod_{i=1}^{n} e^{i\theta - x_i}\, I(x_i \geq i\theta) \\ &= \exp\Big(\theta \sum_{i=1}^{n} i - \sum_{i=1}^{n} x_i\Big) \prod_{i=1}^{n} I\Big(\frac{x_i}{i} \geq \theta\Big) \\ &= e^{\theta n(n+1)/2}\, I\Big(\min_{1 \leq i \leq n} \frac{x_i}{i} \geq \theta\Big) \cdot e^{-\sum_i x_i}, \end{aligned}$$

since $x_i \geq i\theta \iff x_i/i \geq \theta$ (each $i \geq 1 > 0$), and all n of those inequalities hold exactly when the smallest of the ratios does.
The first bracket is a function of θ and of the data only through $T(\mathbf{x}) = \min_i(x_i/i)$, and the second, $h(\mathbf{x}) = e^{-\sum_i x_i}$, is free of θ. By the Factorization Theorem (Theorem 6.2.6), $T = \min_i(X_i/i)$ is sufficient for θ.

Problem 6.3. Let $X_1, \ldots, X_n$ be a random sample from the pdf

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma} e^{-(x-\mu)/\sigma}, \qquad \mu < x < \infty, \quad 0 < \sigma < \infty.$$

Find a two-dimensional sufficient statistic for (μ, σ).

Solution. $T(\mathbf{X}) = \big(X_{(1)}, \sum_{i=1}^{n} X_i\big)$, where $X_{(1)} = \min_i X_i$. Indeed

$$\begin{aligned} f(\mathbf{x} \mid \mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sigma} e^{-(x_i - \mu)/\sigma}\, I(x_i > \mu) \\ &= \frac{1}{\sigma^n} \exp\Big(-\frac{\sum_i x_i - n\mu}{\sigma}\Big)\, I\big(x_{(1)} > \mu\big), \end{aligned}$$

because all n constraints $x_i > \mu$ hold exactly when the smallest observation exceeds μ. The right-hand side depends on $\mathbf{x}$ only through $\big(x_{(1)}, \sum_i x_i\big)$, so taking $h(\mathbf{x}) = 1$ and

$$g(t_1, t_2 \mid \mu, \sigma) = \sigma^{-n} e^{-(t_2 - n\mu)/\sigma}\, I(t_1 > \mu)$$

the Factorization Theorem (Theorem 6.2.6) gives sufficiency.

Problem 6.4. Prove Theorem 6.2.10:
Let $X_1, \ldots, X_n$ be iid observations from a pdf or pmf $f(x \mid \boldsymbol{\theta})$ that belongs to an exponential family given by

$$f(x \mid \boldsymbol{\theta}) = h(x) c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x) \right),$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$, $d \le k$. Then

$$T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j) \right)$$

is a sufficient statistic for $\boldsymbol{\theta}$.

Solution. Multiply the n densities and collect the exponents:

$$\begin{aligned} f(\mathbf{x} \mid \boldsymbol{\theta}) &= \prod_{j=1}^{n} h(x_j) c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x_j) \right) \\ &= \left(\prod_{j=1}^{n} h(x_j) \right) c(\boldsymbol{\theta})^n \exp\left(\sum_{j=1}^{n} \sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x_j) \right) \\ &= \left(\prod_{j=1}^{n} h(x_j) \right) c(\boldsymbol{\theta})^n \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) \sum_{j=1}^{n} t_i(x_j) \right), \end{aligned}$$

the last step being the interchange of two finite sums.
Now set $H(\mathbf{x}) = \prod_{j=1}^{n} h(x_j)$, which does not involve $\boldsymbol{\theta}$, and for $\mathbf{t} = (t_1, \ldots, t_k) \in \mathbb{R}^k$

$$g(\mathbf{t} \mid \boldsymbol{\theta}) = c(\boldsymbol{\theta})^n \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})\, t_i \right).$$

Then $f(\mathbf{x} \mid \boldsymbol{\theta}) = g\big(T(\mathbf{x}) \mid \boldsymbol{\theta}\big) H(\mathbf{x})$ for all $\mathbf{x}$ and all $\boldsymbol{\theta}$, with $T(\mathbf{x}) = \big(\sum_j t_1(x_j), \ldots, \sum_j t_k(x_j)\big)$, so the Factorization Theorem (Theorem 6.2.6) makes $T(\mathbf{X})$ sufficient for $\boldsymbol{\theta}$. ■

Problem 6.5. Let $X_1, \ldots, X_n$ be independent random variables with pdfs

$$f(x_i \mid \theta) = \begin{cases} \dfrac{1}{2i\theta} & -i(\theta - 1) < x_i < i(\theta + 1) \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$. Find a two-dimensional sufficient statistic for θ.

Solution. $T(\mathbf{X}) = \big(\min_i (X_i/i),\ \max_i (X_i/i)\big)$.
The support constraint on x_i is a pair of lower bounds on θ: since $i > 0$,

$$\begin{aligned} x_i > -i(\theta - 1) &\iff i\theta > i - x_i \iff \theta > 1 - \frac{x_i}{i}, \\ x_i < i(\theta + 1) &\iff i\theta > x_i - i \iff \theta > \frac{x_i}{i} - 1. \end{aligned}$$

Imposing both for every i,

$$
\begin{aligned}
f(\mathbf{x} \mid \theta) &= \prod_{i=1}^{n} \frac{1}{2i\theta}\, I\big(-i(\theta-1) < x_i < i(\theta+1)\big) \\
&= \frac{1}{(2\theta)^n n!}\, I\Big(\theta > \max_i\Big(1 - \frac{x_i}{i}\Big)\Big)\, I\Big(\theta > \max_i\Big(\frac{x_i}{i} - 1\Big)\Big) \\
&= \frac{1}{(2\theta)^n n!}\, I\big(\theta > 1 - m(\mathbf{x})\big)\, I\big(\theta > M(\mathbf{x}) - 1\big),
\end{aligned}
$$

where $m(\mathbf{x}) = \min_i(x_i/i)$ and $M(\mathbf{x}) = \max_i(x_i/i)$, and $\prod_{i=1}^n i = n!$.
The data enter only through $\big(m(\mathbf{x}), M(\mathbf{x})\big)$, so with $h(\mathbf{x}) = 1/n!$ and

$$
g(t_1, t_2 \mid \theta) = (2\theta)^{-n} I(\theta > 1 - t_1)\, I(\theta > t_2 - 1)
$$

the Factorization Theorem (Theorem 6.2.6) gives that $T(\mathbf{X}) = \big(\min_i(X_i/i), \max_i(X_i/i)\big)$ is sufficient for θ.

Problem 6.6. Let $X_1, \ldots, X_n$ be a random sample from a gamma(α, β) population. Find a two-dimensional sufficient statistic for (α, β).

Solution. $T(\mathbf{X}) = \big(\sum_{i=1}^n \log X_i,\ \sum_{i=1}^n X_i\big)$.
The gamma pdf is a two-parameter exponential family in the sense of (3.4.2): for $x > 0$,

$$
f(x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} = \frac{1}{x} \cdot \frac{1}{\Gamma(\alpha)\beta^\alpha} \exp\left(\alpha \log x - \frac{1}{\beta} x\right),
$$

which is of the form $h(x)c(\alpha,\beta)\exp\big(w_1 t_1(x) + w_2 t_2(x)\big)$ with $h(x) = x^{-1} I(x > 0)$, $c(\alpha,\beta) = 1/(\Gamma(\alpha)\beta^\alpha)$, $w_1(\alpha,\beta) = \alpha$, $t_1(x) = \log x$, $w_2(\alpha,\beta) = -1/\beta$ and $t_2(x) = x$. Theorem 6.2.10 (proved in Exercise 6.4) therefore makes

$$
T(\mathbf{X}) = \left(\sum_{i=1}^n \log X_i,\ \sum_{i=1}^n X_i\right)
$$

sufficient for (α, β).

Problem 6.7. Let $f(x, y \mid \theta_1, \theta_2, \theta_3, \theta_4)$ be the bivariate pdf for the uniform distribution on the rectangle with lower left corner (θ_1, θ_2) and upper right corner (θ_3, θ_4) in $\mathbb{R}^2$; that is,

$$
f(x, y \mid \boldsymbol{\theta}) = \begin{cases} \dfrac{1}{(\theta_3 - \theta_1)(\theta_4 - \theta_2)} & \theta_1 \le x \le \theta_3,\ \theta_2 \le y \le \theta_4 \\ 0 & \text{otherwise.} \end{cases}
$$

The parameters satisfy $\theta_1 < \theta_3$ and $\theta_2 < \theta_4$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from this pdf. Find a four-dimensional sufficient statistic for $(\theta_1, \theta_2, \theta_3, \theta_4)$.

Solution.

$$
T = \big(X_{(1)},\, Y_{(1)},\, X_{(n)},\, Y_{(n)}\big) = \Big(\min_i X_i,\ \min_i Y_i,\ \max_i X_i,\ \max_i Y_i\Big).
$$

Writing $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$, the joint pdf of the sample is

$$
\begin{aligned}
f(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\theta}) &= \prod_{i=1}^{n} \frac{I(\theta_1 \le x_i \le \theta_3)\, I(\theta_2 \le y_i \le \theta_4)}{(\theta_3 - \theta_1)(\theta_4 - \theta_2)} \\
&= \frac{I\big(x_{(1)} \ge \theta_1\big) I\big(x_{(n)} \le \theta_3\big) I\big(y_{(1)} \ge \theta_2\big) I\big(y_{(n)} \le \theta_4\big)}{\big[(\theta_3 - \theta_1)(\theta_4 - \theta_2)\big]^n},
\end{aligned}
$$

because $\theta_1 \le x_i$ for every i exactly when $\theta_1 \le x_{(1)}$, $x_i \le \theta_3$ for every i exactly when $x_{(n)} \le \theta_3$, and likewise in the y coordinate.

The right-hand side involves the data only through $\big(x_{(1)}, y_{(1)}, x_{(n)}, y_{(n)}\big)$, so with $h(\mathbf{x}, \mathbf{y}) = 1$ and

$$g(t_1, t_2, t_3, t_4 \mid \boldsymbol{\theta}) = \frac{I(t_1 \geq \theta_1) I(t_2 \geq \theta_2) I(t_3 \leq \theta_3) I(t_4 \leq \theta_4)}{\big[(\theta_3 - \theta_1)(\theta_4 - \theta_2)\big]^n}$$

the Factorization Theorem (Theorem 6.2.6) shows that $T = \big(X_{(1)}, Y_{(1)}, X_{(n)}, Y_{(n)}\big)$ is a four-dimensional sufficient statistic for $(\theta_1, \theta_2, \theta_3, \theta_4)$.

6.2 Exercises 6.8–6.14

Problem 6.8. Let $X_1, \ldots, X_n$ be a random sample from a population with location pdf $f(x - \theta)$. Show that the order statistics, $T(X_1, \ldots, X_n) = (X_{(1)}, \ldots, X_{(n)})$, are a sufficient statistic for θ and no further reduction is possible.

Solution. The likelihood is a symmetric function of the sample, so it already factors through the order statistics: for every $\mathbf{x}$,

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f(x_i - \theta) = \prod_{i=1}^{n} f\big(x_{(i)} - \theta\big),$$

and Theorem 6.2.6 applies with $h(\mathbf{x}) = 1$ and $g(T(\mathbf{x}) \mid \theta)$ the right-hand product.
No further reduction is sufficient, f being unspecified. Let S be a function of the order statistics that is not one-to-one, say $S(\mathbf{x}) = S(\mathbf{y})$ for two ordered samples $\mathbf{x} \neq \mathbf{y}$, and take f Cauchy. For that location family the order statistics are minimal sufficient (Exercise 6.9(d), by the ratio criterion of Theorem 6.2.13), so if S were sufficient then T would be a function of S by Definition 6.2.11, giving $T(\mathbf{x}) = T(\mathbf{y})$, i.e. $\mathbf{x} = \mathbf{y}$ — a contradiction. Hence no reduction of the order statistics is sufficient across all location pdfs f.
For a particular f a strict reduction can exist, namely $\bar{X}$ for the normal 6.9(a) and $X_{(1)}$ for the location exponential 6.9(b), so the assertion is to be read exactly as proved above, with f left unspecified.

Problem 6.9. For each of the following distributions let $X_1, \ldots, X_n$ be a random sample. Find a minimal sufficient statistic for θ.
(a) $f(x \mid \theta) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}$, $-\infty < x < \infty$, $-\infty < \theta < \infty$ (normal)
(b) $f(x \mid \theta) = e^{-(x-\theta)}$, $\theta < x < \infty$, $-\infty < \theta < \infty$ (location exponential)
(c) $f(x \mid \theta) = \dfrac{e^{-(x-\theta)}}{\big(1 + e^{-(x-\theta)}\big)^2}$, $-\infty < x < \infty$, $-\infty < \theta < \infty$ (logistic)
(d) $f(x \mid \theta) = \dfrac{1}{\pi\big[1 + (x-\theta)^2\big]}$, $-\infty < x < \infty$, $-\infty < \theta < \infty$ (Cauchy)
(e) $f(x \mid \theta) = \frac{1}{2} e^{-|x-\theta|}$, $-\infty < x < \infty$, $-\infty < \theta < \infty$ (double exponential)

Solution. (a) $\bar{X}$; (b) $X_{(1)}$; (c), (d), (e) the full order statistic $(X_{(1)}, \ldots, X_{(n)})$. In each case apply Theorem 6.2.13: T is minimal sufficient as soon as $f(\mathbf{x} \mid \theta)/f(\mathbf{y} \mid \theta)$ is free of θ exactly when $T(\mathbf{x}) = T(\mathbf{y})$.
(a) Here

$$\frac{f(\mathbf{x} \mid \theta)}{f(\mathbf{y} \mid \theta)} = \exp\Big(-\tfrac{1}{2}\textstyle\sum_i x_i^2 + \tfrac{1}{2}\sum_i y_i^2 + \theta\big(\sum_i x_i - \sum_i y_i\big)\Big),$$

which is free of θ if and only if $\sum_i x_i = \sum_i y_i$.
(b) $f(\mathbf{x} \mid \theta) = e^{n\theta - \sum_i x_i} I_{(-\infty, x_{(1)})}(\theta)$, so

$$\frac{f(\mathbf{x} \mid \theta)}{f(\mathbf{y} \mid \theta)} = e^{\sum_i y_i - \sum_i x_i} \, \frac{I_{(-\infty, x_{(1)})}(\theta)}{I_{(-\infty, y_{(1)})}(\theta)}.$$

The two positivity regions must coincide for the ratio to be a (finite, nonzero) constant, i.e. $x_{(1)} = y_{(1)}$; and when they do the ratio is the constant $e^{\sum y_i - \sum x_i}$.

(c) Put $z = e^{\theta} > 0$, so that $1 + e^{-(x_i-\theta)} = 1 + z a_i$ with $a_i = e^{-x_i}$, and likewise $b_i = e^{-y_i}$. Then

$$\frac{f(\mathbf{x} \mid \theta)}{f(\mathbf{y} \mid \theta)} = e^{\sum_i y_i - \sum_i x_i} \left[\frac{Q(z)}{P(z)}\right]^2, \qquad P(z) = \prod_{i=1}^{n}(1 + z a_i), \;\; Q(z) = \prod_{i=1}^{n}(1 + z b_i).$$

If this is constant on $z > 0$ then $Q^2 = cP^2$ as polynomials; letting $z \downarrow 0$ gives $c = 1$, and since $P, Q > 0$ on $z > 0$ we get $Q \equiv P$. Equal polynomials have equal root multisets, $\{-1/a_i\} = \{-1/b_i\}$, hence $\{x_i\} = \{y_i\}$ as multisets.
(d) The ratio is

$$\frac{f(\mathbf{x} \mid \theta)}{f(\mathbf{y} \mid \theta)} = \frac{\prod_{j=1}^{n}\left[1 + (y_j - \theta)^2\right]}{\prod_{j=1}^{n}\left[1 + (x_j - \theta)^2\right]}.$$

Numerator and denominator are monic polynomials of degree $2n$ in θ; if the ratio equals C for all θ then comparing leading coefficients gives $C = 1$ and the two polynomials are identical. Their root multisets are $\{x_j \pm \sqrt{-1}\}$ and $\{y_j \pm \sqrt{-1}\}$, so $\{x_j\} = \{y_j\}$.
(e) Now

$$\frac{f(\mathbf{x} \mid \theta)}{f(\mathbf{y} \mid \theta)} = \exp\Big(\textstyle\sum_i |y_i - \theta| - \sum_i |x_i - \theta|\Big),$$

so the ratio is free of θ iff $\theta \mapsto \sum_i |x_i - \theta| - \sum_i |y_i - \theta|$ is constant. Each of $\sum_i |x_i - \theta|$ and $\sum_i |y_i - \theta|$ is piecewise linear in θ with slope $-n$ to the left of all the points, and with slope jumping by $+2$ (per tied point) at each $x_{(i)}$, resp. $y_{(i)}$. A constant difference forces identical jump locations and sizes, i.e. $x_{(i)} = y_{(i)}$ for all i.

Problem 6.10. Show that the minimal sufficient statistic for the uniform$(\theta, \theta+1)$, found in Example 6.2.15, is not complete. (In Example 6.2.15, $X_1, \ldots, X_n$ are iid uniform observations on the interval $(\theta, \theta + 1)$, $-\infty < \theta < \infty$, and $T(\mathbf{X}) = \big(X_{(1)}, X_{(n)}\big)$ is shown to be minimal sufficient.)

Solution. Take $g(T) = X_{(n)} - X_{(1)} - \dfrac{n-1}{n+1}$, with $n \geq 2$.
By Example 6.2.17 the range $R = X_{(n)} - X_{(1)}$ has pdf

$$h(r \mid \theta) = n(n-1) r^{n-2}(1 - r), \qquad 0 < r < 1,$$

for every θ — a beta$(n-1, 2)$ density, whose mean is $\dfrac{n-1}{(n-1)+2} = \dfrac{n-1}{n+1}$. Hence

$$\mathrm{E}_\theta\, g(T) = \mathrm{E}_\theta R - \frac{n-1}{n+1} = 0 \qquad \text{for all } \theta,$$

while R is a continuous random variable, so $P_\theta\big(g(T) = 0\big) = 0 \neq 1$. By Definition 6.2.21 the family of distributions of T is not complete.

Problem 6.11. Refer to the pdfs given in Exercise 6.9 (normal, location exponential, logistic, Cauchy and double exponential, each of the location form $f(x - \theta)$). For each, let $X_{(1)} < \cdots < X_{(n)}$ be the ordered sample, and define $Y_i = X_{(n)} - X_{(i)}$, $i = 1, \ldots, n-1$.
(a) For each of the pdfs in Exercise 6.9, verify that the set $(Y_1, \ldots, Y_{n-1})$ is ancillary for θ. Try to prove a general theorem, like Example 6.2.18, that handles all these families at once.
(b) In each case determine whether the set $(Y_1, \ldots, Y_{n-1})$ is independent of the minimal sufficient statistic.

Solution. (a) All five at once: in any location family the vector of order-statistic differences is ancillary. Let $X_1, \ldots, X_n$ be iid with cdf $F(x-\theta)$ and let $Z_1, \ldots, Z_n$ be iid F (the case $\theta = 0$), so that $X_i = Z_i + \theta$ in distribution jointly, by Theorem 3.5.6. The map $t \mapsto t + \theta$ is strictly increasing, hence order preserving,

so $X_{(i)} = Z_{(i)} + \theta$ for every i and

$$Y_i = X_{(n)} - X_{(i)} = \big(Z_{(n)} + \theta\big) - \big(Z_{(i)} + \theta\big) = Z_{(n)} - Z_{(i)}.$$

Therefore, for all $y_1, \dots, y_{n-1}$,

$$P_\theta\big(Y_1 \le y_1, \dots, Y_{n-1} \le y_{n-1}\big) = P\big(Z_{(n)} - Z_{(1)} \le y_1, \dots, Z_{(n)} - Z_{(n-1)} \le y_{n-1}\big),$$

which does not involve θ because the distribution of $Z_1, \dots, Z_n$ does not. By Definition 6.2.16 the set $(Y_1, \dots, Y_{n-1})$ is ancillary. Each of 6.9(a)–(e) is of the form $f(x - \theta)$ — in 6.9(b) with $f(z) = e^{-z} I_{(0,\infty)}(z)$, so the moving support is no obstacle — so all five are covered.
(b) Independent in (a) and (b); not independent in (c), (d), (e).
(i) Normal. The minimal sufficient statistic is $\bar{X}$ (Exercise 6.9(a)). The $\mathrm{n}(\theta, 1)$ family is an exponential family with $t(x) = x$ and $w(\theta) = \theta$ ranging over all of $\mathbb{R}$, which certainly contains an open set in $\mathbb{R}^1$, so $\sum_i X_i$ — equivalently $\bar{X}$ — is complete by Theorem 6.2.25. Being complete and minimal sufficient, $\bar{X}$ is independent of every ancillary statistic by Basu's Theorem 6.2.24, in particular of $(Y_1, \dots, Y_{n-1})$.
(ii) Location exponential. The minimal sufficient statistic is $X_{(1)}$ (Exercise 6.9(b)), whose pdf is $h(t \mid \theta) = n e^{-n(t-\theta)}$, $t > \theta$. It is complete: if $\mathrm{E}_\theta g\big(X_{(1)}\big) = 0$ for all θ then

$$0 = n e^{n\theta} \int_\theta^\infty g(t) e^{-nt}\, dt \implies \int_\theta^\infty g(t) e^{-nt}\, dt = 0 \ \ \forall \theta \implies g(\theta) e^{-n\theta} = 0 \ \ \forall \theta,$$

differentiating in θ; hence $g = 0$ except on a null set, i.e. $P_\theta\big(g(X_{(1)}) = 0\big) = 1$ for every θ. Basu's Theorem 6.2.24 again gives independence of $(Y_1, \dots, Y_{n-1})$ and $X_{(1)}$.
(iii) Logistic, Cauchy, double exponential. Here the minimal sufficient statistic is the full order statistic $T = \big(X_{(1)}, \dots, X_{(n)}\big)$ (Exercise 6.9(c)–(e)), and $(Y_1, \dots, Y_{n-1})$ is a function of T. A statistic independent of itself must be degenerate: $P(Y_1 \in A) = P(Y_1 \in A)^2$ for all A forces every such probability to be 0 or 1. But $Y_1 = X_{(n)} - X_{(1)}$ is a nondegenerate continuous random variable in each of these three models, so $(Y_1, \dots, Y_{n-1})$ is not independent of T. Basu's Theorem 6.2.24 is not contradicted, since T fails to be complete here: for any c with $0 < P(Y_1 \le c) < 1$ the bounded statistic $g(T) = I_{(-\infty, c]}(Y_1) - P(Y_1 \le c)$ has mean zero for every θ — the probability being free of θ by part (a) — without being identically zero.

Problem 6.12. A natural ancillary statistic in most problems is the *sample size*. For example, let N be a random variable taking values $1, 2, \dots$ with known probabilities $p_1, p_2, \dots$, where $\sum_i p_i = 1$. Having observed $N = n$, perform n Bernoulli trials with success probability θ, getting X successes.
(a) Prove that the pair (X, N) is minimal sufficient and N is ancillary for θ. (Note the similarity to some of the hierarchical models discussed in Section 4.4.)
(b) Prove that the estimator X/N is unbiased for θ and has variance $\theta(1 - \theta)\mathrm{E}(1/N)$.

Solution. (a) The raw data are N together with the trial outcomes $T_1, \dots, T_N \in \{0, 1\}$, with joint pmf

$$f\big(t_1, \dots, t_n, n \mid \theta\big) = p_n\, \theta^{\sum_j t_j} (1 - \theta)^{\,n - \sum_j t_j} = \underbrace{p_n}_{h} \ \underbrace{\theta^x (1 - \theta)^{n-x}}_{g((x,n)\mid\theta)}, \qquad x = \textstyle\sum_j t_j.$$

Since p_n is known and free of θ, the Factorization Theorem 6.2.6 shows (X, N) is sufficient.
For minimality use Theorem 6.2.13. Given two data points with values (x, n) and (y, m) of (X, N),

$$\frac{f(\mathbf{t}, n \mid \theta)}{f(\mathbf{s}, m \mid \theta)} = \frac{p_n}{p_m}\, \theta^{x-y}\, (1 - \theta)^{(n-x)-(m-y)}.$$

If $\theta^a (1 - \theta)^b$ is constant on $0 < \theta < 1$, then differentiating its logarithm gives $a/\theta - b/(1 - \theta) \equiv 0$, so $a = b = 0$. Thus the ratio is free of θ if and only if $x = y$ and $n - x = m - y$, i.e. if and only if $(x, n) = (y, m)$. Hence (X, N) is minimal sufficient.
N is ancillary because its marginal pmf is $P_\theta(N = n) = p_n$, which is known and does not involve θ (Definition 6.2.16).

(b) Given $N = n$, $X \sim \text{binomial}(n, \theta)$, so $\mathrm{E}(X \mid N) = N\theta$ and $\mathrm{Var}(X \mid N) = N\theta(1-\theta)$. By Theorem 4.4.3,

$$\mathrm{E}\Big(\frac{X}{N}\Big) = \mathrm{E}\Big[\mathrm{E}\Big(\frac{X}{N}\,\Big|\,N\Big)\Big] = \mathrm{E}\Big[\frac{N\theta}{N}\Big] = \theta,$$

so X/N is unbiased. By the conditional variance identity, Theorem 4.4.7,

$$\begin{aligned}\mathrm{Var}\Big(\frac{X}{N}\Big) &= \mathrm{E}\Big[\mathrm{Var}\Big(\frac{X}{N}\,\Big|\,N\Big)\Big] + \mathrm{Var}\Big[\mathrm{E}\Big(\frac{X}{N}\,\Big|\,N\Big)\Big] \\ &= \mathrm{E}\Big[\frac{N\theta(1-\theta)}{N^2}\Big] + \mathrm{Var}(\theta) = \theta(1-\theta)\,\mathrm{E}\Big(\frac{1}{N}\Big),\end{aligned}$$

the second term vanishing because θ is a constant; $\mathrm{E}(1/N)$ is finite since $0 < 1/N \le 1$.

Problem 6.13. Suppose X_1 and X_2 are iid observations from the pdf $f(x \mid \alpha) = \alpha x^{\alpha-1} e^{-x^\alpha}$, $x > 0$, $\alpha > 0$. Show that $(\log X_1)/(\log X_2)$ is an ancillary statistic.

Solution. Transform to the exponential scale: $Y = X^\alpha \sim \text{exponential}(1)$, free of α. Indeed the cdf of X is

$$F(x \mid \alpha) = \int_0^x \alpha t^{\alpha-1} e^{-t^\alpha}\,dt = 1 - e^{-x^\alpha}, \qquad x > 0,$$

so $P_\alpha(X^\alpha \le y) = P_\alpha\big(X \le y^{1/\alpha}\big) = 1 - e^{-y}$ for $y > 0$.
Writing $X_i = Y_i^{1/\alpha}$ with Y_1, Y_2 iid exponential(1) gives $\log X_i = \alpha^{-1} \log Y_i$, so that

$$\frac{\log X_1}{\log X_2} = \frac{\alpha^{-1}\log Y_1}{\alpha^{-1}\log Y_2} = \frac{\log Y_1}{\log Y_2},$$

whose distribution is that of a fixed function of (Y_1, Y_2) and hence does not involve α (the ratio is well defined a.s., since $P(\log Y_2 = 0) = P(Y_2 = 1) = 0$). By Definition 6.2.16 it is ancillary — this is Example 6.2.19 for the scale family formed by $\log X_1, \log X_2$, whose scale parameter is $1/\alpha$.

Problem 6.14. Let $X_1, \ldots, X_n$ be a random sample from a location family. Show that $M - \bar{X}$ is an ancillary statistic, where M is the sample median.

Solution. Both M and $\bar{X}$ shift with θ, so their difference does not.
Let the sample come from the cdf $F(x-\theta)$ and let $Z_1, \ldots, Z_n$ be iid F, so that $X_i = Z_i + \theta$ (Theorem 3.5.6). Since $t \mapsto t + \theta$ is increasing, $X_{(i)} = Z_{(i)} + \theta$ for every i, and therefore

$$M = \begin{cases} X_{((n+1)/2)} = Z_{((n+1)/2)} + \theta, & n \text{ odd}, \\ \dfrac{X_{(n/2)} + X_{(n/2+1)}}{2} = \dfrac{Z_{(n/2)} + Z_{(n/2+1)}}{2} + \theta, & n \text{ even}, \end{cases}$$

i.e. $M = M_Z + \theta$ where M_Z is the median of the Z_i. Also $\bar{X} = \bar{Z} + \theta$. Hence

$$M - \bar{X} = (M_Z + \theta) - (\bar{Z} + \theta) = M_Z - \bar{Z},$$

so for every t, $P_\theta(M - \bar{X} \le t) = P(M_Z - \bar{Z} \le t)$, which does not involve θ because the distribution of $Z_1, \ldots, Z_n$ does not. By Definition 6.2.16, $M - \bar{X}$ is ancillary.

6.3 Exercises 6.15–6.21

Problem 6.15. Let $X_1, \dots, X_n$ be iid $\mathrm{n}(\theta, a\theta^2)$, where a is a known constant and $\theta > 0$.
(a) Show that the parameter space does not contain a two-dimensional open set.
(b) Show that the statistic $T = (\bar{X}, S^2)$ is a sufficient statistic for θ, but the family of distributions is not complete.

Solution. (a) The natural parameter set is the parabola $w_1 = -\frac{a}{2}w_2^2$, which contains no open ball. Indeed, writing the pdf in the exponential-family form of Definition 3.4.1,

$$\begin{aligned} f(x \mid \theta) &= \frac{1}{\sqrt{2\pi a\theta^2}} \exp\!\left(-\frac{(x-\theta)^2}{2a\theta^2}\right) \\ &= \frac{e^{-1/(2a)}}{\sqrt{2\pi a}\,\theta} \exp\!\left(-\frac{1}{2a\theta^2}\,x^2 + \frac{1}{a\theta}\,x\right), \end{aligned}$$

so $w_1(\theta) = -1/(2a\theta^2)$ and $w_2(\theta) = 1/(a\theta)$ (necessarily $a > 0$, since $a\theta^2$ is a variance). Since $w_2(\theta)^2 = 1/(a^2\theta^2) = -2w_1(\theta)/a$, the parameter set is

$$\Big\{(w_1, w_2) : w_1 = -\tfrac{a}{2}w_2^2,\ w_2 > 0\Big\} \subset \mathbb{R}^2.$$

Every ball around a point (w_1, w_2) of this curve contains the point $(w_1 + \varepsilon, w_2)$, which is off the curve for all small $\varepsilon \neq 0$; hence no two-dimensional open set is contained in the parameter space. (This is a curved exponential family, Definition 3.4.6.)
(b) Sufficiency. The joint pdf is

$$f(\mathbf{x} \mid \theta) = \frac{e^{-n/(2a)}}{(2\pi a)^{n/2}\theta^n} \exp\!\left(-\frac{1}{2a\theta^2}\sum_i x_i^2 + \frac{1}{a\theta}\sum_i x_i\right),$$

which depends on $\mathbf{x}$ only through $\big(\sum x_i, \sum x_i^2\big)$, a one-to-one function of $(\bar{x}, s^2)$. The Factorization Theorem (Theorem 6.2.6) with $h(\mathbf{x}) = 1$ gives sufficiency of $T = (\bar{X}, S^2)$.
Incompleteness. From $\bar{X} \sim \mathrm{n}(\theta, a\theta^2/n)$ and Theorem 5.3.1,

$$\mathrm{E}_\theta \bar{X}^2 = \frac{a\theta^2}{n} + \theta^2, \qquad \mathrm{E}_\theta S^2 = a\theta^2.$$

Hence the nonzero function $g(T) = a\bar{X}^2 - \frac{n+a}{n}S^2$ satisfies, for every $\theta > 0$,

$$\mathrm{E}_\theta g(T) = a\theta^2\Big(\frac{a}{n} + 1\Big) - \frac{n+a}{n}\,a\theta^2 = 0,$$

while $P_\theta\big(g(T) \neq 0\big) = 1$ (the distribution of $(\bar{X}, S^2)$ is continuous and g vanishes only on a curve). So T is not complete.

Problem 6.16. A famous example in genetic modeling (Tanner, 1996 or Dempster, Laird, and Rubin 1977) is a genetic linkage multinomial model, where we observe the multinomial vector (x_1, x_2, x_3, x_4) with cell probabilities given by

$$\big(\tfrac{1}{2} + \tfrac{\theta}{4},\ \tfrac{1}{4}(1-\theta),\ \tfrac{1}{4}(1-\theta),\ \tfrac{\theta}{4}\big)\,.$$

(a) Show that this is a curved exponential family.
(b) Find a sufficient statistic for θ.
(c) Find a minimal sufficient statistic for θ.

Solution. (a) It is a two-parameter exponential family carrying the one-dimensional $\theta \in (0, 1)$, with natural parameters tied by $w_1 = \log(2 + 3e^{w_2})$. With $n = \sum_i x_i$ fixed and $p_1 = \frac{2+\theta}{4}$, $p_2 = p_3 = \frac{1-\theta}{4}$, $p_4 = \frac{\theta}{4}$,

$$f(\mathbf{x}\mid\theta)=\binom{n}{x_1\,x_2\,x_3\,x_4}\exp\Big(\sum_{i=1}^{4}x_i\log p_i\Big)$$
$$=\binom{n}{x_1\,x_2\,x_3\,x_4}\Big(\tfrac{1-\theta}{4}\Big)^n\exp\big(w_1(\theta)x_1+w_2(\theta)x_4\big),$$

where we used $\sum_i x_i\log p_i = n\log p_3+\sum_{i\neq 3}x_i\log(p_i/p_3)$ and $p_2=p_3$, and set

$$w_1(\theta)=\log\frac{2+\theta}{1-\theta},\qquad w_2(\theta)=\log\frac{\theta}{1-\theta}.$$

Here $k=2$ but $d=\dim\Theta=1$. Solving the second equation gives $\theta=e^{w_2}/(1+e^{w_2})$, whence

$$\frac{2+\theta}{1-\theta}=2+3e^{w_2},\qquad\text{i.e.}\qquad w_1=\log\big(2+3e^{w_2}\big).$$

So $\{(w_1(\theta),w_2(\theta)):0<\theta<1\}$ is a smooth curve in $\mathbb{R}^2$, of dimension $1<2=k$: this is a curved exponential family in the sense of Definition 3.4.6.
(b) $T(\mathbf{X})=(X_1,X_4)$. The display in (a) factors as $g(t_1,t_4\mid\theta)h(\mathbf{x})$ with h the multinomial coefficient, so the Factorization Theorem (Theorem 6.2.6) applies. (Note $X_2+X_3=n-X_1-X_4$ is then determined, but X_2 and X_3 separately are discarded.)
(c) (X_1,X_4) is already minimal. By Theorem 6.2.13 it suffices to show that

$$\frac{f(\mathbf{x}\mid\theta)}{f(\mathbf{y}\mid\theta)}=\frac{h(\mathbf{x})}{h(\mathbf{y})}\exp\big(a\,w_1(\theta)+b\,w_2(\theta)\big),\qquad a=x_1-y_1,\ b=x_4-y_4,$$

is free of θ only when $a=b=0$. As θ runs over $(0,1)$, $u=w_2(\theta)$ runs over all of $\mathbb{R}$, and by (a) the exponent is $a\log(2+3e^u)+bu$. If this equals a constant c for all u, then letting $u\to-\infty$ gives $a\log 2+bu\to c$, forcing $b=0$; then letting $u\to+\infty$ gives $au+a\log 3\to c$, forcing $a=0$. Hence $x_1=y_1$ and $x_4=y_4$, and (X_1,X_4) is a minimal sufficient statistic.

Problem 6.17. Let $X_1,\dots,X_n$ be iid with geometric distribution

$$P_\theta(X=x)=\theta(1-\theta)^{x-1},\qquad x=1,2,\dots,\quad 0<\theta<1.$$

Show that $\sum X_i$ is sufficient for θ, and find the family of distributions of $\sum X_i$. Is the family complete?

Solution. $\sum X_i$ is negative binomial, and the family is complete. The joint pmf is

$$f(\mathbf{x}\mid\theta)=\prod_{i=1}^{n}\theta(1-\theta)^{x_i-1}=\underbrace{\theta^n(1-\theta)^{\sum_i x_i-n}}_{g(\sum x_i\mid\theta)}\cdot\underbrace{\prod_{i=1}^{n}I_{\{1,2,\dots\}}(x_i)}_{h(\mathbf{x})},$$

so $T=\sum_i X_i$ is sufficient by the Factorization Theorem (Theorem 6.2.6).
Since T counts the trial on which the nth success occurs, T is negative binomial(n,θ) in the number-of-trials parametrization of Section 3.2:

$$P_\theta(T=t)=\binom{t-1}{n-1}\theta^n(1-\theta)^{t-n},\qquad t=n,n+1,n+2,\dots$$

This is a full one-parameter exponential family,

$$P_\theta(T=t)=\binom{t-1}{n-1}\theta^n\exp\big\{w(\theta)(t-n)\big\},\qquad w(\theta)=\log(1-\theta),$$

and as θ ranges over $(0,1)$ the natural parameter $w(\theta)$ ranges over all of $(-\infty,0)$, which contains an open set in $\mathbb{R}^1$. Theorem 6.2.25 therefore applies: the family of distributions of $\sum X_i$ is complete.

Problem 6.18. Let $X_1, \dots, X_n$ be iid Poisson(λ). Show that the family of distributions of $\sum X_i$ is complete. Prove completeness without using Theorem 6.2.25.

Solution. A power series that vanishes on an interval has all coefficients zero. By Theorem 4.3.2 (or Example 4.6.8), $T = \sum_i X_i \sim \text{Poisson}(n\lambda)$. Suppose g satisfies $\mathrm{E}_\lambda g(T) = 0$ for every $\lambda > 0$; the expectation existing means $\sum_t |g(t)|(n\lambda)^t/t! < \infty$ for every $\lambda > 0$, so the power series below converges for all λ. Then

$$0 = \mathrm{E}_\lambda g(T) = \sum_{t=0}^{\infty} g(t)\frac{(n\lambda)^t e^{-n\lambda}}{t!} \qquad \Longrightarrow \qquad \sum_{t=0}^{\infty} \frac{g(t)\, n^t}{t!}\,\lambda^t = 0$$

for all $\lambda > 0$, after multiplying by $e^{n\lambda} > 0$. The left side is a power series in λ with infinite radius of convergence which is identically 0 on $(0, \infty)$; by uniqueness of power-series coefficients, every coefficient vanishes:

$$\frac{g(t)\, n^t}{t!} = 0 \quad \text{for all } t = 0, 1, 2, \dots \qquad \Longrightarrow \qquad g(t) = 0 \ \text{ for all } t.$$

Hence $P_\lambda\big(g(T) = 0\big) = 1$ for every $\lambda > 0$, and the family $\{\text{Poisson}(n\lambda) : \lambda > 0\}$ is complete.

Problem 6.19. The random variable X takes the values 0, 1, 2 according to one of the following distributions:

	$P(X=0)$	$P(X=1)$	$P(X=2)$	range
Distribution 1	p	$3p$	$1-4p$	$0 < p < \frac{1}{4}$
Distribution 2	p	p^2	$1-p-p^2$	$0 < p < \frac{1}{2}$

In each case determine whether the family of distributions of X is complete.

Solution. Distribution 1 is not complete; Distribution 2 is.
Distribution 1. Take $g(0) = 3$, $g(1) = -1$, $g(2) = 0$. Then for every $p \in (0, \frac{1}{4})$,

$$\mathrm{E}_p g(X) = 3p + (-1)(3p) + 0 \cdot (1 - 4p) = 0,$$

yet $P_p\big(g(X) \neq 0\big) \geq P_p(X = 0) = p > 0$. So the family fails Definition 6.2.21.
Distribution 2. Suppose $\mathrm{E}_p g(X) = 0$ for all $p \in (0, \frac{1}{2})$. Collecting powers of p,

$$\begin{aligned} 0 &= g(0)p + g(1)p^2 + g(2)(1 - p - p^2) \\ &= g(2) + \big[g(0) - g(2)\big]p + \big[g(1) - g(2)\big]p^2. \end{aligned}$$

A polynomial vanishing on the interval $(0, \frac{1}{2})$ has all coefficients zero, so $g(2) = 0$, then $g(0) = g(2) = 0$ and $g(1) = g(2) = 0$. Thus $g \equiv 0$ on $\{0, 1, 2\}$ and the family is complete.

Problem 6.20. For each of the following pdfs let $X_1, \dots, X_n$ be iid observations. Find a complete sufficient statistic, or show that one does not exist.
(a) $f(x \mid \theta) = \dfrac{2x}{\theta^2}$, $0 < x < \theta$, $\theta > 0$
(b) $f(x \mid \theta) = \dfrac{\theta}{(1+x)^{1+\theta}}$, $0 < x < \infty$, $\theta > 0$
(c) $f(x \mid \theta) = \dfrac{(\log\theta)\theta^x}{\theta - 1}$, $0 < x < 1$, $\theta > 1$
(d) $f(x \mid \theta) = e^{-(x-\theta)} \exp\big(-e^{-(x-\theta)}\big)$, $-\infty < x < \infty$, $-\infty < \theta < \infty$
(e) $f(x \mid \theta) = \binom{2}{x}\theta^x(1-\theta)^{2-x}$, $x = 0, 1, 2$, $0 \leq \theta \leq 1$

Solution. Each family has one. In (b)-(e) the density is a full one-parameter exponential family, so Theorem 6.2.25 delivers the answer once the natural parameter set is seen to contain an open interval.
(a) $X_{(n)}$. The joint pdf is

$$f(\mathbf{x} \mid \theta) = \frac{2^n \prod_i x_i}{\theta^{2n}}\, I_{(0,\infty)}(x_{(1)})\, I_{[x_{(n)},\infty)}(\theta),$$

so $X_{(n)}$ is sufficient (Theorem 6.2.6). Its pdf is $f_{(n)}(t \mid \theta) = 2nt^{2n-1}/\theta^{2n}$ on $(0,\theta)$, by Theorem 5.4.4 applied to $F(x \mid \theta) = x^2/\theta^2$. If $\mathrm{E}_\theta g(X_{(n)}) = 0$ for all $\theta > 0$ then $\int_0^\theta g(t)t^{2n-1}\,dt = 0$ for all $\theta > 0$, and differentiating in θ (Fundamental Theorem of Calculus) gives $g(\theta)\theta^{2n-1} = 0$, i.e. $g(\theta) = 0$ for every $\theta > 0$. Complete.
(b) $T = \sum_i \log(1 + X_i)$. Here

$$f(\mathbf{x} \mid \theta) = \theta^n \Big(\prod_i \frac{1}{1+x_i}\Big) \exp\Big\{-\theta \sum_i \log(1+x_i)\Big\},$$

with $w(\theta) = -\theta$ sweeping the open set $(-\infty, 0)$ as θ ranges over $(0,\infty)$. Complete sufficient by Theorem 6.2.25.
(c) $T = \sum_i X_i$. Since

$$f(\mathbf{x} \mid \theta) = \Big(\frac{\log\theta}{\theta - 1}\Big)^n \exp\Big\{(\log\theta)\sum_i x_i\Big\}, \qquad 0 < x_i < 1,$$

the natural parameter $w(\theta) = \log\theta$ covers the open set $(0,\infty)$. Complete sufficient by Theorem 6.2.25.
(d) $T = \sum_i e^{-X_i}$. Multiplying the n densities,

$$f(\mathbf{x} \mid \theta) = e^{n\theta} e^{-\sum_i x_i} \exp\Big\{-e^\theta \sum_i e^{-x_i}\Big\},$$

an exponential family with $t(x) = e^{-x}$ and $w(\theta) = -e^\theta$, which ranges over the open set $(-\infty, 0)$. Complete sufficient by Theorem 6.2.25.
(e) $T = \sum_i X_i \sim \text{binomial}(2n, \theta)$. Writing

$$f(\mathbf{x} \mid \theta) = \Big(\prod_i \binom{2}{x_i}\Big)(1-\theta)^{2n} \exp\Big\{\log\frac{\theta}{1-\theta}\sum_i x_i\Big\}, \qquad 0 < \theta < 1,$$

the natural parameter $\log\big(\theta/(1-\theta)\big)$ covers all of $\mathbb{R}$, so Theorem 6.2.25 gives completeness over $0 < \theta < 1$. Every $t \in \{0, \ldots, 2n\}$ has positive probability there, so $g \equiv 0$ on the whole range, which also settles the endpoints $\theta = 0, 1$.

Problem 6.21. Let X be one observation from the pdf

$$f(x \mid \theta) = \Big(\frac{\theta}{2}\Big)^{|x|} (1-\theta)^{1-|x|}, \qquad x = -1, 0, 1, \quad 0 \le \theta \le 1.$$

(a) Is X a complete sufficient statistic?
(b) Is $|X|$ a complete sufficient statistic?
(c) Does $f(x \mid \theta)$ belong to the exponential class?

Solution. (a) No. X is trivially sufficient, but take $g(-1) = -1$, $g(0) = 0$, $g(1) = 1$. Since $P_\theta(X = \pm 1) = \theta/2$ and $P_\theta(X = 0) = 1 - \theta$,

$$\mathrm{E}_\theta g(X) = (-1)\frac{\theta}{2} + 0 \cdot (1-\theta) + (1)\frac{\theta}{2} = 0 \qquad \text{for every } \theta \in [0,1],$$

yet $P_\theta\big(g(X) \neq 0\big) = \theta > 0$ whenever $\theta > 0$. So Definition 6.2.21 fails.
(b) Yes. Sufficiency is immediate from the Factorization Theorem (Theorem 6.2.6): $f(x \mid \theta) = g(|x| \mid \theta)h(x)$ with $g(t \mid \theta) = (\theta/2)^t(1-\theta)^{1-t}$ and $h(x) = 1$ on $\{-1, 0, 1\}$. And $|X| \sim \text{Bernoulli}(\theta)$, since $P_\theta(|X| = 1) = \theta/2 + \theta/2 = \theta$. If $\mathrm{E}_\theta g(|X|) = 0$ for all θ, then

$$0 = g(0)(1-\theta) + g(1)\theta = g(0) + \big[g(1) - g(0)\big]\theta,$$

a polynomial in θ vanishing on $[0,1]$, so $g(0)=0$ and then $g(1)=0$. Hence $g\equiv 0$ and $|X|$ is a complete sufficient statistic.
(c) Yes, for $0<\theta<1$: with $t(x)=|x|$,

$$f(x\mid\theta)=\underbrace{I_{\{-1,0,1\}}(x)}_{h(x)}\,\underbrace{(1-\theta)}_{c(\theta)}\exp\Big\{\underbrace{\log\frac{\theta}{2(1-\theta)}}_{w(\theta)}\,|x|\Big\},$$

which is the form of Definition 3.4.1 with $k=1$. The endpoints must be dropped: at $\theta=1$ we would need $c(1)e^{w(1)}=\frac12$ with $c(1)=0$, and at $\theta=0$ we would need $w(0)=-\infty$; neither is admissible.

6.4 Exercises 6.22–6.28

Problem 6.22. Let $X_1,\dots,X_n$ be a random sample from a population with pdf

$$f(x\mid\theta)=\theta x^{\theta-1},\qquad 0<x<1,\quad \theta>0.$$

(a) Is $\sum X_i$ sufficient for θ?
(b) Find a complete sufficient statistic for θ.

Solution. (a) No. The minimal sufficient statistic is $\prod X_i$, and it is not a function of $\sum X_i$.
For two sample points in $(0,1)^n$,

$$\frac{f(\mathbf{x}\mid\theta)}{f(\mathbf{y}\mid\theta)}=\left(\frac{\prod_i x_i}{\prod_i y_i}\right)^{\theta-1},$$

which is constant in $\theta>0$ if and only if $\prod_i x_i=\prod_i y_i$. Hence $\prod X_i$ is minimal sufficient by Theorem 6.2.13. If $\sum X_i$ were sufficient, then by Definition 6.2.11 $\prod X_i$ would be a function of $\sum X_i$. It is not (here $n\ge 2$): take $\mathbf{x}$ and $\mathbf{y}$ agreeing in coordinates $3,\dots,n$ with

$$(x_1,x_2)=\left(\tfrac15,\tfrac25\right),\qquad (y_1,y_2)=\left(\tfrac{1}{10},\tfrac12\right),$$

so that $\sum x_i=\sum y_i$ while $x_1x_2=\frac{2}{25}\neq\frac{1}{20}=y_1y_2$. So $\sum X_i$ is not sufficient.
(b) $T=-\sum_{i=1}^n\log X_i$.
The pdf is a one-parameter exponential family,

$$f(x\mid\theta)=\theta\exp\{(\theta-1)\log x\}=\frac1x\,\theta\exp\{\theta\log x\},$$

with $h(x)=1/x$ on $(0,1)$, $c(\theta)=\theta$, $w(\theta)=\theta$, $t(x)=\log x$. The parameter space $\{w(\theta):\theta>0\}=(0,\infty)$ contains an open set in $\mathbb{R}$, so by Theorem 6.2.25 the statistic $\sum_i t(X_i)=\sum_i\log X_i$ is complete and sufficient; so is its one-to-one image $T=-\sum_i\log X_i$.

Problem 6.23. Let $X_1,\dots,X_n$ be a random sample from a uniform distribution on the interval $(\theta,2\theta)$, $\theta>0$. Find a minimal sufficient statistic for θ. Is the statistic complete?

Solution. $T(\mathbf{X})=\big(X_{(1)},X_{(n)}\big)$ is minimal sufficient, and it is not complete.
The joint pdf is

$$f(\mathbf{x}\mid\theta)=\frac{1}{\theta^n}\prod_{i=1}^n I_{(\theta,2\theta)}(x_i)=\frac{1}{\theta^n}\,I\big(\tfrac12 x_{(n)}<\theta<x_{(1)}\big),$$

so for two sample points the ratio $f(\mathbf{x}\mid\theta)/f(\mathbf{y}\mid\theta)$ is 1 where both are positive and is 0 or undefined otherwise. It is therefore constant (and finite, nonzero) as a function of θ exactly when the two indicator sets $\big(\frac12 x_{(n)},x_{(1)}\big)$ and $\big(\frac12 y_{(n)},y_{(1)}\big)$ coincide, that is, when $x_{(1)}=y_{(1)}$ and $x_{(n)}=y_{(n)}$. By Theorem 6.2.13, $\big(X_{(1)},X_{(n)}\big)$ is minimal sufficient.

Completeness fails. Writing $X_i = \theta U_i$ with $U_1, \dots, U_n$ iid uniform$(1, 2)$,

$$\mathrm{E}\, X_{(1)} = \theta\left(1 + \frac{1}{n+1}\right) = \theta\,\frac{n+2}{n+1}, \qquad \mathrm{E}\, X_{(n)} = \theta\left(1 + \frac{n}{n+1}\right) = \theta\,\frac{2n+1}{n+1},$$

since $U_{(1)} - 1$ and $U_{(n)} - 1$ are the extreme order statistics of a uniform$(0, 1)$ sample. Hence

$$g\big(X_{(1)}, X_{(n)}\big) = (2n+1)X_{(1)} - (n+2)X_{(n)}$$

satisfies $\mathrm{E}_\theta\, g(T) = 0$ for every $\theta > 0$, yet $P_\theta(g(T) = 0) = 0$ for $n \geq 2$, since $\big(X_{(1)}, X_{(n)}\big)$ has a continuous distribution. So T is not complete. (For $n = 1$ take instead $g(x) = x^{-1}\sin(2\pi \log_2 x)$: the substitution $u = \log_2 x$ integrates $\sin(2\pi u)$ over one full period.)

Problem 6.24. Consider the following family of distributions:

$$\mathcal{P} = \Big\{P_\lambda(X = x) : P_\lambda(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \\ x = 0, 1, 2, \dots;\ \lambda = 0 \text{ or } 1\Big\}.$$

This is a Poisson family with λ restricted to be 0 or 1. Show that the family $\mathcal{P}$ is not complete, demonstrating that completeness can be dependent on the range of the parameter. (See Exercises 6.15 and 6.18.)

Solution. Take

$$g(x) = \begin{cases} 1, & x = 1, \\ -2, & x = 2, \\ 0, & \text{otherwise.} \end{cases}$$

Then $\mathrm{E}_0\, g(X) = g(0) = 0$ (at $\lambda = 0$ the distribution is point mass at 0), and

$$\mathrm{E}_1\, g(X) = e^{-1}\sum_{x \geq 0} \frac{g(x)}{x!} = e^{-1}\left(\frac{1}{1!} - \frac{2}{2!}\right) = 0.$$

So $\mathrm{E}_\lambda\, g(X) = 0$ for every λ in the parameter set $\{0, 1\}$, while

$$P_1\big(g(X) = 0\big) = 1 - e^{-1}\left(1 + \tfrac{1}{2}\right) = 1 - \tfrac{3}{2}e^{-1} \approx 0.448 < 1.$$

Thus g is not P_λ-a.s. zero for all λ, and $\mathcal{P}$ is not complete.
The range matters: for λ ranging over $(0, \infty)$, $e^\lambda \mathrm{E}_\lambda g(X) = \sum_x g(x)\lambda^x/x!$ is a power series vanishing on an interval, so every coefficient $g(x)/x!$ is 0 (Exercise 6.18) — two parameter values give only the two equations above.

Problem 6.25. We have seen a number of theorems concerning sufficiency and related concepts for exponential families. Theorem 5.2.11 gave the distribution of a statistic whose sufficiency is characterized in Theorem 6.2.10 and completeness in Theorem 6.2.25. But if the family is curved, the open set condition of Theorem 6.2.25 is not satisfied. In such cases, is the sufficient statistic of Theorem 6.2.10 also minimal? By applying Theorem 6.2.13 to $T(\mathbf{x})$ of Theorem 6.2.10, establish the following:
(a) The statistic $\big(\sum X_i, \sum X_i^2\big)$ is sufficient, but not minimal sufficient, in the $n(\mu, \mu)$ family.
(b) The statistic $\sum X_i^2$ is minimal sufficient in the $n(\mu, \mu)$ family.
(c) The statistic $\big(\sum X_i, \sum X_i^2\big)$ is minimal sufficient in the $n(\mu, \mu^2)$ family.
(d) The statistic $\big(\sum X_i, \sum X_i^2\big)$ is minimal sufficient in the $n(\mu, \sigma^2)$ family.

Solution. Take (b) first, since (a) rests on it. Throughout $a = \sum x_i^2 - \sum y_i^2$ and $b = \sum x_i - \sum y_i$ for sample points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and Theorem 6.2.13 is applied to the ratio of joint densities.
(b) In $n(\mu, \mu)$, $\mu > 0$, the pdf is

$$f(x \mid \mu) = \frac{1}{\sqrt{2\pi\mu}} \exp\left\{-\frac{x^2}{2\mu} + x - \frac{\mu}{2}\right\},$$

an exponential family with $t_1(x) = x$, $t_2(x) = x^2$, so Theorem 6.2.10 gives $\left(\sum X_i, \sum X_i^2\right)$ as a sufficient statistic. For the ratio,

$$\frac{f(\mathbf{x} \mid \mu)}{f(\mathbf{y} \mid \mu)} = \exp\left\{-\frac{a}{2\mu} + b\right\},$$

which is constant in $\mu > 0$ if and only if $a = 0$. By Theorem 6.2.13, $\sum X_i^2$ is minimal sufficient.
(a) Sufficiency is Theorem 6.2.10, as just noted. It is not minimal: by (b) the statistic $\sum X_i^2$ is sufficient, so a minimal sufficient statistic must be a function of it (Definition 6.2.11), and $\left(\sum X_i, \sum X_i^2\right)$ is not. Indeed $\mathbf{x} = (1, 2, x_3, \ldots, x_n)$ and $\mathbf{y} = (-1, -2, x_3, \ldots, x_n)$ give the same $\sum X_i^2$ but $\sum x_i - \sum y_i = 6$.
(c) In $n(\mu, \mu^2)$, $\mu \neq 0$,

$$f(x \mid \mu) = \frac{1}{\sqrt{2\pi}\,|\mu|} \exp\left\{-\frac{x^2}{2\mu^2} + \frac{x}{\mu} - \frac{1}{2}\right\},$$

so

$$\frac{f(\mathbf{x} \mid \mu)}{f(\mathbf{y} \mid \mu)} = \exp\left\{-\frac{a}{2\mu^2} + \frac{b}{\mu}\right\}.$$

This is constant in μ iff $h(\mu) := -a/(2\mu^2) + b/\mu$ is constant; since $h(\mu) \to 0$ as $\mu \to \infty$, the constant is 0, so $\mu^2 h(\mu) = -a/2 + b\mu \equiv 0$, forcing $b = 0$ and $a = 0$. By Theorem 6.2.13, $\left(\sum X_i, \sum X_i^2\right)$ is minimal sufficient.
(d) In $n(\mu, \sigma^2)$ with both parameters free,

$$\frac{f(\mathbf{x} \mid \mu, \sigma^2)}{f(\mathbf{y} \mid \mu, \sigma^2)} = \exp\left\{-\frac{a}{2\sigma^2} + \frac{\mu b}{\sigma^2}\right\}.$$

Fixing $\sigma^2 = 1$ and varying μ forces $b = 0$; then varying σ^2 forces $a = 0$. Conversely $a = b = 0$ makes the ratio identically 1. By Theorem 6.2.13, $\left(\sum X_i, \sum X_i^2\right)$ is minimal sufficient (Example 6.2.14, in the equivalent coordinates $(\bar{X}, S^2)$).

Problem 6.26. Use Theorem 6.6.5 to establish that, given a sample $X_1, \ldots, X_n$, the following statistics are minimal sufficient.

	Statistic	Distribution
(a)	$\bar{X}$	$n(\theta, 1)$
(b)	$\sum X_i$	gamma(α, β), α known
(c)	$\max_i X_i$	uniform$(0, \theta)$
(d)	$X_{(1)}, \ldots, X_{(n)}$	Cauchy$(\theta, 1)$
(e)	$X_{(1)}, \ldots, X_{(n)}$	logistic(μ, β)

Theorem 6.6.5 states: if the densities $\{f_0(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$ all have common support, then (a) $T(\mathbf{x}) = (f_1(\mathbf{x})/f_0(\mathbf{x}), \ldots, f_k(\mathbf{x})/f_0(\mathbf{x}))$ is minimal sufficient for $\{f_0, \ldots, f_k\}$; and (b) if $\mathcal{F}$ is a family of densities with common support, $f_i \in \mathcal{F}$ for $i = 0, \ldots, k$, and $T(\mathbf{x})$ is sufficient for $\mathcal{F}$, then $T(\mathbf{x})$ is minimal sufficient for $\mathcal{F}$.

Solution. In each case pick $f_0, \ldots, f_k \in \mathcal{F}$ with the common support that Theorem 6.6.5 requires, show the likelihood-ratio vector $T(\mathbf{x})$ is a one-to-one function of the proposed statistic $S(\mathbf{x})$, and check S is sufficient for $\mathcal{F}$; Theorem 6.6.5(a) then makes S minimal sufficient for $\{f_0, \ldots, f_k\}$ (one-to-one maps preserve minimal sufficiency), and 6.6.5(b) makes it minimal sufficient for $\mathcal{F}$.

(a) Take $k=1$, $\theta_0=0$, $\theta_1=1$. Then

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \exp\left\{-\tfrac{1}{2}\sum_i (x_i-1)^2 + \tfrac{1}{2}\sum_i x_i^2\right\} = \exp\left\{n\bar{x} - \tfrac{n}{2}\right\},$$

a strictly increasing, hence one-to-one, function of $\bar{x}$. The support is $\mathbb{R}^n$ for every θ, and $\bar{X}$ is sufficient for $\mathcal{F}=\{n(\theta,1)\}$ by Theorem 6.2.10. So $\bar{X}$ is minimal sufficient.
(b) Take $k=1$, $\beta_0=1$, $\beta_1=\frac{1}{2}$. With α known,

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \frac{\beta_1^{-n\alpha}}{\beta_0^{-n\alpha}} \exp\left\{-\left(\tfrac{1}{\beta_1}-\tfrac{1}{\beta_0}\right)\sum_i x_i\right\} = 2^{n\alpha} e^{-\sum_i x_i},$$

one-to-one in $\sum x_i$. The support is $(0,\infty)^n$ for every β, and $\sum X_i$ is sufficient for $\mathcal{F}=\{\text{gamma}(\alpha,\beta) : \beta>0\}$ by Theorem 6.2.10. So $\sum X_i$ is minimal sufficient.
(c) Theorem 6.6.5 does not apply here: the support $(0,\theta)^n$ depends on θ, so the common-support hypothesis fails. Argue directly instead, exactly as the theorem's proof does. Let T' be any statistic sufficient for $\mathcal{F}=\{\text{uniform}(0,\theta):\theta>0\}$. By the Factorization Theorem 6.2.6 there are g,h with

$$\theta^{-n} I\big(x_{(n)}<\theta\big) = g\big(T'(\mathbf{x}) \mid \theta\big)\, h(\mathbf{x}), \qquad \mathbf{x}\in(0,\infty)^n,$$

and $h(\mathbf{x})>0$ (take $\theta > x_{(n)}$). If $T'(\mathbf{x})=T'(\mathbf{y})$, then dividing,

$$\frac{I\big(x_{(n)}<\theta\big)}{h(\mathbf{x})} = \frac{I\big(y_{(n)}<\theta\big)}{h(\mathbf{y})} \quad \text{for every } \theta>0,$$

so the two indicators vanish for the same θ, giving $x_{(n)}=y_{(n)}$. Hence $\max_i X_i$ is a function of every sufficient statistic; since it is itself sufficient (factorization, with $h\equiv 1$), it is minimal sufficient by Definition 6.2.11.
(d) Cauchy$(\theta,1)$: $f_\theta(\mathbf{x}) = \pi^{-n}/P_{\mathbf{x}}(\theta)$, where

$$P_{\mathbf{x}}(\theta) = \prod_{j=1}^{n}\big(1+(x_j-\theta)^2\big)$$

is a monic polynomial in θ of degree $2n$, strictly positive. Take $k=2n$, $\theta_0=0$, and $\theta_1,\dots,\theta_{2n}$ distinct and nonzero. Then

$$\frac{f_i(\mathbf{x})}{f_0(\mathbf{x})} = \frac{P_{\mathbf{x}}(0)}{P_{\mathbf{x}}(\theta_i)}, \qquad i=1,\dots,2n,$$

so $T(\mathbf{x})$ determines $Q(\theta) := P_{\mathbf{x}}(\theta)/P_{\mathbf{x}}(0)$ at the $2n+1$ distinct points $0,\theta_1,\dots,\theta_{2n}$ (at $\theta=0$ the value is 1). A polynomial of degree $2n$ is determined by its values at $2n+1$ points, so Q is known; its leading coefficient is $1/P_{\mathbf{x}}(0)$, whence $P_{\mathbf{x}} = P_{\mathbf{x}}(0)\,Q$ is known. The complex roots of $P_{\mathbf{x}}$ are $x_j \pm i$, $j=1,\dots,n$, so the multiset $\{x_1,\dots,x_n\}$ — that is, $\big(x_{(1)},\dots,x_{(n)}\big)$ — is recovered. Conversely $T(\mathbf{x})$ is a function of the order statistics, so the map is one-to-one. The support is $\mathbb{R}^n$ throughout, and the order statistics are sufficient for $\mathcal{F}$ because $f_\theta(\mathbf{x})$ is a symmetric function of $x_1,\dots,x_n$ (factorization with $h\equiv 1$). So the order statistics are minimal sufficient.
(e) Logistic(μ,β): restrict to $\beta=1$ and put $u_j = e^{-x_j}>0$, $c=e^{\mu}>0$. Then

$$f_\mu(\mathbf{x}) = \prod_{j=1}^{n}\frac{c\,u_j}{(1+c\,u_j)^2} = \frac{c^n\prod_j u_j}{g(c)^2}, \qquad g(c) := \prod_{j=1}^{n}(1+c\,u_j),$$

a degree-n polynomial in c with $g(0)=1$. Take $\mu_0=0$ (i.e. $c=1$) and $\mu_i=\log c_i$ for $c_1,\dots,c_n$ distinct and different from 1. Then

$$\frac{f_i(\mathbf{x})}{f_0(\mathbf{x})} = c_i^n\left(\frac{g(1)}{g(c_i)}\right)^2,$$

so $T(\mathbf{x})$ determines $\varphi(c) := g(c)/g(1)>0$ at $c_1,\dots,c_n$, and $\varphi(1)=1$ is known. These are $n+1$ distinct points, so the degree-n polynomial φ is determined; then $g(1)=1/\varphi(0)$ and $g=g(1)\varphi$ are known. The roots of g are $-1/u_j = -e^{x_j}$, so the multiset $\{x_1,\dots,x_n\}$ is recovered, and the map is one-to-one. The support is $\mathbb{R}^n$ for all (μ,β), and the order statistics are sufficient for $\mathcal{F}$ since the joint pdf is symmetric in $x_1,\dots,x_n$. So the order statistics are minimal sufficient.

Problem 6.27. Let $X_1, \ldots, X_n$ be a random sample from the *inverse Gaussian distribution* with pdf

$$f(x \mid \mu, \lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} e^{-\lambda(x-\mu)^2/(2\mu^2 x)}, \qquad 0 < x < \infty.$$

(a) Show that the statistics

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad T = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i} - \frac{1}{\bar{X}}}$$

are sufficient and complete.
(b) For $n = 2$, show that $\bar{X}$ has an inverse Gaussian distribution, $n\lambda/T$ has a χ^2_{n-1} distribution, and they are independent. (Schwarz and Samanta 1991 do the general case.)

Solution. Errata: part (b) is false for T exactly as printed, and the intended denominator is $\sum_{i=1}^{n}\left(1/X_i - 1/\bar{X}\right) = \sum_i 1/X_i - n/\bar{X}$, so read $T = n/\sum_i\left(1/X_i - 1/\bar{X}\right)$ throughout.
(a) Expanding the exponent,

$$-\frac{\lambda(x-\mu)^2}{2\mu^2 x} = -\frac{\lambda x}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2x},$$

so the joint pdf is

$$f(\mathbf{x} \mid \mu, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{n/2} \left(\prod_i x_i\right)^{-3/2} e^{n\lambda/\mu} \exp\left\{-\frac{\lambda}{2\mu^2}\sum_i x_i - \frac{\lambda}{2}\sum_i \frac{1}{x_i}\right\}.$$

This is a two-parameter exponential family with $t_1(x) = x$, $t_2(x) = 1/x$ and natural parameters

$$\eta_1 = -\frac{\lambda}{2\mu^2}, \qquad \eta_2 = -\frac{\lambda}{2}.$$

As (μ, λ) ranges over $(0, \infty)^2$, (η_1, η_2) ranges over all of $(-\infty, 0) \times (-\infty, 0)$, which contains an open subset of $\mathbb{R}^2$. By Theorem 6.2.10 and Theorem 6.2.25, $(\sum_i X_i,\ \sum_i 1/X_i)$ is a complete sufficient statistic.
Now $(\bar{X}, T)$ is a one-to-one function of it:

$$\bar{X} = \frac{1}{n}\sum_i X_i, \qquad \frac{n}{T} = \sum_i \frac{1}{X_i} - \frac{n}{\bar{X}},$$

so $\sum_i X_i = n\bar{X}$ and $\sum_i 1/X_i = n/T + n/\bar{X}$. Sufficiency and completeness are preserved under one-to-one transformations, so $(\bar{X}, T)$ is sufficient and complete.
(b) Let $n = 2$ and set

$$Y = \bar{X} = \frac{X_1 + X_2}{2}, \qquad W = \frac{n\lambda}{T} = \lambda\sum_{i=1}^{2}\left(\frac{1}{X_i} - \frac{1}{\bar{X}}\right) = \frac{\lambda\,(X_1 - X_2)^2}{X_1 X_2 (X_1 + X_2)},$$

the last equality from

$$\frac{1}{x_1} + \frac{1}{x_2} - \frac{4}{x_1 + x_2} = \frac{(x_1 + x_2)^2 - 4x_1 x_2}{x_1 x_2 (x_1 + x_2)}.$$

Write $s = x_1 + x_2 = 2y$ and $p = x_1 x_2$. Since $\sum 1/x_i = s/p$, the definition of w gives $\lambda s/p = w + 2\lambda/y$, and therefore

$$\begin{aligned}\frac{2\lambda}{\mu} - \frac{\lambda s}{2\mu^2} - \frac{\lambda}{2}\cdot\frac{s}{p} &= \frac{2\lambda}{\mu} - \frac{\lambda y}{\mu^2} - \frac{w}{2} - \frac{\lambda}{y} \\ &= -\frac{\lambda}{y}\left(\frac{y}{\mu} - 1\right)^2 - \frac{w}{2} = -\frac{\lambda(y-\mu)^2}{\mu^2 y} - \frac{w}{2}.\end{aligned}$$

For the Jacobian, restrict to $x_1 > x_2$ (the map $(x_1, x_2) \mapsto (y, w)$ is two-to-one, so double at the end). The change $(x_1, x_2) \to (s, p)$ has $|\partial(s,p)/\partial(x_1,x_2)| = |x_1 - x_2|$, and inverting $w = \lambda(s^2 - 4p)/(ps)$ gives

$$p = \frac{2\lambda y^2}{yw + 2\lambda}, \qquad \left|\frac{\partial(s,p)}{\partial(y,w)}\right| = 2\left|\frac{\partial p}{\partial w}\right| = \frac{4\lambda y^3}{(yw+2\lambda)^2}.$$

Also $|x_1 - x_2| = \sqrt{s^2 - 4p} = \sqrt{wps/\lambda}$. Collecting, and using $p^{-2} = (yw + 2\lambda)^2/(4\lambda^2 y^4)$ so that $p^{-2} \cdot 4\lambda y^3/(yw+2\lambda)^2 = 1/(\lambda y)$, the joint pdf of (Y, W) is

$$\begin{aligned} g(y,w) &= 2 \cdot \frac{\lambda}{2\pi}\, p^{-3/2} \cdot \frac{1}{\sqrt{wps/\lambda}} \cdot \frac{4\lambda y^3}{(yw+2\lambda)^2}\, e^{-\lambda(y-\mu)^2/(\mu^2 y)} e^{-w/2} \\ &= \frac{\lambda}{\pi} \cdot \frac{1}{\lambda y}\sqrt{\frac{\lambda}{2yw}}\, e^{-\lambda(y-\mu)^2/(\mu^2 y)} e^{-w/2} \\ &= \frac{1}{\pi}\sqrt{\frac{\lambda}{2y^3 w}}\, e^{-\lambda(y-\mu)^2/(\mu^2 y)} e^{-w/2}. \end{aligned}$$

This factors as

$$g(y,w) = \underbrace{\left(\frac{2\lambda}{2\pi y^3}\right)^{1/2} e^{-2\lambda(y-\mu)^2/(2\mu^2 y)}}_{\mathrm{IG}(\mu, 2\lambda)\ \text{pdf at } y} \cdot \underbrace{\frac{1}{\sqrt{2\pi}}\, w^{-1/2} e^{-w/2}}_{\chi^2_1\ \text{pdf at } w}, \qquad y > 0,\ w > 0.$$

Hence $\bar{X} \sim$ inverse Gaussian$(\mu, n\lambda)$ with $n = 2$, $n\lambda/T \sim \chi^2_1 = \chi^2_{n-1}$, and the two are independent.

Problem 6.28. Prove Theorem 6.6.5. (Hint: First establish that the minimal sufficiency of $T(\mathbf{X})$ in the family $\{f_0(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$ follows from Theorem 6.2.13. Then argue that any statistic that is sufficient in $\mathcal{F}$ must be a function of $T(\mathbf{x})$.)
Theorem 6.6.5 (Minimal sufficient statistics). Suppose that the family of densities $\{f_0(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$ all have common support. Then
a. The statistic

$$T(\mathbf{X}) = \left(\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})}, \frac{f_2(\mathbf{X})}{f_0(\mathbf{X})}, \ldots, \frac{f_k(\mathbf{X})}{f_0(\mathbf{X})}\right)$$

is minimal sufficient for the family $\{f_0(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$.
b. If $\mathcal{F}$ is a family of densities with common support, and
(i) $f_i(\mathbf{x}) \in \mathcal{F}$, $i = 0, 1, \ldots, k$,
(ii) $T(\mathbf{x})$ is sufficient for $\mathcal{F}$,
then $T(\mathbf{x})$ is minimal sufficient for $\mathcal{F}$.

Solution. (a) Apply Theorem 6.2.13 on the common support $\mathcal{X}$, where every f_i is positive so the ratios below are finite. It suffices to show that for $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$\frac{f_i(\mathbf{x})}{f_i(\mathbf{y})} \ \text{ is constant in } i \iff T(\mathbf{x}) = T(\mathbf{y}).$$

($\Rightarrow$) Suppose $f_i(\mathbf{x})/f_i(\mathbf{y}) = c$ for all $i = 0, \ldots, k$. Taking $i = 0$ gives $c = f_0(\mathbf{x})/f_0(\mathbf{y})$, so for each i,

$$\frac{f_i(\mathbf{x})}{f_i(\mathbf{y})} = \frac{f_0(\mathbf{x})}{f_0(\mathbf{y})} \iff \frac{f_i(\mathbf{x})}{f_0(\mathbf{x})} = \frac{f_i(\mathbf{y})}{f_0(\mathbf{y})},$$

which is precisely $T(\mathbf{x}) = T(\mathbf{y})$.
($\Leftarrow$) Conversely, if $T(\mathbf{x}) = T(\mathbf{y})$ then for every $i = 1, \ldots, k$,

$$\frac{f_i(\mathbf{x})}{f_i(\mathbf{y})} = \frac{[f_i(\mathbf{x})/f_0(\mathbf{x})]\, f_0(\mathbf{x})}{[f_i(\mathbf{y})/f_0(\mathbf{y})]\, f_0(\mathbf{y})} = \frac{f_0(\mathbf{x})}{f_0(\mathbf{y})},$$

and the case $i = 0$ is trivial, so the ratio does not depend on i.
By Theorem 6.2.13, $T(\mathbf{X})$ is minimal sufficient for $\{f_0, \ldots, f_k\}$.
(b) Let $T'(\mathbf{X})$ be any statistic sufficient for $\mathcal{F}$. By the Factorization Theorem 6.2.6 there exist $g(\cdot \mid f)$ and h with

$$f(\mathbf{x}) = g\big(T'(\mathbf{x}) \mid f\big)\, h(\mathbf{x}) \qquad \text{for all } f \in \mathcal{F},\ \mathbf{x} \in \mathcal{X}.$$

By hypothesis (i) this factorization holds in particular for $f = f_0, \ldots, f_k$, so T' is sufficient for the subfamily $\{f_0, \ldots, f_k\}$. By part (a), T is minimal sufficient for that subfamily, hence $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ (Definition 6.2.11).
Thus T is a function of every statistic sufficient for $\mathcal{F}$, and by hypothesis (ii) T is itself sufficient for $\mathcal{F}$. By Definition 6.2.11, $T(\mathbf{X})$ is minimal sufficient for $\mathcal{F}$.

6.5 Exercises 6.29–6.35

Problem 6.29. The concept of minimal sufficiency can be extended beyond parametric families of distributions. Show that if $X_1, \ldots, X_n$ are a random sample from a density f that is unknown, then the order statistics are minimal sufficient.
(Hint: Use Theorem 6.6.5, taking the family $\{f_0(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$ to be logistic densities.)

Solution. Take $k = n$ and let $f_j(\mathbf{x}) = \prod_{i=1}^n \ell(x_i \mid \mu_j)$, $j = 0, 1, \ldots, n$, be the n-fold products of logistic$(\mu_j, 1)$ densities

$$\ell(x \mid \mu) \;=\; \frac{e^{-(x-\mu)}}{\big(1 + e^{-(x-\mu)}\big)^2},$$

with $\mu_0 = 0$ and $\mu_1, \ldots, \mu_n$ distinct and nonzero. Let $\mathcal{F}$ be the family of all joint densities $\prod_{i=1}^n f(x_i)$ with f a density that is strictly positive on $\mathbb{R}$, so every member of $\mathcal{F}$ has the common support $\mathbb{R}^n$ demanded by Theorem 6.6.5, and $f_j \in \mathcal{F}$ for each j.
The order statistics are sufficient for $\mathcal{F}$: $\prod_{i=1}^n f(x_i)$ is a symmetric function of $\mathbf{x}$, hence depends on $\mathbf{x}$ only through $\big(x_{(1)}, \ldots, x_{(n)}\big)$, and the Factorization Theorem (Theorem 6.2.10) applies with $h \equiv 1$.
Now compute the ratios of Theorem 6.6.5. Writing $y_i = e^{-x_i}$ and $t_j = e^{-\mu_j}$,

$$f_j(\mathbf{x}) \;=\; \prod_{i=1}^n \frac{y_i/t_j}{(1 + y_i/t_j)^2} \;=\; t_j^n \,\frac{\prod_{i=1}^n y_i}{\prod_{i=1}^n (t_j + y_i)^2},$$

so that, with $P(t) = \prod_{i=1}^n (t + y_i)$ and $t_0 = 1$,

$$\frac{f_j(\mathbf{x})}{f_0(\mathbf{x})} \;=\; t_j^n \left(\frac{P(1)}{P(t_j)}\right)^2, \qquad j = 1, \ldots, n.$$

Let $T(\mathbf{x}) = \big(f_1(\mathbf{x})/f_0(\mathbf{x}), \ldots, f_n(\mathbf{x})/f_0(\mathbf{x})\big)$ be the statistic of Theorem 6.6.5(a). Since $t_j > 0$ and $P > 0$ on $(0, \infty)$, its jth coordinate determines $Q(t_j)$, where $Q(t) := P(t)/P(1)$. With $Q(1) = 1$ this pins the degree-n polynomial Q at the $n + 1$ distinct points $t_1, \ldots, t_n, 1$, hence determines Q; its leading coefficient $1/P(1)$ then yields $P = P(1)Q$, hence the multiset of roots $\{-y_1, \ldots, -y_n\}$, hence ($x \mapsto e^{-x}$ being strictly decreasing) the order statistics $x_{(1)} \le \cdots \le x_{(n)}$. Conversely the order statistics determine P and so determine $T(\mathbf{x})$.
So T and $\big(X_{(1)}, \ldots, X_{(n)}\big)$ are one-to-one functions of each other; in particular $T(\mathbf{X})$ is sufficient for $\mathcal{F}$, so hypotheses (i) and (ii) of Theorem 6.6.5(b) hold and $T(\mathbf{X})$ – equivalently the order statistics – is minimal sufficient for $\mathcal{F}$.

Problem 6.30. Let $X_1, \ldots, X_n$ be a random sample from the pdf $f(x|\mu) = e^{-(x-\mu)}$, where $-\infty < \mu < x < \infty$.
(a) Show that $X_{(1)} = \min_i X_i$ is a complete sufficient statistic.
(b) Use Basu's Theorem to show that $X_{(1)}$ and S^2 are independent.

Solution. (a) The joint pdf factors as

$$f(\mathbf{x} \mid \mu) = \prod_{i=1}^{n} e^{-(x_i-\mu)} I_{(\mu,\infty)}(x_i) = \underbrace{e^{n\mu} I_{(\mu,\infty)}\big(x_{(1)}\big)}_{g(x_{(1)}|\mu)} \cdot \underbrace{e^{-\sum_i x_i}}_{h(\mathbf{x})},$$

so $X_{(1)}$ is sufficient by the Factorization Theorem (Theorem 6.2.10).
For completeness, $P_\mu\big(X_{(1)} > t\big) = \prod_i P_\mu(X_i > t) = e^{-n(t-\mu)}$ for $t > \mu$, so $X_{(1)}$ has pdf $ne^{-n(t-\mu)}$ on (μ, ∞). If $E_\mu g\big(X_{(1)}\big) = 0$ for all μ, then

$$0 = \int_\mu^\infty g(t)\, ne^{-n(t-\mu)}\, dt \quad \Longleftrightarrow \quad \int_\mu^\infty g(t) e^{-nt}\, dt = 0 \ \text{ for all } \mu.$$

Differentiating in μ gives $-g(\mu)e^{-n\mu} = 0$, so $g = 0$ a.e. (at every Lebesgue point), which is what completeness requires. Hence $X_{(1)}$ is a complete sufficient statistic.
(b) Write $X_i = \mu + Z_i$ with $Z_1, \ldots, Z_n$ iid exponential(1), free of μ. Then

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - \bar{X}\big)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(Z_i - \bar{Z}\big)^2,$$

whose distribution does not involve μ; thus S^2 is ancillary (Definition 6.2.16). Since $X_{(1)}$ is complete sufficient by part (a), Basu's Theorem (Theorem 6.2.24) gives that $X_{(1)}$ and S^2 are independent.

Problem 6.31. Boos and Hughes-Oliver (1998) detail a number of instances where application of Basu's Theorem can simplify calculations. Here are a few.
(a) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$, where σ^2 is known.
(i) Show that $\bar{X}$ is complete sufficient for μ, and S^2 is ancillary. Hence by Basu's Theorem, $\bar{X}$ and S^2 are independent.
(ii) Show that this independence carries over even if σ^2 is unknown, as knowledge of σ^2 has no bearing on the distributions. (Compare this proof to the more involved Theorem 5.3.1(a).)
(b) A *Monte Carlo swindle* is a technique for improving variance estimates. Suppose that $X_1, \ldots, X_n$ are iid $\mathrm{n}(\mu, \sigma^2)$ and that we want to compute the variance of the median, M.
(i) Apply Basu's Theorem to show that $\mathrm{Var}(M) = \mathrm{Var}(M - \bar{X}) + \mathrm{Var}(\bar{X})$; thus we only have to simulate the $\mathrm{Var}(M - \bar{X})$ piece of $\mathrm{Var}(M)$ (since $\mathrm{Var}(\bar{X}) = \sigma^2/n$).
(ii) Show that the swindle estimate is more precise by showing that the variance of M is approximately $2[\mathrm{Var}(M)]^2/(N-1)$ and that of $M - \bar{X}$ is approximately $2[\mathrm{Var}(M - \bar{X})]^2/(N-1)$, where N is the number of Monte Carlo samples.
(c) (i) If X/Y and Y are independent random variables, show that

$$\mathrm{E}\left(\frac{X}{Y}\right)^k = \frac{\mathrm{E}(X^k)}{\mathrm{E}(Y^k)}.$$

(ii) Use this result and Basu's Theorem to show that if $X_1, \ldots, X_n$ are iid gamma(α, β), where α is known, then for $T = \sum_i X_i$

$$\mathrm{E}\big(X_{(i)} \mid T\big) = \mathrm{E}\left(\frac{X_{(i)}}{T}\, T \,\middle|\, T\right) = T\,\frac{\mathrm{E}\big(X_{(i)}\big)}{\mathrm{E}T}.$$

Solution. (a)(i) With σ^2 known the model is the one-parameter exponential family

$$f(x \mid \mu) = \underbrace{\frac{e^{-x^2/(2\sigma^2)}}{\sqrt{2\pi}\,\sigma}}_{h(x)} \underbrace{e^{-\mu^2/(2\sigma^2)}}_{c(\mu)} \exp\left\{\frac{\mu}{\sigma^2}\, x\right\},$$

and $\{\mu/\sigma^2 : \mu \in \mathbb{R}\} = \mathbb{R}$ contains an open interval, so by Theorem 6.2.25 $\sum_i X_i$ – equivalently $\bar{X}$ – is a complete sufficient statistic for μ. Writing $X_i = \mu + \sigma Z_i$ with Z_i iid $\mathrm{n}(0,1)$,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\bigl(X_i - \bar{X}\bigr)^2 = \frac{\sigma^2}{n-1}\sum_{i=1}^{n}\bigl(Z_i - \bar{Z}\bigr)^2,$$

whose distribution ($\sigma^2\chi^2_{n-1}/(n-1)$) does not involve μ, so S^2 is ancillary. Basu's Theorem (Theorem 6.2.24) gives $\bar{X} \perp S^2$.
(ii) Independence is a property of the joint distribution of $(\bar{X}, S^2)$ at a single parameter point (μ, σ^2), and part (i) established it at every such point, since σ^2 there was an arbitrary fixed value. Enlarging the model by declaring σ^2 unknown does not alter any of these joint distributions; it only changes which of them are indexed by the parameter. Hence $\bar{X}$ and S^2 are independent under $\mathrm{n}(\mu, \sigma^2)$ for all (μ, σ^2), which is Theorem 5.3.1(a).
(b)(i) With σ^2 fixed, $M - \bar{X}$ is ancillary for μ by Exercise 6.14 (the normal family is a location family), and $\bar{X}$ is complete sufficient by (a)(i). Basu's Theorem gives $M - \bar{X} \perp \bar{X}$, so $\mathrm{Cov}(M - \bar{X}, \bar{X}) = 0$ and

$$\begin{aligned}\mathrm{Var}(M) &= \mathrm{Var}\bigl((M - \bar{X}) + \bar{X}\bigr)\\ &= \mathrm{Var}(M - \bar{X}) + \mathrm{Var}(\bar{X}) + 2\,\mathrm{Cov}(M - \bar{X}, \bar{X})\\ &= \mathrm{Var}(M - \bar{X}) + \frac{\sigma^2}{n}.\end{aligned}$$

(ii) Here "the variance of M" must be read as the variance of the Monte Carlo estimate of $\mathrm{Var}(M)$. Both schemes estimate a variance by the sample variance of N iid replicates, and for $Y_1, \ldots, Y_N$ iid $\mathrm{n}(\theta, \tau^2)$ Theorem 5.3.1(c) gives $(N-1)S_Y^2/\tau^2 \sim \chi^2_{N-1}$, whence $\mathrm{Var}(S_Y^2) = 2\tau^4/(N-1)$. The direct estimator is the sample variance of replicates of M, which is approximately normal, so its variance is approximately

$$\frac{2[\mathrm{Var}(M)]^2}{N-1},$$

while the swindle estimator $\widehat{\mathrm{Var}}(M - \bar{X}) + \sigma^2/n$ adds a constant to the sample variance of the approximately normal replicates of $M - \bar{X}$, so its variance is approximately

$$\frac{2[\mathrm{Var}(M - \bar{X})]^2}{N-1}.$$

By (i), $\mathrm{Var}(M - \bar{X}) = \mathrm{Var}(M) - \sigma^2/n < \mathrm{Var}(M)$, so the swindle estimate has the smaller variance.
(c)(i) Independence of X/Y and Y makes $(X/Y)^k$ and Y^k independent, so

$$\mathrm{E}(X^k) = \mathrm{E}\left[\left(\frac{X}{Y}\right)^k Y^k\right] = \mathrm{E}\left(\frac{X}{Y}\right)^k \mathrm{E}\bigl(Y^k\bigr),$$

and dividing by $\mathrm{E}(Y^k)$ (assumed finite and nonzero) gives the claim.
(ii) With α known, $f(x \mid \beta) = \bigl[x^{\alpha-1}/\Gamma(\alpha)\bigr]\beta^{-\alpha}\exp\{-x/\beta\}$ is a one-parameter exponential family with natural parameter $-1/\beta$ ranging over the open interval $(-\infty, 0)$, so by Theorem 6.2.25 $T = \sum_i X_i$ is complete sufficient for β. Writing $X_j = \beta Z_j$ with Z_j iid gamma$(\alpha, 1)$,

$$\frac{X_{(i)}}{T} = \frac{Z_{(i)}}{\sum_j Z_j},$$

free of β, so $X_{(i)}/T$ is ancillary and Basu's Theorem gives $X_{(i)}/T \perp T$. Hence

$$\begin{aligned}\mathrm{E}\bigl(X_{(i)} \mid T\bigr) &= \mathrm{E}\left(\frac{X_{(i)}}{T}\,T \;\middle|\; T\right) = T\,\mathrm{E}\left(\frac{X_{(i)}}{T} \;\middle|\; T\right)\\ &= T\,\mathrm{E}\left(\frac{X_{(i)}}{T}\right) = T\,\frac{\mathrm{E}\bigl(X_{(i)}\bigr)}{\mathrm{E}T},\end{aligned}$$

the third equality by the independence just established and the fourth by part (i) with $X = X_{(i)}$, $Y = T$, $k = 1$.

Problem 6.32. Prove the Likelihood Principle Corollary. That is, assuming both the Formal Sufficiency Principle and the Conditionality Principle, prove that if $E = (\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$ is an experiment, then $\mathrm{Ev}(E, \mathbf{x})$ should depend on E and $\mathbf{x}$ only through $L(\theta|\mathbf{x})$.

Solution. By Birnbaum's Theorem (Theorem 6.3.6) the Formal Sufficiency Principle and the Conditionality Principle together imply the Formal Likelihood Principle, so it is enough to derive the corollary from the latter.
Call $(E_1, \mathbf{x}_1^*)$ and $(E_2, \mathbf{x}_2^*)$ *likelihood equivalent* if

$$L(\theta \mid \mathbf{x}_2^*) = C\, L(\theta \mid \mathbf{x}_1^*) \qquad \text{for all } \theta$$

for some constant C depending on the sample points but not on θ. This is an equivalence relation (take $C = 1$, C^{-1}, products of constants), and its classes are exactly the proportionality classes of likelihood functions. Condition (6.3.3) is precisely likelihood equivalence, so the Formal Likelihood Principle says that Ev is constant on each class. A function constant on the classes of an equivalence relation factors through the quotient: there is a ψ, defined on proportionality classes of likelihood functions, with

$$\mathrm{Ev}(E, \mathbf{x}) = \psi\big([\, L(\cdot \mid \mathbf{x})\,]\big).$$

Thus $\mathrm{Ev}(E, \mathbf{x})$ depends on E and $\mathbf{x}$ only through $L(\theta \mid \mathbf{x})$, and then only up to a θ-free constant multiple.

Problem 6.33. Fill in the gaps in the proof of Theorem 6.3.6, Birnbaum's Theorem.
Here $E_1 = (\mathbf{X}_1, \theta, \{f_1(\mathbf{x}_1|\theta)\})$ and $E_2 = (\mathbf{X}_2, \theta, \{f_2(\mathbf{x}_2|\theta)\})$ are experiments with the same θ, and $\mathbf{x}_1^*, \mathbf{x}_2^*$ are sample points satisfying the hypothesis (6.3.3) of the Formal Likelihood Principle, namely $L(\theta|\mathbf{x}_2^*) = C\, L(\theta|\mathbf{x}_1^*)$ for all θ, with C free of θ. E^* is the mixed experiment of the Conditionality Principle: J takes the values $1, 2$ with probability $\frac{1}{2}$ each, independent of θ, $\mathbf{X}_1$ and $\mathbf{X}_2$, experiment E_J is then performed, $\mathbf{X}^* = (J, \mathbf{X}_J)$, and $f^*((j, \mathbf{x}_j)|\theta) = \frac{1}{2} f_j(\mathbf{x}_j|\theta)$. The statistic on the sample space of E^* is

$$T(j, \mathbf{x}_j) = \begin{cases} (1, \mathbf{x}_1^*) & \text{if } j = 1,\ \mathbf{x}_1 = \mathbf{x}_1^*, \\ (1, \mathbf{x}_1^*) & \text{if } j = 2,\ \mathbf{x}_2 = \mathbf{x}_2^*, \\ (j, \mathbf{x}_j) & \text{otherwise.} \end{cases}$$

(a) Define $g(\mathbf{t}|\theta) = g((j, \mathbf{x}_j)|\theta) = f^*((j, \mathbf{x}_j)|\theta)$ and

$$h(j, \mathbf{x}_j) = \begin{cases} C & \text{if } (j, \mathbf{x}_j) = (2, \mathbf{x}_2^*), \\ 1 & \text{otherwise.} \end{cases}$$

Show that $T(j, \mathbf{x}_j)$ is a sufficient statistic in the E^* experiment by verifying that

$$g\big(T(j, \mathbf{x}_j) \mid \theta\big)\, h(j, \mathbf{x}_j) = f^*\big((j, \mathbf{x}_j) \mid \theta\big)$$

for all $(j, \mathbf{x}_j)$.
(b) As T is sufficient, show that the Formal Sufficiency Principle implies (6.3.4). Also the Conditionality Principle implies (6.3.5), and hence deduce the Formal Likelihood Principle.
(c) To prove the converse, first let one experiment be the E^* experiment and the other E_j and deduce that $\mathrm{Ev}(E^*, (j, \mathbf{x}_j)) = \mathrm{Ev}(E_j, \mathbf{x}_j)$, the Conditionality Principle. Then, if $T(\mathbf{X})$ is sufficient and $T(\mathbf{x}) = T(\mathbf{y})$, show that the likelihoods are proportional and then use the Formal Likelihood Principle to deduce $\mathrm{Ev}(E, \mathbf{x}) = \mathrm{Ev}(E, \mathbf{y})$, the Formal Sufficiency Principle.

Solution. (a) Two cases exhaust the sample space of E^*.

(i) $(j, \mathbf{x}_j) = (2, \mathbf{x}_2^*)$. Then $T(2, \mathbf{x}_2^*) = (1, \mathbf{x}_1^*)$ and $h = C$, so

$$\begin{aligned} g\big(T(2, \mathbf{x}_2^*) \mid \theta\big) h(2, \mathbf{x}_2^*) &= C \cdot \tfrac{1}{2} f_1(\mathbf{x}_1^* \mid \theta) \\ &= \tfrac{1}{2} f_2(\mathbf{x}_2^* \mid \theta) = f^*\big((2, \mathbf{x}_2^*) \mid \theta\big), \end{aligned}$$

the middle equality being exactly $f_2(\mathbf{x}_2^*|\theta) = C\, f_1(\mathbf{x}_1^*|\theta)$, which is (6.3.3).
(ii) Every other $(j, \mathbf{x}_j)$, including $(1, \mathbf{x}_1^*)$. Then $T(j, \mathbf{x}_j) = (j, \mathbf{x}_j)$ and $h = 1$, so

$$g\big(T(j, \mathbf{x}_j) \mid \theta\big) h(j, \mathbf{x}_j) = g\big((j, \mathbf{x}_j) \mid \theta\big) = f^*\big((j, \mathbf{x}_j) \mid \theta\big).$$

Since h does not involve θ and g involves the sample point only through T, the Factorization Theorem (Theorem 6.2.10) shows $T(J, \mathbf{X}_J)$ is sufficient for θ in E^*.
(b) By construction $T(1, \mathbf{x}_1^*) = T(2, \mathbf{x}_2^*) = (1, \mathbf{x}_1^*)$, and by (a) T is sufficient in E^*. The Formal Sufficiency Principle, applied to the single experiment E^* and the two sample points $(1, \mathbf{x}_1^*)$ and $(2, \mathbf{x}_2^*)$, therefore gives

$$\mathrm{Ev}\big(E^*, (1, \mathbf{x}_1^*)\big) = \mathrm{Ev}\big(E^*, (2, \mathbf{x}_2^*)\big),$$

which is (6.3.4). The Conditionality Principle (6.3.2) applied at $(1, \mathbf{x}_1^*)$ and at $(2, \mathbf{x}_2^*)$ gives (6.3.5),

$$\mathrm{Ev}\big(E^*, (1, \mathbf{x}_1^*)\big) = \mathrm{Ev}(E_1, \mathbf{x}_1^*), \qquad \mathrm{Ev}\big(E^*, (2, \mathbf{x}_2^*)\big) = \mathrm{Ev}(E_2, \mathbf{x}_2^*).$$

Chaining the three equalities,

$$\mathrm{Ev}(E_1, \mathbf{x}_1^*) = \mathrm{Ev}\big(E^*, (1, \mathbf{x}_1^*)\big) = \mathrm{Ev}\big(E^*, (2, \mathbf{x}_2^*)\big) = \mathrm{Ev}(E_2, \mathbf{x}_2^*),$$

which is the Formal Likelihood Principle.
(c) Assume the Formal Likelihood Principle.
Conditionality. Take the two experiments of the Formal Likelihood Principle to be E^* and E_j, with sample points $(j, \mathbf{x}_j)$ and $\mathbf{x}_j$. Their likelihoods satisfy

$$L^*\big(\theta \mid (j, \mathbf{x}_j)\big) = f^*\big((j, \mathbf{x}_j) \mid \theta\big) = \tfrac{1}{2} f_j(\mathbf{x}_j \mid \theta) = \tfrac{1}{2} L_j(\theta \mid \mathbf{x}_j)$$

for all θ, proportional with the θ-free constant $\frac{1}{2}$. Hence the Formal Likelihood Principle yields $\mathrm{Ev}(E^*, (j, \mathbf{x}_j)) = \mathrm{Ev}(E_j, \mathbf{x}_j)$, which is (6.3.2), the Conditionality Principle.
Sufficiency. Let $E = (\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$, let $T(\mathbf{X})$ be sufficient for θ, and let $T(\mathbf{x}) = T(\mathbf{y})$. By the Factorization Theorem there are g and h with $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$ for all $\mathbf{x}, \theta$. We may assume $h(\mathbf{y}) \neq 0$, since otherwise $f(\mathbf{y}|\theta) \equiv 0$ and $\mathbf{y}$ is not an observable sample point. Then for all θ,

$$\begin{aligned} L(\theta \mid \mathbf{x}) &= g\big(T(\mathbf{x}) \mid \theta\big) h(\mathbf{x}) = g\big(T(\mathbf{y}) \mid \theta\big) h(\mathbf{x}) \\ &= \frac{h(\mathbf{x})}{h(\mathbf{y})}\, g\big(T(\mathbf{y}) \mid \theta\big) h(\mathbf{y}) = \frac{h(\mathbf{x})}{h(\mathbf{y})}\, L(\theta \mid \mathbf{y}), \end{aligned}$$

so the likelihoods are proportional with the θ-free constant $C = h(\mathbf{x})/h(\mathbf{y})$. Applying the Formal Likelihood Principle with $E_1 = E_2 = E$ gives $\mathrm{Ev}(E, \mathbf{x}) = \mathrm{Ev}(E, \mathbf{y})$, the Formal Sufficiency Principle.

Problem 6.34. Consider the model in Exercise 6.12. There, N is a random variable taking values $1, 2, \ldots$ with known probabilities $p_1, p_2, \ldots$, where $\sum_n p_n = 1$; having observed $N = n$, one performs n Bernoulli trials with success probability θ, getting X successes. Show that the Formal Likelihood Principle implies that any conclusions about θ should not depend on the fact that the sample size n was chosen randomly. That is, the likelihood for (n, x), a sample point from Exercise 6.12, is proportional to the likelihood for the sample point x, a sample point from a fixed-sample-size binomial(n, θ) experiment.

Solution. The two likelihoods differ by the factor p_n, which is free of θ.

Let E_1 be the random-sample-size experiment of Exercise 6.12, with sample point (n, x) and joint pmf

$$f_1\big((n,x) \mid \theta\big) = P(N=n)\,P(X=x \mid N=n) = p_n \binom{n}{x} \theta^x (1-\theta)^{n-x},$$

and let E_2 be the fixed-sample-size binomial(n, θ) experiment, with sample point x and pmf

$$f_2(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

Then for every $\theta \in [0,1]$,

$$L\big(\theta \mid (n,x)\big) = p_n\, L(\theta \mid x),$$

and the constant $C = p_n$ depends on the sample points but not on θ, which is exactly condition (6.3.3). The Formal Likelihood Principle therefore gives

$$\mathrm{Ev}\big(E_1, (n,x)\big) = \mathrm{Ev}(E_2, x).$$

So the evidence about θ from observing x successes after the sample size n arose at random is identical to the evidence from observing x successes in a binomial experiment in which n was fixed in advance: no conclusion about θ may depend on the randomness of the sample size.

Problem 6.35. A risky experimental treatment is to be given to at most three patients. The treatment will be given to one patient. If it is a success, then it will be given to a second. If it is a success, it will be given to a third patient. Model the outcomes for the patients as independent Bernoulli(p) random variables. Identify the four sample points in this model and show that, according to the Formal Likelihood Principle, the inference about p should not depend on the fact that the sample size was determined by the data.

Solution. The four sample points are F, SF, SSF, SSS (S = success, F = failure), with pmf and likelihood as follows.

sample point	no. of trials m	successes s	$P(\cdot \mid p)$	$L(p \mid \cdot)$
F	1	0	$1-p$	$1-p$
SF	2	1	$p(1-p)$	$p(1-p)$
SSF	3	2	$p^2(1-p)$	$p^2(1-p)$
SSS	3	3	p^3	p^3

These exhaust the sample space: $(1-p) + p(1-p) + p^2(1-p) + p^3 = 1$.
Each likelihood has the form $L(p) = p^s(1-p)^{m-s}$ with (m, s) read from the table. Let E_2 be the fixed-sample-size experiment running exactly m independent Bernoulli(p) trials and recording the *sequence* of outcomes; its sample point with the same successes in the same positions has pmf $p^s(1-p)^{m-s}$, so

$$L\big(p \mid \text{sequential point}\big) = 1 \cdot L\big(p \mid \text{fixed-}m \text{ point}\big)$$

for all p, which is (6.3.3) with $C = 1$. (Recording only the count gives $f_2 = \binom{m}{s} p^s (1-p)^{m-s}$ and $C = \binom{m}{s}^{-1}$; either constant is free of p, which is all (6.3.3) asks.) The Formal Likelihood Principle therefore gives

$$\mathrm{Ev}\big(E_1, \text{sequential point}\big) = \mathrm{Ev}\big(E_2, \text{fixed-}m \text{ point}\big),$$

so the inference about p after observing SSF is the same as after observing S, S, F in three trials fixed in advance: the data-determined sample size contributes only the p-free constant and cannot enter any conclusion about p.

6.6 Exercises 6.36–6.42

Problem 6.36. One advantage of using a minimal sufficient statistic is that unbiased estimators will have smaller variance, as the following exercise will show. Suppose that T_1 is sufficient and T_2 is minimal sufficient, U is an unbiased estimator of θ, and define $U_1 = E(U \mid T_1)$ and $U_2 = E(U \mid T_2)$.
(a) Show that $U_2 = E(U_1 \mid T_2)$.
(b) Now use the conditional variance formula (Theorem 4.4.7) to show that $\operatorname{Var} U_2 \le \operatorname{Var} U_1$.
(See Pena and Rohatgi 1994 for more on the relationship between sufficiency and unbiasedness.)

Solution. (a) T_2 is a function of T_1, so conditioning on T_1 and then on T_2 is conditioning on T_2. Precisely: T_2 minimal sufficient and T_1 sufficient give $T_2 = h(T_1)$ (Definition 6.2.11), so $\sigma(T_2) \subseteq \sigma(T_1)$ and conditioning on T_1 already fixes T_2. Averaging over the conditional distribution of T_1 given T_2 (Theorem 4.4.3 applied within that conditional distribution),

$$E(U_1 \mid T_2) = E\big(E(U \mid T_1) \,\big|\, T_2\big) = E(U \mid T_2) = U_2.$$

Both are genuine estimators, free of θ by sufficiency of T_1 and T_2.
(b) Apply Theorem 4.4.7 to U_1 conditioned on T_2, using part (a):

$$\begin{aligned} \operatorname{Var} U_1 &= \operatorname{Var}\big(E(U_1 \mid T_2)\big) + E\big(\operatorname{Var}(U_1 \mid T_2)\big) \\ &= \operatorname{Var} U_2 + E\big(\operatorname{Var}(U_1 \mid T_2)\big) \\ &\ge \operatorname{Var} U_2, \end{aligned}$$

the last step because $\operatorname{Var}(U_1 \mid T_2) \ge 0$ pointwise.

Problem 6.37. Joshi and Nabar (1989) examine properties of linear estimators for the parameter in the so-called "Problem of the Nile," where (X, Y) has the joint density

$$f(x, y \mid \theta) = \exp\{-(\theta x + y/\theta)\}, \qquad x > 0,\ y > 0.$$

(a) For an iid sample of size n, show that the Fisher information is $I(\theta) = 2n/\theta^2$.
(b) For the estimators

$$T = \sqrt{\textstyle\sum Y_i / \sum X_i} \quad \text{and} \quad U = \sqrt{\textstyle\sum X_i \sum Y_i},$$

show that
(i) the information in T alone is $[2n/(2n+1)]I(\theta)$;
(ii) the information in (T, U) is $I(\theta)$;
(iii) (T, U) is jointly sufficient but not complete.

Solution. The density factors, so X and Y are independent with $X \sim$ exponential$(1/\theta)$ (rate θ) and $Y \sim$ exponential(θ); in particular $EY = \theta$ and the density is already normalized.
(a) For a single pair, $\log f = -\theta x - y/\theta$, so

$$\frac{\partial}{\partial\theta} \log f = -x + \frac{y}{\theta^2}, \qquad \frac{\partial^2}{\partial\theta^2} \log f = -\frac{2y}{\theta^3},$$

whence $I_1(\theta) = 2EY/\theta^3 = 2\theta/\theta^3 = 2/\theta^2$, and by additivity over iid pairs $I(\theta) = 2n/\theta^2$.
(b) Put $S = \sum_{i=1}^n X_i$ and $W = \sum_{i=1}^n Y_i$. These are independent with

$$G_1 := \theta S \sim \text{gamma}(n, 1), \qquad G_2 := W/\theta \sim \text{gamma}(n, 1),$$

and $T = \sqrt{W/S} = \theta\sqrt{G_2/G_1}$, $U = \sqrt{SW} = \sqrt{G_1 G_2}$.
(i) $T = \theta V$ with $V = (G_2/G_1)^{1/2}$ free of θ. The ratio $R = G_2/G_1$ has the beta-prime density $f_R(r) = \frac{\Gamma(2n)}{\Gamma(n)^2}\, r^{n-1}(1+r)^{-2n}$, so $V = \sqrt{R}$ has $f_V(v) = \frac{2\Gamma(2n)}{\Gamma(n)^2}\, v^{2n-1}(1+v^2)^{-2n}$ and

$$f_T(t \mid \theta) = \tfrac{1}{\theta} f_V(t/\theta) = \frac{2\Gamma(2n)}{\Gamma(n)^2}\, \frac{t^{2n-1}\theta^{2n}}{(\theta^2 + t^2)^{2n}}, \qquad t > 0.$$

Differentiating $\log f_T = c + (2n-1)\log t + 2n\log\theta - 2n\log(\theta^2+t^2)$ twice,

$$\frac{\partial^2}{\partial\theta^2}\log f_T = -\frac{2n}{\theta^2} - 4n\,\frac{t^2-\theta^2}{(\theta^2+t^2)^2},$$

so, substituting $T=\theta V$,

$$I_T(\theta) = \frac{2n}{\theta^2}\left(1 + 2\,E\left[\frac{V^2-1}{(1+V^2)^2}\right]\right).$$

Now $B := V^2/(1+V^2) = G_2/(G_1+G_2) \sim \text{beta}(n,n)$, and

$$\frac{V^2-1}{(1+V^2)^2} = B(1-B) - (1-B)^2 = (1-B)(2B-1),$$

whose expectation is $3EB - 1 - 2EB^2 = \frac{3}{2} - 1 - \frac{n+1}{2n+1} = -\frac{1}{2(2n+1)}$, using $EB = \frac{1}{2}$ and $EB^2 = \frac{n+1}{2(2n+1)}$. Therefore

$$I_T(\theta) = \frac{2n}{\theta^2}\left(1 - \frac{1}{2n+1}\right) = \frac{2n}{2n+1}\,I(\theta).$$

(ii) The map is one-to-one: $TU = W$ and $U/T = S$, so (T,U) and (S,W) generate the same information. The joint density of (S,W) is

$$g(s,w\mid\theta) = \frac{\theta^n}{\Gamma(n)}s^{n-1}e^{-\theta s}\cdot\frac{1}{\Gamma(n)\theta^n}w^{n-1}e^{-w/\theta} = \frac{s^{n-1}w^{n-1}}{\Gamma(n)^2}e^{-\theta s - w/\theta},$$

so $\partial_\theta^2 \log g = -2w/\theta^3$ and $I_{(T,U)}(\theta) = 2EW/\theta^3 = 2n/\theta^2 = I(\theta)$.
(iii) Sufficiency: $f(\mathbf{x},\mathbf{y}\mid\theta) = \exp\{-\theta s - w/\theta\}\cdot\mathbf{1}_{\{x_i>0,\,y_i>0\}}$ depends on the sample only through (s,w), hence through the equivalent pair (T,U) (Theorem 6.2.6). Incompleteness: $U = \sqrt{G_1G_2}$ has a distribution free of θ, i.e. U is ancillary, with finite mean

$$EU = \left(E\sqrt{G_1}\right)^2 = \left(\frac{\Gamma(n+\frac{1}{2})}{\Gamma(n)}\right)^2 =: c_n.$$

Then $h(T,U) = U - c_n$ satisfies $E_\theta h(T,U) = 0$ for every $\theta > 0$ while $P_\theta(h(T,U)=0) = 0$, so (T,U) is not complete.

Problem 6.38. In Definition 6.4.2, show that (iii) is implied by (i) and (ii).
Recall Definition 6.4.2: a set of functions $\{g(\mathbf{x}) : g\in\mathcal{G}\}$ from the sample space $\mathcal{X}$ onto $\mathcal{X}$ is called a group of transformations of $\mathcal{X}$ if
(i) (Inverse) For every $g\in\mathcal{G}$ there is a $g'\in\mathcal{G}$ such that $g'(g(\mathbf{x})) = \mathbf{x}$ for all $\mathbf{x}\in\mathcal{X}$.
(ii) (Composition) For every $g\in\mathcal{G}$ and $g'\in\mathcal{G}$ there exists $g''\in\mathcal{G}$ such that $g'(g(\mathbf{x})) = g''(\mathbf{x})$ for all $\mathbf{x}\in\mathcal{X}$.
(iii) (Identity) The identity, $e(\mathbf{x})$, defined by $e(\mathbf{x}) = \mathbf{x}$, is an element of $\mathcal{G}$.

Solution. Compose any element with its inverse.
Let $g\in\mathcal{G}$ (the set is nonempty by hypothesis). By (i) there is $g'\in\mathcal{G}$ with $g'(g(\mathbf{x})) = \mathbf{x}$ for all $\mathbf{x}\in\mathcal{X}$. By (ii) applied to this pair g, g' there is $g''\in\mathcal{G}$ with

$$g''(\mathbf{x}) = g'(g(\mathbf{x})) = \mathbf{x} \qquad \text{for all } \mathbf{x}\in\mathcal{X}.$$

Thus $g'' = e$, and $e\in\mathcal{G}$.

Problem 6.39. Measurement equivariance requires the same inference for two equivalent data points: $\mathbf{x}$, measurements expressed in one scale, and $\mathbf{y}$, *exactly the same measurements* expressed in a different scale. Formal invariance, in the end, leads to a relationship between the inferences at two *different*

data points in the same measurement scale.
Suppose an experimenter wishes to estimate θ, the mean boiling point of water, based on a single observation X, the boiling point measured in degrees Celsius. Because of the altitude and impurities in the water he decides to use the estimate $T(x) = .5x + .5(100)$. If the measurement scale is changed to degrees Fahrenheit, the experimenter would use $T^*(y) = .5y + .5(212)$ to estimate the mean boiling point expressed in degrees Fahrenheit.
(a) The familiar relation between degrees Celsius and degrees Fahrenheit would lead us to convert Fahrenheit to Celsius using the transformation $\frac{5}{9}(T^*(y) - 32)$. Show that this procedure is measurement equivariant in that the same answer will be obtained for the same data; that is, $\frac{5}{9}(T^*(y) - 32) = T(x)$.
(b) Formal invariance would require that $T(x) = T^*(x)$ for all x. Show that the estimators we have defined above do not satisfy this. So they are not equivariant in the sense of the Equivariance Principle.

Solution. (a) The same data in the two scales are x and $y = \frac{9}{5}x + 32$, and substituting this into T^* collapses to T:

$$\begin{aligned}\tfrac{5}{9}\big(T^*(y) - 32\big) &= \tfrac{5}{9}\big(.5(\tfrac{9}{5}x + 32) + 106 - 32\big)\\ &= \tfrac{5}{9}\big(.9x + 16 + 74\big)\\ &= \tfrac{5}{9}(.9x + 90)\\ &= .5x + 50 = T(x).\end{aligned}$$

(b) $T(x) = .5x + 50$ while $T^*(x) = .5x + 106$, so

$$T^*(x) - T(x) = 56 \neq 0 \qquad \text{for every } x,$$

and formal invariance fails at every sample point. Since the Equivariance Principle demands both measurement equivariance and formal invariance, these estimators are not equivariant in its sense.

Problem 6.40. Let $X_1, \ldots, X_n$ be iid observations from a location-scale family. Let $T_1(X_1, \ldots, X_n)$ and $T_2(X_1, \ldots, X_n)$ be two statistics that both satisfy

$$T_i(ax_1 + b, \ldots, ax_n + b) = a\,T_i(x_1, \ldots, x_n)$$

for all values of $x_1, \ldots, x_n$ and b and for any $a > 0$.
(a) Show that T_1/T_2 is an ancillary statistic.
(b) Let R be the sample range and S be the sample standard deviation. Verify that R and S satisfy the above condition so that R/S is an ancillary statistic.

Solution. (a) The scale factor cancels in the ratio, leaving a function of the standardized sample alone. By the representation in Theorem 3.5.6 write $X_i = \mu + \sigma Z_i$, where $Z_1, \ldots, Z_n$ are iid from the standard member $f(z)$ of the family and the distribution of $\mathbf{Z}$ involves neither μ nor σ. Applying the hypothesis with $a = \sigma > 0$ and $b = \mu$,

$$T_i(\mathbf{X}) = T_i(\sigma Z_1 + \mu, \ldots, \sigma Z_n + \mu) = \sigma\,T_i(\mathbf{Z}), \qquad i = 1, 2,$$

so on $\{T_2(\mathbf{Z}) \neq 0\}$,

$$\frac{T_1(\mathbf{X})}{T_2(\mathbf{X})} = \frac{\sigma\,T_1(\mathbf{Z})}{\sigma\,T_2(\mathbf{Z})} = \frac{T_1(\mathbf{Z})}{T_2(\mathbf{Z})}.$$

The right-hand side is a fixed function of $\mathbf{Z}$, whose distribution is free of (μ, σ); hence so is the distribution of T_1/T_2, which is therefore ancillary (Definition 6.2.16).
(b) The map $x_i \mapsto ax_i + b$ with $a > 0$ preserves order, so $(ax + b)_{(j)} = ax_{(j)} + b$ and

$$R(a\mathbf{x} + b) = \big(ax_{(n)} + b\big) - \big(ax_{(1)} + b\big) = aR(\mathbf{x}).$$

For S, the mean transforms as $\overline{a\mathbf{x}+b} = a\bar{x}+b$, so each deviation scales: $(ax_i+b)-(a\bar{x}+b) = a(x_i-\bar{x})$, and

$$S(a\mathbf{x}+b) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} a^2(x_i-\bar{x})^2} = |a|S(\mathbf{x}) = aS(\mathbf{x}).$$

Both satisfy the hypothesis, so by (a) R/S is ancillary (with $S > 0$ a.s. whenever the family is continuous).

Problem 6.41. Suppose that for the model in Example 6.4.6, the inference to be made is an estimate of the mean μ. (In that example $X_1, \ldots, X_n$ are iid $\mathrm{n}(\mu, \sigma^2)$, both μ and σ^2 unknown, and $\mathcal{G} = \{g_a(\mathbf{x}) : -\infty < a < \infty\}$ with $g_a(x_1, \ldots, x_n) = (x_1 + a, \ldots, x_n + a)$.) Let $T(\mathbf{x})$ be the estimate used if $\mathbf{X} = \mathbf{x}$ is observed. If $g_a(\mathbf{X}) = \mathbf{Y} = \mathbf{y}$ is observed, then let $T^*(\mathbf{y})$ be the estimate of $\mu + a$, the mean of each Y_i. If $\mu + a$ is estimated by $T^*(\mathbf{y})$, then μ would be estimated by $T^*(\mathbf{y}) - a$.
(a) Show that measurement equivariance requires that $T(\mathbf{x}) = T^*(\mathbf{y}) - a$ for all $\mathbf{x} = (x_1, \ldots, x_n)$ and all a.
(b) Show that formal invariance requires that $T(\mathbf{x}) = T^*(\mathbf{x})$ and hence the Equivariance Principle requires that $T(x_1, \ldots, x_n) + a = T(x_1 + a, \ldots, x_n + a)$ for all $(x_1, \ldots, x_n)$ and all a.
(c) If $X_1, \ldots, X_n$ are iid $f(x-\theta)$, show that, as long as $E_0 X_1 = 0$, the estimator $W(X_1, \ldots, X_n) = \bar{X}$ is equivariant for estimating θ and satisfies $E_\theta W = \theta$.

Solution. (a) $\mathbf{x}$ and $\mathbf{y} = g_a(\mathbf{x})$ are the same measurements in two scales (the second shifted by a), so the two resulting estimates of the same quantity μ must agree. In the $\mathbf{x}$-scale that estimate is $T(\mathbf{x})$; in the $\mathbf{y}$-scale it is $T^*(\mathbf{y}) - a$, since $T^*(\mathbf{y})$ estimates $\mu + a$. Measurement equivariance is precisely the statement

$$T(\mathbf{x}) = T^*(\mathbf{y}) - a = T^*(x_1 + a, \ldots, x_n + a) - a$$

for all $\mathbf{x}$ and all a.
(b) By Example 6.4.6 the family $\mathcal{F}$ is invariant under $\mathcal{G}$: $\mathbf{Y} = g_a(\mathbf{X})$ is again an iid $\mathrm{n}(\mu+a, \sigma^2)$ sample, with the same sample space $\mathbb{R}^n$ and the same parameter space $\mathbb{R} \times (0, \infty)$. The two inference problems – estimate the mean from an iid normal sample – therefore have identical formal structure, and formal invariance demands the same procedure in both, $T^* = T$. Substituting $T^* = T$ into (a),

$$T(x_1, \ldots, x_n) = T(x_1 + a, \ldots, x_n + a) - a,$$

i.e. $T(x_1, \ldots, x_n) + a = T(x_1 + a, \ldots, x_n + a)$ for all $\mathbf{x}$ and all a.
(c) Equivariance is the identity $\overline{\mathbf{x} + a} = \bar{x} + a$:

$$W(x_1 + a, \ldots, x_n + a) = \frac{1}{n}\sum_{i=1}^{n}(x_i + a) = \bar{x} + a = W(\mathbf{x}) + a,$$

which is exactly the condition in (b). For the mean, $X_1 - \theta$ has density f, so $E_\theta X_1 = \theta + E_0 X_1 = \theta$, and

$$E_\theta W = \frac{1}{n}\sum_{i=1}^{n} E_\theta X_i = \theta.$$

Problem 6.42. Suppose we have a random sample $X_1, \ldots, X_n$ from $\frac{1}{\sigma} f((x-\theta)/\sigma)$, a location-scale pdf. We want to estimate θ, and we have two groups of transformations under consideration:

$$\mathcal{G}_1 = \{g_{a,c}(\mathbf{x}) : -\infty < a < \infty,\ c > 0\},$$

where $g_{a,c}(x_1, \ldots, x_n) = (cx_1 + a, \ldots, cx_n + a)$, and

$$\mathcal{G}_2 = \{g_a(\mathbf{x}) : -\infty < a < \infty\},$$

where $g_a(x_1, \ldots, x_n) = (x_1 + a, \ldots, x_n + a)$.
(a) Show that estimators of the form

$$W(x_1, \ldots, x_n) = \bar{x} + k,$$

where k is a nonzero constant, are equivariant with respect to the group $\mathcal{G}_2$ but are not equivariant with respect to the group $\mathcal{G}_1$.
(b) For each group, under what conditions does an equivariant estimator W satisfy $E_\theta W = \theta$, that is, it is unbiased for estimating θ?

Solution. (a) A shift carries $\bar{x} + k$ along unchanged, but a rescaling multiplies the nuisance constant k too, and only $k = 0$ survives that.
By the argument of Exercise 6.41(a)-(b): if $Y_i = cX_i + a$ then $\mathbf{Y}$ is a random sample from $\frac{1}{c\sigma} f\big((y - (c\theta + a))/(c\sigma)\big)$, so the family is invariant with $\theta' = c\theta + a$, $\sigma' = c\sigma$; measurement equivariance gives $T(\mathbf{x}) = (T^*(\mathbf{y}) - a)/c$ and formal invariance gives $T^* = T$, so the Equivariance Principle requires

$$\begin{aligned} \mathcal{G}_1: &\quad W(cx_1 + a, \ldots, cx_n + a) = cW(\mathbf{x}) + a \quad (c > 0), \\ \mathcal{G}_2: &\quad W(x_1 + a, \ldots, x_n + a) = W(\mathbf{x}) + a. \end{aligned}$$

For $\mathcal{G}_2$:

$$W(x_1 + a, \ldots, x_n + a) = (\bar{x} + a) + k = W(\mathbf{x}) + a,$$

so W is $\mathcal{G}_2$-equivariant for every k. For $\mathcal{G}_1$,

$$W(cx_1 + a, \ldots, cx_n + a) = c\bar{x} + a + k, \qquad cW(\mathbf{x}) + a = c\bar{x} + ck + a,$$

and these agree for all $c > 0$ only if $k = ck$ for all $c > 0$, i.e. $k = 0$. Since $k \neq 0$, W is not $\mathcal{G}_1$-equivariant.
(b) Write $X_i = \theta + \sigma Z_i$ with $Z_1, \ldots, Z_n$ iid from the standard pdf f (Theorem 3.5.6), so that the distribution of $\mathbf{Z}$ is free of (θ, σ). Assume $E|W| < \infty$.
(i) $\mathcal{G}_2$-equivariance, applied with $a = \theta$ to the sample $\sigma\mathbf{Z}$, gives $W(\mathbf{X}) = W(\sigma\mathbf{Z}) + \theta$, hence

$$E_\theta W = \theta + b(\sigma), \qquad b(\sigma) := E\, W(\sigma Z_1, \ldots, \sigma Z_n).$$

So W is unbiased for θ if and only if $b(\sigma) = 0$ for every $\sigma > 0$: a whole family of conditions, one per σ.
(ii) $\mathcal{G}_1$-equivariance, applied with $c = \sigma$ and $a = \theta$, gives $W(\mathbf{X}) = \sigma W(\mathbf{Z}) + \theta$, hence

$$E_\theta W = \theta + \sigma\, b, \qquad b := E\, W(Z_1, \ldots, Z_n),$$

with b a single constant, free of both parameters. So W is unbiased for θ if and only if $b = 0$, that is, $E_{\theta=0,\sigma=1} W = 0$: one condition rather than a family of them.

6.7 Exercises 6.43–6.43

Problem 6.43. Again, suppose we have a random sample $X_1, \ldots, X_n$ from $\frac{1}{\sigma} f((x-\theta)/\sigma)$, a location–scale pdf, but we are now interested in estimating σ^2. We can consider three groups of transformations:

$$\mathcal{G}_1 = \{g_{a,c}(\mathbf{x}) : -\infty < a < \infty,\ c > 0\},$$

where $g_{a,c}(x_1, \ldots, x_n) = (cx_1 + a, \ldots, cx_n + a)$;

$$\mathcal{G}_2 = \{g_a(\mathbf{x}) : -\infty < a < \infty\},$$

where $g_a(x_1, \ldots, x_n) = (x_1 + a, \ldots, x_n + a)$; and

$$\mathcal{G}_3 = \{g_c(\mathbf{x}) : c > 0\},$$

where $g_c(x_1, \ldots, x_n) = (cx_1, \ldots, cx_n)$.

(a) Show that estimators of σ^2 of the form kS^2, where k is a positive constant and S^2 is the sample variance, are invariant with respect to $\mathcal{G}_2$ and equivariant with respect to the other two groups.
(b) Show that the larger class of estimators of σ^2 of the form

$$W(X_1, \ldots, X_n) = \phi\left(\frac{\bar{X}}{S}\right) S^2,$$

where $\phi(x)$ is a function, are equivariant with respect to $\mathcal{G}_3$ but not with respect to either $\mathcal{G}_1$ or $\mathcal{G}_2$, unless $\phi(x)$ is a constant (Brewster and Zidek 1974).
Consideration of estimators of this form led Stein (1964) and Brewster and Zidek (1974) to find improved estimators of variance (see Lehmann and Casella 1998, Section 3.3).

Solution. Everything follows from the two identities

$$\begin{aligned} \bar{x}(g_{a,c}(\mathbf{x})) &= c\,\bar{x}(\mathbf{x}) + a, \\ S^2(g_{a,c}(\mathbf{x})) &= c^2\,S^2(\mathbf{x}) \qquad (c > 0), \end{aligned}$$

the second because every deviation $x_i - \bar{x}$ is scaled by c while the shift a cancels.
Fix the induced action on the parameter first (Definition 6.4.4): if X_i has pdf $\frac{1}{\sigma} f((x-\theta)/\sigma)$ and $Y_i = cX_i + a$ with $c > 0$, then

$$f_{Y_i}(y) = \frac{1}{c\sigma}\, f\left(\frac{y - (c\theta + a)}{c\sigma}\right),$$

so each of $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ leaves the family invariant, with $(\theta, \sigma^2) \mapsto (c\theta + a,\, c^2\sigma^2)$. An estimator W of σ^2 is therefore equivariant for a group exactly when $W(g_{a,c}(\mathbf{x})) = c^2 W(\mathbf{x})$ for all its members; for $\mathcal{G}_2$, where $c = 1$, that requirement is invariance.
(a) For $W = kS^2$ the second identity gives

$$W(g_{a,c}(\mathbf{x})) = k\,c^2 S^2(\mathbf{x}) = c^2 W(\mathbf{x}),$$

equivariance for $\mathcal{G}_1$ and, at $a = 0$, for $\mathcal{G}_3$; at $c = 1$ it reads $W(g_a(\mathbf{x})) = W(\mathbf{x})$, invariance for $\mathcal{G}_2$.
(b) With $S = \sqrt{S^2} > 0$, so $S(g_{a,c}(\mathbf{x})) = cS(\mathbf{x})$,

$$\begin{aligned} W(g_{a,c}(\mathbf{x})) &= \phi\left(\frac{c\bar{x} + a}{cS}\right)(cS)^2 \\ &= c^2\,\phi\Big(\frac{\bar{x}}{S} + \frac{a}{cS}\Big)\, S^2. \end{aligned}$$

(i) $\mathcal{G}_3$ ($a = 0$): the argument of ϕ is untouched, so $W(g_c(\mathbf{x})) = c^2 W(\mathbf{x})$ for every ϕ.
(ii) $\mathcal{G}_2$ ($c = 1$): invariance demands $\phi(\bar{x}/S + a/S) = \phi(\bar{x}/S)$ for all a and all $\mathbf{x}$. Given $t_1, t_2 \in \mathbb{R}$, pick $\mathbf{x}$ with $S(\mathbf{x}) = 1$ and $\bar{x} = t_1$ (translate any $\mathbf{x}_0$ with $S(\mathbf{x}_0) = 1$, which exists as $n \geq 2$) and take $a = t_2 - t_1$: this forces $\phi(t_2) = \phi(t_1)$, so ϕ is constant.
(iii) $\mathcal{G}_1$: the requirement $\phi(\bar{x}/S + a/(cS)) = \phi(\bar{x}/S)$ at $c = 1$ is exactly (ii), so again ϕ is constant.

7 Point Estimation

Contents

7.1 Exercises 7.1–7.7 . . . 226
7.2 Exercises 7.8–7.14 . . . 229
7.3 Exercises 7.15–7.21 . . . 234
7.4 Exercises 7.22–7.28 . . . 241
7.5 Exercises 7.29–7.35 . . . 248
7.6 Exercises 7.36–7.42 . . . 256
7.7 Exercises 7.43–7.49 . . . 260
7.8 Exercises 7.50–7.56 . . . 265
7.9 Exercises 7.57–7.63 . . . 271
7.10 Exercises 7.64–7.66 . . . 275

7.1 Exercises 7.1–7.7

Problem 7.1. One observation is taken on a discrete random variable X with pmf $f(x|\theta)$, where $\theta \in \{1, 2, 3\}$. Find the MLE of θ.

x	$f(x \mid 1)$	$f(x \mid 2)$	$f(x \mid 3)$
0	1/3	1/4	0
1	1/3	1/4	0
2	0	1/4	1/4
3	1/6	1/4	1/2
4	1/6	0	1/4

Solution.

$$\hat{\theta}(x) = \begin{cases} 1 & x = 0, 1, \\ 2 \text{ or } 3 & x = 2, \\ 3 & x = 3, 4. \end{cases}$$

With one observation $L(\theta|x) = f(x|\theta)$ is the row of the table indexed by the observed x, so the MLE is the column label attaining the row maximum. (Check the five rows!) At $x = 2$ the maximum $\frac{1}{4}$ is attained twice, so the MLE is not unique there.

Problem 7.2. Let $X_1, \ldots, X_n$ be a random sample from a gamma(α, β) population.
(a) Find the MLE of β, assuming α is known.
(b) If α and β are both unknown, there is no explicit formula for the MLEs of α and β, but the maximum can be found numerically. The result in part (a) can be used to reduce the problem to the maximization of a univariate function. Find the MLEs for α and β for the data in Exercise 7.10(c), namely the lengths (in millimeters) of $n = 14$ cuckoos' eggs found in hedge sparrow nests:

$$22.0,\ 23.9,\ 20.9,\ 23.8,\ 25.0,\ 24.0,\ 21.7,$$
$$23.8,\ 22.8,\ 23.1,\ 23.1,\ 23.5,\ 23.0,\ 23.0.$$

Solution. (a) $\hat{\beta} = \bar{X}/\alpha$. The log likelihood is

$$\log L(\beta|\mathbf{x}) = -n \log \Gamma(\alpha) - n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \frac{1}{\beta} \sum_{i=1}^{n} x_i,$$

so

$$\frac{\partial}{\partial \beta} \log L = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{n} x_i = 0 \quad \Longrightarrow \quad \hat{\beta} = \frac{\sum_i x_i}{n\alpha} = \frac{\bar{x}}{\alpha}.$$

The derivative is positive for $\beta < \bar{x}/\alpha$ and negative for $\beta > \bar{x}/\alpha$, so this stationary point is the global maximum.
(b) Substituting $\beta = \bar{x}/\alpha$ from (a) leaves the profile log likelihood

$$\ell(\alpha) = -n \log \Gamma(\alpha) - n\alpha \log \frac{\bar{x}}{\alpha} + (\alpha - 1) \sum_{i=1}^{n} \log x_i - n\alpha,$$

whose derivative gives the single equation

$$\log \alpha - \psi(\alpha) = \log \bar{x} - \frac{1}{n} \sum_{i=1}^{n} \log x_i, \qquad \psi = \frac{\Gamma'}{\Gamma}.$$

The left side decreases strictly from $+\infty$ to 0 on $(0, \infty)$, so the root is unique. For the cuckoo data, $\bar{x} = 23.1143$ and $n^{-1} \sum \log x_i = 3.139478$, so the right side is $c = 0.00097244$; solving numerically,

$$\hat{\alpha} = 514.34, \qquad \hat{\beta} = \frac{\bar{x}}{\hat{\alpha}} = 0.04494.$$

Problem 7.3. Given a random sample $X_1, \ldots, X_n$ from a population with pdf $f(x|\theta)$, show that maximizing the likelihood function, $L(\theta|\mathbf{x})$, as a function of θ is equivalent to maximizing $\log L(\theta|\mathbf{x})$.

Solution. Because log is strictly increasing on $(0, \infty)$, it preserves the ordering of likelihood values, hence the set of maximizers.
Fix $\mathbf{x}$ and write $L(\theta) = \prod_i f(x_i|\theta) \geq 0$, with the convention $\log 0 = -\infty$. For any $\theta_1, \theta_2 \in \Theta$,

$$L(\theta_1) \leq L(\theta_2) \iff \log L(\theta_1) \leq \log L(\theta_2),$$

since log and its inverse exp are both strictly increasing. Taking $\theta_2 = \hat{\theta}$ and letting θ_1 range over Θ: $\hat{\theta}$ maximizes L if and only if it maximizes $\log L$, so the two have the same set of maximizers.

Problem 7.4. Prove the assertion in Example 7.2.8. That is, let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, 1)$ where it is known that θ must be nonnegative, and prove that

$$\hat{\theta} = \bar{X} \ \text{ if } \bar{X} \geq 0, \qquad \hat{\theta} = 0 \ \text{ if } \bar{X} < 0$$

is the MLE when the range of θ is restricted to the nonnegative axis.

Solution. $\hat{\theta} = \max\{\bar{x}, 0\}$, because the likelihood factors as

$$L(\theta|\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp\Big(-\tfrac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \Big) \exp\Big(-\tfrac{n}{2} (\bar{x} - \theta)^2 \Big),$$

using $\sum_i (x_i - \theta)^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2$. Only the last factor involves θ, so L strictly increases in θ on $(-\infty, \bar{x}]$ and strictly decreases on $[\bar{x}, \infty)$.
(i) If $\bar{x} \geq 0$, the unconstrained maximizer $\bar{x}$ lies in $\Theta = [0, \infty)$, so $\hat{\theta} = \bar{x}$.
(ii) If $\bar{x} < 0$, then $\theta \geq 0 > \bar{x}$ for every admissible θ, so L is strictly decreasing on all of $[0, \infty)$ and attains its maximum at the left endpoint: $\hat{\theta} = 0$.

Problem 7.5. Consider estimating the binomial parameter k as in Example 7.2.9. There $X_1, \ldots, X_n$ is a random sample from a binomial(k, p) population with p known and k unknown, so that

$$L(k|\mathbf{x}, p) = \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i} (1-p)^{k - x_i},$$

and the MLE is the integer $k \geq \max_i x_i$ satisfying

$$\frac{L(k|\mathbf{x}, p)}{L(k-1|\mathbf{x}, p)} \geq 1 \quad \text{and} \quad \frac{L(k+1|\mathbf{x}, p)}{L(k|\mathbf{x}, p)} < 1,$$

equivalently

$$\big(k(1-p)\big)^n \geq \prod_{i=1}^{n} (k - x_i) \quad \text{and} \quad \big((k+1)(1-p)\big)^n < \prod_{i=1}^{n} (k + 1 - x_i).$$

Dividing by k^n and setting $z = 1/k$ leads to the equation

$$(1-p)^n = \prod_{i=1}^{n} (1 - x_i z), \qquad 0 \leq z \leq 1/\max_i x_i,$$

whose root $\hat{z}$ is unique.

(a) Prove the assertion that the integer k that satisfies the inequalities and is the MLE is the largest integer less than or equal to $1/\hat{z}$.
(b) Let $p = \frac{1}{2}$, $n = 4$, and $X_1 = 0$, $X_2 = 20$, $X_3 = 1$, and $X_4 = 19$. What is $\hat{k}$?

Solution. (a) $\hat{k} = \lfloor 1/\hat{z} \rfloor$. Put

$$h(z) = \prod_{i=1}^{n}(1 - x_i z), \qquad z \in [0,\, 1/\max_i x_i],$$

a strictly decreasing bijection of $[0, 1/\max_i x_i]$ onto $[0, 1]$ (every factor is nonnegative and nonincreasing there, and the one with $x_i = \max_i x_i$ strictly decreases), so $\hat{z}$ solving $h(z) = (1-p)^n$ is unique.
For an integer $k \geq \max_i x_i$, divide the first displayed inequality by k^n and the second by $(k+1)^n$:

$$h\big(\tfrac{1}{k}\big) \leq (1-p)^n = h(\hat{z}) \quad \text{and} \quad h\Big(\tfrac{1}{k+1}\Big) > h(\hat{z}),$$

and both arguments $1/k,\ 1/(k+1)$ lie in $[0, 1/\max_i x_i]$. Applying the strictly decreasing h^{-1} reverses each inequality:

$$\frac{1}{k} \;\geq\; \hat{z} \quad \text{and} \quad \frac{1}{k+1} \;<\; \hat{z}, \qquad \text{i.e.} \qquad \frac{1}{\hat{z}} - 1 \;<\; k \;\leq\; \frac{1}{\hat{z}}.$$

Exactly one integer lies in $(1/\hat{z} - 1,\ 1/\hat{z}\,]$, namely $\lfloor 1/\hat{z} \rfloor$, giving both the uniqueness claimed in Example 7.2.9 and the stated form of the MLE.
(b) $\hat{k} = 26$. Here $\max_i x_i = 20$ and, since $x_1 = 0$ contributes the factor 1,

$$\tfrac{1}{16} = (1 - \tfrac{1}{2})^4 = (1 - 20z)(1 - z)(1 - 19z) = 1 - 40z + 419z^2 - 380z^3,$$

i.e. $6080z^3 - 6704z^2 + 640z - 15 = 0$. Its roots are 0.03817, 0.06464, 0.99982, and only the first lies in $[0, 1/20]$, so $\hat{z} = 0.0381730$ and

$$\frac{1}{\hat{z}} = 26.1965, \qquad \hat{k} = \lfloor 26.1965 \rfloor = 26.$$

Problem 7.6. Let $X_1, \ldots, X_n$ be a random sample from the pdf

$$f(x|\theta) = \theta x^{-2}, \qquad 0 < \theta \leq x < \infty.$$

(a) What is a sufficient statistic for θ?
(b) Find the MLE of θ.
(c) Find the method of moments estimator of θ.

Solution. (a) $X_{(1)} = \min_i X_i$. The joint pdf is

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \theta x_i^{-2} I_{[\theta,\infty)}(x_i) = \underbrace{\theta^n I_{[\theta,\infty)}\big(x_{(1)}\big)}_{g(x_{(1)}|\theta)} \cdot \underbrace{\prod_{i=1}^{n} x_i^{-2}}_{h(\mathbf{x})},$$

since $x_i \geq \theta$ for all i exactly when $x_{(1)} \geq \theta$. The Factorization Theorem (Theorem 6.2.6) applies, so $X_{(1)}$ is sufficient.
(b) $\hat{\theta} = X_{(1)}$. By the display above $L(\theta|\mathbf{x}) = \theta^n h(\mathbf{x})$ for $0 < \theta \leq x_{(1)}$ and 0 for $\theta > x_{(1)}$; the first branch strictly increases in θ, so the maximum is at the largest admissible value $\theta = x_{(1)}$.
(c) It does not exist:

$$\mathrm{E}_\theta X = \int_\theta^\infty x \cdot \theta x^{-2}\, dx = \theta \int_\theta^\infty \frac{dx}{x} = \theta \lim_{b\to\infty} \log \frac{b}{\theta} = \infty.$$

The first population moment is infinite, so there is no equation $\mathrm{E}_\theta X = \bar{X}$ to solve and the method of

moments estimator is undefined.

Problem 7.7. Let $X_1, \dots, X_n$ be iid with one of two pdfs. If $\theta = 0$, then

$$f(x|\theta) = \begin{cases} 1 & \text{if } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

while if $\theta = 1$, then

$$f(x|\theta) = \begin{cases} 1/(2\sqrt{x}) & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the MLE of θ.

Solution.

$$\hat{\theta} = \begin{cases} 1 & \text{if } \prod_{i=1}^n x_i < 4^{-n}, \\ 0 & \text{if } \prod_{i=1}^n x_i \geq 4^{-n}. \end{cases}$$

The parameter space is $\Theta = \{0, 1\}$, so maximizing the likelihood is just comparing two numbers. For $\mathbf{x} \in (0,1)^n$,

$$L(0|\mathbf{x}) = 1, \qquad L(1|\mathbf{x}) = \prod_{i=1}^n \frac{1}{2\sqrt{x_i}} = 2^{-n}\Big(\prod_{i=1}^n x_i\Big)^{-1/2},$$

and therefore

$$L(1|\mathbf{x}) > L(0|\mathbf{x}) \iff \Big(\prod_i x_i\Big)^{1/2} < 2^{-n} \iff \prod_{i=1}^n x_i < 4^{-n}.$$

At the boundary $\prod_i x_i = 4^{-n}$ the two likelihoods tie and either value is an MLE; above it is assigned to $\theta = 0$.

7.2 Exercises 7.8–7.14

Problem 7.8. One observation, X, is taken from a $n(0, \sigma^2)$ population.
(a) Find an unbiased estimator of σ^2.
(b) Find the MLE of σ.
(c) Discuss how the method of moments estimator of σ might be found.

Solution. (a) X^2, since $\mathrm{E}X^2 = \operatorname{Var} X + (\mathrm{E}X)^2 = \sigma^2$.
(b) $\hat{\sigma} = |X|$. With $L(\sigma \mid x) = (2\pi\sigma^2)^{-1/2} e^{-x^2/(2\sigma^2)}$,

$$\log L = -\tfrac{1}{2}\log(2\pi) - \log\sigma - \frac{x^2}{2\sigma^2}, \qquad \frac{\partial \log L}{\partial \sigma} = -\frac{1}{\sigma} + \frac{x^2}{\sigma^3} = \frac{x^2 - \sigma^2}{\sigma^3},$$

which is positive for $\sigma < |x|$ and negative for $\sigma > |x|$, so $\hat{\sigma} = |x|$ is the global maximum on $\sigma > 0$. (At $x = 0$, an event of probability 0, the likelihood is unbounded as $\sigma \to 0$.)
(c) The first moment carries no information, $\mathrm{E}_\sigma X = 0$ for every σ, so the order-one equation is vacuous and one matches the lowest moment that does depend on σ:

$$\mathrm{E}_\sigma X^2 = \sigma^2 = X^2 \quad \Longrightarrow \quad \tilde{\sigma} = |X|,$$

reproducing the MLE. Matching the first absolute moment instead, $\mathrm{E}_\sigma|X| = \sigma\sqrt{2/\pi}$, gives $\tilde{\sigma} = \sqrt{\pi/2}\,|X|$: with the odd moments dead, the answer depends on which moment one elects to match.

Problem 7.9. Let $X_1, \ldots, X_n$ be iid with pdf

$$f(x \mid \theta) = \frac{1}{\theta}, \qquad 0 \le x \le \theta, \qquad \theta > 0.$$

Estimate θ using both the method of moments and maximum likelihood. Calculate the means and variances of the two estimators. Which one should be preferred and why?

Solution. $\tilde{\theta} = 2\bar{X}$ and $\hat{\theta} = X_{(n)}$, and the MLE wins: its mean squared error is $O(n^{-2})$ against $O(n^{-1})$. Method of moments: $\mathrm{E}_\theta X = \theta/2$, so matching $\theta/2 = \bar{X}$ gives $\tilde{\theta} = 2\bar{X}$. Maximum likelihood: $L(\theta \mid \mathbf{x}) = \theta^{-n} I(x_{(n)} \le \theta)$ is zero for $\theta < x_{(n)}$ and strictly decreasing thereafter, so $\hat{\theta} = X_{(n)}$.
Moments of $\tilde{\theta}$. Since $\operatorname{Var} X = \theta^2/12$,

$$\mathrm{E}_\theta \tilde{\theta} = \theta, \qquad \operatorname{Var}_\theta \tilde{\theta} = 4 \cdot \frac{\theta^2/12}{n} = \frac{\theta^2}{3n}.$$

Moments of $\hat{\theta}$. The pdf of $X_{(n)}$ is nx^{n-1}/θ^n on $(0, \theta)$ (Theorem 5.4.4), so $\mathrm{E}_\theta X_{(n)}^k = n\theta^k/(n+k)$ and

$$\begin{aligned} \mathrm{E}_\theta \hat{\theta} &= \frac{n}{n+1}\theta, \\ \operatorname{Var}_\theta \hat{\theta} &= \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} = \frac{n\theta^2}{(n+2)(n+1)^2}. \end{aligned}$$

Comparison. The MLE is biased downward by $\theta/(n+1)$, so

$$\begin{aligned} \mathrm{MSE}_\theta(\hat{\theta}) &= \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2} = \frac{2\theta^2}{(n+1)(n+2)}, \\ \mathrm{MSE}_\theta(\tilde{\theta}) &= \frac{\theta^2}{3n}. \end{aligned}$$

Now $\mathrm{MSE}(\hat{\theta}) \le \mathrm{MSE}(\tilde{\theta})$ iff $6n \le (n+1)(n+2)$, i.e. iff $(n-1)(n-2) \ge 0$, which holds for every $n \ge 1$ with equality exactly at $n = 1, 2$. So $\hat{\theta}$ is preferred: at least as good at every sample size, and better by a factor of order $n/6$ for $n \ge 3$.

Problem 7.10. The independent random variables $X_1, \ldots, X_n$ have the common distribution

$$P(X_i \le x \mid \alpha, \beta) = \begin{cases} 0 & \text{if } x < 0, \\ (x/\beta)^\alpha & \text{if } 0 \le x \le \beta, \\ 1 & \text{if } x > \beta, \end{cases}$$

where the parameters α and β are positive.
(a) Find a two-dimensional sufficient statistic for (α, β).
(b) Find the MLEs of α and β.
(c) The length (in millimeters) of cuckoos' eggs found in hedge sparrow nests can be modeled with this distribution. For the data

22.0	23.9	20.9	23.8	25.0	24.0	21.7
23.8	22.8	23.1	23.1	23.5	23.0	23.0

find the MLEs of α and β.

Solution. (a) $T(\mathbf{X}) = \big(\prod_{i=1}^n X_i,\ X_{(n)}\big)$. Differentiating the cdf, $f(x \mid \alpha, \beta) = \alpha x^{\alpha-1}/\beta^\alpha$ on $0 < x < \beta$, so

$$L(\alpha, \beta \mid \mathbf{x}) = \frac{\alpha^n}{\beta^{n\alpha}} \Big(\prod_{i=1}^n x_i\Big)^{\alpha-1} I\big(x_{(1)} > 0\big)\, I\big(x_{(n)} \le \beta\big),$$

which factors as $g(T(\mathbf{x}) \mid \alpha, \beta) \cdot h(\mathbf{x})$ with $h \equiv I(x_{(1)} > 0)$; the Factorization Theorem (Theorem 6.2.6)

gives sufficiency. Equivalently $\big(\sum_i \log X_i,\ X_{(n)}\big)$.
(b)

$$\hat\beta = X_{(n)}, \qquad \hat\alpha = \frac{n}{n\log X_{(n)} - \sum_{i=1}^n \log X_i} = \Big(\log X_{(n)} - \tfrac{1}{n}\textstyle\sum_i \log X_i\Big)^{-1}.$$

For fixed $\alpha > 0$ the likelihood is 0 for $\beta < x_{(n)}$ and proportional to $\beta^{-n\alpha}$, strictly decreasing, for $\beta \ge x_{(n)}$; hence $\hat\beta = x_{(n)}$ whatever α is, and we may substitute. With $\beta = x_{(n)}$,

$$\log L = n\log\alpha + (\alpha-1)\sum_i \log x_i - n\alpha \log x_{(n)}, \qquad \frac{\partial \log L}{\partial \alpha} = \frac{n}{\alpha} - D,$$

where $D = n\log x_{(n)} - \sum_i \log x_i \ge 0$. Since $\partial^2 \log L/\partial\alpha^2 = -n/\alpha^2 < 0$, $\log L$ is strictly concave in α and the root $\hat\alpha = n/D$ is the unique maximum. ($D > 0$ unless all the x_i are equal, an event of probability 0.)
(c) Here $n = 14$ and $x_{(n)} = 25.0$, so $\hat\beta = 25.0$. For $\hat\alpha$,

$$\begin{aligned} \sum_{i=1}^{14} \log x_i &= 43.95270, \\ 14\log 25.0 &= 45.06426, \\ D &= 45.06426 - 43.95270 = 1.11156, \end{aligned}$$

whence

$$\hat\alpha = \frac{14}{1.11156} = 12.5949.$$

So $\hat\alpha \approx 12.59$ and $\hat\beta = 25.0$ mm.

Problem 7.11. Let $X_1, \dots, X_n$ be iid with pdf

$$f(x \mid \theta) = \theta x^{\theta-1}, \qquad 0 \le x \le 1, \qquad 0 < \theta < \infty.$$

(a) Find the MLE of θ, and show that its variance $\to 0$ as $n \to \infty$.
(b) Find the method of moments estimator of θ.

Solution. (a) $\hat\theta = \dfrac{-n}{\sum_{i=1}^n \log X_i}$, with $\operatorname{Var}_\theta \hat\theta = \dfrac{n^2\theta^2}{(n-1)^2(n-2)} \to 0$.
From $\log L = n\log\theta + (\theta-1)\sum_i \log x_i$,

$$\frac{\partial \log L}{\partial \theta} = \frac{n}{\theta} + \sum_i \log x_i, \qquad \frac{\partial^2 \log L}{\partial \theta^2} = -\frac{n}{\theta^2} < 0,$$

so $\log L$ is strictly concave and the unique root $\hat\theta = -n/\sum_i \log x_i$ is the maximum (the x_i lie in $(0,1)$ with probability one, so $\sum_i \log x_i < 0$).
For the variance, put $Y_i = -\log X_i$. Then for $y > 0$,

$$P(Y_i > y) = P\big(X_i < e^{-y}\big) = \big(e^{-y}\big)^\theta = e^{-\theta y},$$

so $Y_i \sim$ exponential$(1/\theta)$ and $T = \sum_i Y_i \sim$ gamma$(n, 1/\theta)$. Negative moments of a gamma are immediate from the density:

$$\mathrm{E}_\theta T^{-k} = \frac{\theta^n}{\Gamma(n)} \int_0^\infty t^{n-k-1} e^{-\theta t}\, dt = \frac{\theta^k\, \Gamma(n-k)}{\Gamma(n)}, \qquad k < n.$$

Hence for $n > 2$, with $\hat\theta = n/T$,

$$\begin{aligned}
\mathrm{E}_\theta\,\hat\theta &= \frac{n\theta}{n-1},\\
\mathrm{E}_\theta\,\hat\theta^{\,2} &= \frac{n^2\theta^2}{(n-1)(n-2)},\\
\mathrm{Var}_\theta\,\hat\theta &= n^2\theta^2\Big[\frac{1}{(n-1)(n-2)} - \frac{1}{(n-1)^2}\Big] = \frac{n^2\theta^2}{(n-1)^2(n-2)}.
\end{aligned}$$

This is $\theta^2\big(1+O(n^{-1})\big)/n \to 0$ as $n\to\infty$.
(b) $\tilde\theta = \dfrac{\bar X}{1-\bar X}$, from

$$\mathrm{E}_\theta X = \int_0^1 x\,\theta x^{\theta-1}\,dx = \frac{\theta}{\theta+1} = \bar X \quad\Longrightarrow\quad \theta(1-\bar X) = \bar X.$$

Problem 7.12. Let $X_1,\dots,X_n$ be a random sample from a population with pmf

$$P_\theta(X=x) = \theta^x(1-\theta)^{1-x}, \qquad x = 0 \text{ or } 1, \qquad 0 \le \theta \le \tfrac12.$$

(a) Find the method of moments estimator and MLE of θ.
(b) Find the mean squared errors of each of the estimators.
(c) Which estimator is preferred? Justify your choice.

Solution. (a) $\tilde\theta = \bar X$ and $\hat\theta = \min\big(\bar X, \tfrac12\big)$.
Method of moments: $\mathrm{E}_\theta X = \theta$, so matching first moments gives $\tilde\theta = \bar X$ (which may fall outside the parameter space $[0,\tfrac12]$ – the method of moments does not know about the restriction). Maximum likelihood: with $y = \sum_i x_i$,

$$L(\theta \mid \mathbf{x}) = \theta^y(1-\theta)^{n-y}, \qquad \frac{\partial \log L}{\partial\theta} = \frac{y}{\theta} - \frac{n-y}{1-\theta} = \frac{y-n\theta}{\theta(1-\theta)},$$

so L increases on $[0,\bar x]$ and decreases on $[\bar x,1]$. Maximising over the restricted range $[0,\tfrac12]$ therefore gives the interior point $\bar x$ when $\bar x \le \tfrac12$, and the right endpoint $\tfrac12$ when $\bar x > \tfrac12$: $\hat\theta = \min(\bar X, \tfrac12)$.
(b) $\bar X$ is unbiased, so

$$\mathrm{MSE}_\theta(\tilde\theta) = \mathrm{Var}_\theta \bar X = \frac{\theta(1-\theta)}{n}.$$

For the MLE, writing $Y = \sum_i X_i \sim \text{binomial}(n,\theta)$,

$$\mathrm{MSE}_\theta(\hat\theta) = \sum_{y=0}^{n}\Big(\min\big(\tfrac{y}{n},\tfrac12\big) - \theta\Big)^2 \binom{n}{y}\theta^y(1-\theta)^{n-y},$$

which has no simpler closed form; but the difference of the two does:

$$\begin{aligned}
\mathrm{MSE}_\theta(\tilde\theta) - \mathrm{MSE}_\theta(\hat\theta) &= \mathrm{E}_\theta\Big[(\bar X-\theta)^2 - \big(\tfrac12-\theta\big)^2;\ \bar X > \tfrac12\Big]\\
&= \mathrm{E}_\theta\Big[\big(\bar X - \tfrac12\big)\big(\bar X + \tfrac12 - 2\theta\big);\ \bar X > \tfrac12\Big],
\end{aligned}$$

since the two estimators agree on $\{\bar X \le \tfrac12\}$.
(c) The MLE. On the event $\{\bar X > \tfrac12\}$ both factors above are positive: $\bar X - \tfrac12 > 0$, and $\bar X + \tfrac12 - 2\theta > 1 - 2\theta \ge 0$ because $\theta \le \tfrac12$. Hence

$$\mathrm{MSE}_\theta(\hat\theta) \le \mathrm{MSE}_\theta(\tilde\theta) \qquad \text{for every } \theta \in [0,\tfrac12],$$

with strict inequality whenever $P_\theta(\bar X > \tfrac12) > 0$, that is for every $\theta \in (0,\tfrac12]$; at $\theta = 0$ both vanish. So $\hat\theta$ dominates $\tilde\theta$, and $\tilde\theta$ is inadmissible.

Problem 7.13. Let $X_1, \dots, X_n$ be a sample from a population with double exponential pdf

$$f(x \mid \theta) = \tfrac{1}{2} e^{-|x-\theta|}, \qquad -\infty < x < \infty, \qquad -\infty < \theta < \infty.$$

Find the MLE of θ. (Hint: Consider the case of even n separate from that of odd n, and find the MLE in terms of the order statistics. A complete treatment of this problem is given in Norton 1984.)

Solution. The MLE is the sample median: for $n = 2k+1$ odd, $\hat\theta = X_{(k+1)}$ uniquely; for $n = 2m$ even, every $\theta \in [X_{(m)}, X_{(m+1)}]$ is an MLE, and the usual convention takes $\hat\theta = \frac{1}{2}\big(X_{(m)} + X_{(m+1)}\big)$.
Since $L(\theta \mid \mathbf{x}) = 2^{-n} \exp\{-g(\theta)\}$ with

$$g(\theta) = \sum_{i=1}^{n} |x_i - \theta|,$$

maximising L is minimising g. Each summand is convex in θ, so g is convex and piecewise linear with knots at the order statistics; a minimiser exists because $g(\theta) \to \infty$ as $|\theta| \to \infty$. On the open interval $x_{(j)} < \theta < x_{(j+1)}$ exactly j of the x_i lie below θ and $n - j$ above, so

$$g'(\theta) = j - (n - j) = 2j - n, \qquad j = 0, 1, \dots, n$$

(with $x_{(0)} = -\infty$, $x_{(n+1)} = +\infty$). By convexity the minimum is where this slope changes sign.
(i) $n = 2k+1$ odd. Then $2j - n < 0$ exactly for $j \le k$ and $2j - n > 0$ exactly for $j \ge k+1$; the slope is never 0. So g strictly decreases up to $x_{(k+1)}$ and strictly increases after it:

$$\hat\theta = x_{(k+1)} = x_{((n+1)/2)},$$

the unique MLE.
(ii) $n = 2m$ even. Now $2j - n = 0$ at $j = m$, so g is constant on $[x_{(m)}, x_{(m+1)}]$, strictly decreasing before and strictly increasing after. The set of MLEs is the whole interval

$$\big[x_{(m)},\, x_{(m+1)}\big] = \big[x_{(n/2)},\, x_{(n/2+1)}\big],$$

and the MLE is not unique; taking the midpoint gives the sample median.

Problem 7.14. Let X and Y be independent exponential random variables, with

$$f(x \mid \lambda) = \frac{1}{\lambda} e^{-x/\lambda},\ x > 0, \qquad f(y \mid \mu) = \frac{1}{\mu} e^{-y/\mu},\ y > 0.$$

We observe Z and W with

$$Z = \min(X, Y) \qquad \text{and} \qquad W = \begin{cases} 1 & \text{if } Z = X, \\ 0 & \text{if } Z = Y. \end{cases}$$

In Exercise 4.26 the joint distribution of Z and W was obtained. Now assume that (Z_i, W_i), $i = 1, \dots, n$, are n iid observations. Find the MLEs of λ and μ.

Solution.

$$\hat\lambda = \frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n W_i}, \qquad \hat\mu = \frac{\sum_{i=1}^n Z_i}{n - \sum_{i=1}^n W_i}.$$

By Exercise 4.26, Z and W are independent with $Z \sim \text{exponential}\big(\frac{\lambda\mu}{\lambda+\mu}\big)$ and $P(W = 1) = \frac{\mu}{\lambda+\mu}$, so their joint density with respect to (Lebesgue) $\times$ (counting) measure is

$$f(z, w \mid \lambda, \mu) = \Big(\frac{1}{\lambda}\Big)^{w} \Big(\frac{1}{\mu}\Big)^{1-w} \exp\Big\{-z\Big(\frac{1}{\lambda} + \frac{1}{\mu}\Big)\Big\}, \qquad z > 0,\ w \in \{0, 1\}$$

(summing out w returns the density of Z, integrating out z at $w=1$ returns $P(W=1)$. Check!) Reparametrise by the rates $\eta = 1/\lambda$, $\nu = 1/\mu$ and set $S = \sum_i z_i$, $T = \sum_i w_i$. Then

$$\log L(\eta, \nu \mid \mathbf{z}, \mathbf{w}) = T \log \eta + (n - T) \log \nu - S(\eta + \nu),$$

which separates. Each piece is strictly concave, $\partial^2/\partial\eta^2 = -T/\eta^2 < 0$ and likewise in ν, so the stationary point

$$\frac{\partial \log L}{\partial \eta} = \frac{T}{\eta} - S = 0, \qquad \frac{\partial \log L}{\partial \nu} = \frac{n-T}{\nu} - S = 0$$

gives the unique maximum $\hat\eta = T/S$, $\hat\nu = (n-T)/S$. The invariance property of MLEs (Theorem 7.2.10) applied to $\lambda = 1/\eta$, $\mu = 1/\nu$ yields

$$\hat\lambda = \frac{S}{T} = \frac{\sum_i Z_i}{\sum_i W_i}, \qquad \hat\mu = \frac{S}{n-T} = \frac{\sum_i Z_i}{n - \sum_i W_i}.$$

(If $T = 0$ or $T = n$ the profile likelihood is monotone in λ or in μ and no MLE exists in $(0,\infty)^2$.)

7.3 Exercises 7.15–7.21

Problem 7.15. Let $X_1, X_2, \ldots, X_n$ be a sample from the *inverse Gaussian* pdf,

$$f(x \mid \mu, \lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left\{-\lambda(x-\mu)^2/(2\mu^2 x)\right\}, \qquad x > 0.$$

(a) Show that the MLEs of μ and λ are

$$\hat\mu_n = \bar X \qquad \text{and} \qquad \hat\lambda_n = \frac{n}{\sum_{i=1}^n \left(\frac{1}{X_i} - \frac{1}{\bar X}\right)}.$$

(b) Tweedie (1957) showed that $\hat\mu_n$ and $\hat\lambda_n$ are independent, $\hat\mu_n$ having an inverse Gaussian distribution with parameters μ and $n\lambda$, and $n\lambda/\hat\lambda_n$ having a χ^2_{n-1} distribution. Schwarz and Samanta (1991) give a proof of these facts using an induction argument.
(i) Show that $\hat\mu_2$ has an inverse Gaussian distribution with parameters μ and 2λ, $2\lambda/\hat\lambda_2$ has a χ^2_1 distribution, and they are independent.
(ii) Assume the result is true for $n = k$ and that we get a new, independent observation x. Establish the induction step used by Schwarz and Samanta (1991), and transform the pdf $f(x, \hat\mu_k, \hat\lambda_k)$ to $f(x, \hat\mu_{k+1}, \hat\lambda_{k+1})$. Show that this density factors in the appropriate way and that the result of Tweedie follows.

Solution. (a) Since $(x_i - \mu)^2/x_i = x_i - 2\mu + \mu^2/x_i$, the log likelihood is

$$\ell(\mu, \lambda \mid \mathbf{x}) = \frac{n}{2}\log\lambda - \frac{3}{2}\sum_i \log x_i - \frac{n}{2}\log 2\pi \\ - \frac{\lambda}{2\mu^2}\sum_i x_i + \frac{n\lambda}{\mu} - \frac{\lambda}{2}\sum_i \frac{1}{x_i}.$$

Then

$$\frac{\partial \ell}{\partial \mu} = \frac{\lambda}{\mu^2}\left(\frac{\sum_i x_i}{\mu} - n\right),$$

which is > 0 for $\mu < \bar x$ and < 0 for $\mu > \bar x$; hence $\hat\mu_n = \bar X$ for every $\lambda > 0$. Substituting $\mu = \bar x$ into $\partial\ell/\partial\lambda = 0$,

$$\frac{n}{2\lambda} = \frac{\sum_i x_i}{2\bar x^2} - \frac{n}{\bar x} + \frac{1}{2}\sum_i \frac{1}{x_i} = \frac{1}{2}\sum_i\left(\frac{1}{x_i} - \frac{1}{\bar x}\right),$$

giving $\hat\lambda_n = n/\sum_i(1/X_i - 1/\bar X)$, a maximum because ℓ is strictly concave in λ.

(b) One lemma — the Schwarz–Samanta induction step — does everything.
Lemma. Let $a, b > 0$ and let $U \sim \mathrm{IG}(\mu, a\lambda)$ and $V \sim \mathrm{IG}(\mu, b\lambda)$ be independent. Put

$$T = \frac{aU + bV}{a+b}, \qquad Q = \lambda\left(\frac{a}{U} + \frac{b}{V} - \frac{a+b}{T}\right).$$

Then $T \sim \mathrm{IG}(\mu, (a+b)\lambda)$, $Q \sim \chi^2_1$, and $T \perp Q$.
Proof. The exponent separates: $au + bv = (a+b)t$ and $a/u + b/v = (a+b)/t + q/\lambda$ give

$$\begin{aligned}\frac{a\lambda(u-\mu)^2}{2\mu^2 u} + \frac{b\lambda(v-\mu)^2}{2\mu^2 v} &= \frac{\lambda}{2\mu^2}\Big[au + bv - 2\mu(a+b) + \mu^2\Big(\frac{a}{u} + \frac{b}{v}\Big)\Big] \\ &= \frac{(a+b)\lambda(t-\mu)^2}{2\mu^2 t} + \frac{q}{2}.\end{aligned}$$

So the joint density of (T, Q) is $\psi(t,q)\, e^{-(a+b)\lambda(t-\mu)^2/(2\mu^2 t)} e^{-q/2}$, where ψ carries the powers and the Jacobian and does not involve μ. From the displayed definitions,

$$\left|\frac{\partial(t,q)}{\partial(u,v)}\right| = \frac{ab\lambda}{a+b}\left|\frac{1}{u^2} - \frac{1}{v^2}\right|,$$

and the map is two-to-one, so (writing $\sum$ for the sum over the two branches)

$$\psi(t,q) = \sum \frac{\lambda\sqrt{ab}}{2\pi(uv)^{3/2}} \cdot \frac{(a+b)u^2v^2}{ab\lambda\,|u^2 - v^2|} = \frac{a+b}{2\pi\sqrt{ab}} \sum \frac{\sqrt{uv}}{|u^2 - v^2|}.$$

Parametrize the branches by $\alpha = t/u - 1$, $\beta = t/v - 1$. The two constraints become

$$a\alpha + b\beta = (a+b)m, \qquad \alpha\beta = -m, \qquad m := \frac{qt}{(a+b)\lambda},$$

whose solutions are, with $R := \sqrt{(a+b)^2 m^2 + 4abm}$ and $\epsilon = \pm 1$,

$$\alpha_\epsilon = \frac{(a+b)m + \epsilon R}{2a}, \qquad \beta_\epsilon = \frac{(a+b)m - \epsilon R}{2b}.$$

Since $u = t/(1+\alpha)$ and $v = t/(1+\beta)$,

$$\frac{\sqrt{uv}}{|u^2 - v^2|} = \frac{[(1+\alpha)(1+\beta)]^{3/2}}{t\,|\beta - \alpha|\,(2 + \alpha + \beta)}.$$

Set $Y_\epsilon = R + \epsilon(b-a)m > 0$. Then (Check!)

$$\begin{aligned}(1+\alpha_\epsilon)(1+\beta_\epsilon) &= \frac{Y_\epsilon^2}{4abm}, \\ |\beta_\epsilon - \alpha_\epsilon| &= \frac{(a+b)Y_\epsilon}{2ab}, \\ 2 + \alpha_\epsilon + \beta_\epsilon &= \frac{RY_\epsilon}{2abm},\end{aligned}$$

the first because $Y_\epsilon^2 = 2m\big[2ab + (a^2+b^2)m + \epsilon(b-a)R\big]$. Hence each branch contributes $\sqrt{ab}\,Y_\epsilon/\big(2t(a+b)R\sqrt{m}\big)$, and $Y_+ + Y_- = 2R$ gives

$$\psi(t,q) = \frac{a+b}{2\pi\sqrt{ab}} \cdot \frac{\sqrt{ab}}{t(a+b)\sqrt{m}} = \frac{1}{2\pi}\sqrt{\frac{(a+b)\lambda}{t^3 q}}.$$

Therefore

$$f_{T,Q}(t,q) = \left(\frac{(a+b)\lambda}{2\pi t^3}\right)^{1/2} e^{-\frac{(a+b)\lambda(t-\mu)^2}{2\mu^2 t}} \cdot \frac{1}{\sqrt{2\pi q}} e^{-q/2},$$

which is the asserted product of an $\text{IG}(\mu,(a+b)\lambda)$ density and a χ^2_1 density. $\square$
(i) Apply the Lemma with $a=b=1$, $U=X_1$, $V=X_2$: then $T=\hat\mu_2$ and

$$Q=\lambda\left(\frac{1}{X_1}+\frac{1}{X_2}-\frac{2}{\hat\mu_2}\right)=\frac{2\lambda}{\hat\lambda_2},$$

so $\hat\mu_2\sim\text{IG}(\mu,2\lambda)$, $2\lambda/\hat\lambda_2\sim\chi^2_1$, and they are independent.
(ii) Suppose the claim holds at $n=k$, and let $X_{k+1}=x\sim\text{IG}(\mu,\lambda)$ be independent of $X_1,\dots,X_k$. Apply the Lemma with $a=k$, $b=1$, $U=\hat\mu_k\sim\text{IG}(\mu,k\lambda)$, $V=x$: since $T=(k\hat\mu_k+x)/(k+1)=\hat\mu_{k+1}$,

$$\hat\mu_{k+1}\sim\text{IG}(\mu,(k+1)\lambda),\qquad Q=\lambda\left(\frac{k}{\hat\mu_k}+\frac{1}{x}-\frac{k+1}{\hat\mu_{k+1}}\right)\sim\chi^2_1,$$

independent of each other. This is precisely the transformation of $f(x,\hat\mu_k,\hat\lambda_k)$ into $f(x,\hat\mu_{k+1},\hat\lambda_{k+1})$: from part (a), $k/\hat\lambda_k=\sum_{i\le k}X_i^{-1}-k/\hat\mu_k$, so

$$\frac{(k+1)\lambda}{\hat\lambda_{k+1}}=\lambda\sum_{i=1}^{k+1}\frac{1}{X_i}-\frac{(k+1)\lambda}{\hat\mu_{k+1}}=\frac{k\lambda}{\hat\lambda_k}+Q.$$

By the induction hypothesis $k\lambda/\hat\lambda_k\sim\chi^2_{k-1}$ and is independent of $\hat\mu_k$, hence independent of the pair $(\hat\mu_k,x)$ and so of $(\hat\mu_{k+1},Q)$. Thus the joint density factors as

$$f\big(\hat\mu_{k+1}\big)\cdot f_{\chi^2_{k-1}}\left(\frac{k\lambda}{\hat\lambda_k}\right)f_{\chi^2_1}(Q),$$

so $(k+1)\lambda/\hat\lambda_{k+1}$ is the sum of independent χ^2_{k-1} and χ^2_1 variates, i.e. χ^2_k, and it is independent of $\hat\mu_{k+1}$. With the base case (i) at $k=2$, Tweedie's result follows for all $n\ge2$.

Problem 7.16. Berger and Casella (1992) also investigate *power means*, which we have seen in Exercise 4.57. Recall that a power mean is defined as $\left[\frac1n\sum_{i=1}^n x_i^r\right]^{1/r}$. This definition can be further generalized by noting that the power function x^r can be replaced by any continuous, monotone function h, yielding the *generalized mean* $h^{-1}\left(\frac1n\sum_i h(x_i)\right)$.
(a) The least squares problem $\min_a\sum_i(x_i-a)^2$ is sometimes solved using transformed variables, that is, solving $\min_a\sum_i[h(x_i)-h(a)]^2$. Show that the solution to this latter problem is $a=h^{-1}\left((1/n)\sum_i h(x_i)\right)$.
(b) Show that the arithmetic mean is the solution to the untransformed least squares problem, the geometric mean is the solution to the problem transformed by $h(x)=\log x$, and the harmonic mean is the solution to the problem transformed by $h(x)=1/x$.
(c) Show that if the least squares problem is transformed with the Box-Cox Transformation (see Exercise 11.3), $h(x)=(x^\lambda-1)/\lambda$, then the solution is a generalized mean with $h(x)=x^\lambda$.
(d) Let $X_1,\dots,X_n$ be a sample from a lognormal(μ,σ^2) population. Show that the MLE of μ is the geometric mean.
(e) Suppose that $X_1,\dots,X_n$ are a sample from a one-parameter exponential family $f(x\mid\theta)=\exp\{\theta h(x)-H(\theta)\}g(x)$, where $h=H'$ and h is an increasing function.
(i) Show that the MLE of θ is $\hat\theta=h^{-1}\left((1/n)\sum_i h(x_i)\right)$.
(ii) Show that two densities that satisfy $h=H'$ are the normal and the inverted gamma with pdf $f(x\mid\theta)=\theta x^{-2}\exp\{-\theta/x\}$ for $x>0$, and for the normal the MLE is the arithmetic mean and for the inverted gamma it is the harmonic mean.

Solution. (a) Substitute $b=h(a)$: the problem becomes $\min_b\sum_i[h(x_i)-b]^2$, minimized at $b=\frac1n\sum_i h(x_i)$. This value is attainable because h continuous and monotone has an interval range $\mathcal{B}$,

which therefore contains the average of the $h(x_i) \in \mathcal{B}$; and h is one-to-one, so

$$a = h^{-1}\left(\frac{1}{n}\sum_i h(x_i)\right).$$

(b) Take $h(x) = x$, $\log x$, $1/x$ in turn in part (a):

$$\begin{aligned} h(x) = x: &\quad a = \frac{1}{n}\sum_i x_i, \\ h(x) = \log x: &\quad a = \exp\Big(\frac{1}{n}\sum_i \log x_i\Big) = \Big(\prod_i x_i\Big)^{1/n}, \\ h(x) = 1/x: &\quad a = \Big(\frac{1}{n}\sum_i \frac{1}{x_i}\Big)^{-1} = \frac{n}{\sum_i 1/x_i}, \end{aligned}$$

the arithmetic, geometric and harmonic means.
(c) For $h(x) = (x^\lambda - 1)/\lambda$ we have $h^{-1}(y) = (\lambda y + 1)^{1/\lambda}$ and $\frac{1}{n}\sum_i h(x_i) = \frac{1}{\lambda}\big(\frac{1}{n}\sum_i x_i^\lambda - 1\big)$, so by (a)

$$a = \left(\frac{1}{n}\sum_i x_i^\lambda - 1 + 1\right)^{1/\lambda} = \left(\frac{1}{n}\sum_i x_i^\lambda\right)^{1/\lambda},$$

the generalized mean with $h(x) = x^\lambda$ (the power mean of order λ).
(d) The lognormal pdf is $f(x \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2}x^{-1}\exp\{-(\log x - \mu)^2/(2\sigma^2)\}$, so for fixed σ^2, maximizing the likelihood in μ is exactly the least squares problem $\min_\mu \sum_i (\log x_i - \mu)^2$ — i.e. part (a) with $h = \log$ and $\mu = h(a)$. Hence $\hat\mu = \frac{1}{n}\sum_i \log x_i$, and by the invariance property (Theorem 7.2.10) the MLE of the parameter on the original scale, e^μ, is

$$e^{\hat\mu} = \Big(\prod_i x_i\Big)^{1/n},$$

the geometric mean.
(e)(i) The log likelihood is $\ell(\theta) = \theta\sum_i h(x_i) - nH(\theta) + \sum_i \log g(x_i)$, so

$$\ell'(\theta) = \sum_i h(x_i) - nH'(\theta) = \sum_i h(x_i) - nh(\theta),$$

and $\ell''(\theta) = -nh'(\theta) < 0$ since h is increasing. Thus ℓ is strictly concave and the unique root

$$h(\hat\theta) = \frac{1}{n}\sum_i h(x_i), \qquad \text{i.e.} \qquad \hat\theta = h^{-1}\Big(\frac{1}{n}\sum_i h(x_i)\Big),$$

is the MLE.
(ii) *Normal.* For $n(\theta, 1)$,

$$f(x \mid \theta) = \exp\{\theta x - \theta^2/2\}\cdot\frac{1}{\sqrt{2\pi}}e^{-x^2/2},$$

so $h(x) = x$, $H(\theta) = \theta^2/2$, and $H'(\theta) = \theta = h(\theta)$. By (i), $\hat\theta = \frac{1}{n}\sum_i x_i$, the arithmetic mean.
Inverted gamma. For $f(x \mid \theta) = \theta x^{-2}e^{-\theta/x}$, $x > 0$,

$$f(x \mid \theta) = \exp\{\theta(-1/x) - (-\log\theta)\}\cdot x^{-2},$$

so $h(x) = -1/x$, $H(\theta) = -\log\theta$, and $H'(\theta) = -1/\theta = h(\theta)$; also $h(x) = -1/x$ is increasing on $(0, \infty)$. By (i), $-1/\hat\theta = -\frac{1}{n}\sum_i 1/x_i$, so

$$\hat\theta = \frac{n}{\sum_i 1/x_i},$$

the harmonic mean.

Problem 7.17. The Borel Paradox (Miscellanea 4.9.3) can also arise in inference problems. Suppose that X_1 and X_2 are iid exponential(θ) random variables.
(a) If we observe only X_2, show that the MLE of θ is $\hat\theta = X_2$.
(b) Suppose that we instead observe only $Z = (X_2 - 1)/X_1$. Find the joint distribution of (X_1, Z), and integrate out X_1 to get the likelihood function.
(c) Suppose that $X_2 = 1$. Compare the MLEs for θ from parts (a) and (b).
(d) Bayesian analysis is not immune to the Borel Paradox. If $\pi(\theta)$ is a prior density for θ, show that the posterior distributions, at $X_2 = 1$, are different in parts (a) and (b).
(Communicated by L. Mark Berliner, Ohio State University.)

Solution. (a) $\hat\theta = X_2$. With $f(x \mid \theta) = \theta^{-1} e^{-x/\theta}$,

$$\frac{\partial}{\partial\theta}\Big[-\log\theta - \frac{x_2}{\theta}\Big] = -\frac{1}{\theta} + \frac{x_2}{\theta^2} = \frac{x_2 - \theta}{\theta^2},$$

positive for $\theta < x_2$ and negative for $\theta > x_2$.
(b) Fix $X_1 = x_1$ and put $X_2 = 1 + z x_1$; the Jacobian is $\partial x_2/\partial z = x_1$, so

$$f(x_1, z \mid \theta) = \frac{x_1}{\theta^2}\, e^{-(1 + x_1 + z x_1)/\theta}, \qquad x_1 > 0,\ 1 + z x_1 > 0.$$

The constraint $1 + z x_1 > 0$ is vacuous when $z \ge 0$ and reads $x_1 < -1/z$ when $z < 0$. For $z \ge 0$, using $\int_0^\infty x e^{-bx}\, dx = b^{-2}$ with $b = (1+z)/\theta$,

$$L(\theta \mid z) = \frac{e^{-1/\theta}}{\theta^2} \cdot \frac{\theta^2}{(1+z)^2} = \frac{e^{-1/\theta}}{(1+z)^2}, \qquad z \ge 0.$$

For $z < 0$ write $u = -1/z > 0$, so that $(1+z)u = u - 1$, and $\int_0^u x e^{-bx} dx = b^{-2}[1 - e^{-bu}(1 + bu)]$ gives

$$L(\theta \mid z) = \frac{e^{-1/\theta}}{(1+z)^2}\left[1 - e^{-(u-1)/\theta}\Big(1 + \frac{u-1}{\theta}\Big)\right], \qquad z < 0.$$

(c) They disagree, although $\{X_2 = 1\}$ and $\{Z = 0\}$ are the same event. From (a), $\hat\theta = x_2 = 1$; from (b) at $z = 0$,

$$L(\theta \mid 0) = e^{-1/\theta},$$

strictly increasing in θ with $\sup_\theta L = 1$ attained only as $\theta \to \infty$, so no MLE exists. The two likelihoods differ by the factor θ^{-1}: the two conditioning variables approach the common event through different families of neighborhoods.
(d) For a prior $\pi(\theta)$ the two posteriors at $X_2 = 1$ are

$$\pi(\theta \mid x_2 = 1) = \frac{\theta^{-1} e^{-1/\theta}\pi(\theta)}{\int_0^\infty t^{-1} e^{-1/t}\pi(t)\, dt}, \qquad \pi(\theta \mid z = 0) = \frac{e^{-1/\theta}\pi(\theta)}{\int_0^\infty e^{-1/t}\pi(t)\, dt},$$

whenever the normalizing integrals are finite. Their ratio is proportional to θ^{-1}, which is not constant in θ, so the two posteriors differ for every prior π.

Problem 7.18. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid bivariate normal random variables (pairs) where all five parameters are unknown.
(a) Show that the method of moments estimators for $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$ are

$$\begin{aligned} &\hat\mu_X = \bar x, \qquad \hat\mu_Y = \bar y, \\ &\hat\sigma_X^2 = \tfrac{1}{n}\textstyle\sum_i (x_i - \bar x)^2, \qquad \hat\sigma_Y^2 = \tfrac{1}{n}\sum_i (y_i - \bar y)^2, \\ &\hat\rho = \tfrac{1}{n}\textstyle\sum_i (x_i - \bar x)(y_i - \bar y)\big/(\hat\sigma_X \hat\sigma_Y). \end{aligned}$$

(b) Derive the MLEs of the unknown parameters and show that they are the same as the method of moments estimators. (One attack is to write the joint pdf as the product of a conditional and a marginal, that is, write

$$f(x, y \mid \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) = f(y \mid x, \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)\, f(x \mid \mu_X, \sigma_X^2),$$

and argue that the MLEs for μ_X and σ_X^2 are given by $\bar{x}$ and $\frac{1}{n}\sum_i (x_i - \bar{x})^2$. Then, turn things around to get the MLEs for μ_Y and σ_Y^2. Finally, work with the *partially maximized* likelihood function $L(\bar{x}, \bar{y}, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, \rho \mid \mathbf{x}, \mathbf{y})$ to get the MLE for ρ. As might be guessed, this is a difficult problem.)

Solution. Write $s_{xx} = \frac{1}{n}\sum_i (x_i - \bar{x})^2$, $s_{yy} = \frac{1}{n}\sum_i (y_i - \bar{y})^2$, $s_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$, and $r = s_{xy}/\sqrt{s_{xx} s_{yy}}$; assume the data are not collinear, so $|r| < 1$.
(a) Equate the first two population moments and the cross moment to their sample versions. From $EX = \mu_X$ and $EY = \mu_Y$, $\hat{\mu}_X = \bar{x}$ and $\hat{\mu}_Y = \bar{y}$. From $EX^2 = \sigma_X^2 + \mu_X^2$,

$$\hat{\sigma}_X^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2 = s_{xx},$$

and likewise $\hat{\sigma}_Y^2 = s_{yy}$. From $EXY = \rho\sigma_X\sigma_Y + \mu_X\mu_Y$,

$$\hat{\rho}\,\hat{\sigma}_X\hat{\sigma}_Y = \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y} = s_{xy}, \qquad \text{so} \qquad \hat{\rho} = \frac{s_{xy}}{\hat{\sigma}_X\hat{\sigma}_Y} = r.$$

(b) Put $u = \sigma_X^2$, $v = \sigma_Y^2$. The log likelihood is

$$\ell = -n\log(2\pi) - \frac{n}{2}\log u - \frac{n}{2}\log v - \frac{n}{2}\log(1-\rho^2)$$
$$-\frac{1}{2(1-\rho^2)}\sum_i \left[\frac{(x_i-\mu_X)^2}{u} - \frac{2\rho(x_i-\mu_X)(y_i-\mu_Y)}{\sqrt{uv}} + \frac{(y_i-\mu_Y)^2}{v}\right].$$

Means. With (u, v, ρ) fixed, $\partial\ell/\partial\mu_X = 0$ and $\partial\ell/\partial\mu_Y = 0$ read, with $A = \sum_i (x_i - \mu_X)$ and $B = \sum_i (y_i - \mu_Y)$,

$$\frac{A}{u} - \frac{\rho B}{\sqrt{uv}} = 0, \qquad \frac{B}{v} - \frac{\rho A}{\sqrt{uv}} = 0.$$

This is $\Sigma^{-1}(A, B)^\top = 0$, and Σ^{-1} is nonsingular for $|\rho| < 1$, so $A = B = 0$: $\hat{\mu}_X = \bar{x}$ and $\hat{\mu}_Y = \bar{y}$, whatever (u, v, ρ).
Partially maximized likelihood. Substituting $\mu_X = \bar{x}$, $\mu_Y = \bar{y}$ and writing

$$Q = \frac{s_{xx}}{u} - \frac{2\rho s_{xy}}{\sqrt{uv}} + \frac{s_{yy}}{v},$$

we get $\ell = -n\log(2\pi) - \frac{n}{2}\log\big[uv(1-\rho^2)\big] - \dfrac{nQ}{2(1-\rho^2)}$. Then

$$\frac{2u}{n}\frac{\partial\ell}{\partial u} = -1 + \frac{1}{1-\rho^2}\left[\frac{s_{xx}}{u} - \frac{\rho s_{xy}}{\sqrt{uv}}\right],$$
$$\frac{2v}{n}\frac{\partial\ell}{\partial v} = -1 + \frac{1}{1-\rho^2}\left[\frac{s_{yy}}{v} - \frac{\rho s_{xy}}{\sqrt{uv}}\right],$$
$$\frac{(1-\rho^2)^2}{n}\frac{\partial\ell}{\partial\rho} = \rho(1-\rho^2) + \frac{s_{xy}}{\sqrt{uv}}(1-\rho^2) - \rho Q.$$

Adding the first two equations set to zero gives $Q = 2(1-\rho^2)$; subtracting them gives $s_{xx}/u = s_{yy}/v$. Inserting $Q = 2(1-\rho^2)$ into the third equation and cancelling $1 - \rho^2 > 0$,

$$\rho + \frac{s_{xy}}{\sqrt{uv}} - 2\rho = 0 \qquad \Longrightarrow \qquad \rho = \frac{s_{xy}}{\sqrt{uv}}.$$

Feeding this back into the first equation, $1-\rho^2 = s_{xx}/u - \rho^2$, so $u = s_{xx}$, and then $v = s_{yy}$ and $\rho = s_{xy}/\sqrt{s_{xx}s_{yy}} = r$. Thus

$$\hat{\mu}_X = \bar{x}, \quad \hat{\mu}_Y = \bar{y}, \quad \hat{\sigma}_X^2 = s_{xx}, \quad \hat{\sigma}_Y^2 = s_{yy}, \quad \hat{\rho} = r,$$

exactly the method of moments estimators of part (a). This is the only stationary point, and it is the global maximum: with $p = s_{xx}/u$, $q = s_{yy}/v$ we have $Q \geq (1-|r|)(p+q)$ (since $2\sqrt{pq} \leq p+q$), so

$$\ell \leq C + \frac{n}{2}\log(pq) - \frac{n}{2}\log(1-\rho^2) - \frac{n(1-|r|)(p+q)}{2(1-\rho^2)}.$$

For fixed ρ the bound tends to $-\infty$ as p or q tends to 0 or ∞; and maximized over $p, q > 0$ it equals $C - n + \frac{n}{2}\log(1-\rho^2) - n\log(1-|r|)$, which tends to $-\infty$ as $|\rho| \to 1$.

Problem 7.19. Suppose that the random variables $Y_1, \ldots, Y_n$ satisfy

$$Y_i = \beta x_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $x_1, \ldots, x_n$ are fixed constants, and $\epsilon_1, \ldots, \epsilon_n$ are iid $n(0, \sigma^2)$, σ^2 unknown.
(a) Find a two-dimensional sufficient statistic for (β, σ^2).
(b) Find the MLE of β, and show that it is an unbiased estimator of β.
(c) Find the distribution of the MLE of β.

Solution. (a) $T(\mathbf{Y}) = \left(\sum_i x_i Y_i,\ \sum_i Y_i^2\right)$. Indeed

$$f(\mathbf{y} \mid \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\beta\sum_i x_i y_i}{\sigma^2} - \frac{\beta^2\sum_i x_i^2}{2\sigma^2}\right\},$$

which depends on $\mathbf{y}$ only through T; take $h(\mathbf{y}) \equiv 1$ in the Factorization Theorem (Theorem 6.2.6). (This is the two-parameter exponential family of Theorem 6.2.10 with natural statistics $\sum x_i y_i$ and $\sum y_i^2$.)
(b) $\hat{\beta} = \sum_i x_i Y_i / \sum_i x_i^2$. For fixed σ^2, maximizing the likelihood is minimizing $\sum_i (y_i - \beta x_i)^2$, a strictly convex quadratic in β with derivative $-2\sum_i x_i(y_i - \beta x_i)$. Since $EY_i = \beta x_i$,

$$E\hat{\beta} = \frac{\sum_i x_i(\beta x_i)}{\sum_i x_i^2} = \beta.$$

(c) $\hat{\beta}$ is a fixed linear combination of independent normals, hence normal (Corollary 4.6.10), with mean β and

$$\operatorname{Var}\hat{\beta} = \frac{\sum_i x_i^2\sigma^2}{\left(\sum_i x_i^2\right)^2} = \frac{\sigma^2}{\sum_i x_i^2}, \qquad \text{so} \qquad \hat{\beta} \sim n\left(\beta,\ \frac{\sigma^2}{\sum_i x_i^2}\right).$$

Problem 7.20. Consider $Y_1, \ldots, Y_n$ as defined in Exercise 7.19.
(a) Show that $\sum Y_i / \sum x_i$ is an unbiased estimator of β.
(b) Calculate the exact variance of $\sum Y_i / \sum x_i$ and compare it to the variance of the MLE.

Solution. (a) Assume $\sum_i x_i \neq 0$. Then

$$E\left[\frac{\sum_i Y_i}{\sum_i x_i}\right] = \frac{\sum_i \beta x_i}{\sum_i x_i} = \beta.$$

(b) The Y_i are independent with variance σ^2, so

$$\operatorname{Var}\left[\frac{\sum_i Y_i}{\sum_i x_i}\right] = \frac{n\sigma^2}{\left(\sum_i x_i\right)^2}.$$

By the Cauchy-Schwarz Inequality (Theorem 4.7.3) applied to the vectors $(1,\dots,1)$ and $(x_1,\dots,x_n)$,

$$\left(\sum_i x_i\right)^2 \le n\sum_i x_i^2, \qquad \text{hence} \qquad \frac{n\sigma^2}{(\sum_i x_i)^2} \;\ge\; \frac{\sigma^2}{\sum_i x_i^2} = \operatorname{Var}\hat{\beta}.$$

The MLE therefore has the smaller variance, with equality if and only if all the x_i are equal.

Problem 7.21. Again, let $Y_1,\dots,Y_n$ be as defined in Exercise 7.19.
(a) Show that $\left[\sum(Y_i/x_i)\right]/n$ is also an unbiased estimator of β.
(b) Calculate the exact variance of $\left[\sum(Y_i/x_i)\right]/n$ and compare it to the variances of the estimators in the previous two exercises.

Solution. (a) Assume every $x_i \neq 0$. Since $EY_i = \beta x_i$,

$$E\left[\frac{1}{n}\sum_i \frac{Y_i}{x_i}\right] = \frac{1}{n}\sum_i \frac{\beta x_i}{x_i} = \beta.$$

(b) By independence,

$$\operatorname{Var}\left[\frac{1}{n}\sum_i \frac{Y_i}{x_i}\right] = \frac{1}{n^2}\sum_i \frac{\sigma^2}{x_i^2} = \frac{\sigma^2}{n^2}\sum_i \frac{1}{x_i^2}.$$

Cauchy-Schwarz applied to $(x_1,\dots,x_n)$ and $(1/x_1,\dots,1/x_n)$ gives

$$n^2 = \left(\sum_i x_i\cdot\frac{1}{x_i}\right)^2 \le \left(\sum_i x_i^2\right)\left(\sum_i \frac{1}{x_i^2}\right), \qquad \text{so} \qquad \frac{\sigma^2}{n^2}\sum_i \frac{1}{x_i^2} \;\ge\; \frac{\sigma^2}{\sum_i x_i^2},$$

so this estimator is again beaten by the MLE. If in addition the x_i all have the same sign, then $(\sum_i x_i)\left(\sum_i 1/x_i\right) \ge n^2$ and $\left(\sum_i 1/x_i\right)^2 \le n\sum_i 1/x_i^2$ (Cauchy-Schwarz twice), whence

$$\sum_i \frac{1}{x_i^2} \;\ge\; \frac{1}{n}\left(\sum_i \frac{1}{x_i}\right)^2 \;\ge\; \frac{n^3}{(\sum_i x_i)^2},$$

and the three variances are ordered

$$\frac{\sigma^2}{\sum_i x_i^2} \;\le\; \frac{n\sigma^2}{(\sum_i x_i)^2} \;\le\; \frac{\sigma^2}{n^2}\sum_i \frac{1}{x_i^2},$$

with equality throughout iff all the x_i are equal. (Without the common-sign assumption the last two can reverse: for $(x_1,x_2,x_3) = (1,-1,2)$ they are $0.75\sigma^2$ and $0.25\sigma^2$.)

7.4 Exercises 7.22–7.28

Problem 7.22. This exercise will prove the assertions in Example 7.2.16, and more. Let $X_1,\dots,X_n$ be a random sample from a $\mathrm{n}(\theta,\sigma^2)$ population, and suppose that the prior distribution on θ is $\mathrm{n}(\mu,\tau^2)$. Here we assume that σ^2, μ, and τ^2 are all known.
(a) Find the joint pdf of $\bar{X}$ and θ.
(b) Show that $m(\bar{x} \mid \sigma^2,\mu,\tau^2)$, the marginal distribution of $\bar{X}$, is $\mathrm{n}(\mu,(\sigma^2/n)+\tau^2)$.
(c) Show that $\pi(\theta \mid \bar{x},\sigma^2,\mu,\tau^2)$, the posterior distribution of θ, is normal with mean and variance

given by (7.2.10), namely (for a single observation X with variance σ^2)

$$\mathrm{E}(\theta \mid x) = \frac{\tau^2}{\tau^2+\sigma^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\mu,$$
$$\mathrm{Var}(\theta \mid x) = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}.$$

Solution. (a) Since $\bar{X} \mid \theta \sim \mathrm{n}(\theta, \sigma^2/n)$ by Theorem 5.3.1(b), the joint pdf is

$$f(\bar{x},\theta) = \sqrt{\frac{n}{2\pi\sigma^2}}\, e^{-n(\bar{x}-\theta)^2/(2\sigma^2)} \cdot \frac{1}{\sqrt{2\pi\tau^2}}\, e^{-(\theta-\mu)^2/(2\tau^2)}.$$

Everything else is one completion of the square. Write $a = n/\sigma^2$, $b = 1/\tau^2$; then

$$a(\theta-\bar{x})^2 + b(\theta-\mu)^2 = (a+b)\Big(\theta - \tfrac{a\bar{x}+b\mu}{a+b}\Big)^2 + \frac{ab}{a+b}(\bar{x}-\mu)^2,$$

as one expands both sides (Check!). Hence

$$f(\bar{x},\theta) = \underbrace{\frac{\sqrt{ab}}{2\pi}\, e^{-\frac{ab}{2(a+b)}(\bar{x}-\mu)^2}}_{\text{free of } \theta}\; e^{-\frac{a+b}{2}\left(\theta - \frac{a\bar{x}+b\mu}{a+b}\right)^2}.$$

(b) Integrating the display over θ contributes the constant $\sqrt{2\pi/(a+b)}$, so

$$m(\bar{x}) = \frac{1}{\sqrt{2\pi}}\sqrt{\frac{ab}{a+b}}\; e^{-\frac{ab}{2(a+b)}(\bar{x}-\mu)^2},$$

a normal density with mean μ and variance

$$\frac{a+b}{ab} = \frac{1}{a} + \frac{1}{b} = \frac{\sigma^2}{n} + \tau^2.$$

(c) Dividing, $\pi(\theta \mid \bar{x}) = f(\bar{x},\theta)/m(\bar{x})$ is the θ-factor normalized, i.e. normal with

$$\mathrm{E}(\theta \mid \bar{x}) = \frac{a\bar{x}+b\mu}{a+b} = \frac{\tau^2}{\tau^2+\sigma^2/n}\,\bar{x} + \frac{\sigma^2/n}{\sigma^2/n+\tau^2}\,\mu,$$
$$\mathrm{Var}(\theta \mid \bar{x}) = \frac{1}{a+b} = \frac{(\sigma^2/n)\tau^2}{\sigma^2/n+\tau^2},$$

which is (7.2.10) with σ^2 replaced by σ^2/n.

Problem 7.23. If S^2 is the sample variance based on a sample of size n from a normal population, we know that $(n-1)S^2/\sigma^2$ has a χ^2_{n-1} distribution. The conjugate prior for σ^2 is the *inverted gamma* pdf, $\mathrm{IG}(\alpha,\beta)$, given by

$$\pi(\sigma^2) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\,\frac{1}{(\sigma^2)^{\alpha+1}}\, e^{-1/(\beta\sigma^2)}, \qquad 0 < \sigma^2 < \infty,$$

where α and β are positive constants. Show that the posterior distribution of σ^2 is

$$\mathrm{IG}\left(\alpha + \frac{n-1}{2},\ \left[\frac{(n-1)S^2}{2} + \frac{1}{\beta}\right]^{-1}\right).$$

Find the mean of this distribution, the Bayes estimator of σ^2.

Solution. The Bayes estimator is

$$\mathrm{E}(\sigma^2 \mid s^2) = \frac{\dfrac{(n-1)s^2}{2} + \dfrac{1}{\beta}}{\alpha + \dfrac{n-3}{2}}.$$

Transforming $W = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ back to S^2, with Jacobian $dw/ds^2 = (n-1)/\sigma^2$, the χ^2_{n-1} pdf $\propto w^{(n-3)/2}e^{-w/2}$ gives

$$f(s^2 \mid \sigma^2) = c(s^2, n)\,(\sigma^2)^{-(n-1)/2} e^{-(n-1)s^2/(2\sigma^2)},$$

with c free of σ^2. Multiplying by the prior and dropping factors free of σ^2,

$$\pi(\sigma^2 \mid s^2) \propto (\sigma^2)^{-(\alpha + \frac{n-1}{2} + 1)} \exp\left\{ -\frac{1}{\sigma^2}\left[\frac{(n-1)s^2}{2} + \frac{1}{\beta} \right] \right\},$$

which is the $\mathrm{IG}(\alpha', \beta')$ kernel with

$$\alpha' = \alpha + \frac{n-1}{2}, \qquad \beta' = \left[\frac{(n-1)s^2}{2} + \frac{1}{\beta} \right]^{-1}.$$

For the mean, substitute $u = 1/\sigma^2$ in the $\mathrm{IG}(a, b)$ density:

$$\mathrm{E}\,\sigma^2 = \frac{1}{\Gamma(a) b^a} \int_0^\infty (\sigma^2)^{-a} e^{-1/(b\sigma^2)}\, d\sigma^2 = \frac{1}{\Gamma(a) b^a} \int_0^\infty u^{a-2} e^{-u/b}\, du = \frac{1}{(a-1)b},$$

finite for $a > 1$. With $a = \alpha'$, $b = \beta'$ this is the mean displayed above, valid for $\alpha + (n-3)/2 > 0$.

Problem 7.24. Let $X_1, \ldots, X_n$ be iid Poisson(λ), and let λ have a gamma(α, β) distribution, the conjugate family for the Poisson.
(a) Find the posterior distribution of λ.
(b) Calculate the posterior mean and variance.

Solution. (a) The posterior is

$$\lambda \mid \mathbf{x} \sim \mathrm{gamma}\left(\sum_{i=1}^n x_i + \alpha,\ \frac{\beta}{n\beta + 1} \right).$$

Indeed, dropping factors free of λ,

$$\begin{aligned} \pi(\lambda \mid \mathbf{x}) &\propto \left(\prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \lambda^{\alpha - 1} e^{-\lambda/\beta} \\ &\propto \lambda^{\sum x_i + \alpha - 1} e^{-\lambda\left(n + \frac{1}{\beta}\right)}, \end{aligned}$$

and $(n + 1/\beta)^{-1} = \beta/(n\beta + 1)$.
(b) By the gamma moments (Section 3.3),

$$\begin{aligned} \mathrm{E}(\lambda \mid \mathbf{x}) &= \frac{\beta\,(\sum_{i=1}^n x_i + \alpha)}{n\beta + 1} = \frac{n\beta}{n\beta + 1}\,\bar{x} + \frac{1}{n\beta + 1}\,\alpha\beta, \\ \mathrm{Var}(\lambda \mid \mathbf{x}) &= \frac{\beta^2\,(\sum_{i=1}^n x_i + \alpha)}{(n\beta + 1)^2}, \end{aligned}$$

Problem 7.25. We examine a generalization of the hierarchical (Bayes) model considered in Example 7.2.16 and Exercise 7.22. Suppose that we observe $X_1, \dots, X_n$, where

$$\begin{aligned} X_i \mid \theta_i &\sim \mathrm{n}(\theta_i, \sigma^2), \quad i = 1, \dots, n, \quad \text{independent}, \\ \theta_i &\sim \mathrm{n}(\mu, \tau^2), \quad i = 1, \dots, n, \quad \text{independent}. \end{aligned}$$

(a) Show that the marginal distribution of X_i is $\mathrm{n}(\mu, \sigma^2 + \tau^2)$ and that, marginally, $X_1, \dots, X_n$ are iid. (*Empirical Bayes analysis* would use the marginal distribution of the X_is to estimate the prior parameters μ and τ^2. See Miscellanea 7.5.6.)
(b) Show, in general, that if

$$\begin{aligned} X_i \mid \theta_i &\sim f(x \mid \theta_i), \quad i = 1, \dots, n, \quad \text{independent}, \\ \theta_i &\sim \pi(\theta \mid \tau), \quad i = 1, \dots, n, \quad \text{independent}, \end{aligned}$$

then marginally, $X_1, \dots, X_n$ are iid.

Solution. (b) first, since (a) is a special case of it. The full joint density of $(\mathbf{X}, \boldsymbol{\theta})$ factors across i, so Fubini splits the n-fold integral into n single ones:

$$\begin{aligned} m(x_1, \dots, x_n) &= \int \cdots \int \prod_{i=1}^{n} f(x_i \mid \theta_i) \pi(\theta_i \mid \tau) \, d\theta_1 \cdots d\theta_n \\ &= \prod_{i=1}^{n} \int f(x_i \mid \theta_i) \pi(\theta_i \mid \tau) \, d\theta_i = \prod_{i=1}^{n} m(x_i), \end{aligned}$$

where $m(t) = \int f(t \mid \theta) \pi(\theta \mid \tau) \, d\theta$ is one and the same function of t for every i. A joint density that factors into identical marginals is exactly the iid condition (Definition 5.1.1).
(a) Apply this with $f(\cdot \mid \theta_i)$ the $\mathrm{n}(\theta_i, \sigma^2)$ density and $\pi(\cdot \mid \tau)$ the $\mathrm{n}(\mu, \tau^2)$ density: the one-dimensional integral $m(x_i)$ is the marginal computed in Exercise 7.22(b) with sample size 1, namely

$$m(x_i) = \frac{1}{\sqrt{2\pi(\sigma^2 + \tau^2)}} \, e^{-(x_i - \mu)^2 / \left(2(\sigma^2 + \tau^2)\right)},$$

the $\mathrm{n}(\mu, \sigma^2 + \tau^2)$ density. Hence $X_1, \dots, X_n$ are iid $\mathrm{n}(\mu, \sigma^2 + \tau^2)$.

Problem 7.26. In Example 7.2.16 we saw that the normal distribution is its own conjugate family. It is sometimes the case, however, that a conjugate prior does not accurately reflect prior knowledge, and a different prior is sought. Let $X_1, \dots, X_n$ be iid $\mathrm{n}(\theta, \sigma^2)$, and let θ have a double exponential distribution, that is, $\pi(\theta) = e^{-|\theta|/a}/(2a)$, a known. Find the mean of the posterior distribution of θ.

Solution. Write $v = \sigma^2/n$, $s = \sigma/\sqrt{n}$, and

$$\mu_+ = \bar{x} - \frac{v}{a}, \qquad \mu_- = \bar{x} + \frac{v}{a};$$

then, with Φ and φ the standard normal cdf and pdf,

$$\mathrm{E}(\theta \mid \mathbf{x}) = \frac{e^{-\bar{x}/a} \, \mu_+ \Phi(\mu_+/s) + e^{\bar{x}/a} \, \mu_- \Phi(-\mu_-/s)}{e^{-\bar{x}/a} \, \Phi(\mu_+/s) + e^{\bar{x}/a} \, \Phi(-\mu_-/s)}.$$

By sufficiency of $\bar{X}$ (Example 6.2.4) the posterior is proportional to $e^{-(\theta - \bar{x})^2/(2v)} e^{-|\theta|/a}$. Split on the sign of θ and complete the square in each piece:

$$\begin{aligned} -\frac{(\theta - \bar{x})^2}{2v} - \frac{\theta}{a} &= -\frac{(\theta - \mu_+)^2}{2v} - \frac{\bar{x}}{a} + \frac{v}{2a^2}, \quad \theta > 0, \\ -\frac{(\theta - \bar{x})^2}{2v} + \frac{\theta}{a} &= -\frac{(\theta - \mu_-)^2}{2v} + \frac{\bar{x}}{a} + \frac{v}{2a^2}, \quad \theta < 0. \end{aligned}$$

Writing $\phi_v(t) = s^{-1}\varphi(t/s)$ for the $\mathrm{n}(0, v)$ density and discarding the common factor $e^{v/(2a^2)}/(2a)$,

$$\pi(\theta \mid \mathbf{x}) \propto e^{-\bar{x}/a}\phi_v(\theta - \mu_+)I_{(0,\infty)}(\theta) + e^{\bar{x}/a}\phi_v(\theta - \mu_-)I_{(-\infty,0)}(\theta).$$

The two needed truncated moments are, for any m,

$$\int_0^\infty \phi_v(\theta - m)\,d\theta = \Phi(m/s), \qquad \int_0^\infty \theta\,\phi_v(\theta - m)\,d\theta = m\Phi(m/s) + s\varphi(m/s),$$
$$\int_{-\infty}^0 \phi_v(\theta - m)\,d\theta = \Phi(-m/s), \quad \int_{-\infty}^0 \theta\,\phi_v(\theta - m)\,d\theta = m\Phi(-m/s) - s\varphi(m/s),$$

by the substitution $\theta = m + sz$ and $\int z\varphi(z)\,dz = -\varphi(z)$. The two $s\varphi$ terms cancel in the numerator, because

$$e^{-\bar{x}/a}\varphi(\mu_+/s) = \frac{1}{\sqrt{2\pi}}e^{-\bar{x}^2/(2v) - v/(2a^2)} = e^{\bar{x}/a}\varphi(\mu_-/s),$$

both exponents equalling $-\bar{x}^2/(2v) - v/(2a^2)$. Dividing numerator by denominator gives the expression displayed above.

Problem 7.27. Refer to Example 7.2.17: we observe $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, all mutually independent, where $Y_i \sim \text{Poisson}(\beta\tau_i)$ and $X_i \sim \text{Poisson}(\tau_i)$, so that the complete-data pmf (7.2.11) is

$$f\big((x_1, y_1), \ldots, (x_n, y_n) \mid \beta, \tau_1, \ldots, \tau_n\big) = \prod_{i=1}^n \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!}\,\frac{e^{-\tau_i}(\tau_i)^{x_i}}{x_i!}.$$

(a) Show that the likelihood estimators from the complete-data likelihood (7.2.11) are given by (7.2.12), that is,

$$\hat{\beta} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}, \qquad \hat{\tau}_j = \frac{x_j + y_j}{\hat{\beta} + 1}, \quad j = 1, 2, \ldots, n.$$

(b) Show that the limit of the EM sequence in (7.2.23),

$$\hat{\beta}^{(r+1)} = \frac{\sum_{i=1}^n y_i}{\hat{\tau}_1^{(r)} + \sum_{i=2}^n x_i}, \qquad \hat{\tau}_1^{(r+1)} = \frac{\hat{\tau}_1^{(r)} + y_1}{\hat{\beta}^{(r+1)} + 1},$$
$$\hat{\tau}_j^{(r+1)} = \frac{x_j + y_j}{\hat{\beta}^{(r+1)} + 1}, \qquad j = 2, 3, \ldots, n,$$

satisfies the incomplete-data likelihood equations (7.2.16) (the case x_1 missing),

$$\hat{\beta} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n \hat{\tau}_i}, \qquad y_1 = \hat{\tau}_1\hat{\beta}, \qquad x_j + y_j = \hat{\tau}_j(\hat{\beta} + 1), \quad j = 2, \ldots, n.$$

(c) A direct solution of the original (incomplete-data) likelihood equations is possible. Show that the solution to (7.2.16) is given by

$$\hat{\beta} = \frac{\sum_{i=2}^n y_i}{\sum_{i=2}^n x_i}, \qquad \hat{\tau}_1 = \frac{y_1}{\hat{\beta}}, \qquad \hat{\tau}_j = \frac{x_j + y_j}{\hat{\beta} + 1}, \quad j = 2, 3, \ldots, n,$$

and that this is the limit of the EM sequence in (7.2.23).

Solution. (a) The complete-data log likelihood is, up to terms free of the parameters,

$$\ell = \sum_{i=1}^n \Big[-\beta\tau_i + y_i \log\beta - \tau_i + (x_i + y_i)\log\tau_i\Big],$$

so setting $\partial\ell/\partial\beta$ and $\partial\ell/\partial\tau_j$ to zero gives

$$\sum_{i=1}^{n}\tau_i = \frac{\sum_{i=1}^{n} y_i}{\beta}, \qquad \tau_j(\beta+1) = x_j + y_j.$$

Summing the second over j and substituting into the first, $\hat\beta\sum_i(x_i+y_i) = (\hat\beta+1)\sum_i y_i$, i.e. $\hat\beta\sum_i x_i = \sum_i y_i$, which is (7.2.12). (The stationary point is the maximum: ℓ is concave in $\log\beta, \log\tau_j$ and $\to -\infty$ on the boundary.)

(b) The three maps in (7.2.23) are continuous where $\tau_1^{(r)} \geq 0$, so a limit $(\hat\beta, \hat\tau_1, \ldots, \hat\tau_n)$ of the sequence is a fixed point of them. The third recursion is literally the third equation of (7.2.16); the second gives $\hat\tau_1(\hat\beta+1) = \hat\tau_1 + y_1$, i.e. $y_1 = \hat\tau_1\hat\beta$. For the first, write $S_y = \sum_{i=1}^{n} y_i$ and use the first recursion together with $y_1 = \hat\tau_1\hat\beta$:

$$\hat\beta\sum_{i=2}^{n} x_i = S_y - \hat\beta\hat\tau_1 = S_y - y_1 = \sum_{i=2}^{n} y_i.$$

Hence

$$\begin{aligned}\hat\beta\sum_{i=1}^{n}\hat\tau_i &= \hat\beta\hat\tau_1 + \frac{\hat\beta}{\hat\beta+1}\sum_{i=2}^{n}(x_i+y_i)\\ &= y_1 + \frac{\hat\beta}{\hat\beta+1}\left(\frac{1}{\hat\beta}+1\right)\sum_{i=2}^{n} y_i = y_1 + \sum_{i=2}^{n} y_i = S_y,\end{aligned}$$

which is the first equation of (7.2.16).

(c) Substituting $\hat\tau_1 = y_1/\hat\beta$ and $\hat\tau_j = (x_j+y_j)/(\hat\beta+1)$ into $\hat\beta\sum_i\hat\tau_i = S_y$ gives

$$y_1 + \frac{\hat\beta}{\hat\beta+1}\sum_{i=2}^{n}(x_i+y_i) = S_y \iff \hat\beta\sum_{i=2}^{n} x_i = \sum_{i=2}^{n} y_i,$$

the stated $\hat\beta$; the other two displayed formulas are then the remaining two equations of (7.2.16) read off directly.

For the limit, note the EM recursion collapses to a scalar one in $t = \hat\tau_1$. Put $S_x = \sum_{i=2}^{n} x_i$; then $\hat\beta^{(r+1)} = S_y/(t^{(r)} + S_x)$ and

$$t^{(r+1)} = g(t^{(r)}), \qquad g(t) = \frac{(t+y_1)(t+S_x)}{t+S_x+S_y}.$$

On $t \geq 0$,

$$g'(t) = \frac{t^2 + 2t(S_x+S_y) + S_x^2 + S_xS_y + y_1S_y}{(t+S_x+S_y)^2} > 0,$$

so g is increasing, and

$$g(t) - t = \frac{y_1S_x - t\,(S_y - y_1)}{t+S_x+S_y}$$

is positive for $t < t^*$ and negative for $t > t^*$, where

$$t^* = \frac{y_1S_x}{S_y - y_1} = \frac{y_1\sum_{i=2}^{n} x_i}{\sum_{i=2}^{n} y_i} = \frac{y_1}{\hat\beta}$$

is the unique fixed point (assuming $\sum_{i\geq 2} y_i > 0$). An increasing map with a single fixed point drives $\{t^{(r)}\}$ monotonically and boundedly toward t^*, so $t^{(r)} \to t^* = \hat\tau_1$. Consequently

$$\hat\beta^{(r)} \to \frac{S_y}{t^*+S_x} = \frac{S_y(S_y-y_1)}{S_xS_y} = \frac{\sum_{i=2}^{n} y_i}{\sum_{i=2}^{n} x_i} = \hat\beta,$$

and $\hat\tau_j^{(r)} \to (x_j+y_j)/(\hat\beta+1)$ for $j \geq 2$, which is the solution found above.

Problem 7.28. Use the model of Example 7.2.17 ($Y_i \sim \text{Poisson}(\beta\tau_i)$, $X_i \sim \text{Poisson}(\tau_i)$, all mutually independent) on the data in the following table adapted from Lange et al. (1994). These are leukemia counts and the associated populations for a number of areas in New York State.

Counts of leukemia cases

Population	3540	3560	3739	2784	2571	2729	3952	993	1908
Number of cases	3	4	1	1	3	1	2	0	2
Population	948	1172	1047	3138	5485	5554	2943	4969	4828
Number of cases	0	1	3	5	4	6	2	5	4

(a) Fit the Poisson model to these data both to the full data set and to an "incomplete" data set where we suppose that the first population count ($x_1 = 3540$) is missing.
(b) Suppose that instead of having an x value missing, we actually have lost a leukemia count (assume that $y_1 = 3$ is missing). Use the EM algorithm to find the MLEs in this case, and compare your answers to those of part (a).

Solution. Here $n = 18$, $\sum_{i=1}^{18} x_i = 55{,}860$, $\sum_{i=1}^{18} y_i = 47$, and with the first area removed $\sum_{i=2}^{18} x_i = 52{,}320$, $\sum_{i=2}^{18} y_i = 44$. The three fits are

$$\begin{aligned}
\text{full data:}\quad & \hat\beta = \frac{47}{55{,}860} = 8.4139\times10^{-4},\\
x_1\text{ missing:}\quad & \hat\beta = \frac{44}{52{,}320} = 8.4098\times10^{-4},\quad \hat\tau_1 = \frac{y_1}{\hat\beta} = 3567.27,\\
y_1\text{ missing:}\quad & \hat\beta = \frac{44}{52{,}320} = 8.4098\times10^{-4},\quad \hat\tau_1 = x_1 = 3540,
\end{aligned}$$

with $\hat\tau_j = (x_j + y_j)/(\hat\beta + 1)$ in every case for $j \geq 2$.
(a) The full-data fit is (7.2.12); the x_1-missing fit is Exercise 7.27(c), which is the limit of the EM sequence (7.2.23).
(b) With Y_1 unobserved, the complete-data log likelihood

$$\ell = \sum_{i=1}^{n}\big[-\beta\tau_i + y_i\log(\beta\tau_i) - \log y_i!\big] + \sum_{i=1}^{n}\big[-\tau_i + x_i\log\tau_i - \log x_i!\big]$$

is linear in y_1 apart from $-\log y_1!$, which is free of the parameters. Since $Y_1 \mid \beta^{(r)}, \tau_1^{(r)} \sim \text{Poisson}(\beta^{(r)}\tau_1^{(r)})$, the E-step of (7.2.20) therefore amounts to replacing y_1 by $\beta^{(r)}\tau_1^{(r)}$ (exactly as (7.2.22) replaced x_1 by $\tau_1^{(r)}$), and the M-step is (7.2.12) with that substitution:

$$\begin{aligned}
\hat\beta^{(r+1)} &= \frac{\hat\beta^{(r)}\hat\tau_1^{(r)} + \sum_{i=2}^{n} y_i}{\sum_{i=1}^{n} x_i}, \qquad \hat\tau_1^{(r+1)} = \frac{x_1 + \hat\beta^{(r)}\hat\tau_1^{(r)}}{\hat\beta^{(r+1)} + 1},\\
\hat\tau_j^{(r+1)} &= \frac{x_j + y_j}{\hat\beta^{(r+1)} + 1}, \qquad j = 2,\dots,n.
\end{aligned}$$

At a fixed point the second equation gives $\hat\tau_1(\hat\beta + 1) = x_1 + \hat\beta\hat\tau_1$, i.e. $\hat\tau_1 = x_1$, and then the first gives

$$\hat\beta\Big(x_1 + \sum_{i=2}^{n} x_i\Big) = \hat\beta x_1 + \sum_{i=2}^{n} y_i \iff \hat\beta = \frac{\sum_{i=2}^{n} y_i}{\sum_{i=2}^{n} x_i}.$$

The same values come from the incomplete-data likelihood directly: omitting the Y_1 factor leaves $\partial\ell/\partial\tau_1 = -1 + x_1/\tau_1$, so $\hat\tau_1 = x_1$, while the β and τ_j ($j \geq 2$) equations are those of Exercise 7.27(c) restricted to $i \geq 2$. Numerically the EM sequence started at $\hat\beta^{(0)} = \hat\tau_1^{(0)} = 1$ reaches $\hat\beta = 8.4098\times10^{-4}$, $\hat\tau_1 = 3540.00$.
Comparison: losing y_1 and losing x_1 give the identical $\hat\beta$ and identical $\hat\tau_j$ for $j \geq 2$; they differ only in $\hat\tau_1$, which is estimated from whichever of the two counts survives ($x_1 = 3540$ versus $y_1/\hat\beta = 3567.27$). Both are within 1% of the full-data $\hat\tau_1 = 3540.02$.

i	x_i	y_i	$\hat{\tau}_i$ full	$\hat{\tau}_i$, x_1 miss.	$\hat{\tau}_i$, y_1 miss.
1	3540	3	3540.02	3567.27	3540.00
2	3560	4	3561.00	3561.01	3561.01
3	3739	1	3736.86	3736.86	3736.86
4	2784	1	2782.66	2782.66	2782.66
5	2571	3	2571.84	2571.84	2571.84
6	2729	1	2727.70	2727.71	2727.71
7	3952	2	3950.68	3950.68	3950.68
8	993	0	992.17	992.17	992.17
9	1908	2	1908.39	1908.40	1908.40
10	948	0	947.20	947.20	947.20
11	1172	1	1172.01	1172.01	1172.01
12	1047	3	1049.12	1049.12	1049.12
13	3138	5	3140.36	3140.36	3140.36
14	5485	4	5484.39	5484.39	5484.39
15	5554	6	5555.33	5555.33	5555.33
16	2943	2	2942.52	2942.53	2942.53
17	4969	5	4969.82	4969.82	4969.82
18	4828	4	4827.94	4827.94	4827.94

7.5 Exercises 7.29–7.35

Problem 7.29. An alternative to the model of Example 7.2.17 is the following, where we observe (Y_i, X_i), $i = 1, 2, \ldots, n$, where $Y_i \sim \text{Poisson}(m\beta\tau_i)$ and $(X_1, \ldots, X_n) \sim \text{multinomial}(m; \tau)$, where $\tau = (\tau_1, \tau_2, \ldots, \tau_n)$ with $\sum_{i=1}^n \tau_i = 1$. So here, for example, we assume that the population counts are multinomial allocations rather than Poisson counts. (Treat $m = \sum x_i$ as known.)
(a) Show that the joint density of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{X} = (X_1, \ldots, X_n)$ is

$$f(\mathbf{y}, \mathbf{x} \,|\, \beta, \tau) = \prod_{i=1}^n \frac{e^{-m\beta\tau_i}(m\beta\tau_i)^{y_i}}{y_i!}\, m!\, \frac{\tau_i^{x_i}}{x_i!}.$$

(b) If the complete data are observed, show that the MLEs are given by

$$\hat{\beta} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} \quad \text{and} \quad \hat{\tau}_j = \frac{x_j + y_j}{\sum_{i=1}^n x_i + y_i}, \qquad j = 1, 2, \ldots, n.$$

(c) Suppose that x_1 is missing. Use the fact that $X_1 \sim \text{binomial}(m, \tau_1)$ to calculate the expected complete-data log likelihood. Show that the EM sequence is given by

$$\hat{\beta}^{(r+1)} = \frac{\sum_{i=1}^n y_i}{m\hat{\tau}_1^{(r)} + \sum_{i=2}^n x_i}, \qquad \hat{\tau}_j^{(r+1)} = \frac{x_j + y_j}{m\hat{\tau}_1^{(r)} + \sum_{i=2}^n x_i + \sum_{i=1}^n y_i},$$

$j = 1, 2, \ldots, n$.
(d) Use this model to find the MLEs for the data in Exercise 7.28, first assuming that you have all the data and then assuming that $x_1 = 3540$ is missing. The data are leukemia counts y_i and the associated populations x_i for $n = 18$ areas in New York State (adapted from Lange et al. 1994):

i	1	2	3	4	5	6	7	8	9
Population x_i	3540	3560	3739	2784	2571	2729	3952	993	1908
Cases y_i	3	4	1	1	3	1	2	0	2
i	10	11	12	13	14	15	16	17	18
Population x_i	948	1172	1047	3138	5485	5554	2943	4969	4828
Cases y_i	0	1	3	5	4	6	2	5	4

Solution. (a) $\mathbf{Y}$ and $\mathbf{X}$ are independent, so the joint pmf is the product of the n independent Poisson pmfs with the one multinomial pmf,

$$f(\mathbf{y},\mathbf{x}\,|\,\beta,\tau)=\left[\prod_{i=1}^{n}\frac{e^{-m\beta\tau_i}(m\beta\tau_i)^{y_i}}{y_i!}\right]\cdot m!\prod_{i=1}^{n}\frac{\tau_i^{x_i}}{x_i!},$$

which is the printed display (whose single factor $m!$ is displayed inside the product sign but belongs outside it).
(b) Taking logs and using $\sum_i\tau_i=1$,

$$\begin{aligned}\log L(\beta,\tau\,|\,\mathbf{y},\mathbf{x})&=-m\beta\sum_{i=1}^{n}\tau_i+\Big(\sum_{i=1}^{n}y_i\Big)\log(m\beta)+\sum_{i=1}^{n}(x_i+y_i)\log\tau_i+c\\&=-\beta\sum_{i=1}^{n}x_i+\Big(\sum_{i=1}^{n}y_i\Big)\log\beta+\sum_{i=1}^{n}(x_i+y_i)\log\tau_i+c',\end{aligned}$$

where c,c' are free of (β,τ) and the second line writes $m=\sum_i x_i$, keeping the dependence on x_1 explicit for (c). The β part is maximized at

$$-\sum_{i=1}^{n}x_i+\frac{\sum_i y_i}{\beta}=0\quad\Longrightarrow\quad\hat\beta=\frac{\sum_{i=1}^{n}y_i}{\sum_{i=1}^{n}x_i},$$

and, β having dropped out of the τ part, what remains is a multinomial log likelihood with cell counts x_i+y_i; maximizing $\sum_i(x_i+y_i)\log\tau_i$ over the simplex gives the usual multinomial MLE

$$\hat\tau_j=\frac{x_j+y_j}{\sum_{i=1}^{n}(x_i+y_i)}.$$

(Lagrange: $(x_j+y_j)/\tau_j=\lambda$ for all j, so $\tau_j\propto x_j+y_j$.)
(c) The complete-data log likelihood is linear in x_1 apart from the parameter-free term $-\log x_1!$, so the E-step is pure substitution: with $X_1\sim\text{binomial}(m,\tau_1)$, the marginal of the multinomial (the exercise directs us to this unconditional law, since conditioning as well on the known $m=\sum_i x_i$ would recover x_1 exactly),

$$\mathrm{E}\big[X_1\,|\,\hat\tau_1^{(r)}\big]=m\hat\tau_1^{(r)},$$

and

$$\begin{aligned}\mathrm{E}^{(r)}\big[\log L\big]&=-\beta\Big(m\hat\tau_1^{(r)}+\sum_{i=2}^{n}x_i\Big)+\Big(\sum_{i=1}^{n}y_i\Big)\log\beta\\&\quad+\big(m\hat\tau_1^{(r)}+y_1\big)\log\tau_1+\sum_{i=2}^{n}(x_i+y_i)\log\tau_i+c,\end{aligned}$$

where $\mathrm{E}^{(r)}$ denotes expectation given the observed data and $(\hat\beta^{(r)},\hat\tau^{(r)})$. This is exactly the complete-data log likelihood of part (b) with x_1 replaced by $m\hat\tau_1^{(r)}$, so the M-step is (b) with that substitution:

$$\hat\beta^{(r+1)}=\frac{\sum_{i=1}^{n}y_i}{m\hat\tau_1^{(r)}+\sum_{i=2}^{n}x_i},$$

$$\begin{aligned}\hat\tau_1^{(r+1)}&=\frac{m\hat\tau_1^{(r)}+y_1}{m\hat\tau_1^{(r)}+\sum_{i=2}^{n}x_i+\sum_{i=1}^{n}y_i},\\\hat\tau_j^{(r+1)}&=\frac{x_j+y_j}{m\hat\tau_1^{(r)}+\sum_{i=2}^{n}x_i+\sum_{i=1}^{n}y_i},\qquad j=2,\dots,n.\end{aligned}$$

(The printed answer writes numerator x_j+y_j for every j; at $j=1$ the missing x_1 is of course read as $m\hat\tau_1^{(r)}$.)

(d) Here $n = 18$, $m = \sum_i x_i = 55{,}860$, $\sum_i y_i = 47$, $\sum_i (x_i + y_i) = 55{,}907$. With all the data,

$$\hat\beta = \frac{47}{55{,}860} = 8.4139 \times 10^{-4}, \qquad \hat\tau_j = \frac{x_j + y_j}{55{,}907}.$$

With x_1 missing, the fixed point of the τ_1 recursion in (c) solves

$$m\tau_1^2 + \Big(\sum_{i=2}^{n} x_i + \sum_{i=1}^{n} y_i - m\Big)\tau_1 - y_1 = 0,$$

that is $55{,}860\,\tau_1^2 - 3493\,\tau_1 - 3 = 0$, whose positive root is

$$\hat\tau_1 = 0.063379, \qquad m\hat\tau_1 = 3540.33,$$

against the true $x_1 = 3540$. Then

$$\hat\beta = \frac{47}{3540.33 + 52{,}320} = 8.4138 \times 10^{-4}, \qquad \hat\tau_1 = \frac{3543.33}{55{,}907.33},$$

with $\hat\tau_j = (x_j + y_j)/55{,}907.33$ for $j \ge 2$; these reproduce the full-data values tabulated below to five decimals, except $\hat\tau_1 = 0.06338$.

j	1	2	3	4	5	6	7	8	9
$\hat\tau_j$	0.06337	0.06375	0.06690	0.04981	0.04604	0.04883	0.07072	0.01776	0.03416
j	10	11	12	13	14	15	16	17	18
$\hat\tau_j$	0.01696	0.02098	0.01878	0.05622	0.09818	0.09945	0.05268	0.08897	0.08643

Problem 7.30. The EM algorithm is useful in a variety of situations, and the definition of "missing data" can be stretched to accommodate many different models. Suppose that we have a mixture density $pf(x)+(1-p)g(x)$, where p is unknown. If we observe $\mathbf{X} = (X_1, \ldots, X_n)$, the sample density is

$$\prod_{i=1}^{n} \big[pf(x_i) + (1-p)g(x_i)\big],$$

which could be difficult to deal with. (Actually, a mixture of two is not terrible, but consider what the likelihood would look like with a mixture $\sum_{i=1}^{k} p_i f_i(x)$ for large k.) The EM solution is to augment the observed (or incomplete) data with $\mathbf{Z} = (Z_1, \ldots, Z_n)$, where Z_i tells which component of the mixture X_i came from; that is,

$$X_i | z_i = 1 \sim f(x_i) \quad \text{and} \quad X_i | z_i = 0 \sim g(x_i),$$

and $P(Z_i = 1) = p$.
(a) Show that the joint density of $(\mathbf{X}, \mathbf{Z})$ is given by

$$\prod_{i=1}^{n} \big[pf(x_i)\big]^{z_i} \big[(1-p)g(x_i)\big]^{1-z_i}.$$

(b) Show that the missing data distribution, the distribution of $Z_i | x_i, p$, is Bernoulli with success probability $pf(x_i)/(pf(x_i) + (1-p)g(x_i))$.
(c) Calculate the expected complete-data log likelihood, and show that the EM sequence is given by

$$\hat p^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat p^{(r)} f(x_i)}{\hat p^{(r)} f(x_i) + (1 - \hat p^{(r)}) g(x_i)}.$$

Solution. (a) $P(Z_i = z_i) = p^{z_i}(1-p)^{1-z_i}$ and $X_i | z_i$ has density $f(x_i)^{z_i} g(x_i)^{1-z_i}$, so the pair has joint

density

$$f(x_i, z_i \,|\, p) = \big[pf(x_i)\big]^{z_i}\big[(1-p)g(x_i)\big]^{1-z_i},$$

and the n pairs are independent. Summing over $z_i \in \{0,1\}$ returns $pf(x_i) + (1-p)g(x_i)$, so this augmentation does reproduce the stated incomplete-data density.
(b) By Bayes' rule, dividing the joint density of (a) by that marginal,

$$P(Z_i = 1 \,|\, x_i, p) = \frac{pf(x_i)}{pf(x_i) + (1-p)g(x_i)},$$

a Bernoulli distribution since Z_i takes only the values 0 and 1.
(c) The complete-data log likelihood is linear in the z_i,

$$\log L(p \,|\, \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n}\Big[z_i \log p + (1-z_i)\log(1-p)\Big] + \sum_{i=1}^{n}\Big[z_i \log f(x_i) + (1-z_i)\log g(x_i)\Big],$$

so the E-step only replaces each z_i by its conditional mean from (b),

$$\hat{z}_i^{(r)} = \mathrm{E}\big[Z_i \,\big|\, x_i, \hat{p}^{(r)}\big] = \frac{\hat{p}^{(r)} f(x_i)}{\hat{p}^{(r)} f(x_i) + (1-\hat{p}^{(r)})g(x_i)}.$$

Writing $S_r = \sum_{i=1}^{n} \hat{z}_i^{(r)}$, the part of the expected complete-data log likelihood that involves p is

$$S_r \log p + (n - S_r)\log(1-p),$$

a binomial log likelihood, maximized at $p = S_r/n$. Hence

$$\hat{p}^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{p}^{(r)} f(x_i)}{\hat{p}^{(r)} f(x_i) + (1-\hat{p}^{(r)})g(x_i)}.$$

Problem 7.31. Prove Theorem 7.2.20. (Theorem 7.2.20, Monotonic EM sequence: the sequence $\{\hat{\theta}_{(r)}\}$ defined by (7.2.20), $\hat{\theta}^{(r+1)}$ = the value maximizing $\mathrm{E}[\log L(\theta|\mathbf{y}, \mathbf{X})|\hat{\theta}^{(r)}, \mathbf{y}]$, satisfies $L(\hat{\theta}^{(r+1)}|\mathbf{y}) \geq L(\hat{\theta}^{(r)}|\mathbf{y})$, with equality holding if and only if successive iterations yield the same value of the maximized expected complete-data log likelihood. Here $k(\mathbf{x}|\theta, \mathbf{y}) = f(\mathbf{y}, \mathbf{x}|\theta)/g(\mathbf{y}|\theta)$ and (7.2.19) reads $\log L(\theta|\mathbf{y}) = \mathrm{E}[\log L(\theta|\mathbf{y}, \mathbf{X})|\theta', \mathbf{y}] - \mathrm{E}[\log k(\mathbf{X}|\theta, \mathbf{y})|\theta', \mathbf{y}]$.)
(a) Show that, using (7.2.19), we can write

$$\begin{aligned}\log L(\hat{\theta}^{(r)}|\mathbf{y}) = {} & \mathrm{E}\big[\log L(\hat{\theta}^{(r)}|\mathbf{y}, \mathbf{X}) \,\big|\, \hat{\theta}^{(r)}, \mathbf{y}\big] \\ & - \mathrm{E}\big[\log k(\mathbf{X}|\hat{\theta}^{(r)}, \mathbf{y}) \,\big|\, \hat{\theta}^{(r)}, \mathbf{y}\big],\end{aligned}$$

and, since $\hat{\theta}^{(r+1)}$ is a maximum,

$$\mathrm{E}\big[\log L(\hat{\theta}^{(r+1)}|\mathbf{y}, \mathbf{X}) \,\big|\, \hat{\theta}^{(r)}, \mathbf{y}\big] \geq \mathrm{E}\big[\log L(\hat{\theta}^{(r)}|\mathbf{y}, \mathbf{X}) \,\big|\, \hat{\theta}^{(r)}, \mathbf{y}\big].$$

When is the inequality an equality?
(b) Now use Jensen's inequality to show that

$$\mathrm{E}\big[\log k(\mathbf{X}|\hat{\theta}^{(r+1)}, \mathbf{y}) \,\big|\, \hat{\theta}^{(r)}, \mathbf{y}\big] \leq \mathrm{E}\big[\log k(\mathbf{X}|\hat{\theta}^{(r)}, \mathbf{y}) \,\big|\, \hat{\theta}^{(r)}, \mathbf{y}\big],$$

which together with part (a) proves the theorem. (Hint: If f and g are densities, since log is a concave function, Jensen's inequality (4.7.7) implies

$$\int \log\left(\frac{f(x)}{g(x)}\right) g(x)\,dx \leq \log\left(\int \frac{f(x)}{g(x)} g(x)\,dx\right) = \log\left(\int f(x)\,dx\right) = 0.$$

By the property of logs, this in turn implies that $\int \log[f(x)]g(x)\,dx \leq \int \log[g(x)]g(x)\,dx$.)

Solution. (a) Abbreviate the two conditional expectations in (7.2.19), taken with $\theta' = \hat\theta^{(r)}$ fixed, by

$$Q(\theta) = \mathrm{E}\big[\log L(\theta|\mathbf{y},\mathbf{X}) \,\big|\, \hat\theta^{(r)}, \mathbf{y}\big],$$

$$H(\theta) = \mathrm{E}\big[\log k(\mathbf{X}|\theta,\mathbf{y}) \,\big|\, \hat\theta^{(r)}, \mathbf{y}\big],$$

so that (7.2.19) reads $\log L(\theta|\mathbf{y}) = Q(\theta) - H(\theta)$ for every θ; in particular at $\theta = \hat\theta^{(r)}$ this is the first display of the exercise. Subtracting the identity at $\theta = \hat\theta^{(r)}$ from the identity at $\theta = \hat\theta^{(r+1)}$,

$$\log L(\hat\theta^{(r+1)}|\mathbf{y}) - \log L(\hat\theta^{(r)}|\mathbf{y}) = \underbrace{\big[Q(\hat\theta^{(r+1)}) - Q(\hat\theta^{(r)})\big]}_{A} - \underbrace{\big[H(\hat\theta^{(r+1)}) - H(\hat\theta^{(r)})\big]}_{B}.$$

By the definition (7.2.20), $\hat\theta^{(r+1)}$ maximizes Q, and $\hat\theta^{(r)}$ is one competitor in that maximization; hence $A \ge 0$. The inequality is an equality exactly when $\hat\theta^{(r)}$ is itself a maximizer of Q, i.e. when the two successive maximized values coincide.
(b) $B \le 0$: the conditional expectation defining H is taken under the density $k(\cdot|\hat\theta^{(r)},\mathbf{y})$, so the hint applies verbatim with $f = k(\cdot|\hat\theta^{(r+1)},\mathbf{y})$ and $g = k(\cdot|\hat\theta^{(r)},\mathbf{y})$ (both are genuine densities in $\mathbf{x}$, integrating to 1, by (7.2.17)):

$$\begin{aligned} B &= \int \log\left(\frac{k(\mathbf{x}|\hat\theta^{(r+1)},\mathbf{y})}{k(\mathbf{x}|\hat\theta^{(r)},\mathbf{y})}\right) k(\mathbf{x}|\hat\theta^{(r)},\mathbf{y})\,d\mathbf{x} \\ &\le \log \int k(\mathbf{x}|\hat\theta^{(r+1)},\mathbf{y})\,d\mathbf{x} = \log 1 = 0, \end{aligned}$$

Jensen's inequality (4.7.7) applying because log is concave.
Combining, $\log L(\hat\theta^{(r+1)}|\mathbf{y}) - \log L(\hat\theta^{(r)}|\mathbf{y}) = A - B \ge 0$, which is (7.2.24).
Equality in (7.2.24) means $A = B$, and since $A \ge 0 \ge B$ both vanish; $A = 0$ is exactly the stated condition, that successive iterations yield the same maximized expected complete-data log likelihood.

Problem 7.32. The algorithm of Exercise 5.65 (the Sampling/Importance Resampling algorithm: draw $Y_1,\dots,Y_m$ iid from g, form weights $q_i \propto f(Y_i)/g(Y_i)$, and resample from the discrete distribution putting mass q_i on Y_i to get an approximate sample from f) can be adapted to simulate (approximately) a sample from the posterior distribution using only a sample from the prior distribution. Let $X_1,\dots,X_n \sim f(x|\theta)$, where θ has prior distribution π. Generate $\theta_1,\dots,\theta_m$ from π, and calculate $q_i = L(\theta_i|\mathbf{x})/\sum_{j=1}^m L(\theta_j|\mathbf{x})$, where $L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta)$ is the likelihood function.
(a) Generate $\theta_1^*,\dots,\theta_r^*$, where $P(\theta^* = \theta_i) = q_i$. Show that this is an (approximate) sample from the posterior in the sense that $P(\theta^* \le t)$ converges to $\int_{-\infty}^t \pi(\theta|\mathbf{x})\,d\theta$.
(b) Show that the estimator $\sum_{j=1}^r h(\theta_j^*)/r$ converges to $\mathrm{E}[h(\theta)|\mathbf{x}]$, where the expectation is with respect to the posterior distribution.
(c) Ross (1996) suggests that Rao-Blackwellization can improve the estimate in part (b). Show that for any j,

$$\mathrm{E}[h(\theta_j^*)\,|\,\theta_1,\dots,\theta_m] = \frac{1}{\sum_{i=1}^m L(\theta_i|\mathbf{x})}\sum_{i=1}^m h(\theta_i)L(\theta_i|\mathbf{x})$$

has the same mean and smaller variance than the estimator in part (b).

Solution. (a) Condition on $\theta_1,\dots,\theta_m$ and write the resampling cdf as a ratio of two sample averages over the iid prior draws:

$$P(\theta^* \le t\,|\,\theta_1,\dots,\theta_m) = \sum_{i=1}^m q_i I(\theta_i \le t) = \frac{\frac{1}{m}\sum_{i=1}^m L(\theta_i|\mathbf{x})I(\theta_i \le t)}{\frac{1}{m}\sum_{i=1}^m L(\theta_i|\mathbf{x})}.$$

The θ_i are iid π and $L(\theta|\mathbf{x})$ has finite prior variance (the hypothesis of Theorem 5.5.2), so by the WLLN the numerator converges in probability to $\mathrm{E}_\pi[L(\theta|\mathbf{x})I(\theta \le t)] = \int_{-\infty}^t L(\theta|\mathbf{x})\pi(\theta)\,d\theta$ and the denominator to $\mathrm{E}_\pi L(\theta|\mathbf{x}) = m(\mathbf{x}) > 0$, the marginal (7.2.8). By Slutsky's Theorem (5.5.17) the ratio

converges in probability to

$$\frac{\int_{-\infty}^{t} L(\theta|\mathbf{x})\pi(\theta)\,d\theta}{m(\mathbf{x})} = \int_{-\infty}^{t} \pi(\theta|\mathbf{x})\,d\theta,$$

the last equality being the definition (7.2.7) of the posterior; and since the ratio is a cdf, hence bounded by 1, taking expectations gives $P(\theta^* \le t) \to \int_{-\infty}^{t} \pi(\theta|\mathbf{x})\,d\theta$ as well.
(b) Given $\theta_1, \dots, \theta_m$ the θ_j^* are iid with mean $\sum_i q_i h(\theta_i)$, so by the WLLN in r,

$$\frac{1}{r}\sum_{j=1}^{r} h(\theta_j^*) \longrightarrow \sum_{i=1}^{m} q_i h(\theta_i) = \frac{\frac{1}{m}\sum_{i=1}^{m} h(\theta_i)L(\theta_i|\mathbf{x})}{\frac{1}{m}\sum_{i=1}^{m} L(\theta_i|\mathbf{x})},$$

and the right side converges, exactly as in (a) with h in place of $I(\cdot \le t)$, to

$$\frac{\int h(\theta)L(\theta|\mathbf{x})\pi(\theta)\,d\theta}{m(\mathbf{x})} = \mathrm{E}[h(\theta)|\mathbf{x}]$$

(assuming $\mathrm{E}_\pi|h(\theta)|L(\theta|\mathbf{x}) < \infty$, which is $\mathrm{E}[|h(\theta)|\,|\,\mathbf{x}] < \infty$).
(c) The displayed quantity is the conditional mean: since $P(\theta_j^* = \theta_i|\theta_1, \dots, \theta_m) = q_i$,

$$\mathrm{E}[h(\theta_j^*)\,|\,\theta_1, \dots, \theta_m] = \sum_{i=1}^{m} q_i h(\theta_i) = \frac{\sum_{i=1}^{m} h(\theta_i)L(\theta_i|\mathbf{x})}{\sum_{i=1}^{m} L(\theta_i|\mathbf{x})} =: \delta,$$

which is free of j. Write $T = \frac{1}{r}\sum_{j=1}^{r} h(\theta_j^*)$ for the estimator of (b); then $\mathrm{E}[T|\theta_1, \dots, \theta_m] = \delta$ too, so by the tower property (Theorem 4.4.3)

$$\mathrm{E}\,\delta = \mathrm{E}\big[\mathrm{E}[T|\theta_1, \dots, \theta_m]\big] = \mathrm{E}\,T,$$

the same mean. For the variance, the conditional variance identity (Theorem 4.4.7) gives

$$\begin{aligned}\mathrm{Var}\,T &= \mathrm{Var}\big(\mathrm{E}[T|\theta_1, \dots, \theta_m]\big) + \mathrm{E}\big[\mathrm{Var}(T|\theta_1, \dots, \theta_m)\big] \\ &= \mathrm{Var}\,\delta + \frac{1}{r}\,\mathrm{E}\big[\mathrm{Var}(h(\theta_1^*)|\theta_1, \dots, \theta_m)\big],\end{aligned}$$

using conditional independence of the θ_j^*. The second term is nonnegative, so $\mathrm{Var}\,\delta \le \mathrm{Var}\,T$, strictly unless h is constant on $\{\theta_1, \dots, \theta_m\}$ with probability 1.

Problem 7.33. In Example 7.3.5 the MSE of the Bayes estimator, $\hat{p}_\mathrm{B}$, of a success probability was calculated (the estimator was derived in Example 7.2.14). Show that the choice $\alpha = \beta = \sqrt{n/4}$ yields a constant MSE for $\hat{p}_\mathrm{B}$.
For reference: $X_1, \dots, X_n$ are iid Bernoulli(p), $Y = \sum_i X_i$, the prior on p is beta(α, β), $\hat{p}_\mathrm{B} = (Y + \alpha)/(\alpha + \beta + n)$, and Example 7.3.5 gives

$$\mathrm{E}_p(\hat{p}_\mathrm{B} - p)^2 = \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2.$$

Solution. Put the MSE over the common denominator $(\alpha + \beta + n)^2$ and read off the coefficients of p and p^2. Since $np + \alpha - p(\alpha + \beta + n) = \alpha - p(\alpha + \beta)$,

$$\begin{aligned}(\alpha+\beta+n)^2\,\mathrm{E}_p(\hat{p}_\mathrm{B} - p)^2 &= np(1-p) + \big(\alpha - p(\alpha+\beta)\big)^2 \\ &= -p^2\big[n - (\alpha+\beta)^2\big] + p\big[n - 2\alpha(\alpha+\beta)\big] + \alpha^2.\end{aligned}$$

This is free of p if and only if both brackets vanish:

$$(\alpha+\beta)^2 = n \implies \alpha+\beta = \sqrt{n}, \qquad 2\alpha\sqrt{n} = n \implies \alpha = \tfrac{\sqrt{n}}{2} = \sqrt{n/4},$$

and then $\beta = \sqrt{n} - \alpha = \sqrt{n/4}$ as well. With $\alpha = \beta = \sqrt{n}/2$ the surviving constant is $\alpha^2 = n/4$, so

$$\hat{p}_{\mathrm{B}} = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}, \qquad \mathrm{E}_p(\hat{p}_{\mathrm{B}} - p)^2 = \frac{n/4}{(n+\sqrt{n})^2} = \frac{n}{4(n+\sqrt{n})^2},$$

constant in p, as claimed in Example 7.3.5.

Problem 7.34. Let $X_1, \ldots, X_n$ be a random sample from a binomial(n, p). We want to find equivariant point estimators of p using the group described in Example 6.4.1.
(Example 6.4.1, binomial equivariance: with $X \sim$ binomial(n,p), relabelling successes as failures sends X to $Y = n - X \sim$ binomial$(n, 1-p)$, and the group of transformations is $\mathcal{G} = \{g_1, g_2\}$ with $g_1(x) = n - x$ and $g_2(x) = x$; measurement equivariance together with formal invariance requires the estimator T of p to satisfy (6.4.1), $T(x) = 1 - T(n-x)$.)
(a) Find the class of estimators that are equivariant with respect to this group.
(b) Within the class of Bayes estimators of Example 7.2.14, $\hat{p}_{\mathrm{B}} = (y+\alpha)/(\alpha+\beta+n)$ arising from a beta(α, β) prior, find the estimators that are equivariant with respect to this group.
(c) From the equivariant Bayes estimators of part (b), find the one with the smallest MSE.

Solution. (a) The equivariant estimators are exactly those satisfying

$$T(x_1, \ldots, x_n) = 1 - T(n - x_1, \ldots, n - x_n)$$

for all sample points, this being (6.4.1) applied coordinatewise: the transformation $g(x_1, \ldots, x_n) = (n - x_1, \ldots, n - x_n)$ carries the binomial(n,p) model into the binomial$(n, 1-p)$ model, so $\bar{g}(p) = 1 - p$ and $T(g(\mathbf{x})) = \bar{g}(T(\mathbf{x})) = 1 - T(\mathbf{x})$. Equivalently, $T(\mathbf{x}) + T(g(\mathbf{x})) = 1$; any T may be specified freely on half the sample space and is then determined on the other half. Writing $N = n^2$ for the total number of Bernoulli trials, the sufficient statistic is $Y = \sum_i X_i \sim$ binomial(N, p), g sends $y \mapsto N - y$, and the condition reads $T(y) = 1 - T(N - y)$.
(b) Those with $\alpha = \beta$. Working with $Y \sim$ binomial(N,p) as in (a), the Bayes estimator of Example 7.2.14 is $\hat{p}_{\mathrm{B}}(y) = (y+\alpha)/(\alpha+\beta+N)$, and

$$1 - \hat{p}_{\mathrm{B}}(N-y) = 1 - \frac{N - y + \alpha}{\alpha + \beta + N} = \frac{y + \beta}{\alpha + \beta + N},$$

which equals $\hat{p}_{\mathrm{B}}(y)$ for all y if and only if $\alpha = \beta$. (This is the statement that a symmetric beta prior is the invariant one: $p \mapsto 1 - p$ leaves beta(α, α) fixed.)
(c) There is no uniformly best one; the best in the minimax sense is $\alpha = \beta = \sqrt{N}/2$, the constant-MSE estimator of Exercise 7.33. Setting $\alpha = \beta$ in the MSE formula of Example 7.3.5 and writing $u = p(1-p) \in [0, \frac{1}{4}]$,

$$\mathrm{E}_p(\hat{p}_{\mathrm{B}} - p)^2 = \frac{Np(1-p) + \alpha^2(1-2p)^2}{(N+2\alpha)^2} = \frac{\alpha^2 + u\,(N - 4\alpha^2)}{(N+2\alpha)^2},$$

using $(1-2p)^2 = 1 - 4u$. The MSE is thus linear in u with slope $(N - 4\alpha^2)/(N+2\alpha)^2$, so the risk functions of different α always cross and no single α dominates: at $u = 0$ the MSE $\alpha^2/(N+2\alpha)^2$ increases in α, while at $u = \frac{1}{4}$ the MSE $(N/4)/(N+2\alpha)^2$ decreases in α. Maximizing over u,

$$\sup_{0 \le p \le 1} \mathrm{E}_p(\hat{p}_{\mathrm{B}} - p)^2 = \begin{cases} \dfrac{N/4}{(N+2\alpha)^2}, & \alpha \le \frac{\sqrt{N}}{2} \quad \text{(decreasing in } \alpha\text{)}, \\ \dfrac{\alpha^2}{(N+2\alpha)^2}, & \alpha \ge \frac{\sqrt{N}}{2} \quad \text{(increasing in } \alpha\text{)}, \end{cases}$$

so the maximum risk is smallest at $\alpha = \beta = \sqrt{N}/2 = \sqrt{N/4}$, where the slope $N - 4\alpha^2$ vanishes and the MSE is the constant $N/(4(N+\sqrt{N})^2)$ of Exercise 7.33.

Problem 7.35. The Pitman Estimator of Location (see Lehmann and Casella 1998 Section 3.1, or the original paper by Pitman 1939) is given by

$$d_{\mathrm{P}}(\mathbf{X}) = \frac{\int_{-\infty}^{\infty} t \prod_{i=1}^{n} f(x_i - t)\, dt}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} f(x_i - t)\, dt},$$

where we observe a random sample $X_1, \ldots, X_n$ from $f(x-\theta)$. Pitman showed that this estimator is the location-equivariant estimator with smallest mean squared error (that is, it minimizes (7.3.3)). The goals of this exercise are more modest.
(a) Show that $d_{\mathrm{P}}(\mathbf{X})$ is invariant with respect to the location group of Example 7.3.6, that is, the group $\mathcal{G} = \{g_a(\mathbf{x}) : -\infty < a < \infty\}$ with $g_a(x_1, \ldots, x_n) = (x_1 + a, \ldots, x_n + a)$, whose equivariant estimators are characterized by (7.3.2):

$$W(x_1, \ldots, x_n) + a = W(x_1 + a, \ldots, x_n + a).$$

(b) Show that if $f(x-\theta)$ is $\mathrm{n}(\theta, 1)$, then $d_{\mathrm{P}}(\mathbf{X}) = \bar{X}$.
(c) Show that if $f(x-\theta)$ is uniform$(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, then $d_{\mathrm{P}}(\mathbf{X}) = \frac{1}{2}(X_{(1)} + X_{(n)})$.

Solution. (a) Substitute $t = s + a$ in both integrals. (The exercise says "invariant"; what is to be verified is equivariance, condition (7.3.2).) With $\mathbf{x} + a$ denoting $g_a(\mathbf{x})$ and $L(t) = \prod_{i=1}^{n} f(x_i - t)$,

$$\begin{aligned} d_{\mathrm{P}}(\mathbf{x} + a) &= \frac{\int_{-\infty}^{\infty} t \prod_i f(x_i + a - t)\, dt}{\int_{-\infty}^{\infty} \prod_i f(x_i + a - t)\, dt} = \frac{\int_{-\infty}^{\infty} (s + a) L(s)\, ds}{\int_{-\infty}^{\infty} L(s)\, ds} \\ &= \frac{\int_{-\infty}^{\infty} s L(s)\, ds}{\int_{-\infty}^{\infty} L(s)\, ds} + a = d_{\mathrm{P}}(\mathbf{x}) + a, \end{aligned}$$

the translation $t = s + a$ leaving $dt = ds$ and the limits unchanged. This is (7.3.2).
(b) The weight $L(t)$ is, up to a factor free of t, a normal density in t centered at $\bar{x}$: by the usual decomposition $\sum_i (x_i - t)^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - t)^2$,

$$L(t) = (2\pi)^{-n/2} e^{-\frac{1}{2}\sum_i (x_i - \bar{x})^2}\, e^{-\frac{n}{2}(t - \bar{x})^2} \;\propto\; \exp\left\{ -\frac{(t - \bar{x})^2}{2(1/n)} \right\},$$

which is proportional to the $\mathrm{n}(\bar{x}, 1/n)$ density in t. Hence $d_{\mathrm{P}}(\mathbf{x})$ is the mean of that density,

$$d_{\mathrm{P}}(\mathbf{x}) = \bar{x}.$$

(c) Here $f(u) = I(-\frac{1}{2} \le u \le \frac{1}{2})$, so

$$L(t) = \prod_{i=1}^{n} I\big(|x_i - t| \le \tfrac{1}{2}\big) = I\big(x_{(n)} - \tfrac{1}{2} \le t \le x_{(1)} + \tfrac{1}{2}\big),$$

since $|x_i - t| \le \frac{1}{2}$ for every i is the same as $t \ge x_{(n)} - \frac{1}{2}$ and $t \le x_{(1)} + \frac{1}{2}$. The weight is therefore uniform on an interval (nonempty because $x_{(n)} - x_{(1)} \le 1$ for any sample from this density), and d_{P} is its midpoint:

$$d_{\mathrm{P}}(\mathbf{x}) = \frac{1}{2}\left[\big(x_{(n)} - \tfrac{1}{2}\big) + \big(x_{(1)} + \tfrac{1}{2}\big)\right] = \frac{x_{(1)} + x_{(n)}}{2}.$$

7.6 Exercises 7.36–7.42

Problem 7.36. The *Pitman Estimator of Scale* is given by

$$d_P(\mathbf{X}) \;=\; \frac{\int_0^\infty t^{\,n+r-1}\prod_{i=1}^n f(tx_i)\,dt}{\int_0^\infty t^{\,n+2r-1}\prod_{i=1}^n f(tx_i)\,dt},$$

where we observe a random sample $X_1,\dots,X_n$ from $\frac{1}{\sigma}f(x/\sigma)$. Pitman showed that this estimator is the scale-equivariant estimator of σ^r with smallest scaled mean squared error (that is, it minimizes $\mathrm{E}(d-\sigma^r)^2/\sigma^{2r}$).
(a) Show that $d_P(\mathbf{X})$ is equivariant with respect to the scale group, that is, it satisfies

$$d_P(cx_1,\dots,cx_n) \;=\; c^r\,d_P(x_1,\dots,x_n)$$

for any constant $c>0$.
(b) Find the Pitman scale-equivariant estimator for σ^2 if $X_1,\dots,X_n$ are iid $\mathrm{n}(0,\sigma^2)$.
(c) Find the Pitman scale-equivariant estimator for β if $X_1,\dots,X_n$ are iid exponential(β).
(d) Find the Pitman scale-equivariant estimator for θ if $X_1,\dots,X_n$ are iid uniform$(0,\theta)$.

Solution. (a) Substitute $u=ct$ in both integrals. For any exponent m,

$$\int_0^\infty t^{\,m-1}\prod_{i=1}^n f(tcx_i)\,dt = \int_0^\infty (u/c)^{m-1}\prod_{i=1}^n f(ux_i)\,\frac{du}{c}$$
$$= c^{-m}\int_0^\infty u^{\,m-1}\prod_{i=1}^n f(ux_i)\,du.$$

Applying this with $m=n+r$ in the numerator and $m=n+2r$ in the denominator,

$$d_P(c\mathbf{x}) = \frac{c^{-(n+r)}}{c^{-(n+2r)}}\,d_P(\mathbf{x}) = c^r d_P(\mathbf{x}).$$

(b) $\sum X_i^2/(n+2)$. Here $r=2$, $f(x)=(2\pi)^{-1/2}e^{-x^2/2}$, so with $S=\sum_{i=1}^n x_i^2$,

$$\prod_{i=1}^n f(tx_i) = (2\pi)^{-n/2}e^{-t^2S/2}, \qquad t>0.$$

The substitution $v=t^2S/2$ gives, for any $m>-1$,

$$\int_0^\infty t^{\,m}e^{-t^2S/2}\,dt \;=\; \tfrac12\Big(\tfrac{2}{S}\Big)^{(m+1)/2}\Gamma\Big(\tfrac{m+1}{2}\Big).$$

With $m=n+1$ over $m=n+3$ the powers of $2/S$ leave one factor $S/2$ and

$$d_P(\mathbf{X}) = \frac{S}{2}\cdot\frac{\Gamma\big(\frac{n+2}{2}\big)}{\Gamma\big(\frac{n+4}{2}\big)} = \frac{S}{2}\cdot\frac{2}{n+2} = \frac{\sum_{i=1}^n X_i^2}{n+2}.$$

(c) $\sum X_i/(n+1)$. Here $r=1$, $f(x)=e^{-x}I(x>0)$, so $\prod_i f(tx_i)=e^{-tS}$ with $S=\sum x_i$ (all $x_i>0$), and $\int_0^\infty t^{\,m}e^{-tS}dt = m!/S^{m+1}$. Hence

$$d_P(\mathbf{X}) = \frac{n!/S^{\,n+1}}{(n+1)!/S^{\,n+2}} = \frac{S}{n+1} = \frac{\sum_{i=1}^n X_i}{n+1}.$$

(d) $\frac{n+2}{n+1}X_{(n)}$. Here $r=1$ and $f(x)=I(0<x<1)$, so $\prod_i f(tx_i)=1$ exactly when $0<t<1/x_{(n)}$. Thus

$$d_P(\mathbf{X}) = \frac{\int_0^{1/x_{(n)}} t^{\,n}\,dt}{\int_0^{1/x_{(n)}} t^{\,n+1}\,dt} = \frac{x_{(n)}^{-(n+1)}/(n+1)}{x_{(n)}^{-(n+2)}/(n+2)} = \frac{n+2}{n+1}\,X_{(n)}.$$

Problem 7.37. Let $X_1, \ldots, X_n$ be a random sample from a population with pdf

$$f(x|\theta) = \frac{1}{2\theta}, \qquad -\theta < x < \theta, \quad \theta > 0.$$

Find, if one exists, a best unbiased estimator of θ.

Solution. $W = \dfrac{n+1}{n} \max\limits_{1\le i\le n} |X_i|$, with $\mathrm{Var}_\theta W = \dfrac{\theta^2}{n(n+2)}$.
Put $T = \max_i |X_i|$. Since

$$f(\mathbf{x}|\theta) = (2\theta)^{-n} \prod_{i=1}^{n} I(|x_i| < \theta) = (2\theta)^{-n} I(T(\mathbf{x}) < \theta),$$

T is sufficient by the Factorization Theorem 6.2.6 (take $h(\mathbf{x}) \equiv 1$). Each $|X_i|$ is uniform$(0, \theta)$, so T has pdf nt^{n-1}/θ^n on $0 < t < \theta$, and T is complete: $\mathrm{E}_\theta g(T) = 0$ for all $\theta > 0$ forces

$$\int_0^\theta g(t)\, nt^{n-1}\, dt = 0 \quad \forall\, \theta \implies g(\theta)\, n\theta^{n-1} = 0 \quad \forall\, \theta,$$

on differentiating in θ, so $P_\theta(g(T) = 0) = 1$ for every θ.
Now

$$\mathrm{E}_\theta T = \int_0^\theta t\, \frac{nt^{n-1}}{\theta^n}\, dt = \frac{n}{n+1}\theta, \qquad \mathrm{E}_\theta T^2 = \frac{n}{n+2}\theta^2,$$

so $W = \frac{n+1}{n} T$ is an unbiased estimator of θ that is a function of the complete sufficient statistic alone; by Theorem 7.3.23 it is the unique best unbiased estimator of θ. Its variance is

$$\mathrm{Var}_\theta W = \frac{(n+1)^2}{n^2} \left[\frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} \right] = \frac{\theta^2}{n(n+2)}.$$

Problem 7.38. For each of the following distributions, let $X_1, \ldots, X_n$ be a random sample. Is there a function of θ, say $g(\theta)$, for which there exists an unbiased estimator whose variance attains the Cramer–Rao Lower Bound? If so, find it. If not, show why not.
(a) $f(x|\theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad \theta > 0$
(b) $f(x|\theta) = \dfrac{\log(\theta)}{\theta - 1}\, \theta^x, \quad 0 < x < 1, \quad \theta > 1$

Solution. In both cases the answer is yes, and Corollary 7.3.15 produces the pair (g, W) at once: writing $\frac{\partial}{\partial\theta} \log L(\theta|\mathbf{x}) = a(\theta)\,[W(\mathbf{x}) - g(\theta)]$ exhibits W as an unbiased estimator of $g(\theta)$ attaining the bound. (Both are one-parameter exponential families with support free of θ, so the differentiation-under-the-integral hypothesis (7.3.4) of Theorem 7.3.9 holds.)
(a) $g(\theta) = 1/\theta$, attained by $W = -\frac{1}{n}\sum_{i=1}^n \log X_i$. Indeed $\log L = n\log\theta + (\theta - 1)\sum \log x_i$, so

$$\frac{\partial}{\partial\theta} \log L = \frac{n}{\theta} + \sum_{i=1}^{n} \log x_i = -n \left[\left(-\frac{1}{n} \sum_{i=1}^{n} \log x_i \right) - \frac{1}{\theta} \right],$$

which is (7.3.12) with $a(\theta) = -n$. Consistently, $-\log X_i$ is exponential$(1/\theta)$ (since $P(-\log X > y) = P(X < e^{-y}) = e^{-\theta y}$), so $EW = 1/\theta$ and

$$\mathrm{Var}_\theta W = \frac{1}{n\theta^2},$$

which is exactly the bound: $\frac{\partial^2}{\partial\theta^2} \log f = -1/\theta^2$, so by Lemma 7.3.11 the information is $1/\theta^2$ and, with $g'(\theta) = -1/\theta^2$,

$$\frac{[g'(\theta)]^2}{n\, \mathrm{E}_\theta\big[(\frac{\partial}{\partial\theta} \log f)^2\big]} = \frac{\theta^{-4}}{n\theta^{-2}} = \frac{1}{n\theta^2}.$$

(b) $g(\theta) = \dfrac{\theta}{\theta-1} - \dfrac{1}{\log\theta}$, attained by $W = \bar{X}$. Here $\log f(x|\theta) = \log\log\theta - \log(\theta-1) + x\log\theta$, so

$$\frac{\partial}{\partial\theta}\log L = \frac{n}{\theta\log\theta} - \frac{n}{\theta-1} + \frac{1}{\theta}\sum_{i=1}^{n} x_i = \frac{n}{\theta}\left[\bar{x} - \left(\frac{\theta}{\theta-1} - \frac{1}{\log\theta}\right)\right],$$

which is (7.3.12) with $a(\theta) = n/\theta$. That $\mathrm{E}_\theta \bar{X} = g(\theta)$ is the direct computation: with $c = \log\theta$,

$$\mathrm{E}_\theta X = \frac{c}{\theta-1}\int_0^1 x e^{cx}\,dx = \frac{c}{\theta-1}\cdot\frac{\theta(c-1)+1}{c^2} = \frac{\theta}{\theta-1} - \frac{1}{c}.$$

The attained variance is $\mathrm{Var}_\theta \bar{X} = \sigma^2(\theta)/n$ with

$$\sigma^2(\theta) = \frac{\theta(c^2-2c+2)-2}{(\theta-1)c^2} - \left(\frac{\theta}{\theta-1} - \frac{1}{c}\right)^2, \qquad c = \log\theta.$$

Problem 7.39. Prove Lemma 7.3.11: if $f(x|\theta)$ satisfies

$$\frac{d}{d\theta}\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right) = \int \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right) f(x|\theta)\right] dx$$

(true for an exponential family), then

$$\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right) = -\,\mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right).$$

Solution. The left-hand side of the hypothesis is zero, because $\mathrm{E}_\theta\big(\frac{\partial}{\partial\theta}\log f(X|\theta)\big) = 0$ identically in θ by (7.3.8). So it suffices to expand the right-hand integrand. Using the product rule and then $\frac{\partial}{\partial\theta} f(x|\theta) = f(x|\theta)\frac{\partial}{\partial\theta}\log f(x|\theta)$,

$$\begin{aligned}\frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f\right) f\right] &= \left(\frac{\partial^2}{\partial\theta^2}\log f\right) f + \left(\frac{\partial}{\partial\theta}\log f\right)\frac{\partial f}{\partial\theta} \\ &= \left[\frac{\partial^2}{\partial\theta^2}\log f + \left(\frac{\partial}{\partial\theta}\log f\right)^2\right] f.\end{aligned}$$

Integrating over x and invoking the hypothesis,

$$0 = \mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right) + \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right),$$

which is the assertion.

Problem 7.40. Let $X_1, \ldots, X_n$ be iid Bernoulli(p). Show that the variance of $\bar{X}$ attains the Cramer–Rao Lower Bound, and hence $\bar{X}$ is the best unbiased estimator of p.

Solution. $\mathrm{Var}_p \bar{X} = p(1-p)/n$, and this is the bound. With $f(x|p) = p^x(1-p)^{1-x}$ on $x \in \{0,1\}$,

$$\frac{\partial}{\partial p}\log f(x|p) = \frac{x}{p} - \frac{1-x}{1-p} = \frac{x-p}{p(1-p)},$$

so the information number is

$$\mathrm{E}_p\left(\left(\frac{\partial}{\partial p}\log f(X|p)\right)^2\right) = \frac{\mathrm{E}_p(X-p)^2}{p^2(1-p)^2} = \frac{p(1-p)}{p^2(1-p)^2} = \frac{1}{p(1-p)}.$$

The Bernoulli family is an exponential family with support $\{0,1\}$ free of p, so the interchange (7.3.4) is legitimate and Corollary 7.3.10 applies: for $\tau(p) = p$, $\tau'(p) = 1$, any unbiased estimator W of p satisfies

$$\operatorname{Var}_p W \;\geq\; \frac{1}{n/\big(p(1-p)\big)} \;=\; \frac{p(1-p)}{n}.$$

Since $\mathrm{E}_p\bar{X} = p$ and $\operatorname{Var}_p\bar{X} = \frac{1}{n^2}\cdot n\,p(1-p) = p(1-p)/n$, the bound is attained and $\bar{X}$ is the best unbiased estimator of p.
Method (2): by the attainment Corollary 7.3.15,

$$\frac{\partial}{\partial p}\log L(p|\mathbf{x}) = \frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} = \frac{n}{p(1-p)}\,(\bar{x}-p),$$

which is (7.3.12) with $a(p) = n/[p(1-p)]$ and $\tau(p) = p$.

Problem 7.41. Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2.
(a) Show that the estimator $\sum_{i=1}^n a_i X_i$ is an unbiased estimator of μ if $\sum_{i=1}^n a_i = 1$.
(b) Among all unbiased estimators of this form (called *linear unbiased estimators*) find the one with minimum variance, and calculate the variance.

Solution. (a) $\mathrm{E}\Big(\sum_{i=1}^n a_i X_i\Big) = \sum_{i=1}^n a_i\,\mathrm{E}X_i = \mu\sum_{i=1}^n a_i = \mu$ whenever $\sum_i a_i = 1$.
(b) $\bar{X}$, with variance σ^2/n. By independence,

$$\operatorname{Var}\Big(\sum_{i=1}^n a_i X_i\Big) = \sigma^2\sum_{i=1}^n a_i^2,$$

so the problem is to minimize $\sum a_i^2$ subject to $\sum a_i = 1$. Apply the Cauchy–Schwarz Inequality (Theorem 4.7.3) to $X = a_I$ and $Y \equiv 1$, where I is uniform on $\{1,\ldots,n\}$ (the hypotheses are vacuous: any two random variables are allowed):

$$1 = \Big(\sum_{i=1}^n a_i\Big)^2 \leq n\sum_{i=1}^n a_i^2, \qquad \text{so} \qquad \sum_{i=1}^n a_i^2 \geq \frac{1}{n},$$

with equality if and only if the a_i are all equal, i.e. $a_i = 1/n$. Hence the minimum-variance linear unbiased estimator is $\bar{X}$, and $\operatorname{Var}\bar{X} = \sigma^2/n$.

Problem 7.42. Let $W_1, \ldots, W_k$ be unbiased estimators of a parameter θ with $\operatorname{Var} W_i = \sigma_i^2$ and $\operatorname{Cov}(W_i, W_j) = 0$ if $i \neq j$.
(a) Show that, of all estimators of the form $\sum_i a_i W_i$, where the a_i are constants and $\mathrm{E}_\theta\big(\sum_i a_i W_i\big) = \theta$, the estimator

$$W^* = \frac{\sum_i W_i/\sigma_i^2}{\sum_i (1/\sigma_i^2)}$$

has minimum variance.
(b) Show that $\operatorname{Var} W^* = \dfrac{1}{\sum_i (1/\sigma_i^2)}$.

Solution. (a) Unbiasedness forces $\theta = \sum_i a_i\theta$ for all θ, i.e. $\sum_{i=1}^k a_i = 1$; and by uncorrelatedness $\operatorname{Var}\big(\sum_i a_i W_i\big) = \sum_i a_i^2\sigma_i^2$. Apply the Cauchy–Schwarz Inequality (Theorem 4.7.3, with I uniform on

$\{1,\dots,k\}$) to $X = a_I\sigma_I$ and $Y = 1/\sigma_I$:

$$1 = \Big(\sum_{i=1}^{k} a_i\Big)^2 = \Big(\sum_{i=1}^{k}(a_i\sigma_i)\frac{1}{\sigma_i}\Big)^2 \le \Big(\sum_{i=1}^{k} a_i^2\sigma_i^2\Big)\Big(\sum_{i=1}^{k}\frac{1}{\sigma_i^2}\Big),$$

so for every admissible choice of the a_i,

$$\operatorname{Var}\Big(\sum_{i=1}^{k} a_i W_i\Big) = \sum_{i=1}^{k} a_i^2\sigma_i^2 \;\ge\; \frac{1}{\sum_{i=1}^{k}(1/\sigma_i^2)}.$$

Equality holds iff $a_i\sigma_i \propto 1/\sigma_i$, i.e. $a_i \propto 1/\sigma_i^2$; the constraint $\sum_i a_i = 1$ then pins down

$$a_i^* = \frac{1/\sigma_i^2}{\sum_{j=1}^{k}(1/\sigma_j^2)},$$

which is exactly $W^* = \sum_i a_i^* W_i$.
(b)

$$\operatorname{Var} W^* = \sum_{i=1}^{k}(a_i^*)^2\sigma_i^2 = \frac{\sum_{i=1}^{k}\sigma_i^{-4}\sigma_i^2}{\big(\sum_j \sigma_j^{-2}\big)^2} = \frac{\sum_i(1/\sigma_i^2)}{\big(\sum_j(1/\sigma_j^2)\big)^2} = \frac{1}{\sum_{i=1}^{k}(1/\sigma_i^2)},$$

which meets the lower bound of part (a).

7.7 Exercises 7.43–7.49

Problem 7.43. Exercise 7.42 established that the optimal weights are $q_i^* = (1/\sigma_i^2)/(\sum_j 1/\sigma_j^2)$. A result due to Tukey (see Bloch and Moses 1988) states that if $W = \sum_i q_i W_i$ is an estimator based on another set of weights $q_i \ge 0$, $\sum_i q_i = 1$, then

$$\frac{\operatorname{Var} W}{\operatorname{Var} W^*} \le \frac{1}{1-\lambda^2},$$

where λ satisfies $(1+\lambda)/(1-\lambda) = b_{\max}/b_{\min}$, and $b_{\max}$ and $b_{\min}$ are the largest and smallest of $b_i = q_i/q_i^*$.
(a) Prove Tukey's inequality.
(b) Use the inequality to assess the performance of the usual mean $\sum_i W_i/k$ as a function of $\sigma^2_{\max}/\sigma^2_{\min}$.
(Here $W_1,\dots,W_k$ are uncorrelated unbiased estimators of θ with $\operatorname{Var} W_i = \sigma_i^2$, as in Exercise 7.42.)

Solution. (a) The ratio is the second moment of b under the weights q_i^*, and a variable on $[b_{\min}, b_{\max}]$ with mean 1 has second moment at most $1/(1-\lambda^2)$.
Write $S = \sum_j 1/\sigma_j^2$, so $q_i^*\sigma_i^2 = 1/S = \operatorname{Var} W^*$ (Exercise 7.42(b)). With $q_i = b_i q_i^*$,

$$\begin{aligned}\operatorname{Var} W &= \sum_i q_i^2\sigma_i^2 = \sum_i b_i^2\, q_i^*\big(q_i^*\sigma_i^2\big)\\ &= \operatorname{Var} W^* \sum_i q_i^* b_i^2,\end{aligned}$$

while $\sum_i q_i = 1$ says $\sum_i q_i^* b_i = 1$. So if b denotes a random variable taking the value b_i with probability q_i^* (the q_i^* are positive and sum to 1), then

$$\frac{\operatorname{Var} W}{\operatorname{Var} W^*} = \mathrm{E}\, b^2, \qquad \mathrm{E}\, b = 1.$$

Put $m = b_{\min}$, $M = b_{\max}$. Since $(b-m)(M-b) \ge 0$ pointwise,

$$\mathrm{E}\, b^2 \le (m+M)\,\mathrm{E}\, b - mM = m + M - mM.$$

Finally $M/m = (1+\lambda)/(1-\lambda)$ fixes only the ratio, so maximize $g(m) = m + M - mM$ over $m > 0$ along $M = m(1+\lambda)/(1-\lambda)$:

$$g(m) = \frac{2m}{1-\lambda} - \frac{(1+\lambda)m^2}{1-\lambda}, \qquad g'(m) = 0 \iff m = \frac{1}{1+\lambda},$$

and at that point $M = 1/(1-\lambda)$ and

$$g\Big(\tfrac{1}{1+\lambda}\Big) = \frac{2}{1-\lambda^2} - \frac{1}{1-\lambda^2} = \frac{1}{1-\lambda^2}.$$

Hence $\operatorname{Var} W / \operatorname{Var} W^* \le m + M - mM \le 1/(1-\lambda^2)$. (Equivalently $\lambda = (b_{\max} - b_{\min})/(b_{\max} + b_{\min})$.)
(b) For $\overline{W} = \sum_i W_i/k$ we have $q_i = 1/k$, so

$$b_i = \frac{q_i}{q_i^*} = \frac{\sigma_i^2 S}{k}, \qquad \frac{b_{\max}}{b_{\min}} = \frac{\sigma^2_{\max}}{\sigma^2_{\min}} =: r.$$

Then $\lambda = (r-1)/(r+1)$, $1 - \lambda^2 = 4r/(r+1)^2$, and Tukey's inequality reads

$$\frac{\operatorname{Var} \overline{W}}{\operatorname{Var} W^*} \le \frac{(1+r)^2}{4r}.$$

So the simple mean's variance exceeds the optimum by at most the factor

$r = \sigma^2_{\max}/\sigma^2_{\min}$	1	2	4	10
$(1+r)^2/(4r)$	1.000	1.125	1.563	3.025

a two-fold spread in the variances costing it at most 12.5% and a four-fold spread at most 56%.

Problem 7.44. Let $X_1, \dots, X_n$ be iid $\mathrm{n}(\theta, 1)$. Show that the best unbiased estimator of θ^2 is $\bar{X}^2 - (1/n)$. Calculate its variance (use Stein's Identity from Section 3.6), and show that it is greater than the Cramer–Rao Lower Bound.

Solution. $\bar{X}$ is a complete sufficient statistic (Theorem 6.2.25 applied to this one-parameter exponential family), and $\bar{X} \sim \mathrm{n}(\theta, 1/n)$ gives $\mathrm{E}\,\bar{X}^2 = \theta^2 + 1/n$, so $\bar{X}^2 - 1/n$ is an unbiased function of it; by Lehmann–Scheffe (Theorem 7.3.23) it is the best unbiased estimator of θ^2.
For the variance write $Y = \bar{X} \sim \mathrm{n}(\theta, 1/n)$ and apply Stein's Lemma (Lemma 3.6.5), $\mathrm{E}[g(Y)(Y-\theta)] = \frac{1}{n}\,\mathrm{E}\,g'(Y)$, which is legitimate for the polynomials $g(y) = y^2, y^3$ (all moments finite). Taking $g(y) = y^2$,

$$\mathrm{E}\,Y^3 = \theta\,\mathrm{E}\,Y^2 + \tfrac{2}{n}\,\mathrm{E}\,Y = \theta\Big(\theta^2 + \tfrac{1}{n}\Big) + \tfrac{2\theta}{n} = \theta^3 + \tfrac{3\theta}{n},$$

and then $g(y) = y^3$,

$$\begin{aligned}\mathrm{E}\,Y^4 &= \theta\,\mathrm{E}\,Y^3 + \tfrac{3}{n}\,\mathrm{E}\,Y^2 \\ &= \theta^4 + \tfrac{3\theta^2}{n} + \tfrac{3}{n}\Big(\theta^2 + \tfrac{1}{n}\Big) = \theta^4 + \tfrac{6\theta^2}{n} + \tfrac{3}{n^2}.\end{aligned}$$

Hence

$$\begin{aligned}\operatorname{Var}\Big(\bar{X}^2 - \tfrac{1}{n}\Big) &= \mathrm{E}\,Y^4 - \big(\mathrm{E}\,Y^2\big)^2 \\ &= \theta^4 + \tfrac{6\theta^2}{n} + \tfrac{3}{n^2} - \Big(\theta^2 + \tfrac{1}{n}\Big)^2 = \frac{4\theta^2}{n} + \frac{2}{n^2}.\end{aligned}$$

The information in one observation is $\mathrm{E}(X-\theta)^2 = 1$, so for $\tau(\theta) = \theta^2$ the iid Cramer–Rao bound of Corollary 7.3.10 (the normal pdf satisfies the conditions (7.3.4) of Theorem 7.3.9, as used in Example 7.3.14) is

$$\frac{[\tau'(\theta)]^2}{nI(\theta)} = \frac{4\theta^2}{n} < \frac{4\theta^2}{n} + \frac{2}{n^2}.$$

Problem 7.45. Let $X_1, X_2, \ldots, X_n$ be iid from a distribution with mean μ and variance σ^2, and let S^2 be the usual unbiased estimator of σ^2. In Example 7.3.4 we saw that, under normality, the MLE has smaller MSE than S^2. In this exercise we will explore variance estimates some more.
(a) Show that, for any estimator of the form aS^2, where a is a constant,

$$\mathrm{MSE}(aS^2) = \mathrm{E}[aS^2 - \sigma^2]^2 = a^2 \operatorname{Var}(S^2) + (a-1)^2\sigma^4.$$

(b) Show that

$$\operatorname{Var}(S^2) = \frac{1}{n}\left(\kappa - \frac{n-3}{n-1}\right)\sigma^4,$$

where $\kappa = \mathrm{E}[X-\mu]^4/\sigma^4$ is the **kurtosis**. (You may have already done this in Exercise 5.8(b).)
(c) Show that, under normality, the kurtosis is 3 and establish that, in this case, the estimator of the form aS^2 with the minimum MSE is $\frac{n-1}{n+1}S^2$. (Lemma 3.6.5 may be helpful.)
(d) If normality is not assumed, show that $\mathrm{MSE}(aS^2)$ is minimized at

$$a = \frac{n-1}{(n+1) + \frac{(\kappa-3)(n-1)}{n}},$$

which is useless as it depends on a parameter.
(e) Show that
(i) for distributions with $\kappa > 3$, the optimal a will satisfy $a < \frac{n-1}{n+1}$;
(ii) for distributions with $\kappa < 3$, the optimal a will satisfy $\frac{n-1}{n+1} < a < 1$.
See Searls and Intarapanich (1990) for more details.

Solution. (a) $\mathrm{E}(aS^2) = a\sigma^2$, so the bias is $(a-1)\sigma^2$ and the decomposition (7.3.1) gives

$$\mathrm{MSE}(aS^2) = \operatorname{Var}(aS^2) + \big[(a-1)\sigma^2\big]^2 = a^2 \operatorname{Var}(S^2) + (a-1)^2\sigma^4.$$

(b) Take $\mu = 0$ (both sides are location free) and put $T = \sum_i X_i^2$, $U = n\bar{X}^2$, so $(n-1)S^2 = T - U$. All third-moment terms in $\mathrm{E}[TU]$ carry a lone factor $\mathrm{E}\,X = 0$ and drop out, so with $\mathrm{E}\,X^2 = \sigma^2$, $\mathrm{E}\,X^4 = \kappa\sigma^4$,

$$\begin{aligned}
\mathrm{E}\,T^2 &= n\kappa\sigma^4 + n(n-1)\sigma^4, \\
\mathrm{E}[TU] &= \tfrac{1}{n}\big[n\kappa\sigma^4 + n(n-1)\sigma^4\big] = \kappa\sigma^4 + (n-1)\sigma^4, \\
\mathrm{E}\,U^2 &= \tfrac{1}{n^2}\,\mathrm{E}\Big(\sum_i X_i\Big)^4 = \tfrac{1}{n}\big[\kappa + 3(n-1)\big]\sigma^4.
\end{aligned}$$

(Check! each expansion keeps only the terms with every index repeated an even number of times.) Therefore

$$\begin{aligned}
(n-1)^2\,\mathrm{E}\,S^4 &= \mathrm{E}\,T^2 - 2\,\mathrm{E}[TU] + \mathrm{E}\,U^2 \\
&= \sigma^4\left[\frac{(n-1)^2}{n}\kappa + \frac{(n-1)(n^2-2n+3)}{n}\right],
\end{aligned}$$

so that

$$\begin{aligned}
\operatorname{Var}(S^2) = \mathrm{E}\,S^4 - \sigma^4 &= \sigma^4\left[\frac{\kappa}{n} + \frac{n^2-2n+3}{n(n-1)} - 1\right] \\
&= \sigma^4\left[\frac{\kappa}{n} + \frac{3-n}{n(n-1)}\right] = \frac{\sigma^4}{n}\left(\kappa - \frac{n-3}{n-1}\right).
\end{aligned}$$

(c) For $X \sim \mathrm{n}(\mu, \sigma^2)$, Stein's Lemma (Lemma 3.6.5) with $g(x) = (x-\mu)^3$ gives

$$\mathrm{E}(X-\mu)^4 = \sigma^2\,\mathrm{E}\big[3(X-\mu)^2\big] = 3\sigma^4,$$

i.e. $\kappa = 3$. Then (b) becomes

$$\operatorname{Var}(S^2) = \frac{\sigma^4}{n}\cdot\frac{3(n-1)-(n-3)}{n-1} = \frac{2\sigma^4}{n-1},$$

and by (a)

$$\mathrm{MSE}(aS^2) = \sigma^4 \left[\frac{2a^2}{n-1} + (a-1)^2 \right].$$

This is a strictly convex quadratic in a, so its unique minimizer solves half the derivative equation:

$$\frac{2a}{n-1} + (a-1) = 0 \iff a = \frac{n-1}{n+1} < 1.$$

(d) With $\mathrm{Var}(S^2) = c\,\sigma^4$, $c = \frac{1}{n}\big(\kappa - \frac{n-3}{n-1}\big)$, part (a) gives $\mathrm{MSE}(aS^2)/\sigma^4 = a^2 c + (a-1)^2$, minimized at

$$a = \frac{1}{1+c} = \frac{n}{n + \kappa - \frac{n-3}{n-1}}.$$

Writing $\kappa = 3 + (\kappa - 3)$ in the denominator,

$$n + 3 - \frac{n-3}{n-1} + (\kappa - 3) = \frac{n(n+1)}{n-1} + (\kappa - 3),$$

so dividing numerator and denominator by $n/(n-1)$,

$$a = \frac{n-1}{(n+1) + \frac{(\kappa-3)(n-1)}{n}}.$$

It depends on κ, an unknown feature of the sampling distribution.
(e) Both claims read off the denominator $D = (n+1) + (\kappa-3)(n-1)/n$, which is positive in each case.
(i) $\kappa > 3 \Rightarrow D > n+1 \Rightarrow a = (n-1)/D < (n-1)/(n+1)$.
(ii) $\kappa < 3 \Rightarrow D < n+1 \Rightarrow a > (n-1)/(n+1)$. For the upper bound, $\kappa = \mathrm{E}(X-\mu)^4/\sigma^4 \geq \big[\mathrm{E}(X-\mu)^2\big]^2/\sigma^4 = 1$ by Jensen's inequality (Theorem 4.7.7), so

$$(3-\kappa)\frac{n-1}{n} \leq 2 \cdot \frac{n-1}{n} < 2 \quad \Longrightarrow \quad D > n+1-2 = n-1,$$

whence $a = (n-1)/D < 1$.

Problem 7.46. Let X_1, X_2, and X_3 be a random sample of size three from a uniform$(\theta, 2\theta)$ distribution, where $\theta > 0$.
(a) Find the method of moments estimator of θ.
(b) Find the MLE, $\hat{\theta}$, and find a constant k such that $\mathrm{E}_\theta(k\hat{\theta}) = \theta$.
(c) Which of the two estimators can be improved by using sufficiency? How?
(d) Find the method of moments estimate and the MLE of θ based on the data

$$1.29,\ .86,\ 1.33,$$

three observations of average berry sizes (in centimeters) of wine grapes.

Solution. (a) $\tilde{\theta} = \frac{2}{3}\bar{X}$, since $\mathrm{E}_\theta\, X = \frac{\theta+2\theta}{2} = \frac{3}{2}\theta$, and solving $\frac{3}{2}\theta = \bar{X}$.
(b) $\hat{\theta} = X_{(3)}/2$ and $k = 8/7$. The likelihood is

$$L(\theta \mid \mathbf{x}) = \theta^{-3}\, I\big(\theta \leq x_{(1)}\big)\, I\big(x_{(3)} \leq 2\theta\big) = \theta^{-3}\, I\Big(\tfrac{x_{(3)}}{2} \leq \theta \leq x_{(1)}\Big),$$

a nonempty interval with probability 1, and θ^{-3} is decreasing, so L is maximized at the left endpoint $\theta = x_{(3)}/2$. Writing $X_i = \theta(1+U_i)$ with U_i iid uniform$(0,1)$ gives $X_{(3)} = \theta(1+U_{(3)})$ and $\mathrm{E}\,U_{(3)} = 3/4$, so

$$\mathrm{E}_\theta\,\hat{\theta} = \tfrac{1}{2}\theta\big(1 + \tfrac{3}{4}\big) = \tfrac{7}{8}\theta, \qquad k = \tfrac{8}{7}.$$

(c) The method of moments estimator. By the factorization in (b), $T = (X_{(1)}, X_{(3)})$ is sufficient, and $\hat{\theta}$ is already a function of T, so Rao–Blackwellization leaves it unchanged; but $\bar{X}$ is not a function of

T, so Theorem 7.3.17 improves it, strictly because $\operatorname{Var}(\bar X \mid T) > 0$ with positive probability. For a uniform parent, $X_{(2)}$ given $X_{(1)} = a$, $X_{(3)} = b$ is uniform(a, b), hence

$$\mathrm{E}\big[\bar X \mid T\big] = \frac{X_{(1)} + \frac{X_{(1)}+X_{(3)}}{2} + X_{(3)}}{3} = \frac{X_{(1)} + X_{(3)}}{2},$$

and the improved version of $\tilde\theta = \frac{2}{3}\bar X$ is

$$\mathrm{E}\big[\tilde\theta \mid T\big] = \frac{X_{(1)} + X_{(3)}}{3},$$

still unbiased because $\mathrm{E}\, X_{(1)} = \frac{5}{4}\theta$ and $\mathrm{E}\, X_{(3)} = \frac{7}{4}\theta$ sum to 3θ.
(d) Here $\bar x = 3.48/3 = 1.16$ and $x_{(3)} = 1.33$, so

$$\tilde\theta = \tfrac{2}{3}(1.16) = 0.773, \qquad \hat\theta = \tfrac{1.33}{2} = 0.665.$$

Problem 7.47. Suppose that when the radius of a circle is measured, an error is made that has a $\mathrm{n}(0, \sigma^2)$ distribution. If n independent measurements are made, find an unbiased estimator of the area of the circle. Is it best unbiased?

Solution. $\pi\big(\bar X^2 - \sigma^2/n\big)$, and yes, it is best unbiased.
The measurements are $X_i = r + \varepsilon_i$, that is, $X_1, \dots, X_n$ iid $\mathrm{n}(r, \sigma^2)$, and the target is $\tau(r) = \pi r^2$. Since $\bar X \sim \mathrm{n}(r, \sigma^2/n)$,

$$\mathrm{E}_r\left[\pi\Big(\bar X^2 - \frac{\sigma^2}{n}\Big)\right] = \pi\Big(r^2 + \frac{\sigma^2}{n} - \frac{\sigma^2}{n}\Big) = \pi r^2.$$

With σ^2 known this is a function of $\bar X$, which is complete and sufficient for r by Theorem 6.2.25 (the natural parameter space $\{r/\sigma^2 : r > 0\}$ contains an open interval), so Lehmann–Scheffe (Theorem 7.3.23) makes it the unique best unbiased estimator.
(If σ^2 is unknown, replace it by S^2: $\mathrm{E}\, S^2 = \sigma^2$ gives unbiasedness of $\pi(\bar X^2 - S^2/n)$, and $(\bar X, S^2)$ is complete sufficient for (r, σ^2) by Example 6.2.9 and Theorem 6.2.25.)

Problem 7.48. Suppose that X_i, $i = 1, \dots, n$, are iid Bernoulli(p).
(a) Show that the variance of the MLE of p attains the Cramer–Rao Lower Bound.
(b) For $n \ge 4$, show that the product $X_1X_2X_3X_4$ is an unbiased estimator of p^4, and use this fact to find the best unbiased estimator of p^4.

Solution. (a) The MLE is $\hat p = \bar X$ (Example 7.2.7), with $\operatorname{Var}\bar X = p(1-p)/n$. Since

$$\frac{\partial}{\partial p}\log f(x \mid p) = \frac{x}{p} - \frac{1-x}{1-p} = \frac{x-p}{p(1-p)},$$

the information in one observation is

$$I(p) = \mathrm{E}\left[\frac{(X-p)^2}{p^2(1-p)^2}\right] = \frac{1}{p(1-p)},$$

so the iid Cramer–Rao bound of Corollary 7.3.10 (the interchange (7.3.4) is a finite sum over $\{0, 1\}$, hence automatic) is

$$\frac{1}{nI(p)} = \frac{p(1-p)}{n} = \operatorname{Var}\bar X.$$

Method (2): $\frac{\partial}{\partial p}\log L(p \mid \mathbf{x}) = \frac{n}{p(1-p)}(\bar x - p)$ is of the form $a(p)[\bar x - p]$, so Corollary 7.3.15 gives attainment directly.
(b) $\mathrm{E}(X_1X_2X_3X_4) = \prod_{i=1}^4 \mathrm{E}\, X_i = p^4$ by independence, since each X_i is 0 or 1 with $\mathrm{E}\, X_i = p$. Now $T = \sum_{i=1}^n X_i \sim$ binomial(n, p) is complete sufficient (exponential family, Theorem 6.2.25), so by Rao–Blackwell and Lehmann–Scheffe (Theorems 7.3.17 and 7.3.23) the best unbiased estimator of p^4 is

$\mathrm{E}[X_1X_2X_3X_4 \mid T]$. For $t \geq 4$,

$$P(X_1 = \cdots = X_4 = 1, T = t) = p^4 \binom{n-4}{t-4} p^{t-4}(1-p)^{n-t}$$
$$= \binom{n-4}{t-4} p^t (1-p)^{n-t},$$

so, dividing by $P(T = t) = \binom{n}{t} p^t (1-p)^{n-t}$,

$$\mathrm{E}\big[X_1X_2X_3X_4 \mid T = t\big] = \frac{\binom{n-4}{t-4}}{\binom{n}{t}} = \frac{t(t-1)(t-2)(t-3)}{n(n-1)(n-2)(n-3)},$$

a formula that also gives the correct value 0 for $t < 4$. Hence the best unbiased estimator of p^4 is

$$\frac{T(T-1)(T-2)(T-3)}{n(n-1)(n-2)(n-3)}, \qquad T = \sum_{i=1}^{n} X_i.$$

Problem 7.49. Let $X_1, \ldots, X_n$ be iid exponential(λ), with density $f(x \mid \lambda) = \lambda^{-1} e^{-x/\lambda}$, $x > 0$, so that λ is the mean.
(a) Find an unbiased estimator of λ based only on $Y = \min\{X_1, \ldots, X_n\}$.
(b) Find a better estimator than the one in part (a). Prove that it is better.
(c) The following data are high-stress failure times (in hours) of Kevlar/epoxy spherical vessels used in a sustained pressure environment on the space shuttle:

50.1, 70.1, 137.0, 166.9, 170.5, 152.8, 80.5, 123.5, 112.6, 148.5, 160.0, 125.4.

Failure times are often modeled with the exponential distribution. Estimate the mean failure time using the estimators from parts (a) and (b).

Solution. (a) nY. Indeed $P(Y > y) = \big[e^{-y/\lambda}\big]^n = e^{-ny/\lambda}$, so $Y \sim$ exponential(λ/n) and $\mathrm{E}\, Y = \lambda/n$.
(b) $\bar{X}$, with $\mathrm{Var}\, \bar{X} = \lambda^2/n$ against $\mathrm{Var}(nY) = n^2(\lambda/n)^2 = \lambda^2$, a factor-$n$ improvement.
To prove it without computing either variance: $T = \sum_i X_i$ is a complete sufficient statistic (exponential family, Theorem 6.2.25), and $\bar{X} = T/n$ is an unbiased function of T. Since $\mathrm{E}[nY \mid T]$ is also an unbiased function of T, completeness forces

$$\mathrm{E}[nY \mid T] = \bar{X} \quad \text{with probability 1,}$$

so $\bar{X}$ is exactly the Rao–Blackwellization of nY and Theorem 7.3.17 gives $\mathrm{Var}\, \bar{X} \leq \mathrm{Var}(nY)$, with strict inequality for $n \geq 2$ because nY is not a function of T alone.
(c) Here $n = 12$, $\sum_i x_i = 1497.9$ and $x_{(1)} = 50.1$, so

$$ny = 12(50.1) = 601.2 \text{ hours}, \qquad \bar{x} = \frac{1497.9}{12} = 124.825 \text{ hours}.$$

7.8 Exercises 7.50–7.56

Problem 7.50. Let $X_1, \dots, X_n$ be iid $\mathrm{n}(\theta, \theta^2)$, $\theta > 0$. For this model both $\bar{X}$ and cS are unbiased estimators of θ, where

$$c = \frac{\sqrt{n-1}\,\Gamma((n-1)/2)}{\sqrt{2}\,\Gamma(n/2)}.$$

(a) Prove that for any number a the estimator $a\bar{X} + (1-a)(cS)$ is an unbiased estimator of θ.
(b) Find the value of a that produces the estimator with minimum variance.
(c) Show that $(\bar{X}, S^2)$ is a sufficient statistic for θ but it is not a complete sufficient statistic.

Solution. (a) Expectation is linear and both pieces already have mean θ:

$$\mathrm{E}_\theta\big[a\bar{X} + (1-a)cS\big] = a\theta + (1-a)\theta = \theta.$$

(b) $a = \dfrac{n(c^2-1)}{n(c^2-1)+1}$.
By Theorem 5.3.1 $\bar{X}$ and S^2 are independent, hence so are $\bar{X}$ and cS, and

$$\begin{aligned}\mathrm{Var}_\theta \bar{X} &= \frac{\theta^2}{n},\\ \mathrm{Var}_\theta(cS) &= c^2\,\mathrm{E}_\theta S^2 - \theta^2 = (c^2-1)\theta^2,\end{aligned}$$

using $\mathrm{E}_\theta S^2 = \theta^2$ and $\mathrm{E}_\theta(cS) = \theta$. So

$$\mathrm{Var}_\theta\big[a\bar{X} + (1-a)cS\big] = \theta^2\Big[\frac{a^2}{n} + (1-a)^2(c^2-1)\Big],$$

a strictly convex quadratic in a whose derivative vanishes at the inverse-variance weight

$$a = \frac{1/\mathrm{Var}_\theta \bar{X}}{1/\mathrm{Var}_\theta \bar{X} + 1/\mathrm{Var}_\theta(cS)} = \frac{n(c^2-1)}{n(c^2-1)+1},$$

free of θ, as it must be for the answer to be an estimator. (Note $c^2 > 1$ because $\mathrm{Var}_\theta(cS) > 0$.)
(c) Since $\sum_i (x_i - \theta)^2 = (n-1)s^2 + n(\bar{x}-\theta)^2$,

$$f(\mathbf{x} \mid \theta) = (2\pi\theta^2)^{-n/2} \exp\Big\{-\frac{(n-1)s^2 + n(\bar{x}-\theta)^2}{2\theta^2}\Big\},$$

which depends on $\mathbf{x}$ only through $(\bar{x}, s^2)$; by the Factorization Theorem 6.2.6, $(\bar{X}, S^2)$ is sufficient.
It is not complete: $g(\bar{X}, S^2) = \bar{X} - c\sqrt{S^2}$ satisfies $\mathrm{E}_\theta\, g = \theta - \theta = 0$ for every $\theta > 0$, yet $P_\theta(g \neq 0) = 1$.

Problem 7.51. Gleser and Healy (1976) give a detailed treatment of the estimation problem in the $\mathrm{n}(\theta, a\theta^2)$ family, where a is a known constant (of which Exercise 7.50 is a special case). We explore a small part of their results here. Again let $X_1, \dots, X_n$ be iid $\mathrm{n}(\theta, \theta^2)$, $\theta > 0$, and let $\bar{X}$ and cS be as in Exercise 7.50. Define the class of estimators

$$\mathcal{T} = \big\{T :\ T = a_1\bar{X} + a_2(cS)\big\},$$

where we do not assume that $a_1 + a_2 = 1$.
(a) Find the estimator $T \in \mathcal{T}$ that minimizes $\mathrm{E}_\theta(\theta - T)^2$; call it T^*.
(b) Show that the MSE of T^* is smaller than the MSE of the estimator derived in Exercise 7.50(b).
(c) Show that the MSE of $T^{*+} = \max\{0, T^*\}$ is smaller than the MSE of T^*.
(d) Would θ be classified as a location parameter or a scale parameter? Explain.

Solution. (a) $T^* = \dfrac{n(c^2-1)\,\bar{X} + cS}{(n+1)(c^2-1)+1}$.
Write $v_1 = \mathrm{Var}_\theta \bar{X}/\theta^2 = 1/n$ and $v_2 = \mathrm{Var}_\theta(cS)/\theta^2 = c^2 - 1$ from Exercise 7.50(b); $\bar{X}$ and cS are

independent with mean θ, and $\mathrm{E}_\theta T = (a_1+a_2)\theta$, so

$$\mathrm{E}_\theta(\theta - T)^2 = \theta^2 \underbrace{\left[a_1^2 v_1 + a_2^2 v_2 + (a_1+a_2-1)^2\right]}_{=:g(a_1,a_2)}.$$

The factor θ^2 means the minimizer is free of θ. Setting $\partial g/\partial a_i = 0$ gives

$$a_1 v_1 = a_2 v_2 = -(a_1+a_2-1) =: s,$$

so $a_i = s/v_i$ and $s(1/v_1 + 1/v_2) = 1 - s$, i.e.

$$s = \frac{1}{1 + \frac{1}{v_1} + \frac{1}{v_2}}, \qquad a_i = \frac{1/v_i}{1 + \frac{1}{v_1} + \frac{1}{v_2}}.$$

With $1/v_1 = n$, $1/v_2 = 1/(c^2-1)$ this is the displayed T^* (multiply numerator and denominator by $c^2 - 1$). Since g is a positive definite quadratic, this stationary point is the unique global minimum. Its value is

$$g(a_1, a_2) = a_1 s + a_2 s + s^2 = s(1-s) + s^2 = s,$$

so

$$\mathrm{MSE}(T^*) = \frac{\theta^2}{1 + n + \frac{1}{c^2-1}}.$$

(b) The estimator of Exercise 7.50(b) is the minimizer of g over the line $a_1 + a_2 = 1$, where the bias term drops out; its MSE is

$$\min_a \theta^2\left[a^2 v_1 + (1-a)^2 v_2\right] = \frac{\theta^2 v_1 v_2}{v_1 + v_2} = \frac{\theta^2}{n + \frac{1}{c^2-1}}.$$

Comparing denominators, $\mathrm{MSE}(T^*)$ carries the extra $+1$ and is therefore strictly the smaller. Strictness is genuine because the unconstrained minimizer has $a_1 + a_2 = 1 - s < 1$, so it lies off the line $a_1 + a_2 = 1$.
(c) $\theta > 0$, so truncating at 0 moves every sample point no further from θ:

$$|T^{*+} - \theta| = \begin{cases} |T^* - \theta|, & T^* \geq 0, \\ \theta < \theta - T^* = |T^* - \theta|, & T^* < 0. \end{cases}$$

Hence $(T^{*+} - \theta)^2 \leq (T^* - \theta)^2$ pointwise, with strict inequality on $\{T^* < 0\}$. That event has positive probability: $cS > 0$ and, by independence and the full support of $\bar{X}$,

$$P_\theta\big(T^* < 0\big) = P_\theta\Big(\bar{X} < -\tfrac{cS}{n(c^2-1)}\Big) > 0.$$

Taking expectations, $\mathrm{MSE}(T^{*+}) < \mathrm{MSE}(T^*)$.
(d) Scale parameter. The pdf satisfies

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi}\,\theta} e^{-(x-\theta)^2/(2\theta^2)} = \frac{1}{\theta} f_1\Big(\frac{x}{\theta}\Big),$$
$$f_1(z) = \frac{1}{\sqrt{2\pi}} e^{-(z-1)^2/2},$$

exactly the form of Definition 3.5.4 with standard pdf f_1, the $\mathrm{n}(1,1)$ density; equivalently $X/\theta \sim \mathrm{n}(1,1)$, free of θ. It is not a location parameter, since $X - \theta \sim \mathrm{n}(0, \theta^2)$ still depends on θ.

Problem 7.52. Let $X_1, \ldots, X_n$ be iid Poisson(λ), and let $\bar{X}$ and S^2 denote the sample mean and variance, respectively. We now complete Example 7.3.8 in a different way. There we used the Cramer-Rao Bound; now we use completeness.

(a) Prove that $\bar{X}$ is the best unbiased estimator of λ without using the Cramer-Rao Theorem.
(b) Prove the rather remarkable identity $\mathrm{E}(S^2 \mid \bar{X}) = \bar{X}$, and use it to explicitly demonstrate that $\operatorname{Var} S^2 > \operatorname{Var} \bar{X}$.
(c) Using completeness, can a general theorem be formulated for which the identity in part (b) is a special case?

Solution. (a) The Poisson family is a full one-parameter exponential family,

$$f(x \mid \lambda) = \frac{e^{-\lambda}}{x!}\, e^{(\log \lambda)x},$$

with natural parameter $\log \lambda$ ranging over all of $\mathbb{R}$ as λ ranges over $(0, \infty)$; by Theorem 6.2.25 $T = \sum_i X_i$ is a complete sufficient statistic. Now $\bar{X} = T/n$ is a function of T alone with $\mathrm{E}_\lambda \bar{X} = \lambda$, so by Theorem 7.3.23 it is the unique best unbiased estimator of λ.
(b) $\mathrm{E}(S^2 \mid \bar{X})$ is a function of the complete sufficient statistic $T = n\bar{X}$, and by Theorem 5.2.6 and the tower property

$$\mathrm{E}_\lambda\big[\mathrm{E}(S^2 \mid \bar{X}) - \bar{X}\big] = \mathrm{E}_\lambda S^2 - \mathrm{E}_\lambda \bar{X} = \lambda - \lambda = 0 \quad \text{for all } \lambda > 0.$$

Completeness of T forces $\mathrm{E}(S^2 \mid \bar{X}) - \bar{X} = 0$ with probability 1, i.e. $\mathrm{E}(S^2 \mid \bar{X}) = \bar{X}$.
Feeding this into the conditional variance identity (Theorem 4.4.7),

$$\begin{aligned} \operatorname{Var}_\lambda S^2 &= \operatorname{Var}_\lambda\big[\mathrm{E}(S^2 \mid \bar{X})\big] + \mathrm{E}_\lambda\big[\operatorname{Var}(S^2 \mid \bar{X})\big] \\ &= \operatorname{Var}_\lambda \bar{X} + \mathrm{E}_\lambda\big[\operatorname{Var}(S^2 \mid \bar{X})\big]. \end{aligned}$$

The last term is strictly positive for $n \geq 2$: on the positive-probability event $\{T = 2\}$ the statistic S^2 is not constant, since the sample $(2, 0, \ldots, 0)$ gives $S^2 = 4/n$ while $(1, 1, 0, \ldots, 0)$ gives $S^2 = (2n-4)/\big(n(n-1)\big)$, and these agree for no $n \geq 2$. Hence $\operatorname{Var}_\lambda S^2 > \operatorname{Var}_\lambda \bar{X} = \lambda/n$.
(c) Yes. If T is a complete sufficient statistic for θ and W is any unbiased estimator of $\tau(\theta)$, then $\mathrm{E}(W \mid T)$ is the unique best unbiased estimator of $\tau(\theta)$ (Theorem 7.3.23, Exercise 7.56). Consequently, if $\phi(T)$ is any function of T with $\mathrm{E}_\theta \phi(T) = \tau(\theta)$, then

$$\mathrm{E}(W \mid T) = \phi(T) \quad \text{with probability 1,}$$

because both sides are functions of T unbiased for $\tau(\theta)$. Part (b) is the case $T = \sum X_i$, $\tau(\lambda) = \lambda$, $W = S^2$, $\phi(T) = T/n = \bar{X}$.

Problem 7.53. Finish some of the details left out of the proof of Theorem 7.3.20. Suppose W is an unbiased estimator of $\tau(\theta)$, and U is an unbiased estimator of 0. Show that if, for some $\theta = \theta_0$, $\operatorname{Cov}_{\theta_0}(W, U) \neq 0$, then W cannot be the best unbiased estimator of $\tau(\theta)$.

Solution. Take $\phi_a = W + aU$ with $a = -\operatorname{Cov}_{\theta_0}(W, U)/\operatorname{Var}_{\theta_0} U$.
First, $\operatorname{Var}_{\theta_0} U > 0$, so a is well defined: if $\operatorname{Var}_{\theta_0} U = 0$ then $U = \mathrm{E}_{\theta_0} U = 0$ with P_{θ_0}-probability 1, forcing $\operatorname{Cov}_{\theta_0}(W, U) = 0$, contrary to hypothesis. (If $\operatorname{Var}_{\theta_0} W = \infty$ then W is certainly not best unbiased, so assume it finite; then $\operatorname{Cov}_{\theta_0}(W, U)$ is finite by Cauchy-Schwarz.)
Next, ϕ_a is an unbiased estimator of $\tau(\theta)$ for every a, since $\mathrm{E}_\theta \phi_a = \tau(\theta) + a \cdot 0 = \tau(\theta)$ for all θ.
Finally, at θ_0,

$$\begin{aligned} \operatorname{Var}_{\theta_0} \phi_a &= \operatorname{Var}_{\theta_0} W + 2a \operatorname{Cov}_{\theta_0}(W, U) + a^2 \operatorname{Var}_{\theta_0} U \\ &= \operatorname{Var}_{\theta_0} W - \frac{\big[\operatorname{Cov}_{\theta_0}(W, U)\big]^2}{\operatorname{Var}_{\theta_0} U} \\ &< \operatorname{Var}_{\theta_0} W, \end{aligned}$$

the strict inequality because $\operatorname{Cov}_{\theta_0}(W, U) \neq 0$. So ϕ_a is unbiased for $\tau(\theta)$ with strictly smaller variance at θ_0, and W is not best unbiased.

Problem 7.54. Consider the "Problem of the Nile" (see Exercise 6.37): (X, Y) has the joint density

$$f(x, y \mid \theta) = \exp\{-(\theta x + y/\theta)\}, \qquad x > 0,\ y > 0,$$

and $(X_1, Y_1), \ldots, (X_n, Y_n)$ is an iid sample from it. The estimators of Exercise 6.37 are

$$T = \sqrt{\sum_i Y_i \Big/ \sum_i X_i}, \qquad U = \sqrt{\sum_i X_i \sum_i Y_i}.$$

(a) Show that T is the MLE of θ and U is ancillary, and

$$\mathrm{E}T = \frac{\Gamma(n+1/2)\Gamma(n-1/2)}{[\Gamma(n)]^2}\,\theta, \qquad \mathrm{E}T^2 = \frac{\Gamma(n+1)\Gamma(n-1)}{[\Gamma(n)]^2}\,\theta^2.$$

(b) Let $Z_1 = (n-1)\big/\sum_i X_i$ and $Z_2 = \sum_i Y_i/n$. Show that both are unbiased with variances $\theta^2/(n-2)$ and θ^2/n, respectively.
(c) Find the best unbiased estimator of the form $aZ_1 + (1-a)Z_2$, calculate its variance, and compare it to the bias-corrected MLE.

Solution. (a) $\hat\theta = T$.
Write $V = \sum_i X_i$, $W = \sum_i Y_i$. The density factors as $(\theta e^{-\theta x})(\theta^{-1} e^{-y/\theta})$, so $X_i \sim$ exponential$(1/\theta)$ and $Y_i \sim$ exponential(θ) independently (scale parametrization: $\mathrm{E}X_i = 1/\theta$, $\mathrm{E}Y_i = \theta$). Hence V and W are independent with

$$\theta V \sim \text{gamma}(n, 1), \qquad W/\theta \sim \text{gamma}(n, 1),$$

and $\mathrm{E}G^r = \Gamma(n+r)/\Gamma(n)$ for $G \sim \text{gamma}(n,1)$, $r > -n$.
The log likelihood is $\ell(\theta) = -\theta V - W/\theta$, so

$$\ell'(\theta) = -V + \frac{W}{\theta^2} = 0 \iff \theta = \sqrt{W/V} = T, \qquad \ell''(\theta) = -\frac{2W}{\theta^3} < 0,$$

and $\ell(\theta) \to -\infty$ at both ends of $(0, \infty)$, so T is the unique maximum.
U is ancillary because

$$U^2 = VW = (\theta V)\big(W/\theta\big)$$

is the product of two independent gamma$(n,1)$ variables, a distribution free of θ.
By independence of V and W,

$$\mathrm{E}T = \mathrm{E}\big[W^{1/2}\big]\,\mathrm{E}\big[V^{-1/2}\big] = \theta^{1/2}\frac{\Gamma(n+\frac12)}{\Gamma(n)} \cdot \theta^{1/2}\frac{\Gamma(n-\frac12)}{\Gamma(n)},$$
$$\mathrm{E}T^2 = \mathrm{E}[W]\,\mathrm{E}\big[V^{-1}\big] = \theta\,\frac{\Gamma(n+1)}{\Gamma(n)} \cdot \theta\,\frac{\Gamma(n-1)}{\Gamma(n)},$$

which are the two stated formulas (the second needs $n \ge 2$). In particular $\mathrm{E}T^2 = \dfrac{n}{n-1}\theta^2$.
(b) With $\mathrm{E}V^{-k} = \theta^k \Gamma(n-k)/\Gamma(n)$ for $k = 1, 2$,

$$\mathrm{E}Z_1 = (n-1)\frac{\theta}{n-1} = \theta, \qquad \mathrm{E}Z_1^2 = (n-1)^2\frac{\theta^2}{(n-1)(n-2)} = \frac{n-1}{n-2}\theta^2,$$

so $\mathrm{Var}\,Z_1 = \theta^2\big(\frac{n-1}{n-2} - 1\big) = \dfrac{\theta^2}{n-2}$ (requires $n \ge 3$). And $\mathrm{E}Z_2 = n\theta/n = \theta$ with $\mathrm{Var}\,Z_2 = n\theta^2/n^2 = \theta^2/n$.
(c) $a^* = \dfrac{n-2}{2(n-1)}$, with variance $\dfrac{\theta^2}{2(n-1)}$.
Z_1 and Z_2 are independent (functions of V and W), so inverse-variance weighting is optimal:

$$a^* = \frac{1/\mathrm{Var}Z_1}{1/\mathrm{Var}Z_1 + 1/\mathrm{Var}Z_2} = \frac{n-2}{(n-2)+n} = \frac{n-2}{2(n-1)},$$

$$\operatorname{Var}\big(a^*Z_1+(1-a^*)Z_2\big)=\frac{1}{\frac{1}{\operatorname{Var}Z_1}+\frac{1}{\operatorname{Var}Z_2}}=\frac{\theta^2}{2(n-1)}.$$

The bias-corrected MLE is $\tilde\theta = T/b_n$ with $b_n=\Gamma(n+\frac12)\Gamma(n-\frac12)/[\Gamma(n)]^2$; it is unbiased with

$$\operatorname{Var}\tilde\theta=\frac{\mathrm{E}T^2-(\mathrm{E}T)^2}{b_n^2}=\theta^2\Big[\frac{n}{(n-1)b_n^2}-1\Big].$$

The MLE wins for every $n\ge 3$. Writing $r_n=\Gamma(n+\frac12)/\Gamma(n)$, the identity $\Gamma(n+\frac12)=(n-\frac12)\Gamma(n-\frac12)$ gives $b_n=r_n^2/(n-\frac12)$, so

$$\operatorname{Var}\tilde\theta<\frac{\theta^2}{2(n-1)}\iff b_n^2>\frac{2n}{2n-1}\iff r_n^4>n\big(n-\tfrac12\big),$$

and Watson's gamma-ratio bound $r_n^2>n-\frac14$ gives $r_n^4>(n-\frac14)^2>n(n-\frac12)$. In units of θ^2:

n	$\operatorname{Var}\tilde\theta$	$\operatorname{Var}(a^*Z_1+(1-a^*)Z_2)$
3	.2297	.2500
10	.0542	.0556
20	.0260	.0263

Problem 7.55. For each of the following pdfs, let $X_1,\dots,X_n$ be a sample from that distribution. In each case, find the best unbiased estimator of θ^r. (See Guenther 1978 for a complete discussion of this problem.)

(a) $f(x\mid\theta)=\dfrac{1}{\theta},\quad 0<x<\theta,\ r<n$

(b) $f(x\mid\theta)=e^{-(x-\theta)},\quad x>\theta$

(c) $f(x\mid\theta)=\dfrac{e^{-x}}{e^{-\theta}-e^{-b}},\quad \theta<x<b,\ b$ known

Solution. In each part a single order statistic Y is a complete sufficient statistic, and $\mathrm{E}_\theta g(Y)=\theta^r$ is an integral equation that one differentiation in θ solves outright; by Theorem 7.3.23 the resulting $g(Y)$ is **the** best unbiased estimator. In (b) and (c) the parameter ranges over $\mathbb{R}$, so there r is a positive integer.

(a) $\dfrac{n+r}{n}\,Y^r$, where $Y=\max_i X_i$.

Y is a complete sufficient statistic (Example 6.2.23) with pdf $f_Y(y)=ny^{n-1}/\theta^n$ on $(0,\theta)$, so

$$\mathrm{E}_\theta g(Y)=\theta^r\iff\int_0^\theta g(y)y^{n-1}\,dy=\frac{\theta^{n+r}}{n}.$$

Differentiating in θ gives $g(\theta)\theta^{n-1}=\frac{n+r}{n}\theta^{n+r-1}$, i.e. $g(y)=\frac{n+r}{n}y^r$. (Check: $\mathrm{E}_\theta Y^r=\frac{n}{n+r}\theta^r$, valid for any real $r>-n$.)

(b) $Y^r-\dfrac{r}{n}Y^{r-1}$, where $Y=\min_i X_i$.

By the Factorization Theorem 6.2.6 applied to $\prod_i e^{-(x_i-\theta)}=e^{n\theta}e^{-\sum x_i}I(\min x_i>\theta)$, Y is sufficient, with pdf $f_Y(y)=ne^{-n(y-\theta)}$ on (θ,∞). It is complete: if $\mathrm{E}_\theta g(Y)=0$ for all θ then $\int_\theta^\infty g(y)e^{-ny}dy=0$ for all θ, and differentiating gives $g(\theta)e^{-n\theta}=0$, so $g\equiv 0$. Now

$$\mathrm{E}_\theta g(Y)=\theta^r\iff\int_\theta^\infty g(y)e^{-ny}\,dy=\frac{\theta^re^{-n\theta}}{n},$$

and differentiating in θ,

$$-g(\theta)e^{-n\theta}=\frac{r\theta^{r-1}e^{-n\theta}-n\theta^re^{-n\theta}}{n}$$
$$\implies\quad g(\theta)=\theta^r-\frac{r}{n}\theta^{r-1}.$$

(Check for $r=1$: $\mathrm{E}_\theta Y=\theta+1/n$.)

(c) $Y^r - \dfrac{r}{n}Y^{r-1}\big(1 - e^{Y-b}\big)$, where $Y = \min_i X_i$.
The joint pdf is $\big(e^{-\theta} - e^{-b}\big)^{-n} e^{-\sum x_i} I(\min x_i > \theta) I(\max x_i < b)$, so Y is sufficient, with

$$f_Y(y) = \frac{n\,(e^{-y} - e^{-b})^{n-1} e^{-y}}{(e^{-\theta} - e^{-b})^n}, \qquad \theta < y < b.$$

Completeness follows as in (b): $\int_\theta^b g(y)(e^{-y} - e^{-b})^{n-1} e^{-y} dy = 0$ for all $\theta < b$ differentiates to $g(\theta)(e^{-\theta} - e^{-b})^{n-1} e^{-\theta} = 0$, hence $g \equiv 0$. Writing $A(\theta) = (e^{-\theta} - e^{-b})^n$, unbiasedness says

$$\int_\theta^b g(y)(e^{-y} - e^{-b})^{n-1} e^{-y}\, dy = \frac{\theta^r A(\theta)}{n},$$

and differentiating in θ,

$$\begin{aligned} -g(\theta)(e^{-\theta} - e^{-b})^{n-1} e^{-\theta} &= \frac{r\theta^{r-1}}{n} A(\theta) - \theta^r (e^{-\theta} - e^{-b})^{n-1} e^{-\theta} \\ \implies \quad g(\theta) &= \theta^r - \frac{r}{n}\,\theta^{r-1}\,\frac{e^{-\theta} - e^{-b}}{e^{-\theta}} \\ &= \theta^r - \frac{r}{n}\,\theta^{r-1}\big(1 - e^{\theta - b}\big). \end{aligned}$$

Problem 7.56. Prove the assertion made in the text preceding Example 7.3.24: If T is a complete sufficient statistic for a parameter θ, and $h(X_1, \ldots, X_n)$ is any unbiased estimator of $\tau(\theta)$, then $\phi(T) = \mathrm{E}\big(h(X_1, \ldots, X_n) \mid T\big)$ is the best unbiased estimator of $\tau(\theta)$.

Solution. Let W be an arbitrary unbiased estimator of $\tau(\theta)$; we show $\mathrm{Var}_\theta\, \phi(T) \le \mathrm{Var}_\theta W$ for every θ. By the Rao-Blackwell Theorem 7.3.17 applied to h (whose hypotheses hold: h is unbiased for $\tau(\theta)$ and T is sufficient, so $\phi(T) = \mathrm{E}(h \mid T)$ is a genuine statistic, free of θ), $\phi(T)$ is unbiased for $\tau(\theta)$. Applying the same theorem to W, the statistic $\psi(T) := \mathrm{E}(W \mid T)$ is unbiased for $\tau(\theta)$ and

$$\mathrm{Var}_\theta\, \psi(T) \le \mathrm{Var}_\theta W \qquad \text{for all } \theta.$$

Both $\phi(T)$ and $\psi(T)$ are functions of T with

$$\mathrm{E}_\theta\big[\phi(T) - \psi(T)\big] = \tau(\theta) - \tau(\theta) = 0 \qquad \text{for all } \theta,$$

so completeness of T gives $P_\theta\big(\phi(T) = \psi(T)\big) = 1$ for every θ. Therefore

$$\mathrm{Var}_\theta\, \phi(T) = \mathrm{Var}_\theta\, \psi(T) \le \mathrm{Var}_\theta W,$$

and since W was arbitrary, $\phi(T)$ is the best unbiased estimator of $\tau(\theta)$; it is **the** best unbiased estimator because best unbiased estimators are unique (Theorem 7.3.19).

7.9 Exercises 7.57–7.63

Problem 7.57. Let $X_1, \ldots, X_{n+1}$ be iid Bernoulli(p), and define the function $h(p)$ by

$$h(p) = P\left(\sum_{i=1}^n X_i > X_{n+1} \,\middle|\, p\right),$$

the probability that the first n observations exceed the $(n+1)$st.
(a) Show that

$$T(X_1, \ldots, X_{n+1}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n X_i > X_{n+1}, \\ 0 & \text{otherwise} \end{cases}$$

is an unbiased estimator of $h(p)$.
(b) Find the best unbiased estimator of $h(p)$.

Solution. (a) T is the indicator of the very event whose probability defines h, so

$$E_pT = P_p\left(\sum_{i=1}^{n} X_i > X_{n+1}\right) = h(p).$$

(b) The best unbiased estimator is $\phi(T^*)$, where $T^* = \sum_{i=1}^{n+1} X_i$ and

$$\phi(t) = \begin{cases} 0, & t = 0, \\ \dfrac{n}{n+1}, & t = 1, \\ \dfrac{n-1}{n+1}, & t = 2, \\ 1, & t \geq 3. \end{cases}$$

Indeed $T^* \sim \text{binomial}(n+1, p)$ is a complete sufficient statistic, so by the assertion proved in Exercise 7.56 (and Theorem 7.3.23) the Rao–Blackwellized estimator $\phi(T^*) = E(T \mid T^*)$ is the best unbiased estimator of $h(p)$. To evaluate it, note that given $T^* = t$ the t ones occupy a uniformly chosen t-subset of the $n+1$ coordinates, so

$$P(X_{n+1} = 1 \mid T^* = t) = \frac{t}{n+1}, \qquad P(X_{n+1} = 0 \mid T^* = t) = \frac{n+1-t}{n+1}.$$

Conditioning on X_{n+1} and writing $Y = \sum_{i=1}^{n} X_i = T^* - X_{n+1}$:
(i) $X_{n+1} = 0$: the event is $Y = t > 0$, which holds iff $t \geq 1$.
(ii) $X_{n+1} = 1$: the event is $Y = t - 1 > 1$, which holds iff $t \geq 3$.
Hence $\phi(0) = 0$; for $t = 1, 2$ only branch (i) contributes, giving $\phi(t) = (n+1-t)/(n+1)$; and for $t \geq 3$ both branches contribute, giving $\phi(t) = 1$. (Check: $\sum_t \binom{n+1}{t} p^t (1-p)^{n+1-t} \phi(t) = h(p)$.)

Problem 7.58. Let X be an observation from the pdf

$$f(x|\theta) = \left(\frac{\theta}{2}\right)^{|x|} (1-\theta)^{1-|x|}, \qquad x = -1, 0, 1; \quad 0 \leq \theta \leq 1.$$

(a) Find the MLE of θ.
(b) Define the estimator $T(X)$ by

$$T(X) = \begin{cases} 2 & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Show that $T(X)$ is an unbiased estimator of θ.
(c) Find a better estimator than $T(X)$ and prove that it is better.

Solution. (a) $\hat{\theta} = |X|$. The likelihood takes only two values on $0 \leq \theta \leq 1$: $L(\theta|0) = 1 - \theta$, maximized at $\theta = 0$, and $L(\theta| \pm 1) = \theta/2$, maximized at $\theta = 1$.
(b) $P_\theta(X = 1) = \theta/2$, so $E_\theta T = 2 \cdot (\theta/2) = \theta$.
(c) $|X|$ is better. Since $P_\theta(X = \pm 1) = \theta/2$ and $P_\theta(X = 0) = 1 - \theta$, $|X| \sim \text{Bernoulli}(\theta)$; in particular $E_\theta|X| = \theta$, so $|X|$ is unbiased, and both estimators have MSE equal to their variance. Now

$$\begin{aligned} \text{Var}_\theta\, T &= E_\theta T^2 - \theta^2 = 4 \cdot \tfrac{\theta}{2} - \theta^2 = 2\theta - \theta^2, \\ \text{Var}_\theta\, |X| &= \theta(1-\theta) = \theta - \theta^2, \end{aligned}$$

so $\text{Var}_\theta\, |X| < \text{Var}_\theta\, T$ for every $\theta \in (0, 1]$, with equality only at $\theta = 0$. (This is Rao–Blackwell, Theorem

7.3.17: $E(T \mid |X|) = 2 \cdot \frac{1}{2} \cdot 1_{\{|X|=1\}} = |X|$, and since $|X|$ is complete, Theorem 7.3.23 makes it the best unbiased estimator.)

Problem 7.59. Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$. Find the best unbiased estimator of σ^p, where p is a known positive constant, not necessarily an integer.

Solution. The best unbiased estimator is

$$\frac{(n-1)^{p/2}\,\Gamma\left(\frac{n-1}{2}\right)}{2^{p/2}\,\Gamma\left(\frac{n-1+p}{2}\right)}\,S^p.$$

Here $(\bar{X}, S^2)$ is sufficient (Example 6.2.9) and complete, the normal family being a full two-parameter exponential family (Theorem 6.2.25), and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ (Theorem 5.3.1). For $W \sim \chi^2_\nu$ and $r > -\nu/2$,

$$EW^r = \frac{1}{\Gamma(\nu/2)2^{\nu/2}} \int_0^\infty w^{\nu/2+r-1} e^{-w/2}\,dw = \frac{2^r\,\Gamma(\frac{\nu}{2}+r)}{\Gamma(\frac{\nu}{2})}.$$

Taking $\nu = n-1$ and $r = p/2$ (legitimate since $p > 0$),

$$E_\sigma S^p = E\left[\left(\frac{\sigma^2 W}{n-1}\right)^{p/2}\right] = \sigma^p\,\frac{2^{p/2}\,\Gamma\left(\frac{n-1+p}{2}\right)}{(n-1)^{p/2}\,\Gamma\left(\frac{n-1}{2}\right)}.$$

Dividing S^p by this constant gives an unbiased estimator of σ^p that is a function of the complete sufficient statistic alone, so it is the unique best unbiased estimator by Theorem 7.3.23.

Problem 7.60. Let $X_1, \ldots, X_n$ be iid gamma(α, β) with α known. Find the best unbiased estimator of $1/\beta$.

Solution. The best unbiased estimator is

$$\frac{n\alpha - 1}{\sum_{i=1}^n X_i} \qquad (n\alpha > 1).$$

The gamma family with α known is a one-parameter exponential family with natural sufficient statistic $T = \sum_{i=1}^n X_i$, which is complete by Theorem 6.2.25; and $T \sim \text{gamma}(n\alpha, \beta)$. Hence

$$E_\beta\left[\frac{1}{T}\right] = \frac{1}{\Gamma(n\alpha)\beta^{n\alpha}} \int_0^\infty t^{n\alpha-2} e^{-t/\beta}\,dt = \frac{\Gamma(n\alpha-1)\,\beta^{n\alpha-1}}{\Gamma(n\alpha)\,\beta^{n\alpha}} = \frac{1}{(n\alpha-1)\beta},$$

the integral converging precisely when $n\alpha > 1$. So $(n\alpha-1)/T$ is an unbiased estimator of $1/\beta$ based only on the complete sufficient statistic, and Theorem 7.3.23 makes it the unique best unbiased estimator.

Problem 7.61. Show that the log of the likelihood function for estimating σ^2, based on observing $S^2 \sim \sigma^2\chi^2_\nu/\nu$, can be written in the form

$$\log L(\sigma^2|s^2) = K_1\frac{s^2}{\sigma^2} - K_2 \log\frac{s^2}{\sigma^2} + K_3,$$

where K_1, K_2, and K_3 are constants, not dependent on σ^2. Relate the above log likelihood to the loss function discussed in Example 7.3.27. See Anderson (1984a) for a discussion of this relationship.

Solution. $K_1 = K_2 = -\nu/2$. Writing $S^2 = \sigma^2 W/\nu$ with $W \sim \chi^2_\nu$, the change of variables $w = \nu s^2/\sigma^2$

(Jacobian ν/σ^2) gives

$$f(s^2|\sigma^2) = \frac{\nu}{\sigma^2} \cdot \frac{1}{\Gamma(\nu/2)2^{\nu/2}} \left(\frac{\nu s^2}{\sigma^2}\right)^{\nu/2-1} e^{-\nu s^2/(2\sigma^2)}$$
$$= \frac{\nu^{\nu/2}}{\Gamma(\nu/2)2^{\nu/2}} (s^2)^{\nu/2-1} (\sigma^2)^{-\nu/2} e^{-\nu s^2/(2\sigma^2)},$$

so that, with $c = \log\big[\nu^{\nu/2}/(\Gamma(\nu/2)2^{\nu/2})\big]$,

$$\log L(\sigma^2|s^2) = -\frac{\nu}{2}\frac{s^2}{\sigma^2} - \frac{\nu}{2}\log\sigma^2 + \Big(\frac{\nu}{2} - 1\Big)\log s^2 + c.$$

Substituting $-\frac{\nu}{2}\log\sigma^2 = \frac{\nu}{2}\log\frac{s^2}{\sigma^2} - \frac{\nu}{2}\log s^2$ collapses the two $\log s^2$ terms into one and yields

$$\log L(\sigma^2|s^2) = -\frac{\nu}{2}\frac{s^2}{\sigma^2} + \frac{\nu}{2}\log\frac{s^2}{\sigma^2} + \underbrace{\big(c - \log s^2\big)}_{K_3},$$

which is the stated form with $K_1 = -\nu/2$ and $-K_2 = \nu/2$, i.e. $K_1 = K_2 = -\nu/2$.
Stein's loss of Example 7.3.27 is $L(\sigma^2, a) = \dfrac{a}{\sigma^2} - 1 - \log\dfrac{a}{\sigma^2}$, so at the action $a = s^2$ the two displays match term by term:

$$\log L(\sigma^2|s^2) = -\frac{\nu}{2}\Big[L(\sigma^2, s^2) + 1\Big] + K_3 = -\frac{\nu}{2}L(\sigma^2, s^2) + K_3',$$

so Stein's loss is the negative log likelihood up to the positive scale $\nu/2$ and an additive constant: minimizing the loss and maximizing the likelihood are the same problem, both solved at $\sigma^2 = s^2$.

Problem 7.62. Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population, σ^2 known. Consider estimating θ using squared error loss. Let $\pi(\theta)$ be a $\mathrm{n}(\mu, \tau^2)$ prior distribution on θ and let δ^π be the Bayes estimator of θ. Verify the following formulas for the risk function and Bayes risk.
(a) For any constants a and b, the estimator $\delta(\mathbf{x}) = a\bar{X} + b$ has risk function

$$R(\theta, \delta) = a^2\frac{\sigma^2}{n} + \big(b - (1-a)\theta\big)^2.$$

(b) Let $\eta = \sigma^2/(n\tau^2 + \sigma^2)$. The risk function for the Bayes estimator is

$$R(\theta, \delta^\pi) = (1-\eta)^2\frac{\sigma^2}{n} + \eta^2(\theta - \mu)^2.$$

(c) The Bayes risk for the Bayes estimator is

$$B(\pi, \delta^\pi) = \tau^2\eta.$$

Solution. (a) Squared error risk splits as variance plus squared bias (equation (7.3.1)):

$$R(\theta, \delta) = \mathrm{Var}_\theta(a\bar{X} + b) + \big(E_\theta(a\bar{X} + b) - \theta\big)^2$$
$$= a^2\frac{\sigma^2}{n} + \big(a\theta + b - \theta\big)^2 = a^2\frac{\sigma^2}{n} + \big(b - (1-a)\theta\big)^2,$$

using $E_\theta\bar{X} = \theta$ and $\mathrm{Var}_\theta\,\bar{X} = \sigma^2/n$.
(b) Under squared error loss the posterior expected loss is minimized by the posterior mean (the discussion following (7.3.19)). Here $\bar{X}$ is sufficient with $\bar{X} \sim \mathrm{n}(\theta, \sigma^2/n)$, so Example 7.2.16 (with σ^2

replaced by σ^2/n) gives the posterior of θ given $\mathbf{x}$ as

$$\mathrm{n}\left(\frac{\tau^2}{\tau^2+\sigma^2/n}\,\bar{x}+\frac{\sigma^2/n}{\tau^2+\sigma^2/n}\,\mu,\ \frac{\sigma^2\tau^2/n}{\tau^2+\sigma^2/n}\right).$$

Since $\dfrac{\sigma^2/n}{\tau^2+\sigma^2/n}=\dfrac{\sigma^2}{n\tau^2+\sigma^2}=\eta$, the Bayes estimator is

$$\delta^\pi(\mathbf{x})=(1-\eta)\bar{x}+\eta\mu,$$

which is of the form in part (a) with $a=1-\eta$ and $b=\eta\mu$. Substituting,

$$R(\theta,\delta^\pi)=(1-\eta)^2\frac{\sigma^2}{n}+\big(\eta\mu-\eta\theta\big)^2=(1-\eta)^2\frac{\sigma^2}{n}+\eta^2(\theta-\mu)^2.$$

(c) The Bayes risk is the prior average of the risk function (see (7.3.19)), and under π we have $E^\pi(\theta-\mu)^2=\tau^2$. Writing $D=n\tau^2+\sigma^2$, so that $\eta=\sigma^2/D$ and $1-\eta=n\tau^2/D$,

$$\begin{aligned}B(\pi,\delta^\pi)&=(1-\eta)^2\frac{\sigma^2}{n}+\eta^2\tau^2=\frac{n^2\tau^4}{D^2}\cdot\frac{\sigma^2}{n}+\frac{\sigma^4}{D^2}\tau^2\\&=\frac{\tau^2\sigma^2\big(n\tau^2+\sigma^2\big)}{D^2}=\frac{\tau^2\sigma^2}{D}=\tau^2\eta.\end{aligned}$$

Problem 7.63. Let $X\sim\mathrm{n}(\mu,1)$. Let δ^π be the Bayes estimator of μ for squared error loss. Compute and graph the risk functions, $R(\mu,\delta^\pi)$, for $\pi(\mu)\sim\mathrm{n}(0,1)$ and $\pi(\mu)\sim\mathrm{n}(0,10)$. Comment on how the prior affects the risk function of the Bayes estimator.

Solution.

$$R(\mu,\delta^\pi)=\frac{1+\mu^2}{4}\quad\text{for }\pi=\mathrm{n}(0,1),\qquad R(\mu,\delta^\pi)=\frac{100+\mu^2}{121}\quad\text{for }\pi=\mathrm{n}(0,10).$$

This is Exercise 7.62 with $n=1$, $\sigma^2=1$, prior mean $\mu_0=0$ and prior variance τ^2, so $\eta=1/(\tau^2+1)$, $\delta^\pi(x)=(1-\eta)x$, and

$$R(\mu,\delta^\pi)=(1-\eta)^2+\eta^2\mu^2=\frac{\tau^4+\mu^2}{(\tau^2+1)^2}.$$

Putting $\tau^2=1$ gives $\eta=\frac12$ and $R=\frac14(1+\mu^2)$; putting $\tau^2=10$ gives $\eta=\frac{1}{11}$ and $R=(100+\mu^2)/121$. Both risks are upward parabolas in μ centred at the prior mean 0, to be compared with the constant risk $R(\mu,X)=1$ of the MLE:

μ	0	1	1.5	2	3	4	5
$\tau^2=1$	0.250	0.500	0.813	1.250	2.500	4.250	6.500
$\tau^2=10$	0.826	0.835	0.845	0.860	0.901	0.959	1.033

The tighter prior shrinks harder ($x/2$ versus $10x/11$), buying far lower risk near the prior mean (0.250 against 0.826 at $\mu=0$) at the price of a much steeper parabola (leading coefficient $\eta^2=1/4$ against $1/121$). The curves cross at $|\mu|=\sqrt{31/13}\approx1.54$; the $\tau^2=1$ rule beats the MLE only for $|\mu|<\sqrt{3}$, the $\tau^2=10$ rule out to $|\mu|<\sqrt{21}$.

7.10 Exercises 7.64–7.66

Problem 7.64. Let $X_1,\dots,X_n$ be independent random variables, where X_i has cdf $F(x|\theta_i)$. Show that, for $i=1,\dots,n$, if $\delta_i^{\pi_i}(X_i)$ is a Bayes rule for estimating θ_i using loss $L(\theta_i,a_i)$ and prior $\pi_i(\theta_i)$, then

$$\delta^\pi(\mathbf{X})=(\delta_1^{\pi_1}(X_1),\dots,\delta_n^{\pi_n}(X_n))$$

is a Bayes rule for estimating $\theta=(\theta_1,\dots,\theta_n)$ using the loss $\sum_{i=1}^n L(\theta_i,a_i)$ and prior $\pi(\theta)=$

$\prod_{i=1}^n \pi_i(\theta_i)$.

Solution. The posterior factors, and a loss that separates across coordinates is therefore minimized coordinatewise.
By (7.3.19), a Bayes rule is any rule that, for each observed $\mathbf{x}$, minimizes the posterior expected loss. Independence of the X_i given θ together with the product prior gives a product marginal $m(\mathbf{x}) = \prod_{i=1}^n m_i(x_i)$, where $m_i(x_i) = \int f(x_i|\theta_i)\pi_i(\theta_i)\,d\theta_i$, whence

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \frac{\prod_{i=1}^n f(x_i|\theta_i)\,\prod_{i=1}^n \pi_i(\theta_i)}{\prod_{i=1}^n m_i(x_i)} \\ &= \prod_{i=1}^n \pi_i(\theta_i|x_i).\end{aligned}$$

So under the posterior the θ_i are independent, with θ_i distributed as $\pi_i(\cdot|x_i)$ and in particular not depending on x_j, $j \neq i$. Hence for any action $\mathbf{a} = (a_1, \ldots, a_n)$,

$$\begin{aligned}E\Big[\sum_{i=1}^n L(\theta_i, a_i)\,\Big|\,\mathbf{X} = \mathbf{x}\Big] &= \sum_{i=1}^n E\big[L(\theta_i, a_i)\,|\,\mathbf{X} = \mathbf{x}\big] \\ &= \sum_{i=1}^n E\big[L(\theta_i, a_i)\,|\,X_i = x_i\big].\end{aligned}$$

The ith summand involves only the ith coordinate of $\mathbf{a}$, so the sum over the product action space is minimized by minimizing each summand separately; by hypothesis the ith is minimized at $a_i = \delta_i^{\pi_i}(x_i)$. Therefore

$$\min_{\mathbf{a}} E\Big[\sum_{i=1}^n L(\theta_i, a_i)\,\Big|\,\mathbf{x}\Big] = \sum_{i=1}^n E\big[L(\theta_i, \delta_i^{\pi_i}(x_i))\,|\,x_i\big],$$

attained at $\mathbf{a} = \delta^\pi(\mathbf{x})$. Since this holds for every $\mathbf{x}$, δ^π minimizes the Bayes risk (7.3.19) and is a Bayes rule.

Problem 7.65. A loss function investigated by Zellner (1986) is the LINEX (LINear-EXponential) loss, a loss function that can handle asymmetries in a smooth way. The LINEX loss is given by

$$L(\theta, a) = e^{c(a-\theta)} - c(a-\theta) - 1,$$

where c is a positive constant. As the constant c varies, the loss function varies from very asymmetric to almost symmetric.
(a) For $c = .2, .5, 1$, plot $L(\theta, a)$ as a function of $a - \theta$.
(b) If $X \sim F(x|\theta)$, show that the Bayes estimator of θ, using a prior π, is given by

$$\delta^\pi(X) = -\tfrac{1}{c}\log E\big(e^{-c\theta}|X\big).$$

(c) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, \sigma^2)$, where σ^2 is known, and suppose that θ has the noninformative prior $\pi(\theta) = 1$. Show that the Bayes estimator versus LINEX loss is given by $\delta^B(\bar{X}) = \bar{X} - (c\sigma^2/(2n))$.
(d) Calculate the posterior expected loss for $\delta^B(\bar{X})$ and $\bar{X}$ using LINEX loss.
(e) Calculate the posterior expected loss for $\delta^B(\bar{X})$ and $\bar{X}$ using squared error loss.

Solution. (a) Write $d = a - \theta$, so $L = e^{cd} - cd - 1$ with $L'(d) = c(e^{cd} - 1)$ and $L''(d) = c^2 e^{cd} > 0$: each curve is strictly convex with its unique zero minimum at $d = 0$. For $d < 0$ the loss is asymptotically the line $-cd$; for $d > 0$ it explodes exponentially. Near the origin $L(d) \approx c^2 d^2/2$, so small c looks like a shallow symmetric parabola and large c like a sharply one-sided curve. Sample ordinates:

$a-\theta$	-3	-2	-1	0	1	2	3
$c=.2$	0.1488	0.0703	0.0187	0	0.0214	0.0918	0.2221
$c=.5$	0.7231	0.3679	0.1065	0	0.1487	0.7183	1.9817
$c=1$	2.0498	1.1353	0.3679	0	0.7183	4.3891	16.0855

At $c = .2$ the two sides stay within a factor of 1.5 out to $|d| = 3$ (nearly symmetric); at $c = 1$ overestimation by 3 costs about eight times underestimation by 3.

(b) By (7.3.19) the Bayes rule minimizes posterior expected loss at each x:

$$E[L(\theta,a)|x] = e^{ca}E(e^{-c\theta}|x) - c\big(a - E(\theta|x)\big) - 1.$$

Assume $E(e^{-c\theta}|x) < \infty$. Differentiating in a,

$$\frac{\partial}{\partial a}E[L(\theta,a)|x] = c\,e^{ca}E(e^{-c\theta}|x) - c,$$

which vanishes exactly when $e^{ca}E(e^{-c\theta}|x) = 1$, that is at

$$\delta^{\pi}(x) = -\tfrac{1}{c}\log E\big(e^{-c\theta}|x\big).$$

The second derivative $c^2e^{ca}E(e^{-c\theta}|x) > 0$, so this stationary point is the global minimum.

(c) With $\pi(\theta) = 1$ and $\bar{x}$ sufficient, $\pi(\theta|\mathbf{x}) \propto \exp\{-n(\bar{x}-\theta)^2/(2\sigma^2)\}$, i.e. $\theta|\mathbf{x} \sim \mathrm{n}(\bar{x}, \sigma^2/n)$ (a proper posterior, so the formal Bayes rule is defined). By the normal mgf $M(t) = \exp\{\mu t + \tau^2t^2/2\}$ (Table of Common Distributions) evaluated at $t = -c$,

$$E\big(e^{-c\theta}|\mathbf{x}\big) = \exp\{-c\bar{x} + c^2\sigma^2/(2n)\},$$

so part (b) gives

$$\delta^B(\bar{X}) = -\tfrac{1}{c}\Big(-c\bar{X} + \frac{c^2\sigma^2}{2n}\Big) = \bar{X} - \frac{c\sigma^2}{2n}.$$

(d) From the display in (b) with $E(\theta|\mathbf{x}) = \bar{x}$ and the mgf value above,

$$E[L(\theta,a)|\mathbf{x}] = \exp\Big\{c(a-\bar{x}) + \frac{c^2\sigma^2}{2n}\Big\} - c(a-\bar{x}) - 1.$$

Put $t = c^2\sigma^2/(2n)$. At $a = \delta^B(\bar{x})$ we have $c(a-\bar{x}) = -t$, so the exponent is 0 and

$$E[L(\theta,\delta^B)|\mathbf{x}] = 1 + t - 1 = \frac{c^2\sigma^2}{2n};$$

at $a = \bar{x}$ we have $c(a-\bar{x}) = 0$ and

$$E[L(\theta,\bar{X})|\mathbf{x}] = e^t - 1 = \exp\{c^2\sigma^2/(2n)\} - 1.$$

(e) For squared error loss, $E[(\theta-a)^2|\mathbf{x}] = \mathrm{Var}(\theta|\mathbf{x}) + (\bar{x}-a)^2$, so

$$E[(\theta-\bar{X})^2|\mathbf{x}] = \frac{\sigma^2}{n},$$
$$E[(\theta-\delta^B)^2|\mathbf{x}] = \frac{\sigma^2}{n} + \frac{c^2\sigma^4}{4n^2}.$$

Problem 7.66. The jackknife is a general technique for reducing bias in an estimator (Quenouille, 1956). A one-step jackknife estimator is defined as follows. Let $X_1, \ldots, X_n$ be a random sample, and let $T_n = T_n(X_1, \ldots, X_n)$ be some estimator of a parameter θ. In order to "jackknife" T_n we calculate the n statistics $T_n^{(i)}$, $i = 1, \ldots, n$, where $T_n^{(i)}$ is calculated just as T_n but using the $n-1$ observations

with X_i removed from the sample. The jackknife estimator of θ, denoted by $\mathrm{JK}(T_n)$, is given by

$$\mathrm{JK}(T_n) = nT_n - \frac{n-1}{n}\sum_{i=1}^{n} T_n^{(i)}.$$

(In general, $\mathrm{JK}(T_n)$ will have a smaller bias than T_n. See Miller 1974 for a good review of the properties of the jackknife.)
Now, to be specific, let $X_1, \ldots, X_n$ be iid Bernoulli(θ). The object is to estimate θ^2.
(a) Show that the MLE of θ^2, $(\sum_{i=1}^n X_i/n)^2$, is a biased estimator of θ^2.
(b) Derive the one-step jackknife estimator based on the MLE.
(c) Show that the one-step jackknife estimator is an unbiased estimator of θ^2. (In general, jackknifing only reduces bias. In this special case, however, it removes it entirely.)
(d) Is this jackknife estimator the best unbiased estimator of θ^2? If so, prove it. If not, find the best unbiased estimator.

Solution. Write $S = \sum_{i=1}^n X_i \sim \mathrm{binomial}(n, \theta)$, so that $T_n = \bar{X}^2 = S^2/n^2$.
(a) By invariance (Theorem 7.2.10) the MLE of θ^2 is $\bar{X}^2$, and

$$E_\theta \bar{X}^2 = \mathrm{Var}_\theta\, \bar{X} + (E_\theta \bar{X})^2 = \frac{\theta(1-\theta)}{n} + \theta^2,$$

so the bias is $\theta(1-\theta)/n > 0$ for every $0 < \theta < 1$.
(b) Deleting X_i leaves $T_n^{(i)} = (S - X_i)^2/(n-1)^2$. Since each X_i is 0 or 1 we have $X_i^2 = X_i$ and hence $\sum_i X_i^2 = S$, so

$$\begin{aligned}\sum_{i=1}^{n}(S - X_i)^2 &= nS^2 - 2S\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} X_i^2 \\ &= (n-2)S^2 + S.\end{aligned}$$

Therefore

$$\begin{aligned}\mathrm{JK}(T_n) &= \frac{nS^2}{n^2} - \frac{n-1}{n}\cdot\frac{(n-2)S^2 + S}{(n-1)^2} \\ &= \frac{(n-1)S^2 - (n-2)S^2 - S}{n(n-1)} \\ &= \frac{S(S-1)}{n(n-1)}.\end{aligned}$$

(c) The second factorial moment of a binomial(n, θ) variable is $E_\theta[S(S-1)] = n(n-1)\theta^2$, so

$$E_\theta\, \mathrm{JK}(T_n) = \frac{n(n-1)\theta^2}{n(n-1)} = \theta^2.$$

(d) Yes. S is a complete sufficient statistic for θ (Example 6.2.22), and $\mathrm{JK}(T_n) = S(S-1)/[n(n-1)]$ is a function of S alone whose expectation is θ^2 by (c); by Theorem 7.3.23 it is therefore the unique best unbiased estimator of θ^2.

8 Hypothesis Testing

Contents

8.1 Exercises 8.1–8.7 280
8.2 Exercises 8.8–8.14 285
8.3 Exercises 8.15–8.21 290
8.4 Exercises 8.22–8.28 295
8.5 Exercises 8.29–8.35 299
8.6 Exercises 8.36–8.42 304
8.7 Exercises 8.43–8.49 312
8.8 Exercises 8.50–8.56 318
8.9 Exercises 8.57–8.58 325

8.1 Exercises 8.1–8.7

Problem 8.1. In 1,000 tosses of a coin, 560 heads and 440 tails appear. Is it reasonable to assume that the coin is fair? Justify your answer.

Solution. No. Model the tosses as $X_1, \dots, X_n$ iid Bernoulli(p) with $n = 1000$, $y = \sum x_i = 560$, and test $H_0\colon p = 1/2$ versus $H_1\colon p \neq 1/2$. The MLE is $\hat{p} = y/n = .56$, so by Definition 8.2.1 the LRT statistic is

$$\lambda(\mathbf{x}) = \frac{(1/2)^y (1/2)^{n-y}}{\hat{p}^{\,y}(1-\hat{p})^{n-y}},$$

$$\begin{aligned} -2\log\lambda(\mathbf{x}) &= 2n\Big[\hat{p}\log\tfrac{\hat{p}}{1/2} + (1-\hat{p})\log\tfrac{1-\hat{p}}{1/2}\Big] \\ &= 2000\big[.56\log(1.12) + .44\log(.88)\big] \\ &= 14.43. \end{aligned}$$

Under H_0 this is approximately χ^2_1 (one free parameter, restricted to a point), so the p-value is $P(\chi^2_1 \geq 14.43) = .00015$ and H_0 is rejected at any conventional level.
Method (2): under H_0, $Y \sim$ binomial$(1000, 1/2)$ has mean 500 and standard deviation $\sqrt{1000(.25)} = 15.81$, so

$$z = \frac{560 - 500}{15.81} = 3.79, \qquad P(|Z| \geq 3.79) = .00015,$$

and the exact binomial two-sided value is $2P(Y \geq 560) = .00017$.

Problem 8.2. In a given city it is assumed that the number of automobile accidents in a given year follows a Poisson distribution. In past years the average number of accidents per year was 15, and this year it was 10. Is it justified to claim that the accident rate has dropped?

Solution. No. Let $X \sim$ Poisson(θ) be this year's count (writing θ for the rate, since $\lambda(x)$ denotes the LRT statistic) and test $H_0\colon \theta \geq 15$ versus $H_1\colon \theta < 15$, with $x = 10$ observed. The MLE is $\hat{\theta} = x = 10$, and since L is decreasing on $\theta > x$ the restricted MLE is $\hat{\theta}_0 = \max\{x, 15\} = 15$; so by Definition 8.2.1

$$\lambda(x) = \frac{e^{-15}15^x/x!}{e^{-x}x^x/x!} = e^{x-15}\Big(\frac{15}{x}\Big)^x \qquad (x < 15),$$

which is increasing in x on $x < 15$ (its logarithm has derivative $\log(15/x) > 0$). Hence $\{\lambda(x) \leq c\} = \{x \leq b\}$: the LRT rejects for small counts, and the p-value is

$$P_{15}(X \leq 10) = \sum_{k=0}^{10} \frac{e^{-15}15^k}{k!} = .118.$$

At $x = 10$, $\lambda(10) = e^{-5}(1.5)^{10} = .389$, so $-2\log\lambda(10) = 1.89$ and the χ^2_1 approximation agrees. A count of 10 sits $(15-10)/\sqrt{15} = 1.29$ standard deviations below the historical mean: ordinary Poisson variation, not a drop in the rate.

Problem 8.3. Here, the LRT alluded to in Example 8.2.9 will be derived. Suppose that we observe m iid Bernoulli(θ) random variables, denoted by $Y_1, \dots, Y_m$. Show that the LRT of $H_0\colon \theta \leq \theta_0$ versus $H_1\colon \theta > \theta_0$ will reject H_0 if $\sum_{i=1}^m Y_i > b$.

Solution. $\lambda(\mathbf{y})$ is a nonincreasing function of $y = \sum_{i=1}^m y_i$, so $\{\lambda \leq c\}$ is an upper tail in y. Write $\hat{\theta} = y/m$; the likelihood $L(\theta \mid \mathbf{y}) = \theta^y(1-\theta)^{m-y}$ increases on $[0, \hat{\theta}]$ and decreases on $[\hat{\theta}, 1]$, so the

maximum over $\Theta_0 = [0, \theta_0]$ sits at $\hat{\theta}_0 = \min\{\hat{\theta}, \theta_0\}$ and Definition 8.2.1 gives

$$\lambda(\mathbf{y}) = \begin{cases} 1, & \hat{\theta} \le \theta_0, \\ \dfrac{\theta_0^{\,y}(1-\theta_0)^{m-y}}{\hat{\theta}^{\,y}(1-\hat{\theta})^{m-y}}, & \hat{\theta} > \theta_0. \end{cases}$$

On the second branch,

$$\begin{aligned} -\log \lambda(\mathbf{y}) &= m\Big[\hat{\theta}\log\frac{\hat{\theta}}{\theta_0} + (1-\hat{\theta})\log\frac{1-\hat{\theta}}{1-\theta_0}\Big] \\ &=: m\,K(\hat{\theta}, \theta_0), \end{aligned}$$

and, regarding $\hat{\theta} = p$ as continuous,

$$\frac{\partial}{\partial p}K(p, \theta_0) = \log\frac{p(1-\theta_0)}{\theta_0(1-p)} > 0 \quad \text{for } p > \theta_0,$$

since $p \mapsto p/(1-p)$ is increasing. Thus λ is strictly decreasing in y on $\{y > m\theta_0\}$ and equals 1 on $\{y \le m\theta_0\}$: it is nonincreasing in y throughout, and strictly so wherever $\lambda < 1$.
Consequently, for any $c \in (0, 1)$ the rejection region $\{\mathbf{y} : \lambda(\mathbf{y}) \le c\}$ is an upper set in y, i.e.

$$\{\mathbf{y} : \lambda(\mathbf{y}) \le c\} = \Big\{\mathbf{y} : \sum_{i=1}^{m} y_i > b\Big\}, \qquad b = \min\{y : \lambda(y) \le c\} - 1,$$

which is the test used for $H_{02}\colon \theta_2 \le .95$ in the acceptance-sampling intersection-union test of Example 8.2.9.

Problem 8.4. Prove the assertion made in the text after Definition 8.2.1. If $f(x \mid \theta)$ is the pmf of a discrete random variable, then the numerator of $\lambda(\mathbf{x})$, the LRT statistic, is the maximum probability of the observed sample when the maximum is computed over parameters in the null hypothesis. Furthermore, the denominator of $\lambda(\mathbf{x})$ is the maximum probability of the observed sample over all possible parameters.

Solution. The likelihood evaluated at the observed sample is the probability of the observed sample. Indeed, for $X_1, \ldots, X_n$ iid with pmf $f(x \mid \theta)$, independence gives, for the observed $\mathbf{x} = (x_1, \ldots, x_n)$,

$$\begin{aligned} L(\theta \mid \mathbf{x}) &= \prod_{i=1}^{n} f(x_i \mid \theta) = \prod_{i=1}^{n} P_\theta(X_i = x_i) \\ &= P_\theta(X_1 = x_1, \ldots, X_n = x_n) = P_\theta(\mathbf{X} = \mathbf{x}), \end{aligned}$$

the middle equality being the defining property of a pmf and the next one independence; without the iid assumption the same conclusion $L(\theta \mid \mathbf{x}) = f(\mathbf{x} \mid \theta) = P_\theta(\mathbf{X} = \mathbf{x})$ is the definition of the joint pmf. Since $\mathbf{x}$ is held fixed and only θ varies, taking suprema over the two regions of Θ gives

$$\begin{aligned} \sup_{\Theta_0} L(\theta \mid \mathbf{x}) &= \sup_{\theta \in \Theta_0} P_\theta(\mathbf{X} = \mathbf{x}), \\ \sup_{\Theta} L(\theta \mid \mathbf{x}) &= \sup_{\theta \in \Theta} P_\theta(\mathbf{X} = \mathbf{x}), \end{aligned}$$

so the numerator of $\lambda(\mathbf{x})$ in Definition 8.2.1 is the largest probability the null hypothesis can assign to the sample actually observed, and the denominator is the largest probability any $\theta \in \Theta$ can assign to it. When the suprema are attained (for instance whenever a restricted and an unrestricted MLE exist) they are maxima, as asserted.

Problem 8.5. A random sample, $X_1, \ldots, X_n$, is drawn from a Pareto population with pdf

$$f(x \mid \theta, \nu) = \frac{\theta \nu^\theta}{x^{\theta+1}} I_{[\nu,\infty)}(x), \qquad \theta > 0,\ \nu > 0.$$

(a) Find the MLEs of θ and ν.
(b) Show that the LRT of

$$H_0\colon \theta = 1,\ \nu \text{ unknown} \qquad \text{versus} \qquad H_1\colon \theta \neq 1,\ \nu \text{ unknown},$$

has critical region of the form $\{\mathbf{x} : T(\mathbf{x}) \le c_1 \text{ or } T(\mathbf{x}) \ge c_2\}$, where $0 < c_1 < c_2$ and

$$T = \log\left[\frac{\prod_{i=1}^n X_i}{(\min_i X_i)^n}\right].$$

(c) Show that, under H_0, $2T$ has a chi squared distribution, and find the number of degrees of freedom. (Hint: Obtain the joint distribution of the $n-1$ nontrivial terms $X_i/(\min_i X_i)$ conditional on $\min_i X_i$. Put these $n-1$ terms together, and notice that the distribution of T given $\min_i X_i$ does not depend on $\min_i X_i$, so it is the unconditional distribution of T.)

Solution. (a) $\hat\nu = X_{(1)} = \min_i X_i$ and $\hat\theta = n/T$, with $T = \sum_i \log(X_i/X_{(1)})$ the statistic of part (b). The likelihood is

$$L(\theta, \nu \mid \mathbf{x}) = \theta^n \nu^{n\theta} \Big(\prod_i x_i\Big)^{-(\theta+1)} I_{[0,\,x_{(1)}]}(\nu),$$

which for each $\theta > 0$ is strictly increasing in ν on $(0, x_{(1)}]$; hence $\hat\nu = x_{(1)}$, as in Example 8.2.3. Substituting and differentiating,

$$\frac{\partial}{\partial\theta} \log L(\theta, x_{(1)} \mid \mathbf{x}) = \frac{n}{\theta} + n \log x_{(1)} - \sum_i \log x_i = \frac{n}{\theta} - T,$$

which is positive for $\theta < n/T$ and negative for $\theta > n/T$, so $\hat\theta = n/T$ is the maximum. ($T > 0$ with probability 1.)
(b) Under H_0 only ν is free, and the same monotonicity gives the restricted MLE $\hat\nu_0 = x_{(1)}$, so

$$\sup_{\Theta_0} L = x_{(1)}^n \Big(\prod_i x_i\Big)^{-2},$$

$$\sup_{\Theta} L = \hat\theta^{\,n} x_{(1)}^{n\hat\theta} \Big(\prod_i x_i\Big)^{-(\hat\theta+1)}.$$

Dividing, and using $\prod_i x_i / x_{(1)}^n = e^T$ together with $\hat\theta = n/T$,

$$\lambda(\mathbf{x}) = \hat\theta^{\,-n} \left[\frac{\prod_i x_i}{x_{(1)}^n}\right]^{\hat\theta - 1} = \Big(\frac{T}{n}\Big)^n e^{\,T(\hat\theta - 1)}$$
$$= \Big(\frac{T}{n}\Big)^n e^{\,n - T} =: g(T).$$

Now $\log g(t) = n \log t - n \log n + n - t$ has derivative $n/t - 1$, so g increases strictly on $(0, n)$, decreases strictly on (n, ∞), and $g(n) = 1$; also $g(0^+) = 0$ and $g(\infty) = 0$. Hence for each $c \in (0, 1)$ the set $\{g(T) \le c\}$ is exactly $\{T \le c_1\} \cup \{T \ge c_2\}$, where $0 < c_1 < n < c_2$ are the two roots of $g(t) = c$.
(c) $2T \sim \chi^2_{2(n-1)}$.
Since $P_{\theta,\nu}(X > x) = (\nu/x)^\theta$ for $x \ge \nu$, the variables $Z_i = \log(X_i/\nu)$ satisfy $P(Z_i > z) = e^{-\theta z}$, i.e. $Z_1, \ldots, Z_n$ are iid exponential with mean $1/\theta$, and under H_0 ($\theta = 1$) they are iid exponential(1). Because $\log(X_i/X_{(1)}) = Z_i - Z_{(1)}$,

$$T = \sum_{i=1}^n \big(Z_{(i)} - Z_{(1)}\big) = \sum_{i=2}^n (n - i + 1)\big(Z_{(i)} - Z_{(i-1)}\big),$$

the second equality being the count of how many times each gap $Z_{(i)} - Z_{(i-1)}$ appears. Apply to the order-statistic density $n!\, e^{-\sum_i z_i}$ on $z_1 < \cdots < z_n$ (Theorem 5.4.6) the change of variables

$$V = Z_{(1)}, \qquad W_i = (n-i+1)(Z_{(i)} - Z_{(i-1)}), \quad i = 2, \dots, n,$$

whose inverse is triangular with Jacobian $\left[\prod_{i=2}^{n}(n-i+1)\right]^{-1} = 1/(n-1)!$, and which satisfies $\sum_i z_{(i)} = nv + \sum_{i=2}^{n} w_i$. The joint density becomes

$$\frac{n!}{(n-1)!}\, e^{-nv} \prod_{i=2}^{n} e^{-w_i} = \left(ne^{-nv}\right) \prod_{i=2}^{n} e^{-w_i}, \qquad v > 0,\ w_i > 0.$$

So $V = Z_{(1)}$ is independent of $W_2, \dots, W_n$, which are iid exponential(1); this is the hint's statement that the conditional distribution of the nontrivial ratios given $\min_i X_i$ does not depend on $\min_i X_i$. Therefore

$$T = \sum_{i=2}^{n} W_i \sim \text{gamma}(n-1, 1), \qquad 2T \sim \text{gamma}(n-1, 2) = \chi^2_{2(n-1)}.$$

Problem 8.6. Suppose that we have two independent random samples: $X_1, \dots, X_n$ are exponential(θ), and $Y_1, \dots, Y_m$ are exponential(μ).
(a) Find the LRT of $H_0\colon \theta = \mu$ versus $H_1\colon \theta \neq \mu$.
(b) Show that the test in part (a) can be based on the statistic

$$T = \frac{\sum X_i}{\sum X_i + \sum Y_i}.$$

(c) Find the distribution of T when H_0 is true.

Solution. (a) Reject H_0 for small values of

$$\lambda(\mathbf{x}, \mathbf{y}) = \frac{(n+m)^{n+m}}{n^n m^m}\, T^n (1-T)^m, \qquad T = \frac{\sum x_i}{\sum x_i + \sum y_i}.$$

Indeed, with $s = \sum x_i$, $u = \sum y_i$ and the exponential(β) pdf $\beta^{-1}e^{-x/\beta}$,

$$L(\theta, \mu) = \theta^{-n} e^{-s/\theta}\, \mu^{-m} e^{-u/\mu},$$

so the unrestricted MLEs are $\hat\theta = \bar x = s/n$ and $\hat\mu = \bar y = u/m$ (here $\partial \log L/\partial\theta = -n/\theta + s/\theta^2$ changes sign once, from + to −, at $\theta = s/n$; likewise for μ), giving $\sup_\Theta L = (s/n)^{-n}(u/m)^{-m}e^{-(n+m)}$. Under H_0 the two samples pool into one exponential sample of size $n+m$, so $\hat\theta_0 = (s+u)/(n+m)$ and

$$\sup_{\Theta_0} L = \left(\frac{s+u}{n+m}\right)^{-(n+m)} e^{-(s+u)/\hat\theta_0} = \left(\frac{s+u}{n+m}\right)^{-(n+m)} e^{-(n+m)}.$$

The exponential factors cancel and

$$\begin{aligned} \lambda &= \frac{(s/n)^n (u/m)^m}{\big((s+u)/(n+m)\big)^{n+m}} \\ &= \frac{(n+m)^{n+m}}{n^n m^m} \cdot \frac{s^n u^m}{(s+u)^{n+m}}. \end{aligned}$$

(b) The last fraction is $T^n(1-T)^m$, since $s/(s+u) = T$ and $u/(s+u) = 1-T$; so λ depends on the data only through T. Writing $h(t) = t^n(1-t)^m$, we have $h'(t)/h(t) = n/t - m/(1-t)$, positive for $t < n/(n+m)$ and negative for $t > n/(n+m)$, with h vanishing at both endpoints. Hence λ is unimodal in T with maximum value 1 at $T = n/(n+m)$, and the LRT rejection region is

$$\{\lambda \le c\} = \{T \le c_1\} \cup \{T \ge c_2\}, \qquad 0 < c_1 < \frac{n}{n+m} < c_2 < 1,$$

the two roots of $h(t) = c\,n^n m^m/(n+m)^{n+m}$. So the test rejects H_0 when T is too far from $n/(n+m)$ in either direction.
(c) $T \sim \text{beta}(n, m)$.
Under H_0 let θ be the common value. Sums of iid exponentials are gamma, so $S = \sum X_i \sim \text{gamma}(n, \theta)$ and $U = \sum Y_i \sim \text{gamma}(m, \theta)$ are independent, whence $S/\theta \sim \text{gamma}(n, 1)$ and $U/\theta \sim \text{gamma}(m, 1)$ are independent and Exercise 4.24 applies to them:

$$T = \frac{S/\theta}{S/\theta + U/\theta} \sim \text{beta}(n, m),$$

free of the nuisance parameter θ, so c_1, c_2 can be set for exact size α.

Problem 8.7. We have already seen the usefulness of the LRT in dealing with problems with nuisance parameters. We now look at some other nuisance parameter problems.
(a) Find the LRT of

$$H_0\colon \theta \le 0 \qquad \text{versus} \qquad H_1\colon \theta > 0$$

based on a sample $X_1, \ldots, X_n$ from a population with probability density function $f(x \mid \theta, \lambda) = \frac{1}{\lambda} e^{-(x-\theta)/\lambda} I_{[\theta,\infty)}(x)$, where both θ and λ are unknown.
(b) We have previously seen that the exponential pdf is a special case of a gamma pdf. Generalizing in another way, the exponential pdf can be considered as a special case of the Weibull(γ, β). The Weibull pdf, which reduces to the exponential if $\gamma = 1$, is very important in modeling reliability of systems. Suppose that $X_1, \ldots, X_n$ is a random sample from a Weibull population with both γ and β unknown. Find the LRT of $H_0\colon \gamma = 1$ versus $H_1\colon \gamma \ne 1$.

Solution. (a) Reject H_0 for large values of

$$W = \frac{nX_{(1)}}{\sum_i (X_i - X_{(1)})}, \qquad \text{i.e. reject iff } W \ge k > 0.$$

(Throughout this exercise $\Lambda(\mathbf{x})$ denotes the LRT statistic, since λ is the scale parameter in part (a).) The likelihood is

$$L(\theta, \lambda \mid \mathbf{x}) = \lambda^{-n} \exp\Big[-\tfrac{1}{\lambda} \textstyle\sum_i (x_i - \theta) \Big] I_{(-\infty,\, x_{(1)}]}(\theta),$$

increasing in θ on $(-\infty, x_{(1)}]$ for every $\lambda > 0$ (as in Example 8.2.3). Hence the unrestricted maximizer is $\hat\theta = x_{(1)}$, and then $\hat\lambda = \frac{1}{n}\sum_i (x_i - x_{(1)})$ by the exponential-scale computation of Exercise 8.6(a), so $\sup_\Theta L = \hat\lambda^{-n} e^{-n}$.
Under $\Theta_0 = \{\theta \le 0\}$ the same monotonicity gives $\hat\theta_0 = \min\{x_{(1)}, 0\}$, and then $\hat\lambda_0 = \frac{1}{n}\sum_i (x_i - \hat\theta_0)$:
(i) If $x_{(1)} \le 0$ the restricted and unrestricted maxima coincide and $\Lambda(\mathbf{x}) = 1$.
(ii) If $x_{(1)} > 0$ then $\hat\theta_0 = 0$, $\hat\lambda_0 = \bar x$, and $\sup_{\Theta_0} L = \bar x^{-n} e^{-n}$, so

$$\Lambda(\mathbf{x}) = \Big(\frac{\hat\lambda}{\bar x}\Big)^n = \Big(\frac{\bar x - x_{(1)}}{\bar x}\Big)^n = (1 + W)^{-n},$$

using $\bar x/(\bar x - x_{(1)}) = 1 + x_{(1)}/(\bar x - x_{(1)}) = 1 + W$.
Since $W \le 0$ exactly in case (i) and $w \mapsto (1+w)^{-n}$ is strictly decreasing for $w > 0$, we get $\{\Lambda(\mathbf{x}) \le c\} = \{W \ge c^{-1/n} - 1\}$ for every $c \in (0, 1)$, which is the stated form. At the boundary $\theta = 0$ the statistic W is scale free, so k is set from that (size-determining) null distribution.
(b) With the Weibull pdf $f(x \mid \gamma, \beta) = \frac{\gamma}{\beta} x^{\gamma-1} e^{-x^\gamma/\beta}$, $x > 0$, reject H_0 for small values of

$$\Lambda(\mathbf{x}) = \frac{1}{\hat\gamma^{\,n}} \left[\frac{\sum_i x_i^{\hat\gamma}}{\sum_i x_i} \right]^n \Big(\prod_i x_i \Big)^{1-\hat\gamma},$$

where $\hat\gamma$ is the unrestricted MLE of γ, defined implicitly below.

For fixed γ, $\log L = n\log\gamma - n\log\beta + (\gamma-1)\sum_i \log x_i - \frac{1}{\beta}\sum_i x_i^\gamma$ is maximized over β at $\hat\beta(\gamma) = \frac{1}{n}\sum_i x_i^\gamma$, giving the profile likelihood

$$L\big(\gamma, \hat\beta(\gamma)\big) = \gamma^n \Big(\frac{n}{\sum_i x_i^\gamma}\Big)^n \Big(\prod_i x_i\Big)^{\gamma-1} e^{-n}.$$

Its logarithm has derivative

$$\frac{n}{\gamma} + \sum_i \log x_i - n\,\frac{\sum_i x_i^\gamma \log x_i}{\sum_i x_i^\gamma},$$

so $\hat\gamma$ is the root of

$$\frac{1}{\hat\gamma} + \frac{1}{n}\sum_i \log x_i = \frac{\sum_i x_i^{\hat\gamma}\log x_i}{\sum_i x_i^{\hat\gamma}},$$

which has no closed form and must be found numerically. Setting $\gamma = 1$ in the profile likelihood gives the null maximum $\sup_{\Theta_0} L = (n/\sum_i x_i)^n e^{-n}$, since the Weibull reduces to the exponential(β) there with $\hat\beta_0 = \bar x$; dividing by $L(\hat\gamma, \hat\beta(\hat\gamma))$ and cancelling e^{-n} yields the displayed $\Lambda(\mathbf{x})$.

8.2 Exercises 8.8–8.14

Problem 8.8. A special case of a normal family is one in which the mean and the variance are related, the $\mathrm{n}(\theta, a\theta)$ family. If we are interested in testing this relationship, regardless of the value of θ, we are again faced with a nuisance parameter problem.
(a) Find the LRT of $H_0 : a = 1$ versus $H_1 : a \neq 1$ based on a sample $X_1, \ldots, X_n$ from a $\mathrm{n}(\theta, a\theta)$ family, where θ is unknown.
(b) A similar question can be asked about a related family, the $\mathrm{n}(\theta, a\theta^2)$ family. Thus, if $X_1, \ldots, X_n$ are iid $\mathrm{n}(\theta, a\theta^2)$, where θ is unknown, find the LRT of $H_0 : a = 1$ versus $H_1 : a \neq 1$.

Solution. (a) $\lambda(\mathbf{x}) = \Big(\hat\sigma^2/\hat\theta_0\Big)^{n/2} e^{\,n(\bar x - \hat\theta_0)}$, where $\hat\sigma^2 = \frac{1}{n}\sum (x_i - \bar x)^2$ and $\hat\theta_0$ is the positive root of $\theta^2 + \theta = \frac{1}{n}\sum x_i^2$.
In the unrestricted family the mean is $\mu = \theta$ and the variance is $\sigma^2 = a\theta$, and as (a, θ) ranges over $(0,\infty)^2$ the pair (μ, σ^2) ranges over all of $(0,\infty)^2$. So the unrestricted MLEs are the ordinary normal ones, $\hat\theta = \bar x$ and $\widehat{a\theta} = \hat\sigma^2$ (assume $\bar x > 0$, so that this lies in the parameter space), giving

$$\sup_{a,\theta} L = (2\pi\hat\sigma^2)^{-n/2} e^{-n/2}.$$

Under H_0 the model is $\mathrm{n}(\theta, \theta)$ and

$$\log L(\theta) = -\tfrac{n}{2}\log(2\pi\theta) - \frac{\sum x_i^2}{2\theta} + \sum x_i - \frac{n\theta}{2},$$

so $\partial/\partial\theta = 0$ gives, after multiplying by $2\theta^2/n$,

$$\theta^2 + \theta - \tfrac{1}{n}\textstyle\sum x_i^2 = 0, \qquad \hat\theta_0 = \frac{-1 + \sqrt{1 + \frac{4}{n}\sum x_i^2}}{2}.$$

(The other root is negative, and $\log L \to -\infty$ at both ends of $(0,\infty)$, so this is the maximum.) At the root, $\sum x_i^2 = n(\hat\theta_0^2 + \hat\theta_0)$, hence

$$\begin{aligned}\frac{\sum (x_i - \hat\theta_0)^2}{2\hat\theta_0} &= \frac{n(\hat\theta_0^2 + \hat\theta_0) - 2n\hat\theta_0\bar x + n\hat\theta_0^2}{2\hat\theta_0} \\ &= \frac{n(2\hat\theta_0 + 1)}{2} - n\bar x.\end{aligned}$$

Therefore

$$\lambda(\mathbf{x}) = \frac{(2\pi\hat\theta_0)^{-n/2}\exp\big[-\tfrac{n}{2}(2\hat\theta_0+1)+n\bar x\big]}{(2\pi\hat\sigma^2)^{-n/2}e^{-n/2}} = \Big(\frac{\hat\sigma^2}{\hat\theta_0}\Big)^{n/2} e^{\,n(\bar x-\hat\theta_0)},$$

and the LRT rejects H_0 when $\lambda(\mathbf{x}) \le c$.
(b) $\lambda(\mathbf{x}) = \big(\hat\sigma^2/\hat\theta_0^2\big)^{n/2}\exp\Big[\dfrac{n\bar x}{2\hat\theta_0} - \dfrac{n}{2}\Big]$, with

$$\hat\theta_0 = \operatorname{sgn}(\bar x)\,\frac{\sqrt{\bar x^2 + \tfrac{4}{n}\sum x_i^2} - |\bar x|}{2}.$$

Again $(\mu,\sigma^2) = (\theta, a\theta^2)$ ranges over $(\mathbb{R}\setminus\{0\})\times(0,\infty)$, so the denominator is the same $(2\pi\hat\sigma^2)^{-n/2}e^{-n/2}$. Under H_0,

$$\log L(\theta) = -n\log|\theta| - \tfrac{n}{2}\log 2\pi - \frac{\sum x_i^2}{2\theta^2} + \frac{n\bar x}{\theta} - \frac{n}{2},$$

and $\partial/\partial\theta = 0$ gives $n\theta^2 + n\bar x\,\theta - \sum x_i^2 = 0$, whose two roots $\theta_\pm = \tfrac12\big(-\bar x \pm D\big)$, $D = \sqrt{\bar x^2 + \tfrac{4}{n}\sum x_i^2}$, have opposite signs and are the maxima of $\log L$ on $(0,\infty)$ and $(-\infty,0)$ respectively. At a root, $\sum(x_i-\theta)^2/(2\theta^2) = n - n\bar x/(2\theta)$, so with $t = \bar x/D \in (-1,1)$,

$$\frac{\log L(\theta_+) - \log L(\theta_-)}{n} = \log\frac{1+t}{1-t} + \frac{2t}{1-t^2},$$

which has the sign of t, i.e. of $\bar x$. Hence $\hat\theta_0 = \theta_+$ when $\bar x > 0$ and $\hat\theta_0 = \theta_-$ when $\bar x < 0$, which is the displayed formula. Substituting,

$$\lambda(\mathbf{x}) = \frac{(2\pi\hat\theta_0^2)^{-n/2}\exp\big[-n+\frac{n\bar x}{2\hat\theta_0}\big]}{(2\pi\hat\sigma^2)^{-n/2}e^{-n/2}} = \Big(\frac{\hat\sigma^2}{\hat\theta_0^2}\Big)^{n/2}\exp\Big[\frac{n\bar x}{2\hat\theta_0} - \frac{n}{2}\Big],$$

and H_0 is rejected for $\lambda(\mathbf{x}) \le c$.

Problem 8.9. Stefanski (1996) establishes the arithmetic-geometric-harmonic mean inequality (see Example 4.7.8 and Miscellanea 4.9.2) using a proof based on likelihood ratio tests. Suppose that $Y_1,\dots,Y_n$ are independent with pdfs $\lambda_i e^{-\lambda_i y_i}$, and we want to test

$$H_0 : \lambda_1 = \cdots = \lambda_n \qquad \text{vs.} \qquad H_1 : \lambda_i \text{ are not all equal.}$$

(a) Show that the LRT statistic is given by $(\bar Y)^{-n}/\big(\prod_i Y_i\big)^{-1}$ and hence deduce the arithmetic-geometric mean inequality.
(b) Make the transformation $X_i = 1/Y_i$, and show that the LRT statistic based on $X_1,\dots,X_n$ is given by $\big[n/\sum_i(1/X_i)\big]^n/\prod_i X_i$ and hence deduce the geometric-harmonic mean inequality.

Solution. (a) $L(\boldsymbol\lambda \mid \mathbf{y}) = \prod_i \lambda_i e^{-\lambda_i y_i}$ factors into n separate one-parameter problems, so the unrestricted MLEs are $\hat\lambda_i = 1/y_i$ and

$$\sup_{\boldsymbol\lambda} L = \prod_{i=1}^n \frac{1}{y_i}e^{-1} = \frac{e^{-n}}{\prod_i y_i}.$$

Under H_0 the common value λ has $L(\lambda) = \lambda^n e^{-\lambda\sum y_i}$, whose MLE is $\hat\lambda_0 = n/\sum y_i = 1/\bar y$, giving $\sup_{H_0} L = \bar y^{-n}e^{-n}$. Hence

$$\lambda(\mathbf{y}) = \frac{\bar y^{-n}e^{-n}}{e^{-n}\big(\prod_i y_i\big)^{-1}} = \frac{\prod_{i=1}^n y_i}{\bar y^{\,n}}.$$

A likelihood ratio is a restricted supremum divided by an unrestricted one, so $\lambda(\mathbf{y}) \le 1$ identically;

since $\mathbf{y}$ ranges over all of $(0,\infty)^n$, this says

$$\Big(\prod_{i=1}^n y_i\Big)^{1/n} \le \frac{1}{n}\sum_{i=1}^n y_i \qquad \text{for all } y_i > 0,$$

the arithmetic-geometric mean inequality.
(b) With $x_i = 1/y_i$ the Jacobian is x_i^{-2}, so $f(x_i \mid \lambda_i) = \lambda_i e^{-\lambda_i/x_i} x_i^{-2}$ and

$$L(\boldsymbol{\lambda} \mid \mathbf{x}) = \Big(\prod_i x_i^{-2}\Big)\prod_i \lambda_i e^{-\lambda_i/x_i}.$$

The unrestricted MLEs are $\hat\lambda_i = x_i$ and under H_0, $\hat\lambda_0 = n/\sum_i (1/x_i)$; the Jacobian factor cancels in the ratio, as does e^{-n}, leaving

$$\lambda(\mathbf{x}) = \frac{\big[n/\sum_i(1/x_i)\big]^n e^{-n}}{\big(\prod_i x_i\big)e^{-n}} = \frac{\big[n/\sum_i(1/x_i)\big]^n}{\prod_i x_i}.$$

Again $\lambda(\mathbf{x}) \le 1$ for every $\mathbf{x} \in (0,\infty)^n$, i.e.

$$\frac{n}{\sum_{i=1}^n (1/x_i)} \le \Big(\prod_{i=1}^n x_i\Big)^{1/n},$$

the harmonic-geometric mean inequality.

Problem 8.10. Let $X_1, \dots, X_n$ be iid Poisson(λ), and let λ have a gamma(α, β) distribution, the conjugate family for the Poisson. In Exercise 7.24 the posterior distribution of λ was found, including the posterior mean and variance. Now consider a Bayesian test of $H_0 : \lambda \le \lambda_0$ versus $H_1 : \lambda > \lambda_0$.
(a) Calculate expressions for the posterior probabilities of H_0 and H_1.
(b) If $\alpha = \frac{5}{2}$ and $\beta = 2$, the prior distribution is a chi squared distribution with 5 degrees of freedom. Explain how a chi squared table could be used to perform a Bayesian test.

Solution. (a) By Exercise 7.24 the posterior is $\lambda \mid \mathbf{x} \sim \text{gamma}\big(\alpha + \sum x_i,\ \beta/(n\beta+1)\big)$ (the prior kernel $\lambda^{\alpha-1}e^{-\lambda/\beta}$ times the likelihood kernel $\lambda^{\sum x_i} e^{-n\lambda}$ is $\lambda^{\alpha+\sum x_i - 1} e^{-\lambda(n\beta+1)/\beta}$). Writing $a = \alpha + \sum x_i$ and $b = \beta/(n\beta+1)$,

$$P(H_0 \mid \mathbf{x}) = \frac{1}{\Gamma(a) b^a}\int_0^{\lambda_0} \lambda^{a-1} e^{-\lambda/b}\, d\lambda,$$

and $P(H_1 \mid \mathbf{x}) = 1 - P(H_0 \mid \mathbf{x})$. The Bayes test rejects H_0 when $P(H_1 \mid \mathbf{x}) > P(H_0 \mid \mathbf{x})$, i.e. when $P(H_0 \mid \mathbf{x}) < \frac{1}{2}$.
(b) With $\alpha = \frac{5}{2}$, $\beta = 2$ the posterior is $\text{gamma}\big(\frac{5}{2} + \sum x_i,\ \frac{2}{2n+1}\big)$, so rescaling to scale parameter 2,

$$(2n+1)\lambda \mid \mathbf{x} \ \sim\ \text{gamma}\Big(\tfrac{5}{2} + \textstyle\sum x_i,\ 2\Big) = \chi^2_{5+2\sum x_i},$$

an integer degrees of freedom because $\sum x_i$ is a nonnegative integer. Hence

$$P(H_0 \mid \mathbf{x}) = P\big(\chi^2_{5+2\sum x_i} \le (2n+1)\lambda_0\big),$$

read directly off a chi squared table with $5 + 2\sum x_i$ degrees of freedom at the point $(2n+1)\lambda_0$: reject H_0 if that tabulated probability is less than $\frac{1}{2}$, equivalently if $(2n+1)\lambda_0$ falls below the median of the $\chi^2_{5+2\sum x_i}$ distribution.

Problem 8.11. In Exercise 7.23 the posterior distribution of σ^2, the variance of a normal population, given S^2, the sample variance based on a sample of size n, was found using a conjugate prior for σ^2

(the inverted gamma pdf with parameters α and β). Based on observing S^2, a decision about the hypotheses $H_0 : \sigma \le 1$ versus $H_1 : \sigma > 1$ is to be made.
(a) Find the region of the sample space for which $P(\sigma \le 1 \mid s^2) > P(\sigma > 1 \mid s^2)$, the region for which a Bayes test will decide that $\sigma \le 1$.
(b) Compare the region in part (a) with the acceptance region of an LRT. Is there any choice of prior parameters for which the regions agree?

Solution. (a) $\Big\{s^2 :\ s^2 < \dfrac{m_{\,2\alpha+n-1} - 2/\beta}{n-1}\Big\}$, where m_ν denotes the median of the χ^2_ν distribution.
By Exercise 7.23 the posterior is

$$\sigma^2 \mid s^2 \ \sim\ \mathrm{IG}\Big(a,\, b\Big), \quad a = \alpha + \frac{n-1}{2}, \quad b = \Big[\frac{(n-1)s^2}{2} + \frac{1}{\beta}\Big]^{-1},$$

so $1/\sigma^2 \mid s^2 \sim \text{gamma}(a, b)$ and $(2/b)(1/\sigma^2) \mid s^2 \sim \text{gamma}(a, 2) = \chi^2_{2a}$. Since $\{\sigma \le 1\} = \{\sigma^2 \le 1\} = \{1/\sigma^2 \ge 1\}$,

$$\begin{aligned} P(\sigma \le 1 \mid s^2) &= P\Big(\chi^2_{2a} \ \ge\ \tfrac{2}{b}\Big) \\ &= P\Big(\chi^2_{2\alpha+n-1} \ \ge\ (n-1)s^2 + \tfrac{2}{\beta}\Big). \end{aligned}$$

This exceeds $\frac{1}{2}$ exactly when $(n-1)s^2 + 2/\beta$ lies below the median $m_{2\alpha+n-1}$, which is the displayed region. (It is empty unless $\beta > 2/m_{2\alpha+n-1}$.)
(b) The LRT of $H_0 : \sigma \le 1$ has $\lambda(\mathbf{x}) = 1$ when $\hat\sigma^2 = \frac{1}{n}\sum(x_i - \bar x)^2 \le 1$ and is decreasing in $\hat\sigma^2$ thereafter, so its size-γ acceptance region is

$$\big\{s^2 :\ (n-1)s^2 \le \chi^2_{n-1,\gamma}\big\},$$

$\chi^2_{n-1,\gamma}$ being the upper-γ cutoff. Both regions are one-sided intervals $\{s^2 < \text{constant}\}$, so they agree iff

$$m_{2\alpha+n-1} - \frac{2}{\beta} = \chi^2_{n-1,\gamma}.$$

Because $m_\nu \uparrow \infty$ in ν, choose any α with $m_{2\alpha+n-1} > \chi^2_{n-1,\gamma}$ and then

$$\beta = \frac{2}{m_{2\alpha+n-1} - \chi^2_{n-1,\gamma}} > 0.$$

So yes. For instance with $n = 11$ and $\gamma = .05$, $\chi^2_{10,.05} = 18.307$; taking $\alpha = 4.5$ gives $m_{19} = 18.338$ and $\beta = 2/(18.338 - 18.307) = 65.3$.

Problem 8.12. For samples of size $n = 1, 4, 16, 64, 100$ from a normal population with mean μ and known variance σ^2, plot the power function of the following LRTs. Take $\alpha = .05$.
(a) $H_0 : \mu \le 0$ versus $H_1 : \mu > 0$
(b) $H_0 : \mu = 0$ versus $H_1 : \mu \ne 0$

Solution. (a) $\beta(\mu) = P\Big(Z > 1.645 - \dfrac{\sqrt{n}\,\mu}{\sigma}\Big) = 1 - \Phi\Big(1.645 - \dfrac{\sqrt{n}\,\mu}{\sigma}\Big)$.
Here $\lambda(\mathbf{x}) = 1$ for $\bar x \le 0$ and $\exp[-n\bar x^2/(2\sigma^2)]$ for $\bar x > 0$, so the LRT rejects for large $\bar X$, and $\sup_{\mu \le 0}\beta(\mu) = \beta(0) = .05$ forces the cutoff $\bar x > z_{.05}\sigma/\sqrt{n} = 1.645\,\sigma/\sqrt{n}$; then $\sqrt{n}(\bar X - \mu)/\sigma \sim \mathrm{n}(0,1)$ gives the displayed form. It is increasing in μ, tends to 0 as $\mu \to -\infty$ and to 1 as $\mu \to \infty$, and steepens with $\sqrt{n}$:

μ/σ	$n=1$	$n=4$	$n=16$	$n=64$	$n=100$
-1.0	.0041	.0001	.0000	.0000	.0000
-0.5	.0160	.0041	.0001	.0000	.0000
0	.0500	.0500	.0500	.0500	.0500
0.25	.0815	.1261	.2595	.6387	.8037
0.5	.1261	.2595	.6387	.9907	.9996
1.0	.2595	.6387	.9907	1.0000	1.0000
1.5	.4424	.9123	1.0000	1.0000	1.0000
2.0	.6387	.9907	1.0000	1.0000	1.0000

(b) $\beta(\mu) = \Phi\Big(-1.96 - \dfrac{\sqrt{n}\,\mu}{\sigma}\Big) + 1 - \Phi\Big(1.96 - \dfrac{\sqrt{n}\,\mu}{\sigma}\Big)$.
Here the LRT rejects for $|\bar{x}| > z_{.025}\sigma/\sqrt{n} = 1.96\,\sigma/\sqrt{n}$ (Example 8.2.2). The curve is symmetric about $\mu = 0$, where it attains its minimum .05, and rises to 1 in both tails:

\(	μ	/σ\)	$n=1$	$n=4$	$n=16$	$n=64$	$n=100$
0	.0500	.0500	.0500	.0500	.0500		
0.25	.0572	.0791	.1701	.5160	.7054		
0.5	.0791	.1701	.5160	.9793	.9988		
1.0	.1701	.5160	.9793	1.0000	1.0000		
1.5	.3230	.8508	1.0000	1.0000	1.0000		
2.0	.5160	.9793	1.0000	1.0000	1.0000		

Problem 8.13. Let X_1, X_2 be iid uniform$(\theta, \theta+1)$. For testing $H_0 : \theta = 0$ versus $H_1 : \theta > 0$, we have two competing tests:

$$\phi_1(X_1) \ : \ \text{Reject } H_0 \text{ if } X_1 > .95,$$
$$\phi_2(X_1, X_2) \ : \ \text{Reject } H_0 \text{ if } X_1 + X_2 > C.$$

(a) Find the value of C so that ϕ_2 has the same size as ϕ_1.
(b) Calculate the power function of each test. Draw a well-labeled graph of each power function.
(c) Prove or disprove: ϕ_2 is a more powerful test than ϕ_1.
(d) Show how to get a test that has the same size but is more powerful than ϕ_2.

Solution. (a) $C = 2 - \sqrt{.1} = 1.6838$.
ϕ_1 has size $P_0(X_1 > .95) = .05$. Under H_0, $X_1 + X_2$ has the triangular density on $(0,2)$, so for $1 \le C \le 2$, $P_0(X_1 + X_2 > C) = (2-C)^2/2$; setting this to .05 gives $(2-C)^2 = .1$.
(b) Since $X_1 \sim$ uniform$(\theta, \theta+1)$,

$$\beta_1(\theta) = P_\theta(X_1 > .95) = \begin{cases} \theta + .05, & 0 \le \theta \le .95, \\ 1, & \theta > .95, \end{cases}$$

a straight line of slope 1 from $(0, .05)$ to $(.95, 1)$, flat at 1 thereafter. For ϕ_2, $T = X_1 + X_2 - 2\theta$ is triangular on $(0,2)$ with $P(T > t) = (2-t)^2/2$ for $1 \le t \le 2$ and $1 - t^2/2$ for $0 \le t \le 1$, so with $t = C - 2\theta$,

$$\beta_2(\theta) = \begin{cases} \dfrac{(2 - C + 2\theta)^2}{2}, & 0 \le \theta \le \dfrac{C-1}{2} = .3419, \\ 1 - \dfrac{(C - 2\theta)^2}{2}, & .3419 \le \theta \le \dfrac{C}{2} = .8419, \\ 1, & \theta > .8419. \end{cases}$$

This is a convex parabolic arc from $(0, .05)$ up to $(.3419, .5)$, then a concave arc reaching 1 at $\theta = .8419$, flat at 1 thereafter. Plotting both on the same axes (θ horizontal, power vertical):

θ	0	.05	.1	.2	.3419	.5	.7	.8419	.95
β_1	.05	.100	.150	.250	.392	.550	.750	.892	1.000
β_2	.05	.087	.133	.256	.500	.766	.960	1.000	1.000

The two curves start together at $(0, .05)$, β_1 is above β_2 for small θ, they cross at $\theta = \frac{1}{2}\big[1-2(2-C)\big] = .184$, and β_2 is above thereafter.
(c) Disprove. Both tests have size .05, but at $\theta = .1$

$$\beta_2(.1) = \frac{(.3162 + .2)^2}{2} = .1332 < .15 = \beta_1(.1),$$

so ϕ_2 is not more powerful than ϕ_1.
(d) Take

$$\phi_3(X_1, X_2): \text{ Reject } H_0 \text{ if } X_1 + X_2 > C \text{ or } \max(X_1, X_2) > 1.$$

Under H_0 the sample space is $(0, 1)^2$, so $P_0(\max(X_1, X_2) > 1) = 0$ and ϕ_3 has the same size .05 as ϕ_2. Its rejection region contains that of ϕ_2, so $\beta_3(\theta) \geq \beta_2(\theta)$ for every θ, and the inclusion is strict in probability whenever $0 < \theta < C - 1 = .6838$: then

$$\{1 < X_1 < C - \theta,\ \theta < X_2 < C - X_1\}$$

is a nonempty open subset of the support on which ϕ_3 rejects and ϕ_2 does not. For example $\beta_3(.1) = .240 > .133 = \beta_2(.1)$ and $\beta_3(.5) = .800 > .766 = \beta_2(.5)$.

Problem 8.14. For a random sample $X_1, \ldots, X_n$ of Bernoulli(p) variables, it is desired to test

$$H_0 : p = .49 \qquad \text{versus} \qquad H_1 : p = .51.$$

Use the Central Limit Theorem to determine, approximately, the sample size needed so that the two probabilities of error are both about .01. Use a test function that rejects H_0 if $\sum_{i=1}^n X_i$ is large.

Solution. $n \approx 13{,}525$, rejecting H_0 when $\sum_i X_i > n/2 = 6762.5$.
Let the test reject when $\sum_i X_i > c$. By the CLT, $\sum_i X_i$ is approximately $\mathrm{n}\big(np,\, np(1-p)\big)$, and both p values give the same variance factor $p(1-p) = (.49)(.51) = .2499$. The two error requirements are

$$\frac{c - .49n}{\sqrt{.2499\, n}} = z_{.01} = 2.3263,$$
$$\frac{c - .51n}{\sqrt{.2499\, n}} = -z_{.01} = -2.3263.$$

Adding gives $c = .5n$; subtracting gives

$$\frac{.02n}{\sqrt{.2499\, n}} = 2(2.3263) \quad \Longrightarrow \quad \sqrt{n} = \frac{2(2.3263)\sqrt{.2499}}{.02} = 116.29,$$

so $n = 13{,}524.3$, i.e. $n \approx 13{,}525$ and $c = 6762.5$. (Check: $(6762.5 - .49(13525))/\sqrt{.2499(13525)} = 2.327$, so each error probability is .010.)

8.3 Exercises 8.15–8.21

Problem 8.15. Show that for a random sample $X_1, \ldots, X_n$ from a $\mathrm{n}(0, \sigma^2)$ population, the most powerful test of $H_0 \colon \sigma = \sigma_0$ versus $H_1 \colon \sigma = \sigma_1$, where $\sigma_0 < \sigma_1$, is given by

$$\phi\big(\textstyle\sum X_i^2\big) = \begin{cases} 1 & \text{if } \sum X_i^2 > c \\ 0 & \text{if } \sum X_i^2 \leq c. \end{cases}$$

For a given value of α, the size of the Type I Error, show how the value of c is explicitly determined.

Solution. The likelihood ratio is an increasing function of $\sum X_i^2$, so Neyman–Pearson (Theorem 8.3.12) delivers exactly ϕ, and $c = \sigma_0^2 \chi^2_{n,\alpha}$.
With $f(\mathbf{x} \mid \sigma) = (2\pi\sigma^2)^{-n/2} \exp\bigl(-\sum x_i^2/(2\sigma^2)\bigr)$,

$$\frac{f(\mathbf{x} \mid \sigma_1)}{f(\mathbf{x} \mid \sigma_0)} = \left(\frac{\sigma_0}{\sigma_1}\right)^n \exp\left\{\tfrac{1}{2}\left(\tfrac{1}{\sigma_0^2} - \tfrac{1}{\sigma_1^2}\right)\sum x_i^2\right\}.$$

Since $\sigma_0 < \sigma_1$ the bracket $\sigma_0^{-2} - \sigma_1^{-2}$ is strictly positive, so the ratio is strictly increasing in $\sum x_i^2$. Hence for each $k > 0$

$$\frac{f(\mathbf{x} \mid \sigma_1)}{f(\mathbf{x} \mid \sigma_0)} > k \iff \sum x_i^2 > c(k),$$

so $\{\sum x_i^2 > c\}$ satisfies (8.3.1) with ϕ as its test function, and Theorem 8.3.12(a) (equivalently Corollary 8.3.13 applied to the sufficient statistic $T = \sum X_i^2$) makes it most powerful at its size.
Under H_0, $\sum X_i^2/\sigma_0^2 \sim \chi^2_n$ (sum of squares of n iid $\mathrm{n}(0, \sigma_0^2)$ variables, Lemma 5.3.2), so

$$\alpha = P_{\sigma_0}\Bigl(\sum X_i^2 > c\Bigr) = P\left(\chi^2_n > \frac{c}{\sigma_0^2}\right),$$

and since the χ^2_n cdf is continuous and strictly increasing on $(0, \infty)$ this determines c uniquely: $c = \sigma_0^2 \chi^2_{n,\alpha}$, with $\chi^2_{n,\alpha}$ the upper α point of the χ^2_n distribution.

Problem 8.16. One very striking abuse of α levels is to choose them *after* seeing the data and to choose them in such a way as to force rejection (or acceptance) of a null hypothesis. To see what the *true* Type I and Type II Error probabilities of such a procedure are, calculate size and power of the following two trivial tests:
(a) Always reject H_0, no matter what data are obtained (equivalent to the practice of choosing the α level to force rejection of H_0).
(b) Always accept H_0, no matter what data are obtained (equivalent to the practice of choosing the α level to force acceptance of H_0).

Solution. (a) Size 1, power $\beta(\theta) \equiv 1$. The test function is $\phi \equiv 1$, so $\beta(\theta) = E_\theta \phi(\mathbf{X}) = 1$ for every $\theta \in \Theta$, whence

$$\text{size} = \sup_{\theta \in \Theta_0} \beta(\theta) = 1, \qquad P(\text{Type II Error}) = 1 - \beta(\theta) = 0 \ \text{ for all } \theta \in \Theta_0^c.$$

This is the extreme case flagged after Definition 8.3.11: never a Type II Error, always a Type I Error.
(b) Size 0, power $\beta(\theta) \equiv 0$. Here $\phi \equiv 0$, so

$$\text{size} = \sup_{\theta \in \Theta_0} \beta(\theta) = 0, \qquad P(\text{Type II Error}) = 1 - \beta(\theta) = 1 \ \text{ for all } \theta \in \Theta_0^c.$$

So the true error probabilities are the extremes: forcing rejection costs Type I Error 1, forcing acceptance costs Type II Error 1.

Problem 8.17. Suppose that $X_1, \ldots, X_n$ are iid with a beta$(\mu, 1)$ pdf and $Y_1, \ldots, Y_m$ are iid with a beta$(\theta, 1)$ pdf. Also assume that the Xs are independent of the Ys.
(a) Find an LRT of $H_0\colon \theta = \mu$ versus $H_1\colon \theta \neq \mu$.
(b) Show that the test in part (a) can be based on the statistic

$$T = \frac{\sum \log X_i}{\sum \log X_i + \sum \log Y_i}.$$

(c) Find the distribution of T when H_0 is true, and then show how to get a test of size $\alpha = .10$.

Solution. (a) Reject H_0 when $\lambda \le c$, with λ the statistic displayed below, built from $S = -\sum_{i=1}^{n} \log X_i > 0$ and $U = -\sum_{j=1}^{m} \log Y_j > 0$. Since $f(x \mid a) = ax^{a-1}$ on $(0,1)$, the likelihood is

$$L(\mu, \theta) = \mu^n e^{-(\mu-1)S}\, \theta^m e^{-(\theta-1)U},$$

so $\partial \log L/\partial\mu = n/\mu - S$ and $\partial \log L/\partial\theta = m/\theta - U$ give the unrestricted MLEs $\hat\mu = n/S$, $\hat\theta = m/U$ (the log likelihood is strictly concave in each argument). Under H_0 the common value ν has MLE $\hat\nu = (n+m)/(S+U)$. Substituting,

$$\sup_{\Theta_0} L = \left(\tfrac{n+m}{S+U}\right)^{n+m} e^{-(n+m)} e^{S+U},$$
$$\sup_{\Theta} L = \left(\tfrac{n}{S}\right)^{n} \left(\tfrac{m}{U}\right)^{m} e^{-(n+m)} e^{S+U},$$

so the LRT statistic is

$$\lambda(\mathbf{x}, \mathbf{y}) = \frac{(n+m)^{n+m}}{n^n m^m} \cdot \frac{S^n U^m}{(S+U)^{n+m}}.$$

(b) λ is a function of T alone. Both sums of logs are negative, so $T = (-S)/(-S-U) = S/(S+U) \in (0,1)$, and

$$\frac{S^n U^m}{(S+U)^{n+m}} = T^n(1-T)^m, \qquad \text{so} \qquad \lambda = \frac{(n+m)^{n+m}}{n^n m^m}\, T^n (1-T)^m.$$

The map $t \mapsto t^n(1-t)^m$ increases on $(0, \frac{n}{n+m})$ and decreases on $(\frac{n}{n+m}, 1)$, so $\{\lambda \le c\} = \{T \le t_1\} \cup \{T \ge t_2\}$ for a pair $0 < t_1 < \frac{n}{n+m} < t_2 < 1$ with $t_1^n(1-t_1)^m = t_2^n(1-t_2)^m$.
(c) $T \sim \text{beta}(n, m)$ under H_0. If $X \sim \text{beta}(a, 1)$ then $P(-\log X > t) = P(X < e^{-t}) = e^{-at}$, so $-\log X_i \sim \text{exponential}(1/\nu)$ under H_0, and $\nu(-\log X_i) \sim \text{exponential}(1) = \text{gamma}(1,1)$. Adding independent gammas with a common scale (Example 4.6.8), $\nu S \sim \text{gamma}(n, 1)$ and $\nu U \sim \text{gamma}(m, 1)$, independently. Hence

$$T = \frac{\nu S}{\nu S + \nu U} \sim \text{beta}(n, m)$$

by Exercise 4.24 — and this holds for every common value ν, so T is a pivot on H_0 and the test has the same null distribution throughout Θ_0.
For size .10, let F denote the beta(n, m) cdf and choose $t_1 < t_2$ by

$$t_1^n(1-t_1)^m = t_2^n(1-t_2)^m, \qquad F(t_1) + 1 - F(t_2) = .10,$$

rejecting H_0 when $T \le t_1$ or $T \ge t_2$; this is exactly the LRT with c chosen so that $P_{H_0}(\lambda \le c) = .10$.

Problem 8.18. Let $X_1, \ldots, X_n$ be a random sample from a $\text{n}(\theta, \sigma^2)$ population, σ^2 known. An LRT of $H_0\colon \theta = \theta_0$ versus $H_1\colon \theta \ne \theta_0$ is a test that rejects H_0 if $|\bar X - \theta_0|/(\sigma/\sqrt{n}) > c$.
(a) Find an expression, in terms of standard normal probabilities, for the power function of this test.
(b) The experimenter desires a Type I Error probability of .05 and a maximum Type II Error probability of .25 at $\theta = \theta_0 + \sigma$. Find values of n and c that will achieve this.

Solution. (a) With Φ the standard normal cdf and $d = \sqrt{n}(\theta - \theta_0)/\sigma$,

$$\beta(\theta) = 1 - \Phi(c - d) + \Phi(-c - d).$$

Indeed $\bar X \sim \text{n}(\theta, \sigma^2/n)$, so $Z = (\bar X - \theta)/(\sigma/\sqrt{n}) \sim \text{n}(0,1)$ and $(\bar X - \theta_0)/(\sigma/\sqrt{n}) = Z + d$; hence

$$\begin{aligned}\beta(\theta) &= P\big(|Z + d| > c\big) \\ &= P(Z > c - d) + P(Z < -c - d) \\ &= 1 - \Phi\Big(c - \tfrac{\sqrt{n}(\theta-\theta_0)}{\sigma}\Big) + \Phi\Big(-c - \tfrac{\sqrt{n}(\theta-\theta_0)}{\sigma}\Big).\end{aligned}$$

(b) $c = 1.96$ and $n = 7$. The Type I Error probability is $\beta(\theta_0) = 2\big(1 - \Phi(c)\big) = .05$, so $c = z_{.025} = 1.96$.

At $\theta = \theta_0 + \sigma$ we have $d = \sqrt{n}$, and the requirement is $\beta(\theta_0 + \sigma) \geq .75$, i.e.

$$1 - \Phi(1.96 - \sqrt{n}) + \Phi(-1.96 - \sqrt{n}) \;\geq\; .75.$$

Dropping the (negligible) second term, $\Phi(\sqrt{n} - 1.96) \geq .75$ requires $\sqrt{n} \geq 1.96 + .674 = 2.634$, i.e. $n \geq 6.94$. Checking the two candidate integers in the exact expression,

$$\beta(\theta_0 + \sigma) = \begin{cases} .6878 & n = 6, \\ .7536 & n = 7, \end{cases}$$

so $n = 7$ is the smallest sample size meeting both requirements, with $c = 1.96$.

Problem 8.19. The random variable X has pdf $f(x) = e^{-x}$, $x > 0$. One observation is obtained on the random variable $Y = X^\theta$, and a test of $H_0\colon \theta = 1$ versus $H_1\colon \theta = 2$ needs to be constructed. Find the UMP level $\alpha = .10$ test and compute the Type II Error probability.

Solution. Reject H_0 if $Y < .0765$ or $Y > 3.638$; the Type II Error probability is .610.
Since $y = x^\theta$ is increasing on $x > 0$ with $x = y^{1/\theta}$ and $dx/dy = \theta^{-1} y^{1/\theta - 1}$, Theorem 2.1.5 gives

$$f(y \mid \theta) = \frac{1}{\theta}\, y^{\frac{1}{\theta} - 1} e^{-y^{1/\theta}}, \qquad y > 0,$$

so $f(y \mid 1) = e^{-y}$ and $f(y \mid 2) = \frac{1}{2} y^{-1/2} e^{-\sqrt{y}}$. The likelihood ratio is

$$\frac{f(y \mid 2)}{f(y \mid 1)} = \tfrac{1}{2}\, y^{-1/2} \exp\{y - \sqrt{y}\} =: e^{g(y)}, \qquad g(y) = -\log 2 - \tfrac{1}{2} \log y + y - \sqrt{y}.$$

This ratio is *not* monotone: with $u = \sqrt{y}$,

$$g'(y) = 1 - \frac{1}{2y} - \frac{1}{2\sqrt{y}} = 1 - \frac{1}{2u^2} - \frac{1}{2u},$$

which is negative for $u < 1$, zero at $u = 1$, and positive for $u > 1$. So g decreases on $(0, 1)$, increases on $(1, \infty)$, and $g(y) \to \infty$ at both ends. Both tails of Y therefore favour $\theta = 2$, and the Neyman–Pearson region $\{f(y \mid 2) > k f(y \mid 1)\} = \{g(y) > \log k\}$ has the form

$$R = \{y < a\} \cup \{y > b\}, \qquad 0 < a < 1 < b, \qquad g(a) = g(b) = \log k.$$

Since $Y \sim$ exponential(1) under H_0, the size condition is

$$\alpha = \left(1 - e^{-a}\right) + e^{-b} = .10,$$

and together with $g(a) = g(b)$ this determines the pair uniquely. Solving numerically,

$$a = .0765, \qquad b = 3.638, \qquad k = e^{g(a)} = 1.479,$$

with tail masses $1 - e^{-a} = .0737$ and $e^{-b} = .0263$ summing to .10. (Check!) By Theorem 8.3.12 this test is most powerful, and since H_1 is simple it is UMP level .10.
For the Type II Error, under H_1 we have $Y = X^2$, so $\{a \leq Y \leq b\} = \{\sqrt{a} \leq X \leq \sqrt{b}\}$ and

$$P(\text{Type II}) = e^{-\sqrt{a}} - e^{-\sqrt{b}} = e^{-.2767} - e^{-1.9073} = .7583 - .1485 = .610.$$

Problem 8.20. Let X be a random variable whose pmf under H_0 and H_1 is given by

x	1	2	3	4	5	6	7
$f(x \mid H_0)$	.01	.01	.01	.01	.01	.01	.94
$f(x \mid H_1)$	.06	.05	.04	.03	.02	.01	.79

Use the Neyman–Pearson Lemma to find the most powerful test for H_0 versus H_1 with size $\alpha = .04$. Compute the probability of Type II Error for this test.

Solution. Reject H_0 if $x \in \{1, 2, 3, 4\}$; the Type II Error probability is .82.
Tabulate the likelihood ratio $r(x) = f(x \mid H_1)/f(x \mid H_0)$:

x	1	2	3	4	5	6	7
$r(x)$	6	5	4	3	2	1	.840

Theorem 8.3.12 says the most powerful test puts into R every x with $r(x) > k$ and excludes every x with $r(x) < k$. Any $k \in (2, 3)$ forces exactly

$$R = \{1, 2, 3, 4\}, \qquad \alpha = P(X \in R \mid H_0) = .01 + .01 + .01 + .01 = .04,$$

so this R satisfies (8.3.1) and (8.3.2) and is MP of size exactly .04 by Theorem 8.3.12(a).
The Type II Error probability is

$$P(X \in R^c \mid H_1) = f(5 \mid H_1) + f(6 \mid H_1) + f(7 \mid H_1) = .02 + .01 + .79 = .82.$$

Problem 8.21. In the proof of Theorem 8.3.12 (Neyman–Pearson Lemma), it was stated that the proof, which was given for continuous random variables, can easily be adapted to cover discrete random variables. Provide the details; that is, prove the Neyman–Pearson Lemma for discrete random variables. Assume that the α level is attainable.

Solution. Replace $\int \cdots d\mathbf{x}$ by $\sum_{\mathbf{x}\in\mathcal{X}}$ throughout; every step of the printed proof (p. 389) survives verbatim, and the final "except on a null set" caveat disappears.
Let $\mathcal{X}$ be the (countable) sample space, $f(\mathbf{x} \mid \theta_i)$ the pmfs, and let R satisfy (8.3.1) and (8.3.2):

$$\begin{aligned} &\mathbf{x} \in R \text{ if } f(\mathbf{x} \mid \theta_1) > k f(\mathbf{x} \mid \theta_0), \\ &\mathbf{x} \in R^c \text{ if } f(\mathbf{x} \mid \theta_1) < k f(\mathbf{x} \mid \theta_0), \end{aligned} \qquad \alpha = P_{\theta_0}(\mathbf{X} \in R),$$

for some $k \geq 0$. Write $\phi = I_R$ for its test function and let ϕ' be the test function of any level α test, with power functions

$$\beta(\theta) = \sum_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}) f(\mathbf{x} \mid \theta), \qquad \beta'(\theta) = \sum_{\mathbf{x}\in\mathcal{X}} \phi'(\mathbf{x}) f(\mathbf{x} \mid \theta).$$

Since $0 \leq \phi' \leq 1$, and since (8.3.1) makes $\phi(\mathbf{x}) = 1$ wherever $f(\mathbf{x} \mid \theta_1) - k f(\mathbf{x} \mid \theta_0) > 0$ and $\phi(\mathbf{x}) = 0$ wherever it is negative, every term of the sum below is nonnegative:

$$\big(\phi(\mathbf{x}) - \phi'(\mathbf{x})\big)\big(f(\mathbf{x} \mid \theta_1) - k f(\mathbf{x} \mid \theta_0)\big) \geq 0 \qquad \text{for every } \mathbf{x} \in \mathcal{X}.$$

Summing over $\mathcal{X}$ (the two pmfs are summable, so the sum splits) gives the discrete form of (8.3.3):

$$\begin{aligned} 0 &\leq \sum_{\mathbf{x}} \big[\phi(\mathbf{x}) - \phi'(\mathbf{x})\big]\big[f(\mathbf{x} \mid \theta_1) - k f(\mathbf{x} \mid \theta_0)\big] \\ &= \beta(\theta_1) - \beta'(\theta_1) - k\big(\beta(\theta_0) - \beta'(\theta_0)\big). \end{aligned}$$

Sufficiency. ϕ has size α by (8.3.2) and ϕ' has level α, so $\beta(\theta_0) - \beta'(\theta_0) = \alpha - \beta'(\theta_0) \geq 0$. With $k \geq 0$,

$$0 \leq \beta(\theta_1) - \beta'(\theta_1) - k\big(\beta(\theta_0) - \beta'(\theta_0)\big) \leq \beta(\theta_1) - \beta'(\theta_1),$$

so $\beta(\theta_1) \geq \beta'(\theta_1)$. As θ_1 is the only point of Θ_0^c and ϕ' was an arbitrary level α test, ϕ is UMP level α.
Necessity. Assume α is attainable, so a ϕ as above with $k > 0$ exists, and let ϕ' now be any UMP level α test. By sufficiency ϕ is also UMP level α, hence $\beta(\theta_1) = \beta'(\theta_1)$, and the display above becomes

$$0 \leq -k\big(\beta(\theta_0) - \beta'(\theta_0)\big) = -k\big(\alpha - \beta'(\theta_0)\big),$$

so $\beta'(\theta_0) \geq \alpha$; combined with $\beta'(\theta_0) \leq \alpha$ (level α) this gives $\beta'(\theta_0) = \alpha$, i.e. every UMP level α test has size α, which is (8.3.2).
Consequently the inequality is an equality: a sum of nonnegative terms vanishes only if *each* term vanishes, so for every $\mathbf{x} \in \mathcal{X}$,

$$\big(\phi(\mathbf{x}) - \phi'(\mathbf{x})\big)\big(f(\mathbf{x} \mid \theta_1) - k f(\mathbf{x} \mid \theta_0)\big) = 0.$$

At each $\mathbf{x}$, then, either $f(\mathbf{x} \mid \theta_1) = k f(\mathbf{x} \mid \theta_0)$ — where (8.3.1) imposes no requirement — or $\phi'(\mathbf{x}) = \phi(\mathbf{x})$, and ϕ' inherits (8.3.1) there from ϕ. So ϕ' satisfies (8.3.1) at *every* sample point and the exceptional set A of Theorem 8.3.12(b) may be taken empty.

8.4 Exercises 8.22–8.28

Problem 8.22. Let $X_1, \ldots, X_{10}$ be iid Bernoulli(p).
(a) Find the most powerful test of size $\alpha = .0547$ of the hypotheses $H_0 : p = \frac{1}{2}$ versus $H_1 : p = \frac{1}{4}$. Find the power of this test.
(b) For testing $H_0 : p \leq \frac{1}{2}$ versus $H_1 : p > \frac{1}{2}$, find the size and sketch the power function of the test that rejects H_0 if $\sum_{i=1}^{10} X_i \geq 6$.
(c) For what α levels does there exist a UMP test of the hypotheses in part (a)?

Solution. (a) Reject H_0 if $Y = \sum_{i=1}^{10} X_i \leq 2$; its power is $137781/262144 \approx .5256$.
Indeed $Y \sim \text{binomial}(10, p)$ is sufficient, and the likelihood ratio is

$$\frac{f(y \mid \frac{1}{4})}{f(y \mid \frac{1}{2})} = \frac{\binom{10}{y}(\frac{1}{4})^y(\frac{3}{4})^{10-y}}{\binom{10}{y}(\frac{1}{2})^{10}} = \frac{3^{10-y}}{2^{10}},$$

strictly decreasing in y. So by the Neyman–Pearson Lemma (Theorem 8.3.12) a test rejecting for large values of this ratio is a test rejecting for small y, and $\{y \leq k\}$ is the rejection region. The size requirement picks k:

$$P\Big(Y \leq 2 \,\Big|\, p = \tfrac{1}{2}\Big) = \frac{1 + 10 + 45}{2^{10}} = \frac{56}{1024} = .0547.$$

Hence $k = 2$, and

$$\text{power} = P\Big(Y \leq 2 \,\Big|\, p = \tfrac{1}{4}\Big) = \sum_{y=0}^{2} \binom{10}{y} \frac{3^{10-y}}{4^{10}} = \frac{3^8(9 + 30 + 45)}{4^{10}} = \frac{137781}{262144} \approx .5256.$$

(b) The power function is $\beta(p) = P(Y \geq 6 \mid p) = \sum_{y=6}^{10} \binom{10}{y} p^y (1-p)^{10-y}$, which is increasing in p (the binomial family has MLR in Y, Exercise 8.25(c)), so

$$\text{size} = \sup_{p \leq 1/2} \beta(p) = \beta(\tfrac{1}{2}) = \frac{210 + 120 + 45 + 10 + 1}{1024} = \frac{193}{512} \approx .3770.$$

The sketch is the increasing S-shaped curve through

p	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
$\beta(p)$	0	.0001	.0064	.0473	.1662	.3770	.6331	.8497	.9672	.9984	1

(c) Exactly $\alpha = 0$ and the attainable tail probabilities $P(Y \leq k \mid p = \frac{1}{2})$ for $k = 0, 1, \ldots, 10$, namely

$$\alpha \in \{0,\ .0010,\ .0107,\ .0547,\ .1719,\ .3770,\ .6230,\ .8281,\ .9453,\ .9893,\ .9990,\ 1\},$$

since for these and only these values does a nonrandomized Neyman–Pearson test have size exactly α; the hypotheses being simple versus simple, most powerful and UMP coincide here.

Problem 8.23. Suppose X is one observation from a population with beta$(\theta, 1)$ pdf.
(a) For testing $H_0 : \theta \le 1$ versus $H_1 : \theta > 1$, find the size and sketch the power function of the test that rejects H_0 if $X > \frac{1}{2}$.
(b) Find the most powerful level α test of $H_0 : \theta = 1$ versus $H_1 : \theta = 2$.
(c) Is there a UMP test of $H_0 : \theta \le 1$ versus $H_1 : \theta > 1$? If so, find it. If not, prove so.

Solution. (a) Size $= \frac{1}{2}$. With $f(x \mid \theta) = \theta x^{\theta-1}$ on $(0,1)$, so $F(x \mid \theta) = x^\theta$,

$$\begin{aligned}\beta(\theta) = P\Big(X > \tfrac{1}{2} \,\Big|\, \theta\Big) &= 1 - \Big(\tfrac{1}{2}\Big)^\theta \\ &= 1 - 2^{-\theta},\end{aligned}$$

which increases from 0 to 1; hence $\sup_{\theta \le 1} \beta(\theta) = \beta(1) = \frac{1}{2}$. The sketch is the increasing curve through

θ	0	$\frac{1}{2}$	1	2	3	4
$\beta(\theta)$	0	.2929	.5	.75	.875	.9375

(b) Reject H_0 if $X > 1 - \alpha$. The likelihood ratio

$$\frac{f(x \mid 2)}{f(x \mid 1)} = \frac{2x}{1} = 2x$$

is increasing in x, so by the Neyman–Pearson Lemma (Theorem 8.3.12) the most powerful test rejects for large x, say $X > c$; since $X \sim$ uniform$(0,1)$ under H_0, the size condition $1 - c = \alpha$ gives $c = 1 - \alpha$.
(c) Yes: the test of part (b), rejecting when $X > 1 - \alpha$. For $\theta_2 > \theta_1$,

$$\frac{f(x \mid \theta_2)}{f(x \mid \theta_1)} = \frac{\theta_2}{\theta_1}\, x^{\theta_2 - \theta_1}$$

is increasing in x on $(0,1)$, so the family has MLR in X. By the Karlin–Rubin Theorem (Theorem 8.3.17) the test rejecting when $X > c$ is UMP level α for $H_0 : \theta \le 1$ versus $H_1 : \theta > 1$, where $\alpha = P(X > c \mid \theta = 1) = 1 - c$, i.e. $c = 1 - \alpha$.

Problem 8.24. Find the LRT of a **simple** H_0 versus a **simple** H_1. Is this test equivalent to the one obtained from the Neyman–Pearson Lemma? (This relationship is treated in some detail by Solomon 1975.)

Solution. The LRT rejects when $f(\mathbf{x} \mid \theta_1)/f(\mathbf{x} \mid \theta_0) \ge k$, which is exactly the Neyman–Pearson test; so yes, they are equivalent.
Here $\Theta = \{\theta_0, \theta_1\}$ with $\Theta_0 = \{\theta_0\}$, so by Definition 8.2.1

$$\lambda(\mathbf{x}) = \frac{f(\mathbf{x} \mid \theta_0)}{\max\{f(\mathbf{x} \mid \theta_0),\, f(\mathbf{x} \mid \theta_1)\}} = \min\left\{1,\ \frac{f(\mathbf{x} \mid \theta_0)}{f(\mathbf{x} \mid \theta_1)}\right\},$$

and the LRT rejects H_0 when $\lambda(\mathbf{x}) \le c$ for some $c \in [0,1]$. For $c < 1$ the minimum can only be attained by the second entry, so

$$\lambda(\mathbf{x}) \le c \iff \frac{f(\mathbf{x} \mid \theta_0)}{f(\mathbf{x} \mid \theta_1)} \le c \iff \frac{f(\mathbf{x} \mid \theta_1)}{f(\mathbf{x} \mid \theta_0)} \ge \frac{1}{c},$$

which is the rejection region of Theorem 8.3.12 with $k = 1/c$. As c runs over $[0,1)$ the constant $k = 1/c$ runs over $(1, \infty)$, and conversely every Neyman–Pearson test with $k > 1$ is the LRT with $c = 1/k$; the remaining value $c = 1$ rejects everywhere, matching $k = 0$.

Problem 8.25. Show that each of the following families has an MLR.
(a) $n(\theta, \sigma^2)$ family with σ^2 known
(b) Poisson(θ) family
(c) binomial(n, θ) family with n known

Solution. Each is a one-parameter exponential family whose natural parameter $w(\theta)$ is increasing, so Exercise 8.27 gives the MLR at once; here are the ratios, for $\theta_2 > \theta_1$, increasing in x as Definition 8.3.16 requires.
(a)

$$\frac{f(x \mid \theta_2)}{f(x \mid \theta_1)} = \exp\left\{\frac{(\theta_2 - \theta_1)x}{\sigma^2} - \frac{\theta_2^2 - \theta_1^2}{2\sigma^2}\right\},$$

increasing in x because $\theta_2 - \theta_1 > 0$.
(b) For $x = 0, 1, 2, \ldots$,

$$\frac{f(x \mid \theta_2)}{f(x \mid \theta_1)} = e^{-(\theta_2 - \theta_1)} \left(\frac{\theta_2}{\theta_1}\right)^x,$$

increasing in x because $\theta_2/\theta_1 > 1$.
(c) For $x = 0, 1, \ldots, n$,

$$\frac{f(x \mid \theta_2)}{f(x \mid \theta_1)} = \left(\frac{\theta_2/(1-\theta_2)}{\theta_1/(1-\theta_1)}\right)^x \left(\frac{1-\theta_2}{1-\theta_1}\right)^n,$$

and $\theta/(1-\theta)$ is increasing on $(0, 1)$, so the base of the xth power exceeds 1 and the ratio is increasing in x.

Problem 8.26. (a) Show that if a family of pdfs $\{f(x \mid \theta) : \theta \in \Theta\}$ has an MLR, then the corresponding family of cdfs is **stochastically increasing** in θ. (See the Miscellanea section.)
(b) Show that the converse of part (a) is false; that is, give an example of a family of cdfs that is stochastically increasing in θ for which the corresponding family of pdfs does not have an MLR.

Solution. (a) MLR forces $F(x_0 \mid \theta_2) \le F(x_0 \mid \theta_1)$ for all x_0 and all $\theta_1 < \theta_2$, which is stochastic ordering. Indeed Definition 8.3.16 says $f(x \mid \theta_2)/f(x \mid \theta_1)$ is nondecreasing in x, which in cross-multiplied form is

$$f(x \mid \theta_2)\, f(y \mid \theta_1) \;\le\; f(x \mid \theta_1)\, f(y \mid \theta_2) \qquad \text{for all } x \le y$$

(if $f(x \mid \theta_1) = 0 < f(x \mid \theta_2)$ the ratio is $+\infty$ at x, hence at every $y \ge x$, forcing $f(y \mid \theta_1) = 0$ and the left side to vanish). Integrate this over $x \le x_0$ and $y > x_0$:

$$F(x_0 \mid \theta_2)\big(1 - F(x_0 \mid \theta_1)\big) \;\le\; F(x_0 \mid \theta_1)\big(1 - F(x_0 \mid \theta_2)\big).$$

The cross terms $F(x_0 \mid \theta_1)F(x_0 \mid \theta_2)$ cancel, leaving $F(x_0 \mid \theta_2) \le F(x_0 \mid \theta_1)$ for every x_0. (Sums in place of integrals give the discrete case.)
(b) Take the Cauchy location family

$$f(x \mid \theta) = \frac{1}{\pi\big(1 + (x - \theta)^2\big)}, \qquad -\infty < x, \theta < \infty.$$

It is a location family, so $F(x \mid \theta) = F_0(x - \theta)$ is decreasing in θ for each fixed x: stochastically increasing. But it has no MLR (Exercise 8.29(a)): for $\theta_1 = 0$, $\theta_2 = 1$,

$$\frac{f(x \mid 1)}{f(x \mid 0)} = \frac{1 + x^2}{1 + (x-1)^2}$$

equals $\frac{1}{2}$ at $x = 0$, equals 2 at $x = 1$, and tends to 1 as $x \to \infty$, so it is not nondecreasing in x.

Problem 8.27. Suppose $g(t \mid \theta) = h(t)c(\theta)e^{w(\theta)t}$ is a one-parameter exponential family for the random variable T. Show that this family has an MLR if $w(\theta)$ is an increasing function of θ. Give three examples of such a family.

Solution. For $\theta_2 > \theta_1$ the factor $h(t)$ cancels and

$$\frac{g(t \mid \theta_2)}{g(t \mid \theta_1)} = \frac{c(\theta_2)}{c(\theta_1)} \exp\big\{\big(w(\theta_2) - w(\theta_1)\big)t\big\},$$

a positive constant times e^{at} with $a = w(\theta_2) - w(\theta_1) > 0$ because w is increasing. Hence the ratio is increasing in t on $\{t : h(t) > 0\}$, which is the MLR property (Definition 8.3.16).
Three examples:
(i) $T \sim n(\theta, \sigma^2)$, σ^2 known: $h(t) = e^{-t^2/(2\sigma^2)}/\sqrt{2\pi\sigma^2}$, $c(\theta) = e^{-\theta^2/(2\sigma^2)}$, $w(\theta) = \theta/\sigma^2$, increasing.
(ii) $T \sim \text{Poisson}(\theta)$: $h(t) = 1/t!$, $c(\theta) = e^{-\theta}$, $w(\theta) = \log\theta$, increasing on $\theta > 0$.
(iii) $T \sim \text{binomial}(n, \theta)$, n known: $h(t) = \binom{n}{t}$, $c(\theta) = (1-\theta)^n$, $w(\theta) = \log\big(\theta/(1-\theta)\big)$, increasing on $0 < \theta < 1$.

Problem 8.28. Let $f(x \mid \theta)$ be the logistic location pdf

$$f(x \mid \theta) = \frac{e^{(x-\theta)}}{\big(1 + e^{(x-\theta)}\big)^2}, \qquad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

(a) Show that this family has an MLR.
(b) Based on one observation, X, find the most powerful size α test of $H_0 : \theta = 0$ versus $H_1 : \theta = 1$. For $\alpha = .2$, find the size of the Type II Error.
(c) Show that the test in part (b) is UMP size α for testing $H_0 : \theta \le 0$ versus $H_1 : \theta > 0$. What can be said about UMP tests in general for the logistic location family?

Solution. (a) For $\theta_2 > \theta_1$,

$$\frac{f(x \mid \theta_2)}{f(x \mid \theta_1)} = e^{\theta_1 - \theta_2}\left(\frac{1 + e^{x-\theta_1}}{1 + e^{x-\theta_2}}\right)^2,$$

and writing $F_0(u) = e^u/(1+e^u)$ (the standard logistic cdf, which is increasing),

$$\frac{d}{dx}\log\frac{f(x \mid \theta_2)}{f(x \mid \theta_1)} = 2\big[F_0(x-\theta_1) - F_0(x-\theta_2)\big] > 0$$

since $x - \theta_1 > x - \theta_2$. So the ratio is increasing in x: MLR in X.
(b) Reject H_0 if $X > \log\frac{1-\alpha}{\alpha}$. By (a) the ratio $f(x \mid 1)/f(x \mid 0)$ is increasing in x, so the Neyman–Pearson region $\{f(x \mid 1)/f(x \mid 0) \ge k\}$ is an upper tail $\{X > c\}$, and since $F(x \mid \theta) = e^{x-\theta}/(1+e^{x-\theta})$,

$$\alpha = P(X > c \mid \theta = 0) = \frac{1}{1+e^c} \quad \Longrightarrow \quad c = \log\frac{1-\alpha}{\alpha}.$$

For $\alpha = .2$, $c = \log 4$ and

$$P(\text{Type II Error}) = P(X \le \log 4 \mid \theta = 1) = \frac{e^{\log 4 - 1}}{1 + e^{\log 4 - 1}}$$
$$= \frac{4}{4+e} \approx .5954.$$

(c) By part (a) the family has MLR in X, so the Karlin–Rubin Theorem (Theorem 8.3.17) applies directly: the test rejecting H_0 when $X > c$ is UMP level α for $H_0 : \theta \le 0$ versus $H_1 : \theta > 0$, where $\alpha = P(X > c \mid \theta = 0)$ – the same $c = \log\frac{1-\alpha}{\alpha}$ as in (b), and $\sup_{\theta \le 0}\beta(\theta) = \beta(0) = \alpha$ because β is increasing. In general, for one observation from the logistic location family and any θ_0, the test rejecting when

$$X > \theta_0 + \log\frac{1-\alpha}{\alpha}$$

is UMP level α for $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$, and the mirror-image lower-tail test is UMP for the opposite one-sided problem. No UMP test exists against $H_1 : \theta \neq \theta_0$: by (a) the most powerful test against $\theta_1 < \theta_0$ rejects for small X, so no single region is most powerful on both sides.

8.5 Exercises 8.29–8.35

Problem 8.29. Let X be one observation from a Cauchy(θ) distribution, that is, with pdf

$$f(x \mid \theta) = \frac{1}{\pi}\,\frac{1}{1+(x-\theta)^2}, \qquad -\infty < x < \infty,\ -\infty < \theta < \infty.$$

(a) Show that this family does not have an MLR.
(b) Show that the test

$$\phi(x) = \begin{cases} 1 & \text{if } 1 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

is most powerful of its size for testing $H_0 : \theta = 0$ versus $H_1 : \theta = 1$. Calculate the Type I and Type II Error probabilities.
(c) Prove or disprove: The test in part (b) is UMP for testing $H_0 : \theta \le 0$ versus $H_1 : \theta > 0$. What can be said about UMP tests in general for the Cauchy location family?

Solution. (a) No: for $\theta_1 = 0 < \theta_2 = 1$ the ratio

$$g(x) = \frac{f(x \mid 1)}{f(x \mid 0)} = \frac{1+x^2}{1+(x-1)^2}$$

satisfies $g(0) = \frac{1}{2}$, $g(1) = 2$, and $g(x) \to 1$ as $x \to \pm\infty$. A function that starts at 1, dips to $\frac{1}{2}$, rises to 2 and returns to 1 is not monotone, so Definition 8.3.16 fails.
(b) Take $k = 2$ in the Neyman–Pearson Lemma. Since

$$(1+x^2) - 2\big(1+(x-1)^2\big) = -x^2 + 4x - 3 = -(x-1)(x-3),$$

we get

$$f(x \mid 1) > 2f(x \mid 0) \iff 1 < x < 3, \qquad f(x \mid 1) < 2f(x \mid 0) \iff x \notin [1,3].$$

Thus ϕ rejects exactly where $f(x \mid 1) > 2f(x \mid 0)$ and accepts exactly where $f(x \mid 1) < 2f(x \mid 0)$ (the boundary $\{1,3\}$ is null), which is (8.3.1), so ϕ is MP of its size by Theorem 8.3.12a. The error probabilities are

$$\alpha = P_0(1 < X < 3) = \frac{\tan^{-1} 3 - \tan^{-1} 1}{\pi} = .1476,$$

$$\beta = 1 - P_1(1 < X < 3) = 1 - \frac{\tan^{-1} 2}{\pi} = .6476.$$

(c) Disprove. Because $\tan^{-1} 2 + \tan^{-1} 3 = \frac{3\pi}{4}$, the test $\phi^*(x) = I(x > 2)$ has exactly the same size,

$$P_0(X > 2) = \tfrac{1}{2} - \frac{\tan^{-1} 2}{\pi} = \frac{\tan^{-1} 3 - \tan^{-1} 1}{\pi} = \alpha,$$

but its power $P_\theta(X > 2) = \frac{1}{2} + \pi^{-1}\tan^{-1}(\theta - 2) \to 1$ as $\theta \to \infty$, whereas

$$\beta_\phi(\theta) = \frac{\tan^{-1}(3-\theta) - \tan^{-1}(1-\theta)}{\pi} \longrightarrow 0.$$

So ϕ^* beats ϕ at all large θ, and ϕ is not UMP.
Indeed no UMP test of this size exists. Both β_ϕ and β_{ϕ^*} are largest at $\theta = 0$ over $\theta \le 0$, so both are size-α tests of the composite H_0. If ψ were UMP size α for it, then $\beta_\psi(1) \ge \beta_\phi(1)$, while ψ is also level α for the simple $H_0 : \theta = 0$ and so $\beta_\psi(1) \le \beta_\phi(1)$ by part (b); hence ψ is MP level α there, and the necessity half of the Neyman–Pearson Lemma (Theorem 8.3.12b, whose hypothesis holds because

ϕ satisfies (8.3.1) and (8.3.2) with $k = 2 > 0$) forces $\psi = \phi$ off the null set $\{1, 3\}$ – which ϕ^* beats. The same obstruction occurs at every level: against θ_1 the MP test rejects on

$$\Big\{x : \ \frac{1+x^2}{1+(x-\theta_1)^2} > k\Big\} = \Big\{x : \ (k-1)x^2 - 2k\theta_1 x + k(1+\theta_1^2) < 1\Big\},$$

a set delimited by the roots of a quadratic that moves with θ_1, while Theorem 8.3.12b makes each of these tests essentially unique. So the Cauchy location family admits no UMP test of $H_0 : \theta \le 0$ versus $H_1 : \theta > 0$.

Problem 8.30. Let $f(x \mid \theta)$ be the Cauchy *scale* pdf

$$f(x \mid \theta) = \frac{\theta}{\pi}\,\frac{1}{\theta^2 + x^2}, \qquad -\infty < x < \infty, \ \theta > 0.$$

(a) Show that this family does not have an MLR.
(b) If X is one observation from $f(x \mid \theta)$, show that $|X|$ is sufficient for θ and that the distribution of $|X|$ *does* have an MLR.

Solution. (a) No: the ratio depends on x only through x^2, so it is an even function of x and cannot be monotone. Explicitly, for $\theta_2 > \theta_1$,

$$r(x) = \frac{f(x \mid \theta_2)}{f(x \mid \theta_1)} = \frac{\theta_2}{\theta_1}\cdot\frac{\theta_1^2 + x^2}{\theta_2^2 + x^2}, \qquad r(0) = \frac{\theta_1}{\theta_2} < 1 < \frac{\theta_2}{\theta_1} = \lim_{|x|\to\infty} r(x),$$

so r decreases on $(-\infty, 0)$ and increases on $(0, \infty)$; Definition 8.3.16 fails.
(b) $f(x \mid \theta) = \dfrac{\theta}{\pi}\dfrac{1}{\theta^2 + |x|^2} = g\big(|x| \ \big|\ \theta\big)\cdot 1$, so by the Factorization Theorem (Theorem 6.2.6) $|X|$ is sufficient for θ. By symmetry of f about 0, $Y = |X|$ has pdf

$$f_Y(y \mid \theta) = \frac{2\theta}{\pi}\,\frac{1}{\theta^2 + y^2}, \qquad y > 0.$$

For $\theta_2 > \theta_1$,

$$\frac{f_Y(y \mid \theta_2)}{f_Y(y \mid \theta_1)} = \frac{\theta_2}{\theta_1}\cdot\frac{\theta_1^2 + y^2}{\theta_2^2 + y^2}, \qquad \frac{d}{dy}\frac{\theta_1^2 + y^2}{\theta_2^2 + y^2} = \frac{2y(\theta_2^2 - \theta_1^2)}{(\theta_2^2 + y^2)^2} > 0$$

for all $y > 0$. The ratio is increasing in y, so the family of $|X|$ has an MLR.

Problem 8.31. Let $X_1, \ldots, X_n$ be iid Poisson(λ).
(a) Find a UMP test of $H_0 : \lambda \le \lambda_0$ versus $H_1 : \lambda > \lambda_0$.
(b) Consider the specific case $H_0 : \lambda \le 1$ versus $H_1 : \lambda > 1$. Use the Central Limit Theorem to determine the sample size n so a UMP test satisfies $P(\text{reject } H_0 \mid \lambda = 1) = .05$ and $P(\text{reject } H_0 \mid \lambda = 2) = .9$.

Solution. (a) Reject H_0 if $T = \sum_{i=1}^n X_i > k$. Indeed

$$f(\mathbf{x} \mid \lambda) = \frac{1}{\prod x_i!}\, e^{-n\lambda}\exp\Big\{(\log\lambda)\sum_{i=1}^n x_i\Big\}$$

is an exponential family in the sufficient statistic T, whose pmf Poisson($n\lambda$) is again of the form $h(t)c(\lambda)e^{w(\lambda)t}$ with $w(\lambda) = \log\lambda$ increasing, so T has an MLR by Exercise 8.27. Karlin–Rubin (Theorem 8.3.17) then makes the test UMP level α for $\alpha = P_{\lambda_0}(T > k)$. Being discrete, T attains only these

sizes; an arbitrary α requires the randomized version

$$\phi(t) = \begin{cases} 1, & t > k \\ \gamma, & t = k \\ 0, & t < k \end{cases} \qquad \text{with} \qquad P_{\lambda_0}(T > k) + \gamma P_{\lambda_0}(T = k) = \alpha.$$

(b) $T \sim \text{Poisson}(n\lambda)$ has mean and variance $n\lambda$, so the CLT gives

$$\frac{k - n}{\sqrt{n}} = z_{.05} = 1.645, \qquad \frac{k - 2n}{\sqrt{2n}} = -z_{.10} = -1.282.$$

Eliminating k from $k = n + 1.645\sqrt{n}$ and $k = 2n - 1.282\sqrt{2n} = 2n - 1.813\sqrt{n}$,

$$n = (1.645 + 1.813)\sqrt{n} = 3.458\sqrt{n} \quad \Longrightarrow \quad \sqrt{n} = 3.458, \quad n = 11.96.$$

Take $n = 12$, whence $k = 12 + 1.645\sqrt{12} = 17.7$: reject H_0 if $\sum X_i \geq 18$.

Problem 8.32. Let $X_1, \ldots, X_n$ be iid $\text{n}(\theta, 1)$, and let θ_0 be a specified value of θ.
(a) Find the UMP, size α, test of $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$.
(b) Show that there does not exist a UMP, size α, test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

Solution. (a) Reject H_0 if $\bar{X} < \theta_0 - z_\alpha/\sqrt{n}$. By Exercise 8.25(a) the $\text{n}(\theta, 1)$ family has an MLR in the sufficient statistic $\bar{X}$ (Theorem 6.2.10), so $T = -\bar{X}$ has an MLR in $\eta = -\theta$, and in these parameters the problem reads $H_0 : \eta \leq -\theta_0$ versus $H_1 : \eta > -\theta_0$, which is Karlin–Rubin form (Theorem 8.3.17). Rejecting for large T, i.e. for small $\bar{X}$, is therefore UMP level α with

$$\alpha = P_{\theta_0}\Big(\bar{X} < \theta_0 - \tfrac{z_\alpha}{\sqrt{n}}\Big) = P(Z < -z_\alpha),$$

since $\bar{X} \sim \text{n}(\theta_0, 1/n)$ there. The power function is decreasing in θ, so the supremum over $\theta \geq \theta_0$ is attained at θ_0 and the size is α.
(b) Fix $\theta_1 > \theta_0$ and $\theta_2 < \theta_0$. The likelihood ratio for the simple problem θ_0 versus θ_1 is

$$\frac{f(\mathbf{x} \mid \theta_1)}{f(\mathbf{x} \mid \theta_0)} = \exp\Big\{n(\theta_1 - \theta_0)\bar{x} - \tfrac{n}{2}(\theta_1^2 - \theta_0^2)\Big\},$$

increasing in $\bar{x}$, so the size-α test satisfying (8.3.1) and (8.3.2) is $\phi_+ = I\big(\bar{X} > \theta_0 + z_\alpha/\sqrt{n}\big)$, with $k > 0$; symmetrically, against θ_2 it is $\phi_- = I\big(\bar{X} < \theta_0 - z_\alpha/\sqrt{n}\big)$. Neither depends on which θ_1 or θ_2 was chosen.
Now suppose ϕ were UMP size α for $H_1 : \theta \neq \theta_0$. Then ϕ is in particular UMP level α for θ_0 versus θ_1, so by the necessity half of the Neyman–Pearson Lemma (Theorem 8.3.12b, whose hypothesis $k > 0$ was just checked) $\phi = \phi_+$ a.e.; the same argument at θ_2 gives $\phi = \phi_-$ a.e. But

$$\beta_{\phi_+}(\theta_2) < \alpha < \beta_{\phi_-}(\theta_2)$$

because the power function of ϕ_+ increases and that of ϕ_- decreases in θ, both equalling α at θ_0. So $\phi_+ \neq \phi_-$ on a set of positive probability, a contradiction.

Problem 8.33. Let $X_1, \ldots, X_n$ be a random sample from the uniform$(\theta, \theta + 1)$ distribution. To test $H_0 : \theta = 0$ versus $H_1 : \theta > 0$, use the test

$$\text{reject } H_0 \text{ if } Y_n \geq 1 \text{ or } Y_1 \geq k,$$

where k is a constant, $Y_1 = \min\{X_1, \ldots, X_n\}$, $Y_n = \max\{X_1, \ldots, X_n\}$.
(a) Determine k so that the test will have size α.

(b) Find an expression for the power function of the test in part (a).
(c) Prove that the test is UMP size α.
(d) Find values of n and k so that the UMP .10 level test will have power at least .8 if $\theta > 1$.

Solution. (a) $k = 1 - \alpha^{1/n}$. Under $\theta = 0$ the X_i are iid uniform$(0, 1)$, so $P_0(Y_n \geq 1) = 0$ and the size is

$$P_0(Y_1 \geq k) = P_0(X_1 \geq k, \ldots, X_n \geq k) = (1 - k)^n = \alpha.$$

(b) Write $\beta(\theta)$ for the power. For $\theta \geq k$ every $X_i \geq \theta \geq k$, so $Y_1 \geq k$ with probability one and $\beta(\theta) = 1$. For $0 \leq \theta < k$ (note $k < 1$) the acceptance event is $\{Y_n < 1\} \cap \{Y_1 < k\} = \{Y_n < 1\} \setminus \{\text{all } X_i \in [k, 1)\}$, and since $(\theta, \theta + 1) \supset [k, 1)$,

$$P_\theta(\text{accept}) = (1 - \theta)^n - (1 - k)^n.$$

Hence

$$\beta(\theta) = \begin{cases} 1 - (1 - \theta)^n + \alpha, & 0 \leq \theta < k, \\ 1, & \theta \geq k, \end{cases} \qquad k = 1 - \alpha^{1/n},$$

which is continuous (both branches give 1 at $\theta = k$) and increasing, with $\beta(0) = \alpha$.
(c) Fix $\theta_1 > 0$ and let ϕ be the given test. With

$$f(\mathbf{x} \mid \theta) = I\big(Y_1 > \theta\big) I\big(Y_n < \theta + 1\big),$$

the likelihood ratio is $\lambda(\mathbf{x}) = f(\mathbf{x} \mid \theta_1)/f(\mathbf{x} \mid 0)$, taking only the values 0, 1, ∞.
(i) $0 < \theta_1 \leq k$: take $k' = 1$ in Theorem 8.3.12a. Then $\lambda > 1$ means $f(\mathbf{x} \mid \theta_1) > 0 = f(\mathbf{x} \mid 0)$, which (as $Y_1 > \theta_1 > 0$) forces $Y_n \geq 1$, and ϕ rejects. And $\lambda < 1$ means $f(\mathbf{x} \mid 0) > 0 = f(\mathbf{x} \mid \theta_1)$; since $Y_n < 1 < \theta_1 + 1$ this forces $Y_1 \leq \theta_1 \leq k$, and ϕ accepts (a.e., as $P(Y_1 = \theta_1) = 0$). So ϕ is MP of its size α against θ_1.
(ii) $\theta_1 > k$: now ϕ rejects on all of $\{f(\mathbf{x} \mid \theta_1) > 0\} = \{Y_1 > \theta_1\} \cap \{Y_n < \theta_1 + 1\}$, because $Y_1 > \theta_1 > k$ there. Hence $\beta_\phi(\theta_1) = 1$, the largest possible power, and ϕ is trivially MP.
In both cases ϕ has size α and maximal power at θ_1; since $\theta_1 > 0$ was arbitrary, ϕ is UMP size α.
(d) As printed the requirement is vacuous: by part (b), $\beta(\theta) = 1$ for every $\theta \geq k$, and $k = 1 - (.10)^{1/n} < 1$, so *every* $n \geq 1$ with $k = 1 - (.10)^{1/n}$ already gives power 1 (hence $\geq .8$) at every $\theta > 1$. The intended threshold is $\theta \geq .1$; solving that, and using that β is increasing so $\theta = .1$ binds,

$$1 - (1 - .1)^n + .10 \;\geq\; .8 \iff (.9)^n \leq .3 \iff n \;\geq\; \frac{\log .3}{\log .9} = 11.43,$$

so $n = 12$ and

$$k = 1 - (.10)^{1/12} = .1746,$$

with resulting power $1 - (.9)^{12} + .10 = .818$ at $\theta = .1$.

Problem 8.34. In each of the following two situations, show that for any number c, if $\theta_1 \leq \theta_2$, then

$$P_{\theta_1}(T > c) \;\leq\; P_{\theta_2}(T > c).$$

(a) θ is a location parameter in the distribution of the random variable T.
(b) The family of pdfs of T, $\{g(t \mid \theta) : \theta \in \Theta\}$, has an MLR.

Solution. (a) $g(t \mid \theta) = g(t - \theta)$ for a fixed pdf g (Definition 3.5.2), so with G its cdf,

$$P_\theta(T > c) = \int_c^\infty g(t - \theta)\, dt = \int_{c-\theta}^\infty g(u)\, du = 1 - G(c - \theta).$$

Since $\theta_1 \leq \theta_2$ gives $c - \theta_2 \leq c - \theta_1$ and G is nondecreasing, $1 - G(c - \theta_1) \leq 1 - G(c - \theta_2)$.

(b) Let $\theta_1 < \theta_2$ (equality is trivial) and put

$$D(t) = g(t \mid \theta_2) - g(t \mid \theta_1), \qquad t_0 = \inf\{t : D(t) > 0\}.$$

D changes sign at most once, from $-$ to $+$: for $t < t_0$, $D(t) \le 0$ by definition of the infimum; and if $t > t_0$, pick $s \in (t_0, t)$ with $D(s) > 0$. If $g(s \mid \theta_1) > 0$ then $r(s) = g(s \mid \theta_2)/g(s \mid \theta_1) > 1$, and the MLR (Definition 8.3.16) gives $r(t) \ge r(s) > 1$ whenever $g(t \mid \theta_1) > 0$, so $D(t) > 0$; if instead $g(s \mid \theta_1) = 0 < g(s \mid \theta_2)$ then $r(s) = \infty$ and monotonicity forces $g(t \mid \theta_1) = 0$, so $D(t) = g(t \mid \theta_2) \ge 0$. Either way $D(t) \ge 0$ for $t > t_0$.
Since both are pdfs, $\int_{-\infty}^{\infty} D(t)\,dt = 0$, so

$$P_{\theta_2}(T > c) - P_{\theta_1}(T > c) = \int_c^{\infty} D(t)\,dt = -\int_{-\infty}^{c} D(t)\,dt.$$

If $c \ge t_0$ the middle expression integrates a nonnegative function; if $c < t_0$ the right expression is minus the integral of a nonpositive function. Either way the difference is ≥ 0.

Problem 8.35. The usual t distribution, as derived in Section 5.3.2, is also known as a *central t* distribution. It can be thought of as the pdf of a random variable of the form $T = \mathrm{n}(0,1)/\sqrt{\chi^2_\nu/\nu}$, where the normal and the chi squared random variables are independent. A generalization of the t distribution, the *noncentral t*, is of the form $T' = \mathrm{n}(\mu,1)/\sqrt{\chi^2_\nu/\nu}$, where the normal and the chi squared random variables are independent and we can have $\mu \ne 0$. (We have already seen a noncentral pdf, the noncentral chi squared, in (4.4.3).) Formally, if $X \sim \mathrm{n}(\mu,1)$ and $Y \sim \chi^2_\nu$, independent of X, then $T' = X/\sqrt{Y/\nu}$ has a noncentral t distribution with ν degrees of freedom and noncentrality parameter $\delta = \sqrt{\mu^2}$.
(a) Calculate the mean and variance of T'.
(b) The pdf of T' is given by

$$f_{T'}(t \mid \delta) = \frac{e^{-\delta^2/2}}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})\sqrt{\nu}} \sum_{k=0}^{\infty} \frac{(2/\nu)^{k/2}(\delta t)^k}{k!} \frac{\Gamma\big[(\nu+k+1)/2\big]}{\big(1+(t^2/\nu)\big)^{(\nu+k+1)/2}}.$$

Show that this pdf reduces to that of a central t if $\delta = 0$.
(c) Show that the pdf of T' has an MLR in its noncentrality parameter.

Solution. Throughout, $\delta = \mu$: the series in (b) contains odd powers of δt, so the noncentrality parameter must be the *signed* mean, and the printed $\delta = \sqrt{\mu^2}$ is to be read that way.
(a) Write $T' = \sqrt{\nu}\, X\, Y^{-1/2}$ with $X \perp Y$. From the χ^2_ν pdf,

$$\mathrm{E}\, Y^r = \frac{2^r \Gamma\big(\frac{\nu}{2} + r\big)}{\Gamma\big(\frac{\nu}{2}\big)}, \qquad r > -\tfrac{\nu}{2},$$

so $\mathrm{E}\, Y^{-1/2} = 2^{-1/2}\Gamma(\frac{\nu-1}{2})/\Gamma(\frac{\nu}{2})$ and $\mathrm{E}\, Y^{-1} = 1/(\nu-2)$. Independence gives

$$\mathrm{E}\, T' = \sqrt{\nu}\, \mathrm{E}X\, \mathrm{E}\, Y^{-1/2} = \delta\sqrt{\frac{\nu}{2}}\, \frac{\Gamma\big(\frac{\nu-1}{2}\big)}{\Gamma\big(\frac{\nu}{2}\big)} \qquad (\nu > 1),$$

$$\mathrm{E}\, T'^2 = \nu\, \mathrm{E}X^2\, \mathrm{E}\, Y^{-1} = \frac{\nu(1+\delta^2)}{\nu-2} \qquad (\nu > 2),$$

and therefore

$$\mathrm{Var}\, T' = \frac{\nu(1+\delta^2)}{\nu-2} - \frac{\nu\delta^2}{2}\left[\frac{\Gamma\big(\frac{\nu-1}{2}\big)}{\Gamma\big(\frac{\nu}{2}\big)}\right]^2, \qquad \nu > 2.$$

(b) At $\delta = 0$ every term with $k \ge 1$ carries the factor $\delta^k = 0$, leaving only $k = 0$:

$$f_{T'}(t \mid 0) = \frac{1}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})\sqrt{\nu}}\, \frac{\Gamma\big(\frac{\nu+1}{2}\big)}{(1+t^2/\nu)^{(\nu+1)/2}} = \frac{\Gamma\big(\frac{\nu+1}{2}\big)}{\Gamma\big(\frac{\nu}{2}\big)\sqrt{\nu\pi}}\, \frac{1}{(1+t^2/\nu)^{(\nu+1)/2}},$$

using $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. This is the Student's t_ν pdf of Section 5.3.2.
(c) Write $T' = XW$, where $X \sim \mathrm{n}(\delta, 1)$ and $W = \sqrt{\nu/Y}$ is independent of X and free of δ; by the transformation $Y = \nu/W^2$ applied to the χ^2_ν pdf,

$$h(w) = c_\nu \, w^{-\nu-1} e^{-\nu/(2w^2)}, \qquad w > 0.$$

Substituting $x = t/w$ in $f(t \mid \delta) = \int_0^\infty w^{-1} \varphi(t/w - \delta) h(w)\, dw$ (φ the standard normal pdf) gives, for $t \neq 0$,

$$\begin{aligned} f(t \mid \delta) &= \int_{-\infty}^{\infty} L(t,x)\, \varphi(x - \delta)\, dx, \\ L(t,x) &= \frac{1}{|x|} h\left(\frac{t}{x}\right) I(tx > 0) \\ &= c_\nu \frac{|x|^\nu}{|t|^{\nu+1}} e^{-\nu x^2/(2t^2)} I(tx > 0). \end{aligned}$$

Both kernels are TP_2. For the normal one, $\partial^2 \log \varphi(x - \delta)/\partial x\, \partial\delta = 1 > 0$, so for $x < y$ and $\delta_1 < \delta_2$,

$$\Delta_\varphi = \varphi(x - \delta_1)\varphi(y - \delta_2) - \varphi(x - \delta_2)\varphi(y - \delta_1) \;\geq\; 0.$$

For L, fix $t_1 < t_2$ and $x < y$ and compute

$$\Delta_L = L(t_1, x)L(t_2, y) - L(t_1, y)L(t_2, x).$$

(i) t_1, t_2 of the same sign: on the open quadrant $\{tx > 0\}$,

$$\frac{\partial^2}{\partial t\, \partial x} \log L(t,x) = \frac{\partial}{\partial t}\left(\frac{\nu}{x} - \frac{\nu x}{t^2}\right) = \frac{2\nu x}{t^3} > 0,$$

so $\Delta_L \geq 0$ when x, y both lie in that quadrant, and any x or y outside it zeroes a whole column of the determinant. (ii) $t_1 < 0 < t_2$: then $L(t_1, y)L(t_2, x) \neq 0$ would need $y < 0$ and $x > 0$, contradicting $x < y$, so $\Delta_L = L(t_1, x)L(t_2, y) \geq 0$.
The basic composition formula (Cauchy–Binet for kernels) now gives, for $t_1 < t_2$ and $\delta_1 < \delta_2$,

$$\begin{aligned} &f(t_1 \mid \delta_1) f(t_2 \mid \delta_2) - f(t_1 \mid \delta_2) f(t_2 \mid \delta_1) \\ &\qquad = \iint_{x<y} \Delta_L \, \Delta_\varphi \, dx\, dy \;\geq\; 0. \end{aligned}$$

Rearranging,

$$\frac{f(t_2 \mid \delta_2)}{f(t_2 \mid \delta_1)} \;\geq\; \frac{f(t_1 \mid \delta_2)}{f(t_1 \mid \delta_1)},$$

so $f(t \mid \delta_2)/f(t \mid \delta_1)$ is nondecreasing in t (on $t \neq 0$, hence everywhere by continuity of f). The noncentral t family has an MLR in δ.

8.6 Exercises 8.36–8.42

Problem 8.36. We have one observation from a beta$(1, \theta)$ population.
(a) To test $H_0 : \theta_1 \leq \theta \leq \theta_2$ versus $H_1 : \theta < \theta_1$ or $\theta > \theta_2$, where $\theta_1 = 1$ and $\theta_2 = 2$, a test satisfies $\mathrm{E}_{\theta_1}\phi = .5$ and $\mathrm{E}_{\theta_2}\phi = .3$. Find a test that is as good, and explain why it is as good.
(b) For testing $H_0 : \theta = \theta_1$ versus $H_1 : \theta \neq \theta_1$, with $\theta_1 = 1$, find a two-sided test (other than $\phi \equiv .1$) that satisfies $\mathrm{E}_{\theta_1}\phi = .1$ and $\frac{d}{d\theta}\mathrm{E}_\theta\phi\big|_{\theta=\theta_1} = 0$.

Solution. (a) Take the nonrandomized test

$$\psi(x) = \begin{cases} 1 & x < .05 \text{ or } x > .55, \\ 0 & \text{otherwise.} \end{cases}$$

The beta$(1,\theta)$ pdf is $f(x|\theta)=\theta(1-x)^{\theta-1}$, so $P_\theta(X>c)=(1-c)^\theta$ and

$$\beta_\psi(\theta)=1-(.95)^\theta+(.45)^\theta,$$

giving $\beta_\psi(1)=1-.95+.45=.5$ and $\beta_\psi(2)=1-.9025+.2025=.3$: the same two error probabilities as ϕ. It is as good because it has at least as much power at every θ in H_1.
Fix $\theta\notin[1,2]$ and write $u=1-x$, so that $f(x|1)=1$, $f(x|2)=2u$ and $f(x|\theta)=\theta u^{\theta-1}$. Choose $k_1,k_2\in\mathbb{R}$ (uniquely; the 2×2 system is nonsingular) so that

$$h(x):=\theta u^{\theta-1}-k_1-2k_2u=0 \quad \text{at } u=.45 \text{ and } u=.95.$$

Every $\theta\in H_1$ has $\theta<1$ or $\theta>2$, hence $(\theta-1)(\theta-2)>0$ and $\frac{d^2}{du^2}u^{\theta-1}=(\theta-1)(\theta-2)u^{\theta-3}>0$ on $(0,1)$; subtracting the affine part leaves h strictly convex in u with exactly those two zeros, so

$$h>0 \text{ on } \{u<.45\}\cup\{u>.95\}, \qquad h<0 \text{ on } (.45,.95).$$

In terms of x that positivity set is $\{x>.55\}\cup\{x<.05\}$, which is exactly where $\psi=1$, and $\psi=0$ exactly where $h<0$. Hence $(\psi-\phi)h\geq 0$ pointwise, and

$$\begin{aligned}0&\leq\int_0^1(\psi-\phi)h\,dx\\&=\mathrm{E}_\theta\psi-\mathrm{E}_\theta\phi-k_1(\mathrm{E}_1\psi-\mathrm{E}_1\phi)-k_2(\mathrm{E}_2\psi-\mathrm{E}_2\phi)\\&=\mathrm{E}_\theta\psi-\mathrm{E}_\theta\phi,\end{aligned}$$

the last two brackets vanishing because $\mathrm{E}_1\psi=.5=\mathrm{E}_1\phi$ and $\mathrm{E}_2\psi=.3=\mathrm{E}_2\phi$. So $\beta_\psi\geq\beta_\phi$ throughout H_1 while the two constraints are met exactly.
(b) Reject H_0 if $X<.0804$ or $X>.9804$.
For a two-sided region $\{X<c_2\}\cup\{X>c_1\}$ put $a=1-c_1$, $b=1-c_2$, so that

$$\beta(\theta)=1-b^\theta+a^\theta, \qquad \beta'(\theta)=a^\theta\log a-b^\theta\log b.$$

The two requirements are therefore

$$b-a=.9 \qquad \text{and} \qquad a\log a=b\log b.$$

Substituting $b=a+.9$, the function $a\log a-(a+.9)\log(a+.9)$ equals $+.0948$ at $a=0$ and $-.2303$ at $a=.1$, so it has a root in between; numerically $a=.019602$, $b=.919602$ (then $a\log a=b\log b=-.0771$). Thus $c_1=1-a=.9804$, $c_2=1-b=.0804$, and

$$\beta(\theta)=1-(.9196)^\theta+(.0196)^\theta$$

satisfies $\beta(1)=.1$ and $\beta'(1)=0$. (Check!)

Problem 8.37. Let $X_1,\ldots,X_n$ be a random sample from a n(θ,σ^2) population. Consider testing

$$H_0:\theta\leq\theta_0 \qquad \text{versus} \qquad H_1:\theta>\theta_0.$$

(a) If σ^2 is known, show that the test that rejects H_0 when

$$\bar{X}>\theta_0+z_\alpha\sqrt{\sigma^2/n}$$

is a test of size α. Show that the test can be derived as an LRT.
(b) Show that the test in part (a) is a UMP test.
(c) If σ^2 is unknown, show that the test that rejects H_0 when

$$\bar{X}>\theta_0+t_{n-1,\alpha}\sqrt{S^2/n}$$

is a test of size α. Show that the test can be derived as an LRT.

Solution. (a) Since $\bar{X} \sim \mathrm{n}(\theta, \sigma^2/n)$,

$$\beta(\theta) = P_\theta\Big(\bar{X} > \theta_0 + z_\alpha \tfrac{\sigma}{\sqrt{n}}\Big) = P\Big(Z > z_\alpha + \tfrac{\sqrt{n}(\theta_0 - \theta)}{\sigma}\Big),$$

which is increasing in θ; hence the size is $\sup_{\theta \le \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha$.
For the LRT, the unrestricted MLE is $\bar{x}$ and the MLE restricted to $\theta \le \theta_0$ is $\min\{\bar{x}, \theta_0\}$, so

$$\lambda(\mathbf{x}) = \begin{cases} 1 & \bar{x} \le \theta_0, \\ \exp\Big(-\dfrac{n(\bar{x} - \theta_0)^2}{2\sigma^2}\Big) & \bar{x} > \theta_0, \end{cases}$$

using $\sum(x_i - \theta_0)^2 = \sum(x_i - \bar{x})^2 + n(\bar{x} - \theta_0)^2$. Thus $\lambda < c$ if and only if $\bar{x} - \theta_0 > \sigma\sqrt{-2\log c}/\sqrt{n}$, and $c = e^{-z_\alpha^2/2}$ returns exactly the given region.
(b) The $\mathrm{n}(\theta, \sigma^2)$ family with σ^2 known has MLR in the sufficient statistic $\bar{X}$ (Exercise 8.25(a)). The test rejects for $\bar{X} > t_0$ with $t_0 = \theta_0 + z_\alpha\sigma/\sqrt{n}$ and $P_{\theta_0}(\bar{X} > t_0) = \alpha$, so by the Karlin-Rubin Theorem 8.3.17 it is UMP level α; by part (a) its size is α, so it is UMP size α.
(c) By Theorem 5.3.1 and Section 5.3.2, $(\bar{X} - \theta)/\sqrt{S^2/n} \sim t_{n-1}$ for every (θ, σ^2). For $\theta \le \theta_0$,

$$\begin{aligned} P_\theta\Big(\bar{X} > \theta_0 + t_{n-1,\alpha}\sqrt{S^2/n}\Big) &\le P_\theta\Big(\bar{X} > \theta + t_{n-1,\alpha}\sqrt{S^2/n}\Big) \\ &= P(T_{n-1} > t_{n-1,\alpha}) = \alpha, \end{aligned}$$

with equality at $\theta = \theta_0$; the bound does not depend on σ^2, so the size is α.
For the LRT both θ and σ^2 are free. Maximizing over σ^2 first gives $\sup L = (2\pi\hat{\sigma}^2)^{-n/2}e^{-n/2}$ with $\hat{\sigma}^2$ the MLE of σ^2 at the maximizing θ, so with $\hat{\sigma}^2 = \sum(x_i - \bar{x})^2/n$ and, when $\bar{x} > \theta_0$, $\hat{\sigma}_0^2 = \sum(x_i - \theta_0)^2/n$,

$$\lambda(\mathbf{x}) = \Big(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\Big)^{n/2} = \left(\frac{\sum(x_i - \bar{x})^2}{\sum(x_i - \theta_0)^2}\right)^{n/2} = \Big(1 + \frac{t^2}{n-1}\Big)^{-n/2},$$

where $t = (\bar{x} - \theta_0)/\sqrt{s^2/n}$, and $\lambda = 1$ when $\bar{x} \le \theta_0$. Hence $\lambda < c$ if and only if $\bar{x} > \theta_0$ and t exceeds a constant; taking that constant to be $t_{n-1,\alpha}$ gives the size-α test above.

Problem 8.38. Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, \sigma^2)$, where θ_0 is a specified value of θ and σ^2 is unknown. We are interested in testing

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \ne \theta_0.$$

(a) Show that the test that rejects H_0 when

$$|\bar{X} - \theta_0| > t_{n-1,\alpha/2}\sqrt{S^2/n}$$

is a test of size α.
(b) Show that the test in part (a) can be derived as an LRT.

Solution. (a) By Theorem 5.3.1 ($\bar{X}$ and S^2 are independent, $\bar{X} \sim \mathrm{n}(\theta, \sigma^2/n)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$) and the definition of Student's t in Section 5.3.2,

$$T = \frac{\bar{X} - \theta_0}{\sqrt{S^2/n}} \sim t_{n-1} \quad \text{when } \theta = \theta_0,$$

whatever the value of σ^2. Since H_0 consists only of the points (θ_0, σ^2),

$$\text{size} = \sup_{\sigma^2 > 0} P_{\theta_0,\sigma^2}(|T| > t_{n-1,\alpha/2}) = \alpha.$$

(b) The unrestricted MLEs are $\bar{x}$ and $\hat{\sigma}^2 = \sum(x_i - \bar{x})^2/n$; under H_0 the MLE of σ^2 is $\hat{\sigma}_0^2 = \sum(x_i -$

$\theta_0)^2/n$. Since $\sup L = (2\pi\hat{\sigma}^2)^{-n/2}e^{-n/2}$ in each case,

$$\lambda(\mathbf{x}) = \left(\frac{\sum(x_i-\bar{x})^2}{\sum(x_i-\theta_0)^2}\right)^{n/2} = \left(1+\frac{t^2}{n-1}\right)^{-n/2},$$

by $\sum(x_i-\theta_0)^2 = \sum(x_i-\bar{x})^2 + n(\bar{x}-\theta_0)^2$, with $t = (\bar{x}-\theta_0)/\sqrt{s^2/n}$. This is strictly decreasing in $|t|$, so $\lambda < c$ if and only if $|t| > \sqrt{(n-1)(c^{-2/n}-1)}$, and the choice

$$c = \left(1+\frac{t^2_{n-1,\alpha/2}}{n-1}\right)^{-n/2}$$

reproduces the test of part (a).

Problem 8.39. Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a random sample from a bivariate normal distribution with parameters $\mu_X,\mu_Y,\sigma^2_X,\sigma^2_Y,\rho$. We are interested in testing

$$H_0: \mu_X = \mu_Y \qquad \text{versus} \qquad H_1: \mu_X \neq \mu_Y.$$

(a) Show that the random variables $W_i = X_i - Y_i$ are iid $\mathrm{n}(\mu_W,\sigma^2_W)$.
(b) Show that the above hypothesis can be tested with the statistic

$$T_W = \frac{\bar{W}}{\sqrt{\frac{1}{n}S^2_W}},$$

where $\bar{W} = \frac{1}{n}\sum_{i=1}^n W_i$ and $S^2_W = \frac{1}{n-1}\sum_{i=1}^n (W_i-\bar{W})^2$. Furthermore, show that, under H_0, $T_W \sim$ Student's t with $n-1$ degrees of freedom. (This test is known as the paired-sample t test.)

Solution. (a) Each W_i is the linear combination $aX_i + bY_i$ with $a=1$, $b=-1$, so by fact (d) following Definition 4.5.10,

$$W_i \sim \mathrm{n}\big(\underbrace{\mu_X-\mu_Y}_{\mu_W},\ \underbrace{\sigma^2_X+\sigma^2_Y-2\rho\sigma_X\sigma_Y}_{\sigma^2_W}\big),$$

and $\sigma^2_W > 0$ because $|\rho| < 1$. The W_i are identically distributed functions of the iid pairs (X_i,Y_i), hence iid.
(b) H_0 is exactly $\mu_W = 0$, and by (a) $W_1,\ldots,W_n$ is a random sample from $\mathrm{n}(\mu_W,\sigma^2_W)$ with both parameters unknown. So this is the one-sample problem of Exercise 8.38 with $\theta_0 = 0$: under H_0, Theorem 5.3.1 and Section 5.3.2 give

$$T_W = \frac{\bar{W}-0}{\sqrt{S^2_W/n}} \sim t_{n-1},$$

and by Exercise 8.38 the size-α LRT rejects H_0 when $|T_W| > t_{n-1,\alpha/2}$.

Problem 8.40. Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a random sample from a bivariate normal distribution with parameters $\mu_X,\mu_Y,\sigma^2_X,\sigma^2_Y,\rho$.
(a) Derive the LRT of

$$H_0: \mu_X = \mu_Y \qquad \text{versus} \qquad H_1: \mu_X \neq \mu_Y,$$

where σ^2_X,σ^2_Y, and ρ are unspecified and unknown.
(b) Show that the test derived in part (a) is equivalent to the paired t test of Exercise 8.39.
(Hint: Straightforward maximization of the bivariate likelihood is possible but somewhat nasty. Filling in the gaps of the following argument gives a more elegant proof.) Make the transformation

$u = x - y$, $v = x + y$. Let $f(x, y)$ denote the bivariate normal pdf, and write

$$f(x, y) = g(v|u)h(u),$$

where $g(v|u)$ is the conditional pdf of V given U, and $h(u)$ is the marginal pdf of U. Argue that (1) the likelihood can be equivalently factored and (2) the piece involving $g(v|u)$ has the same maximum whether or not the means are restricted. Thus, it can be ignored (since it will cancel) and the LRT is based only on $h(u)$. However, $h(u)$ is a normal pdf with mean $\mu_X - \mu_Y$, and the LRT is the usual one-sample t test, as derived in Exercise 8.38.

Solution. (a) The LRT rejects H_0 when $|T_W| > t_{n-1,\alpha/2}$, where $W_i = X_i - Y_i$ and $T_W = \bar{W}/\sqrt{S_W^2/n}$; equivalently

$$\lambda(\mathbf{x}, \mathbf{y}) = \Big(1 + \frac{T_W^2}{n-1}\Big)^{-n/2}.$$

Filling in the three gaps of the hint:
(i) *Reparametrize.* Put $U_i = X_i - Y_i$, $V_i = X_i + Y_i$. This is a nonsingular linear map, so the (U_i, V_i) are iid bivariate normal with

$$\begin{aligned} \mu_U &= \mu_X - \mu_Y, & \mu_V &= \mu_X + \mu_Y, \\ \sigma_U^2 &= \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y, & \sigma_V^2 &= \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y, \end{aligned}$$

and $\mathrm{Cov}(U, V) = \sigma_X^2 - \sigma_Y^2$. This parameter map is one-to-one onto: invert it by $\sigma_X^2 = \frac{1}{4}(\sigma_U^2 + \sigma_V^2) + \frac{1}{2}\mathrm{Cov}(U, V)$, $\sigma_Y^2 = \frac{1}{4}(\sigma_U^2 + \sigma_V^2) - \frac{1}{2}\mathrm{Cov}(U, V)$, $\rho = (\sigma_V^2 - \sigma_U^2)/(4\sigma_X\sigma_Y)$, the constraint $|\rho| < 1$ being positive definiteness of either covariance matrix (the linear map is nonsingular). Under this correspondence H_0 reads $\mu_U = 0$, with the other four parameters unrestricted.
(ii) *Factor the likelihood.* Since $(u, v) \mapsto (x, y)$ has constant Jacobian, $f(x, y) = 2g(v|u)h(u)$, so

$$L(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho \,|\, \mathbf{x}, \mathbf{y}) = 2^n \prod_{i=1}^n g(v_i|u_i) \prod_{i=1}^n h(u_i).$$

By Definition 4.5.10 and the conditional distributions listed after it, $g(v|u)$ is the $\mathrm{n}(\gamma_0 + \gamma_1 u,\ \tau^2)$ pdf and $h(u)$ is the $\mathrm{n}(\mu_U, \sigma_U^2)$ pdf, where

$$\gamma_1 = \rho_{UV}\frac{\sigma_V}{\sigma_U}, \qquad \gamma_0 = \mu_V - \gamma_1\mu_U, \qquad \tau^2 = \sigma_V^2(1 - \rho_{UV}^2).$$

The five numbers $(\mu_U, \sigma_U^2, \gamma_0, \gamma_1, \tau^2)$ are again a one-to-one reparametrization, and they vary freely over $\mathbb{R} \times (0, \infty) \times \mathbb{R} \times \mathbb{R} \times (0, \infty)$.
(iii) *Cancel the conditional piece.* Because $(\gamma_0, \gamma_1, \tau^2)$ is variation independent of (μ_U, σ_U^2) and is unrestricted both under H_0 and under the full model,

$$\begin{aligned} \sup_{\Theta} L &= \Big(\sup_{\gamma_0,\gamma_1,\tau^2} \textstyle\prod_i g\Big)\Big(\sup_{\mu_U,\sigma_U^2} \textstyle\prod_i h\Big), \\ \sup_{\Theta_0} L &= \Big(\sup_{\gamma_0,\gamma_1,\tau^2} \textstyle\prod_i g\Big)\Big(\sup_{\sigma_U^2} \textstyle\prod_i h\big|_{\mu_U=0}\Big), \end{aligned}$$

and the common first factor cancels in λ.
(b) What remains is the LRT of $\mu_U = 0$ against $\mu_U \neq 0$ based on the random sample $u_1, \ldots, u_n$ from $\mathrm{n}(\mu_U, \sigma_U^2)$ with σ_U^2 unknown. That is Exercise 8.38 with $\theta_0 = 0$:

$$\lambda = \left(\frac{\sum(u_i - \bar{u})^2}{\sum u_i^2}\right)^{n/2} = \Big(1 + \frac{t^2}{n-1}\Big)^{-n/2}, \qquad t = \frac{\bar{u}}{\sqrt{s_u^2/n}}.$$

Since $u_i = x_i - y_i = w_i$, this t is T_W and the LRT rejects for $|T_W| > t_{n-1,\alpha/2}$: the paired t test of Exercise 8.39.

Problem 8.41. Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu_X, \sigma_X^2)$, and let $Y_1, \ldots, Y_m$ be an independent random sample from a $\mathrm{n}(\mu_Y, \sigma_Y^2)$. We are interested in testing

$$H_0 : \mu_X = \mu_Y \qquad \text{versus} \qquad H_1 : \mu_X \neq \mu_Y$$

with the assumption that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.
(a) Derive the LRT for these hypotheses. Show that the LRT can be based on the statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}},$$

where

$$S_p^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2\right).$$

(The quantity S_p^2 is sometimes referred to as a pooled variance estimate. This type of estimate will be used extensively in Section 11.2.)
(b) Show that, under H_0, $T \sim t_{n+m-2}$. (This test is known as the two-sample t test.)
(c) Samples of wood were obtained from the core and periphery of a certain Byzantine church. The date of the wood was determined, giving the following data.

Core		Periphery	
1294	1251	1284	1274
1279	1248	1272	1264
1274	1240	1256	1256
1264	1232	1254	1250
1263	1220	1242	
1254	1218		
1251	1210		

Use the two-sample t test to determine if the mean age of the core is the same as the mean age of the periphery.

Solution. (a) $\lambda = \left(1 + \dfrac{T^2}{n+m-2}\right)^{-(n+m)/2}$, so the LRT rejects for large $|T|$.
The likelihood is

$$L = (2\pi\sigma^2)^{-(n+m)/2} \exp\left(-\frac{\sum_i (x_i - \mu_X)^2 + \sum_i (y_i - \mu_Y)^2}{2\sigma^2}\right),$$

so maximizing over σ^2 gives $\sup L = (2\pi\hat{\sigma}^2)^{-(n+m)/2} e^{-(n+m)/2}$ with $(n+m)\hat{\sigma}^2$ the minimized sum of squares. Unrestricted, $\hat{\mu}_X = \bar{x}$, $\hat{\mu}_Y = \bar{y}$ and

$$(n+m)\hat{\sigma}^2 = \sum_i (x_i - \bar{x})^2 + \sum_i (y_i - \bar{y})^2 = (n+m-2)s_p^2.$$

Under H_0 the common mean has MLE $\hat{\mu}_0 = (n\bar{x} + m\bar{y})/(n+m)$, and since $\bar{x} - \hat{\mu}_0 = \frac{m(\bar{x}-\bar{y})}{n+m}$ and $\bar{y} - \hat{\mu}_0 = -\frac{n(\bar{x}-\bar{y})}{n+m}$,

$$\begin{aligned}(n+m)\hat{\sigma}_0^2 &= \sum_i (x_i - \hat{\mu}_0)^2 + \sum_i (y_i - \hat{\mu}_0)^2 \\ &= (n+m-2)s_p^2 + \frac{nm^2 + mn^2}{(n+m)^2}(\bar{x} - \bar{y})^2 \\ &= (n+m-2)s_p^2 + \frac{(\bar{x} - \bar{y})^2}{\frac{1}{n} + \frac{1}{m}}.\end{aligned}$$

Hence

$$\lambda = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{(n+m)/2} = \left(1 + \frac{t^2}{n+m-2}\right)^{-(n+m)/2},$$

a strictly decreasing function of $|t|$, so $\lambda < c$ if and only if $|t| > t_{n+m-2,\alpha/2}$ for the appropriate c.
(b) Under H_0 the numerator $\bar{X} - \bar{Y} \sim \mathrm{n}\big(0, \sigma^2(\frac{1}{n} + \frac{1}{m})\big)$, so

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathrm{n}(0,1).$$

By Theorem 5.3.1(c), $(n-1)S_X^2/\sigma^2 \sim \chi^2_{n-1}$ and $(m-1)S_Y^2/\sigma^2 \sim \chi^2_{m-1}$, independent (the two samples are independent), so by Lemma 5.3.2(b)

$$\frac{(n+m-2)S_p^2}{\sigma^2} \sim \chi^2_{n+m-2},$$

and it is independent of $\bar{X} - \bar{Y}$ by Theorem 5.3.1(a). Therefore

$$T = \frac{Z}{\sqrt{\frac{(n+m-2)S_p^2/\sigma^2}{n+m-2}}} \sim t_{n+m-2}$$

by the definition of Student's t in Section 5.3.2.
(c) $T = -1.29$: do not reject H_0. With $n = 14$ core and $m = 9$ periphery observations,

$$\bar{x} = \frac{17498}{14} = 1249.86, \qquad \bar{y} = \frac{11352}{9} = 1261.33,$$

$$\sum (x_i - \bar{x})^2 = 7687.71, \qquad \sum (y_i - \bar{y})^2 = 1408.00,$$

so

$$s_p^2 = \frac{7687.71 + 1408.00}{21} = 433.13, \qquad s_p\sqrt{\tfrac{1}{14} + \tfrac{1}{9}} = 8.892,$$

$$t = \frac{1249.86 - 1261.33}{8.892} = \frac{-11.48}{8.892} = -1.291.$$

Since $|t| = 1.29 < t_{21,.025} = 2.080$ (two-sided p-value = .21), there is no evidence that the mean age of the core differs from that of the periphery.

Problem 8.42. The assumption of equal variances, which was made in Exercise 8.41, is not always tenable. In such a case, the distribution of the statistic is no longer a t. Indeed, there is doubt as to the wisdom of calculating a pooled variance estimate. (This problem, of making inference on means when variances are unequal, is, in general, quite a difficult one. It is known as the Behrens-Fisher Problem.) A natural test to try is the following modification of the two-sample t test: Test

$$H_0 : \mu_X = \mu_Y \qquad \text{versus} \qquad H_1 : \mu_X \neq \mu_Y,$$

where we do not assume that $\sigma_X^2 = \sigma_Y^2$, using the statistic

$$T' = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)}},$$

where

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{and} \quad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2.$$

The exact distribution of T' is not pleasant, but we can approximate the distribution using Satterthwaite's approximation (Example 7.2.3).

(a) Show that

$$\frac{\dfrac{S_X^2}{n}+\dfrac{S_Y^2}{m}}{\dfrac{\sigma_X^2}{n}+\dfrac{\sigma_Y^2}{m}} \sim \frac{\chi_\nu^2}{\nu} \qquad \text{(approximately)},$$

where ν can be estimated with

$$\hat{\nu} = \frac{\left(\dfrac{S_X^2}{n}+\dfrac{S_Y^2}{m}\right)^2}{\dfrac{S_X^4}{n^2(n-1)}+\dfrac{S_Y^4}{m^2(m-1)}}.$$

(b) Argue that the distribution of T' can be approximated by a t distribution with $\hat{\nu}$ degrees of freedom.
(c) Re-examine the data from Exercise 8.41 using the approximate t test of this exercise; that is, test if the mean age of the core is the same as the mean age of the periphery using the T' statistic.
(d) Is there any statistical evidence that the variance of the data from the core may be different from the variance of the data from the periphery? (Recall Example 5.4.1.)

Solution. (a) Apply Example 7.2.3 with $k = 2$,

$$Y_1 = \frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi_{n-1}^2, \qquad Y_2 = \frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi_{m-1}^2$$

(independent, by Theorem 5.3.1(c) and independence of the two samples), and, writing $c = \sigma_X^2/n + \sigma_Y^2/m$,

$$a_1 = \frac{\sigma_X^2}{c\,n(n-1)}, \qquad a_2 = \frac{\sigma_Y^2}{c\,m(m-1)}.$$

Then $a_1Y_1 + a_2Y_2 = \left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)/c$ is the ratio in question, and Satterthwaite's first-moment constraint holds:

$$\sum_i a_i r_i = a_1(n-1) + a_2(m-1) = \frac{\sigma_X^2/n + \sigma_Y^2/m}{c} = 1,$$

so matching second moments as in Example 7.2.3 gives $a_1Y_1 + a_2Y_2 \sim \chi_\nu^2/\nu$ approximately, with Satterthwaite's estimator

$$\hat{\nu} = \frac{(\sum_i a_i Y_i)^2}{\sum_i a_i^2 Y_i^2/r_i}.$$

Here $a_1Y_1 = (S_X^2/n)/c$ and $a_2Y_2 = (S_Y^2/m)/c$, so every c cancels and

$$\hat{\nu} = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}.$$

(b) Under H_0, $\bar{X} - \bar{Y} \sim \mathrm{n}(0, \sigma_X^2/n + \sigma_Y^2/m)$ exactly, so dividing numerator and denominator of T' by $\sqrt{c}$,

$$T' = \frac{(\bar{X}-\bar{Y})/\sqrt{c}}{\sqrt{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)/c}} \approx \frac{\mathrm{n}(0,1)}{\sqrt{\chi_{\hat{\nu}}^2/\hat{\nu}}},$$

by part (a). The numerator is independent of the denominator, since $(\bar{X}, \bar{Y})$ is independent of (S_X^2, S_Y^2) by Theorem 5.3.1(a) applied within each sample plus independence across samples. By the definition of Student's t in Section 5.3.2, $T' \approx t_{\hat{\nu}}$.
(c) $T' = -1.46$ on $\hat{\nu} = 20.6$ degrees of freedom: do not reject H_0. From Exercise 8.41(c), $s_X^2 = 591.36$ $(n = 14)$ and $s_Y^2 = 176.00$ $(m = 9)$, so

$$\frac{s_X^2}{n} + \frac{s_Y^2}{m} = 42.240 + 19.556 = 61.796, \qquad \sqrt{61.796} = 7.861,$$

$$t' = \frac{1249.86 - 1261.33}{7.861} = -1.460,$$

$$\hat{\nu} = \frac{(61.796)^2}{\frac{(42.240)^2}{13} + \frac{(19.556)^2}{8}} = \frac{3818.7}{137.25 + 47.80} = 20.64.$$

Since $|t'| = 1.46 < t_{20.6,.025} = 2.08$ (two-sided p-value $= .16$), the conclusion of Exercise 8.41 is unchanged: no evidence of a difference in mean age.
(d) $F = s_X^2/s_Y^2 = 591.36/176.00 = 3.36$, which is not significant at level .05. If $\sigma_X^2 = \sigma_Y^2$ then $S_X^2/S_Y^2 \sim F_{13,8}$ by the variance-ratio distribution of Example 5.3.5 (the printed cross-reference to Example 5.4.1 is a misprint), and the level-.05 acceptance region is

$$\big(F_{13,8,.975},\ F_{13,8,.025}\big) = (0.295,\ 4.162) \ni 3.36.$$

The two-sided p-value is .09: suggestive, but not significant at level .05.

8.7 Exercises 8.43–8.49

Problem 8.43. Sprott and Farewell (1993) note that in the two-sample t test, a valid t statistic can be derived as long as the ratio of variances is known. Let $X_1, \ldots, X_{n_1}$ be a sample from a $n(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_{n_2}$ a sample from a $n(\mu_2, \rho^2\sigma^2)$, where ρ^2 is known. Show that

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{1}{n_1} + \dfrac{\rho^2}{n_2}}\ \sqrt{\dfrac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2/\rho^2}{n_1 + n_2 - 2}}}$$

has Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom and $\dfrac{s_Y^2}{\rho^2 s_X^2}$ has an F distribution with $n_2 - 1$ and $n_1 - 1$ degrees of freedom.
Sprott and Farewell also note that the t statistic is maximized at

$$\rho^2 = \frac{n_1\sqrt{n_1 - 1}\, s_X^2}{n_2\sqrt{n_2 - 1}\, s_Y^2},$$

and they suggest plotting the statistic for plausible values of ρ^2, possibly those in a confidence interval.

Solution. Rescale the second sample: $Y_j^* = Y_j/\rho$ is a random sample from $n(\mu_2/\rho, \sigma^2)$, so both samples now have the common variance σ^2 and $S_{Y^*}^2 = S_Y^2/\rho^2$.
By Theorem 5.3.1(b) and Corollary 4.6.10 (independent normals combine linearly),

$$Z := \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{\rho^2}{n_2}}} \sim n(0, 1),$$

while Theorem 5.3.1(c) on each sample and Lemma 5.3.2(b) (independent chi squareds add) give

$$W := \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2/\rho^2}{\sigma^2} = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_{Y^*}^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}.$$

Theorem 5.3.1(a) makes $\bar{X} \perp S_X^2$ and $\bar{Y} \perp S_Y^2$, and the samples are independent, so $Z \perp W$. The displayed statistic is exactly $Z/\sqrt{W/(n_1 + n_2 - 2)}$ – the unknown σ cancels – which is Student's t with $n_1 + n_2 - 2$ degrees of freedom (Section 5.3.2).
The same two chi squareds give

$$\frac{s_Y^2}{\rho^2 s_X^2} = \frac{\big[(n_2 - 1)s_Y^2/(\rho^2\sigma^2)\big]/(n_2 - 1)}{\big[(n_1 - 1)s_X^2/\sigma^2\big]/(n_1 - 1)} = \frac{\chi^2_{n_2-1}/(n_2 - 1)}{\chi^2_{n_1-1}/(n_1 - 1)},$$

a ratio of independent chi squareds divided by their degrees of freedom, which is $F_{n_2-1,\,n_1-1}$ (Definition 5.3.6, with the Y^* sample in the numerator).

Problem 8.44. Verify that Test 3 in Example 8.3.20 is an unbiased level α test. In that example $X_1, \ldots, X_n$ are iid $n(\theta, \sigma^2)$ with σ^2 known, the hypotheses are $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, and Test 3 rejects H_0 if and only if

$$\bar{X} > \theta_0 + \frac{\sigma z_{\alpha/2}}{\sqrt{n}} \qquad \text{or} \qquad \bar{X} < \theta_0 - \frac{\sigma z_{\alpha/2}}{\sqrt{n}}.$$

Solution. The power function is

$$\beta_3(\theta) = \Phi(\delta - z_{\alpha/2}) + \Phi(-\delta - z_{\alpha/2}), \qquad \delta := \frac{\sqrt{n}\,(\theta - \theta_0)}{\sigma},$$

since with $Z = \sqrt{n}(\bar{X} - \theta)/\sigma \sim n(0,1)$ the two (disjoint) rejection events are $Z > z_{\alpha/2} - \delta$ and $Z < -z_{\alpha/2} - \delta$.
Level (indeed size) α: $\Theta_0 = \{\theta_0\}$ and $\delta = 0$ there, so

$$\beta_3(\theta_0) = 2\Phi(-z_{\alpha/2}) = \alpha,$$

whence $\sup_{\theta \in \Theta_0} \beta_3(\theta) = \alpha$ (Definitions 8.3.5 and 8.3.6).
Unbiasedness: write $h(\delta) = \Phi(\delta - z_{\alpha/2}) + \Phi(-\delta - z_{\alpha/2})$, so $\beta_3(\theta) = h(\delta)$ with h even. For $\delta > 0$,

$$h'(\delta) = \varphi(\delta - z_{\alpha/2}) - \varphi(\delta + z_{\alpha/2}) > 0,$$

because $|\delta - z_{\alpha/2}| < \delta + z_{\alpha/2}$ and φ is strictly decreasing in $|\cdot|$. Hence h is strictly increasing on $[0, \infty)$ and, being even, attains its minimum only at $\delta = 0$:

$$\beta_3(\theta) = h(\delta) > h(0) = \alpha = \beta_3(\theta_0) \qquad \text{for every } \theta \neq \theta_0.$$

So $\beta_3(\theta') \geq \beta_3(\theta'')$ for all $\theta' \in \Theta_0^c$ and $\theta'' \in \Theta_0$, which is Definition 8.3.9.

Problem 8.45. Let $X_1, \ldots, X_n$ be a random sample from a $n(\theta, \sigma^2)$ population. Consider testing

$$H_0 : \theta \leq \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0.$$

Let $\bar{X}_m$ denote the sample mean of the first m observations, $X_1, \ldots, X_m$, for $m = 1, \ldots, n$. If σ^2 is known, show that for each $m = 1, \ldots, n$, the test that rejects H_0 when

$$\bar{X}_m > \theta_0 + z_\alpha \sqrt{\sigma^2/m}$$

is an unbiased size α test. Graph the power function for each of these tests if $n = 4$.

Solution. Each of the n tests has power function

$$\beta_m(\theta) = P_\theta\left(\bar{X}_m > \theta_0 + z_\alpha \frac{\sigma}{\sqrt{m}}\right) = P\left(Z > z_\alpha - \frac{\sqrt{m}\,(\theta - \theta_0)}{\sigma}\right) = \Phi\left(\frac{\sqrt{m}\,(\theta - \theta_0)}{\sigma} - z_\alpha\right),$$

because $\bar{X}_m \sim n(\theta, \sigma^2/m)$ (Theorem 5.3.1(b) applied to the first m observations).
Size α: Φ is strictly increasing, so β_m is strictly increasing in θ, and

$$\sup_{\theta \leq \theta_0} \beta_m(\theta) = \beta_m(\theta_0) = \Phi(-z_\alpha) = \alpha$$

(Definition 8.3.5).
Unbiased: for $\theta' > \theta_0 \geq \theta''$, monotonicity gives $\beta_m(\theta') > \beta_m(\theta_0) \geq \beta_m(\theta'')$, which is Definition 8.3.9.

Graph ($n = 4$). With $\sigma = 1$, $\theta_0 = 0$, $\alpha = 0.05$ (so $z_{.05} = 1.645$), the four curves are $\beta_m(\theta) = \Phi(\sqrt{m}\,\theta - 1.645)$: strictly increasing ogives from 0 to 1 crossing at the common point $(0, 0.05)$, steeper the larger m is, and nested, since $\sqrt{m}\,\theta$ increases in m for $\theta > 0$ and decreases in m for $\theta < 0$.

θ	$m = 1$	$m = 2$	$m = 3$	$m = 4$
-1.0	0.0041	0.0011	0.0004	0.0001
-0.5	0.0160	0.0093	0.0060	0.0041
0.0	0.0500	0.0500	0.0500	0.0500
0.5	0.1261	0.1742	0.2180	0.2595
1.0	0.2595	0.4088	0.5347	0.6388
1.5	0.4424	0.6831	0.8298	0.9123
2.0	0.6388	0.8817	0.9656	0.9907
2.5	0.8038	0.9707	0.9964	0.9996

Problem 8.46. Let $X_1, \ldots, X_n$ be a random sample from a $n(\theta, \sigma^2)$ population. Consider testing

$$H_0 : \theta_1 \le \theta \le \theta_2 \qquad \text{versus} \qquad H_1 : \theta < \theta_1 \ \text{ or } \ \theta > \theta_2.$$

(a) Show that the test

$$\text{reject } H_0 \text{ if } \ \bar{X} > \theta_2 + t_{n-1,\alpha/2}\sqrt{S^2/n} \quad \text{or} \quad \bar{X} < \theta_1 - t_{n-1,\alpha/2}\sqrt{S^2/n}$$

is not a size α test.
(b) Show that, for an appropriately chosen constant k, a size α test is given by

$$\text{reject } H_0 \text{ if } \ |\bar{X} - \bar{\theta}| > k\sqrt{S^2/n},$$

where $\bar{\theta} = (\theta_1 + \theta_2)/2$.
(c) Show that the tests in parts (a) and (b) are unbiased of their size. (Assume that the noncentral t distribution has an MLR.)

Solution. Sizes here are computed with σ held fixed (equivalently, σ^2 known); if σ were a free nuisance parameter then letting $\sigma \to 0$ in (b) drives the rejection probability at $\theta = \theta_1$ to 1, so no constant k could exist.
Throughout write $t = t_{n-1,\alpha/2} > 0$ and put

$$W = \frac{\sqrt{n}(\bar{X} - \theta)}{\sigma} \sim n(0,1), \qquad R = \frac{S}{\sigma}, \qquad \bar{\theta} = \frac{\theta_1 + \theta_2}{2},$$

$$\delta_0 = \frac{\sqrt{n}\,(\theta_2 - \theta_1)}{2\sigma} > 0, \qquad v = \frac{\sqrt{n}\,(\theta - \bar{\theta})}{\sigma},$$

so that $W \perp R$ (Theorem 5.3.1(a)), the law of R does not depend on (θ, σ), and H_0 says exactly $|v| \le \delta_0$. Two facts are used repeatedly:

$$\begin{aligned} E\big[\Phi(-tR)\big] &= P(W < -tR) \\ &= P\left(\frac{\sqrt{n}(\bar{X} - \theta)}{S} < -t\right) = \frac{\alpha}{2}, \end{aligned}$$

and, for each fixed $c > 0$, the function

$$\psi_c(v) = \Phi(v - c) + \Phi(-v - c), \qquad \psi_c'(v) = \varphi(v - c) - \varphi(v + c) > 0 \ \ (v > 0),$$

is even and strictly increasing in $|v|$ (the computation of Exercise 8.44).
(a) Since $\bar{X} = \theta + \sigma W/\sqrt{n}$ and $S = \sigma R$, the two rejection events are $W > (\delta_0 - v) + tR$ and $W < -(\delta_0 + v) - tR$, and they are disjoint; conditioning on R,

$$\begin{aligned} \beta_a(\theta, \sigma) &= E\Big[\Phi\big(v - \delta_0 - tR\big) + \Phi\big(-v - \delta_0 - tR\big)\Big] \\ &= E\big[\psi_{\delta_0 + tR}(v)\big]. \end{aligned}$$

This is even in v and strictly increasing in $|v|$, so on $\Theta_0 = \{|v| \le \delta_0\}$ it is largest at $|v| = \delta_0$, i.e. at $\theta = \theta_1$ or $\theta = \theta_2$, where

$$\beta_a = E\big[\Phi(-tR)\big] + E\big[\Phi(-2\delta_0 - tR)\big] = \frac{\alpha}{2} + E\big[\Phi(-2\delta_0 - tR)\big] < \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha,$$

the inequality being strict because $\Phi(-2\delta_0 - tR) < \Phi(-tR)$ pointwise. So $\sup_{\Theta_0} \beta_a < \alpha$: the test is conservative, not of size α.
(b) Here $\bar{X} - \bar{\theta} = \sigma(W + v)/\sqrt{n}$, so the test rejects iff $|W + v| > kR$, giving

$$\beta_b(\theta, \sigma) = E\big[\Phi(v - kR) + \Phi(-v - kR)\big] = E\big[\psi_{kR}(v)\big],$$

again even and strictly increasing in $|v|$. Hence

$$\sup_{\Theta_0} \beta_b = \beta_b\big|_{|v|=\delta_0} = P\big(|W + \delta_0| > kR\big) = P\big(|T_{n-1}(\delta_0)| > k\big),$$

where $T_{n-1}(\delta_0) = (W + \delta_0)/R = \sqrt{n}(\bar{X} - \bar{\theta})/S$ has a noncentral t distribution with $n - 1$ degrees of freedom and noncentrality δ_0. The map $k \mapsto P(|T_{n-1}(\delta_0)| > k)$ is continuous and strictly decreasing from 1 (at $k = 0$) to 0, so there is a unique k with

$$P\big(|T_{n-1}(\delta_0)| > k\big) = \alpha,$$

and for that k the test has size exactly α.
(c) Both power functions have the form $E[\psi_{c(R)}(v)]$ with $c(R) > 0$, so both depend on θ only through v, are even in v and strictly increasing in $|v|$ – the assumed MLR of the noncentral t in its noncentrality parameter, applied to the two tails at once. Since $\Theta_0 = \{|v| \le \delta_0\}$ and $\Theta_0^c = \{|v| > \delta_0\}$, for $\theta' \in \Theta_0^c$ and $\theta'' \in \Theta_0$,

$$\beta(\theta') > \beta\big|_{|v|=\delta_0} = \sup_{\Theta_0} \beta \;\ge\; \beta(\theta''),$$

so each test is unbiased (Definition 8.3.9) and its power exceeds its own size off Θ_0.

Problem 8.47. Consider two independent normal samples with equal variances, as in Exercise 8.41: $X_1, \dots, X_n$ iid $n(\mu_X, \sigma^2)$ and, independently, $Y_1, \dots, Y_m$ iid $n(\mu_Y, \sigma^2)$, with pooled variance estimate

$$S_p^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2}{n + m - 2}.$$

Consider testing $H_0 : \mu_X - \mu_Y \le -\delta$ or $\mu_X - \mu_Y \ge \delta$ versus $H_1 : -\delta < \mu_X - \mu_Y < \delta$, where δ is a specified positive constant. (This is called an *equivalence testing* problem.)
(a) Show that the size α LRT of $H_0^- : \mu_X - \mu_Y \le -\delta$ versus $H_1^- : \mu_X - \mu_Y > -\delta$ rejects H_0^- if

$$T^- = \frac{\bar{X} - \bar{Y} - (-\delta)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \;\ge\; t_{n+m-2,\alpha}.$$

(b) Find the size α LRT of $H_0^+ : \mu_X - \mu_Y \ge \delta$ versus $H_1^+ : \mu_X - \mu_Y < \delta$.
(c) Explain how the tests in (a) and (b) can be combined into a level α test of H_0 versus H_1.
(d) Show that the test in (c) is a size α test. (Hint: Consider $\sigma \to 0$.)
This procedure is sometimes known as the *two one-sided tests procedure* and was derived by Schuirmann (1987) (see also Westlake 1981) for the problem of testing *bioequivalence*. See also the review article by Berger and Hsu (1996) and Exercise 9.33 for a confidence interval counterpart.

Solution. Write $\eta = \mu_X - \mu_Y$, $c = \frac{1}{n} + \frac{1}{m}$, $t = t_{n+m-2,\alpha}$, and

$$T^- = \frac{\bar{X} - \bar{Y} + \delta}{\sqrt{S_p^2 c}}, \qquad T^+ = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{S_p^2 c}}.$$

(a) With $\theta = (\mu_X, \mu_Y, \sigma^2)$ the likelihood is

$$L = (2\pi\sigma^2)^{-(n+m)/2} \exp\left\{ -\frac{1}{2\sigma^2}\left[\textstyle\sum_i (x_i - \mu_X)^2 + \sum_j (y_j - \mu_Y)^2 \right] \right\}.$$

Unrestricted, $\hat\mu_X = \bar x$, $\hat\mu_Y = \bar y$ and $\hat\sigma^2 = (n+m-2)s_p^2/(n+m)$. Under H_0^-: if $\bar x - \bar y \le -\delta$ the unrestricted maximizer is admissible and $\lambda = 1$, so H_0^- is not rejected. Otherwise the constrained maximizer lies on $\mu_X - \mu_Y = -\delta$, and minimizing the exponent subject to that constraint gives, with $d = \bar x - \bar y + \delta > 0$,

$$\hat\mu_{X,0} = \bar x - \frac{m}{n+m} d, \qquad \hat\mu_{Y,0} = \bar y + \frac{n}{n+m} d,$$

$$n(\bar x - \hat\mu_{X,0})^2 + m(\bar y - \hat\mu_{Y,0})^2 = \frac{nm}{n+m} d^2 = \frac{d^2}{c}, \qquad \hat\sigma_0^2 = \hat\sigma^2 + \frac{d^2}{c(n+m)}.$$

Hence

$$\lambda(\mathbf{x}, \mathbf{y}) = \left(\frac{\hat\sigma^2}{\hat\sigma_0^2} \right)^{(n+m)/2} = \left(1 + \frac{(T^-)^2}{n+m-2} \right)^{-(n+m)/2},$$

a strictly decreasing function of T^- on the region $d > 0$ where rejection can occur. So $\lambda < c^*$ if and only if $T^- \ge k$ for some $k > 0$. Under $\eta = -\delta$, T^- is the usual two-sample t statistic and $T^- \sim t_{n+m-2}$ (Exercise 8.41(b)); in general T^- is noncentral t with noncentrality $(\eta + \delta)/(\sigma\sqrt{c})$, whose MLR makes $P_\eta(T^- \ge k)$ increasing in η. Therefore

$$\sup_{\eta \le -\delta} P_\eta(T^- \ge k) = P_{-\delta}(T^- \ge k) = \alpha \quad \Longleftrightarrow \quad k = t_{n+m-2,\alpha}.$$

(b) Replacing (X, Y) by (Y, X) and δ by $-\delta$ reverses every inequality: the size α LRT rejects H_0^+ if

$$T^+ = \frac{\bar X - \bar Y - \delta}{\sqrt{S_p^2 c}} \;\le\; -t_{n+m-2,\alpha}.$$

(c) H_0 is the union $\{\eta \le -\delta\} \cup \{\eta \ge \delta\}$, so the intersection-union test is appropriate: reject H_0 if and only if both H_0^- and H_0^+ are rejected, that is

$$R = \left\{ T^- \ge t_{n+m-2,\alpha} \right\} \cap \left\{ T^+ \le -t_{n+m-2,\alpha} \right\}.$$

Each component test has size α, so by Theorem 8.3.23 the IUT is a level α test of H_0 versus H_1. (Equivalently: reject when the interval $\bar X - \bar Y \pm t_{n+m-2,\alpha}\sqrt{S_p^2 c}$ lies inside $(-\delta, \delta)$.)

(d) Apply Theorem 8.3.24 with $R^- = \{T^- \ge t\}$, $R^+ = \{T^+ \le -t\}$, along the sequence of null points $\eta = -\delta$, $\sigma = \sigma_\ell \downarrow 0$. Writing $\bar X - \bar Y = -\delta + \sigma\sqrt{c}\, Z$ with $Z \sim n(0,1)$ and $S_p = \sigma R$ with R distribution free,

$$T^- = \frac{Z}{R}, \qquad T^+ = \frac{Z}{R} - \frac{2\delta}{\sigma R \sqrt{c}}.$$

Thus $P(R^-) = P(t_{n+m-2} \ge t) = \alpha$ for every σ_ℓ, while $T^+ \to -\infty$ in probability as $\sigma_\ell \to 0$, so $P(R^+) \to 1$. Theorem 8.3.24 therefore gives

$$\sup_{\eta \in H_0} P(R) = \alpha,$$

so the two one-sided tests procedure has size exactly α.

Problem 8.48. Prove the assertion in Example 8.3.30 that the conditional distribution of S_1 given S is hypergeometric. (In that example $S_1 \sim$ binomial(n_1, p_1) and $S_2 \sim$ binomial(n_2, p_2) are independent, $H_0 : p_1 = p_2$ is tested against $H_1 : p_1 > p_2$, and $S = S_1 + S_2$ is sufficient under H_0, where p denotes the common value of $p_1 = p_2$.)

Solution. Under H_0, $S_1 \mid S = s$ is hypergeometric$(n_1 + n_2,\, n_1,\, s)$, i.e.

$$P(S_1 = s_1 \mid S = s) = \frac{\binom{n_1}{s_1}\binom{n_2}{s - s_1}}{\binom{n_1 + n_2}{s}}, \qquad \max\{0, s - n_2\} \le s_1 \le \min\{n_1, s\}.$$

Indeed, with p the common success probability, Example 8.3.30 gives the joint pmf

$$f(s_1, s_2 \mid p) = \binom{n_1}{s_1}\binom{n_2}{s_2} p^{s_1+s_2}(1-p)^{n_1+n_2-(s_1+s_2)},$$

and $S = S_1 + S_2 \sim$ binomial(n_1+n_2, p), since by Theorem 4.6.7 its mgf is $[(1-p)+pe^t]^{n_1}[(1-p)+pe^t]^{n_2} = [(1-p)+pe^t]^{n_1+n_2}$. Hence

$$\begin{aligned} P(S_1 = s_1 \mid S = s) &= \frac{f(s_1,\, s - s_1 \mid p)}{P(S = s)} \\ &= \frac{\binom{n_1}{s_1}\binom{n_2}{s - s_1} p^s (1-p)^{n_1+n_2-s}}{\binom{n_1 + n_2}{s} p^s (1-p)^{n_1+n_2-s}} = \frac{\binom{n_1}{s_1}\binom{n_2}{s - s_1}}{\binom{n_1 + n_2}{s}}, \end{aligned}$$

the factor $p^s(1-p)^{n_1+n_2-s}$ cancelling. This is the pmf (3.2.2) of a hypergeometric random variable with population size $N = n_1 + n_2$, $M = n_1$ marked items, and sample size $K = s$.

Problem 8.49. In each of the following situations, calculate the p-value of the observed data.
(a) For testing $H_0 : \theta \le \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$, 7 successes are observed out of 10 Bernoulli trials.
(b) For testing $H_0 : \lambda \le 1$ versus $H_1 : \lambda > 1$, $X = 3$ are observed, where $X \sim$ Poisson(λ).
(c) For testing $H_0 : \lambda \le 1$ versus $H_1 : \lambda > 1$, $X_1 = 3$, $X_2 = 5$, and $X_3 = 1$ are observed, where $X_i \sim$ Poisson(λ), independent.

Solution. In each case the p-value (8.3.9) is $p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \ge W(\mathbf{x}))$, and in each case the relevant family has an MLR in the natural statistic W, so the supremum is attained at the boundary of Θ_0.
(a) $W = X \sim$ binomial$(10, \theta)$, observed $W = 7$; the binomial family has an MLR (Exercise 8.25(c)), hence is stochastically increasing (Exercise 8.26(a)), so $P_\theta(X \ge 7)$ increases in θ and

$$p = P_{1/2}(X \ge 7) = \frac{1}{2^{10}}\left[\binom{10}{7} + \binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right] = \frac{120 + 45 + 10 + 1}{1024} = \frac{176}{1024} = 0.1719.$$

(b) $W = X \sim$ Poisson(λ), observed $W = 3$; the Poisson family has an MLR (Exercise 8.25(b)), so $P_\lambda(X \ge 3)$ increases in λ and

$$p = P_1(X \ge 3) = 1 - e^{-1}\left(1 + 1 + \tfrac{1}{2}\right) = 1 - \tfrac{5}{2}e^{-1} = 0.0803.$$

(c) $W = \sum_{i=1}^3 X_i$ is sufficient and $\sum X_i \sim$ Poisson(3λ); the observed value is $3 + 5 + 1 = 9$, so

$$p = P_{\lambda=1}\left(\sum_{i=1}^3 X_i \ge 9\right) = 1 - e^{-3}\sum_{k=0}^{8}\frac{3^k}{k!} = 1 - 0.99620 = 0.0038.$$

8.8 Exercises 8.50–8.56

Problem 8.50. Let $X_1, \dots, X_n$ be iid $\mathrm{n}(\theta, \sigma^2)$, σ^2 known, and let θ have a double exponential distribution, that is,

$$\pi(\theta) = \frac{e^{-|\theta|/a}}{2a}, \qquad a \text{ known.}$$

A Bayesian test of the hypotheses $H_0\colon \theta \le 0$ versus $H_1\colon \theta > 0$ will decide in favor of H_1 if its posterior probability is large.
(a) For a given constant K, calculate the posterior probability that $\theta > K$, that is, $P(\theta > K \mid x_1, \dots, x_n, a)$.
(b) Find an expression for $\lim_{a\to\infty} P(\theta > K \mid x_1, \dots, x_n, a)$.
(c) Compare your answer in part (b) to the p-value associated with the classical hypothesis test.

Solution. Write $\sigma_n^2 = \sigma^2/n$ and

$$\mu_+ = \bar{x} - \frac{\sigma_n^2}{a}, \qquad \mu_- = \bar{x} + \frac{\sigma_n^2}{a},$$

and let Φ be the standard normal cdf. Then for $K \ge 0$

$$P(\theta > K \mid \mathbf{x}, a) = \frac{e^{-\bar{x}/a}\,\Phi\big((\mu_+ - K)/\sigma_n\big)}{e^{-\bar{x}/a}\,\Phi(\mu_+/\sigma_n) + e^{\bar{x}/a}\,\Phi(-\mu_-/\sigma_n)}.$$

(a) Since $\bar{X}$ is sufficient, $f(\mathbf{x} \mid \theta) \propto \exp\{-n(\theta - \bar{x})^2/(2\sigma^2)\}$ as a function of θ, so

$$\pi(\theta \mid \mathbf{x}, a) \propto \exp\left\{-\frac{(\theta - \bar{x})^2}{2\sigma_n^2} - \frac{|\theta|}{a}\right\}.$$

Completing the square separately on each half-line, and discarding the factor $\exp\{\sigma_n^2/(2a^2)\}$ common to both,

$$\theta > 0: \quad \exp\left\{-\frac{(\theta - \mu_+)^2}{2\sigma_n^2}\right\} e^{-\bar{x}/a},$$
$$\theta < 0: \quad \exp\left\{-\frac{(\theta - \mu_-)^2}{2\sigma_n^2}\right\} e^{\bar{x}/a}.$$

Thus the posterior is a two-piece normal mixture with weights $w_+ = e^{-\bar{x}/a}$, $w_- = e^{\bar{x}/a}$, and normalizing constant

$$m(\mathbf{x}, a) = w_+\Phi(\mu_+/\sigma_n) + w_-\Phi(-\mu_-/\sigma_n).$$

For $K \ge 0$ only the $\theta > 0$ piece contributes, giving the display above. For $K < 0$ both pieces contribute and

$$P(\theta > K \mid \mathbf{x}, a) = \frac{N_K}{m(\mathbf{x}, a)}, \quad N_K = w_+\Phi(\mu_+/\sigma_n) + w_-\big[\Phi(-\mu_-/\sigma_n) - \Phi((K - \mu_-)/\sigma_n)\big].$$

(b) As $a \to \infty$ we have $\sigma_n^2/a \to 0$, hence $\mu_\pm \to \bar{x}$ and $w_\pm \to 1$; the denominator tends to $\Phi(\bar{x}/\sigma_n) + \Phi(-\bar{x}/\sigma_n) = 1$, and both expressions in (a) collapse to the same limit

$$\lim_{a\to\infty} P(\theta > K \mid \mathbf{x}, a) = \Phi\left(\frac{\bar{x} - K}{\sigma/\sqrt{n}}\right) = P\left(Z > \frac{K - \bar{x}}{\sigma/\sqrt{n}}\right),$$

$Z \sim \mathrm{n}(0,1)$. (For $K < 0$ use $\Phi(\bar{x}/\sigma_n) + \Phi(-\bar{x}/\sigma_n) - \Phi((K - \bar{x})/\sigma_n) = 1 - \Phi((K - \bar{x})/\sigma_n)$.)
(c) Take $K = 0$, the boundary of H_0. The classical UMP test rejects for large $\bar{x}$ (Example 8.3.15), so by Example 8.3.29 its p-value at $\mathbf{x}$ is

$$p(\mathbf{x}) = P\left(Z \ge \frac{\bar{x}}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{\sqrt{n}\,\bar{x}}{\sigma}\right).$$

By (b), $\lim_{a\to\infty} P(\theta \le 0 \mid \mathbf{x}, a) = 1 - \Phi(\sqrt{n}\,\bar{x}/\sigma) = p(\mathbf{x})$: as the double exponential prior flattens, the posterior probability of H_0 is exactly the one-sided p-value.

Problem 8.51. Here is another common interpretation of p-values. Consider a problem of testing H_0 versus H_1. Let $W(\mathbf{X})$ be a test statistic. Suppose that for each α, $0 \le \alpha \le 1$, a critical value c_α can be chosen so that $\{\mathbf{x}\colon W(\mathbf{x}) \ge c_\alpha\}$ is the rejection region of a size α test of H_0. Using this family of tests, show that the usual p-value $p(\mathbf{x})$, defined by (8.3.9), is the smallest α level at which we could reject H_0, having observed the data $\mathbf{x}$.

Solution. $p(\mathbf{x}) = \inf A(\mathbf{x})$, where $A(\mathbf{x}) = \{\alpha\colon W(\mathbf{x}) \ge c_\alpha\}$ is the set of levels at which $\mathbf{x}$ leads to rejection. Here, by (8.3.9),

$$p(\mathbf{x}) = \sup_{\theta\in\Theta_0} P_\theta(W(\mathbf{X}) \ge W(\mathbf{x})),$$

and by hypothesis $\sup_{\theta\in\Theta_0} P_\theta(W(\mathbf{X}) \ge c_\alpha) = \alpha$ for every α.
$p(\mathbf{x})$ is a lower bound for $A(\mathbf{x})$: if $\alpha \in A(\mathbf{x})$ then $W(\mathbf{x}) \ge c_\alpha$, so $\{W \ge W(\mathbf{x})\} \subseteq \{W \ge c_\alpha\}$ and

$$p(\mathbf{x}) = \sup_{\theta\in\Theta_0} P_\theta\big(W \ge W(\mathbf{x})\big) \le \sup_{\theta\in\Theta_0} P_\theta\big(W \ge c_\alpha\big) = \alpha.$$

$p(\mathbf{x})$ is the greatest lower bound: let $\alpha > p(\mathbf{x})$ and suppose $\alpha \notin A(\mathbf{x})$, i.e. $W(\mathbf{x}) < c_\alpha$. Then $\{W \ge c_\alpha\} \subseteq \{W \ge W(\mathbf{x})\}$, whence

$$\alpha = \sup_{\theta\in\Theta_0} P_\theta\big(W \ge c_\alpha\big) \le \sup_{\theta\in\Theta_0} P_\theta\big(W \ge W(\mathbf{x})\big) = p(\mathbf{x}),$$

contradicting $\alpha > p(\mathbf{x})$. Hence $(p(\mathbf{x}), 1] \subseteq A(\mathbf{x})$, and $\inf A(\mathbf{x}) = p(\mathbf{x})$.

Problem 8.52. Consider testing $H_0\colon \theta \in \bigcup_{j=1}^k \Theta_j$. For each $j = 1, \dots, k$, let $p_j(\mathbf{x})$ denote a valid p-value for testing $H_{0j}\colon \theta \in \Theta_j$. Let $p(\mathbf{x}) = \max_{1\le j\le k} p_j(\mathbf{x})$.
(a) Show that $p(\mathbf{X})$ is a valid p-value for testing H_0.
(b) Show that the α level test defined by $p(\mathbf{X})$ is the same as an α level IUT defined in terms of individual tests based on the $p_j(\mathbf{x})$s.

Solution. (a) Fix $\theta \in \Theta_0 = \bigcup_{j=1}^k \Theta_j$ and $0 \le \alpha \le 1$. Then $\theta \in \Theta_{j_0}$ for some j_0, and since $p(\mathbf{x}) \le \alpha$ forces $p_{j_0}(\mathbf{x}) \le \alpha$,

$$P_\theta\big(p(\mathbf{X}) \le \alpha\big) \le P_\theta\big(p_{j_0}(\mathbf{X}) \le \alpha\big) \le \alpha,$$

the last inequality because p_{j_0} is a valid p-value for H_{0j_0} (Definition 8.3.26, (8.3.8)) and $\theta \in \Theta_{j_0}$. Also $0 \le p(\mathbf{x}) \le 1$ as a maximum of such numbers. So $p(\mathbf{X})$ is a valid p-value for H_0.
(b) The test based on p has rejection region

$$R = \{\mathbf{x}\colon p(\mathbf{x}) \le \alpha\} = \Big\{\mathbf{x}\colon \max_{1\le j\le k} p_j(\mathbf{x}) \le \alpha\Big\} = \bigcap_{j=1}^k R_j, \quad R_j = \{\mathbf{x}\colon p_j(\mathbf{x}) \le \alpha\},$$

because the maximum is $\le \alpha$ exactly when every p_j is. This is precisely the intersection–union form: H_0 is rejected if and only if each individual H_{0j} is rejected by its own p-value test. Each R_j is a level α test of H_{0j} (validity of p_j), so by Theorem 8.3.23 the IUT with rejection region $\bigcap_j R_j$ is a level α test of H_0 — the same conclusion (a) gives for the p-value test.

Problem 8.53. In Example 8.2.7 we saw an example of a one-sided Bayesian hypothesis test. Now we will consider a similar situation, but with a two-sided test. We want to test

$$H_0\colon \theta = 0 \qquad \text{versus} \qquad H_1\colon \theta \ne 0,$$

and we observe $X_1, \dots, X_n$, a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population, σ^2 known. A type of prior distribution that is often used in this situation is a mixture of a point mass on $\theta = 0$ and a pdf spread out over H_1. A typical choice is to take $P(\theta = 0) = \frac{1}{2}$, and if $\theta \ne 0$, take the prior distribution to be

$\frac{1}{2}\,\mathrm{n}(0,\tau^2)$, where τ^2 is known.
(a) Show that the prior defined above is proper, that is, $P(-\infty < \theta < \infty) = 1$.
(b) Calculate the posterior probability that H_0 is true, $P(\theta = 0 \mid x_1, \ldots, x_n)$.
(c) Find an expression for the p-value corresponding to a value of $\bar{x}$.
(d) For the special case $\sigma^2 = \tau^2 = 1$, compare $P(\theta = 0 \mid x_1, \ldots, x_n)$ and the p-value for a range of values of $\bar{x}$. In particular,
(i) For $n = 9$, plot the p-value and posterior probability as a function of $\bar{x}$, and show that the Bayes probability is greater than the p-value for moderately large values of $\bar{x}$.
(ii) Now, for $\alpha = .05$, set $\bar{x} = z_{\alpha/2}/\sqrt{n}$, fixing the p-value at α for all n. Show that the posterior probability at $\bar{x} = z_{\alpha/2}/\sqrt{n}$ goes to 1 as $n \to \infty$. This is Lindley's Paradox.
Note that small values of $P(\theta = 0 \mid x_1, \ldots, x_n)$ are evidence against H_0, and thus this quantity is similar in spirit to a p-value. The fact that these two quantities can have very different values was noted by Lindley (1957) and is also examined by Berger and Sellke (1987).

Solution. Write $\sigma_n^2 = \sigma^2/n$ and $Z = \sqrt{n}\,\bar{x}/\sigma = \bar{x}/\sigma_n$. The answer to (b) is

$$P(\theta = 0 \mid \mathbf{x}) = \left[1 + \left(\frac{\sigma_n^2}{\sigma_n^2 + \tau^2}\right)^{1/2} \exp\left\{\frac{\tau^2 Z^2}{2(\sigma_n^2 + \tau^2)}\right\}\right]^{-1}.$$

(a) The prior assigns $P(\theta = 0) = \frac{1}{2}$ and, off 0, total mass $\frac{1}{2}\int_{-\infty}^{\infty}(2\pi\tau^2)^{-1/2}e^{-\theta^2/(2\tau^2)}\,d\theta = \frac{1}{2}$; the two disjoint pieces sum to 1.
(b) Since $\sum_i (x_i - \theta)^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2$, the factor $C(\mathbf{x}) = (2\pi\sigma^2)^{-n/2}e^{-\sum(x_i-\bar{x})^2/(2\sigma^2)}\sqrt{2\pi\sigma_n^2}$ is common to both pieces and

$$f(\mathbf{x} \mid 0) = C(\mathbf{x})\,\varphi(\bar{x}; 0, \sigma_n^2),$$
$$\int f(\mathbf{x} \mid \theta)\,\mathrm{n}(\theta \mid 0, \tau^2)\,d\theta = C(\mathbf{x})\,\varphi(\bar{x}; 0, \sigma_n^2 + \tau^2),$$

where $\varphi(\cdot; 0, v)$ is the $\mathrm{n}(0, v)$ pdf; the second line is the convolution $\bar{X} \mid \theta \sim \mathrm{n}(\theta, \sigma_n^2)$ with $\theta \sim \mathrm{n}(0, \tau^2)$, giving $\bar{X} \sim \mathrm{n}(0, \sigma_n^2 + \tau^2)$. Hence, with the $\frac{1}{2}$s and $C(\mathbf{x})$ cancelling,

$$P(\theta = 0 \mid \mathbf{x}) = \left[1 + \frac{\varphi(\bar{x}; 0, \sigma_n^2 + \tau^2)}{\varphi(\bar{x}; 0, \sigma_n^2)}\right]^{-1},$$

and since

$$\frac{\varphi(\bar{x}; 0, \sigma_n^2 + \tau^2)}{\varphi(\bar{x}; 0, \sigma_n^2)} = \sqrt{\frac{\sigma_n^2}{\sigma_n^2 + \tau^2}}\,\exp\left\{\frac{\bar{x}^2}{2} \cdot \frac{\tau^2}{\sigma_n^2(\sigma_n^2 + \tau^2)}\right\},$$

substituting $\bar{x}^2 = \sigma_n^2 Z^2$ gives the boxed formula.
(c) The two-sided test rejects for large $|\bar{x}|$, so by Example 8.3.28 (with σ known, so the null distribution is standard normal)

$$p(\mathbf{x}) = 2\,P\!\left(Z' \geq \frac{\sqrt{n}\,|\bar{x}|}{\sigma}\right) = 2\big(1 - \Phi(|Z|)\big).$$

(d) With $\sigma^2 = \tau^2 = 1$ we have $\sigma_n^2 = 1/n$, $\sigma_n^2 + \tau^2 = (n+1)/n$, so

$$P(\theta = 0 \mid \mathbf{x}) = \left[1 + \frac{1}{\sqrt{n+1}}\,e^{\,nZ^2/(2(n+1))}\right]^{-1}, \qquad Z = \sqrt{n}\,\bar{x}.$$

(i) For $n = 9$, $Z = 3\bar{x}$ and $P(\theta = 0 \mid \mathbf{x}) = [1 + 10^{-1/2}e^{9Z^2/20}]^{-1}$:

$\bar{x}$	$Z = 3\bar{x}$	p-value	$P(\theta = 0 \mid \mathbf{x})$
0.0	0.00	1.000	.760
0.1	0.30	.764	.752
0.2	0.60	.549	.729
0.3	0.90	.368	.687
0.4	1.20	.230	.623
0.5	1.50	.134	.535
0.653	1.96	.050	.360
0.8	2.40	.016	.191
1.0	3.00	.003	.052

The two curves cross at $Z \approx .32$ ($\bar{x} \approx .105$); beyond that the posterior probability exceeds the p-value, and the gap is largest in the range that classical practice calls significant: at the 5% cutoff $\bar{x} = .653$ the p-value is .05 while the posterior probability of H_0 is .36, seven times larger.
(ii) Fix $\alpha = .05$ and $\bar{x} = z_{.025}/\sqrt{n}$, so $Z = z_{.025} = 1.96$ and the p-value equals .05 for every n. Then

$$P(\theta = 0 \mid \mathbf{x}) = \left[1 + \frac{1}{\sqrt{n+1}}\, e^{n(1.96)^2/(2(n+1))}\right]^{-1}.$$

The exponent increases to the finite limit $(1.96)^2/2 = 1.921$, while $(n+1)^{-1/2} \to 0$, so the second term in the bracket tends to $0 \cdot e^{1.921} = 0$ and

$$\lim_{n\to\infty} P(\theta = 0 \mid \mathbf{x}) = 1.$$

Numerically: .33 at $n = 5$, .42 at $n = 20$, .60 at $n = 100$, .82 at $n = 1000$, .94 at $n = 10{,}000$.

Problem 8.54. The discrepancies between p-values and Bayes posterior probabilities are not as dramatic in the one-sided problem, as is discussed by Casella and Berger (1987) and also mentioned in the Miscellanea section. Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population, and suppose that the hypotheses to be tested are

$$H_0\colon \theta \le 0 \qquad \text{versus} \qquad H_1\colon \theta > 0.$$

The prior distribution on θ is $\mathrm{n}(0, \tau^2)$, τ^2 known, which is symmetric about the hypotheses in the sense that $P(\theta \le 0) = P(\theta > 0) = \frac{1}{2}$.
(a) Calculate the posterior probability that H_0 is true, $P(\theta \le 0 \mid x_1, \ldots, x_n)$.
(b) Find an expression for the p-value corresponding to a value of $\bar{x}$, using tests that reject for large values of $\bar{X}$.
(c) For the special case $\sigma^2 = \tau^2 = 1$, compare $P(\theta \le 0 \mid x_1, \ldots, x_n)$ and the p-value for values of $\bar{x} > 0$. Show that the Bayes probability is always greater than the p-value.
(d) Using the expressions derived in parts (a) and (b), show that

$$\lim_{\tau^2\to\infty} P(\theta \le 0 \mid x_1, \ldots, x_n) = \text{p-value},$$

an equality that does not occur in the two-sided problem.

Solution. With $\sigma_n^2 = \sigma^2/n$ and $Z = \sqrt{n}\,\bar{x}/\sigma$,

$$P(\theta \le 0 \mid \mathbf{x}) = \Phi\left(-Z\,\frac{\tau}{\sqrt{\tau^2 + \sigma_n^2}}\right),$$

which exceeds the p-value $\Phi(-Z)$ for $\bar{x} > 0$ and converges to it as $\tau^2 \to \infty$.
(a) $\bar{X}$ is sufficient with $\bar{X} \mid \theta \sim \mathrm{n}(\theta, \sigma_n^2)$, so by Example 7.2.16 (with $\mu = 0$) the posterior is normal with

$$E(\theta \mid \mathbf{x}) = \frac{\tau^2}{\tau^2 + \sigma_n^2}\,\bar{x}, \qquad \mathrm{Var}(\theta \mid \mathbf{x}) = \frac{\tau^2\sigma_n^2}{\tau^2 + \sigma_n^2}.$$

Hence

$$P(\theta \le 0 \mid \mathbf{x}) = \Phi\left(\frac{-E(\theta \mid \mathbf{x})}{\sqrt{\mathrm{Var}(\theta \mid \mathbf{x})}}\right)$$
$$= \Phi\left(-\frac{\bar{x}\,\tau^2/(\tau^2+\sigma_n^2)}{\tau\sigma_n/\sqrt{\tau^2+\sigma_n^2}}\right)$$
$$= \Phi\left(-\frac{\bar{x}}{\sigma_n}\cdot\frac{\tau}{\sqrt{\tau^2+\sigma_n^2}}\right) = \Phi\left(-Z\,\frac{\tau}{\sqrt{\tau^2+\sigma_n^2}}\right).$$

(b) The test rejects for large $\bar{X}$, and by Example 8.3.29 (with σ known the supremum over $\theta \le 0$ is attained at $\theta = 0$) the p-value is

$$p(\mathbf{x}) = P\left(Z' \ge \frac{\sqrt{n}\,\bar{x}}{\sigma}\right) = 1 - \Phi(Z) = \Phi(-Z).$$

(c) For $\sigma^2 = \tau^2 = 1$, $\sigma_n^2 = 1/n$ and $\tau/\sqrt{\tau^2+\sigma_n^2} = \sqrt{n/(n+1)} < 1$, so for $\bar{x} > 0$ (i.e. $Z > 0$)

$$-Z\sqrt{\tfrac{n}{n+1}} > -Z,$$
$$P(\theta \le 0 \mid \mathbf{x}) = \Phi\left(-Z\sqrt{\tfrac{n}{n+1}}\right) > \Phi(-Z) = p(\mathbf{x}),$$

strictly, by strict monotonicity of Φ. The two agree at $Z = 0$ (both $\frac{1}{2}$) and the discrepancy stays small; for $n = 9$:

Z	p-value	$P(\theta \le 0 \mid \mathbf{x})$
0.00	.500	.500
0.50	.309	.318
1.00	.159	.171
1.645	.050	.059
1.96	.025	.032
2.50	.006	.009
3.00	.001	.002

(d) $\tau/\sqrt{\tau^2+\sigma_n^2} = (1+\sigma_n^2/\tau^2)^{-1/2} \to 1$ as $\tau^2 \to \infty$, so by continuity of Φ,

$$\lim_{\tau^2\to\infty} P(\theta \le 0 \mid \mathbf{x}) = \Phi(-Z) = p(\mathbf{x}).$$

In the two-sided problem of Exercise 8.53 the corresponding limit is $\lim_{\tau^2\to\infty} P(\theta = 0 \mid \mathbf{x}) = 1$ for every $\bar{x}$, since the ratio in 8.53(b) tends to 0.

Problem 8.55. Let X have a $\mathrm{n}(\theta, 1)$ distribution, and consider testing $H_0\colon \theta \ge \theta_0$ versus $H_1\colon \theta < \theta_0$. Use the loss function (8.3.13),

$$L(\theta, a_0) = \begin{cases} 0 & \theta \ge \theta_0, \\ b(\theta_0 - \theta) & \theta < \theta_0, \end{cases}$$
$$L(\theta, a_1) = \begin{cases} c(\theta - \theta_0)^2 & \theta \ge \theta_0, \\ 0 & \theta < \theta_0, \end{cases}$$

and investigate the three tests that reject H_0 if $X < -z_\alpha + \theta_0$ for $\alpha = .1$, $.3$, and $.5$.
(a) For $b = c = 1$, graph and compare their risk functions.
(b) For $b = 3$, $c = 1$, graph and compare their risk functions.
(c) Graph and compare the power functions of the three tests to the risk functions in parts (a) and (b).

Solution. Writing $t = \theta - \theta_0$, the power function of the test with cutoff $-z_\alpha + \theta_0$ is

$$\beta_\alpha(\theta) = P_\theta(X < -z_\alpha + \theta_0) = \Phi(-z_\alpha - t),$$

and by (8.3.14) applied to the loss (8.3.13),

$$R(\theta, \delta_\alpha) = \begin{cases} c\, t^2\, \Phi(-z_\alpha - t), & t \geq 0 \quad (\theta \in \Theta_0), \\ b\, |t|\, \Phi(z_\alpha + t), & t < 0 \quad (\theta \in \Theta_0^c), \end{cases}$$

with $z_{.1} = 1.282$, $z_{.3} = .524$, $z_{.5} = 0$.
(a) $b = c = 1$:

$t = \theta - \theta_0$	$\alpha = .1$	$\alpha = .3$	$\alpha = .5$
-3.0	.129	.020	.004
-2.0	.472	.140	.046
-1.5	.620	.247	.100
-1.0	.611	.317	.159
-0.5	.391	.255	.154
0.0	0	0	0
0.5	.009	.038	.077
1.0	.011	.064	.159
1.5	.006	.048	.150
2.0	.002	.023	.091
3.0	.000	.002	.012

No test dominates: on Θ_0 ($t > 0$) the risks are ordered $R_{.1} < R_{.3} < R_{.5}$, on Θ_0^c ($t < 0$) exactly the reverse. Each risk is unimodal on either side and vanishes at $t = 0$ and as $|t| \to \infty$ — the factor t^2 (resp. $|t|$) kills the discontinuity at θ_0 seen in Figure 8.3.4 under generalized 0–1 loss, because at θ_0 both decisions cost nothing. The maximum risks are

$$\sup_\theta R(\theta, \delta_{.1}) = .641, \quad \sup_\theta R(\theta, \delta_{.3}) = .319, \quad \sup_\theta R(\theta, \delta_{.5}) = .170,$$

attained at $t = -1.27,\ -0.93,\ -0.75$; so $\delta_{.5}$ is the minimax choice of the three, the $|t|$-weighted Type II loss being far more expensive than the t^2-weighted Type I loss near θ_0.
(b) $b = 3$, $c = 1$: the $t \geq 0$ half is unchanged and the $t < 0$ half is tripled.

$t = \theta - \theta_0$	$\alpha = .1$	$\alpha = .3$	$\alpha = .5$
-3.0	.386	.060	.012
-2.0	1.417	.420	.137
-1.5	1.861	.741	.301
-1.0	1.833	.952	.476
-0.5	1.174	.765	.463
0.0	0	0	0
0.5	.009	.038	.077
1.0	.011	.064	.159
1.5	.006	.048	.150
2.0	.002	.023	.091
3.0	.000	.002	.012

The qualitative picture is the same but the case for a large α is stronger: $\sup_\theta R$ becomes 1.923, .956, .510 at the same three locations, and the region where $\delta_{.1}$ is preferable ($t > 0$) now buys at most .16 of risk while costing over 1.3 at $t = -1$.
(c) The power functions $\beta_\alpha(\theta) = \Phi(-z_\alpha - t)$ are decreasing in t, equal α at $t = 0$, and are ordered $\beta_{.1} < \beta_{.3} < \beta_{.5}$ uniformly:

$t = \theta - \theta_0$	$\beta_{.1}$	$\beta_{.3}$	$\beta_{.5}$
-3.0	.957	.993	.999
-2.0	.764	.930	.977
-1.0	.389	.683	.841
0.0	.100	.300	.500
1.0	.011	.064	.159
2.0	.001	.006	.023
3.0	.000	.000	.001

The risk functions are exactly these curves reweighted: $R = c\,t^2\beta_\alpha$ on Θ_0 and $R = b|t|(1-\beta_\alpha)$ on Θ_0^c. Hence the uniform ordering of the power functions is what produces the reversed orderings of the risks on the two halves; and because the weights t^2 and $b|t|$ vanish at θ_0, the risk curves — unlike the power curves — carry no information about the size $\alpha = \beta_\alpha(\theta_0)$ of the tests.

Problem 8.56. Consider testing $H_0\colon p \le \frac{1}{3}$ versus $H_1\colon p > \frac{1}{3}$, where $X \sim \text{binomial}(5, p)$, using 0–1 loss. Graph and compare the risk functions for the following two tests. Test I rejects H_0 if $X = 0$ or 1. Test II rejects H_0 if $X = 4$ or 5.

Solution. Test II is the sensible one: its risk stays below .955, while Test I has risk 1 at $p = 0$, the worst possible value, and at the parameter point most favourable to H_0.
Under 0–1 loss, (8.3.12) with $c_{\text{I}} = c_{\text{II}} = 1$ gives

$$R(p,\delta) = \begin{cases} \beta(p), & p \le \frac{1}{3}, \\ 1-\beta(p), & p > \frac{1}{3}, \end{cases}$$

where the two power functions are

$$\beta_{\text{I}}(p) = P_p(X \le 1) = (1-p)^5 + 5p(1-p)^4,$$
$$\beta_{\text{II}}(p) = P_p(X \ge 4) = 5p^4(1-p) + p^5.$$

β_{I} is strictly decreasing and β_{II} strictly increasing, with $\beta_{\text{I}}(p) = \beta_{\text{II}}(1-p)$.

p	β_{I}	β_{II}	R_{I}	R_{II}
0	1.000	.000	1.000	.000
.1	.919	.001	.919	.001
.2	.737	.007	.737	.007
.3	.528	.031	.528	.031
1/3	.461	.045	.461	.045
.4	.337	.087	.663	.913
.5	.188	.188	.812	.812
.6	.087	.337	.913	.663
.8	.007	.737	.993	.263
1	.000	1.000	1.000	.000

Both risks jump at $p = \frac{1}{3}$, where the loss switches from Type I to Type II error: R_{I} falls from 1 to .461 on $[0, \frac{1}{3}]$, jumps up to .539, then rises to 1; R_{II} rises from 0 to .045 on $[0, \frac{1}{3}]$, jumps to .955, then falls to 0.
Neither dominates, since the risks cross at $p = \frac{1}{2}$:

$$\begin{aligned} p \le \tfrac{1}{3}&: \quad R_{\text{II}} = \beta_{\text{II}} < \beta_{\text{I}} = R_{\text{I}}, \\ \tfrac{1}{3} < p < \tfrac{1}{2}&: \quad R_{\text{I}} = 1-\beta_{\text{I}} < 1-\beta_{\text{II}} = R_{\text{II}}, \\ p > \tfrac{1}{2}&: \quad R_{\text{II}} < R_{\text{I}}, \end{aligned}$$

using $\beta_{\text{I}}(p) \gtrless \beta_{\text{II}}(p)$ according as $p \lessgtr \frac{1}{2}$. But Test I's advantage is confined to the strip $(\frac{1}{3}, \frac{1}{2})$, of width $\frac{1}{6}$, while its excess risk equals 1 at both $p = 0$ and $p = 1$; comparing maxima, $\sup_p R_{\text{I}} = 1$ against $\sup_p R_{\text{II}} = .955$. Test I rejects H_0 precisely when the data most strongly support it, so it is worthless.

8.9 Exercises 8.57–8.58

Problem 8.57. Consider testing $H_0 : \mu \le 0$ versus $H_1 : \mu > 0$ using 0-1 loss, where $X \sim \mathrm{n}(\mu, 1)$. Let δ_c be the test that rejects H_0 if $X > c$. For every test in this problem, there is a δ_c in the class of tests $\{\delta_c, -\infty \le c \le \infty\}$ that has a uniformly smaller (in μ) risk function. Let δ be the test that rejects H_0 if $1 < X < 2$. Find a test δ_c that is better than δ. (Either prove that the test is better or graph the risk functions for δ and δ_c and carefully explain why the proposed test should be better.)

Solution. Take δ_{c^*} with $c^* = 1.0989$, the unique c matching the two rejection probabilities at $\mu = 0$:

$$\Phi(-c^*) \;=\; \Phi(2) - \Phi(1) \;=\; 0.1359 \;=:\; \alpha_0.$$

Under 0-1 loss the risk of a test with power function β is, by (8.3.12) with $c_{\mathrm{I}} = c_{\mathrm{II}} = 1$,

$$R(\mu, \delta) = \begin{cases} \beta(\mu), & \mu \le 0, \\ 1 - \beta(\mu), & \mu > 0, \end{cases}$$

so dominating δ means having smaller power for $\mu \le 0$ and larger power for $\mu > 0$. Here

$$\beta_\delta(\mu) = \Phi(2-\mu) - \Phi(1-\mu),$$
$$\beta_{c^*}(\mu) = P_\mu(X > c^*) = \Phi(\mu - c^*),$$

and $\beta_\delta(0) = \beta_{c^*}(0) = \alpha_0$ by the choice of c^*. Since $\beta_\delta'(\mu) = \phi(1-\mu) - \phi(2-\mu) > 0$ for $\mu < \frac{3}{2}$, both tests have size $\sup_{\mu \le 0} \beta = \alpha_0$.
$\mu > 0$. The family $\mathrm{n}(\mu, 1)$ is a one-parameter exponential family in which the natural parameter μ is increasing, so X has an MLR; δ_{c^*} rejects for large X and has size α_0, hence is UMP level α_0 by Theorem 8.3.17 (Karlin-Rubin). Therefore

$$\beta_{c^*}(\mu) \ge \beta_\delta(\mu) \quad \Longrightarrow \quad R(\mu, \delta_{c^*}) \le R(\mu, \delta).$$

$\mu < 0$. Fix $\mu_1 < 0$ and test $H_0 : \mu = 0$ versus $H_1 : \mu = \mu_1$. The ratio

$$\frac{f(x \mid \mu_1)}{f(x \mid 0)} = e^{\mu_1 x - \mu_1^2/2}$$

is **decreasing** in x, so $\{x : x < c^*\}$ is exactly a Neyman-Pearson region $\{f(x \mid \mu_1) > k f(x \mid 0)\}$, and it has P_0-probability $1 - \alpha_0$. By Theorem 8.3.12 (Neyman-Pearson Lemma) it is therefore most powerful of level $1 - \alpha_0$: for every set A with $P_0(A) = \alpha_0$,

$$P_{\mu_1}(X < c^*) \ge P_{\mu_1}(A^c), \quad \text{i.e.} \quad \beta_{c^*}(\mu_1) \le P_{\mu_1}(A).$$

Taking $A = \{1 < X < 2\}$, which has $P_0(A) = \alpha_0$, gives $\beta_{c^*}(\mu_1) \le \beta_\delta(\mu_1)$; the inequality is strict by part (b) (necessity) of Theorem 8.3.12, since A^c is not of the form $\{x < c\}$ up to a null set and so cannot be most powerful. Hence $R(\mu, \delta_{c^*}) < R(\mu, \delta)$ for $\mu < 0$, while at $\mu = 0$ both risks equal α_0. Thus δ_{c^*} dominates δ, strictly for every $\mu \ne 0$.

Problem 8.58. Consider the hypothesis testing problem and loss function given in Example 8.3.31, and let $\sigma = n = 1$. Consider tests that reject H_0 if $X < -z_\alpha + \theta_0$. Find the value of α that minimizes the maximum value of the risk function, that is, that yields a **minimax** test.
(Example 8.3.31: $X_1, \ldots, X_n$ is a random sample from an $\mathrm{n}(\theta, \sigma^2)$ population with σ^2 known, and one tests $H_0 : \theta \ge \theta_0$ versus $H_1 : \theta < \theta_0$ under the generalized 0-1 loss (8.3.11),

$$L(\theta, a_0) = \begin{cases} 0, & \theta \in \Theta_0, \\ c_{\mathrm{II}}, & \theta \in \Theta_0^c, \end{cases} \qquad L(\theta, a_1) = \begin{cases} c_{\mathrm{I}}, & \theta \in \Theta_0, \\ 0, & \theta \in \Theta_0^c, \end{cases}$$

with $c_{\mathrm{I}} = 8$ and $c_{\mathrm{II}} = 3$, where $\Theta_0 = \{\theta : \theta \geq \theta_0\}$.)

Solution.

$$\alpha^* = \frac{c_{\mathrm{II}}}{c_{\mathrm{I}} + c_{\mathrm{II}}} = \frac{3}{11} = .2727,$$

i.e. reject H_0 if $X < \theta_0 - .6046$, with minimax risk $24/11 = 2.182$.
With $\sigma = n = 1$ the test has power function

$$\beta(\theta) = P_\theta\big(X < \theta_0 - z_\alpha\big) = \Phi\big(\theta_0 - \theta - z_\alpha\big),$$

so by (8.3.12) the risk is

$$R(\theta, \delta_\alpha) = \begin{cases} c_{\mathrm{I}}\,\Phi(\theta_0 - \theta - z_\alpha), & \theta \geq \theta_0, \\ c_{\mathrm{II}}\big[1 - \Phi(\theta_0 - \theta - z_\alpha)\big], & \theta < \theta_0. \end{cases}$$

On $\theta \geq \theta_0$ the argument $\theta_0 - \theta - z_\alpha$ decreases as θ grows, so that branch is maximized at $\theta = \theta_0$; on $\theta < \theta_0$ the same argument decreases as $\theta \uparrow \theta_0$, so that branch increases to its supremum there. Hence

$$\begin{aligned} \sup_{\theta \geq \theta_0} R &= c_{\mathrm{I}}\Phi(-z_\alpha) = c_{\mathrm{I}}\,\alpha, \\ \sup_{\theta < \theta_0} R &= c_{\mathrm{II}}\big[1 - \Phi(-z_\alpha)\big] = c_{\mathrm{II}}(1 - \alpha), \end{aligned}$$

the second a supremum not attained (this is the discontinuity at θ_0 visible in Figure 8.3.4). Therefore

$$\sup_\theta R(\theta, \delta_\alpha) = \max\big\{c_{\mathrm{I}}\alpha,\ c_{\mathrm{II}}(1 - \alpha)\big\}.$$

The first term increases in α from 0, the second decreases to 0, so the maximum of the two is minimized where they cross:

$$c_{\mathrm{I}}\alpha = c_{\mathrm{II}}(1 - \alpha) \iff \alpha^* = \frac{c_{\mathrm{II}}}{c_{\mathrm{I}} + c_{\mathrm{II}}}.$$

With $c_{\mathrm{I}} = 8$, $c_{\mathrm{II}} = 3$ this gives $z_{\alpha^*} = .6046$ and minimax value $8 \cdot \frac{3}{11} = \frac{24}{11}$.

9 Interval Estimation

Contents

9.1 Exercises 9.1–9.7 . . . 328
9.2 Exercises 9.8–9.14 . . . 334
9.3 Exercises 9.15–9.21 . . . 339
9.4 Exercises 9.22–9.28 . . . 345
9.5 Exercises 9.29–9.35 . . . 352
9.6 Exercises 9.36–9.42 . . . 358
9.7 Exercises 9.43–9.49 . . . 362
9.8 Exercises 9.50–9.56 . . . 368
9.9 Exercises 9.57–9.58 . . . 374

9.1 Exercises 9.1–9.7

Problem 9.1. If $L(x)$ and $U(x)$ satisfy $P_\theta(L(X) \le \theta) = 1 - \alpha_1$ and $P_\theta(U(X) \ge \theta) = 1 - \alpha_2$, and $L(x) \le U(x)$ for all x, show that $P_\theta(L(X) \le \theta \le U(X)) = 1 - \alpha_1 - \alpha_2$.

Solution. The two one-sided events cover the whole sample space: if $L(x) > \theta$ then $U(x) \ge L(x) > \theta$, so $\{L(X) \le \theta\}^c \subset \{U(X) \ge \theta\}$ and

$$P_\theta\big(\{L(X) \le \theta\} \cup \{U(X) \ge \theta\}\big) = 1.$$

Inclusion-exclusion on these two events therefore gives

$$\begin{aligned} P_\theta(L(X) \le \theta \le U(X)) &= P_\theta(L(X) \le \theta) + P_\theta(U(X) \ge \theta) - 1 \\ &= (1 - \alpha_1) + (1 - \alpha_2) - 1 \\ &= 1 - \alpha_1 - \alpha_2. \end{aligned}$$

Problem 9.2. Let $X_1, \dots, X_n$ be iid $\mathrm{n}(\theta, 1)$. A 95% confidence interval for θ is $\bar{x} \pm 1.96/\sqrt{n}$. Let p denote the probability that an additional independent observation, X_{n+1}, will fall in this interval. Is p greater than, less than, or equal to .95? Prove your answer.

Solution. Less than .95, for every n; indeed

$$p = 2\Phi\left(\frac{1.96}{\sqrt{n+1}}\right) - 1.$$

Since X_{n+1} is independent of $\bar{X}$ and $\mathrm{Var}(X_{n+1} - \bar{X}) = 1 + 1/n = (n+1)/n$, the standardized difference $Z = (X_{n+1} - \bar{X})/\sqrt{(n+1)/n}$ is $\mathrm{n}(0,1)$, so

$$\begin{aligned} p &= P\Big(\bar{X} - \tfrac{1.96}{\sqrt{n}} \le X_{n+1} \le \bar{X} + \tfrac{1.96}{\sqrt{n}}\Big) \\ &= P\left(|Z| \le \frac{1.96/\sqrt{n}}{\sqrt{(n+1)/n}}\right) = P\left(|Z| \le \frac{1.96}{\sqrt{n+1}}\right). \end{aligned}$$

For every $n \ge 1$ we have $1.96/\sqrt{n+1} \le 1.96/\sqrt{2} < 1.96$, and $P(|Z| \le z)$ is strictly increasing in z, so $p < P(|Z| \le 1.96) = .95$.

Problem 9.3. The independent random variables $X_1, \dots, X_n$ have the common distribution

$$P(X_i \le x \mid \alpha, \beta) = \begin{cases} 0 & \text{if } x \le 0, \\ (x/\beta)^\alpha & \text{if } 0 < x < \beta, \\ 1 & \text{if } x \ge \beta, \end{cases}$$

where the parameters α and β are positive.
(a) In Exercise 7.10 the MLEs of α and β were found. If α is a known constant α_0, find an upper confidence limit for β with confidence coefficient .95.
(b) Use the data of Exercise 7.10 to construct an interval estimate for β. Assume that α is known and equal to its MLE.
For reference, Exercise 7.10 gives the MLEs

$$\hat{\beta} = X_{(n)} = \max_i X_i, \qquad \hat{\alpha} = \frac{n}{\sum_{i=1}^n \log(\hat{\beta}/X_i)},$$

and its part (c) data are the lengths (in millimetres) of $n = 14$ cuckoos' eggs found in hedge sparrow nests:

22.0	23.9	20.9	23.8	25.0	24.0	21.7
23.8	22.8	23.1	23.1	23.5	23.0	23.0

Solution. (a) The .95 upper confidence limit is

$$U(\mathbf{X}) = X_{(n)}\, 20^{1/(n\alpha_0)}, \qquad X_{(n)} = \max_i X_i.$$

The largest order statistic, rescaled, is a pivot: for $0 < t < 1$,

$$P_\beta\left(\frac{X_{(n)}}{\beta} \le t\right) = \prod_{i=1}^{n} P_\beta(X_i \le t\beta) = \left(t^{\alpha_0}\right)^n = t^{n\alpha_0},$$

a distribution free of β (Definition 9.2.6). Hence with $c = (.05)^{1/(n\alpha_0)}$,

$$P_\beta\left(\frac{X_{(n)}}{\beta} \ge c\right) = 1 - c^{n\alpha_0} = .95,$$

and $X_{(n)}/\beta \ge c \iff \beta \le X_{(n)}/c = X_{(n)}(.05)^{-1/(n\alpha_0)}$, which is the stated U since $(.05)^{-1} = 20$. Because $\beta \ge X_{(n)}$ with probability one, the associated interval estimate is $[X_{(n)}, U(\mathbf{X})]$, also with coverage .95.
(b) For the $n = 14$ egg lengths, $\hat\beta = X_{(14)} = 25.0$ and

$$\sum_{i=1}^{14} \log\frac{25.0}{x_i} = 1.11156, \qquad \hat\alpha = \frac{14}{1.11156} = 12.5949.$$

Taking $\alpha_0 = \hat\alpha$ gives $n\alpha_0 = 14(12.5949) = 176.33$ and

$$20^{1/(n\alpha_0)} = \exp\left(\frac{\log 20}{176.33}\right) = e^{.016989} = 1.01713,$$

so

$$U(\mathbf{x}) = 25.0 \times 1.01713 = 25.43,$$

and the interval estimate for β is [25.0, 25.43] millimetres. (Substituting $\hat\alpha$ for the unknown α makes the stated .95 coefficient approximate rather than exact.)

Problem 9.4. Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(0, \sigma_X^2)$, and let $Y_1, \ldots, Y_m$ be a random sample from a $\mathrm{n}(0, \sigma_Y^2)$, independent of the Xs. Define $\lambda = \sigma_Y^2/\sigma_X^2$.
(a) Find the level α LRT of $H_0\colon \lambda = \lambda_0$ versus $H_1\colon \lambda \ne \lambda_0$.
(b) Express the rejection region of the LRT of part (a) in terms of an F random variable.
(c) Find a $1 - \alpha$ confidence interval for λ.

Solution. Here Λ is the likelihood ratio statistic (kept apart from the parameter λ), and

$$T_X = \sum_{i=1}^{n} X_i^2, \qquad T_Y = \sum_{j=1}^{m} Y_j^2, \qquad N = n + m.$$

(a) The LRT rejects H_0 for both small and large values of

$$F = \frac{T_Y/(m\lambda_0)}{T_X/n}.$$

Since the means are known to be 0, the unrestricted MLEs are $\hat\sigma_X^2 = T_X/n$ and $\hat\sigma_Y^2 = T_Y/m$, giving

$$\sup L = (2\pi)^{-N/2}(T_X/n)^{-n/2}(T_Y/m)^{-m/2}e^{-N/2}.$$

Under H_0, $\sigma_Y^2 = \lambda_0 \sigma_X^2$, so with $\theta = \sigma_X^2$,

$$\log L = -\tfrac{N}{2}\log(2\pi) - \tfrac{n}{2}\log\theta - \tfrac{m}{2}\log(\lambda_0\theta) - \frac{T_X + T_Y/\lambda_0}{2\theta},$$

which is maximized at $\hat\theta_0 = (T_X + T_Y/\lambda_0)/N$, whence $\sup_{H_0} L = (2\pi)^{-N/2}\lambda_0^{-m/2}\hat\theta_0^{-N/2}e^{-N/2}$. Therefore

$$\Lambda = \lambda_0^{-m/2}\,\hat\theta_0^{-N/2}(T_X/n)^{n/2}(T_Y/m)^{m/2}.$$

Substituting $T_Y/\lambda_0 = (m/n)F\,T_X$ and $T_Y/m = \lambda_0 F\,T_X/n$ collapses every appearance of T_X and λ_0:

$$\Lambda = \left(\frac{N}{n}\right)^{N/2} \frac{F^{m/2}}{\big(1 + mF/n\big)^{N/2}}.$$

So Λ depends on the data only through F. The function $g(F) = F^{m/2}(1 + mF/n)^{-N/2}$ vanishes at $F = 0$ and as $F \to \infty$, and $g'(F) = 0$ forces $1/F + m/n = N/n$, i.e. $F = 1$; hence g is unimodal with peak at $F = 1$. Consequently

$$\{\Lambda \le c\} = \{F \le c_1\} \cup \{F \ge c_2\}, \qquad 0 < c_1 < 1 < c_2,\ g(c_1) = g(c_2).$$

(b) $T_X/\sigma_X^2 \sim \chi^2_n$ and, under H_0, $T_Y/(\lambda_0\sigma_X^2) = T_Y/\sigma_Y^2 \sim \chi^2_m$, independently. Hence

$$F = \frac{T_Y/(\lambda_0\sigma_X^2)/m}{T_X/\sigma_X^2/n} \sim F_{m,n},$$

and the rejection region is $\{F \le c_1\} \cup \{F \ge c_2\}$ with c_1, c_2 determined by

$$P(c_1 \le F_{m,n} \le c_2) = 1 - \alpha \quad \text{together with} \quad g(c_1) = g(c_2).$$

The customary (very slightly non-LRT) choice is the equal-tailed one, $c_1 = F_{m,n,1-\alpha/2}$ and $c_2 = F_{m,n,\alpha/2}$.

(c) Inverting the acceptance region (Theorem 9.2.2): $\mathbf{x}, \mathbf{y} \in A(\lambda)$ iff $c_1 \le \dfrac{nT_Y}{m\lambda T_X} \le c_2$, and solving for λ,

$$C(\mathbf{x}, \mathbf{y}) = \left\{\lambda\colon \frac{1}{c_2}\cdot\frac{T_Y/m}{T_X/n} \;\le\; \lambda \;\le\; \frac{1}{c_1}\cdot\frac{T_Y/m}{T_X/n}\right\},$$

a $1 - \alpha$ confidence interval for λ. With the equal-tailed constants this reads

$$\left[\frac{T_Y/m}{T_X/n}\cdot\frac{1}{F_{m,n,\alpha/2}},\ \frac{T_Y/m}{T_X/n}\cdot\frac{1}{F_{m,n,1-\alpha/2}}\right].$$

Problem 9.5. In Example 9.2.5 a lower confidence bound was put on p, the success probability from a sequence of Bernoulli trials. This exercise will derive an upper confidence bound. That is, observing $X_1, \ldots, X_n$, where $X_i \sim$ Bernoulli(p), we want an interval of the form $[0, U(x_1, \ldots, x_n))$, where $P_p\big(p \in [0, U(X_1, \ldots, X_n))\big) \ge 1 - \alpha$.

(a) Show that inversion of the acceptance region of the test

$$H_0\colon p = p_0 \qquad \text{versus} \qquad H_1\colon p < p_0$$

will give a confidence interval of the desired confidence level and form.

(b) Find equations, similar to those given in (9.2.8), that can be used to construct the confidence interval.

For reference, Example 9.2.5 inverts the tests of $H_0\colon p = p_0$ versus $H_1\colon p > p_0$, whose UMP acceptance regions are $A(p_0) = \{t\colon t \le k(p_0)\}$ with $T = \sum_i X_i \sim$ binomial(n, p) and $k(p_0)$ the integer

between 0 and n satisfying

$$\sum_{y=0}^{k(p_0)} \binom{n}{y} p_0^y (1-p_0)^{n-y} \geq 1-\alpha,$$

$$\sum_{y=0}^{k(p_0)-1} \binom{n}{y} p_0^y (1-p_0)^{n-y} < 1-\alpha,$$

(these two inequalities are (9.2.8)), yielding the lower bound $C(t) = (k^{-1}(t), 1]$ with $k^{-1}(t) = \sup\{p\colon \sum_{y=0}^{t-1} \binom{n}{y} p^y (1-p)^{n-y} \geq 1-\alpha\}$.

Solution. (a) The tests now reject for *small* $T = \sum_i X_i$, so the confidence set is bounded above rather than below. The binomial family has MLR in T (stated after Definition 8.3.16), so by the mirror image of the Karlin-Rubin Theorem 8.3.17 recorded in the remark after its proof, the UMP level α test of $H_0\colon p = p_0$ versus $H_1\colon p < p_0$ rejects for $T < k^*(p_0)$, with acceptance region

$$A(p_0) = \{t\colon t \geq k^*(p_0)\}.$$

Theorem 9.2.2 then makes $C(t) = \{p_0\colon t \in A(p_0)\} = \{p_0\colon k^*(p_0) \leq t\}$ a $1-\alpha$ confidence set (the inequality $\geq 1-\alpha$, rather than equality, because discreteness forces each test to have size at most α). Its form is the one required. For fixed k, $P_p(T \geq k) = \sum_{y=k}^{n} \binom{n}{y} p^y (1-p)^{n-y}$ is increasing in p (MLR implies stochastic ordering; Exercise 8.26 for the complementary sum), and $k^*(p)$ is the largest k with $P_p(T \geq k) \geq 1-\alpha$; hence k^* is a nondecreasing step function of p, with $k^*(0) = 0$. Therefore $\{p_0\colon k^*(p_0) \leq t\}$ is an interval containing 0, i.e. of the form $[0, U(t))$, open on the right for the mirror image of the reason given after (9.2.9): at $p_0 = U(t)$ one has $P_{p_0}(T \geq t+1) = 1-\alpha$, so $k^*(p_0) = t+1 > t$ and $p_0 \notin C(t)$.
(b) Replacing "$\leq k$" by "$\geq k$" throughout (9.2.8), $k^*(p_0)$ is the integer between 0 and n that simultaneously satisfies

$$\sum_{y=k^*(p_0)}^{n} \binom{n}{y} p_0^y (1-p_0)^{n-y} \geq 1-\alpha,$$

$$\sum_{y=k^*(p_0)+1}^{n} \binom{n}{y} p_0^y (1-p_0)^{n-y} < 1-\alpha,$$

and the upper endpoint is the analogue of (9.2.10),

$$U(t) = \inf\left\{p\colon \sum_{y=t+1}^{n} \binom{n}{y} p^y (1-p)^{n-y} \geq 1-\alpha\right\}, \qquad C(t) = [0, U(t)).$$

Equivalently, because the sum is continuous and increasing in p, $U(t)$ is for $0 \leq t < n$ the unique root of

$$\sum_{y=0}^{t} \binom{n}{y} U^y (1-U)^{n-y} = \alpha,$$

and $U(n) = 1$ (the sum over $y > n$ is empty, so $C(n) = [0, 1]$).

Problem 9.6. (a) Derive a confidence interval for a binomial p by inverting the LRT of $H_0\colon p = p_0$ versus $H_1\colon p \neq p_0$.
(b) Show that the interval is a highest density region from $p^y(1-p)^{n-y}$ and is not equal to the interval in (10.4.4).

For reference, (10.4.4) is the score ("Wilson") interval for a binomial proportion,

$$\left\{p\colon \left|\frac{\hat p - p}{\sqrt{p(1-p)/n}}\right| \le z_{\alpha/2}\right\}, \qquad \hat p = y/n.$$

Solution. (a) With $Y = \sum_{i=1}^n X_i \sim \text{binomial}(n,p)$ and $\hat p = Y/n$ the MLE, the LRT statistic is

$$\Lambda(y) = \frac{p_0^y(1-p_0)^{n-y}}{\hat p^{\,y}(1-\hat p)^{n-y}},$$

with acceptance region $A(p_0) = \{y\colon \Lambda(y) \ge c\}$. Inverting it (Theorem 9.2.2) gives exactly the shape (9.2.7),

$$C(y) = \left\{p\colon p^y(1-p)^{n-y} \;\ge\; c\,\hat p^{\,y}(1-\hat p)^{n-y}\right\},$$

the constant c being chosen for $1-\alpha$ coverage. Since the exact null distribution of Λ is discrete and depends on p_0, the usable calibration is the χ^2 one of Theorem 10.3.1: $-2\log\Lambda \le \chi^2_{1,\alpha}$, i.e. $c = e^{-\chi^2_{1,\alpha}/2}$, which is the set of Example 10.4.3,

$$C(y) = \left\{p\colon -2\log\left(\frac{p^y(1-p)^{n-y}}{\hat p^{\,y}(1-\hat p)^{n-y}}\right) \le \chi^2_{1,\alpha}\right\}.$$

(b) Put $h(p) = p^y(1-p)^{n-y}$ and $k' = c\,\hat p^{\,y}(1-\hat p)^{n-y}$, so $C(y) = \{p\colon h(p) \ge k'\}$: the set collects precisely the parameter values at which h is largest, which is the definition of a highest density region for h. It is an interval, because for $0 < y < n$

$$h'(p) = p^{y-1}(1-p)^{n-y-1}\big(y - np\big)$$

is positive for $p < \hat p$ and negative for $p > \hat p$, with $h(0) = h(1) = 0$; hence $\{h \ge k'\}$ is $[L(y), U(y)]$ where the endpoints are the two roots of

$$L^y(1-L)^{n-y} = U^y(1-U)^{n-y} = k'.$$

(For $y = 0$ the interval is $[0, U]$ and for $y = n$ it is $[L, 1]$.) Normalizing, $h(p)/B(y+1, n-y+1)$ is the beta$(y+1, n-y+1)$ density, so $C(y)$ is literally an HPD region of that density.
The interval of (10.4.4) is a different set. Its endpoints are the roots of the quadratic $(\hat p - p)^2 = z^2_{\alpha/2}\,p(1-p)/n$, so they satisfy

$$\frac{(\hat p - L)^2}{L(1-L)} = \frac{(\hat p - U)^2}{U(1-U)} = \frac{z^2_{\alpha/2}}{n},$$

an equal-standardized-distance condition, not the equal-likelihood condition above; the two coincide only by accident. For $n = 10$, $y = 3$, $\alpha = .05$ (so $\chi^2_{1,.05} = z^2_{.025} = 3.8415$),

$$\begin{aligned} \text{LRT interval} &= [.0846,\ .6065],\\ \text{interval (10.4.4)} &= [.1078,\ .6032]. \end{aligned}$$

Problem 9.7. (a) Find the $1-\alpha$ confidence set for a that is obtained by inverting the LRT of $H_0\colon a = a_0$ versus $H_1\colon a \ne a_0$ based on a sample $X_1, \dots, X_n$ from a $\text{n}(\theta, a\theta)$ family, where θ is unknown.
(b) A similar question can be asked about the related family, the $\text{n}(\theta, a\theta^2)$ family. If $X_1, \dots, X_n$ are iid $\text{n}(\theta, a\theta^2)$, where θ is unknown, find the $1-\alpha$ confidence set based on inverting the LRT of $H_0\colon a = a_0$ versus $H_1\colon a \ne a_0$.

Solution. Both models share the unrestricted maximum

$$\sup_{\theta,a} L = (2\pi\hat\sigma^2)^{-n/2}e^{-n/2}, \qquad \hat\sigma^2 = \tfrac{1}{n}\textstyle\sum_i (x_i-\bar x)^2,$$

since for fixed $\theta > 0$ the MLE of a is $\hat a(\theta) = \sum_i (x_i-\theta)^2/(n\theta^r)$ ($r = 1$ in (a), $r = 2$ in (b)), which makes $a\theta^r = \frac{1}{n}\sum_i (x_i-\theta)^2$ and leaves the profile likelihood $\big(\frac{2\pi}{n}\sum_i (x_i-\theta)^2\big)^{-n/2}e^{-n/2}$ in both cases, maximized at $\hat\theta = \bar x$ (take $\bar x > 0$). Write $W = \sum_i x_i^2$.
(a) The confidence set is

$$C(\mathbf{x}) = \left\{ a\colon \left(\frac{\hat\sigma^2}{a\,\hat\theta_0(a)}\right)^{n/2} \exp\!\left(\frac{n\,(\bar x - \hat\theta_0(a))}{a}\right) \ge c \right\},$$

$$\hat\theta_0(a) = \tfrac{1}{2}\left(\sqrt{a^2 + 4W/n} - a\right).$$

Indeed, with a fixed,

$$\log L = -\tfrac{n}{2}\log(2\pi a\theta) - \frac{W - 2n\bar x\theta + n\theta^2}{2a\theta},$$

and setting $\partial/\partial\theta = 0$ and multiplying by $2a\theta^2$ gives

$$n\theta^2 + na\theta - W = 0,$$

whose positive root is $\hat\theta_0(a)$ above (the log-likelihood is $-\infty$ at $\theta \downarrow 0$ and $\theta \to \infty$, so this stationary point is the maximum). The same relation $W = n\hat\theta_0(\hat\theta_0 + a)$ simplifies the exponent:

$$\sum_i (x_i - \hat\theta_0)^2 = W - 2n\bar x\hat\theta_0 + n\hat\theta_0^2 = n\hat\theta_0\big(2\hat\theta_0 + a - 2\bar x\big),$$

so that

$$\sup_{H_0} L = (2\pi a\hat\theta_0)^{-n/2}\exp\!\left(-\frac{n(2\hat\theta_0 + a - 2\bar x)}{2a}\right),$$

and dividing by $\sup L$ gives $\Lambda(a)$ as displayed, since

$$\frac{n}{2} - \frac{n(2\hat\theta_0 + a - 2\bar x)}{2a} = \frac{n(\bar x - \hat\theta_0)}{a}.$$

Inverting the acceptance regions $\{\Lambda(a_0) \ge c\}$ (Theorem 9.2.2) yields $C(\mathbf{x})$. One caveat about c: writing $\Lambda(a;\mathbf{x}) = \Lambda(1;\mathbf{x}/a)$ and noting $X_i/a \sim \mathrm{n}(\theta/a, \theta/a)$, the null distribution of $\Lambda(a_0)$ still depends on the nuisance parameter through θ/a_0. There is no exact pivot here; c must be taken as the smallest cutoff valid for every θ, or, as usual, from $-2\log\Lambda \le \chi^2_{1,\alpha}$ (Theorem 10.3.1) for an asymptotic $1-\alpha$ set.
(b) The confidence set is

$$C(\mathbf{x}) = \left\{ a\colon \left(\frac{\hat\sigma^2}{a\,\hat\theta_0(a)^2}\right)^{n/2} \exp\!\left(\frac{n\,(\bar x - \hat\theta_0(a))}{2a\,\hat\theta_0(a)}\right) \ge c(a) \right\},$$

$$\hat\theta_0(a) = \frac{\sqrt{\bar x^2 + 4aW/n} - \bar x}{2a}.$$

Here, with a fixed,

$$\log L = -\tfrac{n}{2}\log(2\pi a) - n\log\theta - \frac{1}{2a}\left(\frac{W}{\theta^2} - \frac{2n\bar x}{\theta} + n\right),$$

and $\partial/\partial\theta = 0$, multiplied by $a\theta^3$, gives

$$na\theta^2 + n\bar x\theta - W = 0,$$

with positive root $\hat\theta_0(a)$. Using $W = na\hat\theta_0^2 + n\bar x\hat\theta_0$,

$$\sum_i (x_i - \hat\theta_0)^2 = n\hat\theta_0^2(a+1) - n\bar x\hat\theta_0,$$

so

$$\sup_{H_0} L = (2\pi a\hat\theta_0^2)^{-n/2}\exp\left(-\frac{n\big((a+1)\hat\theta_0 - \bar x\big)}{2a\hat\theta_0}\right),$$

and the ratio to $\sup L$ is the displayed $\Lambda(a)$, because

$$\frac{n}{2} - \frac{n\big((a+1)\hat\theta_0 - \bar x\big)}{2a\hat\theta_0} = \frac{n(\bar x - \hat\theta_0)}{2a\hat\theta_0}.$$

Unlike in (a), the cutoff here is exact. Replacing $\mathbf{x}$ by $b\mathbf{x}$ ($b > 0$) sends $\bar x \mapsto b\bar x$, $W \mapsto b^2W$, $\hat\theta_0(a) \mapsto b\hat\theta_0(a)$ and $\hat\sigma^2 \mapsto b^2\hat\sigma^2$, leaving $\Lambda(a)$ unchanged; and $X_i/\theta \sim \mathrm{n}(1, a)$. So the null distribution of $\Lambda(a_0)$ is that of $\Lambda(a_0)$ computed from an iid $\mathrm{n}(1, a_0)$ sample, free of θ, and $c(a_0)$ is determined by $P(\Lambda(a_0) \ge c(a_0)) = 1 - \alpha$ under that distribution, giving exact $1 - \alpha$ coverage.

9.2 Exercises 9.8–9.14

Problem 9.8. Given a sample $X_1, \dots, X_n$ from a pdf of the form $\frac{1}{\sigma} f\big(\frac{x-\theta}{\sigma}\big)$, list at least five different pivotal quantities.

Solution. Every function of $Z_1, \dots, Z_n$ alone is a pivot, where $Z_i = (X_i - \theta)/\sigma$. Indeed the joint pdf of $\mathbf{X}$ is $\prod_i \sigma^{-1} f((x_i - \theta)/\sigma)$, and the change of variables $z_i = (x_i - \theta)/\sigma$ (Jacobian σ^n) gives

$$f_{\mathbf{Z}}(z_1, \dots, z_n) = \prod_{i=1}^{n} f(z_i),$$

which involves neither θ nor σ. Writing $\bar Z$, S_Z, $Z_{(i)}$ for the mean, standard deviation and order statistics of the Z_i, and using $X_{(i)} = \theta + \sigma Z_{(i)}$, $\bar X = \theta + \sigma\bar Z$, $S = \sigma S_Z$:

(i) $\dfrac{\bar X - \theta}{\sigma} = \bar Z.$

(ii) $\dfrac{S}{\sigma} = S_Z.$

(iii) $\dfrac{\bar X - \theta}{S} = \dfrac{\bar Z}{S_Z}.$

(iv) $\dfrac{X_{(1)} - \theta}{\sigma} = Z_{(1)}$ (likewise $X_{(n)}$, or the sample median M).

(v) $\dfrac{X_{(n)} - X_{(1)}}{\sigma} = Z_{(n)} - Z_{(1)}$ (the standardized range).

(vi) $\dfrac{X_{(n)} - X_{(1)}}{S} = \dfrac{Z_{(n)} - Z_{(1)}}{S_Z}.$

(vii) $\dfrac{M - \theta}{S} = \dfrac{M_Z}{S_Z}$, with M_Z the median of the Z_i.

Each right-hand side is a fixed function of $\mathbf{Z}$ alone, so its distribution is determined by f and is free of (θ, σ): each is a pivot by Definition 9.2.6.

Problem 9.9. Show that each of the three quantities listed in Example 9.2.7 is a pivot. That is, let $X_1, \dots, X_n$ be a random sample from the indicated pdf, and let $\bar X$ and S be the sample mean and standard deviation; verify that the entries of Table 9.2.1 have distributions free of all parameters.

Form of pdf	Type of pdf	Pivotal quantity
$f(x-\mu)$	Location	$\bar{X}-\mu$
$\frac{1}{\sigma}f\left(\frac{x}{\sigma}\right)$	Scale	$\bar{X}/\sigma$
$\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$	Location–scale	$\dfrac{\bar{X}-\mu}{S}$

Solution. In each case the standardized sample $\mathbf{Z}$ has a pdf free of parameters, and the tabled quantity is a function of $\mathbf{Z}$ alone.
(i) *Location.* $Z_i = X_i - \mu$ has Jacobian 1, so

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} f(z_i),$$

free of μ, and $\bar{X} - \mu = \frac{1}{n}\sum_i (X_i - \mu) = \bar{Z}$.
(ii) *Scale.* $Z_i = X_i/\sigma$ has Jacobian σ^n, so

$$f_{\mathbf{Z}}(\mathbf{z}) = \sigma^n \prod_{i=1}^{n} \frac{1}{\sigma} f(z_i) = \prod_{i=1}^{n} f(z_i),$$

free of σ, and $\bar{X}/\sigma = \bar{Z}$.
(iii) *Location–scale.* $Z_i = (X_i - \mu)/\sigma$ has Jacobian σ^n and the same pdf $\prod_i f(z_i)$ as in (ii). Since $X_i = \mu + \sigma Z_i$,

$$\bar{X} = \mu + \sigma\bar{Z}, \qquad S^2 = \frac{\sigma^2}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2 = \sigma^2 S_Z^2,$$

so (with $\sigma > 0$) $\dfrac{\bar{X}-\mu}{S} = \dfrac{\sigma\bar{Z}}{\sigma S_Z} = \dfrac{\bar{Z}}{S_Z}$.
Each right-hand side is a fixed function of $\mathbf{Z}$, whose law involves only f, so each is a pivot by Definition 9.2.6.

Problem 9.10. (a) Suppose that T is a real-valued statistic. Suppose that $Q(t,\theta)$ is a monotone function of t for each value of $\theta \in \Theta$. Show that if the pdf of T, $f(t|\theta)$, can be expressed in the form (9.2.11),

$$f(t \mid \theta) = g\big(Q(t,\theta)\big)\left|\frac{\partial}{\partial t}Q(t,\theta)\right|,$$

for some function g, then $Q(T,\theta)$ is a pivot.
(b) Show that (9.2.11) is satisfied by taking $g \equiv 1$ and $Q(t,\theta) = F_\theta(t)$, the cdf of T. (This is the Probability Integral Transform.)

Solution. (a) The pdf of $Q(T,\theta)$ is g, for every θ. Fix θ. Since $q = Q(t,\theta)$ is monotone in t, it is invertible on the support of T; write $t = Q^{-1}(q,\theta)$. Theorem 2.1.5 applies (the monotone-transformation formula) and gives, for q in the range of $Q(\cdot,\theta)$,

$$\begin{aligned} f_{Q(T,\theta)}(q \mid \theta) &= f\big(Q^{-1}(q,\theta) \mid \theta\big)\left|\frac{\partial}{\partial q}Q^{-1}(q,\theta)\right| \\ &= g(q)\left|\tfrac{\partial}{\partial t}Q(t,\theta)\right| \cdot \left|\tfrac{\partial}{\partial t}Q(t,\theta)\right|^{-1} \\ &= g(q), \end{aligned}$$

where both derivatives are evaluated at $t = Q^{-1}(q,\theta)$: the second equality substitutes (9.2.11) and the inverse-function rule $\partial Q^{-1}/\partial q = (\partial Q/\partial t)^{-1}$, legitimate because monotonicity of Q in t forces $\partial Q/\partial t \neq 0$ there. The density g does not involve θ, so $Q(T,\theta)$ is a pivot by Definition 9.2.6.

(b) $F_\theta(t)$ is nondecreasing in t, hence monotone as required, and T continuous with pdf $f(t \mid \theta)$ gives

$$\frac{\partial}{\partial t}Q(t,\theta) = \frac{\partial}{\partial t}F_\theta(t) = f(t \mid \theta) \ \geq 0.$$

Therefore

$$g\big(Q(t,\theta)\big)\left|\frac{\partial}{\partial t}Q(t,\theta)\right| = 1 \cdot f(t \mid \theta) = f(t \mid \theta),$$

which is (9.2.11) with $g \equiv 1$. By part (a), $F_\theta(T)$ is a pivot with density $g \equiv 1$ on $(0,1)$, that is, $F_\theta(T) \sim \text{uniform}(0,1)$.

Problem 9.11. If T is a continuous random variable with cdf $F_T(t \mid \theta)$ and $\alpha_1 + \alpha_2 = \alpha$, show that an α level acceptance region of the hypothesis $H_0 : \theta = \theta_0$ is $\{t : \alpha_1 \leq F_T(t \mid \theta_0) \leq 1 - \alpha_2\}$, with associated confidence $1 - \alpha$ set $\{\theta : \alpha_1 \leq F_T(t \mid \theta) \leq 1 - \alpha_2\}$.

Solution. By Exercise 9.10(b), $F_T(T \mid \theta) \sim \text{uniform}(0,1)$ for every θ (the Probability Integral Transform; T is continuous, as required). Hence for the test with acceptance region

$$A(\theta_0) = \{t : \alpha_1 \leq F_T(t \mid \theta_0) \leq 1 - \alpha_2\},$$

the Type I Error probability is

$$\begin{aligned} P_{\theta_0}\big(T \notin A(\theta_0)\big) &= 1 - P_{\theta_0}\big(\alpha_1 \leq F_T(T \mid \theta_0) \leq 1 - \alpha_2\big) \\ &= 1 - \big[(1 - \alpha_2) - \alpha_1\big] = \alpha_1 + \alpha_2 = \alpha, \end{aligned}$$

so the test is a level α test of $H_0 : \theta = \theta_0$ (indeed of exact size α).
Inverting this family of acceptance regions by Theorem 9.2.2 gives

$$C(t) = \{\theta : t \in A(\theta)\} = \{\theta : \alpha_1 \leq F_T(t \mid \theta) \leq 1 - \alpha_2\},$$

and for each θ,

$$P_\theta\big(\theta \in C(T)\big) = P_\theta\big(T \in A(\theta)\big) = 1 - \alpha,$$

by the same uniform calculation with θ_0 replaced by θ. So $C(T)$ is a $1 - \alpha$ confidence set.

Problem 9.12. Find a pivotal quantity based on a random sample of size n from a $\text{n}(\theta, \theta)$ population, where $\theta > 0$. Use the pivotal quantity to set up a $1 - \alpha$ confidence interval for θ.

Solution. Take

$$Q(\mathbf{X}, \theta) = \frac{\sqrt{n}\,(\bar{X} - \theta)}{\sqrt{\theta}} \sim \text{n}(0,1),$$

a pivot by Definition 9.2.6, since $\bar{X} \sim \text{n}(\theta, \theta/n)$ makes the standardization exactly standard normal for every $\theta > 0$.
With $z = z_{\alpha/2}$, $P_\theta(-z \leq Q \leq z) = 1 - \alpha$, and inverting as in (9.2.13),

$$C(\mathbf{x}) = \Big\{\theta > 0 : \ n(\bar{x} - \theta)^2 \leq z^2\theta\Big\} = \Big\{\theta > 0 : \ n\theta^2 - (2n\bar{x} + z^2)\theta + n\bar{x}^2 \leq 0\Big\}.$$

The quadratic in θ opens upward with discriminant

$$(2n\bar{x} + z^2)^2 - 4n^2\bar{x}^2 = z^2\big(4n\bar{x} + z^2\big),$$

so provided $\bar{x} \geq -z^2/(4n)$ it has the two real roots

$$\theta_\pm = \bar{x} + \frac{z^2}{2n} \pm \frac{z}{2n}\sqrt{4n\bar{x} + z^2},$$

and the confidence set is the interval between them:

$$\left[\ \bar{x} + \frac{z^2}{2n} - \frac{z\sqrt{4n\bar{x} + z^2}}{2n},\ \bar{x} + \frac{z^2}{2n} + \frac{z\sqrt{4n\bar{x} + z^2}}{2n}\ \right],$$

which has coverage probability exactly $1-\alpha$. Both roots are nonnegative here (their product is $\bar{x}^2 \geq 0$ and their sum $2\bar{x} + z^2/n \geq z^2/(2n) > 0$), so the interval does lie in $\theta > 0$ even when $\bar{x} < 0$; only for $\bar{x} < -z^2/(4n)$ is the quadratic everywhere positive and $C(\mathbf{x}) = \emptyset$.
Method (2): $\dfrac{(n-1)S^2}{\theta} = \dfrac{\sum_i (X_i - \bar{X})^2}{\theta} \sim \chi^2_{n-1}$ by Theorem 5.3.1 is also a pivot, giving the interval

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}},\ \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right].$$

Problem 9.13. Let X be a single observation from the beta$(\theta, 1)$ pdf.
(a) Let $Y = -(\log X)^{-1}$. Evaluate the confidence coefficient of the set $[y/2, y]$.
(b) Find a pivotal quantity and use it to set up a confidence interval having the same confidence coefficient as the interval in part (a).
(c) Compare the two confidence intervals.

Solution. (a) The confidence coefficient is $e^{-1/2} - e^{-1} \approx .2387$, free of θ.
Write $W = -\log X$. Since $f(x \mid \theta) = \theta x^{\theta-1}$ on $(0,1)$ gives $F_X(x \mid \theta) = x^\theta$,

$$P_\theta(W > w) = P_\theta\big(X < e^{-w}\big) = e^{-\theta w}, \qquad w > 0,$$

so $W \sim$ exponential$(1/\theta)$ and $Y = 1/W$. Therefore

$$\begin{aligned} P_\theta\big(Y/2 \leq \theta \leq Y\big) &= P_\theta\big(\theta \leq 1/W \leq 2\theta\big) \\ &= P_\theta\Big(\tfrac{1}{2\theta} \leq W \leq \tfrac{1}{\theta}\Big) \\ &= e^{-\theta/(2\theta)} - e^{-\theta/\theta} = e^{-1/2} - e^{-1}. \end{aligned}$$

The value does not depend on θ, so the confidence coefficient is $e^{-1/2} - e^{-1} = .2387$.
(b) $Q(X, \theta) = -\theta \log X = \theta W = \theta/Y \sim$ exponential(1) is a pivot: its distribution is free of θ by the display above (Definition 9.2.6). For $0 \leq a < b$,

$$P_\theta(a \leq Q \leq b) = e^{-a} - e^{-b},$$

and inverting $\{a \leq \theta/y \leq b\}$ gives the interval $[ay, by]$. Requiring the confidence coefficient of part (a) means

$$e^{-a} - e^{-b} = e^{-1/2} - e^{-1} = .2387,$$

a one-parameter family; $(a, b) = (1/2, 1)$ returns exactly the interval $[y/2, y]$ of part (a). The shortest member takes $a = 0$: along the constraint curve $\frac{db}{da} = e^{b-a} > 1$, so the length $(b-a)y$ is strictly increasing in a. With $a = 0$,

$$1 - e^{-b} = .2387 \implies b = -\log(.7613) = .2727,$$

giving

$$C(x) = \big(0,\ .2727\, y\big], \qquad y = -(\log x)^{-1}.$$

(c) Both sets have confidence coefficient .2387. Part (a) has length $y - y/2 = .5000\, y$; part (b) has length $.2727\, y$, about 55% as long. So the interval of (b) dominates: the same confidence at a little over half the length, at the cost of being one-sided (it places no positive lower bound on θ).

Problem 9.14. Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$, where both parameters are unknown. Simultaneous inference on both μ and σ can be made using the Bonferroni Inequality in a number of ways.
(a) Using the Bonferroni Inequality, combine the two confidence sets

$$\left\{\mu:\ \bar{x} - \frac{ks}{\sqrt{n}} \le \mu \le \bar{x} + \frac{ks}{\sqrt{n}}\right\} \quad \text{and} \quad \left\{\sigma^2:\ \frac{(n-1)s^2}{b} \le \sigma^2 \le \frac{(n-1)s^2}{a}\right\}$$

into one confidence set for (μ, σ). Show how to choose a, b, and k to make the simultaneous set a $1-\alpha$ confidence set.
(b) Using the Bonferroni Inequality, combine the two confidence sets

$$\left\{\mu:\ \bar{x} - \frac{k\sigma}{\sqrt{n}} \le \mu \le \bar{x} + \frac{k\sigma}{\sqrt{n}}\right\} \quad \text{and} \quad \left\{\sigma^2:\ \frac{(n-1)s^2}{b} \le \sigma^2 \le \frac{(n-1)s^2}{a}\right\}$$

into one confidence set for (μ, σ). Show how to choose a, b, and k to make the simultaneous set a $1-\alpha$ confidence set.
(c) Compare the confidence sets in parts (a) and (b).

Solution. Split $\alpha = \alpha_1 + \alpha_2$ and give each marginal set coverage $1 - \alpha_i$; Bonferroni's Inequality (1.2.9), $P(A \cap B) \ge P(A) + P(B) - 1$, then delivers $1 - \alpha$ for the intersection.
(a) Take

$$k = t_{n-1,\alpha_1/2}, \qquad a = \chi^2_{n-1,1-\alpha_2/2}, \qquad b = \chi^2_{n-1,\alpha_2/2},$$

and let $C(\mathbf{x}) \subset \mathbb{R} \times (0, \infty)$ be the set of pairs (μ, σ) satisfying both inequalities. By Theorem 5.3.1, $(\bar{X} - \mu)/(S/\sqrt{n}) \sim t_{n-1}$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, so with

$$A = \left\{|\bar{X} - \mu| \le \frac{kS}{\sqrt{n}}\right\}, \qquad B = \left\{a \le \frac{(n-1)S^2}{\sigma^2} \le b\right\},$$

we have $P_{\mu,\sigma}(A) = 1 - \alpha_1$ and $P_{\mu,\sigma}(B) = 1 - \alpha_2$. Hence

$$\begin{aligned} P_{\mu,\sigma}\big((\mu,\sigma) \in C(\mathbf{X})\big) &= P_{\mu,\sigma}(A \cap B) \\ &\ge P_{\mu,\sigma}(A) + P_{\mu,\sigma}(B) - 1 \\ &= (1 - \alpha_1) + (1 - \alpha_2) - 1 = 1 - \alpha. \end{aligned}$$

The symmetric choice $\alpha_1 = \alpha_2 = \alpha/2$ is the usual one.
(b) Identical, except that the μ-statement now uses the known-variance pivot $(\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim \mathrm{n}(0,1)$, so take

$$k = z_{\alpha_1/2}, \qquad a = \chi^2_{n-1,1-\alpha_2/2}, \qquad b = \chi^2_{n-1,\alpha_2/2},$$

again with $\alpha_1 + \alpha_2 = \alpha$. The first set is not by itself a confidence set for μ (it involves the unknown σ), but the event $A' = \{|\bar{X} - \mu| \le k\sigma/\sqrt{n}\}$ still has probability $1 - \alpha_1$, and the intersection $A' \cap B$ is an observable subset of the (μ, σ) plane; Bonferroni gives the same bound $1 - \alpha$.
(c) Both are $1 - \alpha$ sets, but their shapes differ. In (a) the μ-section has the constant half-width $t_{n-1,\alpha_1/2}\, s/\sqrt{n}$ for every σ, so $C(\mathbf{x})$ is a rectangle. In (b) the half-width is $z_{\alpha_1/2}\, \sigma/\sqrt{n}$, growing linearly in σ, so $C(\mathbf{x})$ is a trapezoidal wedge truncated to $\sigma \in [s\sqrt{(n-1)/b},\ s\sqrt{(n-1)/a}]$; it is narrower than (a) for small σ and wider for large σ.
For $n = 10$, $\alpha_1 = \alpha_2 = .05$ the σ-range is $[.688s, 1.826s]$ in both, the half-widths are $.715s$ in (a) and $.620\,\sigma \in [.426s, 1.132s]$ in (b), and the areas are

$$\begin{aligned} &\text{(a)}:\ 2(.715\, s)(1.826 - .688)s = 1.628\, s^2, \\ &\text{(b)}:\ \int_{.688s}^{1.826s} 2(.620\,\sigma)\, d\sigma = 1.772\, s^2, \end{aligned}$$

so here (a) is the smaller set.

9.3 Exercises 9.15–9.21

Problem 9.15. Solve for the roots of the quadratic equation that defines Fieller's confidence set for the ratio of normal means (see Miscellanea 9.5.3). Find conditions on the random variables for which
(a) the parabola opens upward (the confidence set is an interval).
(b) the parabola opens downward (the confidence set is the complement of an interval).
(c) the parabola has no real roots.
In each case, give an interpretation of the meaning of the confidence set. For example, what would you tell an experimenter if, for his data, the parabola had no real roots?
Miscellanea 9.5.3 (Fieller's Theorem), for reference. Given a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from a bivariate normal distribution with parameters $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, a confidence set on $\theta = \mu_Y/\mu_X$ is obtained from $Z_{\theta i} = Y_i - \theta X_i$: the variable $\bar{Z}_\theta = \bar{Y} - \theta\bar{X}$ is normal with mean 0 and variance $V_\theta = \frac{1}{n}(\sigma_Y^2 - 2\theta\rho\sigma_Y\sigma_X + \theta^2\sigma_X^2)$, estimated by

$$\hat{V}_\theta = \frac{1}{n(n-1)}\sum_{i=1}^{n}(Z_{\theta i} - \bar{Z}_\theta)^2 = \frac{1}{n-1}\left(S_Y^2 - 2\theta S_{YX} + \theta^2 S_X^2\right),$$

where $S_Y^2 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$, $S_X^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$ and $S_{YX} = \frac{1}{n}\sum(Y_i - \bar{Y})(X_i - \bar{X})$. Since $\bar{Z}_\theta/\sqrt{\hat{V}_\theta} \sim t_{n-1}$, the set $\{\theta : \bar{Z}_\theta^2/\hat{V}_\theta \le t_{n-1,\alpha/2}^2\}$ is a $1-\alpha$ confidence set; written in the original variables it is

$$\left\{\theta : \left(\bar{x}^2 - \tfrac{t^2}{n-1}S_x^2\right)\theta^2 - 2\left(\bar{x}\bar{y} - \tfrac{t^2}{n-1}S_{yx}\right)\theta + \left(\bar{y}^2 - \tfrac{t^2}{n-1}S_y^2\right) \le 0\right\}, \qquad t^2 = t_{n-1,\alpha/2}^2.$$

Solution. The roots are

$$\theta_\pm = \frac{b \pm \sqrt{b^2 - ac}}{a},$$

where, writing $k = t_{n-1,\alpha/2}^2/(n-1)$,

$$a = \bar{x}^2 - kS_x^2, \qquad b = \bar{x}\bar{y} - kS_{yx}, \qquad c = \bar{y}^2 - kS_y^2,$$

and the confidence set is $\{\theta : a\theta^2 - 2b\theta + c \le 0\}$.
The leading coefficient is the one-sample t-statistic for $\mu_X = 0$: with $s_x^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = \frac{n}{n-1}S_x^2$,

$$a > 0 \iff \frac{(n-1)\bar{x}^2}{S_x^2} > t_{n-1,\alpha/2}^2 \iff \left|\frac{\sqrt{n}\,\bar{x}}{s_x}\right| > t_{n-1,\alpha/2}.$$

Expanding, with $\hat{\Sigma} = \begin{pmatrix} S_x^2 & S_{yx} \\ S_{yx} & S_y^2 \end{pmatrix}$,

$$\begin{aligned} b^2 - ac &= k\left(\bar{x}^2 S_y^2 - 2\bar{x}\bar{y}S_{yx} + \bar{y}^2 S_x^2\right) - k^2\left(S_x^2 S_y^2 - S_{yx}^2\right) \\ &= k\left(S_x^2 S_y^2 - S_{yx}^2\right)\left[(\bar{x},\bar{y})\hat{\Sigma}^{-1}(\bar{x},\bar{y})^\top - k\right], \end{aligned}$$

so, writing $Q = (\bar{x},\bar{y})\hat{\Sigma}^{-1}(\bar{x},\bar{y})^\top$ and using that $\hat{\Sigma}$ is positive definite,

$$\begin{aligned} b^2 - ac \ge 0 &\iff Q \ge k \\ &\iff (n-1)\,Q \ge t_{n-1,\alpha/2}^2 : \end{aligned}$$

there are real roots exactly when $(\bar{x},\bar{y})$ is far from the origin in Mahalanobis distance.
(a) *Upward.* $a > 0$, i.e. $\sqrt{n}\,|\bar{x}|/s_x > t_{n-1,\alpha/2}$: the denominator mean μ_X is significantly different from 0 at level α. Real roots are then automatic, because

$$Q - \frac{\bar{x}^2}{S_x^2} = \frac{(\bar{y}S_x^2 - \bar{x}S_{yx})^2}{S_x^2(S_x^2 S_y^2 - S_{yx}^2)} \ge 0,$$

while $a > 0$ says exactly $\bar{x}^2/S_x^2 > k$, so $Q > k$. The confidence set is the bounded interval $[\theta_-, \theta_+]$, the usual interval estimate of the ratio.
(b) *Downward.* $a < 0$, i.e. $\sqrt{n}|\bar{x}|/s_x < t_{n-1,\alpha/2}$: μ_X is not distinguishable from 0. If in addition $b^2 - ac \geq 0$ the set is

$$(-\infty, \theta_+] \cup [\theta_-, \infty), \qquad \theta_+ < \theta_-,$$

the complement of a bounded interval. This is honest: since μ_X may be arbitrarily close to 0, the ratio μ_Y/μ_X may be arbitrarily large in either direction, and all the data can do is exclude a middle band of ratios.
(c) *No real roots.* $b^2 - ac < 0$ forces $a < 0$: by (a) $a > 0$ gives real roots, and $a = 0$ gives $b^2 - ac = b^2 \geq 0$. Then $a\theta^2 - 2b\theta + c < 0$ for every θ and the confidence set is all of $\mathbf{R}$. Tell the experimenter that at level α the data say nothing whatever about μ_Y/μ_X: the mean vector $(\bar{x}, \bar{y})$ is not significantly different from $(0, 0)$, so every ratio is consistent with the data and a larger sample is needed.

Problem 9.16. Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, \sigma^2)$, where σ^2 is known. For each of the following hypotheses, write out the acceptance region of a level α test and the $1 - \alpha$ confidence interval that results from inverting the test.
(a) $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$
(b) $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$
(c) $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$

Solution. (a) Reject for large $|\bar{x} - \theta_0|$ — the LRT of Example 8.2.2. Writing $\sigma_n = \sigma/\sqrt{n}$, so that $\bar{X} \sim \mathrm{n}(\theta, \sigma_n^2)$,

$$A(\theta_0) = \{\mathbf{x} : |\bar{x} - \theta_0| \leq z_{\alpha/2}\sigma_n\},$$
$$C(\mathbf{x}) = [\bar{x} - z_{\alpha/2}\sigma_n,\ \bar{x} + z_{\alpha/2}\sigma_n].$$

(b) Reject for small $\bar{x}$, the UMP level α test by Karlin–Rubin (Theorem 8.3.17), the hypothesis of a monotone likelihood ratio in $\bar{x}$ holding for the normal family; $\sup_{\theta \geq \theta_0} P_\theta(\bar{X} < \theta_0 - z_\alpha\sigma_n) = \alpha$ is attained at $\theta = \theta_0$, so the test has size α and

$$A(\theta_0) = \{\mathbf{x} : \bar{x} \geq \theta_0 - z_\alpha\sigma_n\},$$
$$C(\mathbf{x}) = \{\theta : \theta \leq \bar{x} + z_\alpha\sigma_n\} = (-\infty,\ \bar{x} + z_\alpha\sigma_n].$$

(c) Reject for large $\bar{x}$, again UMP by Theorem 8.3.17:

$$A(\theta_0) = \{\mathbf{x} : \bar{x} \leq \theta_0 + z_\alpha\sigma_n\},$$
$$C(\mathbf{x}) = \{\theta : \theta \geq \bar{x} - z_\alpha\sigma_n\} = [\bar{x} - z_\alpha\sigma_n,\ \infty).$$

In each case $\theta \in C(\mathbf{x}) \iff \mathbf{x} \in A(\theta)$ and the test has size α at θ, so by Theorem 9.2.2 the coverage is exactly $1 - \alpha$.

Problem 9.17. Find a $1 - \alpha$ confidence interval for θ, given $X_1, \ldots, X_n$ iid with pdf
(a) $f(x|\theta) = 1$, $\theta - \frac{1}{2} < x < \theta + \frac{1}{2}$.
(b) $f(x|\theta) = 2x/\theta^2$, $0 < x < \theta$, $\theta > 0$.

Solution. (a) With $M = \left(X_{(1)} + X_{(n)}\right)/2$ the midrange,

$$\left[M - \tfrac{1-\alpha^{1/n}}{2},\ M + \tfrac{1-\alpha^{1/n}}{2}\right].$$

Indeed θ is a location parameter, so $M - \theta$ is a pivot. Put $V_i = X_i - \theta + \frac{1}{2}$, iid uniform$(0, 1)$, and $W = \left(V_{(1)} + V_{(n)}\right)/2 = M - \theta + \frac{1}{2}$. The joint pdf of $\left(V_{(1)}, V_{(n)}\right)$ is $n(n-1)(v-u)^{n-2}$ on $0 < u < v < 1$ (Theorem 5.4.6); changing to $W = (u + v)/2$, $R = v - u$ (Jacobian 1) and integrating out R over

$0 < r < 2w$ gives, for $0 < w \le \frac{1}{2}$,

$$f_W(w) = \int_0^{2w} n(n-1)r^{n-2}\,dr = n(2w)^{n-1},$$

and f_W is symmetric about $\frac{1}{2}$. Hence $P(W \le w) = \frac{1}{2}(2w)^n$ for $w \le \frac{1}{2}$, and the equal-tailed cutoffs solve $\frac{1}{2}(2w)^n = \frac{\alpha}{2}$, i.e. $w = \frac{1}{2}\alpha^{1/n}$ and $1 - \frac{1}{2}\alpha^{1/n}$:

$$1 - \alpha = P\left(\tfrac{\alpha^{1/n}}{2} \le W \le 1 - \tfrac{\alpha^{1/n}}{2}\right),$$

which rearranges to the interval above.
(b) In terms of the largest observation,

$$\left[X_{(n)}(1-\alpha/2)^{-1/(2n)},\ X_{(n)}(\alpha/2)^{-1/(2n)}\right].$$

Here θ is a scale parameter and $F(x|\theta) = x^2/\theta^2$ on $(0,\theta)$, so $T = X_{(n)}/\theta$ is a pivot with

$$P(T \le t) = \left(t^2\right)^n = t^{2n}, \qquad 0 < t < 1.$$

Any $0 < a < b \le 1$ with $b^{2n} - a^{2n} = 1 - \alpha$ gives the $1-\alpha$ interval $\left[X_{(n)}/b,\ X_{(n)}/a\right]$; the equal-tailed choice $a^{2n} = \alpha/2$, $b^{2n} = 1 - \alpha/2$ yields the display.

Problem 9.18. In this exercise we will investigate some more properties of binomial confidence sets and the Sterne (1954) construction in particular. As in Example 9.2.11, we will again consider the binomial$(3, p)$ distribution.
(a) Draw, as a function of p, a graph of the four probability functions $P_p(X = x)$, $x = 0, \ldots, 3$. Identify the maxima of $P_p(X = 1)$ and $P_p(X = 2)$.
(b) Show that for small ϵ, $P_p(X = 0) > P_p(X = 2)$ for $p = \frac{1}{3} + \epsilon$.
(c) Show that the *most probable* construction is to blame for the difficulties with the Sterne sets by showing that the following acceptance regions can be inverted to obtain a $1 - \alpha = .442$ confidence *interval*:

p	Acceptance region
$[.000, .238]$	$\{0\}$
$(.238, .305)$	$\{0, 1\}$
$[.305, .362]$	$\{1\}$
$(.362, .634)$	$\{1, 2\}$
$[.634, .695]$	$\{2\}$
$(.695, .762)$	$\{2, 3\}$
$[.762, 1.00]$	$\{3\}$

(This is essentially Crow's 1956 modification of Sterne's construction; see Miscellanea 9.5.2.)
Background (Example 9.2.11). Sterne's construction fills $A(p)$ with the most probable x values until its probability is at least $1 - \alpha$. For $X \sim$ binomial$(3, p)$ and $1 - \alpha = .442$ it produces $A(p) = \{0\}$, $\{0, 1\}$, $\{1\}$, $\{0, 1\}$, $\{1, 2\}$, $\{2, 3\}$, $\{2\}$, $\{2, 3\}$, $\{3\}$ on the successive intervals $[.000, .238]$, $(.238, .305)$, $[.305, .362]$, $(.362, .366)$, $[.366, .634]$, $(.634, .638)$, $[.638, .695]$, $(.695, .762)$, $[.762, 1.00]$ (Table 9.2.2), whose inversion gives the non-interval sets $C(0) = [.000, .305) \cup (.362, .366)$ and $C(3) = (.634, .638) \cup (.695, 1.00]$.

Solution. (a) The four curves are

$$P_p(X = 0) = (1-p)^3, \qquad P_p(X = 1) = 3p(1-p)^2,$$
$$P_p(X = 2) = 3p^2(1-p), \quad P_p(X = 3) = p^3.$$

$P_p(X = 0)$ decreases from 1 to 0 and $P_p(X = 3)$ is its mirror image, increasing from 0 to 1; the two

middle curves are unimodal humps, mirror images of each other in $p \mapsto 1-p$. Differentiating,

$$\frac{d}{dp}\,3p(1-p)^2 = 3(1-p)(1-3p) = 0 \text{ at } p = \tfrac{1}{3}, \qquad P_{1/3}(X=1) = \tfrac{4}{9},$$

and by symmetry $P_p(X=2)$ is maximized at $p=\frac{2}{3}$ with $P_{2/3}(X=2)=\frac{4}{9}$. The crossings that matter for the ordering of the pmf are $P_p(X=0)=P_p(X=1)$ at $p=\frac{1}{4}$, $P_p(X=1)=P_p(X=2)$ at $p=\frac{1}{2}$, and the one computed in part (b).
(b) At $p=\frac{1}{3}$ exactly,

$$P_{1/3}(X=0) = \left(\tfrac{2}{3}\right)^3 = \tfrac{8}{27} > \tfrac{6}{27} = 3\left(\tfrac{1}{3}\right)^2\left(\tfrac{2}{3}\right) = P_{1/3}(X=2),$$

so by continuity the strict inequality persists on a neighbourhood of $\frac{1}{3}$. Exactly: for $0<p<1$,

$$(1-p)^3 > 3p^2(1-p) \iff (1-p)^2 > 3p^2 \iff 1-p > \sqrt{3}\,p \iff p < \frac{1}{1+\sqrt{3}} = .36603,$$

so $P_p(X=0) > P_p(X=2)$ for $p = \frac{1}{3}+\epsilon$ whenever $\epsilon < .03270$.
This is the whole trouble with the most probable construction. Just above $p=\frac{1}{3}$ the ordering of the pmf is $P(X=1)>P(X=0)>P(X=2)$, so after $x=1$ the value $x=0$ is admitted to $A(p)$ before $x=2$; the acceptance region jumps back to $\{0,1\}$ on $(.362,.366)$ after having been $\{1\}$, and $C(0)$ picks up the stray piece $(.362,.366)$.
(c) *Erratum: both occurrences of* .634 *in the table should be* .638 — the $\{2\}$-row has to be the reflection $[.638,.695]$ of the $\{1\}$-row $[.305,.362]$ under $p\mapsto 1-p$, and at $p=.634$ the singleton $\{2\}$ carries only $3(.634)^2(.366) = .44135 < .442$; the number $.634 = 1-(1+\sqrt{3})^{-1}$ is the part (b) crossing point, which belongs to Sterne's Table 9.2.2 and not to Crow's. Inverting the corrected family by $C(x)=\{p : x\in A(p)\}$, i.e. reading down the table column by column,

$$\begin{aligned}
C(0) &= [.000,.238]\cup(.238,.305) = [.000,.305),\\
C(1) &= (.238,.305)\cup[.305,.362]\cup(.362,.638) = (.238,.638),\\
C(2) &= (.362,.638)\cup[.638,.695]\cup(.695,.762) = (.362,.762),\\
C(3) &= (.695,.762)\cup[.762,1.00] = (.695,1.00].
\end{aligned}$$

Each is an interval, because the values of x enter and leave $A(p)$ once each as p increases: the acceptance regions are *consecutive* runs $\{x : x_1(p)\le x\le x_2(p)\}$ with $x_1(p) = 0,0,1,1,2,2,3$ and $x_2(p) = 0,1,1,2,2,3,3$ down the table, both nondecreasing in p, which is exactly the property Sterne's regions lack.
It remains to check the level, $P_p(X\in A(p))\ge .442$ for all p. The binding cases are the singletons, and on each the minimum is at an endpoint, since $(1-p)^3$ and p^3 are monotone and the two humps peak inside their intervals:

$$\begin{aligned}
p\in[.000,.238]: &\quad \min(1-p)^3 = (.762)^3 = .4425,\\
p\in[.305,.362]: &\quad \min 3p(1-p)^2 = 3(.305)(.695)^2 = .4420,\\
p\in[.638,.695]: &\quad \min 3p^2(1-p) = 3(.695)^2(.305) = .4420,\\
p\in[.762,1.00]: &\quad \min p^3 = (.762)^3 = .4425,
\end{aligned}$$

the two .4420 entries sitting exactly on the level because the table rounds the roots of $3p(1-p)^2=.442$, namely .30518 and .36230 and their reflections .63770 and .69482, to three places. On the doubleton intervals the coverage is far larger: $(1-p)^3+3p(1-p)^2\ge .7777$ on $(.238,.305)$, $3p(1-p)^2+3p^2(1-p) = 3p(1-p)\ge 3(.362)(.638) = .6929$ on $(.362,.638)$, and symmetrically on $(.695,.762)$. So every $A(p)$ has probability at least .442 and, by Theorem 9.2.2, each $C(x)$ above is a .442 confidence interval.

Problem 9.19. Prove part (b) of Theorem 9.2.12.
Theorem 9.2.12 (Pivoting a continuous cdf). Let T be a statistic with continuous cdf $F_T(t|\theta)$. Let $\alpha_1 + \alpha_2 = \alpha$ with $0 < \alpha < 1$ be fixed values. Suppose that for each $t \in \mathcal{T}$, the functions $\theta_L(t)$ and $\theta_U(t)$ can be defined as follows.
(i) If $F_T(t|\theta)$ is a decreasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by

$$F_T(t|\theta_U(t)) = \alpha_1, \qquad F_T(t|\theta_L(t)) = 1 - \alpha_2.$$

(ii) If $F_T(t|\theta)$ is an increasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by

$$F_T(t|\theta_U(t)) = 1 - \alpha_2, \qquad F_T(t|\theta_L(t)) = \alpha_1.$$

Then the random interval $[\theta_L(T), \theta_U(T)]$ is a $1 - \alpha$ confidence interval for θ. (Part (i) is proved in the text; part (b) here is part (ii), the increasing case.)

Solution. By Exercise 9.11 — whose hypothesis, that T be continuous, is part of Theorem 9.2.12 — the set $\{t : \alpha_1 \le F_T(t|\theta_0) \le 1 - \alpha_2\}$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$, with associated confidence set

$$C(t) = \{\theta : \alpha_1 \le F_T(t|\theta) \le 1 - \alpha_2\}.$$

Fix t. Since $\alpha_1 < 1 - \alpha_2$ and $F_T(t|\cdot)$ is increasing, the values $\theta_L(t)$ and $\theta_U(t)$ solving the two equations in (ii) are unique and satisfy $\theta_L(t) \le \theta_U(t)$. Monotonicity in θ gives

$$\begin{aligned} F_T(t|\theta) < \alpha_1 &\iff \theta < \theta_L(t), \\ F_T(t|\theta) > 1 - \alpha_2 &\iff \theta > \theta_U(t), \end{aligned}$$

and hence

$$C(t) = \{\theta : \alpha_1 \le F_T(t|\theta) \le 1 - \alpha_2\} = \{\theta : \theta_L(t) \le \theta \le \theta_U(t)\}.$$

Finally F_T is continuous, so $F_T(T|\theta) \sim \text{uniform}(0, 1)$ by the Probability Integral Transform (Theorem 2.1.10), and therefore

$$\begin{aligned} P_\theta\left(\theta_L(T) \le \theta \le \theta_U(T)\right) &= P_\theta\left(\alpha_1 \le F_T(T|\theta) \le 1 - \alpha_2\right) \\ &= (1 - \alpha_2) - \alpha_1 = 1 - \alpha. \end{aligned}$$

Problem 9.20. Some of the details of the proof of Theorem 9.2.14 need to be filled in, and the second part of the theorem needs to be proved.
(a) Show that if $F_T(T|\theta)$ is stochastically greater than or equal to a uniform random variable, then so is $\bar{F}_T(T|\theta)$. That is, if $P_\theta(F_T(T|\theta) \le x) \le x$ for every x, $0 \le x \le 1$, then $P_\theta(\bar{F}_T(T|\theta) \le x) \le x$ for every x, $0 \le x \le 1$.
(b) Show that for $\alpha_1 + \alpha_2 = \alpha$, the set $\{\theta : F_T(T|\theta) \le \alpha_1 \text{ and } \bar{F}_T(T|\theta) \le \alpha_2\}$ is a $1 - \alpha$ confidence set.
(c) If the cdf $F_T(t|\theta)$ is a decreasing function of θ for each t, show that the function $\bar{F}_T(t|\theta)$ defined by $\bar{F}_T(t|\theta) = P(T \ge t|\theta)$ is a nondecreasing function of θ for each t.
(d) Prove part (b) of Theorem 9.2.14.
Theorem 9.2.14 (Pivoting a discrete cdf). Let T be a discrete statistic with cdf $F_T(t|\theta) = P(T \le t|\theta)$. Let $\alpha_1 + \alpha_2 = \alpha$ with $0 < \alpha < 1$ be fixed values. Suppose that for each $t \in \mathcal{T}$, $\theta_L(t)$ and $\theta_U(t)$ can be defined as follows.
(i) If $F_T(t|\theta)$ is a decreasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by

$$P(T \le t|\theta_U(t)) = \alpha_1, \qquad P(T \ge t|\theta_L(t)) = \alpha_2.$$

(ii) If $F_T(t|\theta)$ is an increasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by

$$P(T \ge t|\theta_U(t)) = \alpha_1, \qquad P(T \le t|\theta_L(t)) = \alpha_2.$$

Then the random interval $[\theta_L(T), \theta_U(T)]$ is a $1-\alpha$ confidence interval for θ. (The text sketches part (i); the part (b) referred to in (d) above is part (ii).)

Solution. (a) Apply Exercise 2.10 to the statistic $S = -T$. Its cdf is

$$F_S(s|\theta) = P(-T \le s|\theta) = P(T \ge -s|\theta) = \bar{F}_T(-s|\theta),$$

so $F_S(S|\theta) = \bar{F}_T(T|\theta)$, and Exercise 2.10, whose hypothesis is that S be discrete, says exactly that $P_\theta(F_S(S|\theta) \le x) \le x$ for $0 \le x \le 1$.
(b) *Erratum: both inequalities must be reversed.* As printed the set is empty, since $F_T(t|\theta) + \bar{F}_T(t|\theta) = 1 + P(T = t|\theta) \ge 1 > \alpha_1 + \alpha_2$ for every t; the set intended here, and the one used in the proof of Theorem 9.2.14, is

$$C(T) = \{\theta : F_T(T|\theta) \ge \alpha_1 \text{ and } \bar{F}_T(T|\theta) \ge \alpha_2\}.$$

Write $F = F_T(T|\theta)$ and $\bar{F} = \bar{F}_T(T|\theta)$. Boole's inequality together with Exercise 2.10 (for F) and part (a) (for $\bar{F}$) gives

$$\begin{aligned} P_\theta(\theta \notin C(T)) &= P_\theta(F < \alpha_1 \text{ or } \bar{F} < \alpha_2) \\ &\le P_\theta(F < \alpha_1) + P_\theta(\bar{F} < \alpha_2) \\ &\le \alpha_1 + \alpha_2 = \alpha, \end{aligned}$$

so $P_\theta(\theta \in C(T)) \ge 1 - \alpha$ for every θ.
(c) $P(T < t|\theta) = \lim_{s \uparrow t} F_T(s|\theta)$, and each $F_T(s|\cdot)$ is decreasing, so the pointwise limit $P(T < t|\cdot)$ is nonincreasing; hence

$$\bar{F}_T(t|\theta) = 1 - P(T < t|\theta)$$

is nondecreasing in θ. (Concretely, if T has support $t_1 < t_2 < \cdots$ free of θ, then $\bar{F}_T(t_j|\theta) = 1 - F_T(t_{j-1}|\theta)$ for $j \ge 2$ and $\bar{F}_T(t_1|\theta) \equiv 1$.)
(d) Assume $F_T(t|\theta)$ is increasing in θ for each t; then by the argument of (c) applied with the inequalities reversed, $\bar{F}_T(t|\theta)$ is nonincreasing in θ for each t, strictly so at each support point t_j, $j \ge 2$, where $\bar{F}_T(t_j|\theta) = 1 - F_T(t_{j-1}|\theta)$; at t_1, $\bar{F}_T(t_1|\theta) \equiv 1 > \alpha_1$, so $\theta_U(t_1) = \sup \Theta$ and the second implication below is vacuous. Apply part (b) with α_1 and α_2 interchanged: the set

$$C(t) = \{\theta : F_T(t|\theta) \ge \alpha_2 \text{ and } \bar{F}_T(t|\theta) \ge \alpha_1\}$$

has coverage probability at least $1 - \alpha$. Now $\theta_L(t)$ and $\theta_U(t)$ of part (ii) satisfy $F_T(t|\theta_L(t)) = \alpha_2$ and $\bar{F}_T(t|\theta_U(t)) = \alpha_1$, so the monotonicities give

$$\begin{aligned} \theta < \theta_L(t) &\Rightarrow F_T(t|\theta) < F_T(t|\theta_L(t)) = \alpha_2, \\ \theta > \theta_U(t) &\Rightarrow \bar{F}_T(t|\theta) < \bar{F}_T(t|\theta_U(t)) = \alpha_1, \end{aligned}$$

and in either case $\theta \notin C(t)$. Hence $C(t) \subseteq [\theta_L(t), \theta_U(t)]$ and

$$P_\theta\big(\theta_L(T) \le \theta \le \theta_U(T)\big) \ge P_\theta\big(\theta \in C(T)\big) \ge 1 - \alpha,$$

so $[\theta_L(T), \theta_U(T)]$ is a $1 - \alpha$ confidence interval for θ.

Problem 9.21. In Example 9.2.15 it was shown that a confidence interval for a Poisson parameter can be expressed in terms of chi squared cutoff points. Use a similar technique to show that if $X \sim$ binomial(n, p), then a $1 - \alpha$ confidence interval for p is

$$\frac{1}{1 + \frac{n-x+1}{x} F_{2(n-x+1),2x,\alpha/2}} \;\le\; p \;\le\; \frac{\frac{x+1}{n-x} F_{2(x+1),2(n-x),\alpha/2}}{1 + \frac{x+1}{n-x} F_{2(x+1),2(n-x),\alpha/2}},$$

where $F_{\nu_1,\nu_2,\alpha}$ is the upper α cutoff from an F distribution with ν_1 and ν_2 degrees of freedom, and we make the endpoint adjustment that the lower endpoint is 0 if $x = 0$ and the upper endpoint is 1

if $x = n$. These are the Clopper and Pearson (1934) intervals.
(Hint: Recall the following identity from Exercise 2.40, which can be interpreted in the following way. If $X \sim$ binomial(n, θ), then $P_\theta(X \geq x) = P(Y \leq \theta)$, where $Y \sim$ beta$(x, n - x + 1)$. Use the properties of the F and beta distributions from Chapter 5.)

Solution. The endpoints are the solutions of the two Theorem 9.2.14(i) equations

$$P(X \geq x \mid p_L) = \frac{\alpha}{2}, \qquad P(X \leq x \mid p_U) = \frac{\alpha}{2},$$

which apply because the binomial family is stochastically increasing in p, so that $F_X(x|p)$ is a decreasing function of p for each x (see Exercises 3.41–3.43). Both equations are solved in closed form by the Exercise 2.40 identity.
Lower endpoint. By the hint, $P(X \geq x|p) = P(Y \leq p)$ with $Y \sim$ beta$(x, n - x + 1)$, so p_L is the $\alpha/2$ quantile of that beta. By Theorem 5.3.8(c), if $W \sim F_{2x,\,2(n-x+1)}$ then

$$Y = \frac{\frac{x}{n-x+1}W}{1 + \frac{x}{n-x+1}W} \sim \text{beta}(x, n - x + 1),$$

an increasing function of W, so quantiles correspond: p_L is obtained from the lower $\alpha/2$ point of $F_{2x,2(n-x+1)}$. By Theorem 5.3.8(a) that point is $1/F_{2(n-x+1),2x,\alpha/2}$, whence

$$p_L = \frac{\frac{x}{n-x+1} \cdot \frac{1}{F_{2(n-x+1),2x,\alpha/2}}}{1 + \frac{x}{n-x+1} \cdot \frac{1}{F_{2(n-x+1),2x,\alpha/2}}} = \frac{1}{1 + \frac{n-x+1}{x} F_{2(n-x+1),2x,\alpha/2}}.$$

Upper endpoint. Applying the identity with $x + 1$ in place of x,

$$P(X \leq x \mid p) = 1 - P(X \geq x + 1 \mid p) = 1 - P(Y' \leq p) = P(Y' \geq p),$$

where $Y' \sim$ beta$(x + 1, n - x)$; so p_U is the $1 - \alpha/2$ quantile of Y'. Theorem 5.3.8(c) with $W' \sim F_{2(x+1),\,2(n-x)}$ gives $Y' = \frac{\frac{x+1}{n-x}W'}{1+\frac{x+1}{n-x}W'}$, again increasing in W', and the $1-\alpha/2$ quantile of W' is the upper cutoff $F_{2(x+1),2(n-x),\alpha/2}$. Hence

$$p_U = \frac{\frac{x+1}{n-x} F_{2(x+1),2(n-x),\alpha/2}}{1 + \frac{x+1}{n-x} F_{2(x+1),2(n-x),\alpha/2}},$$

and $[p_L, p_U]$ is a $1 - \alpha$ confidence interval by Theorem 9.2.14(i).
The adjustments are forced by the equations themselves: at $x = 0$, $P(X \geq 0|p) \equiv 1$ has no solution, and the confidence set is $\{p : P(X \leq 0|p) \geq \alpha/2\} = [0, p_U]$; at $x = n$, $P(X \leq n|p) \equiv 1$ and the set is $[p_L, 1]$.

9.4 Exercises 9.22–9.28

Problem 9.22. If $X \sim$ negative binomial(r, p), use the relationship between the binomial and negative binomial to show that a $1 - \alpha$ confidence interval for p is given by

$$\frac{1}{1 + \frac{x+1}{r} F_{2(x+1),2r,\alpha/2}} \;\leq\; p \;\leq\; \frac{\frac{r}{x} F_{2r,2x,\alpha/2}}{1 + \frac{r}{x} F_{2r,2x,\alpha/2}},$$

with a suitable modification if $x = 0$.

Solution. The endpoints are the $\alpha/2$ quantile of beta$(r, x + 1)$ and the $1 - \alpha/2$ quantile of beta(r, x), rewritten by the beta–F relation.
Here X counts failures before the rth success, $P(X = x \mid p) = \binom{r+x-1}{x} p^r (1 - p)^x$. The binomial–

negative binomial relationship is the statement that the rth success occurs on or before trial $r+x$ iff the first $r+x$ trials contain at least r successes:

$$P(X \le x \mid p) = P(Y \ge r \mid p), \qquad Y \sim \text{binomial}(r+x, p).$$

Now apply the identity of Exercise 2.40, $P_\theta(Y \ge r) = P(Z \le \theta)$ with $Z \sim \text{beta}(r, n-r+1)$, at $n = r+x$:

$$\begin{aligned} P(X \le x \mid p) &= P(Z_x \le p), & Z_x &\sim \text{beta}(r, x+1), \\ P(X \ge x \mid p) &= 1 - P(Z_{x-1} \le p), & Z_{x-1} &\sim \text{beta}(r, x). \end{aligned}$$

In particular $P(X \le x \mid p)$ is increasing in p, so Theorem 9.2.14(ii) applies with $\alpha_1 = \alpha_2 = \alpha/2$: the endpoints solve

$$P(X \ge x \mid p_U) = \tfrac{\alpha}{2}, \qquad P(X \le x \mid p_L) = \tfrac{\alpha}{2}.$$

By the two displays these say $P(Z_x \le p_L) = \alpha/2$ and $P(Z_{x-1} \le p_U) = 1 - \alpha/2$, i.e. p_L is the $\alpha/2$ quantile of $\text{beta}(r, x+1)$ and p_U the $(1-\alpha/2)$ quantile of $\text{beta}(r, x)$.
Convert with the beta–F relation of Chapter 5: if $Z \sim \text{beta}(a, b)$ then $\frac{b}{a}\frac{Z}{1-Z} \sim F_{2a,2b}$. For the lower endpoint,

$$\frac{x+1}{r} \cdot \frac{p_L}{1-p_L} = F_{2r,2(x+1),1-\alpha/2} = \frac{1}{F_{2(x+1),2r,\alpha/2}},$$

so that

$$\frac{p_L}{1-p_L} = \frac{r}{(x+1)F_{2(x+1),2r,\alpha/2}} \implies p_L = \frac{1}{1 + \frac{x+1}{r}F_{2(x+1),2r,\alpha/2}}.$$

For the upper endpoint,

$$\frac{x}{r} \cdot \frac{p_U}{1-p_U} = F_{2r,2x,\alpha/2} \implies p_U = \frac{\frac{r}{x}F_{2r,2x,\alpha/2}}{1 + \frac{r}{x}F_{2r,2x,\alpha/2}},$$

which is the stated interval.
Modification at $x = 0$: $P(X \ge 0 \mid p) = 1$ for every p, so the upper equation has no solution and one sets $p_U = 1$, leaving the one-sided

$$\left[\left(1 + \tfrac{1}{r}F_{2,2r,\alpha/2}\right)^{-1},\ 1 \right]$$

(all the error now sits in one tail, so α may replace $\alpha/2$).

Problem 9.23. (a) Let $X_1, \ldots, X_n$ be a random sample from a Poisson population with parameter λ and define $Y = \sum X_i$. In Example 9.2.15 a confidence interval for λ was found using the method of Section 9.2.3. Construct another interval for λ by inverting an LRT, and compare the intervals.
(b) The following data, the number of aphids per row in nine rows of a potato field, can be assumed to follow a Poisson distribution:

155 104 66 50 36 40 30 35 42

Use these data to construct a 90% LRT confidence interval for the mean number of aphids per row. Also, construct an interval using the method of Example 9.2.15.

Solution. (a) The inverted LRT is the likelihood interval

$$C(y) = \left\{ \lambda : \ e^{\,y-n\lambda}\left(\tfrac{n\lambda}{y}\right)^y \ge k \right\},$$

whose endpoints are the two roots of $(n\lambda)^y e^{-n\lambda} = k\, y^y e^{-y}$.
Indeed $Y = \sum X_i \sim \text{Poisson}(n\lambda)$ is sufficient, $L(\lambda \mid y) \propto e^{-n\lambda}\lambda^y$ with MLE $\hat\lambda = y/n$, so the LRT statistic for $H_0 : \lambda = \lambda_0$ is

$$\Lambda(y) = \frac{e^{-n\lambda_0}\lambda_0^{\,y}}{e^{-y}(y/n)^y} = e^{\,y-n\lambda_0}\left(\frac{n\lambda_0}{y}\right)^y,$$

and inverting $\{y : \Lambda(y) \ge k\}$ gives $C(y)$ (Section 9.2.1). Since $\log\Lambda$ is strictly concave in λ_0 with maximum at $\lambda_0 = y/n$, $C(y)$ is an interval. The exact k for coverage $1-\alpha$ must be found numerically

(Y is discrete); the standard choice takes $k = e^{-\chi^2_{1,\alpha}/2}$, since $-2\log\Lambda(Y) \to \chi^2_1$ by Theorem 10.3.1 (iid sampling from a one-parameter family, λ interior, regularity conditions met), so that

$$C(y) = \Big\{\lambda :\ 2\Big[y\log\frac{y}{n\lambda} - (y - n\lambda)\Big] \le \chi^2_{1,\alpha}\Big\}.$$

Contrast with Example 9.2.15, whose interval $\big[\frac{1}{2n}\chi^2_{2y,1-\alpha/2},\ \frac{1}{2n}\chi^2_{2(y+1),\alpha/2}\big]$ is exact (coverage $\ge 1-\alpha$ for every λ, the conservatism forced by discreteness) and splits the error equally between the tails. The LRT interval is only asymptotically $1-\alpha$, but shorter, being centred at the mode of the likelihood rather than tail-balanced; at $y = 0$ it is $[0, \chi^2_{1,\alpha}/(2n)]$ against $[0, \chi^2_{2,\alpha/2}/(2n)]$.

(b) Here $n = 9$ and $y = \sum x_i = 558$, so $\hat\lambda = \bar x = 62$.
LRT interval: with $\chi^2_{1,.10} = 2.7055$ (so $k = e^{-1.3528} = .2585$), solve

$$2\Big[558\log\frac{558}{9\lambda} - (558 - 9\lambda)\Big] = 2.7055$$

for the two roots, giving

$$57.78 \ \le\ \lambda \ \le\ 66.42 \qquad \text{(length 8.64)}.$$

Example 9.2.15 interval: $2y = 1116$, $2(y+1) = 1118$, $2n = 18$, so

$$\frac{\chi^2_{1116,.95}}{18} \le \lambda \le \frac{\chi^2_{1118,.05}}{18}, \quad \text{i.e.} \quad 57.75 \ \le\ \lambda \ \le\ 66.49 \qquad \text{(length 8.75)}.$$

The two agree to within 0.1 aphid at each end, the LRT interval sitting just inside the other, as predicted in (a).

Problem 9.24. For $X \sim \text{Poisson}(\lambda)$, show that the coverage probability of the confidence interval $[L(X), U(X)]$ in Example 9.2.15 is given by

$$P_\lambda\big(\lambda \in [L(X), U(X)]\big) = \sum_{x=0}^{\infty} I_{[L(x),U(x)]}(\lambda)\,\frac{e^{-\lambda}\lambda^x}{x!}$$

and that we can define functions $x_l(\lambda)$ and $x_u(\lambda)$ so that

$$P_\lambda\big(\lambda \in [L(X), U(X)]\big) = \sum_{x=x_l(\lambda)}^{x_u(\lambda)} \frac{e^{-\lambda}\lambda^x}{x!}.$$

Hence, explain why the graph of the coverage probability of the Poisson intervals given in Figure 9.2.5 has jumps occurring at the endpoints of the different confidence intervals.

Solution. The first display is just the definition of the expectation of an indicator,

$$P_\lambda\big(\lambda \in [L(X), U(X)]\big) = E_\lambda\Big[I_{[L(X),U(X)]}(\lambda)\Big] = \sum_{x=0}^{\infty} I_{[L(x),U(x)]}(\lambda)\frac{e^{-\lambda}\lambda^x}{x!}.$$

For the second, recall from (9.2.17) with $n = 1$ that $L(x) = \frac{1}{2}\chi^2_{2x,1-\alpha/2}$ and $U(x) = \frac{1}{2}\chi^2_{2(x+1),\alpha/2}$, both strictly increasing in x with $L(0) = 0$ and $L(x), U(x) \to \infty$. Hence for fixed λ the summation index set is an interval of integers: put

$$x_u(\lambda) = \max\{x : L(x) \le \lambda\}, \qquad x_l(\lambda) = \min\{x : U(x) \ge \lambda\},$$

both finite, and

$$\{x : \lambda \in [L(x), U(x)]\} = \{x_l(\lambda), x_l(\lambda) + 1, \ldots, x_u(\lambda)\}.$$

This set is nonempty because $U(x) > L(x+1)$ for every x (the upper cutoff exceeds the lower one at the same degrees of freedom, $\chi^2_{\nu,\alpha/2} > \chi^2_{\nu,1-\alpha/2}$ with $\nu = 2(x+1)$), so $\lambda < L(x_u + 1) < U(x_u)$ and

$x_l(\lambda) \le x_u(\lambda)$. Therefore

$$P_\lambda\big(\lambda \in [L(X), U(X)]\big) = \sum_{x=x_l(\lambda)}^{x_u(\lambda)} \frac{e^{-\lambda}\lambda^x}{x!}.$$

The jumps: each summand $e^{-\lambda}\lambda^x/x!$ is continuous in λ, so the only discontinuities of the coverage probability come from the step functions x_l and x_u, and these are nondecreasing step functions that jump by exactly 1 at

$$\text{(i) } \lambda = L(x) : \ x_u \text{ increases,} \qquad \text{(ii) } \lambda = U(x) : \ x_l \text{ increases.}$$

At (i) the term $e^{-\lambda}\lambda^x/x!$ is added to the sum and the coverage probability jumps up by that amount; at (ii) the term $e^{-\lambda}\lambda^{x_l}/x_l!$ is dropped and it jumps down. Both kinds of jump occur precisely at the values $L(x), U(x)$, that is, at the endpoints of the confidence intervals, which is the sawtooth pattern of Figure 9.2.5.

Problem 9.25. If $X_1, \dots, X_n$ are iid with pdf $f(x \mid \mu) = e^{-(x-\mu)} I_{[\mu,\infty)}(x)$, then $Y = \min\{X_1, \dots, X_n\}$ is sufficient for μ with pdf

$$f_Y(y \mid \mu) = n e^{-n(y-\mu)} I_{[\mu,\infty)}(y).$$

In Example 9.2.13 a $1-\alpha$ confidence interval for μ was found using the method of Section 9.2.3. Compare that interval to $1-\alpha$ intervals obtained by likelihood and pivotal methods.

Solution. The likelihood and (shortest) pivotal methods give the same interval,

$$\Big[Y + \tfrac{1}{n}\log\alpha, \ Y\Big],$$

of length $\frac{1}{n}\log\frac{1}{\alpha}$, which is shorter than the Example 9.2.13 interval $\big[Y + \frac{1}{n}\log\frac{\alpha}{2}, \ Y + \frac{1}{n}\log(1-\frac{\alpha}{2})\big]$ of length $\frac{1}{n}\log\frac{2-\alpha}{\alpha}$.

Pivotal method (Section 9.2.2). $T = Y - \mu$ has pdf $f_T(t) = ne^{-nt} I_{(0,\infty)}(t)$, free of μ, so it is a pivot; any $[a, b]$ with $P(a \le T \le b) = 1-\alpha$ inverts to $\{\mu : Y - b \le \mu \le Y - a\}$, of length $b - a$. Since f_T is strictly decreasing, translating $[a, b]$ leftward with its length held fixed strictly increases $\int_a^b f_T$; hence the shortest such interval has $a = 0$, and then $1 - e^{-nb} = 1 - \alpha$ gives $b = -\frac{1}{n}\log\alpha$. (Theorem 9.3.2 does not apply directly: its condition $f(a) = f(b) > 0$ is unattainable for a monotone pdf, and the optimum sits on the boundary of the support.)

Likelihood method. From the sufficient statistic, $L(\mu \mid y) = ne^{-n(y-\mu)} I_{(-\infty,y]}(\mu)$ is increasing in μ on $\mu \le y$, so $\hat\mu = y$ and the likelihood region is

$$\Big\{\mu : \frac{L(\mu \mid y)}{L(\hat\mu \mid y)} \ge k\Big\} = \{\mu : e^{-n(y-\mu)} \ge k, \ \mu \le y\} = \Big[y + \tfrac{1}{n}\log k, \ y\Big].$$

Its coverage is $P\big(0 \le T \le -\frac{1}{n}\log k\big) = 1 - k$, so $k = \alpha$ gives exactly $1-\alpha$ and reproduces the pivotal interval. (Equivalently: this is the inverted LRT, and here the LRT statistic is a monotone function of the pivot.)

Comparison of lengths. Since $2 - \alpha > 1$,

$$\frac{1}{n}\log\frac{1}{\alpha} \ < \ \frac{1}{n}\log\frac{2-\alpha}{\alpha},$$

e.g. at $\alpha = .10$ the lengths are $2.303/n$ versus $2.944/n$, a 22% saving. Splitting α equally between the two tails, as Section 9.2.3 does, is wasteful here because the pivot's pdf is monotone: all of the error belongs in the upper tail of T.

Problem 9.26. Let $X_1, \dots, X_n$ be iid observations from a beta$(\theta, 1)$ pdf and assume that θ has a gamma(r, λ) prior pdf. Find a $1-\alpha$ Bayes credible set for θ.

Solution. With $T = -\sum_{i=1}^n \log X_i$, the posterior is gamma$\big(n+r,\ (T+1/\lambda)^{-1}\big)$, and a $1-\alpha$ credible set is

$$\Big\{\theta:\ \frac{\chi^2_{2(n+r),1-\alpha/2}}{2\big(T+\frac{1}{\lambda}\big)}\ \le\ \theta\ \le\ \frac{\chi^2_{2(n+r),\alpha/2}}{2\big(T+\frac{1}{\lambda}\big)}\Big\}.$$

The beta$(\theta,1)$ pdf is $f(x\mid\theta)=\theta x^{\theta-1}$, $0<x<1$, so

$$L(\theta\mid \mathbf{x}) = \theta^n\Big(\prod x_i\Big)^{\theta-1} \propto \theta^n e^{-\theta T},$$

and with the gamma(r,λ) prior $\pi(\theta)\propto \theta^{r-1}e^{-\theta/\lambda}$,

$$\pi(\theta\mid\mathbf{x})\ \propto\ \theta^{\,n+r-1}\exp\Big\{-\theta\Big(T+\frac{1}{\lambda}\Big)\Big\},$$

a gamma$\big(n+r,(T+1/\lambda)^{-1}\big)$ density (so the gamma family is conjugate here). Hence, under the posterior, $2(T+1/\lambda)\theta\sim\chi^2_{2(n+r)}$, and splitting α equally between the tails of that chi squared gives the set displayed above, which has posterior probability exactly $1-\alpha$.
For the shortest set replace the equal-tailed cutoffs by the HPD cutoffs of Corollary 9.3.10: $\theta_L<\theta_U$ solving

$$\theta_L^{n+r-1}e^{-\theta_L(T+1/\lambda)} = \theta_U^{n+r-1}e^{-\theta_U(T+1/\lambda)}, \qquad \int_{\theta_L}^{\theta_U}\pi(\theta\mid\mathbf{x})\,d\theta = 1-\alpha,$$

the posterior being unimodal for $n+r\ge 1$.

Problem 9.27. (a) Let $X_1,\dots,X_n$ be iid observations from an exponential(λ) pdf, where λ has the conjugate IG(a,b) prior, an inverted gamma with pdf

$$\pi(\lambda\mid a,b) = \frac{1}{\Gamma(a)b^a}\Big(\frac{1}{\lambda}\Big)^{a+1}e^{-1/(b\lambda)}, \qquad 0<\lambda<\infty.$$

Show how to find a $1-\alpha$ Bayes HPD credible set for λ.
(b) Find a $1-\alpha$ Bayes HPD credible set for σ^2, the variance of a normal distribution, based on the sample variance s^2 and using a conjugate IG(a,b) prior for σ^2.
(c) Starting with the interval from part (b), find the limiting $1-\alpha$ Bayes HPD credible set for σ^2 obtained as $a\to 0$ and $b\to\infty$.

Solution. (a) The posterior is IG(a',b') with

$$a' = n+a, \qquad b' = \Big(\sum_{i=1}^n x_i + \frac{1}{b}\Big)^{-1},$$

and the HPD set is the interval $[\lambda_L,\lambda_U]$ solving the pair of equations below.
Indeed $L(\lambda\mid\mathbf{x})=\lambda^{-n}e^{-\sum x_i/\lambda}$, so

$$\pi(\lambda\mid\mathbf{x})\ \propto\ \Big(\frac{1}{\lambda}\Big)^{n+a+1}\exp\Big\{-\frac{1}{\lambda}\Big(\sum x_i+\frac{1}{b}\Big)\Big\},$$

which is IG(a',b') as claimed (so IG is conjugate). This density is unimodal, with mode $b'^{-1}/(a'+1)$, so Corollary 9.3.10 says the shortest $1-\alpha$ credible set is $\{\lambda:\pi(\lambda\mid\mathbf{x})\ge k\}=[\lambda_L,\lambda_U]$ where $\lambda_L<\lambda_U$ solve

$$\Big(\frac{1}{\lambda_L}\Big)^{a'+1}e^{-1/(b'\lambda_L)} = \Big(\frac{1}{\lambda_U}\Big)^{a'+1}e^{-1/(b'\lambda_U)},$$
$$P\Big(\frac{2}{b'\lambda_U}\le\chi^2_{2a'}\le\frac{2}{b'\lambda_L}\Big)=1-\alpha,$$

the second equation using that $1/\lambda \mid \mathbf{x} \sim \text{gamma}(a', b')$, i.e. $2/(b'\lambda) \mid \mathbf{x} \sim \chi^2_{2a'}$. Two equations in two unknowns; solve numerically.
(b) Only s^2 is used, and $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ gives

$$f(s^2 \mid \sigma^2) \;\propto\; (\sigma^2)^{-(n-1)/2}(s^2)^{(n-3)/2}e^{-(n-1)s^2/(2\sigma^2)},$$

so with the $\text{IG}(a,b)$ prior

$$\pi(\sigma^2 \mid s^2) \;\propto\; \Big(\frac{1}{\sigma^2}\Big)^{\frac{n-1}{2}+a+1} \exp\Big\{-\frac{1}{\sigma^2}\Big(\frac{(n-1)s^2}{2}+\frac{1}{b}\Big)\Big\},$$

that is, $\sigma^2 \mid s^2 \sim \text{IG}(A,B)$ with

$$A = a + \frac{n-1}{2}, \qquad B = \Big(\frac{(n-1)s^2}{2}+\frac{1}{b}\Big)^{-1}.$$

By part (a) the HPD set is $[\sigma^2_L, \sigma^2_U]$ solving

$$\Big(\frac{1}{\sigma^2_L}\Big)^{A+1} e^{-1/(B\sigma^2_L)} = \Big(\frac{1}{\sigma^2_U}\Big)^{A+1} e^{-1/(B\sigma^2_U)},$$
$$P\Big(\frac{2}{B\sigma^2_U} \le \chi^2_{2A} \le \frac{2}{B\sigma^2_L}\Big) = 1-\alpha.$$

(c) As $a \to 0$ and $b \to \infty$, $A \to (n-1)/2$ and $B \to 2/[(n-1)s^2]$, so

$$\sigma^2 \mid s^2 \;\sim\; \text{IG}\Big(\frac{n-1}{2},\; \frac{2}{(n-1)s^2}\Big), \qquad \text{i.e.} \qquad \frac{(n-1)s^2}{\sigma^2}\Big| s^2 \;\sim\; \chi^2_{n-1},$$

the posterior distribution of the classical pivot. Writing $c = (n-1)s^2$ and $\sigma^2 = c/u$, the limiting HPD set is

$$\Big[\frac{(n-1)s^2}{u_U},\; \frac{(n-1)s^2}{u_L}\Big],$$

where $u_L < u_U$ solve

$$u_L^{(n+1)/2}e^{-u_L/2} = u_U^{(n+1)/2}e^{-u_U/2}, \qquad P\big(u_L \le \chi^2_{n-1} \le u_U\big) = 1-\alpha.$$

The first condition says $f_{n+3}(u_L) = f_{n+3}(u_U)$ for a χ^2_{n+3} density, equivalently $u_L^2 f_{n-1}(u_L) = u_U^2 f_{n-1}(u_U)$ — which is exactly the stationarity condition for minimising $c(1/u_L - 1/u_U)$ subject to $\int_{u_L}^{u_U} f_{n-1} = 1-\alpha$. So the limiting HPD set is the shortest classical $1-\alpha$ interval for σ^2 based on the χ^2_{n-1} pivot.

Problem 9.28. Let $X_1, \ldots, X_n$ be iid $\text{n}(\theta, \sigma^2)$, where both θ and σ^2 are unknown, but there is only interest on inference about θ. Consider the prior pdf

$$\pi(\theta, \sigma^2 \mid \mu, \tau^2, a, b) = \frac{1}{\sqrt{2\pi\tau^2\sigma^2}} e^{-(\theta-\mu)^2/(2\tau^2\sigma^2)} \cdot \frac{1}{\Gamma(a)b^a}\Big(\frac{1}{\sigma^2}\Big)^{a+1} e^{-1/(b\sigma^2)},$$

a $\text{n}(\mu, \tau^2\sigma^2)$ multiplied by an $\text{IG}(a,b)$.
(a) Show that this prior is a conjugate prior for this problem.
(b) Find the posterior distribution of θ and use it to construct a $1-\alpha$ credible set for θ.
(c) The classical $1-\alpha$ confidence set for θ can be expressed as

$$\Big\{\theta :\; |\theta - \bar{x}|^2 \le F_{1,n-1,\alpha}\frac{s^2}{n}\Big\}.$$

Is there any (limiting) sequence of τ^2, a, and b that would allow this set to be approached by a Bayes set from part (b)?

Solution. (a) The posterior is again $\mathrm{n}(\mu', \tau'^2\sigma^2) \times \mathrm{IG}(a', b')$, with

$$\mu' = \frac{n\tau^2\bar{x} + \mu}{n\tau^2 + 1}, \qquad \tau'^2 = \frac{\tau^2}{n\tau^2 + 1},$$
$$a' = a + \frac{n}{2}, \qquad \frac{1}{b'} = \frac{1}{b} + \frac{(n-1)s^2}{2} + \frac{n(\bar{x}-\mu)^2}{2(1+n\tau^2)}.$$

To see it, write the likelihood through the sufficient statistics $(\bar{x}, s^2)$,

$$f(\mathbf{x} \mid \theta, \sigma^2) \;\propto\; (\sigma^2)^{-n/2} \exp\Big\{ -\frac{(n-1)s^2 + n(\bar{x}-\theta)^2}{2\sigma^2} \Big\},$$

multiply by the prior, and complete the square in θ:

$$n(\theta - \bar{x})^2 + \frac{(\theta-\mu)^2}{\tau^2} = \Big(n + \frac{1}{\tau^2}\Big)(\theta - \mu')^2 + \frac{n(\bar{x}-\mu)^2}{1+n\tau^2}.$$

Then

$$\pi(\theta, \sigma^2 \mid \mathbf{x}) \;\propto\; \underbrace{(\sigma^2)^{-1/2} e^{-(\theta-\mu')^2/(2\tau'^2\sigma^2)}}_{\mathrm{n}(\mu', \tau'^2\sigma^2)} \cdot \underbrace{\Big(\frac{1}{\sigma^2}\Big)^{a'+1} e^{-1/(b'\sigma^2)}}_{\mathrm{IG}(a', b')},$$

with $\tau'^2 = (n + 1/\tau^2)^{-1}$ and a', b' as displayed. Same family, so the prior is conjugate.
(b) $\theta \mid \mathbf{x}$ is a scaled Student's t on $2a'$ degrees of freedom:

$$\frac{\theta - \mu'}{\tau'/\sqrt{a'b'}} \;\Big|\; \mathbf{x} \;\sim\; t_{2a'}.$$

Indeed $(\theta - \mu')/(\tau'\sigma) \mid \sigma^2, \mathbf{x} \sim \mathrm{n}(0,1)$ while $V = 2/(b'\sigma^2) \mid \mathbf{x} \sim \chi^2_{2a'}$, so

$$\frac{\theta - \mu'}{\tau'}\sqrt{a'b'} = Z\sqrt{\frac{2a'}{V}} = \frac{Z}{\sqrt{V/(2a')}} \;\sim\; t_{2a'}.$$

Hence a $1-\alpha$ credible set (an HPD set, since the t density is symmetric and unimodal — Corollary 9.3.10) is

$$\Big\{\theta :\; |\theta - \mu'| \;\le\; t_{2a',\alpha/2}\,\frac{\tau'}{\sqrt{a'b'}}\Big\}.$$

(c) Yes, but only by leaving the IG parameter space: take $\tau^2 \to \infty$, $b \to \infty$, $a \to -\frac{1}{2}$.
Let $\tau^2 \to \infty$ and $b \to \infty$. Then

$$\mu' \to \bar{x}, \qquad \tau'^2 \to \frac{1}{n}, \qquad b' \to \frac{2}{(n-1)s^2}, \qquad a'b' \to \frac{2a+n}{(n-1)s^2},$$

so the credible set of (b) tends to

$$\Big\{\theta :\; |\theta - \bar{x}| \le t_{2a+n,\alpha/2}\,\sqrt{\frac{n-1}{2a+n}}\;\frac{s}{\sqrt{n}}\Big\}.$$

The classical set is $\{\theta : |\theta - \bar{x}| \le t_{n-1,\alpha/2}\, s/\sqrt{n}\}$, since $F_{1,n-1,\alpha} = t^2_{n-1,\alpha/2}$. The two agree exactly iff $2a + n = n - 1$, i.e. $a = -\frac{1}{2}$ (which makes the scale factor $\sqrt{(n-1)/(2a+n)} = 1$ as well). At $a = -\frac{1}{2}$, $\tau^2 \to \infty$, $b \to \infty$ the prior degenerates to

$$\pi(\theta, \sigma^2) \;\propto\; (\sigma^2)^{-1/2}(\sigma^2)^{-a-1} = \frac{1}{\sigma^2},$$

the usual improper noninformative prior. Since a proper IG requires $a > 0$, no sequence with $a > 0$ reaches the classical set; the best such sequence, $a \to 0$ with $\tau^2, b \to \infty$, gives

$$\Big\{\theta : |\theta - \bar{x}| \le t_{n,\alpha/2}\sqrt{\tfrac{n-1}{n}}\;\tfrac{s}{\sqrt{n}}\Big\},$$

which agrees with the classical set only in the limit $n \to \infty$.

9.5 Exercises 9.29–9.35

Problem 9.29. Let $X_1, \dots, X_n$ be a sequence of n Bernoulli(p) trials.
(a) Calculate a $1-\alpha$ credible set for p using the conjugate beta(a, b) prior.
(b) Using the relationship between the beta and F distributions, write the credible set in a form that is comparable to the form of the intervals in Exercise 9.21. Compare the intervals.
(Exercise 9.21 showed that for $X \sim$ binomial(n, p) the Clopper–Pearson $1-\alpha$ confidence set is

$$\frac{1}{1+\frac{n-x+1}{x}F_{2(n-x+1),2x,\alpha/2}} \;\le\; p \;\le\; \frac{\frac{x+1}{n-x}F_{2(x+1),2(n-x),\alpha/2}}{1+\frac{x+1}{n-x}F_{2(x+1),2(n-x),\alpha/2}},$$

with the endpoint adjustments $p_L = 0$ if $x = 0$ and $p_U = 1$ if $x = n$.)

Solution. Write $y = \sum_i x_i$, $A = y + a$, $B = n - y + b$; the posterior is beta(A, B) and the equal-tailed credible set is $\{p : B_{A,B,1-\alpha/2} \le p \le B_{A,B,\alpha/2}\}$, where $B_{A,B,\gamma}$ is the upper-γ beta(A, B) cutoff.
(a) $Y = \sum_i X_i \sim$ binomial(n, p) is sufficient, so

$$\pi(p \mid y) \;\propto\; p^y(1-p)^{n-y} \cdot p^{a-1}(1-p)^{b-1} \;=\; p^{A-1}(1-p)^{B-1},$$

the beta(A, B) kernel (the conjugate updating of Section 7.2.3). Splitting α equally between the tails, as in (9.2.20), gives the set displayed above.
(b) By Theorem 5.3.8(c), if $F \sim F_{2A,2B}$ then $\frac{(A/B)F}{1+(A/B)F} \sim$ beta(A, B); equivalently, under the posterior

$$\frac{B}{A} \cdot \frac{p}{1-p} \;\sim\; F_{2A,2B}.$$

The map $t \mapsto t/(1+t)$ is increasing, so the credible set becomes

$$\frac{\frac{A}{B}F_{2A,2B,1-\alpha/2}}{1+\frac{A}{B}F_{2A,2B,1-\alpha/2}} \;\le\; p \;\le\; \frac{\frac{A}{B}F_{2A,2B,\alpha/2}}{1+\frac{A}{B}F_{2A,2B,\alpha/2}}.$$

Applying $F_{\nu_1,\nu_2,1-\gamma} = 1/F_{\nu_2,\nu_1,\gamma}$ (Theorem 5.3.8(a)) to the lower endpoint puts it in Clopper–Pearson shape:

$$\frac{1}{1+\frac{B}{A}F_{2B,2A,\alpha/2}} \;\le\; p \;\le\; \frac{\frac{A}{B}F_{2A,2B,\alpha/2}}{1+\frac{A}{B}F_{2A,2B,\alpha/2}}, \qquad A = y + a,\; B = n - y + b.$$

Comparison. The two intervals have literally the same functional form; only the beta parameters differ. Exercise 9.21 is the case

$$\begin{aligned}(A, B) &= (y,\; n-y+1) \quad \text{at the lower endpoint,}\\ (A, B) &= (y+1,\; n-y) \quad \text{at the upper endpoint,}\end{aligned}$$

i.e. the classical interval behaves like a Bayes interval that uses $(a, b) = (0, 1)$ below and $(a, b) = (1, 0)$ above, switching priors between endpoints. Since a beta quantile is increasing in A and decreasing in B, the flat-prior ($a = b = 1$) credible set, which uses $(A, B) = (y+1, n-y+1)$ at both ends, is strictly contained in the Clopper–Pearson set: the classical set pays for its guaranteed $\ge 1-\alpha$ coverage at every p with extra length. For $n = 10$, $y = 3$, $\alpha = .10$ the flat-prior credible set is $[.135, .564]$ while Exercise 9.21 gives $[.087, .607]$.

Problem 9.30. Complete the credible probability calculation needed in Example 9.2.17.
(In Example 9.2.16, $X_1, \dots, X_n$ are iid Poisson(λ) with prior $\lambda \sim$ gamma(a, b), so by (9.2.19) the posterior is

$$\pi\Big(\lambda \;\Big|\; \textstyle\sum X = \sum x\Big) = \text{gamma}\Big(a + \sum x,\; [\,n + (1/b)\,]^{-1}\Big),$$

and the credible probability of the classical confidence set of Example 9.2.15 is

$$(9.2.21) \qquad P\Big\{ \tfrac{1}{2n}\chi^2_{2\Sigma x,\,1-\alpha/2} \le \lambda \le \tfrac{1}{2n}\chi^2_{2(\Sigma x+1),\,\alpha/2} \Big\},$$

where λ has the posterior distribution (9.2.19).)
(a) Assume that a is an integer, and show that $T = \frac{2(nb+1)}{b}\lambda \sim \chi^2_{2(a+\Sigma x)}$.
(b) Show that

$$\frac{\chi^2_\nu - \nu}{\sqrt{2\nu}} \;\to\; \mathrm{n}(0,1)$$

as $\nu \to \infty$. (Use moment generating functions. The limit is difficult to evaluate—take logs and then do a Taylor expansion. Alternatively, see Example A.0.8 in Appendix A.)
(c) Standardize the random variable T of part (a), and write the credible probability (9.2.21) in terms of this variable. Show that the standardized lower cutoff point $\to \infty$ as $\sum x_i \to \infty$, and hence the credible probability goes to 0.

Solution. (a) The posterior scale is $\beta = [n + (1/b)]^{-1} = b/(nb+1)$, and rescaling a gamma rescales its scale parameter, so

$$T = \frac{2}{\beta}\lambda \;\sim\; \mathrm{gamma}\big(a + \textstyle\sum x,\; 2\big) = \chi^2_{2(a+\Sigma x)},$$

the last equality being the definition of the chi squared pdf (Section 3.3), legitimate because a integer makes $2(a + \sum x)$ a positive integer degrees of freedom.
(b) The χ^2_ν mgf is $(1-2t)^{-\nu/2}$ for $t < 1/2$, so $Z_\nu = (\chi^2_\nu - \nu)/\sqrt{2\nu}$ has

$$M_{Z_\nu}(t) = e^{-t\sqrt{\nu/2}}\,(1-u)^{-\nu/2}, \qquad u = t\sqrt{2/\nu}.$$

Taking logs and expanding $\log(1-u) = -u - \frac{u^2}{2} - \frac{u^3}{3} - \cdots$ for $|u| < 1$,

$$\begin{aligned} \log M_{Z_\nu}(t) &= -t\sqrt{\nu/2} + \frac{\nu}{2}\Big(u + \frac{u^2}{2} + O(u^3)\Big) \\ &= -t\sqrt{\nu/2} + t\sqrt{\nu/2} + \frac{t^2}{2} + O(\nu^{-1/2}) \;\longrightarrow\; \frac{t^2}{2}, \end{aligned}$$

using $\frac{\nu}{2}u = t\sqrt{\nu/2}$ and $\frac{\nu}{4}u^2 = t^2/2$. Thus $M_{Z_\nu}(t) \to e^{t^2/2}$ for each t in a neighborhood of 0, and Theorem 2.3.12 gives $Z_\nu \to \mathrm{n}(0,1)$ in distribution.
(c) Write $m = \sum x_i$, $\nu = 2(a+m)$ and $k = \frac{2(nb+1)}{b} \cdot \frac{1}{2n} = 1 + \frac{1}{nb} > 1$. Multiplying the three members of the event in (9.2.21) by $2(nb+1)/b$ and then standardizing T by part (a),

$$(9.2.21) = P\big(L_m \le Z \le U_m\big), \qquad Z = \frac{T-\nu}{\sqrt{2\nu}},$$

$$L_m = \frac{k\,\chi^2_{2m,1-\alpha/2} - \nu}{\sqrt{2\nu}}, \qquad U_m = \frac{k\,\chi^2_{2(m+1),\alpha/2} - \nu}{\sqrt{2\nu}}.$$

Part (b) gives the quantile expansion $\chi^2_{\nu,\gamma} = \nu + z_\gamma\sqrt{2\nu} + o(\sqrt{\nu})$ (convergence in distribution to a continuous strictly increasing limit cdf carries quantiles along), so with $\sqrt{2\cdot 2m} = 2\sqrt{m}$,

$$\begin{aligned} L_m &= \frac{k\big(2m - 2z_{\alpha/2}\sqrt{m} + o(\sqrt{m})\big) - 2(a+m)}{2\sqrt{a+m}} \\ &= \frac{2m(k-1) - 2kz_{\alpha/2}\sqrt{m} - 2a + o(\sqrt{m})}{2\sqrt{a+m}} \;\sim\; (k-1)\sqrt{m} = \frac{\sqrt{m}}{nb} \;\longrightarrow\; \infty. \end{aligned}$$

Hence for every M, eventually $L_m \ge M$, so

$$(9.2.21) \le P(Z \ge L_m) \le P(Z \ge M) \longrightarrow 1 - \Phi(M),$$

by part (b); letting $M \to \infty$ shows the credible probability tends to 0.

Problem 9.31. Complete the coverage probability calculation needed in Example 9.2.17. (There the $1-\alpha$ credible set (9.2.20) for the Poisson mean is

$$\Big\{\lambda:\ \tfrac{b}{2(nb+1)}\chi^2_{2(\Sigma x+a),1-\alpha/2} \le \lambda \le \tfrac{b}{2(nb+1)}\chi^2_{2(\Sigma x+a),\alpha/2}\Big\},$$

and writing $\lambda = \frac{\lambda}{\chi^2_{2Y}}\chi^2_{2Y}$, where $Y \sim \text{Poisson}(n\lambda)$ and $\lambda/\chi^2_{2Y} \to 1/(2n)$, its coverage probability as $\lambda \to \infty$ becomes

$$(9.2.22) \qquad P\Big(\tfrac{nb}{nb+1}\chi^2_{2(Y+a),1-\alpha/2} \le \chi^2_{2Y} \le \tfrac{nb}{nb+1}\chi^2_{2(Y+a),\alpha/2}\Big).$$

Here χ^2_{2Y} has $2Y$ degrees of freedom.)
(a) If χ^2_{2Y} is a chi squared random variable with $Y \sim \text{Poisson}(\lambda)$, show that $\text{E}(\chi^2_{2Y}) = 2\lambda$, $\text{Var}(\chi^2_{2Y}) = 8\lambda$, the mgf of χ^2_{2Y} is given by $\exp\big(-\lambda + \frac{\lambda}{1-2t}\big)$, and

$$\frac{\chi^2_{2Y} - 2\lambda}{\sqrt{8\lambda}} \ \to\ \text{n}(0,1)$$

as $\lambda \to \infty$. (Use moment generating functions.)
(b) Now evaluate (9.2.22) as $\lambda \to \infty$ by first standardizing χ^2_{2Y}. Show that the standardized upper limit goes to $-\infty$ as $\lambda \to \infty$, and hence the coverage probability goes to 0.

Solution. (a) Condition on Y, for which χ^2_{2Y} has mean $2Y$, variance $4Y$ and mgf $(1-2t)^{-Y}$. Theorems 4.4.3 and 4.4.7 (the conditional mean and variance identities) give

$$\text{E}(\chi^2_{2Y}) = \text{E}(2Y) = 2\lambda, \qquad \text{Var}(\chi^2_{2Y}) = \text{E}(4Y) + \text{Var}(2Y) = 4\lambda + 4\lambda = 8\lambda,$$

and, for $t < 1/2$ so that $s = (1-2t)^{-1}$ is in the domain of the Poisson pgf $\text{E}(s^Y) = e^{\lambda(s-1)}$,

$$M(t) = \text{E}\big[(1-2t)^{-Y}\big] = \exp\Big\{\lambda\Big(\tfrac{1}{1-2t} - 1\Big)\Big\} = \exp\Big(-\lambda + \tfrac{\lambda}{1-2t}\Big).$$

For $W_\lambda = (\chi^2_{2Y} - 2\lambda)/\sqrt{8\lambda}$ put $u = 2t/\sqrt{8\lambda} = t/\sqrt{2\lambda}$, so that $\lambda u = 2\lambda t/\sqrt{8\lambda}$ and $\lambda u^2 = t^2/2$:

$$\begin{aligned}\log M_{W_\lambda}(t) &= -\frac{2\lambda t}{\sqrt{8\lambda}} + \lambda \cdot \frac{u}{1-u} \\ &= -\lambda u + \lambda u + \lambda u^2 + O(\lambda u^3) \ \longrightarrow\ \frac{t^2}{2},\end{aligned}$$

since $\lambda u^3 = O(\lambda^{-1/2})$. Theorem 2.3.12 then gives $W_\lambda \to \text{n}(0,1)$.
(b) Put $\Lambda = n\lambda$, so $Y \sim \text{Poisson}(\Lambda)$, and $c = \frac{nb}{nb+1} < 1$. Choose $\delta \in \big(0, \frac{1}{nb}\big)$, which makes $c(1+\delta) = 1-\eta$ for some $\eta > 0$. On the event $A_\Lambda = \{Y \le \Lambda(1+\delta)\}$, whose complement has probability at most $\text{Var}(Y)/(\delta\Lambda)^2 = 1/(\delta^2\Lambda) \to 0$ by Chebychev, the monotonicity of $\chi^2_{\nu,\alpha/2}$ in ν bounds the upper limit by the deterministic

$$q_\Lambda = \chi^2_{2(\lfloor \Lambda(1+\delta)\rfloor + a),\,\alpha/2} = 2\Lambda(1+\delta) + O(\sqrt{\Lambda}),$$

the expansion coming from the quantile form of 9.30(b). Hence the standardized upper limit satisfies, on A_Λ,

$$\frac{c\,\chi^2_{2(Y+a),\alpha/2} - 2\Lambda}{\sqrt{8\Lambda}} \ \le\ \frac{c\,q_\Lambda - 2\Lambda}{\sqrt{8\Lambda}} = \frac{-2\eta\Lambda + O(\sqrt{\Lambda})}{\sqrt{8\Lambda}} \ \longrightarrow\ -\infty.$$

Therefore, discarding the lower constraint,

$$(9.2.22) \ \le\ P(A_\Lambda^c) + P\Big(W_\Lambda \le \frac{c\,q_\Lambda - 2\Lambda}{\sqrt{8\Lambda}}\Big) \ \longrightarrow\ 0,$$

the second term because $W_\Lambda \to \text{n}(0,1)$ by (a) while its cutoff tends to $-\infty$. The credible set has classical coverage tending to 0.

Problem 9.32. In this exercise we will calculate the classical coverage probability of the HPD region in (9.2.23), that is, the coverage probability of the Bayes HPD region using the probability model $\bar{X} \sim \mathrm{n}(\theta, \sigma^2/n)$.
(In Example 9.2.18, $X_1, \dots, X_n$ are iid $\mathrm{n}(\theta, \sigma^2)$ and θ has the prior $\mathrm{n}(\mu, \tau^2)$, with μ, σ, τ known, so that $\pi(\theta \mid \bar{x})$ is $\mathrm{n}\big(\delta^{\mathrm{B}}(\bar{x}), \mathrm{Var}(\theta \mid \bar{x})\big)$ with

$$\delta^{\mathrm{B}}(\bar{x}) = \frac{\sigma^2}{\sigma^2 + n\tau^2}\mu + \frac{n\tau^2}{\sigma^2 + n\tau^2}\bar{x}, \qquad \mathrm{Var}(\theta \mid \bar{x}) = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2},$$

and (9.2.23) is the $1-\alpha$ credible set

$$\delta^{\mathrm{B}}(\bar{x}) - z_{\alpha/2}\sqrt{\mathrm{Var}(\theta \mid \bar{x})} \le \theta \le \delta^{\mathrm{B}}(\bar{x}) + z_{\alpha/2}\sqrt{\mathrm{Var}(\theta \mid \bar{x})},$$

shown in Example 9.3.12 to be an HPD region. Throughout, $\gamma = \sigma^2/(n\tau^2)$.)
(a) Using the definitions given in Example 9.3.12, prove that

$$\begin{aligned} &P_\theta\Big(|\theta - \delta^{\mathrm{B}}(\bar{X})| \le z_{\alpha/2}\sqrt{\mathrm{Var}(\theta \mid \bar{X})}\Big) \\ &\quad = P_\theta\Big[-\sqrt{1+\gamma}\, z_{\alpha/2} + \frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}} \le Z \le \sqrt{1+\gamma}\, z_{\alpha/2} + \frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}}\Big]. \end{aligned}$$

(b) Show that the above set, although a $1-\alpha$ credible set, is not a $1-\alpha$ confidence set. (Fix $\theta \neq \mu$, let $\tau = \sigma/\sqrt{n}$, so that $\gamma = 1$. Prove that as $\sigma^2/n \to 0$, the above probability goes to 0.)
(c) If $\theta = \mu$, however, prove that the coverage probability is bounded away from 0. Find the minimum and maximum of this coverage probability.
(d) Now we will look at the other side. The usual $1-\alpha$ confidence set for θ is $\{\theta : |\theta - \bar{x}| \le z_{\alpha/2}\sigma/\sqrt{n}\}$. Show that the credible probability of this set is given by

$$\begin{aligned} &P_{\bar{x}}\big(|\theta - \bar{x}| \le z_{\alpha/2}\sigma/\sqrt{n}\big) \\ &\quad = P_{\bar{x}}\Big[-\sqrt{1+\gamma}\, z_{\alpha/2} + \frac{\gamma(\bar{x}-\mu)}{\sqrt{1+\gamma}\,\sigma/\sqrt{n}} \le Z \le \sqrt{1+\gamma}\, z_{\alpha/2} + \frac{\gamma(\bar{x}-\mu)}{\sqrt{1+\gamma}\,\sigma/\sqrt{n}}\Big] \end{aligned}$$

and that this probability is not bounded away from 0. Hence, the $1-\alpha$ confidence set is not a $1-\alpha$ credible set.

Solution. (a) In terms of $\gamma = \sigma^2/(n\tau^2)$ the two posterior quantities are

$$\delta^{\mathrm{B}}(\bar{x}) = \frac{\gamma}{1+\gamma}\mu + \frac{1}{1+\gamma}\bar{x}, \qquad \mathrm{Var}(\theta \mid \bar{x}) = \frac{\sigma^2}{n(1+\gamma)}$$

(divide numerator and denominator of each by $n\tau^2$). Splitting $\theta = \frac{\gamma}{1+\gamma}\theta + \frac{1}{1+\gamma}\theta$,

$$\theta - \delta^{\mathrm{B}}(\bar{X}) = \frac{\gamma}{1+\gamma}(\theta - \mu) + \frac{1}{1+\gamma}(\theta - \bar{X}).$$

Multiply the event $|\theta - \delta^{\mathrm{B}}(\bar{X})| \le z_{\alpha/2}\sqrt{\mathrm{Var}(\theta \mid \bar{X})}$ through by the positive constant $(1+\gamma)\sqrt{n}/\sigma$. The right side becomes $\sqrt{1+\gamma}\, z_{\alpha/2}$, and with $Z = \sqrt{n}(\bar{X} - \theta)/\sigma \sim \mathrm{n}(0,1)$ under P_θ the left side becomes $\big|\frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}} - Z\big|$. Undoing the absolute value gives exactly the displayed interval for Z.
(b) With $\tau = \sigma/\sqrt{n}$ we get $\gamma = 1$, so by (a) the coverage probability is

$$\Phi\big(\sqrt{2}\, z_{\alpha/2} + d\big) - \Phi\big(-\sqrt{2}\, z_{\alpha/2} + d\big), \qquad d = \frac{\theta - \mu}{\sigma/\sqrt{n}}.$$

The Z-interval has fixed length $2\sqrt{2}\, z_{\alpha/2}$ but is centered at d, and $|d| \to \infty$ as $\sigma^2/n \to 0$ since $\theta \neq \mu$; a fixed-length interval pushed to $\pm\infty$ has normal probability $\to 0$. So the credible set is not a confidence set.

(c) At $\theta = \mu$ the centering term vanishes and the coverage probability is

$$\Phi\big(\sqrt{1+\gamma}\, z_{\alpha/2}\big) - \Phi\big(-\sqrt{1+\gamma}\, z_{\alpha/2}\big) = 2\Phi\big(\sqrt{1+\gamma}\, z_{\alpha/2}\big) - 1,$$

free of σ, n except through γ and strictly increasing in $\gamma \in (0, \infty)$. Hence

$$\inf_{\gamma} = \lim_{\gamma \to 0} = 2\Phi(z_{\alpha/2}) - 1 = 1 - \alpha, \qquad \sup_{\gamma} = \lim_{\gamma \to \infty} = 1,$$

neither attained. In particular the coverage is always at least $1 - \alpha > 0$.
(d) Under the posterior, $Z = \big(\theta - \delta^{\mathrm{B}}(\bar{x})\big) \big/ \big(\sigma/\sqrt{n(1+\gamma)}\big) \sim \mathrm{n}(0,1)$, and from the form of δ^{B} in (a),

$$\delta^{\mathrm{B}}(\bar{x}) - \bar{x} = \frac{\gamma}{1+\gamma}(\mu - \bar{x}), \qquad \text{so} \qquad \theta - \bar{x} = \frac{\sigma\, Z}{\sqrt{n(1+\gamma)}} - \frac{\gamma(\bar{x} - \mu)}{1+\gamma}.$$

Multiplying $|\theta - \bar{x}| \le z_{\alpha/2}\sigma/\sqrt{n}$ by $\sqrt{n(1+\gamma)}/\sigma$ turns it into

$$\Big| Z - \frac{\gamma(\bar{x} - \mu)}{\sqrt{1+\gamma}\,\sigma/\sqrt{n}} \Big| \le \sqrt{1+\gamma}\, z_{\alpha/2},$$

which is the displayed credible probability. Again the Z-interval has fixed length $2\sqrt{1+\gamma}\, z_{\alpha/2}$ and center $\gamma(\bar{x} - \mu)/(\sqrt{1+\gamma}\,\sigma/\sqrt{n})$, which runs off to $\pm\infty$ for any fixed $\bar{x} \neq \mu$ as $\sigma/\sqrt{n} \to 0$ with γ held fixed (or, for fixed σ, n, τ, as $|\bar{x} - \mu| \to \infty$). So the credible probability is not bounded away from 0, and the classical $1 - \alpha$ confidence set is not a $1 - \alpha$ credible set.

Problem 9.33. Let $X \sim \mathrm{n}(\mu, 1)$ and consider the confidence interval

$$C_a(x) = \big\{\mu : \ \min\{0, (x-a)\} \le \mu \le \max\{0, (x+a)\}\big\}.$$

(a) For $a = 1.645$, prove that the coverage probability of $C_a(x)$ is exactly .95 for all μ, with the exception of $\mu = 0$, where the coverage probability is 1.
(b) Now consider the so-called noninformative prior $\pi(\mu) = 1$. Using this prior and again taking $a = 1.645$, show that the posterior credible probability of $C_a(x)$ is exactly .90 for $-1.645 \le x \le 1.645$ and increases to .95 as $|x| \to \infty$.
This type of interval arises in the problem of *bioequivalence*, where the objective is to decide if two treatments (different formulations of a drug, different delivery systems of a treatment) produce the same effect. The formulation of the problem results in "turning around" the roles of the null and alternative hypotheses (see Exercise 8.47), resulting in some interesting statistics. See Berger and Hsu (1996) for a review of bioequivalence and Brown, Casella, and Hwang (1995) for generalizations of the confidence set.

Solution. The interval always contains 0, since $\min\{0, x-a\} \le 0 \le \max\{0, x+a\}$ for every x; that single observation settles the case $\mu = 0$ and reduces the others to a one-sided event. Write $a = 1.645 = z_{.05}$, so $\Phi(a) = .95$.
(a) For $\mu = 0$ the containment just noted gives coverage 1. For $\mu > 0$ the lower constraint $\min\{0, X - a\} \le \mu$ holds automatically (its left side is $\le 0 < \mu$), so

$$P_\mu\big(\mu \in C_a(X)\big) = P_\mu\big(\mu \le \max\{0, X + a\}\big) = P_\mu(X + a \ge \mu) = \Phi(a) = .95,$$

the middle equality because $\mu > 0$ forces the maximum to be attained at $X + a$. For $\mu < 0$ the upper constraint is automatic and symmetrically

$$P_\mu\big(\mu \in C_a(X)\big) = P_\mu(X - a \le \mu) = \Phi(a) = .95.$$

(b) With $\pi(\mu) = 1$ the posterior is $\mu \mid x \sim \mathrm{n}(x, 1)$, so with $Z = \mu - x \sim \mathrm{n}(0, 1)$ under the posterior we read off the three cases from the locations of the min and max.
(i) $|x| \le a$: then $x - a \le 0 \le x + a$, so $C_a(x) = [x - a, x + a]$ and

$$P(\mu \in C_a(x) \mid x) = P(-a \le Z \le a) = 2\Phi(a) - 1 = .90.$$

(ii) $x > a$: then $C_a(x) = [0, x + a]$ and

$$P(\mu \in C_a(x) \mid x) = P(-x \le Z \le a) = \Phi(a) - \big(1 - \Phi(x)\big),$$

which equals .90 at $x = a$, increases in x, and $\to \Phi(a) = .95$ as $x \to \infty$.
(iii) $x < -a$: then $C_a(x) = [x - a, 0]$ and, by the symmetry $x \mapsto -x$,

$$P(\mu \in C_a(x) \mid x) = P(-a \le Z \le -x) = \Phi(|x|) - \big(1 - \Phi(a)\big),$$

again .90 at $x = -a$ and increasing to .95 as $x \to -\infty$.

Problem 9.34. Suppose that $X_1, \dots, X_n$ is a random sample from a $\mathrm{n}(\mu, \sigma^2)$ population.
(a) If σ^2 is known, find a minimum value for n to guarantee that a .95 confidence interval for μ will have length no more than $\sigma/4$.
(b) If σ^2 is unknown, find a minimum value for n to guarantee, with probability .90, that a .95 confidence interval for μ will have length no more than $\sigma/4$.

Solution. (a) $n = 246$. The interval is $\bar{x} \pm z_{.025}\sigma/\sqrt{n}$, of fixed length $2z_{.025}\sigma/\sqrt{n}$, so

$$\frac{2(1.96)\sigma}{\sqrt{n}} \le \frac{\sigma}{4} \iff \sqrt{n} \ge 8(1.96) = 15.68 \iff n \ge 245.86,$$

and n must be an integer.
(b) $n = 276$. Now the interval is $\bar{x} \pm t_{n-1,.025}S/\sqrt{n}$, whose length is random, and

$$\frac{2t_{n-1,.025}S}{\sqrt{n}} \le \frac{\sigma}{4} \iff \frac{(n-1)S^2}{\sigma^2} \le \frac{n(n-1)}{64\, t^2_{n-1,.025}}.$$

Since $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ (Theorem 5.3.1), the requirement is

$$P\Big(\chi^2_{n-1} \le \frac{n(n-1)}{64\, t^2_{n-1,.025}}\Big) \ge .90,$$

and the left side increases in n. At $n = 275$, $t_{274,.025} = 1.96866$ gives cutoff 303.78 and $P(\chi^2_{274} \le 303.78) = .8956 < .90$; at $n = 276$, $t_{275,.025} = 1.96863$ gives cutoff 306.01 and $P(\chi^2_{275} \le 306.01) = .9038 \ge .90$. (Check!)

Problem 9.35. Let $X_1, \dots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ population. Compare expected lengths of $1 - \alpha$ confidence intervals for μ that are computed assuming
(a) σ^2 is known.
(b) σ^2 is unknown.

Solution. The unknown-variance interval is longer in expectation for every n and every α, by the factor

$$\frac{\mathrm{E}L_t}{\mathrm{E}L_z} = \frac{t_{n-1,\alpha/2}}{z_{\alpha/2}} \cdot \sqrt{\frac{2}{n-1}}\, \frac{\Gamma(n/2)}{\Gamma\big(\frac{n-1}{2}\big)} \;>\; 1.$$

(a) $L_z = 2z_{\alpha/2}\sigma/\sqrt{n}$ is a constant, so $\mathrm{E}L_z = 2z_{\alpha/2}\sigma/\sqrt{n}$.
(b) $L_t = 2t_{n-1,\alpha/2}S/\sqrt{n}$, and since $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$,

$$\mathrm{E}S = \frac{\sigma}{\sqrt{n-1}}\, \mathrm{E}\sqrt{\chi^2_{n-1}} = \frac{\sigma}{\sqrt{n-1}} \cdot \frac{\sqrt{2}\,\Gamma(n/2)}{\Gamma\big(\frac{n-1}{2}\big)},$$

the middle expectation being the $\frac{1}{2}$-moment of a χ^2_{n-1}; hence $\mathrm{E}L_t = 2t_{n-1,\alpha/2}\,\mathrm{E}S/\sqrt{n}$ and the ratio above.

The ratio is a product of two competing factors: $t_{n-1,\alpha/2}/z_{\alpha/2} > 1$ inflates the interval to pay for estimating σ, while $c_n = \sqrt{2/(n-1)}\,\Gamma(n/2)/\Gamma(\frac{n-1}{2}) < 1$ deflates it because $\mathrm{E}S < \sigma$ (Jensen, Theorem 4.7.7, applied to the concave $\sqrt{\cdot}$). The t factor wins at every n; at $\alpha = .05$,

n	t/z	c_n	ratio
5	1.417	.940	1.332
10	1.154	.973	1.123
30	1.044	.991	1.035
100	1.012	.997	1.010

and both factors tend to 1, so the expected lengths agree asymptotically.

9.6 Exercises 9.36–9.42

Problem 9.36. Let $X_1, \ldots, X_n$ be independent with pdfs

$$f_{X_i}(x \mid \theta) = e^{i\theta - x} I_{[i\theta,\infty)}(x).$$

Prove that $T = \min_i(X_i/i)$ is a sufficient statistic for θ. Based on T, find the $1-\alpha$ confidence interval for θ of the form $[T+a, T+b]$ which is of minimum length.

Solution. The shortest interval is $[T - \frac{2}{n(n+1)}\log(1/\alpha),\ T]$, i.e. $a = \frac{2}{n(n+1)}\log\alpha$ and $b = 0$.
Sufficiency is the Factorization Theorem (Theorem 6.2.6): since $x_i \ge i\theta \iff x_i/i \ge \theta$, writing $m = \sum_{i=1}^n i = n(n+1)/2$,

$$\begin{aligned} f(\mathbf{x} \mid \theta) &= \prod_{i=1}^n e^{i\theta - x_i} I_{[i\theta,\infty)}(x_i) \\ &= e^{m\theta} I_{[\theta,\infty)}\big(T(\mathbf{x})\big) \cdot e^{-\sum_i x_i} \\ &= g\big(T(\mathbf{x}) \mid \theta\big)\, h(\mathbf{x}). \end{aligned}$$

For the interval, note $T - \theta$ is a pivot with an exponential$(1/m)$ distribution:

$$\begin{aligned} P_\theta(T > t) &= \prod_{i=1}^n P_\theta(X_i > it) = \prod_{i=1}^n e^{i\theta - it} \\ &= e^{-m(t-\theta)}, \qquad t \ge \theta, \end{aligned}$$

so $U = T - \theta$ has pdf $f_U(u) = me^{-mu}$ on $(0,\infty)$. Now

$$\theta \in [T+a, T+b] \iff -b \le U \le -a,$$

and the interval length is $b - a = (-a) - (-b)$. So minimizing length subject to $P(-b \le U \le -a) = 1-\alpha$ is exactly the problem of Exercise 9.41(a) for the strictly decreasing pdf f_U: the shortest such interval starts at 0. Hence $-b = 0$ and

$$\int_0^{-a} me^{-mu}\,du = 1 - e^{ma} = 1 - \alpha \quad\Longrightarrow\quad a = \frac{\log\alpha}{m}.$$

With $m = n(n+1)/2$ the minimum-length interval is

$$\Big[T + \tfrac{2}{n(n+1)}\log\alpha,\ T\Big], \qquad \text{length } \frac{2}{n(n+1)}\log\frac{1}{\alpha}.$$

Problem 9.37. Let $X_1, \ldots, X_n$ be iid uniform$(0,\theta)$. Let Y be the largest order statistic. Prove that Y/θ is a pivotal quantity and show that the interval

$$\Big\{\theta:\ y \le \theta \le \frac{y}{\alpha^{1/n}}\Big\}$$

is the shortest $1-\alpha$ pivotal interval.

Solution. $Y/\theta \sim \text{beta}(n,1)$, free of θ. Indeed $f_Y(y \mid \theta) = ny^{n-1}/\theta^n$ on $(0,\theta)$ (Theorem 5.4.4), so $U = Y/\theta$ has

$$f_U(u) = nu^{n-1}, \qquad 0 < u < 1,$$

which does not involve θ: U is a pivot, with $P(a \le U \le b) = b^n - a^n$.
Any pivotal interval comes from $a \le U \le b$, and replacing b by $\min(b,1)$ changes no probability while shortening the interval, so we may take $0 < a \le b \le 1$; then

$$\frac{y}{b} \le \theta \le \frac{y}{a}, \qquad \text{length } \; y\Big(\frac{1}{a} - \frac{1}{b}\Big),$$

so we must minimize $L = \frac{1}{a} - \frac{1}{b}$ subject to $b^n - a^n = 1-\alpha$ and $b \le 1$. The constraint gives $b(a) = (1-\alpha+a^n)^{1/n}$, which forces $a \le \alpha^{1/n}$, and $b'(a) = a^{n-1}/b^{n-1}$. Hence

$$L'(a) = -\frac{1}{a^2} + \frac{b'(a)}{b^2} = -\frac{1}{a^2} + \frac{a^{n-1}}{b^{n+1}} = \frac{a^{n+1} - b^{n+1}}{a^2 b^{n+1}} < 0,$$

since $a < b$. So L is strictly decreasing on $(0, \alpha^{1/n}]$ and is minimized at the right endpoint $a = \alpha^{1/n}$, $b = 1$, giving

$$\Big\{\theta : \; y \le \theta \le \frac{y}{\alpha^{1/n}}\Big\}, \qquad \text{length } \; y\big(\alpha^{-1/n} - 1\big).$$

Problem 9.38. If, in Theorem 9.3.2, we assume that f is continuous, then we can simplify the proof. For fixed c, consider the integral $\int_a^{a+c} f(x)\,dx$.
(a) Show that $\dfrac{d}{da}\displaystyle\int_a^{a+c} f(x)\,dx = f(a+c) - f(a)$.
(b) Prove that the unimodality of f implies that $\int_a^{a+c} f(x)\,dx$ is maximized when a satisfies $f(a+c) - f(a) = 0$.
(c) Suppose that, given α, we choose c^* and a^* to satisfy $\int_{a^*}^{a^*+c^*} f(x)\,dx = 1-\alpha$ and $f(a^*+c^*) - f(a^*) = 0$. Prove that this is the shortest $1-\alpha$ interval.

Solution. (a) Write $h(a) = \int_a^{a+c} f(x)\,dx = F(a+c) - F(a)$. Since f is continuous, the Fundamental Theorem of Calculus gives $F' = f$, so

$$h'(a) = f(a+c) - f(a).$$

(b) Let x^* be a mode, so f is nondecreasing on $(-\infty, x^*]$ and nonincreasing on $[x^*, \infty)$. Then

$$\begin{aligned} a \le x^* - c \Rightarrow a < a + c \le x^* \Rightarrow h'(a) = f(a+c) - f(a) \ge 0, \\ a \ge x^* \Rightarrow x^* \le a < a + c \Rightarrow h'(a) = f(a+c) - f(a) \le 0. \end{aligned}$$

So h is nondecreasing up to $x^* - c$ and nonincreasing from x^* onward; being continuous, it attains its maximum over all of $\mathbb{R}$ at some $\hat{a} \in [x^* - c, x^*]$. Since h is differentiable everywhere (part (a)) and $\hat{a}$ is a global maximum, Fermat's rule gives $h'(\hat{a}) = 0$, i.e. $f(\hat{a}+c) = f(\hat{a})$.
(c) First, a^* maximizes h for $c = c^*$. We may assume $a^* \le x^* \le a^* + c^*$: if both endpoints lay on one side of x^*, monotonicity together with $f(a^*) = f(a^*+c^*)$ would make f constant on $[a^*, a^*+c^*]$, and we may replace a^* by the stationary maximizer of part (b), which lies in $[x^* - c^*, x^*]$. Now for $a < a^*$,

$$\begin{aligned} a + c^* \le x^* \Rightarrow f(a) \le f(a + c^*), \\ a + c^* > x^* \Rightarrow f(a) \le f(a^*) = f(a^* + c^*) \le f(a + c^*), \end{aligned}$$

by $f \uparrow$ on $(-\infty, x^*]$ for the first line and for $f(a) \le f(a^*)$, and by $f \downarrow$ on $[x^*, \infty)$ for the last step, where $x^* < a + c^* < a^* + c^*$. So $h' \ge 0$ on $(-\infty, a^*]$, and the mirror argument gives $h' \le 0$ on $[a^*, \infty)$.

Thus

$$M(c^*) := \max_a \int_a^{a+c^*} f(x)\,dx = 1-\alpha.$$

Now M is strictly increasing wherever $M(c) < 1$. Let $c < c'$ and let $a_c \in [x^* - c, x^*]$ maximize for length c, so that

$$M(c') \;\geq\; M(c) + \max\Big\{ \int_{a_c+c}^{a_c+c'} f,\; \int_{a_c-(c'-c)}^{a_c} f \Big\}.$$

Since $M(c) < 1$, mass lies outside $[a_c, a_c + c]$, and either way the bracket is positive:
(i) if $\int_{-\infty}^{a_c} f > 0$ then $f(y) > 0$ for some $y < a_c \leq x^*$, so $f \geq f(y) > 0$ on $[y, a_c]$, whence $\int_{a_c-(c'-c)}^{a_c} f > 0$;
(ii) if $\int_{a_c+c}^{\infty} f > 0$ then $f(y) > 0$ for some $y > a_c + c \geq x^*$, so $f \geq f(y) > 0$ on $[a_c + c, y]$, whence $\int_{a_c+c}^{a_c+c'} f > 0$.
Therefore $M(c) < M(c^*) = 1 - \alpha$ for every $c < c^*$, so no interval of length less than c^* has probability $1 - \alpha$, and $[a^*, a^* + c^*]$ is the shortest $1 - \alpha$ interval.

Problem 9.39. Prove a special case of Theorem 9.3.2. Let $X \sim f(x)$, where f is a symmetric unimodal pdf. For a fixed value of $1 - \alpha$, of all intervals $[a, b]$ that satisfy $\int_a^b f(x)\,dx = 1 - \alpha$, the shortest is obtained by choosing a and b so that $\int_{-\infty}^{a} f(x)\,dx = \alpha/2$ and $\int_b^{\infty} f(x)\,dx = \alpha/2$.

Solution. Take f symmetric about 0 (otherwise translate), so $f(x) = f(|x|)$ with f nonincreasing on $[0, \infty)$, and let z solve $\int_{-z}^{z} f = 1 - \alpha$; by symmetry this $[-z, z]$ is exactly the equal-tailed interval, $\int_{-\infty}^{-z} f = \int_z^{\infty} f = \alpha/2$. We show every competitor is at least as long.
Let $[a, b]$ satisfy $\int_a^b f = 1 - \alpha$ and put $t = (a + b)/2$, $h = (b - a)/2$; by symmetry we may take $t \geq 0$. Cancelling the common part of $[-h, h]$ and $[t - h, t + h]$,

$$\int_{-h}^{h} f - \int_{t-h}^{t+h} f = \int_A f - \int_B f,$$

where $A = [-h, h] \setminus [t - h, t + h]$ and $B = [t - h, t + h] \setminus [-h, h]$ are intervals of equal length (t if $t \leq 2h$, otherwise $2h$) satisfying $|x| \leq h \leq |y|$ for every $x \in A$, $y \in B$. Since $f(x) = f(|x|)$ is nonincreasing in $|x|$, each such pair has $f(x) \geq f(y)$, so $\int_A f \geq \int_B f$ and

$$\int_{-h}^{h} f(x)\,dx \;\geq\; \int_{t-h}^{t+h} f(x)\,dx = 1 - \alpha.$$

If $h < z$ this forces $\int_h^z f = 0$; but $\int_{-z}^{z} f = 1 - \alpha < 1$ puts mass outside $[-z, z]$, so $f(y) > 0$ for some $y > z$, and f nonincreasing on $[0, \infty)$ gives $f \geq f(y) > 0$ on $[h, z]$ — a contradiction. Therefore $h \geq z$, i.e.

$$b - a \;\geq\; 2z,$$

with equality for the equal-tailed interval $[-z, z]$.

Problem 9.40. Building on Exercise 9.39, show that if f is symmetric, the optimal interval is of the form $m \pm k$, where m is the mode of f and k is a constant. Hence, show that (a) symmetric likelihood functions produce likelihood regions that are symmetric about the MLE if k' does not depend on the parameter (see (9.2.7)), and (b) symmetric posterior densities produce HPD regions that are symmetric about the posterior mean.

Solution. A symmetric unimodal f has its point of symmetry m as its mode, so Exercise 9.39 (translated to centre m) says the shortest $1 - \alpha$ interval is

$$[m - k,\; m + k], \qquad k \text{ solving } \int_{m-k}^{m+k} f(x)\,dx = 1 - \alpha.$$

Equivalently, that interval is the level set $\{x : f(x) \geq f(m+k)\}$: symmetry plus unimodality make every level set of f an interval centred at m.
(a) With k' free of θ, the likelihood region (9.2.7) is the level set

$$C(\mathbf{x}) = \{\theta : \ L(\theta \mid \mathbf{x}) \geq k'(\mathbf{x})\}.$$

If $L(\cdot \mid \mathbf{x})$ is symmetric, its centre of symmetry is its maximizer, the MLE $\hat{\theta}$, and $L(\hat{\theta}+t \mid \mathbf{x}) = L(\hat{\theta}-t \mid \mathbf{x})$ gives

$$\hat{\theta} + t \in C(\mathbf{x}) \iff \hat{\theta} - t \in C(\mathbf{x}),$$

so $C(\mathbf{x}) = \hat{\theta} \pm k$. (If k' depended on θ the defining inequality would no longer be invariant under $t \mapsto -t$, and the region need not be symmetric.)
(b) An HPD region is by definition a level set $\{\theta : \pi(\theta \mid \mathbf{x}) \geq k'\}$ of the posterior. If $\pi(\cdot \mid \mathbf{x})$ is symmetric about m then, whenever the posterior mean exists, $E(\theta \mid \mathbf{x}) = m$; the same reflection argument as in (a) gives

$$\{\theta : \ \pi(\theta \mid \mathbf{x}) \geq k'\} = \big\{\theta : \ |\theta - E(\theta \mid \mathbf{x})| \leq k\big\},$$

symmetric about the posterior mean.

Problem 9.41. (a) Prove the following, which is related to Theorem 9.3.2. Let $X \sim f(x)$, where f is a strictly decreasing pdf on $[0, \infty)$. For a fixed value of $1-\alpha$, of all intervals $[a, b]$ that satisfy $\int_a^b f(x)\,dx = 1-\alpha$, the shortest is obtained by choosing $a = 0$ and b so that

$$\int_0^b f(x)\,dx = 1-\alpha.$$

(b) Use the result of part (a) to find the shortest $1-\alpha$ confidence interval in Example 9.2.13.
In Example 9.2.13, $X_1, \ldots, X_n$ are iid with the location exponential pdf

$$f(x \mid \mu) = e^{-(x-\mu)} I_{[\mu,\infty)}(x),$$

so $Y = \min\{X_1, \ldots, X_n\}$ is sufficient, with pdf $ne^{-n(y-\mu)} I_{[\mu,\infty)}(y)$. Pivoting the cdf with an equal split gives the $1-\alpha$ interval

$$C(Y) = \Big\{\mu : \ Y + \tfrac{1}{n}\log\tfrac{\alpha}{2} \ \leq \ \mu \ \leq \ Y + \tfrac{1}{n}\log\big(1 - \tfrac{\alpha}{2}\big)\Big\}.$$

Solution. (a) Shift the competitor to the origin and compare mass. Let b_0 satisfy $\int_0^{b_0} f = 1-\alpha$, and let $0 \leq a < b$ with $\int_a^b f = 1-\alpha$; suppose $a > 0$. Substituting $x \mapsto x - a$ and using $f(x-a) > f(x)$ for $x > a$,

$$\begin{aligned}\int_0^{b-a} f(x)\,dx = \int_a^b f(x-a)\,dx &> \int_a^b f(x)\,dx \\ &= 1-\alpha = \int_0^{b_0} f(x)\,dx.\end{aligned}$$

A pdf strictly decreasing on all of $[0, \infty)$ is strictly positive there, so $F(t) = \int_0^t f$ is strictly increasing; hence $b - a > b_0$. Thus $a = 0$, $b = b_0$ gives the unique shortest interval.
(b) $n(Y-\mu)$ is a pivot with pdf e^{-w} on $(0, \infty)$, strictly decreasing, since $P_\mu(Y > y) = e^{-n(y-\mu)}$ for $y \geq \mu$. An interval $a \leq n(Y-\mu) \leq b$ inverts to

$$Y - \frac{b}{n} \ \leq \ \mu \ \leq \ Y - \frac{a}{n}, \qquad \text{length } \frac{b-a}{n},$$

so by part (a) take $a = 0$ and $\int_0^b e^{-w}\,dw = 1 - e^{-b} = 1-\alpha$, i.e. $b = \log(1/\alpha)$. The shortest $1-\alpha$ interval is

$$\Big\{\mu : \ Y + \tfrac{1}{n}\log\alpha \ \leq \ \mu \ \leq \ Y\Big\}, \qquad \text{length } \frac{1}{n}\log\frac{1}{\alpha}.$$

Problem 9.42. Referring to Example 9.3.4, to find the shortest pivotal interval for a gamma scale parameter, we had to solve a constrained minimization problem. (In Example 9.3.4, $X \sim \text{gamma}(k, \beta)$ and $Y = X/\beta \sim \text{gamma}(k, 1)$ is a pivot; a choice of $a < b$ with $P(a \le Y \le b) = 1 - \alpha$ inverts to $\{\beta : x/b \le \beta \le x/a\}$, whose length is proportional to $\frac{1}{a} - \frac{1}{b}$, so one must minimize $\frac{1}{a} - \frac{1}{b(a)}$ subject to $\int_a^{b(a)} f_Y(y)\,dy = 1 - \alpha$.)
(a) Show that the solution is given by the a and b that satisfy $\int_a^b f_Y(y)\,dy = 1-\alpha$ and $f(b)b^2 = f(a)a^2$.
(b) With one observation from a gamma(k, β) pdf with known shape parameter k, find the shortest $1 - \alpha$ (pivotal) confidence interval of the form $\{\beta : x/b \le \beta \le x/a\}$.

Solution. (a) Differentiate both equations. The constraint $\int_a^{b(a)} f_Y(y)\,dy = 1 - \alpha$ gives, by the Fundamental Theorem of Calculus,

$$f\big(b(a)\big)\frac{db}{da} - f(a) = 0 \quad \Longrightarrow \quad \frac{db}{da} = \frac{f(a)}{f(b)}.$$

Setting the derivative of the objective to zero gives

$$\frac{d}{da}\Big(\frac{1}{a} - \frac{1}{b(a)}\Big) = -\frac{1}{a^2} + \frac{1}{b^2}\frac{db}{da} = 0 \quad \Longrightarrow \quad \frac{db}{da} = \frac{b^2}{a^2}.$$

Equating the two expressions for db/da yields $f(b)b^2 = f(a)a^2$, as claimed.
That this stationary point is the global minimum is Theorem 9.3.2 in disguise: put $V = 1/Y$, with pdf $f_V(v) = f(1/v)/v^2$. Then $a \le Y \le b \iff 1/b \le V \le 1/a$, the probability constraint is unchanged, and the quantity to be minimized, $\frac{1}{a} - \frac{1}{b}$, is exactly the length of $[1/b, 1/a]$. Theorem 9.3.2 applies whenever f_V is unimodal, and its condition $f_V(1/a) = f_V(1/b)$ reads $f(a)a^2 = f(b)b^2$.
(b) Here $f(y) = y^{k-1}e^{-y}/\Gamma(k)$, so $f(y)y^2 \propto y^{k+1}e^{-y}$ and the pair of equations is

$$\int_a^b \frac{y^{k-1}e^{-y}}{\Gamma(k)}\,dy = 1 - \alpha,$$
$$a^{k+1}e^{-a} = b^{k+1}e^{-b},$$

the second equivalently $(k+1)\log a - a = (k+1)\log b - b$. Here $f_V(v) \propto v^{-(k+1)}e^{-1/v}$ is inverted gamma, with

$$(\log f_V)'(v) = \frac{1 - (k+1)v}{v^2},$$

so f_V is unimodal with mode $1/(k+1)$ and the hypothesis of Theorem 9.3.2 holds; the pair above therefore has a unique solution, with $a < k + 1 < b$, and it gives the minimum. The shortest pivotal interval is

$$\Big\{\beta : \ \frac{x}{b} \le \beta \le \frac{x}{a}\Big\}.$$

9.7 Exercises 9.43–9.49

Problem 9.43. Juola (1993) makes the following observation. If we have a pivot $Q(X, \theta)$, a $1 - \alpha$ confidence interval involves finding a and b so that $P(a < Q < b) = 1 - \alpha$. Typically the length of the interval on θ will be some function of a and b like $b - a$ or $1/b^2 - 1/a^2$. If Q has density f and the length can be expressed as $\int_a^b g(t)\,dt$, the shortest pivotal interval is the solution to

$$\min_{\{a,b\}} \int_a^b g(t)\,dt \quad \text{subject to} \quad \int_a^b f(t)\,dt = 1 - \alpha,$$

or, more generally,

$$\min_{C} \int_C g(t)\,dt \quad \text{subject to} \quad \int_C f(t)\,dt \ge 1 - \alpha.$$

(a) Prove that the solution is $C = \{t : g(t) < \lambda f(t)\}$, where λ is chosen so that $\int_C f(t)\,dt = 1-\alpha$. (Hint: You can adapt the proof of Theorem 8.3.12, the Neyman-Pearson Lemma.)
(b) Apply the result in part (a) to get the shortest intervals in Exercises 9.37 and 9.42.
For reference, Exercise 9.37 has $X_1, \dots, X_n$ iid uniform$(0,\theta)$ with Y the largest order statistic, and asserts that $\{\theta : y \le \theta \le y/\alpha^{1/n}\}$ is the shortest $1-\alpha$ pivotal interval; Exercise 9.42(b) takes one observation from a gamma(k,β) pdf with known shape k and asks for the shortest $1-\alpha$ pivotal interval of the form $\{\beta : x/b \le \beta \le x/a\}$, part (a) of that exercise having shown that the optimal a, b satisfy $\int_a^b f_Y(y)\,dy = 1-\alpha$ and $f(b)b^2 = f(a)a^2$.

Solution. (a) Let $C = \{t : g(t) < \lambda f(t)\}$ with $\lambda > 0$ chosen so that $\int_C f = 1-\alpha$, and let C' be any competitor with $\int_{C'} f \ge 1-\alpha$. Splitting both integrals over $C \cap C'$, which cancels,

$$\begin{aligned}\int_C g - \int_{C'} g &= \int_{C\cap C'^c} g \; - \int_{C^c\cap C'} g \\ &\le \lambda \int_{C\cap C'^c} f \; - \lambda \int_{C^c\cap C'} f \\ &= \lambda\left(\int_C f - \int_{C'} f\right) \\ &\le \lambda\big((1-\alpha)-(1-\alpha)\big) = 0,\end{aligned}$$

the middle inequality because $g < \lambda f$ on C and $g \ge \lambda f$ on C^c, and the last because $\int_{C'} f \ge 1-\alpha$. This is the Neyman-Pearson exchange of Theorem 8.3.12 with g in the role of the alternative density.
(b) In both applications the interval is a scale interval $\{\theta : x/b \le \theta \le x/a\}$, whose length is

$$x\left(\frac{1}{a}-\frac{1}{b}\right) = x\int_a^b \frac{dt}{t^2}, \qquad \text{so } g(t) = t^{-2}.$$

Exercise 9.37. Here $Q = Y/\theta$ has pdf $f(t) = nt^{n-1}$ on $(0,1)$, and $a < Q < b$ inverts to $\{\theta : y/b \le \theta \le y/a\}$. By part (a),

$$C = \{t : t^{-2} < \lambda n t^{n-1}\} = \{t \in (0,1) : t^{n+1} > (\lambda n)^{-1}\} = (c, 1),$$

and $1-\alpha = \int_c^1 nt^{n-1}\,dt = 1-c^n$ forces $c = \alpha^{1/n}$. Thus $a = \alpha^{1/n}$, $b = 1$, and the shortest interval is $\{\theta : y \le \theta \le y/\alpha^{1/n}\}$, as claimed in Exercise 9.37.
Exercise 9.42. Here $Q = X/\beta \sim$ gamma$(k,1)$, with $f(t) = t^{k-1}e^{-t}/\Gamma(k)$ on $(0,\infty)$, and again $g(t) = t^{-2}$. By part (a),

$$C = \{t : t^{-2} < \lambda t^{k-1}e^{-t}/\Gamma(k)\} = \{t : t^{k+1}e^{-t} > \Gamma(k)/\lambda\}.$$

Since $t \mapsto t^{k+1}e^{-t}$ increases on $(0, k+1)$ and decreases thereafter, C is an interval (a,b) whose endpoints satisfy

$$a^{k+1}e^{-a} = b^{k+1}e^{-b} \iff a^2 f(a) = b^2 f(b), \qquad \int_a^b f(t)\,dt = 1-\alpha.$$

These are exactly the two conditions of Exercise 9.42(a), and the shortest interval is $\{\beta : x/b \le \beta \le x/a\}$ for that pair (a,b).

Problem 9.44. (a) Let $X_1, \dots, X_n$ be iid Poisson(λ). Find a UMA $1-\alpha$ confidence interval based on inverting the UMP level α test of $H_0\colon \lambda = \lambda_0$ versus $H_1\colon \lambda > \lambda_0$.
(b) Let $f(x|\theta)$ be the logistic$(\theta, 1)$ location pdf. Based on one observation, x, find the UMA one-sided $1-\alpha$ confidence interval of the form $\{\theta : \theta \le U(x)\}$.

Solution. (a) $C^*(\mathbf{x}) = \big\{\lambda : \lambda > \chi^2_{2t,\,1-\alpha}/(2n)\big\}$, where $t = \sum x_i$ and, in the book's upper-tail notation, $P(\chi^2_{2t} > \chi^2_{2t,\,1-\alpha}) = 1-\alpha$.
The Poisson family has MLR in $T = \sum X_i \sim$ Poisson$(n\lambda)$, so by the Karlin-Rubin Theorem 8.3.17

the UMP level α test rejects for large T: reject iff $T \geq k(\lambda_0)$, where $k(\lambda_0)$ is the smallest integer with $P_{\lambda_0}(T \geq k) \leq \alpha$. Since $P_\lambda(T \geq k)$ decreases in k, the acceptance region is

$$A(\lambda_0) = \{\mathbf{x} : P_{\lambda_0}(T \geq t) > \alpha\}, \qquad t = \sum x_i,$$

and inverting gives $C^*(\mathbf{x}) = \{\lambda : P_\lambda(T \geq t) > \alpha\}$. The Poisson-gamma identity

$$P_\lambda(T \geq t) = \sum_{j=t}^{\infty} \frac{e^{-n\lambda}(n\lambda)^j}{j!} = P\big(\chi^2_{2t} \leq 2n\lambda\big)$$

shows the coverage function is strictly increasing in λ, so

$$C^*(\mathbf{x}) = \big\{\lambda : 2n\lambda > \chi^2_{2t,\,1-\alpha}\big\} = \left(\frac{\chi^2_{2t,\,1-\alpha}}{2n},\ \infty\right)$$

(with $\chi^2_{0,1-\alpha} = 0$, so $t = 0$ gives the whole parameter space). By Theorem 9.3.5 this lower confidence bound is UMA: it is obtained by inverting the UMP acceptance regions for $H_1\colon \lambda > \lambda_0$.
(b) $C^*(x) = \big\{\theta : \theta \leq x + \log\frac{1-\alpha}{\alpha}\big\}$.
The logistic$(\theta, 1)$ pdf and cdf are

$$f(x|\theta) = \frac{e^{-(x-\theta)}}{\big(1 + e^{-(x-\theta)}\big)^2}, \qquad F(x|\theta) = \frac{1}{1 + e^{-(x-\theta)}}.$$

For $\theta_2 > \theta_1$, writing $u = e^{-x}$,

$$\frac{f(x|\theta_2)}{f(x|\theta_1)} = e^{\theta_2 - \theta_1}\left(\frac{1 + e^{\theta_1}u}{1 + e^{\theta_2}u}\right)^2,$$

which decreases in u and hence increases in x: the family has MLR. An upper confidence bound comes from inverting tests against $H_1\colon \theta < \theta_0$, for which Theorem 8.3.17 gives the UMP level α test rejecting for small x, say for $x < k(\theta_0)$ with $F(k(\theta_0)|\theta_0) = \alpha$. Solving,

$$\frac{1}{1 + e^{-(k-\theta_0)}} = \alpha \implies k(\theta_0) = \theta_0 + \log\frac{\alpha}{1-\alpha}.$$

Hence $\mathbf{x} \in A(\theta_0)$ iff $x \geq \theta_0 + \log\frac{\alpha}{1-\alpha}$, i.e. iff $\theta_0 \leq x + \log\frac{1-\alpha}{\alpha}$, and inverting gives

$$C^*(x) = \left(-\infty,\ x + \log\frac{1-\alpha}{\alpha}\right],$$

which is UMA by the mirror image of Theorem 9.3.5 (it minimizes $P_\theta(\theta' \in C(X))$ over all $\theta' > \theta$).

Problem 9.45. Let $X_1, \ldots, X_n$ be iid exponential(λ).
(a) Find a UMP size α hypothesis test of $H_0\colon \lambda = \lambda_0$ versus $H_1\colon \lambda < \lambda_0$.
(b) Find a UMA $1-\alpha$ confidence interval based on inverting the test in part (a). Show that the interval can be expressed as

$$C^*(x_1, \ldots, x_n) = \left\{\lambda : 0 \leq \lambda \leq \frac{2\sum x_i}{\chi^2_{2n,\alpha}}\right\}.$$

(c) Find the expected length of $C^*(x_1, \ldots, x_n)$.
(d) Madansky (1962) exhibited a $1-\alpha$ interval whose expected length is shorter than that of the UMA interval. In general, Madansky's interval is difficult to calculate, but in the following situation calculation is relatively simple. Let $1 - \alpha = .3$ and $n = 120$. Madansky's interval is

$$C^{\mathrm{M}}(x_1, \ldots, x_n) = \left\{\lambda : 0 \leq \lambda \leq -\frac{x_{(1)}}{\log(.99)}\right\},$$

which is a 30% confidence interval. Use the fact that $\chi^2_{240,.7} = 251.046$ to show that the 30% UMA interval satisfies

$$\mathrm{E}\left[\text{Length}\left(C^*(x_1,\ldots,x_n)\right)\right] = .956\lambda > \mathrm{E}\left[\text{Length}\left(C^{\mathrm{M}}(x_1,\ldots,x_n)\right)\right] = .829\lambda.$$

Solution. (a) Reject H_0 if $\sum x_i < \lambda_0\chi^2_{2n,\alpha}/2$. (Here exponential($\lambda$) has mean λ, and the book's own value $\chi^2_{240,.7} = 251.046$ in part (d) forces $\chi^2_{2n,\alpha}$ to mean the lower-tail point, $P(\chi^2_{2n} < \chi^2_{2n,\alpha}) = \alpha$, contrary to the convention of Table 1 in the appendix.)
This is an exponential family with natural parameter $-1/\lambda$, increasing in λ, hence MLR in $T = \sum X_i$; by the Karlin-Rubin Theorem 8.3.17 the UMP level α test against $H_1\colon \lambda < \lambda_0$ rejects for small T. Since $2T/\lambda \sim \chi^2_{2n}$, the cutoff giving exact size α is

$$P_{\lambda_0}\left(\frac{2T}{\lambda_0} < \chi^2_{2n,\alpha}\right) = \alpha.$$

(b) The acceptance region is $A(\lambda_0) = \{\mathbf{x} : 2\sum x_i \geq \lambda_0\chi^2_{2n,\alpha}\}$, and $2\sum x_i \geq \lambda_0\chi^2_{2n,\alpha} \iff \lambda_0 \leq 2\sum x_i/\chi^2_{2n,\alpha}$, so inverting gives

$$C^*(x_1,\ldots,x_n) = \left\{\lambda : 0 \leq \lambda \leq \frac{2\sum x_i}{\chi^2_{2n,\alpha}}\right\},$$

UMA by Theorem 9.3.5 (in its upper-bound form, since the tests inverted are UMP against $H_1\colon \lambda < \lambda_0$).
(c) $\mathrm{E}_\lambda X_i = \lambda$, so

$$\mathrm{E}_\lambda\left[\text{Length}(C^*)\right] = \frac{2\,\mathrm{E}_\lambda[\sum X_i]}{\chi^2_{2n,\alpha}} = \frac{2n\lambda}{\chi^2_{2n,\alpha}}.$$

(d) With $n = 120$ and $\chi^2_{240,.7} = 251.046$, part (c) gives

$$\mathrm{E}_\lambda\left[\text{Length}(C^*)\right] = \frac{240\lambda}{251.046} = .9560\lambda.$$

For Madansky's interval, $X_{(1)} \sim$ exponential(λ/n), so with $-\log(.99) > 0$,

$$\begin{aligned} P_\lambda\left(\lambda \leq -\frac{X_{(1)}}{\log(.99)}\right) &= P_\lambda\big(X_{(1)} \geq -\lambda\log(.99)\big) \\ &= e^{\,n\log(.99)} = (.99)^{120} = .2994, \end{aligned}$$

confirming the 30% coverage (exactly $.99^n$, free of λ). Its expected length is

$$\mathrm{E}_\lambda\left[\text{Length}(C^{\mathrm{M}})\right] = \frac{\mathrm{E}_\lambda X_{(1)}}{-\log(.99)} = \frac{\lambda/120}{.0100503} = .8292\lambda.$$

Hence $.956\lambda > .829\lambda$: the UMA interval is the longer of the two.

Problem 9.46. Show that if $A(\theta_0)$ is an unbiased level α acceptance region of a test of $H_0\colon \theta = \theta_0$ versus $H_1\colon \theta \neq \theta_0$ and $C(\mathbf{x})$ is the $1-\alpha$ confidence set formed by inverting the acceptance regions, then $C(\mathbf{x})$ is an unbiased $1-\alpha$ confidence set.

Solution. By Theorem 9.2.2 the inverted set satisfies $\theta' \in C(\mathbf{x}) \iff \mathbf{x} \in A(\theta')$, so every probability statement about the tests transfers verbatim. Unbiasedness of the test with null value θ' says its power at each $\theta \neq \theta'$ is at least α, i.e. $P_\theta(\mathbf{X} \in A(\theta')) \leq 1-\alpha$; hence for all $\theta \neq \theta'$

$$P_\theta\big(\theta' \in C(\mathbf{X})\big) = P_\theta\big(\mathbf{X} \in A(\theta')\big) \leq 1-\alpha,$$

which is Definition 9.3.7. Coverage is inherited the same way: $P_\theta(\theta \in C(\mathbf{X})) = P_\theta(\mathbf{X} \in A(\theta)) \geq 1-\alpha$ because each test has level α.

Problem 9.47. Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population, where σ^2 is known. Show that the usual one-sided $1-\alpha$ upper confidence bound $\{\theta : \theta \le \bar{x} + z_\alpha \sigma/\sqrt{n}\}$ is unbiased, and so is the corresponding lower confidence bound.

Solution. Write $Z = \sqrt{n}(\bar{X} - \theta)/\sigma \sim \mathrm{n}(0,1)$ and $\delta = \sqrt{n}(\theta' - \theta)/\sigma$. For a one-sided bound Definition 9.3.7 is imposed only on the side the bound is built to exclude – an upper bound $(-\infty, U(\bar{X})]$ covers every $\theta' < \theta$ whenever it covers θ – which is also the alternative region of the one-sided test being inverted (Theorem 9.3.5).
Upper bound, $\theta' > \theta$, so $\delta > 0$:

$$\begin{aligned} P_\theta\left(\theta' \le \bar{X} + z_\alpha \tfrac{\sigma}{\sqrt{n}}\right) &= P(Z \ge \delta - z_\alpha) \\ &= \Phi(z_\alpha - \delta) \;<\; \Phi(z_\alpha) = 1 - \alpha. \end{aligned}$$

Lower bound $C(\bar{x}) = \{\theta : \theta \ge \bar{x} - z_\alpha \sigma/\sqrt{n}\}$, $\theta' < \theta$, so $\delta < 0$:

$$\begin{aligned} P_\theta\left(\theta' \ge \bar{X} - z_\alpha \tfrac{\sigma}{\sqrt{n}}\right) &= P(Z \le \delta + z_\alpha) \\ &= \Phi(z_\alpha + \delta) \;<\; \Phi(z_\alpha) = 1 - \alpha. \end{aligned}$$

Each true-coverage probability is the case $\delta = 0$, namely $\Phi(z_\alpha) = 1 - \alpha$.

Problem 9.48. Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\theta, \sigma^2)$ population, where σ^2 is unknown.
(a) Show that the interval $\theta \le \bar{x} + t_{n-1,\alpha}\frac{s}{\sqrt{n}}$ can be derived by inverting the acceptance region of an LRT.
(b) Show that the corresponding two-sided interval in (9.2.14), namely $\left\{\theta : \bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \le \theta \le \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right\}$, can also be derived by inverting the acceptance region of an LRT.
(c) Show that the intervals in parts (a) and (b) are unbiased intervals.

Solution. Write $(n-1)s^2 = \sum(x_i - \bar{x})^2$ and $t = (\bar{x} - \theta_0)/(s/\sqrt{n})$. Both parts rest on the single algebraic identity

$$\sum_i (x_i - \theta_0)^2 = (n-1)s^2 + n(\bar{x} - \theta_0)^2 = (n-1)s^2\left(1 + \frac{t^2}{n-1}\right).$$

(a) Take the LRT of $H_0\colon \theta \ge \theta_0$ versus $H_1\colon \theta < \theta_0$. The unrestricted MLEs are $\hat{\theta} = \bar{x}$, $\hat{\sigma}^2 = \sum(x_i - \bar{x})^2/n$; under H_0 they are the same when $\bar{x} \ge \theta_0$ (so $\lambda = 1$) and are θ_0, $\hat{\sigma}_0^2 = \sum(x_i - \theta_0)^2/n$ when $\bar{x} < \theta_0$. Since the σ^2-profiled likelihood is $(2\pi\hat{\sigma}^2)^{-n/2}e^{-n/2}$,

$$\lambda(\mathbf{x}) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} = \left(1 + \frac{t^2}{n-1}\right)^{-n/2} \qquad (\bar{x} < \theta_0).$$

So $\lambda(\mathbf{x}) < c$ iff $\bar{x} < \theta_0$ and $t^2 > k$, i.e. iff $t < -\sqrt{k}$; the size is α when $\sqrt{k} = t_{n-1,\alpha}$, because $t \sim t_{n-1}$ when $\theta = \theta_0$. The acceptance region is therefore

$$A(\theta_0) = \left\{\mathbf{x} : \frac{\bar{x} - \theta_0}{s/\sqrt{n}} \ge -t_{n-1,\alpha}\right\} = \left\{\mathbf{x} : \theta_0 \le \bar{x} + t_{n-1,\alpha}\frac{s}{\sqrt{n}}\right\},$$

and inverting gives $C(\mathbf{x}) = \{\theta : \theta \le \bar{x} + t_{n-1,\alpha}s/\sqrt{n}\}$.
(b) For $H_0\colon \theta = \theta_0$ versus $H_1\colon \theta \ne \theta_0$ the restricted maximization is over σ^2 only for every $\bar{x}$, so the same computation gives $\lambda(\mathbf{x}) = (1 + t^2/(n-1))^{-n/2}$ for all $\mathbf{x}$. Now $\lambda < c \iff |t| > t_{n-1,\alpha/2}$ for size α, so

$$A(\theta_0) = \left\{\mathbf{x} : |\bar{x} - \theta_0| \le t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right\},$$

which inverts to precisely the interval (9.2.14).
(c) The two-sided case is immediate: the two-sided t test of (b) is unbiased (Exercise 8.38), so by Exercise 9.46 the interval (9.2.14) is unbiased – this is the assertion of Example 9.3.8.
For the one-sided bound of (a), fix $\theta' > \theta$ and set

$$Z = \frac{\sqrt{n}(\bar{X} - \theta)}{\sigma}, \quad V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \delta = \frac{\sqrt{n}(\theta - \theta')}{\sigma} < 0,$$

with $Z \sim \mathrm{n}(0,1)$ independent of V (Theorem 5.3.1). Then $\theta' \in C(\mathbf{X})$ iff $(\bar{X} - \theta')/(S/\sqrt{n}) \geq -t_{n-1,\alpha}$, and that statistic equals $(Z + \delta)/\sqrt{V/(n-1)}$. Hence, conditioning on V,

$$\begin{aligned} P_\theta\big(\theta' \in C(\mathbf{X})\big) &= \mathrm{E}_V\left[P\left(Z \geq -t_{n-1,\alpha}\sqrt{\tfrac{V}{n-1}} - \delta \,\middle|\, V\right)\right] \\ &= \mathrm{E}_V\left[\Phi\left(t_{n-1,\alpha}\sqrt{V/(n-1)} + \delta\right)\right], \end{aligned}$$

which is increasing in δ. Since $\delta < 0$, it is strictly less than its value at $\delta = 0$, which is $P(T_{n-1} \geq -t_{n-1,\alpha}) = 1 - \alpha$, the true-coverage probability. So the one-sided interval is unbiased in the sense appropriate to an upper bound (Exercise 9.47).

Problem 9.49. (Cox's Paradox) We are to test

$$H_0 \colon \theta = \theta_0 \qquad \text{versus} \qquad H_1 \colon \theta > \theta_0,$$

where θ is the mean of one of two normal distributions and θ_0 is a fixed but arbitrary value of θ. We observe the random variable X with distribution

$$X \sim \begin{cases} \mathrm{n}(\theta, 100) & \text{with probability } p \\ \mathrm{n}(\theta, 1) & \text{with probability } 1 - p. \end{cases}$$

(a) Show that the test given by

$$\text{reject } H_0 \text{ if } X > \theta_0 + z_\alpha \sigma,$$

where $\sigma = 1$ or 10 depending on which population is sampled, is a level α test. Derive a $1 - \alpha$ confidence set by inverting the acceptance region of this test.
(b) Show that a more powerful level α test (for $\alpha > p$) is given by

$$\text{reject } H_0 \text{ if } X > \theta_0 + z_{(\alpha-p)/(1-p)} \text{ and } \sigma = 1; \text{ otherwise always reject } H_0.$$

Derive a $1 - \alpha$ confidence set by inverting the acceptance region of this test, and show that it is the empty set with positive probability. (Cox's Paradox states that classic optimal procedures sometimes ignore the information about conditional distributions and provide us with a procedure that, while optimal, is somehow unreasonable; see Cox 1958 or Cornfield 1969.)

Solution. (a) The value $\sigma \in \{1, 10\}$ is observed along with X, and given σ we have $X \sim \mathrm{n}(\theta, \sigma^2)$, so conditioning on which population was sampled,

$$\begin{aligned} P_{\theta_0}(\text{reject}) &= p\, P(X > \theta_0 + 10 z_\alpha \mid \sigma = 10) \\ &\quad + (1-p)\, P(X > \theta_0 + z_\alpha \mid \sigma = 1) \\ &= p\,\alpha + (1-p)\,\alpha = \alpha, \end{aligned}$$

since each conditional probability is $P(Z > z_\alpha) = \alpha$. The acceptance region is $A(\theta_0) = \{x : x \leq \theta_0 + z_\alpha\sigma\} = \{x : \theta_0 \geq x - z_\alpha\sigma\}$, so inverting gives the lower bound

$$C(x) = [\, x - z_\alpha\sigma,\ \infty \,),$$

which is the usual $1 - \alpha$ interval computed with whichever σ was in force.

(b) Write $c = z_{(\alpha-p)/(1-p)}$, which is well defined because $0 < (\alpha-p)/(1-p) < 1$ when $p < \alpha < 1$. The size is

$$P_{\theta_0}(\text{reject}) = p \cdot 1 + (1-p) \cdot \frac{\alpha - p}{1-p} = p + (\alpha - p) = \alpha,$$

so the test has level α. Writing $d = \theta - \theta_0$, its power and that of (a) are

$$\begin{aligned} \beta_b(d) &= p + (1-p)\,[1 - \Phi(c-d)]\,, \\ \beta_a(d) &= p\,\big[1 - \Phi(z_\alpha - \tfrac{d}{10})\big] + (1-p)\,[1 - \Phi(z_\alpha - d)]\,, \end{aligned}$$

so that

$$\beta_b(d) - \beta_a(d) = p\,\Phi\big(z_\alpha - \tfrac{d}{10}\big) - (1-p)\,[\Phi(c-d) - \Phi(z_\alpha - d)]\,.$$

For $d > c\ (> z_\alpha > 0)$ both arguments of the bracket are negative, where φ increases, so by the mean value theorem the bracket is at most $(c - z_\alpha)\varphi(c-d)$, of order $e^{-(d-c)^2/2}$, while the first term is of order $e^{-(d/10 - z_\alpha)^2/2}$ and so dominates for all large d. (The book overstates: the domination is not uniform, since $p = .01$, $\alpha = .05$, $d = 1.7$ give $\beta_a = .518 > \beta_b = .487$, so the claim holds only against distant alternatives – all the paradox needs.)
Inverting is where the paradox appears. If $\sigma = 10$ the test rejects $H_0\colon \theta = \theta_0$ for every θ_0, so $A(\theta_0) = \emptyset$ for all θ_0 and

$$C(x) = \{\theta_0 : x \in A(\theta_0)\} = \emptyset.$$

If $\sigma = 1$, acceptance means $x \le \theta_0 + c$, i.e. $\theta_0 \ge x - c$, so $C(x) = [\,x - c,\ \infty)$. Thus

$$C(X) = \begin{cases} \emptyset & \text{with probability } p, \\ [\,X - z_{(\alpha-p)/(1-p)},\ \infty) & \text{with probability } 1-p, \end{cases}$$

and the empty set occurs with probability $p > 0$. Coverage is nonetheless correct:

$$P_\theta(\theta \in C(X)) = (1-p)\,P(X \le \theta + c \mid \sigma = 1) = (1-p) - (\alpha - p) = 1 - \alpha.$$

9.8 Exercises 9.50–9.56

Problem 9.50. Let $X \sim f(x|\theta)$, and suppose that the interval $\{\theta :\ a(X) \le \theta \le b(X)\}$ is a UMA confidence set for θ.
(a) Find a UMA confidence set for $1/\theta$. Note that if $a(x) < 0 < b(x)$, this set is

$$\{1/\theta :\ 1/b(x) \le 1/\theta\} \cup \{1/\theta :\ 1/\theta \le 1/a(x)\}.$$

Hence it is possible for the UMA confidence set to be neither an interval nor bounded.
(b) Show that, if h is a strictly increasing function, the set $\{h(\theta) :\ h(a(X)) \le h(\theta) \le h(b(X))\}$ is a UMA confidence set for $h(\theta)$. Can the condition on h be relaxed?

Solution. Take the image set $C_h(x) = h(C(x))$, where $C(x) = [a(x), b(x)]$; for (a), $h(\theta) = 1/\theta$.
A one-to-one h transports the false-coverage events (9.3.2) unchanged: for every θ',

$$h(\theta') \in h(C(x)) \iff \theta' \in C(x),$$

so $P_\theta\big(h(\theta') \in C_h(X)\big) = P_\theta(\theta' \in C(X))$ for all θ, θ'. Taking $\theta' = \theta$ makes C_h a $1-\alpha$ confidence set for $h(\theta)$. For any competing $1-\alpha$ set $D(X)$ for $h(\theta)$, the set $h^{-1}(D(X))$ is a $1-\alpha$ set for θ, and

$$\begin{aligned} P_\theta\big(h(\theta') \in C_h(X)\big) &= P_\theta\big(\theta' \in C(X)\big) \\ &\le P_\theta\big(\theta' \in h^{-1}(D(X))\big) \\ &= P_\theta\big(h(\theta') \in D(X)\big), \end{aligned}$$

the middle step being the UMA property of C (p. 445). Hence C_h minimizes the probability of false coverage over the transformed class.

(a) $h(\theta) = 1/\theta$ is one-to-one on $\theta \neq 0$, so the UMA set for $1/\theta$ is

$$C_h(x) = \{1/\theta : \ a(x) \le \theta \le b(x), \ \theta \neq 0\}.$$

If $0 < a(x)$ this is the interval $[1/b(x), 1/a(x)]$. If $a(x) < 0 < b(x)$, then $\theta \in (0, b(x)]$ contributes $1/\theta \ge 1/b(x)$ and $\theta \in [a(x), 0)$ contributes $1/\theta \le 1/a(x)$, giving the displayed union, which is neither an interval nor bounded.
(b) For h strictly increasing, $a(x) \le \theta \le b(x) \iff h(a(x)) \le h(\theta) \le h(b(x))$, so the displayed set is exactly $h(C(x))$ and the argument above applies. The condition can be relaxed to h one-to-one, since only invertibility was used; what changes is the shape, not the optimality. For h strictly decreasing $h(C(x)) = [h(b(x)), h(a(x))]$, and for h one-to-one but not monotone, as in (a), $h(C(x))$ need not be an interval. It cannot be relaxed further: a non-injective h puts values $h(\theta')$ with $\theta' \notin C(X)$ into the image set, inflating false coverage.

Problem 9.51. If $X_1, \ldots, X_n$ are iid from a location pdf $f(x - \theta)$, show that the confidence set

$$C(x_1, \ldots, x_n) = \{\theta : \ \bar{x} - k_1 \le \theta \le \bar{x} + k_2\},$$

where k_1 and k_2 are constants, has constant coverage probability.
(Hint: The pdf of $\bar{X}$ is of the form $f_{\bar{X}}(\bar{x} - \theta)$.)

Solution. The coverage probability is $P_0(-k_2 \le \bar{Z} \le k_1)$, free of θ, where $\bar{Z}$ is the sample mean of n iid draws from f.
Write $X_i = Z_i + \theta$ with $Z_1, \ldots, Z_n$ iid with pdf f, which is what a location pdf $f(x - \theta)$ means. Then $\bar{X} = \bar{Z} + \theta$, so $\bar{X}$ has pdf $f_{\bar{Z}}(\bar{x} - \theta)$ (the hint) and $\bar{X} - \theta = \bar{Z}$ has a distribution free of θ. Hence

$$\begin{aligned} P_\theta\big(\theta \in C(X)\big) &= P_\theta\big(\bar{X} - k_1 \le \theta \le \bar{X} + k_2\big) \\ &= P_\theta\big(- k_2 \le \bar{X} - \theta \le k_1\big) \\ &= P\big(- k_2 \le \bar{Z} \le k_1\big) = \int_{-k_2}^{k_1} f_{\bar{Z}}(t)\, dt, \end{aligned}$$

which is the same number for every θ.

Problem 9.52. Let $X_1, \ldots, X_n$ be a random sample from a $n(\mu, \sigma^2)$ population, where both μ and σ^2 are unknown. Each of the following methods of finding confidence intervals for σ^2 results in intervals of the form

$$\left\{\sigma^2 : \ \frac{(n-1)s^2}{b} \le \sigma^2 \le \frac{(n-1)s^2}{a}\right\},$$

but in each case a and b will satisfy different constraints. The intervals given in this exercise are derived by Tate and Klett (1959), who also tabulate some cutoff points.
Define $f_p(t)$ to be the pdf of a χ^2_p random variable with p degrees of freedom. In order to have a $1 - \alpha$ confidence interval, a and b must satisfy

$$\int_a^b f_{n-1}(t)\, dt = 1 - \alpha,$$

but additional constraints are required to define a and b uniquely. Verify that each of the following constraints can be derived as stated.
(a) *The likelihood ratio interval:* The $1 - \alpha$ confidence interval obtained by inverting the LRT of $H_0 : \sigma = \sigma_0$ versus $H_1 : \sigma \neq \sigma_0$ is of the above form, where a and b also satisfy $f_{n+2}(a) = f_{n+2}(b)$.
(b) *The minimum length interval:* For intervals of the above form, the $1 - \alpha$ confidence interval obtained by minimizing the interval length constrains a and b to satisfy $f_{n+3}(a) = f_{n+3}(b)$.
(c) *The shortest unbiased interval:* For intervals of the above form, the $1 - \alpha$ confidence interval obtained by minimizing the probability of false coverage among all unbiased intervals constrains a

and b to satisfy $f_{n+1}(a) = f_{n+1}(b)$. This interval can also be derived by minimizing the ratio of the endpoints.
(d) *The equal-tail interval:* For intervals of the above form, the $1-\alpha$ confidence interval obtained by requiring that the probability above and below the interval be equal constrains a and b to satisfy

$$\int_0^a f_{n-1}(t)\,dt = \frac{\alpha}{2}, \qquad \int_b^\infty f_{n-1}(t)\,dt = \frac{\alpha}{2}.$$

(This interval, although very common, is clearly nonoptimal no matter what length criterion is used.)
(e) For $\alpha = .1$ and $n = 3$, find the numerical values of a and b for each of the above cases. Compare the length of these intervals.

Solution. Everything follows from the one identity

$$t^k f_p(t) \;=\; \big[p(p+2)\cdots(p+2k-2)\big]\, f_{p+2k}(t),$$

immediate from $f_p(t) = t^{p/2-1}e^{-t/2}/\{\Gamma(p/2)2^{p/2}\}$. Throughout, $T = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ (Theorem 5.3.1), so the acceptance event $a \le T \le b$ inverts to the displayed interval and coverage is $\int_a^b f_{n-1}(t)\,dt$.
(a) With μ a nuisance parameter the MLEs are $\hat\mu = \bar x$ and $\hat\sigma^2 = (n-1)s^2/n$, so with $t = (n-1)s^2/\sigma_0^2$,

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{(2\pi\sigma_0^2)^{-n/2}e^{-(n-1)s^2/(2\sigma_0^2)}}{(2\pi\hat\sigma^2)^{-n/2}e^{-n/2}} \\ &= \Big(\frac{t}{n}\Big)^{n/2} e^{-t/2+n/2} \;\propto\; t^{n/2}e^{-t/2}.\end{aligned}$$

Since $t^{n/2}e^{-t/2} \propto f_{n+2}(t)$, the LRT rejects for small $f_{n+2}(t)$, so the acceptance region is $\{t : f_{n+2}(t) \ge k\} = [a,b]$, an interval because f_{n+2} is unimodal. Its endpoints are the two solutions of $f_{n+2}(a) = f_{n+2}(b)\,(= k)$, together with $\int_a^b f_{n-1} = 1-\alpha$.
(b) The length is $(n-1)s^2\,(1/a - 1/b)$, so minimize $1/a - 1/b$ subject to $\int_a^{b(a)} f_{n-1}(t)\,dt = 1-\alpha$. Differentiating the constraint gives $b'(a) = f_{n-1}(a)/f_{n-1}(b)$, and setting the derivative of the objective to zero gives

$$-\frac{1}{a^2} + \frac{b'(a)}{b^2} = 0 \iff b'(a) = \frac{b^2}{a^2}.$$

Equating the two expressions for $b'(a)$ yields $a^2 f_{n-1}(a) = b^2 f_{n-1}(b)$, and $t^2 f_{n-1}(t) \propto f_{n+3}(t)$ gives $f_{n+3}(a) = f_{n+3}(b)$. (This is Example 9.3.4 and Exercise 9.42, p. 444, with $f_Y = f_{n-1}$.)
(c) For true variance σ^2 and a false value τ^2, put $r = \tau^2/\sigma^2$; then

$$P_{\sigma^2}\big(\tau^2 \in C(\mathbf{X})\big) = P(ar \le T \le br) =: g(r),$$

with $g(1) = 1-\alpha$. By Definition 9.3.7 the interval is unbiased iff $g(r) \le 1-\alpha$ for all $r > 0$, i.e. iff $r = 1$ maximizes g; since g is differentiable this forces $g'(1) = 0$, and

$$g'(r) = b f_{n-1}(br) - a f_{n-1}(ar) \implies a f_{n-1}(a) = b f_{n-1}(b).$$

As $t f_{n-1}(t) \propto f_{n+1}(t)$, this is $f_{n+1}(a) = f_{n+1}(b)$. Together with the coverage equation it pins down (a,b), so this is the unique unbiased member of the class and hence trivially the false-coverage minimizer among them.
For the ratio characterization, the endpoint ratio is $\{(n-1)s^2/a\}/\{(n-1)s^2/b\} = b/a$, and

$$\frac{d}{da}\Big(\frac{b(a)}{a}\Big) = \frac{a\,b'(a) - b}{a^2} = 0 \iff b'(a) = \frac{b}{a},$$

which with $b'(a) = f_{n-1}(a)/f_{n-1}(b)$ is again $a f_{n-1}(a) = b f_{n-1}(b)$.
(d) The interval misses σ^2 on the high side when $\sigma^2 > (n-1)S^2/a$, i.e. $T < a$, and on the low side when $\sigma^2 < (n-1)S^2/b$, i.e. $T > b$. Equating these two error probabilities to $\alpha/2$ is exactly the stated pair of equations. (Check!)

(e) Here $n-1=2$, so $f_2(t)=\frac{1}{2}e^{-t/2}$ and the coverage constraint is $e^{-a/2}-e^{-b/2}=.9$; the side conditions become

$$\text{(a) } f_5: \quad a^{3/2}e^{-a/2}=b^{3/2}e^{-b/2},$$
$$\text{(b) } f_6: \quad a^{2}e^{-a/2}=b^{2}e^{-b/2},$$
$$\text{(c) } f_4: \quad a\,e^{-a/2}=b\,e^{-b/2},$$

while (d) has the closed form $a=-2\log(.95)=.1026$, $b=2\log 20=5.9915$. Solving numerically, with length $=2s^2(1/a-1/b)$:

interval	a	b	$1/a-1/b$	length
(a) LRT	.2065	12.5214	4.7632	$9.526\,s^2$
(b) min length	.2104	18.0077	4.6962	$9.392\,s^2$
(c) unbiased	.1676	7.8643	5.8384	$11.677\,s^2$
(d) equal-tail	.1026	5.9915	9.5810	$19.162\,s^2$

The minimum-length interval is shortest, as it must be; the LRT interval is within 1.4% of it; the unbiased interval pays about 24% extra length for unbiasedness; and the equal-tail interval is more than twice the optimal length.

Problem 9.53. Let $X\sim n(\mu,\sigma^2)$, σ^2 known. For each $c\ge 0$, define an interval estimator for μ by $C(x)=[x-c\sigma,\ x+c\sigma]$ and consider the loss in (9.3.4), namely $L(\theta,C)=b\,\text{Length}(C)-I_C(\theta)$.
(a) Show that the risk function, $R(\mu,C)$, is given by

$$R(\mu,C)=b(2c\sigma)-P(-c\le Z\le c).$$

(b) Using the Fundamental Theorem of Calculus, show that

$$\frac{d}{dc}R(\mu,C)=2b\sigma-\frac{2}{\sqrt{2\pi}}e^{-c^2/2},$$

and, hence, the derivative is an increasing function of c for $c\ge 0$.
(c) Show that if $b\sigma>1/\sqrt{2\pi}$, the derivative is positive for all $c\ge 0$ and, hence, $R(\mu,C)$ is minimized at $c=0$. That is, the best interval estimator is the point estimator $C(x)=[x,x]$.
(d) Show that if $b\sigma\le 1/\sqrt{2\pi}$, the c that minimizes the risk is $c=\sqrt{-2\log(b\sigma\sqrt{2\pi})}$. Hence, if b is chosen so that $c=z_{\alpha/2}$ for some α, then the interval estimator that minimizes the risk is just the usual $1-\alpha$ confidence interval.

Solution. (a) $\text{Length}(C(X))=2c\sigma$ is constant, and with $Z=(X-\mu)/\sigma\sim n(0,1)$,

$$P_\mu(\mu\in C(X))=P_\mu(|X-\mu|\le c\sigma)=P(-c\le Z\le c),$$

so $R(\mu,C)=E_\mu L(\mu,C(X))=b(2c\sigma)-P(-c\le Z\le c)$, which is free of μ: this is (9.3.5) of Example 9.3.13.
(b) Writing $P(-c\le Z\le c)=\int_{-c}^{c}\phi(t)\,dt$ with $\phi(t)=(2\pi)^{-1/2}e^{-t^2/2}$, the Fundamental Theorem of Calculus gives

$$\frac{d}{dc}\int_{-c}^{c}\phi(t)\,dt=\phi(c)+\phi(-c)=2\phi(c),$$

hence

$$\frac{d}{dc}R(\mu,C)=2b\sigma-2\phi(c)=2b\sigma-\frac{2}{\sqrt{2\pi}}e^{-c^2/2}.$$

For $c\ge 0$, $e^{-c^2/2}$ is decreasing, so $-2\phi(c)$ — and therefore the derivative — is increasing in c.
(c) At $c=0$ the derivative equals $2b\sigma-2/\sqrt{2\pi}>0$ when $b\sigma>1/\sqrt{2\pi}$; being increasing by (b), it stays positive for all $c\ge 0$. Thus $R(\mu,C)$ is strictly increasing on $[0,\infty)$ and is minimized at $c=0$, i.e. $C(x)=[x,x]$.
(d) If $b\sigma\le 1/\sqrt{2\pi}$ the derivative is ≤ 0 at $c=0$ and tends to $2b\sigma>0$ as $c\to\infty$, so by the monotonicity

in (b) it has a unique zero, which is the minimizer:

$$2b\sigma = \frac{2}{\sqrt{2\pi}} e^{-c^2/2} \iff e^{-c^2/2} = b\sigma\sqrt{2\pi}$$
$$\iff c = \sqrt{-2\log\big(b\sigma\sqrt{2\pi}\big)},$$

the logarithm being ≤ 0 because $b\sigma\sqrt{2\pi} \le 1$. Inverting this relation, the choice

$$b = \frac{1}{\sigma\sqrt{2\pi}} e^{-z_{\alpha/2}^2/2} = \frac{\phi(z_{\alpha/2})}{\sigma}$$

gives minimizing $c = z_{\alpha/2}$, so the optimal estimator is $C(x) = [x - z_{\alpha/2}\sigma,\ x + z_{\alpha/2}\sigma]$, the usual $1-\alpha$ interval.

Problem 9.54. Let $X \sim n(\mu, \sigma^2)$, but now consider σ^2 unknown. For each $c \ge 0$, define an interval estimator for μ by $C(x) = [x - cs,\ x + cs]$, where s^2 is an estimator of σ^2 independent of X, $\nu S^2/\sigma^2 \sim \chi^2_\nu$ (for example, the usual sample variance). Consider a modification of the loss in (9.3.4),

$$L\big((\mu,\sigma), C\big) = \frac{b}{\sigma}\,\text{Length}(C) - I_C(\mu).$$

(a) Show that the risk function, $R((\mu,\sigma), C)$, is given by

$$R\big((\mu,\sigma), C\big) = b(2cM) - \big[2P(T \le c) - 1\big],$$

where $T \sim t_\nu$ and $M = ES/\sigma$.
(b) If $b \le 1/\sqrt{2\pi}$, show that the c that minimizes the risk satisfies

$$b = \frac{1}{\sqrt{2\pi}} \left(\frac{\nu}{\nu + c^2}\right)^{(\nu+1)/2}.$$

(c) Reconcile this problem with the known σ^2 case. Show that as $\nu \to \infty$, the solution here converges to the solution in the known σ^2 problem. (Be careful of the rescaling done to the loss function.)

Solution. (a) $E\big[(b/\sigma)\,\text{Length}(C(X))\big] = (2bc)\,E(S/\sigma) = 2bcM$, and since $(X-\mu)/\sigma \sim n(0,1)$ is independent of $S/\sigma = \sqrt{W/\nu}$ with $W = \nu S^2/\sigma^2 \sim \chi^2_\nu$, Definition 5.3.4 gives $(X-\mu)/S \sim t_\nu$. Hence

$$P\big(\mu \in C(X)\big) = P\Big(\Big|\frac{X-\mu}{S}\Big| \le c\Big) = 2P(T \le c) - 1,$$

and $R((\mu,\sigma), C) = b(2cM) - [2P(T \le c) - 1]$, constant in (μ,σ).
(b) Differentiating, with f_ν the t_ν pdf,

$$\frac{d}{dc} R = 2bM - 2f_\nu(c),$$

which is increasing in $c \ge 0$ because f_ν is decreasing there; so the minimizing c solves $bM = f_\nu(c)$. Now

$$M = E\sqrt{W/\nu} = \sqrt{\tfrac{2}{\nu}}\,\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})},$$
$$f_\nu(c) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\Big(1 + \tfrac{c^2}{\nu}\Big)^{-\frac{\nu+1}{2}},$$

the first from $EW^{1/2} = \sqrt{2}\,\Gamma(\frac{\nu+1}{2})/\Gamma(\frac{\nu}{2})$. The gamma factors cancel in $bM = f_\nu(c)$, leaving

$$b\sqrt{\tfrac{2}{\nu}} = \frac{1}{\sqrt{\nu\pi}}\Big(1 + \tfrac{c^2}{\nu}\Big)^{-\frac{\nu+1}{2}},$$

that is,

$$b = \frac{1}{\sqrt{2\pi}} \left(\frac{\nu}{\nu + c^2} \right)^{\frac{\nu+1}{2}}.$$

The right side decreases from $1/\sqrt{2\pi}$ at $c = 0$ to 0, so a root $c \geq 0$ exists exactly when $b \leq 1/\sqrt{2\pi}$, the analogue of $b\sigma \leq 1/\sqrt{2\pi}$ in Exercise 9.53(d); it is a minimum because dR/dc crosses from negative to positive there.
(c) The b here plays the role of $b\sigma$ in Exercise 9.53, because dividing Length by σ replaces 9.53's coefficient b by b/σ. Indeed

$$\Big(\frac{\nu}{\nu + c^2} \Big)^{\frac{\nu+1}{2}} = \exp\left\{ -\frac{\nu+1}{2} \log\Big(1 + \frac{c^2}{\nu}\Big) \right\} \xrightarrow[\nu\to\infty]{} e^{-c^2/2},$$

so part (b) becomes $b = (2\pi)^{-1/2} e^{-c^2/2}$, i.e. $c = \sqrt{-2\log(b\sqrt{2\pi})}$, which is the answer of Exercise 9.53(d) under that identification. Consistently $M = E(S/\sigma) \to 1$ and $t_\nu \to \mathrm{n}(0,1)$, so the risk of (a) converges to $2bc - P(-c \leq Z \leq c)$, which is (9.3.5) with $b\sigma$ read as b.

Problem 9.55. The decision theoretic approach to set estimation can be quite useful (see Exercise 9.56) but it can also give some unsettling results, showing the need for thoughtful implementation. Consider again the case of $X \sim n(\mu, \sigma^2)$, σ^2 unknown, and suppose that we have an interval estimator for μ by $C(x) = [x - cs,\ x + cs]$, where s^2 is an estimator of σ^2 independent of X, $\nu S^2/\sigma^2 \sim \chi^2_\nu$. This is, of course, the usual t interval, one of the great statistical procedures that has stood the test of time. Consider the loss

$$L\big((\mu,\sigma), C\big) = b\,\mathrm{Length}(C) - I_C(\mu),$$

similar to that used in Exercise 9.54, but without scaling the length. Construct another procedure C' as

$$C' = \begin{cases} [x - cs,\ x + cs] & \text{if } s < K, \\ \emptyset & \text{if } s \geq K, \end{cases}$$

where K is a positive constant. Notice that C' does exactly the wrong thing. When s^2 is big and there is a lot of uncertainty, we would want the interval to be wide. But C' is empty! Show that we can find a value of K so that

$$R\big((\mu,\sigma), C'\big) \leq R\big((\mu,\sigma), C\big) \quad \text{for every } (\mu,\sigma)$$

with strict inequality for some (μ,σ).

Solution. Any $K \geq 1/(2bc)$ works, and in fact makes the inequality strict at every (μ,σ).
C and C' differ only on the event $\{S \geq K\}$, where C' is empty and so contributes neither length nor coverage. Hence

$$\begin{aligned} R\big((\mu,\sigma), C\big) - R\big((\mu,\sigma), C'\big) &= 2bc\,E\big[S\,I(S \geq K)\big] \\ &\quad - P\big(\mu \in C(X),\, S \geq K\big). \end{aligned}$$

Because X and S are independent, conditioning on S gives

$$P\big(\mu \in C(X), S \geq K\big) = E\Big[\Big(2\Phi\big(cS/\sigma\big) - 1\Big) I(S \geq K)\Big] < P(S \geq K),$$

the inequality being strict since $2\Phi(cS/\sigma) - 1 < 1$ with probability one and $P(S \geq K) > 0$ for every $\sigma > 0$ ($S = \sigma\sqrt{W/\nu}$ with $W \sim \chi^2_\nu$ supported on $(0,\infty)$). On the other hand

$$2bc\,E\big[S\,I(S \geq K)\big] \ \geq\ 2bcK\,P(S \geq K) \ \geq\ P(S \geq K)$$

as soon as $K \geq 1/(2bc)$. Combining the two displays,

$$R\big((\mu,\sigma), C\big) - R\big((\mu,\sigma), C'\big) > 0 \quad \text{for all } (\mu,\sigma),$$

so C' strictly dominates the t interval under this loss.

Problem 9.56. Let $X \sim f(x|\theta)$ and suppose that we want to estimate θ with an interval estimator C using the loss in (9.3.4), $L(\theta, C) = b\,\text{Length}(C) - I_C(\theta)$. If θ has the prior pdf $\pi(\theta)$, show that the Bayes rule is given by

$$C^\pi = \{\theta : \ \pi(\theta|x) \geq b\}.$$

(Hint: Write $\text{Length}(C) = \int_C 1\,d\theta$ and use the Neyman–Pearson Lemma.)

Solution. Keep exactly those θ whose posterior ordinate pays for the length they cost. The Bayes rule minimizes the posterior expected loss for each x separately, and with $\text{Length}(C) = \int_C 1\,d\theta$ that expected loss is a single integral:

$$\begin{aligned} E\big[L(\theta, C) \,\big|\, x\big] &= b\int_C 1\,d\theta - \int_C \pi(\theta|x)\,d\theta \\ &= \int_C \big[b - \pi(\theta|x)\big]\,d\theta. \end{aligned}$$

An integral $\int_C g$ over a variable set C is minimized by taking C to be exactly the set where the integrand is nonpositive, i.e. $C^\pi = \{\theta : b - \pi(\theta|x) \leq 0\} = \{\theta : \pi(\theta|x) \geq b\}$. Formally, for any competing set C,

$$\begin{aligned} \int_C \big[b - \pi(\theta|x)\big]d\theta &= \int_{C\cap C^\pi} \big[b - \pi\big] + \int_{C\setminus C^\pi} \big[b - \pi\big] \\ &\geq \int_{C\cap C^\pi} \big[b - \pi\big] \\ &\geq \int_{C\cap C^\pi} \big[b - \pi\big] + \int_{C^\pi\setminus C} \big[b - \pi\big] \\ &= \int_{C^\pi} \big[b - \pi(\theta|x)\big]d\theta, \end{aligned}$$

the first inequality because the integrand is > 0 off C^π and the second because it is ≤ 0 on C^π. Hence C^π minimizes the posterior expected loss for every x, and therefore minimizes the Bayes risk: it is the Bayes rule.

This is the form the hint intends: the display is the Lagrangian of the Neyman–Pearson Lemma (Theorem 8.3.12), with $d\theta$ in the role of the null density and $\pi(\theta|x)$ in the role of the alternative, so the optimal C is a likelihood-ratio region $\{\theta : \pi(\theta|x) \geq k\}$ and b is the multiplier k.

9.9 Exercises 9.57–9.58

Problem 9.57. (The following two problems relate to Miscellanea 9.5.4.)
Let $X_1, \ldots, X_n$ be iid $n(\mu, \sigma^2)$, where σ^2 is known. We know that a $1-\alpha$ confidence interval for μ is $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
(a) Show that a $1-\alpha$ prediction interval for X_{n+1} is

$$\bar{x} \pm z_{\alpha/2}\,\sigma\sqrt{1 + \tfrac{1}{n}}.$$

(b) Show that a $1-\alpha$ tolerance interval for $100p\%$ of the underlying population is given by

$$\bar{x} \pm \sigma\left(z_{p/2} + \frac{z_{\alpha/2}}{\sqrt{n}}\right).$$

(c) Find a $1-\alpha$ prediction interval for X_{n+1} if σ^2 is unknown.
(If σ^2 is unknown, the $1-\alpha$ tolerance interval is quite an involved calculation.)

Solution. (a) X_{n+1} is independent of $X_1, \dots, X_n$, hence of $\bar{X} \sim n(\mu, \sigma^2/n)$ (Theorem 5.3.1b), so

$$X_{n+1} - \bar{X} \;\sim\; n\Big(0,\; \sigma^2\Big(1 + \tfrac{1}{n}\Big)\Big).$$

Writing $\sigma_p = \sigma\sqrt{1 + 1/n}$,

$$\begin{aligned} P\big(\bar{X} - z_{\alpha/2}\sigma_p \le X_{n+1} \le \bar{X} + z_{\alpha/2}\sigma_p\big) &= P\left(\left|\frac{X_{n+1} - \bar{X}}{\sigma_p}\right| \le z_{\alpha/2}\right) \\ &= 1 - \alpha, \end{aligned}$$

which is requirement (ii) of Miscellanea 9.5.4.
(b) Read the printed $z_{p/2}$ as $z_{(1-p)/2}$: in the book's convention $P(Z > z_\gamma) = \gamma$, so it is $\mu \pm z_{(1-p)/2}\sigma$, not $\mu \pm z_{p/2}\sigma$, that carries population proportion p. Put

$$a \;=\; z_{(1-p)/2} + \frac{z_{\alpha/2}}{\sqrt{n}}, \qquad C(\mathbf{X}) = \big[\bar{X} - \sigma a,\; \bar{X} + \sigma a\big].$$

Setting $W = (\bar{X} - \mu)/\sigma$, the population proportion covered is

$$F(u(\mathbf{X}) \mid \mu) - F(l(\mathbf{X}) \mid \mu) = \Phi(W + a) - \Phi(W - a) \;=:\; g(W).$$

g is even, and $g'(w) = \phi(w+a) - \phi(w-a) < 0$ for $w > 0$ (since $a > 0$), so g decreases in $|w|$ and hence $\{g(W) \ge p\} \supseteq \{|W| \le c\}$ for any c with $g(c) \ge p$. Take $c = z_{\alpha/2}/\sqrt{n}$, so that $a = z_{(1-p)/2} + c$:

$$\begin{aligned} g(c) &= \Phi\big(z_{(1-p)/2} + 2c\big) - \Phi\big(-z_{(1-p)/2}\big) \\ &\ge \Phi\big(z_{(1-p)/2}\big) - \Phi\big(-z_{(1-p)/2}\big) \\ &= \Big(1 - \tfrac{1-p}{2}\Big) - \tfrac{1-p}{2} \;=\; p. \end{aligned}$$

Since $\sqrt{n}\,W \sim n(0,1)$, $P(|W| \le c) = P(|Z| \le z_{\alpha/2}) = 1 - \alpha$, and therefore

$$P_\mu[F(u(\mathbf{X}) \mid \mu) - F(l(\mathbf{X}) \mid \mu) \ge p] \;\ge\; 1 - \alpha,$$

which is requirement (iii) of Miscellanea 9.5.4.
(c) $\bar{x} \pm t_{n-1,\alpha/2}\, s\sqrt{1 + \frac{1}{n}}$.
Indeed $X_{n+1} - \bar{X} \sim n(0, \sigma^2(1+1/n))$ as in (a), and it is independent of S^2: S^2 is a function of $X_1, \dots, X_n$ alone, hence independent of X_{n+1}, and independent of $\bar{X}$ by Theorem 5.3.1a. Since $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ (Theorem 5.3.1c),

$$\frac{X_{n+1} - \bar{X}}{S\sqrt{1 + \frac{1}{n}}} = \frac{\big(X_{n+1} - \bar{X}\big) \,/\, \Big(\sigma\sqrt{1 + \frac{1}{n}}\Big)}{\sqrt{\left[(n-1)S^2/\sigma^2\right]/(n-1)}} \;\sim\; t_{n-1}$$

by Definition 5.3.4, and inverting $|T| \le t_{n-1,\alpha/2}$ gives the interval.

Problem 9.58. Let $X_1, \dots, X_n$ be iid observations from a population with median m. Distribution-free intervals can be based on the order statistics $X_{(1)} \le \cdots \le X_{(n)}$ in the following way.
(a) Show that the one-sided intervals $(-\infty, x_{(n)}]$ and $[x_{(1)}, \infty)$ are each confidence intervals for m with confidence coefficient $1 - (1/2)^n$, and the confidence coefficient of the interval $[x_{(1)}, x_{(n)}]$ is $1 - 2(1/2)^n$.
(b) Show that the one-sided intervals of part (a) are prediction intervals with coefficient $n/(n+1)$ and the two-sided interval is a prediction interval with coefficient $(n-1)/(n+1)$.
(c) The intervals in part (a) can also be used as tolerance intervals for proportion p of the underlying population. Show that, when considered as tolerance intervals, the one-sided intervals have coefficient

$1 - p^n$ and the two-sided interval has coefficient $1 - p^n - n(1-p)p^{n-1}$. Vardeman (1992) refers to this last calculation as a "nice exercise in order statistics."

Solution. (a) $\{X_{(n)} < m\} = \bigcap_{i=1}^{n} \{X_i < m\}$, an intersection of independent events each of probability $1/2$ (the cdf F is continuous, so $P(X_i = m) = 0$ and $P(X_i < m) = P(X_i > m) = 1/2$ by the definition of the median), so

$$P\big(m \le X_{(n)}\big) = 1 - \left(\tfrac{1}{2}\right)^n, \qquad P\big(X_{(1)} \le m\big) = 1 - \left(\tfrac{1}{2}\right)^n$$

(the second by the mirror argument on $\{X_{(1)} > m\} = \bigcap_i \{X_i > m\}$). The two exceptional events are disjoint, so

$$P\big(X_{(1)} \le m \le X_{(n)}\big) = 1 - \left(\tfrac{1}{2}\right)^n - \left(\tfrac{1}{2}\right)^n = 1 - 2\left(\tfrac{1}{2}\right)^n.$$

(b) $X_1, \ldots, X_{n+1}$ are iid and a.s. distinct, hence exchangeable, so each of the $n+1$ indices is equally likely to carry the largest value:

$$P\big(X_{n+1} > X_{(n)}\big) = P\left(X_{n+1} = \max_{1 \le i \le n+1} X_i\right) = \frac{1}{n+1}.$$

Thus $P(X_{n+1} \le X_{(n)}) = n/(n+1)$, and identically $P(X_{(1)} \le X_{n+1}) = n/(n+1)$. The two failure events are disjoint, so

$$P\big(X_{(1)} \le X_{n+1} \le X_{(n)}\big) = 1 - \frac{2}{n+1} = \frac{n-1}{n+1}.$$

(c) By the probability integral transform (Theorem 2.1.10) $U_i = F(X_i)$ are iid uniform$(0,1)$, and F is nondecreasing, so $F(X_{(i)}) = U_{(i)}$. For $(-\infty, X_{(n)}]$ the covered proportion is $F(X_{(n)}) - F(-\infty) = U_{(n)}$, and

$$P\big(U_{(n)} \ge p\big) = 1 - P(\textstyle\bigcap_i \{U_i < p\}) = 1 - p^n.$$

For $[X_{(1)}, \infty)$ the covered proportion is $1 - U_{(1)}$, and

$$P\big(1 - U_{(1)} \ge p\big) = P\big(U_{(1)} \le 1 - p\big) = 1 - p^n$$

since $P(U_{(1)} > 1 - p) = \prod_i P(U_i > 1 - p) = p^n$.
For $[X_{(1)}, X_{(n)}]$ the covered proportion is the uniform range $R = U_{(n)} - U_{(1)}$, which by Example 5.4.7 (with $a = 1$) is beta$(n-1, 2)$, $f_R(r) = n(n-1)r^{n-2}(1-r)$ on $(0,1)$. Hence

$$\begin{aligned} P(R \ge p) &= \int_p^1 n(n-1)r^{n-2}(1-r)\,dr \\ &= \left[nr^{n-1} - (n-1)r^n\right]_{r=p}^{1} \\ &= 1 - np^{n-1} + (n-1)p^n \\ &= 1 - p^n - n(1-p)p^{n-1}. \end{aligned}$$

10 Asymptotic Evaluations

Contents

10.1 Exercises 10.1–10.7 378
10.2 Exercises 10.8–10.14 381
10.3 Exercises 10.15–10.21 390
10.4 Exercises 10.22–10.28 397
10.5 Exercises 10.29–10.35 403
10.6 Exercises 10.36–10.42 412
10.7 Exercises 10.43–10.48 418

10.1 Exercises 10.1–10.7

Problem 10.1. A random sample $X_1, \ldots, X_n$ is drawn from a population with pdf

$$f(x|\theta) = \tfrac{1}{2}(1+\theta x), \qquad -1 < x < 1, \quad -1 < \theta < 1.$$

Find a consistent estimator of θ and show that it is consistent.

Solution. Take $W_n = 3\bar{X}_n$. Since

$$\mathrm{E}_\theta X = \int_{-1}^{1} x\,\tfrac{1}{2}(1+\theta x)\,dx = \tfrac{\theta}{2}\int_{-1}^{1} x^2\,dx = \tfrac{\theta}{3},$$
$$\mathrm{E}_\theta X^2 = \int_{-1}^{1} x^2\,\tfrac{1}{2}(1+\theta x)\,dx = \tfrac{1}{2}\cdot\tfrac{2}{3} = \tfrac{1}{3},$$

(the θ-term in the second integral is odd, hence vanishes), we get $\mathrm{E}_\theta W_n = 3\cdot\theta/3 = \theta$ and

$$\mathrm{Var}_\theta W_n = \frac{9\,\mathrm{Var}_\theta X}{n} = \frac{9}{n}\Big(\frac{1}{3} - \frac{\theta^2}{9}\Big) = \frac{3-\theta^2}{n}.$$

So $\mathrm{Bias}_\theta W_n = 0$ and $\mathrm{Var}_\theta W_n \to 0$ for every $\theta \in (-1,1)$; by Theorem 10.1.3, $W_n = 3\bar{X}_n$ is a consistent sequence of estimators of θ.

Problem 10.2. Prove Theorem 10.1.5. That is: let W_n be a consistent sequence of estimators of a parameter θ, and let $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ be sequences of constants satisfying (i) $\lim_{n\to\infty} a_n = 1$ and (ii) $\lim_{n\to\infty} b_n = 0$. Show that $U_n = a_n W_n + b_n$ is a consistent sequence of estimators of θ.

Solution. Everything follows from the decomposition $a_n W_n + b_n - \theta = a_n(W_n - \theta) + (a_n - 1)\theta + b_n$, i.e.

$$|U_n - \theta| \le |a_n|\,|W_n - \theta| + |a_n - 1|\,|\theta| + |b_n|.$$

Fix $\theta \in \Theta$ and $\epsilon > 0$. By (i) and (ii) there is an N such that for all $n \ge N$ both $|a_n| \le 2$ and $|a_n - 1|\,|\theta| + |b_n| < \epsilon/2$. For such n the display above gives the inclusion of events

$$\{|U_n - \theta| \ge \epsilon\} \subset \big\{\,|W_n - \theta| \ge \tfrac{\epsilon}{4}\,\big\},$$

since $|U_n - \theta| \ge \epsilon$ forces $2|W_n - \theta| \ge \epsilon/2$. Hence for $n \ge N$,

$$P_\theta\big(|U_n - \theta| \ge \epsilon\big) \le P_\theta\Big(|W_n - \theta| \ge \tfrac{\epsilon}{4}\Big) \xrightarrow[n\to\infty]{} 0$$

by the consistency of W_n in the form (10.1.2). As θ and ϵ were arbitrary, U_n is a consistent sequence of estimators of θ.

Problem 10.3. A random sample $X_1, \ldots, X_n$ is drawn from a population that is $\mathrm{n}(\theta, \theta)$, where $\theta > 0$.
(a) Show that the MLE of θ, $\hat{\theta}$, is a root of the quadratic equation $\theta^2 + \theta - W = 0$, where $W = (1/n)\sum_{i=1}^n X_i^2$, and determine which root equals the MLE.
(b) Find the approximate variance of $\hat{\theta}$ using the techniques of Section 10.1.3.

Solution. (a) $\hat\theta = \big(-1+\sqrt{1+4W}\big)/2$, the positive root. Dropping terms free of θ, the log likelihood is

$$\begin{aligned} l(\theta|\mathbf{x}) &= -\frac{n}{2}\log(2\pi\theta) - \frac{1}{2\theta}\sum_{i=1}^{n}(x_i-\theta)^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\theta - \frac{nW}{2\theta} + \sum_i x_i - \frac{n\theta}{2}, \end{aligned}$$

using $\sum(x_i-\theta)^2 = nW - 2\theta\sum x_i + n\theta^2$. Differentiating,

$$l'(\theta|\mathbf{x}) = -\frac{n}{2\theta} + \frac{nW}{2\theta^2} - \frac{n}{2} = \frac{n}{2\theta^2}\big(W-\theta-\theta^2\big),$$

so every stationary point solves $\theta^2+\theta-W=0$, whose roots are $\big(-1\pm\sqrt{1+4W}\big)/2$. Only the + root lies in $\Theta=(0,\infty)$ (as $W>0$ with probability 1), and since $\theta\mapsto\theta^2+\theta$ is increasing on $(0,\infty)$, $l'>0$ for $\theta<\hat\theta$ and $l'<0$ for $\theta>\hat\theta$. Thus $\hat\theta$ is the unique global maximum, i.e. the MLE.
(b) By (10.1.7) with $h(\theta)=\theta$, $\mathrm{Var}_\theta\hat\theta \approx 1/I_n(\theta)$, where $I_n(\theta) = \mathrm{E}_\theta(-l''(\theta|\mathbf{X}))$ by Lemma 7.3.11 (the support $(-\infty,\infty)$ does not depend on θ, so the regularity conditions hold). From the expression for l',

$$l''(\theta|\mathbf{x}) = \frac{n}{2\theta^2} - \frac{nW}{\theta^3}, \qquad \mathrm{E}_\theta X^2 = \theta+\theta^2,$$

whence

$$I_n(\theta) = -\frac{n}{2\theta^2} + \frac{n(\theta+\theta^2)}{\theta^3} = \frac{n}{2\theta^2} + \frac{n}{\theta} = \frac{n(1+2\theta)}{2\theta^2}.$$

Therefore

$$\mathrm{Var}_\theta\hat\theta \approx \frac{2\theta^2}{n(1+2\theta)}, \qquad \widehat{\mathrm{Var}}_{\hat\theta}\,\hat\theta = \frac{2\hat\theta^{\,2}}{n(1+2\hat\theta)}.$$

Problem 10.4. A variation of the model in Exercise 7.19 (there $Y_i=\beta x_i+\epsilon_i$ with $x_1,\dots,x_n$ fixed constants) is to let the random variables $Y_1,\dots,Y_n$ satisfy

$$Y_i = \beta X_i + \epsilon_i, \qquad i=1,\dots,n,$$

where $X_1,\dots,X_n$ are independent $\mathrm{n}(\mu,\tau^2)$ random variables, $\epsilon_1,\dots,\epsilon_n$ are iid $\mathrm{n}(0,\sigma^2)$, and the Xs and ϵs are independent. Exact variance calculations become quite difficult, so we might resort to approximations. In terms of μ, τ^2, and σ^2, find approximate means and variances for
(a) $\sum X_iY_i/\sum X_i^2$
(b) $\sum Y_i/\sum X_i$
(c) $\sum(Y_i/X_i)/n$

Solution. All three have approximate mean β, with approximate variances

$$\frac{\sigma^2}{n(\mu^2+\tau^2)}, \qquad \frac{\sigma^2}{n\mu^2}, \qquad \frac{\sigma^2}{n\mu^2}$$

respectively; throughout we use the ratio approximations of Example 5.5.27,

$$\mathrm{E}\Big(\frac{U}{V}\Big) \approx \frac{\mu_U}{\mu_V}, \quad \mathrm{Var}\Big(\frac{U}{V}\Big) \approx \frac{\mathrm{Var}\,U}{\mu_V^2} + \frac{\mu_U^2\,\mathrm{Var}\,V}{\mu_V^4} - \frac{2\mu_U\,\mathrm{Cov}(U,V)}{\mu_V^3},$$

which need $\mu_V\neq 0$, hence $\mu\neq 0$ in (b) and (c).
(a) Here

$$\frac{\sum X_iY_i}{\sum X_i^2} = \beta + \frac{N_a}{D_a}, \qquad N_a = \frac{1}{n}\sum_i X_i\epsilon_i, \quad D_a = \frac{1}{n}\sum_i X_i^2.$$

Now $\mathrm{E}N_a = \mathrm{E}X\,\mathrm{E}\epsilon = 0$ and $\mathrm{E}D_a = \mu^2+\tau^2$, so by Example 5.5.27, $\mathrm{E}(N_a/D_a) \approx 0/(\mu^2+\tau^2) = 0$ and, the numerator mean being zero, the two remaining terms of the variance formula drop out:

$$\operatorname{Var}\Big(\frac{N_a}{D_a}\Big) \approx \frac{\operatorname{Var} N_a}{(\mathrm{E}D_a)^2} = \frac{\sigma^2(\mu^2+\tau^2)/n}{(\mu^2+\tau^2)^2} = \frac{\sigma^2}{n(\mu^2+\tau^2)},$$

using $\operatorname{Var}(X\epsilon) = \mathrm{E}X^2\,\mathrm{E}\epsilon^2 = \sigma^2(\mu^2+\tau^2)$. So the estimator has approximate mean β and approximate variance $\sigma^2/[n(\mu^2+\tau^2)]$.
(b) Here $\sum Y_i/\sum X_i = \bar{Y}/\bar{X} = \beta + \bar{\epsilon}/\bar{X}$, with $\mathrm{E}\bar{\epsilon} = 0$, $\operatorname{Var}\bar{\epsilon} = \sigma^2/n$, $\mathrm{E}\bar{X} = \mu$, and $\operatorname{Cov}(\bar{\epsilon},\bar{X}) = 0$. Example 5.5.27 again gives mean $\approx \beta$ and

$$\operatorname{Var}\Big(\frac{\bar{\epsilon}}{\bar{X}}\Big) \approx \frac{\operatorname{Var}\bar{\epsilon}}{\mu^2} = \frac{\sigma^2}{n\mu^2}.$$

(c) Here the n summands Y_i/X_i are iid ratios, so apply Example 5.5.27 to a single pair $(U,V) = (Y,X)$: $\mu_U = \beta\mu$, $\mu_V = \mu$, $\operatorname{Var}Y = \beta^2\tau^2+\sigma^2$, $\operatorname{Var}X = \tau^2$, and $\operatorname{Cov}(X,Y) = \beta\operatorname{Var}X = \beta\tau^2$. Thus $\mathrm{E}(Y/X) \approx \beta$ and

$$\begin{aligned}\operatorname{Var}\Big(\frac{Y}{X}\Big) &\approx \frac{\beta^2\tau^2+\sigma^2}{\mu^2} + \frac{\beta^2\mu^2\tau^2}{\mu^4} - \frac{2\beta\mu\cdot\beta\tau^2}{\mu^3} \\ &= \frac{\beta^2\tau^2+\sigma^2+\beta^2\tau^2-2\beta^2\tau^2}{\mu^2} = \frac{\sigma^2}{\mu^2},\end{aligned}$$

so the average of n of them has approximate mean β and approximate variance $\sigma^2/(n\mu^2)$.

Problem 10.5. For the situation of Example 10.1.8 ($\bar{X}_n$ the mean of n iid normal observations with $\mathrm{E}X = \mu$ and $\operatorname{Var}X = \sigma^2$) show that for $T_n = \sqrt{n}/\bar{X}_n$:
(a) $\operatorname{Var}(T_n) = \infty$.
(b) If $\mu \neq 0$ and we delete the interval $(-\delta,\delta)$ from the sample space, then $\operatorname{Var}(T_n) < \infty$.
(c) If $\mu \neq 0$, the probability content of the interval $(-\delta,\delta)$ approaches 0 as $n \to \infty$.

Solution. (a) The variance is infinite because the density of $\bar{X}_n$ is bounded away from 0 near the origin while $1/\bar{x}^2$ is not integrable there. Writing g_n for the $\mathrm{n}(\mu,\sigma^2/n)$ density and putting $c = \min_{|x|\le 1} g_n(x) > 0$,

$$\mathrm{E}\,T_n^2 = n\int_{-\infty}^{\infty} \frac{1}{x^2}\,g_n(x)\,dx \ge nc\int_{-1}^{1}\frac{dx}{x^2} = \infty.$$

Since $(t-c)^2 \ge \tfrac{1}{2}t^2 - c^2$, this gives $\mathrm{E}(T_n-c)^2 = \infty$ for every finite centering c, so $\operatorname{Var}(T_n) = \infty$.
(b) Deleting $(-\delta,\delta)$ leaves $|\bar{X}_n| \ge \delta$, on which T_n is bounded: $|T_n| = \sqrt{n}/|\bar{X}_n| \le \sqrt{n}/\delta$. Hence

$$\operatorname{Var}(T_n) \le \mathrm{E}\big(T_n^2 \mid |\bar{X}_n| \ge \delta\big) \le \frac{n}{\delta^2} < \infty,$$

the conditioning being legitimate since $P(|\bar{X}_n| \ge \delta) > 0$ for the normal distribution of $\bar{X}_n$.
(c) Take $0 < \delta < |\mu|$ (which is possible exactly because $\mu \neq 0$). Then $|\bar{x}| < \delta$ forces $|\bar{x}-\mu| > |\mu| - \delta > 0$, so

$$P\big(|\bar{X}_n| < \delta\big) \le P\big(|\bar{X}_n - \mu| > |\mu| - \delta\big) \xrightarrow[n\to\infty]{} 0$$

by the consistency of $\bar{X}_n$ for μ (Example 10.1.4, via the Weak Law Theorem 5.5.2).

Problem 10.6. For the situation of Example 10.1.10, in which

$$Y_n \mid W_n = w_n \sim \mathrm{n}\big(0,\ w_n + (1-w_n)\sigma_n^2\big), \qquad W_n \sim \text{Bernoulli}(p_n),$$

show that
(a) $\mathrm{E}Y_n = 0$ and $\operatorname{Var}(Y_n) = p_n + (1-p_n)\sigma_n^2$.

(b) $P(Y_n < a) \to P(Z < a)$, and hence $Y_n \to \mathrm{n}(0,1)$ (recall that $p_n \to 1$, $\sigma_n \to \infty$, and $(1-p_n)\sigma_n^2 \to \infty$).

Solution. (a) Both claims are the conditional expectation and conditional variance identities (Theorems 4.4.3 and 4.4.7) applied to the conditioning variable W_n. Since $\mathrm{E}(Y_n|W_n) = 0$ identically,

$$\mathrm{E}Y_n = \mathrm{E}\big[\mathrm{E}(Y_n|W_n)\big] = 0,$$

and, the second term of the variance identity vanishing because $\mathrm{E}(Y_n|W_n)$ is the constant 0,

$$\begin{aligned}\mathrm{Var}(Y_n) &= \mathrm{E}\big[\mathrm{Var}(Y_n|W_n)\big] + \mathrm{Var}\big[\mathrm{E}(Y_n|W_n)\big] \\ &= \mathrm{E}\big[W_n + (1-W_n)\sigma_n^2\big] + 0 = p_n + (1-p_n)\sigma_n^2,\end{aligned}$$

using $\mathrm{E}W_n = p_n$.
(b) Conditioning on W_n and using $Y_n|W_n = 1 \sim \mathrm{n}(0,1)$, $Y_n|W_n = 0 \sim \mathrm{n}(0,\sigma_n^2)$,

$$P(Y_n < a) = p_n P(Z < a) + (1-p_n)P(Z < a/\sigma_n), \qquad Z \sim \mathrm{n}(0,1).$$

Hence, bounding both probabilities in the discrepancy by 1,

$$\big|P(Y_n < a) - P(Z < a)\big| = (1-p_n)\big|P(Z < a/\sigma_n) - P(Z < a)\big| \le 1 - p_n \to 0$$

because $p_n \to 1$. So $P(Y_n < a) \to P(Z < a)$ at every a, that is, $Y_n \to \mathrm{n}(0,1)$ in distribution.

Problem 10.7. In the proof of Theorem 10.1.12 it was shown that the MLE $\hat\theta$ is an asymptotically efficient estimator of θ. Show that if $\tau(\theta)$ is a continuous and differentiable function of θ, then $\tau(\hat\theta)$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$.

Solution. Consistency is Theorem 10.1.6 itself, whose conclusion is stated for $\tau(\hat\theta)$ with τ any continuous function of θ. For efficiency, apply the Delta Method (Theorem 5.5.24) with $g = \tau$ to the conclusion

$$\sqrt{n}\,(\hat\theta - \theta) \to \mathrm{n}\Big(0, \frac{1}{I(\theta)}\Big), \qquad I(\theta) = \mathrm{E}_\theta\Big(\Big(\tfrac{\partial}{\partial\theta}\log f(X|\theta)\Big)^2\Big),$$

established for the MLE in the proof of Theorem 10.1.12. Theorem 5.5.24 requires only that $\tau'(\theta)$ exist and be nonzero at the true θ (differentiability is hypothesized; if $\tau'(\theta) = 0$ then the Cramer-Rao bound for τ is 0 and the second-order Theorem 5.5.26 applies instead), and it yields

$$\sqrt{n}\,\big[\tau(\hat\theta) - \tau(\theta)\big] \to \mathrm{n}\left(0,\ \frac{[\tau'(\theta)]^2}{\mathrm{E}_\theta\big(\big(\frac{\partial}{\partial\theta}\log f(X|\theta)\big)^2\big)}\right).$$

The asymptotic variance displayed on the right is exactly the quantity $v(\theta)$ of Definition 10.1.11 for the parametric function $\tau(\theta)$, so $\tau(\hat\theta)$ is asymptotically efficient for $\tau(\theta)$.

10.2 Exercises 10.8–10.14

Problem 10.8. Finish the proof of Theorem 10.1.6 by establishing the two convergence results in (10.1.6).
(a) Show that

$$\frac{1}{\sqrt{n}}\,l'(\theta_0|\mathbf{X}) = \sqrt{n}\left[\frac{1}{n}\sum_i W_i\right],$$

where

$$W_i \;=\; \left.\frac{\frac{\partial}{\partial\theta}f(X_i|\theta)}{f(X_i|\theta)}\right|_{\theta=\theta_0}$$

has mean 0 and variance $I(\theta_0)$. Now use the Central Limit Theorem to establish the convergence to $\mathrm{n}[0, I(\theta_0)]$.
(b) Show that

$$-\frac{1}{n}l''(\theta_0|\mathbf{X}) = \frac{1}{n}\sum_i W_i^2 \;-\; \frac{1}{n}\sum_i \left.\frac{\frac{\partial^2}{\partial\theta^2}f(X_i|\theta)}{f(X_i|\theta)}\right|_{\theta=\theta_0},$$

and that the mean of the first piece is $I(\theta_0)$ and the mean of the second piece is 0. Apply the WLLN.

Solution. (Throughout, Theorem 10.1.6 is a misprint for Theorem 10.1.12, in whose proof the display (10.1.6) sits.)
(a) Differentiating $l(\theta|\mathbf{x}) = \sum_i \log f(x_i|\theta)$ once gives $l'(\theta_0|\mathbf{X}) = \sum_i W_i$ directly, so

$$\frac{1}{\sqrt{n}}l'(\theta_0|\mathbf{X}) = \frac{1}{\sqrt{n}}\sum_i W_i = \sqrt{n}\left[\frac{1}{n}\sum_i W_i\right].$$

The W_i are iid, and the regularity conditions of Miscellanea 10.6.2 permit differentiation under the integral sign (Theorem 2.4.3), so

$$\mathrm{E}_{\theta_0}W_1 = \int \frac{\frac{\partial}{\partial\theta}f(x|\theta_0)}{f(x|\theta_0)}f(x|\theta_0)\,dx = \frac{d}{d\theta}\int f(x|\theta)\,dx\Big|_{\theta_0} = \frac{d}{d\theta}1 = 0,$$

$$\mathrm{Var}_{\theta_0}W_1 = \mathrm{E}_{\theta_0}W_1^2 = \mathrm{E}_{\theta_0}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\Big|_{\theta_0}\right)^2\right] = I(\theta_0),$$

the last line being the definition of the information number for one observation. Since $0 < I(\theta_0) < \infty$, the Central Limit Theorem (Theorem 5.5.14) applies to the iid sequence $W_1, W_2, \ldots$ and yields

$$\frac{1}{\sqrt{n}}l'(\theta_0|\mathbf{X}) = \sqrt{n}\left[\frac{1}{n}\sum_i W_i - 0\right] \longrightarrow \mathrm{n}[0, I(\theta_0)]$$

in distribution (the limit law is symmetric about 0, so the sign carried by the numerator of (10.1.5) is immaterial).
(b) For a single observation, the quotient rule gives

$$\frac{\partial^2}{\partial\theta^2}\log f(x|\theta) = \frac{\frac{\partial^2}{\partial\theta^2}f(x|\theta)}{f(x|\theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right)^2,$$

so summing over i, evaluating at θ_0, and multiplying by $-1/n$,

$$-\frac{1}{n}l''(\theta_0|\mathbf{X}) = \frac{1}{n}\sum_i W_i^2 - \frac{1}{n}\sum_i V_i, \qquad V_i := \left.\frac{\frac{\partial^2}{\partial\theta^2}f(X_i|\theta)}{f(X_i|\theta)}\right|_{\theta_0}.$$

By part (a), $\mathrm{E}_{\theta_0}W_1^2 = I(\theta_0)$, and differentiating twice under the integral sign,

$$\mathrm{E}_{\theta_0}V_1 = \int \frac{\partial^2}{\partial\theta^2}f(x|\theta)\Big|_{\theta_0}dx = \frac{d^2}{d\theta^2}\int f(x|\theta)\,dx\Big|_{\theta_0} = 0.$$

Each of $\{W_i^2\}$ and $\{V_i\}$ is an iid sequence with finite variance under the regularity conditions, so the Weak Law of Large Numbers (Theorem 5.5.2) gives $\frac{1}{n}\sum_i W_i^2 \to I(\theta_0)$ and $\frac{1}{n}\sum_i V_i \to 0$, both in probability. Hence

$$-\frac{1}{n}l''(\theta_0|\mathbf{X}) \longrightarrow I(\theta_0) - 0 = I(\theta_0)$$

in probability, which is the second statement of (10.1.6). Feeding both limits into (10.1.5) and applying Slutsky's Theorem (Theorem 5.5.17) completes the proof of Theorem 10.1.12.

Problem 10.9. Suppose that $X_1, \ldots, X_n$ are iid Poisson(λ). Find the best unbiased estimator of
(a) $e^{-\lambda}$, the probability that $X = 0$.
(b) $\lambda e^{-\lambda}$, the probability that $X = 1$.
(c) For the best unbiased estimators of parts (a) and (b), calculate the asymptotic relative efficiency with respect to the MLE. Which estimators do you prefer? Why?
(d) A preliminary test of a possible carcinogenic compound can be performed by measuring the mutation rate of microorganisms exposed to the compound. An experimenter places the compound in 15 petri dishes and records the following number of mutant colonies:

10	7	8	13	8
9	5	7	6	8
3	6	6	3	5

Estimate $e^{-\lambda}$, the probability that no mutant colonies emerge, and $\lambda e^{-\lambda}$, the probability that one mutant colony will emerge. Calculate both the best unbiased estimator and the MLE.

Solution. (a) $\left(\frac{n-1}{n}\right)^T$, where $T = \sum_i X_i$.
The Poisson family is a full exponential family, so $T \sim \text{Poisson}(n\lambda)$ is a complete sufficient statistic (Theorem 6.2.25); by Theorem 7.3.23 any function of T is the unique best unbiased estimator of its expectation. Start from the unbiased estimator $h(\mathbf{X}) = I(X_1 = 0)$ and condition, using

$$P(X_1 = x \,|\, T = t) = \frac{P(X_1 = x)\, P\Big(\sum_{i\geq 2} X_i = t - x\Big)}{P(T = t)} = \binom{t}{x}\left(\frac{1}{n}\right)^x \left(\frac{n-1}{n}\right)^{t-x},$$

i.e. $X_1 | T = t \sim \text{binomial}(t, 1/n)$, free of λ as it must be. Hence

$$\mathrm{E}\big[I(X_1 = 0) \,\big|\, T\big] = \left(\frac{n-1}{n}\right)^T.$$

(b) Same conditioning with $h(\mathbf{X}) = I(X_1 = 1)$:

$$\mathrm{E}\big[I(X_1 = 1) \,\big|\, T\big] = \binom{T}{1}\frac{1}{n}\left(\frac{n-1}{n}\right)^{T-1} = \bar{X}\left(\frac{n-1}{n}\right)^{T-1}.$$

(c) Both AREs equal 1. The MLEs are $e^{-\bar{X}}$ and $\bar{X}e^{-\bar{X}}$ by invariance (Theorem 7.2.10), and the Delta Method (Theorem 5.5.24) with $\mathrm{Var}X = \lambda$ gives

$$\sqrt{n}\left(e^{-\bar{X}} - e^{-\lambda}\right) \to \mathrm{n}\Big[0,\ \lambda e^{-2\lambda}\Big],$$
$$\sqrt{n}\left(\bar{X}e^{-\bar{X}} - \lambda e^{-\lambda}\right) \to \mathrm{n}\Big[0,\ \lambda(1-\lambda)^2 e^{-2\lambda}\Big].$$

For the estimator of (a), write $T = n\bar{X}$ and expand $n\log\left(1 - \frac{1}{n}\right) = -1 - \frac{1}{2n} + O(n^{-2})$:

$$\left(\frac{n-1}{n}\right)^T = \exp\big\{\bar{X}\, n\log\big(1 - \tfrac{1}{n}\big)\big\} = e^{-\bar{X}}\exp\left\{-\frac{\bar{X}}{2n} + O(n^{-2})\right\},$$

so that

$$\sqrt{n}\left[\left(\frac{n-1}{n}\right)^T - e^{-\bar{X}}\right] = -\frac{\bar{X}e^{-\bar{X}}}{2\sqrt{n}}\left(1 + o_P(1)\right) \longrightarrow 0$$

in probability (using $\bar{X} \to \lambda$). The same expansion, together with $\left(1 - \frac{1}{n}\right)^{-1} \to 1$, gives the identical conclusion for (b). By Slutsky's Theorem the best unbiased estimators therefore have exactly the asymptotic distributions displayed above, and

$$\mathrm{ARE}\left(\left(\tfrac{n-1}{n}\right)^T, e^{-\bar{X}}\right) = 1, \qquad \mathrm{ARE}\left(\bar{X}\left(\tfrac{n-1}{n}\right)^{T-1}, \bar{X}e^{-\bar{X}}\right) = 1.$$

Since asymptotics cannot separate them, the choice rests on finite-sample grounds, where the best unbiased estimators win: they are exactly unbiased at every n and cost no more to compute, whereas $e^{-\bar X}$ is biased upward at every n by Jensen's inequality (e^{-t} is convex).
(d) Here $n = 15$ and $T = \sum_i x_i = 104$, so $\bar x = 104/15 = 6.9\overline{3}$ and

$$\widehat{e^{-\lambda}}_{\text{MLE}} = e^{-104/15} = 0.000975, \qquad \widehat{e^{-\lambda}}_{\text{UMVUE}} = \left(\tfrac{14}{15}\right)^{104} = 0.000765,$$
$$\widehat{\lambda e^{-\lambda}}_{\text{MLE}} = \tfrac{104}{15} e^{-104/15} = 0.00676, \quad \widehat{\lambda e^{-\lambda}}_{\text{UMVUE}} = \tfrac{104}{15}\left(\tfrac{14}{15}\right)^{103} = 0.00568.$$

Problem 10.10. Continue the calculations of Example 10.1.15, where the properties of the estimator of $p(1-p)$ were examined.
(a) Show that, if $p \neq 1/2$, the MLE $\hat p(1-\hat p)$ is asymptotically efficient.
(b) If $p = 1/2$, use Theorem 5.5.26 to find a limiting distribution of $\hat p(1-\hat p)$.
(c) Calculate the exact expression for $\text{Var}[\hat p(1-\hat p)]$. Is the reason for the failure of the approximations any clearer?

Solution. (a) Take $h(p) = p(1-p)$, so $h'(p) = 1 - 2p \neq 0$ for $p \neq 1/2$. Since $\sqrt{n}(\hat p - p) \to \text{n}[0, p(1-p)]$ (Example 10.1.14), the Delta Method (Theorem 5.5.24) gives

$$\sqrt{n}\,[\hat p(1-\hat p) - p(1-p)] \longrightarrow \text{n}\big[0,\ (1-2p)^2 p(1-p)\big]\,.$$

For one Bernoulli observation the information is

$$\text{E}_p\left(\frac{\partial}{\partial p}\log f(X|p)\right)^2 = \frac{1}{p(1-p)},$$

so the bound of Definition 10.1.11 is

$$v(p) = \frac{[h'(p)]^2}{1/[p(1-p)]} = (1-2p)^2 p(1-p),$$

which is exactly the asymptotic variance above. Hence $\hat p(1-\hat p)$ is asymptotically efficient whenever $p \neq 1/2$.
(b) $n\left[\hat p(1-\hat p) - \frac14\right] \to -\frac14\chi_1^2$.
At $p = 1/2$ we have $h'(1/2) = 0$ and $h''(1/2) = -2 \neq 0$, so the hypotheses of Theorem 5.5.26 hold with $\sigma^2 = p(1-p) = 1/4$, and (5.5.13) reads

$$n\left[\hat p(1-\hat p) - \tfrac14\right] \longrightarrow \frac14 \cdot \frac{-2}{2}\,\chi_1^2 = -\frac14\,\chi_1^2.$$

(c) Write $q = 1-p$ and $D = \hat p - p$, so that $\hat p(1-\hat p) = pq + (q-p)D - D^2$ and therefore

$$\text{Var}\,[\hat p(1-\hat p)] = (q-p)^2\mu_2 - 2(q-p)\mu_3 + \mu_4 - \mu_2^2,$$

where μ_k is the kth central moment of $\hat p = Y/n$, $Y \sim$ binomial(n,p):

$$\mu_2 = \frac{pq}{n}, \quad \mu_3 = \frac{pq(q-p)}{n^2}, \quad \mu_4 = \frac{3p^2q^2}{n^2} + \frac{pq(1-6pq)}{n^3}.$$

Substituting and collecting,

$$\text{Var}\,[\hat p(1-\hat p)] = \frac{pq(q-p)^2}{n} - \frac{2pq(q-p)^2}{n^2} + \frac{2p^2q^2}{n^2} + \frac{pq(1-6pq)}{n^3}.$$

(Check: at $n = 24$, $p = 1/4$ this gives .001911, and summing $(\hat p(1-\hat p))^2$ against the binomial pmf gives .001911.) At $p = 1/2$ every term containing $q - p$ drops out and

$$\text{Var}\,[\hat p(1-\hat p)]\,\Big|_{p=1/2} = \frac{1}{8n^2} - \frac{1}{8n^3} = \frac{n-1}{8n^3}.$$

Yes, the reason is now transparent. The Delta Method estimate $\hat{p}(1-\hat{p})(1-2\hat{p})^2/n$ of Example 10.1.15 is estimating the leading n^{-1} term $pq(q-p)^2/n$, and that term is identically zero at $p = 1/2$: the first-order approximation is not wrong, it is merely blind, because the true variance at $p = 1/2$ is of order n^{-2}. The second-order limit of part (b) sees it, $\mathrm{Var}\left(-\frac{1}{4}\chi^2_1\right) = \frac{1}{8}$, matching $n^2\mathrm{Var}[\hat{p}(1-\hat{p})] \to \frac{1}{8}$.

Problem 10.11. This problem will look at some details and extensions of the calculation in Example 10.1.18, where $X_1, \ldots, X_n$ are a random sample from the gamma(α, β) density

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \qquad x > 0,$$

reparameterized so that the mean $\mu = \alpha\beta$ is explicit, and where Figure 10.1.1 plots $\mathrm{ARE}(\hat{\mu}, \bar{X})$ against the gamma mean μ over $1 \le \mu \le 10$ as four curves corresponding to the scale values $\beta = 1, 3, 5, 10$, the higher curves belonging to the higher values of β.
(a) Reproduce Figure 10.1.1, calculating the ARE for known β. (You can follow the calculations in Example A.0.7, or do your own programming.)
(b) Verify that the $\mathrm{ARE}(\bar{X}, \hat{\mu})$ comparison is the same whether β is known or unknown.
(c) For estimation of β with known μ, show that the method of moment estimate and MLEs are the same. (It may be easier to use the (α, β) parameterization.)
(d) For estimation of β with unknown μ, the method of moment estimate and MLEs are not the same. Compare these estimates using asymptotic relative efficiency, and produce a figure like Figure 10.1.1, where the different curves correspond to different values of μ.

Solution. Throughout, $\psi(t) = \frac{d}{dt}\log\Gamma(t)$ is the digamma function and ψ' the trigamma function, and $\alpha = \mu/\beta$.
(a) $\mathrm{ARE}(\hat{\mu}, \bar{X}) = \alpha\,\psi'(\alpha)$ with $\alpha = \mu/\beta$.
In the (μ, β) parameterization,

$$\log f(x|\mu,\beta) = -\log\Gamma\left(\tfrac{\mu}{\beta}\right) - \tfrac{\mu}{\beta}\log\beta + \left(\tfrac{\mu}{\beta} - 1\right)\log x - \tfrac{x}{\beta},$$

so that

$$\frac{\partial^2}{\partial\mu^2}\log f(x|\mu,\beta) = -\frac{1}{\beta^2}\psi'\left(\frac{\mu}{\beta}\right),$$

free of x. Hence the information for one observation is $\psi'(\alpha)/\beta^2$, and since $\mathrm{Var}X = \alpha\beta^2 = \mu\beta$, the formula of Example 10.1.18 gives

$$\mathrm{ARE}(\hat{\mu}, \bar{X}) = [\beta\mu]\left[\mathrm{E}\left(-\frac{d^2}{d\mu^2} l(\mu,\beta|X)\right)\right] = \mu\beta \cdot \frac{\psi'(\alpha)}{\beta^2} = \alpha\,\psi'(\alpha).$$

Since $\psi'(\alpha) = \frac{1}{\alpha} + \frac{1}{2\alpha^2} + O(\alpha^{-3})$, the ARE is $1 + \frac{1}{2\alpha} + O(\alpha^{-2})$: always above 1, decreasing in $\alpha = \mu/\beta$, and tending to 1. This reproduces Figure 10.1.1:

β	$\mu = 1$	$\mu = 2$	$\mu = 4$	$\mu = 6$	$\mu = 8$	$\mu = 10$
1	1.645	1.290	1.135	1.088	1.065	1.052
3	3.365	2.043	1.461	1.290	1.210	1.165
5	5.253	2.910	1.840	1.521	1.373	1.290
10	10.143	5.253	2.910	2.182	1.840	1.645

(b) As printed the two comparisons differ: with β unknown the MLE of μ is $\bar{X}$ itself, so $\mathrm{ARE}(\bar{X}, \hat{\mu}) = 1$, whereas with β known it is $1/[\alpha\psi'(\alpha)] < 1$. What does hold in both problems, and is the content of the part, is that the asymptotic variance of $\bar{X}$ is $\mu\beta$ – the same whether β is known or estimated.
In the (α, β) parameterization the likelihood equation for β is

$$\frac{\partial l}{\partial\beta} = -\frac{n\alpha}{\beta} + \frac{\sum_i x_i}{\beta^2} = 0 \quad \Longrightarrow \quad \hat{\mu} = \hat{\alpha}\hat{\beta} = \bar{x},$$

so the MLE of the mean is the sample mean exactly. The asymptotic version of the same statement: with

$$\mathrm{I}(\alpha,\beta)=\begin{pmatrix}\psi'(\alpha) & 1/\beta\\ 1/\beta & \alpha/\beta^2\end{pmatrix},\qquad \nabla\mu=(\beta,\ \alpha)^\top,$$

we get $\det \mathrm{I}=[\alpha\psi'(\alpha)-1]/\beta^2$ and

$$\begin{aligned}\nabla\mu^\top \mathrm{I}^{-1}\nabla\mu &= \frac{\beta^2}{\alpha\psi'(\alpha)-1}\left[\beta^2\tfrac{\alpha}{\beta^2}-2\alpha\beta\tfrac{1}{\beta}+\alpha^2\psi'(\alpha)\right]\\ &= \frac{\beta^2\alpha\,[\alpha\psi'(\alpha)-1]}{\alpha\psi'(\alpha)-1}=\alpha\beta^2=\mu\beta,\end{aligned}$$

which is exactly $n\,\mathrm{Var}(\bar X)$. (In the (μ,β) coordinates $I_{\mu\beta}=[1-\alpha\psi'(\alpha)]/\beta^2\neq 0$, so the cancellation is genuine and not an artifact of orthogonality.)
(c) Both equal $\bar X/\alpha$. (As printed the part cannot be right – with the mean μ known, the MLE of β solves a transcendental digamma equation while the moment estimate is $(\overline{X^2}-\mu^2)/\mu$; the hint names the intended reading, namely β estimated with the shape α known.) With α known, $\mathrm{E}X=\alpha\beta$ gives the moment estimate $\tilde\beta=\bar X/\alpha$, while

$$\frac{\partial l}{\partial\beta}=-\frac{n\alpha}{\beta}+\frac{\sum_i X_i}{\beta^2}=0\quad\Longrightarrow\quad\hat\beta=\frac{\bar X}{\alpha},$$

the unique maximum since $\partial l/\partial\beta>0$ for $\beta<\bar X/\alpha$ and <0 for $\beta>\bar X/\alpha$.
(d) With $\alpha=\mu/\beta$,

$$\mathrm{ARE}(\tilde\beta,\hat\beta)=\frac{\alpha\psi'(\alpha)}{(2\alpha+3)\,[\alpha\psi'(\alpha)-1]}.$$

Matching the first two moments, $\bar X=\alpha\beta$ and $\overline{X^2}=\alpha(\alpha+1)\beta^2$, gives

$$\tilde\beta=\frac{\overline{X^2}-\bar X^2}{\bar X}=\frac{\frac{1}{n}\sum_i(X_i-\bar X)^2}{\bar X}.$$

Apply the Multivariate Delta Method (Theorem 5.5.28) to $g(a,b)=b/a-a$ at $(a,b)=(\alpha\beta,\alpha(\alpha+1)\beta^2)$, where $g_a=-(2\alpha+1)/\alpha$ and $g_b=1/(\alpha\beta)$, using the gamma moments $\mathrm{Var}X=\alpha\beta^2$, $\mathrm{Cov}(X,X^2)=2\alpha(\alpha+1)\beta^3$ and $\mathrm{Var}(X^2)=\alpha(\alpha+1)(4\alpha+6)\beta^4$:

$$\begin{aligned}\tau^2=\frac{\beta^2}{\alpha}\Big[&(2\alpha+1)^2-4(2\alpha+1)(\alpha+1)\\ &+(\alpha+1)(4\alpha+6)\Big]=\frac{\beta^2(2\alpha+3)}{\alpha},\end{aligned}$$

the α^2 terms cancelling. For the MLE, the (β,β) entry of I^{-1} computed in part (b) is

$$\mathrm{Var}_{\mathrm{asym}}(\hat\beta)=\frac{\beta^2\psi'(\alpha)}{\alpha\psi'(\alpha)-1},$$

so Definition 10.1.16 gives the ARE displayed above. (Check at $\alpha=2.3,\beta=1.7$: $\tau^2=9.550$ against 9.485 from simulation with $n=4000$.) The ARE depends on (μ,β) only through $\alpha=\mu/\beta$, it increases from $\frac{1}{3}$ (as $\alpha\to 0$) to 1 (as $\alpha\to\infty$, where the gamma is nearly normal and the sample variance is nearly efficient), so plotted against β the curves for larger μ lie uniformly higher:

μ	$\beta=1$	$\beta=2$	$\beta=4$	$\beta=6$	$\beta=8$	$\beta=10$
1	.510	.420	.372	.357	.351	.347
3	.712	.581	.467	.420	.396	.382
5	.798	.678	.548	.482	.444	.420
10	.885	.798	.678	.601	.548	.510

Problem 10.12. Verify that the superefficient estimator d_n of Miscellanea 10.6.1 is asymptotically normal with variance $v(\theta) = 1$ when $\theta \neq 0$ and $v(\theta) = a^2$ when $\theta = 0$. (See Lehmann and Casella 1998, Section 6.2, for more on superefficient estimators.)
Miscellanea 10.6.1 takes $X_1, \ldots, X_n$ iid $\mathrm{n}(\theta, 1)$, for which the Cramer–Rao Lower Bound for unbiased estimators of θ is $1/n$, and defines Hodges' estimator

$$d_n = \begin{cases} \bar{X} & \text{if } |\bar{X}| \geq n^{-1/4}, \\ a\bar{X} & \text{if } |\bar{X}| < n^{-1/4}, \end{cases}$$

with $a < 1$ a fixed constant; the claim is that $\sqrt{n}\,(d_n - \theta) \to \mathrm{n}[0, v(\theta)]$ in distribution.

Solution. In both cases the estimator agrees with a single one of its two branches with probability tending to 1, and that branch is exactly normal.
Write $A_n = \{|\bar{X}| < n^{-1/4}\}$, and recall $\sqrt{n}(\bar{X} - \theta) \sim \mathrm{n}(0, 1)$ exactly, so that $\sqrt{n}\,\bar{X} \sim \mathrm{n}(\sqrt{n}\theta, 1)$. The only tool needed is that if $P(B_n) \to 1$ and $Y_n = Y_n'$ on B_n, then $|P(Y_n \leq x) - P(Y_n' \leq x)| \leq P(B_n^c) \to 0$, so Y_n and Y_n' have the same limiting distribution.
(i) $\theta \neq 0$. With $Z \sim \mathrm{n}(0, 1)$,

$$P(A_n) = P\left(\left|Z + \sqrt{n}\theta\right| < n^{1/4}\right) \leq P\Big(|Z| > \sqrt{n}|\theta| - n^{1/4}\Big) \longrightarrow 0,$$

since $\sqrt{n}|\theta| - n^{1/4} = \sqrt{n}\left(|\theta| - n^{-1/4}\right) \to \infty$. Hence $d_n = \bar{X}$ on A_n^c, an event of probability tending to 1, and

$$\sqrt{n}\,(d_n - \theta) \longrightarrow \mathrm{n}[0, 1], \qquad v(\theta) = 1.$$

(ii) $\theta = 0$. Now $\sqrt{n}\,\bar{X} = Z$ and

$$P(A_n^c) = P\left(|Z| \geq n^{1/4}\right) \longrightarrow 0,$$

so $d_n = a\bar{X}$ with probability tending to 1, and on that event $\sqrt{n}\,(d_n - 0) = a\sqrt{n}\,\bar{X} = aZ$. Hence

$$\sqrt{n}\,(d_n - 0) \longrightarrow \mathrm{n}[0, a^2], \qquad v(0) = a^2.$$

Problem 10.13. Refer to Example 10.1.19, in which the $n^n = 4^4 = 256$ resamples drawn with replacement from the sample $2, 4, 9, 12$ are used to form the nonparametric bootstrap estimate

$$\mathrm{Var}^*(\bar{X}) = \frac{1}{n^n - 1} \sum_{i=1}^{n^n} (\bar{x}_i^* - \bar{\bar{x}}^*)^2\,, \qquad \bar{\bar{x}}^* = \frac{1}{n^n} \sum_{i=1}^{n^n} \bar{x}_i^*.$$

(a) Verify that the bootstrap mean and variance of the sample $2, 4, 9, 12$ are 6.75 and 3.94, respectively.
(b) Verify that 6.75 is the mean of the original sample.
(c) Verify that, when we divide by n instead of $n - 1$, the bootstrap variance of the mean, and the usual estimate of the variance of the mean are the same.
(d) Show how to calculate the bootstrap mean and standard error using the $\binom{4+4-1}{4} = 35$ distinct possible resamples.
(e) Establish parts (b) and (c) for a general sample $X_1, X_2, \ldots, X_n$.

Solution. (a) $\bar{\bar{x}}^* = 6.75$ and $\mathrm{Var}^*(\bar{X}) = 3.9373$. From $\bar{x} = 27/4 = 6.75$ and

$$\sum_i (x_i - \bar{x})^2 = (-4.75)^2 + (-2.75)^2 + (2.25)^2 + (5.25)^2 = 62.75,$$

part (e) below gives $\frac{1}{n^n} \sum_i (\bar{x}_i^* - 6.75)^2 = \frac{62.75}{16} = 3.921875$, and the divisor $n^n - 1 = 255$ in (10.1.9)

inflates this by 256/255:

$$\mathrm{Var}^*(\bar X) = \frac{256}{255}(3.921875) = 3.9373 \approx 3.94.$$

(b) $\bar x = \frac{2+4+9+12}{4} = \frac{27}{4} = 6.75$, the value in part (a).
(c) $\frac{1}{n^n}\sum_i (\bar x_i^* - \bar x)^2 = 3.921875$, and the usual estimate of $\mathrm{Var}(\bar X) = \sigma^2/n$ with σ^2 estimated by $\hat\sigma^2 = \frac{1}{n}\sum_i (x_i - \bar x)^2 = 62.75/4 = 15.6875$ is $\hat\sigma^2/4 = 3.921875$ as well.
(d) Index a resample by its count vector $\mathbf{k} = (k_1, k_2, k_3, k_4)$, $\sum_j k_j = 4$, of which there are $\binom{4+4-1}{4} = 35$; the resample mean is $\bar x^*(\mathbf{k}) = \frac{1}{4}\sum_j k_j x_j$ and, because the 256 ordered resamples are equally likely, $\mathbf{k}$ is multinomial with 4 trials and equal cell probabilities $\frac{1}{4}$:

$$p(\mathbf{k}) = \frac{1}{256}\binom{4}{k_1, k_2, k_3, k_4}.$$

The 35 vectors fall into five types, and the weights account for all 256 ordered resamples:

type of $\mathbf{k}$	number of $\mathbf{k}$	$\binom{4}{\mathbf{k}}$	total
$(4,0,0,0)$	4	1	4
$(3,1,0,0)$	12	4	48
$(2,2,0,0)$	6	6	36
$(2,1,1,0)$	12	12	144
$(1,1,1,1)$	1	24	24

Then

$$\bar{\bar x}^* = \sum_{\mathbf{k}} p(\mathbf{k})\,\bar x^*(\mathbf{k}) = 6.75, \qquad \sum_{\mathbf{k}} p(\mathbf{k})\,(\bar x^*(\mathbf{k}) - 6.75)^2 = 3.921875,$$

so the bootstrap standard error is $\sqrt{3.921875} = 1.980$ (or $\sqrt{(256/255)(3.921875)} = 1.984$ with the divisor of (10.1.9)). The 35 distinct resamples are not equiprobable, which is exactly why they cannot be treated as a random sample.
(e) A resample is $X_1^*, \ldots, X_n^*$ iid from the empirical distribution $\hat F$, which puts mass $1/n$ at each x_j; averaging a function over all n^n ordered resamples is precisely taking E^*. Hence

$$\mathrm{E}^* X_1^* = \frac{1}{n}\sum_j x_j = \bar x, \qquad \mathrm{Var}^* X_1^* = \frac{1}{n}\sum_j (x_j - \bar x)^2 = \hat\sigma^2,$$

and since the X_j^* are independent,

$$\frac{1}{n^n}\sum_{i=1}^{n^n} \bar x_i^* = \mathrm{E}^*\bar X^* = \bar x,$$

which is (b), and

$$\frac{1}{n^n}\sum_{i=1}^{n^n} (\bar x_i^* - \bar x)^2 = \mathrm{Var}^*\bar X^* = \frac{\hat\sigma^2}{n} = \frac{1}{n^2}\sum_j (x_j - \bar x)^2,$$

which is (c): the bootstrap variance of the mean computed with divisor n^n is identical to $\hat\sigma^2/n$, the usual variance estimate of $\bar X$ formed with divisor n.

Problem 10.14. In each of the following situations we will look at the parametric and nonparametric bootstrap. Compare the estimates, and discuss advantages and disadvantages of the methods.
(a) Referring to Example 10.1.22, estimate the variance of S^2 using a nonparametric bootstrap. (There the sample of size $n = 9$ is

-1.81 0.63 2.22 2.41 2.95 4.16 4.24 4.53 5.09

with $\bar x = 2.71$ and $s^2 = 4.82$; a parametric bootstrap drawing $B = 1000$ samples from $\mathrm{n}(2.71, 4.82)$ gave $\mathrm{Var}_B^*(S^2) = 4.33$, against the normal-theory plug-in $2(4.82)^2/8 = 5.81$ and the true value $2(4)^2/8 = 4.00$, the data having been generated with variance 4.)
(b) In Example 5.6.6 we essentially did a parametric bootstrap of the distribution of S^2 from a Poisson

sample. Use the nonparametric bootstrap to provide an alternative histogram of the distribution. (The data there are the bay anchovy larvae counts

19 32 29 13 8 12 16 20 14 17 22 18 23

with $n = 13$, $\bar{x} = 18.69$ and $s^2 = 44.90$; of 5000 simulated Poisson(18.69) samples of size 13, only 27 gave $S^2 > 44.90$, so $P(S^2 > 44.90 \,|\, \lambda = 18.69) \approx .0054$.)
(c) In Example 10.1.18 we looked at the problem of estimating a gamma mean. Suppose that we have a random sample

0.28 0.98 1.36 1.38 2.4 7.42

from a gamma(α, β) distribution. Estimate the mean and variance of the distribution using maximum likelihood and bootstrapping.

Solution. (a) $\mathrm{Var}^*(S^2) = 4.52$. Resampling is unnecessary here: the resamples are iid from $\hat{F}$, so the usual identity for the variance of S^2 applied to the empirical central moments $m_2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = 4.2850$ and $m_4 = \frac{1}{n}\sum(x_i - \bar{x})^4 = 54.461$ gives, with $n = 9$,

$$\mathrm{Var}^*(S^{2*}) = \frac{1}{n}\left[m_4 - \frac{n-3}{n-1}m_2^2\right] = \frac{54.461 - 0.75(4.2850)^2}{9} = 4.52,$$

which $B = 10{,}000$ resamples reproduce as 4.5. Lining the four numbers up,

method	estimate of $\mathrm{Var}(S^2)$
nonparametric bootstrap	4.52
parametric bootstrap (10.1.22)	4.33
normal-theory plug-in $2s^4/8$	5.81
true value $2\sigma^4/8$	4.00

both bootstraps beat the plug-in, and the parametric one is the better of the two – as it should be, since the normal model it assumes is the model that actually generated the data. The nonparametric bootstrap pays for its freedom from that assumption by resampling a distribution whose variance is $\mathrm{E}^* S^{2*} = \frac{n-1}{n}s^2 = 4.29$ rather than 4.82, and by the extra Monte Carlo noise of a finite B.
(b) Resampling with replacement from the 13 counts gives a histogram of S^{2*} centred near $\mathrm{E}^* S^{2*} = \frac{12}{13}(44.90) = 41.44$; below, $B = 200{,}000$ resamples, against the same number of Poisson(18.69) samples:

statistic of S^{2*}	nonparametric	parametric Poisson
mean	41.44	18.69
standard deviation	15.61	7.72
5th percentile	16.8	8.1
median	40.8	17.6
95th percentile	68.3	32.9
$P(S^{2*} > 44.90)$	.40	.005

The two histograms sit in completely different places, and this is the whole lesson. The parametric bootstrap draws from Poisson(18.69), so it describes the distribution of S^2 under the Poisson model, and against it the observed 44.90 is extreme – the evidence against the Poisson assumption reported in Example 5.6.6. The nonparametric bootstrap draws from the empirical distribution, which already carries the observed overdispersion, so it is centred at 44.90 (up to the $\frac{n-1}{n}$ factor) and 44.90 is an unremarkable value in it. The nonparametric histogram is the honest description of the sampling variability of S^2 when no model is assumed; it is useless for checking the Poisson assumption, which is what the parametric bootstrap is for.
(c) MLE: $\hat{\mu} = 2.303$ and $\widehat{\mathrm{Var}}\, X = 4.367$. With $n = 6$, $\bar{x} = 2.3033$ and $\overline{\log x} = 0.36934$, the likelihood equations reduce to

$$\log\hat{\alpha} - \psi(\hat{\alpha}) = \log\bar{x} - \overline{\log x} = 0.46502, \qquad \hat{\beta} = \bar{x}/\hat{\alpha},$$

solved numerically by $\hat{\alpha} = 1.2148$, $\hat{\beta} = 1.8960$. Then $\hat{\mu} = \hat{\alpha}\hat{\beta} = \bar{x} = 2.3033$ (as it must be, by Exercise 10.11(b)) and $\hat{\alpha}\hat{\beta}^2 = 4.3672$. For comparison the moment estimates of Exercise 10.11(d), built from $m_2 = 5.6295$, are $\tilde{\beta} = m_2/\bar{x} = 2.4441$ and $\tilde{\alpha} = \bar{x}/\tilde{\beta} = 0.9424$, while $s^2 = 6.755$.
Bootstrapping with $B = 30{,}000$:

quantity	parametric	nonparametric
estimate of the mean	2.303	2.303
standard error of the mean	0.85	0.97
estimate of the variance	4.367	6.755
standard error of variance	4.26	4.06

The parametric resamples are drawn from gamma$(1.2148, 1.8960)$ and the MLE is recomputed on each; its standard error 0.85 agrees with the asymptotic value $\sqrt{\hat\alpha\hat\beta^2/n} = \sqrt{4.3672/6} = 0.853$. The nonparametric resamples give a slightly larger standard error for the mean and, for the variance, a much larger point estimate, because s^2 is dominated by the single observation 7.42 while the fitted gamma spreads that information over the whole right tail. With $n = 6$ the parametric bootstrap is the more stable of the two – it borrows strength from the assumed shape – but it is only as good as the gamma assumption, which six observations cannot check; the nonparametric bootstrap makes no such commitment but, resampling from six points, has very little to work with.

10.3 Exercises 10.15–10.21

Problem 10.15. (a) Show that $\mathrm{Var}^*_B(\hat\theta)$ of (10.1.11) converges to $\mathrm{Var}^*(\hat\theta)$ of (10.1.10) as $B \to \infty$. (b) For fixed B and $i = 1, 2, \ldots$, calculate the bootstrap variance $\mathrm{Var}^*_{B_i}(\hat\theta)$. Use the Law of Large Numbers to show that $(1/m)\sum_{i=1}^m \mathrm{Var}^*_{B_i}(\hat\theta) \to \mathrm{Var}^*(\hat\theta)$ as $m \to \infty$.
Here, for a sample $\mathbf{x} = (x_1, \ldots, x_n)$ and estimator $\hat\theta(\mathbf{x})$,

$$\mathrm{Var}^*(\hat\theta) \;=\; \frac{1}{n^n - 1}\sum_{i=1}^{n^n}\big(\hat\theta^*_i - \bar{\hat\theta}^*\big)^2 \qquad (10.1.10)$$

is computed from all n^n resamples, while

$$\mathrm{Var}^*_B(\hat\theta) \;=\; \frac{1}{B - 1}\sum_{i=1}^{B}\big(\hat\theta^*_i - \bar{\hat\theta}^*\big)^2 \qquad (10.1.11)$$

is computed from B resamples drawn at random.

Solution. Both limits are the Strong Law (Theorem 5.5.9) applied to iid draws from the bootstrap distribution. With the data $\mathbf{x}$ held fixed, the n^n resamples are equally likely, so a resample produces $\hat\theta^*$ carrying mass n^{-n} at each of $\hat\theta^*_1, \ldots, \hat\theta^*_{n^n}$; writing E_*, Var_* for moments under it,

$$E_*\hat\theta^* = \frac{1}{n^n}\sum_{i=1}^{n^n}\hat\theta^*_i = \bar{\hat\theta}^*,$$

$$\mathrm{Var}_*(\hat\theta^*) = \frac{1}{n^n}\sum_{i=1}^{n^n}\big(\hat\theta^*_i - \bar{\hat\theta}^*\big)^2 = \Big(1 - \tfrac{1}{n^n}\Big)\,\mathrm{Var}^*(\hat\theta).$$

The support is finite, so all moments exist and the Strong Law applies below.
(a) The B resamples are drawn independently, so $\hat\theta^*_1, \ldots, \hat\theta^*_B$ are iid from this distribution. In moment form, applying Theorem 5.5.9 to the first and second sample moments,

$$\begin{aligned}\mathrm{Var}^*_B(\hat\theta) &= \frac{B}{B-1}\Big[\frac{1}{B}\sum_{i=1}^{B}(\hat\theta^*_i)^2 - \Big(\frac{1}{B}\sum_{i=1}^{B}\hat\theta^*_i\Big)^2\Big] \\ &\xrightarrow[B\to\infty]{\text{a.s.}} E_*(\hat\theta^*)^2 - \big(E_*\hat\theta^*\big)^2 = \mathrm{Var}_*(\hat\theta^*),\end{aligned}$$

since $B/(B-1) \to 1$ and $t \mapsto t^2$ is continuous. So the limit is $(1 - n^{-n})\,\mathrm{Var}^*(\hat\theta)$, the factor being only the divisor $n^n - 1$ against n^n in (10.1.10); reading (10.1.10) as the variance of the bootstrap distribution, the limit is exactly $\mathrm{Var}^*(\hat\theta)$.

(b) Run the B-resample bootstrap independently m times, giving $\mathrm{Var}^*_{B_1}(\hat\theta), \mathrm{Var}^*_{B_2}(\hat\theta), \ldots$, iid across i. Each is the sample variance of B iid draws, hence unbiased for their variance (Theorem 5.2.6),

$$E\big[\mathrm{Var}^*_{B_i}(\hat\theta)\big] = \mathrm{Var}_*(\hat\theta^*),$$

and bounded (a function of finitely many bounded variables), so the Strong Law applies to $\{\mathrm{Var}^*_{B_i}(\hat\theta)\}$ at any fixed $B \geq 2$:

$$\frac{1}{m}\sum_{i=1}^{m} \mathrm{Var}^*_{B_i}(\hat\theta) \xrightarrow[m\to\infty]{\text{a.s.}} \mathrm{Var}_*(\hat\theta^*) = \Big(1 - \tfrac{1}{n^n}\Big)\,\mathrm{Var}^*(\hat\theta).$$

Problem 10.16. For the situation of Example 10.1.21, if we observed that $\hat p = 1/2$, we might use a variance estimate from Theorem 5.5.26. Show that this variance estimate would be equal to $2[\mathrm{Var}(\hat p)]^2$.
(a) If we observe $\hat p = 11/24$, verify that this variance estimate is .00007.
(b) Using simulation, calculate the "exact variance" of $\hat p(1-\hat p)$ when $n = 24$ and $p = 11/24$. Verify that it is equal to .00529.
(c) Why do you think the Delta Method is so bad in this case? Might the second-order Delta Method do any better? What about the bootstrap estimate?

Solution. At $p = 1/2$ the second-order Delta Method returns exactly $2[\mathrm{Var}(\hat p)]^2 = 1/(8n^2)$.
Take $g(p) = p(1-p)$, so $g'(p) = 1-2p$ and $g''(p) = -2$. At $p = 1/2$ we have $g'(1/2) = 0$ and $g''(1/2) = -2 \neq 0$, which is precisely the hypothesis of Theorem 5.5.26; and $\sqrt{n}(\hat p - p) \to \mathrm{n}(0, \sigma^2)$ with $\sigma^2 = p(1-p) = 1/4$. Hence (5.5.13) gives

$$n\big[\hat p(1-\hat p) - \tfrac14\big] \;\to\; \sigma^2 \frac{g''(1/2)}{2}\chi^2_1 = -\tfrac14\chi^2_1.$$

Since $\mathrm{Var}\,\chi^2_1 = 2$,

$$\mathrm{Var}\big(\hat p(1-\hat p)\big) \approx \frac{1}{n^2}\cdot\frac{1}{16}\cdot 2 = \frac{1}{8n^2} = 2\Big[\frac{(1/2)(1/2)}{n}\Big]^2 = 2\big[\mathrm{Var}(\hat p)\big]^2.$$

(a) $\hat p = 11/24 \neq 1/2$, so by the prescription of Example 10.1.21 the **first-order** estimate of Example 10.1.15 is the one in force:

$$\widehat{\mathrm{Var}}\big(\hat p(1-\hat p)\big) = \frac{\hat p(1-\hat p)(1-2\hat p)^2}{n} = \frac{\frac{11}{24}\cdot\frac{13}{24}\cdot\big(\frac{1}{12}\big)^2}{24} = \frac{143}{1{,}990{,}656} = .0000718,$$

which rounds to the printed .00007. (The second-order value $2[\widehat{\mathrm{Var}}(\hat p)]^2 = 2(143/13824)^2 = .000214$ is a different number, so .00007 can only be the first-order estimate.)
(b) Writing $u = \hat p - p$ and $q = 1-p$, $g(\hat p) = pq + (1-2p)u - u^2$ exactly, and the binomial central moments $Eu^2 = pq/n$, $Eu^3 = pq(q-p)/n^2$, $Eu^4 = 3p^2q^2/n^2 + pq(1-6pq)/n^3$ give the closed form

$$\mathrm{Var}\big(\hat p(1-\hat p)\big) = \frac{(1-2p)^2pq}{n}\Big(1-\frac{2}{n}\Big) + \frac{2p^2q^2}{n^2} + \frac{pq(1-6pq)}{n^3}.$$

At $n = 24$, $p = 11/24$ the three terms are .0000658, .0002140, −.0000088, so

$$\mathrm{Var}\big(\hat p(1-\hat p)\big) = .000271.$$

Direct summation of $\sum_{k=0}^{24}\big(g(k/24) - Eg\big)^2\binom{24}{k}p^kq^{24-k}$ returns .00027107 and 2×10^6 simulated binomials return .000272, so the printed .00529 is a book erratum – as are the True and Bootstrap rows of Table 10.1.1, whose Delta Method row alone reproduces.
(c) Because $g'(p) = 1-2p$ nearly vanishes at $p = 11/24$: the first-order term $(1-2p)^2pq/n = .0000718$ that the Delta Method keeps is dwarfed by the term it discards, $2p^2q^2/n^2 = 2[\mathrm{Var}(\hat p)]^2 = .000214$,

three times larger. So the estimate is low by a factor of 3.8 here, and by a factor of ∞ at p exactly $1/2$.
Yes, the second-order Delta Method does much better: adding that term gives $.0000718 + .000214 = .000286$ against the exact .000271, an error of 6% rather than 74%. The bootstrap needs no repair at all, since it measures the spread of $\hat{p}^*(1-\hat{p}^*)$ directly rather than through $g'(\hat{p})$.

Problem 10.17. Efron (1982) analyzes data on law school admission, with the object being to examine the correlation between the LSAT (Law School Admission Test) score and the first-year GPA (grade point average). For each of 15 law schools, we have the pair of data points (average LSAT, average GPA):

(576, 3.39) (635, 3.30) (558, 2.81) (578, 3.03) (666, 3.44)
(580, 3.07) (555, 3.00) (661, 3.43) (651, 3.36) (605, 3.13)
(653, 3.12) (575, 2.74) (545, 2.76) (572, 2.88) (594, 2.96)

(a) Calculate the correlation coefficient between LSAT score and GPA.
(b) Use the nonparametric bootstrap to estimate the standard deviation of the correlation coefficient. Use $B = 1000$ resamples, and also plot them in a histogram.
(c) Use the parametric bootstrap to estimate the standard deviation of the correlation coefficient. Assume that (LSAT, GPA) has a bivariate normal distribution, and estimate the five parameters. Then generate 1000 samples of 15 pairs from this bivariate normal distribution.
(d) If (X, Y) are bivariate normal with correlation coefficient ρ and sample correlation r, then the Delta Method can be used to show that

$$\sqrt{n}\,(r-\rho) \;\to\; \mathrm{n}\big(0, (1-\rho^2)^2\big).$$

Use this fact to estimate the standard deviation of r. How does it compare to the bootstrap estimates? Draw an approximate pdf of r.
(e) Fisher's z-transformation is a variance-stabilizing transformation for the correlation coefficient (see Exercise 11.4). If (X, Y) are bivariate normal with correlation coefficient ρ and sample correlation r, then

$$\frac{1}{2}\Big[\log\Big(\frac{1+r}{1-r}\Big) - \log\Big(\frac{1+\rho}{1-\rho}\Big)\Big]$$

is approximately normal. Use this fact to draw an approximate pdf of r.
(Establishing the normality result in part (d) involves some tedious matrix calculations; see Lehmann and Casella 1998, Example 6.5. The z-transformation of part (e) yields faster convergence to normality than the Delta Method of part (d). Diaconis and Holmes 1994 do an exhaustive bootstrap for this problem, enumerating all 77,558,760 correlation coefficients.)

Solution. (a) $r = .776$. From the 15 pairs, $\bar{x} = 600.27$, $\bar{y} = 3.0947$, and

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{110.62}{\sqrt{24{,}455 \times .8302}} = .7764.$$

(b) $\widehat{\mathrm{sd}}(r) = .13$. Resample 15 pairs with replacement from the 15 observed pairs (pairs kept intact, so the dependence is resampled), $B = 1000$ times, recomputing r each time; this gives $\mathrm{Var}_B^*(r)^{1/2} = .125$ (repeated runs give .12 to .14). The histogram of the 1000 values r^* is strongly left-skewed, piling up against the boundary $r = 1$:

bin	count
$[-1, .3)$	2
$[.3, .4)$	4
$[.4, .5)$	26
$[.5, .6)$	60
$[.6, .7)$	165
$[.7, .8)$	266
$[.8, .9)$	304
$[.9, 1.0]$	173

with median .79 and a 2.5%–97.5% spread of $(.49, .96)$.
(c) $\widehat{\mathrm{sd}}(r) = .12$. The five MLEs of the bivariate normal parameters (the variance MLEs use divisor n) are

$$\hat\mu_X = 600.27, \quad \hat\mu_Y = 3.0947,$$
$$\hat\sigma_X^2 = 1630.3, \quad \hat\sigma_Y^2 = .05534, \quad \hat\rho = .7764.$$

Drawing 1000 samples of $n = 15$ pairs from $\mathrm{n}_2(\hat\mu, \hat\Sigma)$ and recomputing r gives $\mathrm{Var}_B^*(r)^{1/2} = .117$ (runs give .11 to .13), with the same left skew (median .79, 2.5%–97.5% spread $(.48, .92)$).
(d) $\widehat{\mathrm{sd}}(r) = .103$. Substituting $\hat\rho = r$ in the limit variance,

$$\widehat{\mathrm{sd}}(r) = \frac{1 - r^2}{\sqrt{n}} = \frac{1 - (.7764)^2}{\sqrt{15}} = \frac{.3972}{3.873} = .1026.$$

This sits 10% to 20% below both bootstrap estimates, and the approximate pdf it produces,

$$r \sim \mathrm{n}\big(.776, (.1026)^2\big),$$

is symmetric about .776 and assigns probability .014 to the impossible region $r > 1$. That is its defect: it cannot see the boundary.
(e) With $\zeta(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho} = \tanh^{-1}\rho$,

$$\zeta(r) \;\approx\; \mathrm{n}\Big(\zeta(\rho), \frac{1}{n-3}\Big), \qquad \zeta(r) = \tanh^{-1}(.7764) = 1.036, \quad \frac{1}{\sqrt{12}} = .2887.$$

(The Delta Method applied to (d) with $\zeta'(\rho) = 1/(1-\rho^2)$ gives variance $(1-\rho^2)^{-2}(1-\rho^2)^2/n = 1/n$ exactly – the stabilization – and $1/(n-3)$ is the standard refinement.) Inverting $r = \tanh\zeta$ gives the approximate pdf of r on $(-1, 1)$:

$$f(r) = \frac{\sqrt{n-3}}{\sqrt{2\pi}\,(1-r^2)} \exp\Big\{-\frac{n-3}{2}\big(\tanh^{-1} r - 1.036\big)^2\Big\},$$

a left-skewed density supported on $(-1, 1)$, peaking near $r = .83$, with central 95% interval $\tanh(1.036 \pm 1.96(.2887)) = (.44, .92)$ and implied standard deviation $(1-r^2)/\sqrt{n-3} = .115$.

Problem 10.18. For the situation of Exercise 10.2.1, that is, if $X_1, X_2, \ldots, X_n$ are iid, where $X_i \sim \mathrm{n}(\mu, \sigma^2)$ with probability $1 - \delta$ and $X_i \sim f(x)$ with probability δ, where $f(x)$ is any density with mean θ and variance τ^2, show that

$$\mathrm{Var}(\bar X) = (1-\delta)\frac{\sigma^2}{n} + \delta\frac{\tau^2}{n} + \frac{\delta(1-\delta)(\theta-\mu)^2}{n}.$$

Also deduce that contamination with a Cauchy pdf will always result in an infinite variance. (Hint: Write this mixture model as a hierarchical model. Let $Y = 0$ with probability $1 - \delta$ and $Y = 1$ with probability δ. Then $\mathrm{Var}(X_i) = E[\mathrm{Var}(X_i|Y)] + \mathrm{Var}(E[X_i|Y])$.)

Solution. The conditional variance identity (Theorem 4.4.7) gives it in one line; the reference is to Example 10.2.1, the δ-contamination model. In the hint's hierarchy $Y \sim \text{Bernoulli}(\delta)$, $X_i|Y=0 \sim \text{n}(\mu,\sigma^2)$, $X_i|Y=1 \sim f$, the two conditional moments are the two-valued random variables

$$E[X_i|Y] = \begin{cases} \mu, & Y=0 \\ \theta, & Y=1 \end{cases} \qquad \text{Var}(X_i|Y) = \begin{cases} \sigma^2, & Y=0 \\ \tau^2, & Y=1, \end{cases}$$

so

$$\begin{aligned} \text{Var}(X_i) &= E\big[\text{Var}(X_i|Y)\big] + \text{Var}\big(E[X_i|Y]\big) \\ &= \big[(1-\delta)\sigma^2 + \delta\tau^2\big] + \delta(1-\delta)(\theta-\mu)^2, \end{aligned}$$

the last term because a random variable taking μ with probability $1-\delta$ and θ with probability δ has variance $\delta(1-\delta)(\theta-\mu)^2$. The X_i are iid, so $\text{Var}(\bar X) = \text{Var}(X_1)/n$, which is the displayed formula.
For Cauchy contamination the formula is unavailable because $\tau^2=\infty$, and the conclusion is that $\bar X$ has infinite second moment for every $\delta > 0$. Condition on $\mathbf{Y} = (Y_1,\dots,Y_n)$: given $\mathbf{Y}$ the X_i are independent, and on the event $A = \{\text{at least one } Y_i = 1\}$ the conditional law of $n\bar X$ is the convolution of a normal and at least one Cauchy, hence itself has Cauchy tails, so

$$E\big[\bar X^2 \,\big|\, \mathbf{Y}\big] = \infty \quad \text{on } A, \qquad P(A) = 1-(1-\delta)^n > 0.$$

Therefore $E[\bar X^2] = E\big[E[\bar X^2|\mathbf{Y}]\big] = \infty$, and the same computation with $\bar X$ replaced by $\bar X - a$ shows the second moment about any center is infinite, so $\text{Var}(\bar X) = \infty$.

Problem 10.19. Another way in which underlying assumptions can be violated is if there is correlation in the sampling, which can seriously affect the properties of the sample mean. Suppose we introduce correlation in the case discussed in Exercise 10.2.1; that is, we observe $X_1,\dots,X_n$, where $X_i \sim \text{n}(\theta,\sigma^2)$, but the X_is are no longer independent.
(a) For the equicorrelated case, that is, $\text{Corr}(X_i,X_j) = \rho$ for $i \neq j$, show that

$$\text{Var}(\bar X) = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2,$$

so $\text{Var}(\bar X) \not\to 0$ as $n \to \infty$.
(b) If the X_is are observed through time (or distance), it is sometimes assumed that the correlation decreases with time (or distance), with one specific model being $\text{Corr}(X_i,X_j) = \rho^{|i-j|}$. Show that in this case

$$\text{Var}(\bar X) = \frac{\sigma^2}{n} + \frac{2\sigma^2}{n^2}\,\frac{\rho}{1-\rho}\Big(n - \frac{1-\rho^n}{1-\rho}\Big),$$

so $\text{Var}(\bar X) \to 0$ as $n \to \infty$. (See Miscellanea 5.8.2 for another effect of correlation.)
(c) The correlation structure in part (b) arises in an autoregressive AR(1) model, where we assume that $X_{i+1} = \rho X_i + \delta_i$, with δ_i iid $\text{n}(0,1)$. If $|\rho| < 1$ and we define $\sigma^2 = 1/(1-\rho^2)$, show that $\text{Corr}(X_1,X_i) = \rho^{i-1}$.

Solution. Everything comes from $\text{Var}(\bar X) = n^{-2}\big[n\sigma^2 + \sum_{i\neq j}\text{Cov}(X_i,X_j)\big]$ with $\text{Cov}(X_i,X_j) = \sigma^2\,\text{Corr}(X_i,X_j)$.
(a) There are $n(n-1)$ ordered pairs $i \neq j$, each contributing $\rho\sigma^2$:

$$\text{Var}(\bar X) = \frac{1}{n^2}\big[n\sigma^2 + n(n-1)\rho\sigma^2\big] = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2 \;\longrightarrow\; \rho\sigma^2 \neq 0.$$

(b) Group the off-diagonal terms by the lag $k = |i-j|$; lag k occurs in $2(n-k)$ ordered pairs, so

$$\text{Var}(\bar X) = \frac{\sigma^2}{n} + \frac{2\sigma^2}{n^2}\sum_{k=1}^{n-1}(n-k)\rho^k.$$

Since $n-k=\#\{m: k\le m\le n-1\}$, interchanging the order of summation gives

$$\sum_{k=1}^{n-1}(n-k)\rho^k=\sum_{m=1}^{n-1}\sum_{k=1}^{m}\rho^k=\frac{\rho}{1-\rho}\sum_{m=1}^{n-1}\bigl(1-\rho^m\bigr)$$
$$=\frac{\rho}{1-\rho}\Bigl[(n-1)-\sum_{m=1}^{n-1}\rho^m\Bigr]=\frac{\rho}{1-\rho}\Bigl(n-\frac{1-\rho^n}{1-\rho}\Bigr),$$

the last step because $1+\sum_{m=1}^{n-1}\rho^m=(1-\rho^n)/(1-\rho)$. This is the stated formula. As $n\to\infty$ the bracket is $n+O(1)$, so

$$\operatorname{Var}(\bar X)=\frac{\sigma^2}{n}+\frac{2\sigma^2\rho}{n(1-\rho)}+O(n^{-2})=\frac{\sigma^2}{n}\cdot\frac{1+\rho}{1-\rho}+O(n^{-2})\ \to\ 0.$$

(c) The AR(1) recursion is variance-stationary at $\sigma^2=1/(1-\rho^2)$:

$$\operatorname{Var}(X_{i+1})=\rho^2\operatorname{Var}(X_i)+1=\frac{\rho^2}{1-\rho^2}+1=\frac{1}{1-\rho^2}=\sigma^2,$$

so if $\operatorname{Var}(X_1)=\sigma^2$ then $\operatorname{Var}(X_i)=\sigma^2$ for all i by induction. Iterating the recursion,

$$X_i=\rho^{\,i-1}X_1+\sum_{j=1}^{i-1}\rho^{\,i-1-j}\delta_j,$$

and the δ_j, $j\ge 1$, are independent of X_1, so all the sum's covariances with X_1 vanish:

$$\operatorname{Cov}(X_1,X_i)=\rho^{\,i-1}\operatorname{Var}(X_1)=\rho^{\,i-1}\sigma^2,\qquad \operatorname{Corr}(X_1,X_i)=\frac{\rho^{\,i-1}\sigma^2}{\sigma\cdot\sigma}=\rho^{\,i-1}.$$

Problem 10.20. Refer to Definition 10.2.2 about breakdown values: for an ordered sample $X_{(1)}<\cdots<X_{(n)}$ and a statistic T_n based on it, T_n has breakdown value b, $0\le b\le 1$, if for every $\epsilon>0$,

$$\lim_{X_{(\{(1-b)n\})}\to\infty}T_n<\infty,\qquad \lim_{X_{(\{(1-(b+\epsilon))n\})}\to\infty}T_n=\infty,$$

where $\{\cdot\}$ in a subscript is rounding to the nearest integer (Definition 5.4.2).
(a) If $T_n=\bar X_n$, the sample mean, show that $b=0$.
(b) If $T_n=M_n$, the sample median, show that $b=.5$.
An estimator that "splits the difference" between the mean and the median in terms of sensitivity is the α-trimmed mean, $0<\alpha<\frac12$, defined as follows. $\bar X_n^{\alpha}$, the α-trimmed mean, is computed by deleting the αn smallest observations and the αn largest observations, and taking the arithmetic mean of the remaining observations.
(c) If $T_n=\bar X_n^{\alpha}$, the α-trimmed mean of the sample, $0<\alpha<\frac12$, show that $0<b<\frac12$.

Solution. Write "corrupting a fraction γ" for driving every order statistic of rank exceeding $(1-\gamma)n$ to infinity with the rest held fixed; in Definition 10.2.2 the breakdown value b is the supremum of the γ that leave T_n bounded.
(a) $b=0$. Corrupting $\gamma=0$ moves nothing, so $\bar X_n<\infty$. For any $\gamma>0$ at least the top order statistic is corrupted, say $X_{(n)}=x\to\infty$, and

$$\bar X_n=\frac{x}{n}+\frac{1}{n}\sum_{i=1}^{n-1}X_{(i)}\ \longrightarrow\ \infty$$

because the remaining sum is fixed.
(b) $b=\frac12$. Take n odd, so $M_n=X_{((n+1)/2)}$. Two cases.

(i) $\gamma < \frac{1}{2}$: the corrupted ranks are those exceeding $(1-\gamma)n$, and $(n+1)/2 \le (1-\gamma)n$ whenever $\gamma \le \frac{1}{2} - \frac{1}{2n}$, so for all large n the median is an **uncorrupted** order statistic and M_n is unchanged, hence finite.
(ii) $\gamma > \frac{1}{2}$: now $(1-\gamma)n < n/2 < (n+1)/2$, so M_n is itself one of the corrupted observations and $M_n \to \infty$.
Hence $b = \frac{1}{2}$.
(c) $b = \alpha$, and $0 < \alpha < \frac{1}{2}$ gives $0 < b < \frac{1}{2}$ at once. The trimmed mean averages the order statistics of ranks $\alpha n + 1, \dots, n - \alpha n$.
(i) $\gamma < \alpha$: every corrupted observation has rank $> (1-\gamma)n > (1-\alpha)n$, so all of them lie among the αn largest and are trimmed away. $\bar{X}_n^{\alpha}$ is an average of fixed numbers, hence finite.
(ii) $\gamma > \alpha$: the largest retained order statistic $X_{(n-\alpha n)}$ has rank $n - \alpha n > (1-\gamma)n$, so it is corrupted, and

$$\bar{X}_n^{\alpha} \;\ge\; \frac{X_{(n-\alpha n)} + \big[(1-2\alpha)n - 1\big] X_{(\alpha n+1)}}{(1-2\alpha)n} \;\longrightarrow\; \infty,$$

since one term of the average tends to ∞ and each of the others is bounded below by the fixed $X_{(\alpha n+1)}$.

Problem 10.21. The breakdown performance of the mean and the median continues with their scale estimate counterparts. For a sample $X_1, \dots, X_n$:
(a) Show that the breakdown value of the sample variance $S^2 = \sum (X_i - \bar{X})^2/(n-1)$ is 0.
(b) A robust alternative is the *median absolute deviation*, or MAD, the median of $|X_1 - M|, |X_2 - M|, \dots, |X_n - M|$, where M is the sample median. Show that this estimator has a breakdown value of 50%.

Solution. (a) $b = 0$: one observation destroys S^2. Corrupt the single largest, $X_{(n)} = x \to \infty$, the others fixed with $\sum_{i<n} X_{(i)} = nc$. Then $\bar{X} = x/n + c$ and, keeping only the top term of the sum,

$$S^2 \;\ge\; \frac{(X_{(n)} - \bar{X})^2}{n-1} = \frac{1}{n-1}\Big(x\,\frac{n-1}{n} - c\Big)^2 \;\longrightarrow\; \infty.$$

Since corrupting a fraction 0 leaves S^2 finite, $b = 0$.
(b) $b = \frac{1}{2}$. Take n odd and corrupt a fraction γ.
(i) $\gamma < \frac{1}{2}$: by 10.20(b) the median M is an uncorrupted order statistic, so $M = m$ is fixed. The $n - \gamma n > n/2$ uncorrupted observations are fixed, so their deviations satisfy $|X_i - m| \le C$ for a constant C not depending on the corruption. Thus strictly more than half of the n numbers $|X_i - M|$ lie in $[0, C]$, and their median – the $\frac{n+1}{2}$-th smallest – is therefore at most C:

$$\text{MAD} \le C < \infty.$$

(ii) $\gamma > \frac{1}{2}$: let $k = \gamma n > n/2$ and drive the top k observations to infinity at separated rates, $X_{(n-k+j)} = jx$ with $x \to \infty$, $j = 1, \dots, k$. Now $M = X_{((n+1)/2)} = j_0 x$ for some $1 \le j_0 \le k$, and every deviation but one tends to infinity:

$$|X_i - M| = \begin{cases} |j - j_0|\, x \to \infty, & X_i = jx,\ j \ne j_0, \\ j_0 x - O(1) \to \infty, & X_i \text{ uncorrupted}, \\ 0, & j = j_0. \end{cases}$$

At most one of the n deviations stays bounded, so the median of them tends to ∞, and MAD $= \infty$. Hence the MAD tolerates any fraction below $\frac{1}{2}$ and breaks above it: $b = .5$.

10.4 Exercises 10.22–10.28

Problem 10.22. This exercise will look at some of the details of Example 10.2.3. There $X_1, \ldots, X_n$ is a sample from a population with pdf f and cdf F (assumed differentiable) with $P(X_i \le \mu) = 1/2$, so μ is the population median; M_n is the sample median; and for fixed a one sets

$$Y_i = \begin{cases} 1 & \text{if } X_i \le \mu + a/\sqrt{n}, \\ 0 & \text{otherwise,} \end{cases}$$

and $p_n = F(\mu + a/\sqrt{n})$.
(a) Verify that, if n is odd, then

$$P\left(\sqrt{n}(M_n - \mu) \le a\right) = P\left(\frac{\sum_i Y_i - np_n}{\sqrt{np_n(1-p_n)}} \ge \frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}}\right).$$

(b) Verify that $p_n \to p = F(\mu) = 1/2$ and

$$\frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}} \longrightarrow -2aF'(\mu) = -2af(\mu).$$

(**Hint:** Establish that $\frac{(n+1)/2 - np_n}{\sqrt{n}}$ is the limit form of a derivative.)
(c) Explain how to go from the statement that

$$P\left(\sqrt{n}(M_n - \mu) \le a\right) \to P\left(Z \ge -2af(\mu)\right)$$

to the conclusion that $\sqrt{n}(M_n - \mu)$ is asymptotically normal with mean 0 and variance $1/[2f(\mu)]^2$. (Note that the CLT would directly apply only if p_n did not depend on n. As it does, more work needs to be done to rigorously conclude limiting normality. When the work is done, the result is as expected.)

Solution. (a) For n odd, $M_n = X_{((n+1)/2)}$, and the median is at most t exactly when at least $(n+1)/2$ of the observations are at most t. With $t = \mu + a/\sqrt{n}$ this is the event $\{\sum_i Y_i \ge (n+1)/2\}$, so

$$P\left(\sqrt{n}(M_n - \mu) \le a\right) = P\left(\sum_i Y_i \ge \tfrac{n+1}{2}\right),$$

and subtracting np_n and dividing by $\sqrt{np_n(1-p_n)} > 0$ on both sides of the inner inequality gives the stated form. (The Y_i are iid Bernoulli(p_n), so $\mathrm{E}\sum Y_i = np_n$ and $\mathrm{Var}\sum Y_i = np_n(1-p_n)$.)
(b) F is differentiable, hence continuous, so $p_n = F(\mu + a/\sqrt{n}) \to F(\mu) = 1/2$. Split the ratio as

$$\frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}} = \frac{(n+1)/2 - np_n}{\sqrt{n}} \cdot \frac{1}{\sqrt{p_n(1-p_n)}},$$

and write the first factor, using $F(\mu) = 1/2$ (and $a \ne 0$, the case $a = 0$ being trivial), as

$$\begin{aligned} \frac{(n+1)/2 - np_n}{\sqrt{n}} &= \frac{1}{2\sqrt{n}} + \sqrt{n}\left(\tfrac{1}{2} - p_n\right) \\ &= \frac{1}{2\sqrt{n}} - a\,\frac{F(\mu + a/\sqrt{n}) - F(\mu)}{a/\sqrt{n}} \\ &\longrightarrow 0 - aF'(\mu) = -af(\mu), \end{aligned}$$

the difference quotient converging because F is differentiable at μ and $a/\sqrt{n} \to 0$. The second factor tends to $(1/4)^{-1/2} = 2$, so the product tends to $-2af(\mu)$.
(c) Read the right side as a normal cdf evaluated at a. (The given limit does need Lindeberg rather than Theorem 5.5.14, since p_n moves with n: the Y_i are uniformly bounded and $np_n(1-p_n) \sim n/4 \to \infty$,

so the standardized sum of (a) tends to Z, and (b) moves the cutoff.) By symmetry of $Z \sim \mathrm{n}(0,1)$,

$$P\left(Z \geq -2af(\mu)\right) = P\left(Z \leq 2af(\mu)\right) = P\left(\frac{Z}{2f(\mu)} \leq a\right),$$

and $Z/[2f(\mu)] \sim \mathrm{n}\big(0, 1/[2f(\mu)]^2\big)$. So the cdf of $\sqrt{n}(M_n - \mu)$ converges at every a to the (everywhere continuous) cdf of $\mathrm{n}(0, 1/[2f(\mu)]^2)$, which is precisely convergence in distribution (Definition 5.5.10):

$$\sqrt{n}(M_n - \mu) \to \mathrm{n}\left(0, \frac{1}{[2f(\mu)]^2}\right).$$

Problem 10.23. In this exercise we will further explore the ARE of the median to the mean, $\mathrm{ARE}(M_n, \bar{X})$.
(a) Verify the three AREs given in Example 10.2.4, namely the table

Distribution	Normal	Logistic	Double exponential
$\mathrm{ARE}(M_n, \bar{X})$	.64	.82	2

(b) Show that $\mathrm{ARE}(M_n, \bar{X})$ is unaffected by scale changes. That is, it does not matter whether the underlying pdf is $f(x)$ or $(1/\sigma)f(x/\sigma)$.
(c) Calculate $\mathrm{ARE}(M_n, \bar{X})$ when the underlying distribution is Student's t with ν degrees of freedom, for $\nu = 3, 5, 10, 25, 50, \infty$. What can you conclude about the ARE and the tails of the distribution?
(d) Calculate $\mathrm{ARE}(M_n, \bar{X})$ when the underlying pdf is the **Tukey model**

$$X \sim \begin{cases} \mathrm{n}(0,1) & \text{with probability } 1-\delta, \\ \mathrm{n}(0,\sigma^2) & \text{with probability } \delta. \end{cases}$$

Calculate the ARE for a range of δ and σ. What can you conclude about the relative performance of the mean and the median?

Solution. Throughout, for a symmetric density f with median (= mean) μ and variance $\sigma^2 < \infty$, Example 10.2.3 gives $\sqrt{n}(M_n - \mu) \to \mathrm{n}(0, 1/[2f(\mu)]^2)$ and the CLT gives $\sqrt{n}(\bar{X} - \mu) \to \mathrm{n}(0, \sigma^2)$, so by Definition 10.1.16

$$\mathrm{ARE}(M_n, \bar{X}) = \frac{\sigma^2}{1/[2f(\mu)]^2} = 4\sigma^2 f(\mu)^2.$$

(a) Each is one substitution into this formula.
(i) $\mathrm{n}(0,1)$: $\sigma^2 = 1$, $f(0) = 1/\sqrt{2\pi}$, so $\mathrm{ARE} = 4 \cdot \frac{1}{2\pi} = \frac{2}{\pi} = .6366$.
(ii) Logistic, $f(x) = e^{-x}/(1+e^{-x})^2$: $\sigma^2 = \pi^2/3$, $f(0) = 1/4$, so $\mathrm{ARE} = 4 \cdot \frac{\pi^2}{3} \cdot \frac{1}{16} = \frac{\pi^2}{12} = .8225$.
(iii) Double exponential, $f(x) = \frac{1}{2}e^{-|x|}$: $\sigma^2 = 2$, $f(0) = 1/2$, so $\mathrm{ARE} = 4 \cdot 2 \cdot \frac{1}{4} = 2$.
(b) If X has pdf f with median μ and variance σ_f^2, then σX has pdf $f_\sigma(x) = (1/\sigma)f(x/\sigma)$, median $\sigma\mu$, and variance $\sigma^2\sigma_f^2$, while $f_\sigma(\sigma\mu) = f(\mu)/\sigma$. Hence

$$4\left(\sigma^2\sigma_f^2\right)\left(\frac{f(\mu)}{\sigma}\right)^2 = 4\sigma_f^2 f(\mu)^2,$$

the same value as for f.
(c) For t_ν ($\nu > 2$), $\sigma^2 = \nu/(\nu-2)$ and

$$f(0) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}, \qquad \mathrm{ARE}(M_n, \bar{X}) = \frac{4\nu}{\nu-2}\, f(0)^2.$$

ν	3	5	10	25	50	∞
ARE	1.62	.96	.76	.68	.66	.64

($\nu = \infty$ is the normal value $2/\pi$.) The ARE decreases monotonically to $2/\pi$ as ν increases: the heavier the tails, the better the median does relative to the mean, and for $\nu \leq 4$ (the crossover lies between $\nu = 4$, ARE $= 1.13$, and $\nu = 5$) the median is actually the more efficient estimator.

(d) The mixture is symmetric about 0, with

$$\sigma^2 = (1-\delta) + \delta\sigma_*^2, \qquad f(0) = \frac{1}{\sqrt{2\pi}}\left[(1-\delta) + \frac{\delta}{\sigma_*}\right],$$

writing σ_* for the contaminating scale. Therefore

$$\mathrm{ARE}(M_n, \bar{X}) = \frac{2}{\pi}\left[(1-\delta) + \delta\sigma_*^2\right]\left[(1-\delta) + \frac{\delta}{\sigma_*}\right]^2.$$

$\delta\backslash\sigma_*$	2	3	5	10
.00	.64	.64	.64	.64
.01	.65	.68	.78	1.24
.05	.70	.83	1.29	3.46
.10	.75	1.00	1.83	5.75
.25	.85	1.33	2.85	9.85

The mean's advantage at the exact normal model is slim (ARE = .64) and fragile: the contaminating variance enters the first bracket linearly through $\delta\sigma_*^2$ while the second bracket, the density at 0, is barely moved, so 1% contamination at scale 10 (ARE 1.24) or 10% at scale 3 (ARE 1.00) already erases it. The median costs at most about 36% efficiency and wins under any appreciable heavy-tailed contamination.

Problem 10.24. Assuming that θ_0 satisfies $\mathrm{E}_{\theta_0}\psi(X-\theta_0) = 0$, show that (10.2.4) and (10.2.5) imply (10.2.6). Here $\hat{\theta}_M$ is the M-estimator solving $\sum_i \psi(x_i - \theta) = 0$, and the three displays of Section 10.2.2 are

$$\frac{-1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i - \theta_0) = \sqrt{n}\left[\frac{-1}{n}\sum_{i=1}^n \psi(X_i - \theta_0)\right] \to \mathrm{n}\big(0, \mathrm{E}_{\theta_0}\psi(X-\theta_0)^2\big)$$

in distribution (10.2.4),

$$\frac{1}{n}\sum_{i=1}^n \psi'(x_i - \theta_0) \to \mathrm{E}_{\theta_0}\psi'(X-\theta_0)$$

in probability (10.2.5), and (10.2.6) is

$$\sqrt{n}\,(\hat{\theta}_M - \theta_0) \to \mathrm{n}\left(0, \frac{\mathrm{E}_{\theta_0}\psi(X-\theta_0)^2}{\left[\mathrm{E}_{\theta_0}\psi'(X-\theta_0)\right]^2}\right).$$

Solution. Slutsky's Theorem (Theorem 5.5.17) applied to the ratio derived on page 486,

$$\sqrt{n}\,(\hat{\theta}_M - \theta_0) = \frac{\dfrac{-1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i - \theta_0)}{\dfrac{1}{n}\sum_{i=1}^n \psi'(X_i - \theta_0)}.$$

Call the numerator U_n and the denominator V_n. The assumption $\mathrm{E}_{\theta_0}\psi(X-\theta_0) = 0$ is exactly what centers U_n, so (10.2.4) says $U_n \to U \sim \mathrm{n}(0, \mathrm{E}_{\theta_0}\psi(X-\theta_0)^2)$ in distribution, and (10.2.5) says $V_n \to c := \mathrm{E}_{\theta_0}\psi'(X-\theta_0)$ in probability. Slutsky's hypotheses hold with $c \neq 0$: ψ is assumed monotone increasing with a unique root (page 485), so $\psi' \geq 0$ and is not a.e. zero, giving $c > 0$; hence $1/V_n \to 1/c$ in probability by Theorem 5.5.4, and $U_n/V_n \to U/c$ in distribution, where

$$\frac{U}{c} \sim \mathrm{n}\left(0, \frac{\mathrm{E}_{\theta_0}\psi(X-\theta_0)^2}{\left[\mathrm{E}_{\theta_0}\psi'(X-\theta_0)\right]^2}\right),$$

which is (10.2.6).

Problem 10.25. If $f(x)$ is a pdf symmetric around 0 and ρ is a symmetric function, show that $\int \psi(x-\theta)f(x-\theta)\,dx = 0$, where $\psi = \rho'$ is an odd function. Show that this then implies that if $X_1, \ldots, X_n$ are iid from $f(x-\theta)$ and $\hat{\theta}_M$ is the minimizer of $\sum_i \rho(x_i - \theta)$, then $\hat{\theta}_M$ is asymptotically normal with mean equal to the true value of θ.

Solution. The integrand is odd, so the integral vanishes. In detail: $\rho(-x) = \rho(x)$ gives, on differentiating, $-\psi(-x) = \psi(x)$, so ψ is odd; f is even; hence $y \mapsto \psi(y)f(y)$ is odd, and substituting $y = x - \theta$,

$$\int_{-\infty}^{\infty} \psi(x-\theta)f(x-\theta)\,dx = \int_{-\infty}^{\infty} \psi(y)f(y)\,dy = 0$$

(the integral being absolutely convergent, as is needed for the variance in (10.2.6) to be finite).
Now let θ be the true value, so $X_i \sim f(x-\theta)$ and the left-hand integral is exactly $\mathrm{E}_\theta \psi(X-\theta)$. What we have just shown is

$$\mathrm{E}_\theta \psi(X-\theta) = 0,$$

which says the true θ is a root of the defining equation $\mathrm{E}_{\theta_0}\psi(X-\theta_0) = 0$ for θ_0; since ψ is monotone increasing (the standing assumption of Section 10.2.2) the root is unique, so $\theta_0 = \theta$. Exercise 10.24 (that is, (10.2.4) and (10.2.5) fed through Slutsky) therefore applies with $\theta_0 = \theta$ and gives

$$\sqrt{n}\,(\hat{\theta}_M - \theta) \to \mathrm{n}\left(0, \frac{\mathrm{E}_\theta \psi(X-\theta)^2}{[\mathrm{E}_\theta \psi'(X-\theta)]^2}\right),$$

so $\hat{\theta}_M$ is asymptotically normal with mean the true θ.

Problem 10.26. Here we look at some details in the calculations in Example 10.2.6, where $X_1, \ldots, X_n$ are iid from $f(x-\theta)$ with f symmetric around 0, and ρ is the Huber criterion (10.2.2), so that

$$\psi(x) = \begin{cases} x & \text{if } |x| \le k, \\ k & \text{if } x > k, \\ -k & \text{if } x < -k. \end{cases}$$

(a) Verify the expressions for $\mathrm{E}_\theta \psi'(X-\theta)$ and $\mathrm{E}_\theta[\psi(X-\theta)]^2$, and hence verify the formula for the variance of $\hat{\theta}_M$.
(b) When calculating the expected value of ψ', we noted that ψ was not differentiable, but we could work with the differentiable portion. Another approach is to realize that the expected value of ψ is differentiable, and that in (10.2.5) we could write

$$\frac{1}{n}\sum_{i=1}^{n} \psi'(x_i - \theta_0) \longrightarrow \frac{d}{d\theta}\,\mathrm{E}_{\theta_0}\psi(X-\theta)\Big|_{\theta=\theta_0}.$$

Show that this is the same limit as in (10.2.5).

Solution. (a) $\psi'(x) = 1$ for $|x| < k$ and $\psi'(x) = 0$ for $|x| > k$ (the two kink points $x = \pm k$ carry no mass, F being continuous), so with $y = x - \theta$

$$\mathrm{E}_\theta \psi'(X-\theta) = \int_{\theta-k}^{\theta+k} f(x-\theta)\,dx = \int_{-k}^{k} f(y)\,dy = P_0(|X| \le k).$$

Splitting the second expectation at the same two points and substituting $y = x - \theta$ in each piece,

$$\begin{aligned}\mathrm{E}_\theta[\psi(X-\theta)]^2 &= \int_{\theta-k}^{\theta+k} (x-\theta)^2 f(x-\theta)\,dx \\ &\quad + k^2\int_{\theta+k}^{\infty} f(x-\theta)\,dx + k^2\int_{-\infty}^{\theta-k} f(x-\theta)\,dx \\ &= \int_{-k}^{k} y^2 f(y)\,dy + k^2\int_{k}^{\infty} f(y)\,dy + k^2\int_{-\infty}^{-k} f(y)\,dy \\ &= \int_{-k}^{k} y^2 f(y)\,dy + 2k^2\int_{k}^{\infty} f(y)\,dy,\end{aligned}$$

the last step by symmetry of f. Substituting both into (10.2.6),

$$\mathrm{Var}(\hat\theta_M) \approx \frac{1}{n}\cdot\frac{\int_{-k}^{k} x^2 f(x)\,dx + 2k^2\int_{k}^{\infty} f(x)\,dx}{[P_0(|X|\le k)]^2}.$$

(The boxed form on page 487 prints $2k^2P_0(|X|>k)$ in the numerator, a slip for $2k^2P_0(X>k) = k^2P_0(|X|>k)$; with the expression above and $k=1.5$, $f=$ standard normal, the asymptotic variance is 1.0371, giving the ARE values .96 versus the mean and 1.51 versus the median recorded in Example 10.2.7.)

(b) The kink contributions cancel. Put $\Delta = \theta - \theta_0$ and substitute $u = x - \theta_0$, so that

$$h(\theta) := \mathrm{E}_{\theta_0}\psi(X-\theta) = \int_{-\infty}^{\infty}\psi(u-\Delta)f(u)\,du,$$

which for the Huber ψ is

$$h(\theta) = \int_{\Delta-k}^{\Delta+k}(u-\Delta)f(u)\,du + k\int_{\Delta+k}^{\infty} f(u)\,du - k\int_{-\infty}^{\Delta-k} f(u)\,du.$$

Each piece is differentiable in Δ, and Leibniz's rule gives

$$\begin{aligned}\frac{dh}{d\Delta} &= \Big[kf(\Delta+k) + kf(\Delta-k) - \int_{\Delta-k}^{\Delta+k} f(u)\,du\Big] \\ &\quad - kf(\Delta+k) - kf(\Delta-k) \\ &= -\int_{\Delta-k}^{\Delta+k} f(u)\,du.\end{aligned}$$

So h is differentiable everywhere even though ψ is not, the boundary terms $kf(\Delta\pm k)$ created by the kinks cancelling exactly, and at $\Delta = 0$

$$\left.\frac{dh}{d\Delta}\right|_{\Delta=0} = -\int_{-k}^{k} f(u)\,du = -P_0(|X|\le k) = -\mathrm{E}_{\theta_0}\psi'(X-\theta_0),$$

which is the limit in (10.2.5) up to the sign that differentiating in θ rather than in the argument of ψ introduces, since $\partial\psi(x-\theta)/\partial\theta = -\psi'(x-\theta)$ (the display on page 487); the magnitude, which is what enters the squared denominator of (10.2.6), is the same.

Problem 10.27. Consider the situation of Example 10.6.2. There F is a continuous cdf with pdf f, μ is the population mean and $m = F^{-1}(1/2)$ the population median, and, following Definition 10.6.1, the influence function of a statistic $T = T(F_n)$ at a point x is

$$IF(T,x) = \lim_{\delta\to 0}\frac{1}{\delta}\left[T(F_\delta) - T(F)\right],$$

where

$$X \sim \begin{cases} F & \text{with probability } 1-\delta, \\ x & \text{with probability } \delta. \end{cases}$$

(a) Verify that $IF(\bar{X}, x) = x - \mu$.
(b) For the median we have $T(F) = m$ if $P(X \le m) = 1/2$ or $m = F^{-1}(1/2)$. If $X \sim F_\delta$, show that

$$P(X \le a) = \begin{cases} (1-\delta)F(a) & \text{if } x > a, \\ (1-\delta)F(a) + \delta & \text{otherwise,} \end{cases}$$

and thus

$$T(F_\delta) = \begin{cases} F^{-1}\left(\frac{1}{2(1-\delta)}\right) & \text{if } x > F^{-1}\left(\frac{1}{2(1-\delta)}\right), \\ F^{-1}\left(\frac{1/2-\delta}{1-\delta}\right) & \text{otherwise.} \end{cases}$$

(c) Show that

$$\frac{1}{\delta}\left[F^{-1}\left(\frac{1}{2(1-\delta)}\right) - F^{-1}\left(\frac{1}{2}\right)\right] \longrightarrow \frac{1}{2f(m)},$$

and complete the argument to calculate $IF(M, x)$. (**Hint:** Write $a_\delta = F^{-1}\left(\frac{1}{2(1-\delta)}\right)$, and argue that the limit is $a'_\delta|_{\delta=0}$. This latter quantity can be calculated using implicit differentiation and the fact that $(1-\delta)^{-1} = 2F(a_\delta)$.)

Solution. (a) The mean of a mixture is the mixture of the means, so $T(F_\delta) = (1-\delta)\mu + \delta x$ and

$$IF(\bar{X}, x) = \lim_{\delta \to 0} \frac{1}{\delta}\left[(1-\delta)\mu + \delta x - \mu\right] = \lim_{\delta \to 0} \frac{\delta(x-\mu)}{\delta} = x - \mu.$$

(b) F_δ puts mass $1-\delta$ on F and mass δ at the point x, so

$$P(X \le a) = (1-\delta)F(a) + \delta\, I(x \le a),$$

which is $(1-\delta)F(a)$ when $x > a$ and $(1-\delta)F(a)+\delta$ otherwise. The median of F_δ solves $P(X \le a) = 1/2$, and there are two branches.
(i) The root lies below the contaminating point ($a < x$): then $(1-\delta)F(a) = 1/2$, so $F(a) = \frac{1}{2(1-\delta)}$ and $a = a_\delta := F^{-1}\left(\frac{1}{2(1-\delta)}\right)$. This is the consistent branch exactly when $x > a_\delta$.
(ii) Otherwise ($a \ge x$): $(1-\delta)F(a) + \delta = 1/2$, so $F(a) = \frac{1/2-\delta}{1-\delta}$ and $a = b_\delta := F^{-1}\left(\frac{1/2-\delta}{1-\delta}\right)$.
(c) The displayed ratio is $\frac{a_\delta - a_0}{\delta - 0}$ with $a_0 = F^{-1}(1/2) = m$, so its limit is $a'_\delta|_{\delta=0}$. Differentiating the defining identity $2F(a_\delta) = (1-\delta)^{-1}$ implicitly,

$$2f(a_\delta)\, a'_\delta = \frac{1}{(1-\delta)^2}, \qquad \text{so} \qquad a'_\delta = \frac{1}{2f(a_\delta)(1-\delta)^2},$$

and putting $\delta = 0$ (where $a_0 = m$, and $f(m) > 0$) gives $a'_0 = \frac{1}{2f(m)}$. The same implicit differentiation applied to $F(b_\delta) = \frac{1/2-\delta}{1-\delta}$ gives

$$f(b_\delta)\, b'_\delta = \frac{-(1-\delta) + (1/2-\delta)}{(1-\delta)^2} = \frac{-1/2}{(1-\delta)^2}, \qquad b'_0 = -\frac{1}{2f(m)}.$$

To finish, fix x and let $\delta \to 0$. If $x > m$, then $a_\delta \to a_0 = m < x$, so $x > a_\delta$ for all sufficiently small δ and branch (i) of (b) is in force: $T(F_\delta) = a_\delta$ and

$$IF(M, x) = \lim_{\delta \to 0} \frac{a_\delta - m}{\delta} = a'_0 = \frac{1}{2f(m)}.$$

If $x \le m$, then $\frac{1}{2(1-\delta)} > \frac{1}{2}$ forces $a_\delta > m \ge x$ for every $\delta > 0$, so branch (ii) is in force: $T(F_\delta) = b_\delta$, with $b_0 = F^{-1}(1/2) = m$, and

$$IF(M, x) = b_0' = -\frac{1}{2f(m)}.$$

That is,

$$IF(M, x) = \begin{cases} \dfrac{1}{2f(m)} & \text{if } x > m, \\ -\dfrac{1}{2f(m)} & \text{otherwise,} \end{cases}$$

the formula quoted in Example 10.6.2.

Problem 10.28. Show that if ρ is defined by (10.2.2), that is,

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le k, \\ k|x| - \frac{1}{2}k^2 & \text{if } |x| \ge k, \end{cases}$$

then both ρ and ρ' are continuous.

Solution. Both pieces agree in value and in slope at $|x| = k$.
Value: at $x = \pm k$ the first formula gives $\frac{1}{2}k^2$ and the second gives $k|\pm k| - \frac{1}{2}k^2 = \frac{1}{2}k^2$, so the two stated cases agree on their overlap $|x| = k$; off $\pm k$ each formula is a polynomial ($-kx - \frac{1}{2}k^2$ for $x < -k$). Hence ρ is continuous.
Slope: the one-sided derivatives at $x = k$ are

$$\frac{d}{dx}\frac{x^2}{2}\bigg|_{x=k} = k \qquad \text{and} \qquad \frac{d}{dx}\left(kx - \tfrac{k^2}{2}\right)\bigg|_{x=k} = k,$$

so $\rho'(k) = k$, and the mirror computation gives $\rho'(-k) = -k$. Hence ρ is differentiable everywhere, with

$$\rho'(x) = \psi(x) = \max\{-k, \min(k, x)\} = \begin{cases} x & \text{if } |x| \le k, \\ k & \text{if } x > k, \\ -k & \text{if } x < -k, \end{cases}$$

which is (10.2.7), and $\rho' = \psi$ is continuous, being a composition of the continuous maps $\min(k, \cdot)$ and $\max(-k, \cdot)$.

10.5 Exercises 10.29–10.35

Problem 10.29. From (10.2.9) we know that an M-estimator can never be more efficient than a maximum likelihood estimator. However, we also know when it can be as efficient. Recall that (10.2.9) reads

$$\mathrm{ARE}(\hat{\theta}_M, \hat{\theta}) = \frac{[\mathrm{E}_\theta \psi(X - \theta_0)\, l'(\theta|X)]^2}{\mathrm{E}_\theta \psi(X-\theta)^2 \; \mathrm{E}_\theta l'(\theta|X)^2} \le 1.$$

(a) Show that (10.2.9) is an equality if we choose $\psi(x - \theta) = c\, l'(\theta|x)$, where l is the log likelihood and c is a constant.
(b) For each of the following distributions, verify that the corresponding ψ functions give asymptotically efficient M-estimators.
(i) Normal: $f(x) = e^{-x^2/2}/(\sqrt{2\pi})$, $\psi(x) = x$.
(ii) Logistic: $f(x) = e^{-x}/(1 + e^{-x})^2$, $\psi(x) = \tanh(x)$, where $\tanh(x)$ is the hyperbolic tangent.
(iii) Cauchy: $f(x) = [\pi(1 + x^2)]^{-1}$, $\psi(x) = 2x/(1 + x^2)$.

(iv) Least informative distribution:

$$f(x) = \begin{cases} Ce^{-x^2/2} & |x| \le c \\ Ce^{-c|x|+c^2/2} & |x| > c \end{cases}$$

with $\psi(x) = \max\{-c, \min(c, x)\}$ and C and c are constants.
(See Huber 1981, Section 3.5, for more details.)

Solution. (a) The Cauchy–Schwarz inequality that produced (10.2.9) is an equality exactly when the two variables are proportional, and $\psi(X - \theta_0) = c\, l'(\theta|X)$ is that proportionality:

$$\frac{[\mathrm{E}(c\, l' \cdot l')]^2}{\mathrm{E}(c\, l')^2\, \mathrm{E}(l')^2} = \frac{c^2\, [\mathrm{E}(l')^2]^2}{c^2\, \mathrm{E}(l')^2\, \mathrm{E}(l')^2} = 1.$$

(b) All four are location families $f(x - \theta)$, so $l(\theta|x) = \log f(x - \theta)$ and

$$l'(\theta|x) = \frac{\partial}{\partial \theta} \log f(x - \theta) = -\frac{f'(x - \theta)}{f(x - \theta)}.$$

By part (a) the M-estimator is asymptotically efficient if and only if

$$\psi(u) = -a\, \frac{f'(u)}{f(u)} = -a\, \frac{d}{du} \log f(u)$$

for some constant a (renamed from the c of part (a), since c is the tuning constant in (iv)). So in each case it suffices to compute $-(\log f)'$.
(i) $\log f(u) = -u^2/2 - \frac{1}{2}\log(2\pi)$, so $-(\log f)'(u) = u = \psi(u)$ with $a = 1$.
(ii) $\log f(u) = -u - 2\log(1 + e^{-u})$, so

$$-(\log f)'(u) = 1 - \frac{2e^{-u}}{1 + e^{-u}} = \frac{1 - e^{-u}}{1 + e^{-u}} = \tanh(u/2).$$

Errata: the printed $\psi(x) = \tanh(x)$ should read $\psi(x) = \tanh(x/2)$, since $\tanh(x)$ is not a constant multiple of $\tanh(x/2)$.
(iii) $\log f(u) = -\log \pi - \log(1 + u^2)$, so $-(\log f)'(u) = 2u/(1 + u^2) = \psi(u)$ with $a = 1$.
(iv) The density is continuous at $|u| = c$ (both branches give $Ce^{-c^2/2}$), and

$$-(\log f)'(u) = \begin{cases} -c & u < -c \\ u & |u| \le c \\ c & u > c \end{cases} = \max\{-c, \min(c, u)\} = \psi(u),$$

with $a = 1$.

Problem 10.30. For M-estimators there is a connection between the ψ function and the breakdown value. The details are rather involved (Huber 1981, Section 3.2) but they can be summarized as follows: If ψ is a bounded function, then the breakdown value of the associated M-estimator is given by

$$b^* = \frac{\eta}{1 + \eta}, \qquad \eta = \min\left\{ -\frac{\psi(-\infty)}{\psi(\infty)}, -\frac{\psi(\infty)}{\psi(-\infty)} \right\}.$$

(a) Calculate the breakdown value of the efficient M-estimators of Exercise 10.29. Which ones are both efficient and robust?
(b) Calculate the breakdown value of these other M-estimators.
(i) The Huber estimator given by (10.2.1); its ψ function is (10.2.7), namely $\psi(x) = x$ for $|x| \le k$, $\psi(x) = k$ for $x > k$, and $\psi(x) = -k$ for $x < -k$.

(ii) Tukey's biweight: $\psi(x) = x(c^2 - x^2)$ for $|x| \le c$ and 0 otherwise, where c is a constant.
(iii) Andrew's sine wave: $\psi(x) = c\sin(x/c)$ for $|x| \le c\pi$ and 0 otherwise.
(c) Evaluate the AREs of the estimators in part (b) with respect to the MLE when the underlying distribution is (i) normal and (ii) double exponential.

Solution. (a) $b^* = 0$ for the normal $\psi(x) = x$ and $b^* = \frac{1}{2}$ for the other three. Every ψ in Exercise 10.29 is odd, so whenever the limits are finite and nonzero, $\psi(-\infty) = -\psi(\infty)$ gives

$$\eta = \min\{1, 1\} = 1, \qquad b^* = \frac{1}{1+1} = \tfrac{1}{2}.$$

(i) Normal, $\psi(x) = x$: ψ is unbounded, so the formula does not apply. Here the M-estimator is $\bar{X}$, whose breakdown value is 0 (Section 10.2.1): driving one of n observations to ∞ drives the root of $\sum(x_i - \theta) = 0$ to ∞.
(ii) Logistic, $\psi(x) = \tanh(x/2)$: $\psi(\pm\infty) = \pm 1$, so $\eta = 1$ and $b^* = \frac{1}{2}$.
(iii) Cauchy, $\psi(x) = 2x/(1 + x^2)$: ψ redescends, $\psi(\pm\infty) = 0$, and both printed ratios read 0/0, so the quoted summary (proved in Huber 1981, Section 3.2, for bounded *monotone* ψ) assigns no value directly. Two observations settle the case. First, since ψ is odd and $\psi(t) > 0$ for every finite $t > 0$, the ratios equal 1 *before* the limit is taken: $-\psi(-t)/\psi(t) = 1$ for all $t \in (0, \infty)$. Reading the display as the limit of the ratios (the only reading that extends the monotone case continuously) therefore gives $\eta = 1$ and $b^* = \frac{1}{2}$. Second, $b^* = \frac{1}{2}$ can be proved outright for this estimator, the Cauchy location MLE $\hat{\theta} = \operatorname{argmin}_t \sum_i \rho(x_i - t)$ with $\rho(x) = \log(1 + x^2)$. Keep $n - m$ good observations $g_i \in [-K, K]$ and let the other m observations be arbitrary. Expanding $(y - t + t)^2 \le 2(y - t)^2 + 2t^2$ gives $1 + y^2 \le 2\left[1 + (y - t)^2\right](1 + t^2)$, hence $\rho(y - t) \ge \rho(y) - \rho(t) - \log 2$ for all y, t; applying this to the bad points and $\rho(g_i - t) \ge \rho(|t| - K)$ to the good ones, for $|t| \ge K$,

$$\sum_i \rho(x_i - t) - \sum_i \rho(x_i - 0) \ge (n - 2m)\rho(|t| - K) - m\left[\rho(|t|) - \rho(|t| - K)\right] - m\log 2 - (n - m)\rho(K),$$

where $\rho(|t|) - \rho(|t| - K) \to 0$ and $\rho(|t| - K) \to \infty$ as $|t| \to \infty$. So if $m < n/2$ the right side tends to $+\infty$, and there is an L depending only on n, m, K — *not* on the contaminants — with $|\hat{\theta}| < L$: no fraction below $\frac{1}{2}$ breaks the estimator down. Conversely no translation-equivariant location estimator withstands $m = \lceil n/2 \rceil$ replacements: for any sample $\mathbf{x} = (x_1, \ldots, x_n)$, the sample $\mathbf{z} = (x_1, \ldots, x_{n-m}, x_1 + \tau, \ldots, x_m + \tau)$ is an m-replacement of $\mathbf{x}$ and an $(n - m \le m)$-replacement of $\mathbf{x} + \tau$, so a uniform bound B on the perturbed estimate's distance from the unperturbed one would force $T(\mathbf{z})$ to lie within $B + |T(\mathbf{x})|$ of both 0 and τ — impossible for τ large. Thus $b^* = \frac{1}{2}$ exactly.
(iv) Least informative, $\psi(x) = \max\{-c, \min(c, x)\}$: $\psi(\pm\infty) = \pm c$, so $\eta = 1$ and $b^* = \frac{1}{2}$.
Thus the logistic, Cauchy and least-informative ψ's are simultaneously efficient (each at its own model) and maximally robust; the normal ψ buys efficiency at zero breakdown.
(b) (i) Huber's ψ is bounded and monotone with $\psi(\pm\infty) = \pm k \ne 0$ — exactly the case in which the quoted formula is proved — so $\eta = \min\{k/k,\, k/k\} = 1$ and $b^* = \frac{1}{2}$.
(ii)–(iii) The biweight and the sine wave are bounded, so the exercise's premise is met, but both vanish identically off a compact set: $\psi(\pm\infty) = 0$, both ratios are 0/0, and here even the limit-of-the-ratio reading of (a)(iii) fails, since $-\psi(-t)/\psi(t)$ is already 0/0 for every $|t| > c$ (resp. $|t| > c\pi$). The value recorded for every bounded odd ψ — and the answer this exercise intends — is nevertheless $b^* = \frac{1}{2}$, by the convention $\eta = 1$ of part (a). It reflects genuine outlier rejection: once an observation is moved far enough out, $\psi(x_i - t) = 0$ for every t in a fixed bounded range, so gross outliers drop out of $\sum_i \psi(x_i - \theta) = 0$ entirely, as if deleted.
What is exactly provable depends on how the redescending estimator is defined, and the honest statement is that $\frac{1}{2}$ is the supremal value. Take the standard definition $T(H) = \operatorname{argmin}_t \mathrm{E}_H \rho(X - t)$ with $\rho(x) = \int_0^{|x|} \psi(u)\, du$, which here is even, continuous, nondecreasing in $|x|$ and constant off the support, with $\rho_{\max} = c^4/4$ (biweight) and $\rho_{\max} = 2c^2$ (sine wave). Contaminate a model F to $H_\varepsilon = (1-\varepsilon)F + \varepsilon G$ and put $A = \inf_t \mathrm{E}_F \rho(X - t)$. Given $\delta > 0$, choose M with $F([-M, M]) \ge 1 - \delta$; every $|t|$ beyond M plus the support radius then puts all of that F-mass a full support-width from t, so $\mathrm{E}_F \rho(X - t) \ge (1 - \delta)\rho_{\max}$

there. Hence if $(1-\varepsilon)(\rho_{\max}-A)>\varepsilon\rho_{\max}$, taking δ small enough gives, for all such t and every G,

$$\mathrm{E}_{H_\varepsilon}\rho(X-t)\ge(1-\varepsilon)(1-\delta)\rho_{\max}>(1-\varepsilon)A+\varepsilon\rho_{\max}\ge\inf_s\mathrm{E}_{H_\varepsilon}\rho(X-s),$$

so every minimizer stays in a fixed compact set; while if $(1-\varepsilon)(\rho_{\max}-A)<\varepsilon\rho_{\max}$, the point mass $G=\delta_y$ with $y\to\infty$ reverses the comparison and drags the minimizer out to y. The exact contamination breakdown is therefore

$$b^*(F)=\frac{\rho_{\max}-A}{2\rho_{\max}-A}<\tfrac12\quad\text{for every nondegenerate } F,$$

increasing to $\frac12$ as $A/\rho_{\max}\to0$, e.g., as $c\to\infty$ (for the biweight, $A/\rho_{\max}\approx 2\,\mathrm{Var}_F X/c^2$ for large c) or as F concentrates; Huber (1984, *Annals of Statistics*) gives the matching finite-sample analysis. At the standard normal the exact values for the tuning constants of part (c) are $.449,.470,.481$ (biweight, $c=3,4,5$) and $.445,.468,.485$ (sine wave, $c=1,1.339,2$). So all three estimators in (b) are reported as $b^*=\frac12$: exact for Huber's monotone ψ, and for the two redescenders the limiting value, with the true breakdown $(\rho_{\max}-A)/(2\rho_{\max}-A)$ just below it.
(c) By (10.2.6) the asymptotic variance of $\hat\theta_M$ is $\mathrm{E}\psi^2/[\mathrm{E}\psi']^2$, and the MLE attains $1/I(\theta)$ with $I=1$ for both the $\mathrm{n}(0,1)$ and the double exponential $f(x)=\frac12e^{-|x|}$ location families. Hence

$$\mathrm{ARE}(\hat\theta_M,\hat\theta)=\frac{[\mathrm{E}\psi'(X)]^2}{\mathrm{E}\psi(X)^2}.$$

For the Huber ψ this is explicit. At the normal, with Φ,φ the standard normal cdf and pdf,

$$\begin{aligned}\mathrm{E}\psi'&=P(|X|\le k)=2\Phi(k)-1,\\ \mathrm{E}\psi^2&=2\Phi(k)-1-2k\varphi(k)+2k^2\left[1-\Phi(k)\right],\end{aligned}$$

and at the double exponential,

$$\mathrm{E}\psi'=1-e^{-k},\qquad \mathrm{E}\psi^2=2-2(k+1)e^{-k},$$

so that $\mathrm{ARE}=(1-e^{-k})^2/\left[2-2(k+1)e^{-k}\right]$ there; at $k=1.5$ these give .964 and .683, the .96 and .68 of Example 10.2.7. The biweight and sine-wave integrals are evaluated numerically over their compact supports.

estimator	tuning constant	normal	double exponential
Huber	$k=1.0$	.903	.756
Huber	$k=1.5$	.964	.683
Huber	$k=2.0$	.990	.629
biweight	$c=3$	.881	.686
biweight	$c=4$	.966	.671
biweight	$c=5$	.988	.645
sine wave	$c=1.0$	.856	.711
sine wave	$c=1.339$	.950	.696
sine wave	$c=2.0$	.990	.644

Problem 10.31. Binomial data gathered from more than one population are often presented in a *contingency table*. For the case of two populations, the table might look like this:

	Population 1	Population 2	Total
Successes	S_1	S_2	$S=S_1+S_2$
Failures	F_1	F_2	$F=F_1+F_2$
Total	n_1	n_2	$n=n_1+n_2$

where Population 1 is binomial(n_1,p_1), with S_1 successes and F_1 failures, and Population 2 is

binomial(n_2, p_2), with S_2 successes and F_2 failures. A hypothesis that is usually of interest is

$$H_0 : p_1 = p_2 \qquad \text{versus} \qquad H_1 : p_1 \neq p_2.$$

(a) Show that a test can be based on the statistic

$$T = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})},$$

where $\hat{p}_1 = S_1/n_1$, $\hat{p}_2 = S_2/n_2$, and $\hat{p} = (S_1 + S_2)/(n_1 + n_2)$. Also, show that as $n_1, n_2 \to \infty$, the distribution of T approaches χ^2_1. (This is a special case of a test known as a *chi squared test of independence*.)
(b) Another way of measuring departure from H_0 is by calculating an *expected frequency table*. This table is constructed by conditioning on the marginal totals and filling in the table according to $H_0 : p_1 = p_2$, that is,

Expected frequencies	1	2	Total
Successes	$\frac{n_1 S}{n_1+n_2}$	$\frac{n_2 S}{n_1+n_2}$	$S = S_1 + S_2$
Failures	$\frac{n_1 F}{n_1+n_2}$	$\frac{n_2 F}{n_1+n_2}$	$F = F_1 + F_2$
Total	n_1	n_2	$n = n_1 + n_2$

Using the expected frequency table, a statistic T^* is computed by going through the cells of the tables and computing

$$T^* = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \frac{\left(S_1 - \frac{n_1 S}{n_1+n_2}\right)^2}{\frac{n_1 S}{n_1+n_2}} + \cdots + \frac{\left(F_2 - \frac{n_2 F}{n_1+n_2}\right)^2}{\frac{n_2 F}{n_1+n_2}}.$$

Show, algebraically, that $T^* = T$ and hence that T^* is asymptotically chi squared.
(c) Another statistic that could be used to test equality of p_1 and p_2 is

$$T^{**} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

Show that, under H_0, T^{**} is asymptotically n$(0, 1)$, and hence its square is asymptotically χ^2_1. Furthermore, show that $(T^{**})^2 \neq T^*$.
(d) Under what circumstances is one statistic preferable to the other?
(e) A famous medical experiment was conducted by Joseph Lister in the late 1800s. Mortality associated with surgery was quite high, and Lister conjectured that the use of a disinfectant, carbolic acid, would help. Over a period of several years Lister performed 75 amputations with and without using carbolic acid. The data are given here:

Patient lived?	Carbolic acid used? Yes	Carbolic acid used? No
Yes	34	19
No	6	16

Use these data to test whether the use of carbolic acid is associated with patient mortality.

Solution. (a) Under H_0 write p for the common value. The two samples are independent, $\mathrm{E}(\hat{p}_1 - \hat{p}_2) = 0$ and

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Let $a_n = (1/n_1 + 1/n_2)^{-1/2}$. By the Central Limit Theorem applied to each sample, $\sqrt{n_i}(\hat{p}_i - p) \to$ n$(0, p(1-p))$ independently, so if $n_1, n_2 \to \infty$ with $n_1/(n_1+n_2) \to \lambda \in (0,1)$,

$$\frac{a_n(\hat{p}_1 - \hat{p}_2)}{\sqrt{p(1-p)}} \to \mathrm{n}(0,1).$$

Since $\hat{p}$ is the MLE of p under H_0 and is consistent, $\hat{p}(1-\hat{p})/[p(1-p)] \to 1$ in probability, and Slutsky's Theorem (Theorem 5.5.17) gives

$$\frac{a_n(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})}} \to \mathrm{n}(0,1), \qquad \text{so} \qquad T \to \chi^2_1.$$

The level α test rejects H_0 when $T > \chi^2_{1,\alpha}$.
(b) All four deviations have the same absolute value. Writing $D = (n_2S_1 - n_1S_2)/n$,

$$S_1 - \frac{n_1 S}{n} = \frac{nS_1 - n_1(S_1+S_2)}{n} = D, \quad S_2 - \frac{n_2 S}{n} = -D,$$
$$F_1 - \frac{n_1 F}{n} = -\left(S_1 - \frac{n_1 S}{n}\right) = -D, \qquad F_2 - \frac{n_2 F}{n} = D,$$

where the third equality uses $F_1 = n_1 - S_1$ and $n_1F/n = n_1 - n_1S/n$. Hence

$$\begin{aligned} T^* &= D^2 \frac{n}{S}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) + D^2 \frac{n}{F}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \\ &= D^2 \frac{n^2}{n_1 n_2}\left(\frac{1}{S} + \frac{1}{F}\right) = D^2 \frac{n^3}{n_1 n_2 SF}. \end{aligned}$$

Now $\hat{p}_1 - \hat{p}_2 = (n_2S_1 - n_1S_2)/(n_1n_2) = nD/(n_1n_2)$, so $D = (\hat{p}_1 - \hat{p}_2)n_1n_2/n$ and

$$T^* = (\hat{p}_1 - \hat{p}_2)^2 \frac{n_1^2 n_2^2}{n^2} \cdot \frac{n^3}{n_1 n_2 SF} = \frac{(\hat{p}_1 - \hat{p}_2)^2\, n\, n_1 n_2}{SF}.$$

On the other side, $\hat{p} = S/n$ and $1 - \hat{p} = F/n$, so the denominator of T is $(n/(n_1n_2))(SF/n^2) = SF/(n\,n_1n_2)$, and T equals the same expression. Thus $T^* = T \to \chi^2_1$.
(c) Under H_0, $\hat{p}_1 \to p$ and $\hat{p}_2 \to p$ in probability, so

$$\frac{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}{p(1-p)(1/n_1 + 1/n_2)} \to 1$$

in probability, and Slutsky applied to the display in part (a) gives $T^{**} \to \mathrm{n}(0,1)$, hence $(T^{**})^2 \to \chi^2_1$. The two statistics have the same numerator and different denominators. With $w_i = n_i/n$, $A_i = \hat{p}_i(1-\hat{p}_i)$ and $d = \hat{p}_1 - \hat{p}_2$, the identity $\hat{p} = w_1\hat{p}_1 + w_2\hat{p}_2$ gives

$$\hat{p}(1-\hat{p}) = w_1A_1 + w_2A_2 + w_1w_2d^2,$$

so the pooled and unpooled variance estimates differ by

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p}) - \left(\frac{A_1}{n_1} + \frac{A_2}{n_2}\right) = (A_1 - A_2)\left(\frac{1}{n_2} - \frac{1}{n_1}\right) + \frac{d^2}{n}.$$

This is not 0 in general, so $(T^{**})^2 \neq T^*$. In particular, when $n_1 = n_2$ the difference is $d^2/n \geq 0$, so the pooled denominator is the larger and $(T^{**})^2 \geq T^*$, with equality only in the degenerate case $\hat{p}_1 = \hat{p}_2$.
(d) For testing $H_0 : p_1 = p_2$ the pooled statistic $T = T^*$ is preferable: under H_0 the pooled $\hat{p}$ is the MLE of the single common parameter, so it estimates the null variance more precisely than $A_1/n_1 + A_2/n_2$ does, and the χ^2_1 approximation is correspondingly better in moderate samples. The statistic T^{**} uses a variance estimate that remains valid when $p_1 \neq p_2$; it is therefore the one to use for power calculations under alternatives and the one to invert to obtain a confidence interval for $p_1 - p_2$.
(e) Reject H_0: carbolic acid is associated with lower mortality ($T = 8.50$, p-value $= .004$). Taking "success" to be that the patient lived, Population 1 is the carbolic acid group with $n_1 = 34 + 6 = 40$, $S_1 = 34$, and Population 2 the control with $n_2 = 19 + 16 = 35$, $S_2 = 19$, so

$$\hat{p}_1 = \tfrac{34}{40} = .850, \qquad \hat{p}_2 = \tfrac{19}{35} = .543, \qquad \hat{p} = \tfrac{53}{75} = .707.$$

Then

$$T = \frac{(.850 - .543)^2}{\left(\frac{1}{40} + \frac{1}{35}\right)(.707)(.293)} = \frac{.09434}{.011105} = 8.50,$$

and $P(\chi^2_1 > 8.50) = .0036 < .05 = \alpha$.

Problem 10.32. (a) Let $(X_1, \ldots, X_n) \sim \text{multinomial}(m, p_1, \ldots, p_n)$. Consider testing $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$. A test that is often used, called *McNemar's Test*, rejects H_0 if

$$\frac{(X_1 - X_2)^2}{X_1 + X_2} > \chi^2_{1,\alpha}.$$

Show that this test statistic has the form (as in Exercise 10.31)

$$\sum_i \frac{(\text{observed} - \text{expected})^2}{\text{expected}},$$

where the X_is are the observed cell frequencies and the expected cell frequencies are the MLEs of mp_i, under the assumption that $p_1 = p_2$.
(b) McNemar's Test is often used in the following type of problem. Subjects are asked if they agree or disagree with a statement. Then they read some information about the statement and are asked again if they agree or disagree. The numbers of responses in each category are summarized in a 2×2 table like this:

		Before: Agree	Before: Disagree
After	Agree	X_3	X_2
	Disagree	X_1	X_4

The hypothesis $H_0 : p_1 = p_2$ states that the proportion of people who change from agree to disagree is the same as the proportion of people who change from disagree to agree. Another hypothesis that might be tested is that the proportion of those who initially agree and then change is the same as the proportion of those who initially disagree and then change. Express this hypothesis in terms of conditional probabilities and show that it is different from the above H_0. (This hypothesis can be tested with a χ^2 test like those in Exercise 10.31.)

Solution. (a) Under H_0 put $p_1 = p_2 = q$; the log likelihood is

$$(X_1 + X_2)\log q + \sum_{j \geq 3} X_j \log p_j, \qquad 2q + \sum_{j \geq 3} p_j = 1.$$

A Lagrange multiplier λ gives $(X_1 + X_2)/q = 2\lambda$ and $X_j/p_j = \lambda$ for $j \geq 3$; summing $X_1 + X_2 = 2\lambda q$ and $X_j = \lambda p_j$ over all cells yields $m = \lambda$, so

$$\hat{q} = \frac{X_1 + X_2}{2m}, \qquad \hat{p}_j = \frac{X_j}{m} \ (j \geq 3).$$

The expected frequencies $m\hat{p}_i$ are therefore $(X_1 + X_2)/2$ in cells 1 and 2 and X_j in every other cell, so all terms with $j \geq 3$ contribute 0 and

$$\sum_i \frac{(\text{obs} - \text{exp})^2}{\text{exp}} = 2 \cdot \frac{\left(\frac{X_1 - X_2}{2}\right)^2}{\frac{X_1 + X_2}{2}} = \frac{(X_1 - X_2)^2}{X_1 + X_2},$$

using $X_1 - (X_1 + X_2)/2 = (X_1 - X_2)/2 = -\,[X_2 - (X_1 + X_2)/2]$.
(b) In the table, cell 1 is (before agree, after disagree), cell 2 is (before disagree, after agree), cell 3 is (agree, agree) and cell 4 is (disagree, disagree). Those who initially agree are cells 1 and 3, and among them the changers are cell 1; those who initially disagree are cells 2 and 4, with changers cell 2. The hypothesis is therefore

$$H_0' : \quad P(\text{after disagree} \mid \text{before agree}) = P(\text{after agree} \mid \text{before disagree}),$$

that is,

$$\frac{p_1}{p_1 + p_3} = \frac{p_2}{p_2 + p_4} \iff p_1 p_4 = p_2 p_3.$$

This is not $H_0 : p_1 = p_2$: for example $p_1 = p_2 = .1$, $p_3 = .5$, $p_4 = .3$ satisfies H_0 but gives $p_1 p_4 = .03 \neq .05 = p_2 p_3$; conversely $(p_1, p_2, p_3, p_4) = (\frac{2}{9}, \frac{1}{9}, \frac{4}{9}, \frac{2}{9})$ gives $p_1 p_4 = p_2 p_3 = \frac{4}{81}$ with $p_1 \neq p_2$. The two coincide only when $p_1 = p_2$ and $p_3 = p_4$, or $p_1 = p_2 = 0$.

Problem 10.33. Fill in the gap in Theorem 10.3.1. Use Theorem 10.1.12 and Slutsky's Theorem (Theorem 5.5.17) to show that

$$\frac{\theta-\hat{\theta}}{\sqrt{-l''(\hat{\theta}|\mathbf{x})}} \to \mathrm{n}(0,1),$$

and therefore $-2\log\lambda(\mathbf{X}) \to \chi^2_1$. (Theorem 10.3.1 tests $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ with $X_1, \ldots, X_n$ iid $f(x|\theta)$, $\hat{\theta}$ the MLE of θ, and $f(x|\theta)$ satisfying the regularity conditions in Miscellanea 10.6.2; its proof Taylor expands $l(\theta|\mathbf{x}) = \log L(\theta|\mathbf{x})$ about $\hat{\theta}$ to get

$$-2\log\lambda(\mathbf{x}) \approx \frac{(\theta-\hat{\theta})^2}{-l''(\hat{\theta}|\mathbf{x})},$$

using $l'(\hat{\theta}|\mathbf{x}) = 0$, and then appeals to $\frac{1}{n}\hat{I}_n(\hat{\theta}) \to I(\theta_0)$ for $\hat{I}_n(\theta) = -l''(\theta|\mathbf{x})$.)

Solution. Multiply, do not divide: the quantity that is asymptotically standard normal is

$$(\hat{\theta}-\theta_0)\sqrt{-l''(\hat{\theta}|\mathbf{x})} = \frac{\hat{\theta}-\theta_0}{\sqrt{-1/l''(\hat{\theta}|\mathbf{x})}},$$

since $-l''(\hat{\theta}|\mathbf{x}) = \hat{I}_n(\hat{\theta})$ is of order n, so the ratio as printed tends to 0; the same slip sits in the display in the proof of Theorem 10.3.1, which should read $-2\log\lambda(\mathbf{x}) \approx (\theta_0-\hat{\theta})^2\left(-l''(\hat{\theta}|\mathbf{x})\right)$.
By Theorem 10.1.12, the regularity conditions of Miscellanea 10.6.2 give

$$\sqrt{n}\,(\hat{\theta}-\theta_0) \to \mathrm{n}\Big(0, \tfrac{1}{I(\theta_0)}\Big), \qquad \text{i.e.} \qquad \sqrt{nI(\theta_0)}\,(\hat{\theta}-\theta_0) \to \mathrm{n}(0,1).$$

Next, the observed information number $\hat{I}_n(\hat{\theta}) = -l''(\hat{\theta}|\mathbf{x})$ of Section 10.3.2 satisfies $\hat{I}_n(\hat{\theta})/n \to I(\theta_0)$ in probability: at the true value,

$$\begin{aligned}\frac{-l''(\theta_0|\mathbf{X})}{n} &= \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)\right)_{\theta_0} \\ &\to \mathrm{E}_{\theta_0}\left(-\frac{\partial^2}{\partial\theta^2}\log f(X|\theta_0)\right)\end{aligned}$$

in probability by the Weak Law, and the right side is $I(\theta_0)$ by Lemma 7.3.11; since $\hat{\theta} \to \theta_0$ in probability (consistency of the MLE, Theorem 10.1.6) and Miscellanea 10.6.2 supplies a third-derivative bound dominating l''' in a neighborhood of θ_0, the same limit holds with $\hat{\theta}$ in place of θ_0. Therefore

$$\sqrt{\frac{-l''(\hat{\theta}|\mathbf{X})}{nI(\theta_0)}} \to 1 \quad \text{in probability,}$$

and Slutsky's Theorem (Theorem 5.5.17) applied to the product gives

$$(\hat{\theta}-\theta_0)\sqrt{-l''(\hat{\theta}|\mathbf{X})} = \sqrt{nI(\theta_0)}\,(\hat{\theta}-\theta_0)\cdot\sqrt{\frac{-l''(\hat{\theta}|\mathbf{X})}{nI(\theta_0)}} \to \mathrm{n}(0,1).$$

Squaring and applying the Continuous Mapping Theorem,

$$\begin{aligned}-2\log\lambda(\mathbf{X}) &= (\hat{\theta}-\theta_0)^2\left(-l''(\hat{\theta}|\mathbf{X})\right) + o_P(1) \\ &\to \chi^2_1,\end{aligned}$$

the $o_P(1)$ being the Taylor remainder, which is $O_P\Big(n|\hat{\theta}-\theta_0|^3\Big) = O_P(n^{-1/2})$ under the same domination condition.

Problem 10.34. For testing $H_0 : p = p_0$ versus $H_1 : p \neq p_0$, suppose we observe $X_1, \ldots, X_n$ iid Bernoulli(p).
(a) Derive an expression for $-2\log\lambda(\mathbf{x})$, where $\lambda(\mathbf{x})$ is the LRT statistic.
(b) As in Example 10.3.2, simulate the distribution of $-2\log\lambda(\mathbf{x})$ and compare it to the χ^2 approximation.

Solution. (a) With $y = \sum_i x_i$ and $\hat{p} = y/n = \bar{x}$ the unrestricted MLE,

$$\lambda(\mathbf{x}) = \frac{p_0^y(1-p_0)^{n-y}}{\hat{p}^y(1-\hat{p})^{n-y}},$$

so

$$-2\log\lambda(\mathbf{x}) = 2n\left[\bar{x}\log\frac{\bar{x}}{p_0} + (1-\bar{x})\log\frac{1-\bar{x}}{1-p_0}\right],$$

with the convention $0\log 0 = 0$ covering $\bar{x} = 0$ and $\bar{x} = 1$.
(b) No simulation is needed: $-2\log\lambda$ is a function of y alone and $y \sim \text{binomial}(n, p_0)$ under H_0, so the exact null distribution is obtained by summing binomial probabilities over the $n+1$ attainable values. The exact percentiles below are the smallest attained values whose cdf reaches the stated level.

n, p_0	.80	.90	.95	.99	$P(-2\log\lambda > 3.841)$
25, $p_0 = .5$	1.99	3.31	3.31	7.10	.043
25, $p_0 = .2$	1.99	2.76	3.44	5.53	.045
100, $p_0 = .2$	1.68	2.84	3.67	6.76	.045
χ^2_1	1.64	2.71	3.84	6.63	.050

The nominal .05 test rejecting when $-2\log\lambda > \chi^2_{1,.05} = 3.841$ has true level .043–.045 in all three cases, agreement of the same quality as in Example 10.3.2. The only visible defect is discreteness: at $n = 25$, $p_0 = .5$ the statistic takes only 13 distinct values, and the atom at 3.314 serves as both the .90 and the .95 percentile.

Problem 10.35. Let $X_1, \ldots, X_n$ be a random sample from a $\text{n}(\mu, \sigma^2)$ population.
(a) If μ is unknown and σ^2 is known, show that $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma$ is a Wald statistic for testing $H_0 : \mu = \mu_0$.
(b) If σ^2 is unknown and μ is known, find a Wald statistic for testing $H_0 : \sigma = \sigma_0$.

Solution. (a) Take $W_n = \bar{X}$, the MLE of μ, in the Wald form $Z_n = (W_n - \theta_0)/S_n$ of Section 10.3.2, with standard error $S_n = 1/\sqrt{\hat{I}_n(W_n)}$ built from the observed information number. From

$$l(\mu|\mathbf{x}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2, \qquad -\frac{\partial^2 l}{\partial\mu^2} = \frac{n}{\sigma^2},$$

so $\hat{I}_n(\bar{X}) = I_n(\mu) = n/\sigma^2$ and $S_n = \sigma/\sqrt{n}$, giving

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma},$$

which is the stated Wald statistic.
(b) Take $\theta = \sigma$ and $W_n = \hat{\sigma}$, the MLE. From

$$l(\sigma|\mathbf{x}) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2,$$

setting $\partial l/\partial\sigma = -n/\sigma + \sum_i (x_i - \mu)^2/\sigma^3 = 0$ gives

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2.$$

Differentiating again,

$$-\frac{\partial^2 l}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{3}{\sigma^4}\sum_i (x_i-\mu)^2, \qquad \hat{I}_n(\hat\sigma) = -\frac{n}{\hat\sigma^2} + \frac{3n\hat\sigma^2}{\hat\sigma^4} = \frac{2n}{\hat\sigma^2},$$

which agrees with the expected information $I_n(\sigma) = 2n/\sigma^2$ evaluated at $\hat\sigma$. Hence $S_n = \hat\sigma/\sqrt{2n}$ and the Wald statistic is

$$Z = \frac{\hat\sigma - \sigma_0}{\hat\sigma/\sqrt{2n}} = \frac{\sqrt{2n}\,(\hat\sigma - \sigma_0)}{\hat\sigma},$$

rejecting H_0 at asymptotic level α when $|Z| > z_{\alpha/2}$.

10.6 Exercises 10.36–10.42

Problem 10.36. Let $X_1, \ldots, X_n$ be a random sample from a gamma(α, β) population. Assume α is known and β is unknown. Consider testing $H_0\colon \beta = \beta_0$.
(a) What is the MLE of β?
(b) Derive a Wald statistic for testing H_0, using the MLE in both the numerator and denominator of the statistic.
(c) Repeat part (b) but using the sample standard deviation in the standard error.

Solution. $\hat\beta = \bar X/\alpha$, and the two Wald statistics are

$$Z_W = \sqrt{n\alpha}\left(1 - \frac{\alpha\beta_0}{\bar X}\right), \qquad Z'_W = \frac{\sqrt{n}\,(\bar X - \alpha\beta_0)}{S}.$$

(a) With α known, $\log L(\beta \mid \mathbf{x}) = -n\alpha\log\beta - \sum_i x_i/\beta + c(\mathbf{x})$, so

$$\frac{\partial}{\partial\beta}\log L = -\frac{n\alpha}{\beta} + \frac{\sum_i x_i}{\beta^2} = \frac{n}{\beta^2}\,(\bar x - \alpha\beta),$$

which is positive for $\beta < \bar x/\alpha$ and negative for $\beta > \bar x/\alpha$; hence $\hat\beta = \bar X/\alpha$ is the unique maximum.
(b) The Wald statistic is $(\hat\beta - \beta_0)$ divided by an estimated standard error, the estimate being $1/\sqrt{I_n(\hat\beta)}$ (Section 10.3.2). Since $\mathrm{E}_\beta \sum_i X_i = n\alpha\beta$,

$$\begin{aligned} I_n(\beta) &= -\mathrm{E}_\beta\left(\frac{\partial^2}{\partial\beta^2}\log L\right) = -\mathrm{E}_\beta\left(\frac{n\alpha}{\beta^2} - \frac{2\sum_i X_i}{\beta^3}\right) \\ &= -\frac{n\alpha}{\beta^2} + \frac{2n\alpha\beta}{\beta^3} = \frac{n\alpha}{\beta^2}. \end{aligned}$$

(The gamma family with α known is an exponential family, so the interchange of derivative and integral needed for the information equality is valid by Theorem 7.3.10.) Evaluating at $\hat\beta$ gives $\mathrm{se}(\hat\beta) = \hat\beta/\sqrt{n\alpha}$ and

$$Z_W = \frac{\hat\beta - \beta_0}{\hat\beta/\sqrt{n\alpha}} = \sqrt{n\alpha}\left(1 - \frac{\beta_0}{\hat\beta}\right) = \sqrt{n\alpha}\left(1 - \frac{\alpha\beta_0}{\bar X}\right),$$

which is asymptotically $\mathrm{n}(0,1)$ under H_0 by Theorem 10.1.12 and Slutsky's Theorem; reject H_0 at level α^* if $|Z_W| > z_{\alpha^*/2}$.
(c) Here $\hat\beta = \bar X/\alpha$ is a multiple of a sample mean, so $\mathrm{Var}_\beta(\hat\beta) = \mathrm{Var}_\beta(X_1)/(n\alpha^2)$; estimating $\mathrm{Var}_\beta(X_1)$ by S^2 gives $\mathrm{se}(\hat\beta) = S/(\alpha\sqrt{n})$ and

$$Z'_W = \frac{\bar X/\alpha - \beta_0}{S/(\alpha\sqrt{n})} = \frac{\sqrt{n}\,(\bar X - \alpha\beta_0)}{S}.$$

This is again asymptotically $\mathrm{n}(0,1)$ under H_0, now by the CLT and Slutsky's Theorem alone, since $S^2 \to \mathrm{Var}_\beta(X_1) = \alpha\beta^2$ in probability.

Problem 10.37. Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ population.
(a) If μ is unknown and σ^2 is known, show that $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma$ is a score statistic for testing $H_0\colon \mu = \mu_0$.
(b) If σ^2 is unknown and μ is known, find a score statistic for testing $H_0\colon \sigma = \sigma_0$.

Solution. (a) The score statistic of Section 10.3.2 is $Z_S = S(\mu_0)/\sqrt{I_n(\mu_0)}$ with $S(\mu) = \frac{\partial}{\partial \mu} \log L(\mu \mid \mathbf{X})$. Here

$$\log L(\mu \mid \mathbf{x}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 + c(\mathbf{x}), \qquad S(\mu) = \frac{n(\bar{x} - \mu)}{\sigma^2},$$

and since $\frac{\partial^2}{\partial \mu^2} \log L = -n/\sigma^2$ is constant, $I_n(\mu) = n/\sigma^2$. Hence

$$Z_S = \frac{n(\bar{X} - \mu_0)/\sigma^2}{\sqrt{n}/\sigma} = \frac{\sqrt{n}\,(\bar{X} - \mu_0)}{\sigma} = Z.$$

(b) Write the likelihood in σ (with μ known):

$$\log L(\sigma \mid \mathbf{x}) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 + c,$$

so that, with $T = \sum_i (x_i - \mu)^2$,

$$S(\sigma) = -\frac{n}{\sigma} + \frac{T}{\sigma^3}, \qquad \frac{\partial^2}{\partial \sigma^2} \log L = \frac{n}{\sigma^2} - \frac{3T}{\sigma^4}.$$

Since $\mathrm{E}_\sigma T = n\sigma^2$ (Theorem 7.3.10 applies: this is an exponential family),

$$I_n(\sigma) = -\mathrm{E}_\sigma \left(\frac{n}{\sigma^2} - \frac{3T}{\sigma^4} \right) = -\frac{n}{\sigma^2} + \frac{3n}{\sigma^2} = \frac{2n}{\sigma^2}.$$

Therefore

$$Z_S = \frac{S(\sigma_0)}{\sqrt{I_n(\sigma_0)}} = \frac{-n/\sigma_0 + T/\sigma_0^3}{\sqrt{2n}/\sigma_0} = \frac{1}{\sqrt{2n}} \left(\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma_0^2} - n \right),$$

and H_0 is rejected at level α when $|Z_S| > z_{\alpha/2}$. Under H_0, $T/\sigma_0^2 \sim \chi^2_n$, so $Z_S = (\chi^2_n - n)/\sqrt{2n}$ exactly.

Problem 10.38. Let $X_1, \ldots, X_n$ be a random sample from a gamma(α, β) population. Assume α is known and β is unknown. Consider testing $H_0\colon \beta = \beta_0$. Derive a score statistic for testing H_0.

Solution.

$$Z_S = \sqrt{\frac{n}{\alpha}} \left(\frac{\bar{X}}{\beta_0} - \alpha \right) = \sqrt{n\alpha} \left(\frac{\bar{X}}{\alpha\beta_0} - 1 \right).$$

From Exercise 10.36 the score function and information number are

$$S(\beta) = -\frac{n\alpha}{\beta} + \frac{\sum_i X_i}{\beta^2}, \qquad I_n(\beta) = \frac{n\alpha}{\beta^2},$$

so the score statistic $Z_S = S(\beta_0)/\sqrt{I_n(\beta_0)}$ of Section 10.3.2 is

$$Z_S = \frac{-n\alpha/\beta_0 + \sum_i X_i/\beta_0^2}{\sqrt{n\alpha}/\beta_0} = \frac{\sum_i X_i/\beta_0 - n\alpha}{\sqrt{n\alpha}} = \sqrt{\frac{n}{\alpha}} \left(\frac{\bar{X}}{\beta_0} - \alpha \right).$$

Reject H_0 at level α^* if $|Z_S| > z_{\alpha^*/2}$; the asymptotic $\mathrm{n}(0,1)$ null distribution is Theorem 10.1.12 with Slutsky's Theorem, the regularity conditions holding because the gamma family with α known is an exponential family.

Problem 10.39. Expand the comparisons made in Example 10.3.7.
(a) Another test based on Huber's M-estimator would be one that used a variance estimate, based on (10.3.6). Examine the performance of such a test statistic, and comment on its desirability (or lack of) as an alternative to either (10.3.8) or (10.3.9).
(b) Another test based on Huber's M-estimator would be one that used a variance from a bootstrap calculation. Examine the performance of such a test statistic.
(c) A robust competitor to $\hat\theta_M$ is the median. Examine the performance of tests of a location parameter based on the median.
For reference, $X_1, \dots, X_n$ are iid from $f(x-\theta)$ with f symmetric about 0, $\hat\theta_M$ solves $\sum_i \psi(x_i - \hat\theta_M) = 0$ with Huber's

$$\psi(x) = \begin{cases} x & \text{if } |x| \le k, \\ k & \text{if } x > k, \\ -k & \text{if } x < -k, \end{cases}$$

the asymptotic variance (10.3.7) is $\big[\int_{-k}^{k} x^2 f(x)\,dx + k^2 P_0(|X| > k)\big] / [P_0(|X| \le k)]^2$, and the three variance estimates are:

$$\widehat{\mathrm{Var}}_1(\hat\theta_M) = \frac{\frac{1}{n}\sum_{i=1}^n [\psi(x_i - \hat\theta_M)]^2}{\big[\frac{1}{n}\sum_{i=1}^n \psi'(x_i - \hat\theta_M)\big]^2},$$

which is (10.3.6);

$$\widehat{\mathrm{Var}}_2(\hat\theta_M) = \frac{\frac{1}{n}\sum_i (x_i - \hat\theta_M)^2 I(|x_i - \hat\theta_M| < k) + k^2\big(\frac{1}{n}\sum_i I(|x_i - \hat\theta_M| > k)\big)}{\big(1 - \frac{1}{n}\sum_i I(|x_i - \hat\theta_M| < k)\big)^2},$$

which is (10.3.8); and

$$\widehat{\mathrm{Var}}_3(\hat\theta_M) = \frac{1}{n}\sum_{i=1}^n (x_i - \hat\theta_M)^2,$$

which is (10.3.9). These give the statistics $Z_{GW} = \sqrt{n}(\hat\theta_M - \theta_0)/\sqrt{\widehat{\mathrm{Var}}_2}$ and $Z_N = \sqrt{n}(\hat\theta_M - \theta_0)/\sqrt{\widehat{\mathrm{Var}}_3}$ compared in Table 10.3.1.

Solution. (a) is not a new test: (10.3.6) is algebraically identical to (10.3.8), so $Z_1 \equiv Z_{GW}$. Indeed, for Huber's ψ,

$$[\psi(u)]^2 = u^2 I(|u| \le k) + k^2 I(|u| > k), \qquad \psi'(u) = I(|u| \le k),$$

(the two points $u = \pm k$ form a null set), and substituting these into (10.3.6) reproduces (10.3.8) term for term. (The denominator of (10.3.8) is misprinted in the text as $\big(1 - \frac{1}{n}\sum_i I(|x_i - \hat\theta_M| < k)\big)^2$; it must be $\big(\frac{1}{n}\sum_i I(|x_i - \hat\theta_M| < k)\big)^2$ to estimate $[P_0(|X| \le k)]^2$ in (10.3.7), and it is the corrected form that is used throughout.) So (10.3.6) is neither desirable nor undesirable as an alternative; it is the same test, and its only value is as a check that (10.3.8) is exactly the plug-in estimate of (10.3.7).
For (b) and (c), a simulation of the design of Table 10.3.1 ($n = 15$, nominal $\alpha = .1$, reject when $|Z| > z_{.05} = 1.645$, $k = 1.5$, shifts measured in units of the sampling pdf's own σ) gives the following power. Columns Z_1, Z_{GW}, Z_N use 10,000 simulations; Z_B (bootstrap variance of $\hat\theta_M$, $B = 200$ resamples) and $Z_{\mathrm{med}} = \hat\theta_{\mathrm{med}}/\sqrt{\widehat{\mathrm{Var}}_B(\hat\theta_{\mathrm{med}})}$ use 2,000.

pdf	shift	Z_1	Z_{GW}	Z_N	Z_B	Z_{med}
Normal	θ_0	.128	.128	.136	.115	.107
Normal	$+.25\sigma$	.281	.281	.296	.298	.214
Normal	$+.50\sigma$	.621	.621	.640	.603	.449
Normal	$+1.0\sigma$	.981	.981	.986	.980	.877
t_5	θ_0	.126	.126	.114	.120	.107
t_5	$+.25\sigma$	.326	.326	.300	.312	.236
t_5	$+.50\sigma$	.689	.689	.656	.671	.546
t_5	$+1.0\sigma$	.991	.991	.986	.989	.947
Logistic	θ_0	.135	.135	.124	.136	.125
Logistic	$+.25\sigma$	.301	.301	.289	.287	.218
Logistic	$+.50\sigma$	.651	.651	.650	.663	.537
Logistic	$+1.0\sigma$	.978	.978	.985	.979	.928
Double exp.	θ_0	.122	.122	.098	.120	.081
Double exp.	$+.25\sigma$	.344	.344	.293	.351	.294
Double exp.	$+.50\sigma$	.724	.724	.666	.710	.637
Double exp.	$+1.0\sigma$	.989	.989	.987	.988	.976

(All five tests have power 1.000 at $\theta_0 + 2\sigma$.) The Z_1 and Z_{GW} columns agree to every reported digit, as part (a) requires, and the Z_{GW}, Z_N columns reproduce the qualitative findings of Table 10.3.1: both sizes exceed the nominal .1 because $z_{\alpha/2}$ ignores the variability of the variance estimate, and Z_N loses ground to Z_{GW} as the tails get heavier.

(b) The bootstrap test is indistinguishable from Z_{GW}: every entry agrees with the Z_{GW} column to within simulation error ($\pm.02$ at 2,000 replications), and its size is inflated in the same way and for the same reason. Since it costs $B = 200$ M-estimations per sample to buy nothing, it is not worth using when (10.3.7) is available in closed form; its value is for ψ functions whose asymptotic variance cannot be evaluated.

(c) The median test holds its nominal level better than either Huber test (sizes $.107, .107, .125, .081$ against $.128, .126, .135, .122$) but is uniformly and substantially less powerful. The loss is largest for the normal, where at $\theta_0 + .5\sigma$ the median attains .449 against .621, and persists even for the double exponential (.637 against .724 at $\theta_0 + .5\sigma$), where the median is the MLE and its ARE against the Huber estimator is $1/.68 \approx 1.47$ by Example 10.2.7. The explanation is the standard error, not the estimator: the asymptotic variance $1/[4nf^2(0)]$ of the median depends on the unknown density at a point, and any estimate of it from $n = 15$ observations, bootstrap included, is far noisier than $\widehat{\text{Var}}_2$. The median's asymptotic advantage for sharply peaked densities is therefore not realized at $n = 15$.

Problem 10.40. In Example 10.4.5 we saw that the Poisson assumption, together with the Central Limit Theorem, could be used to form an approximate interval based on the fact that

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \to \text{n}(0, 1).$$

Show that this approximation is optimal according to Wilks (1938). That is, show that

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} = \frac{\frac{\partial}{\partial\lambda} \log L(\lambda \mid \mathbf{X})}{\sqrt{-\text{E}_\lambda \left(\frac{\partial^2}{\partial\lambda^2} \log L(\lambda \mid \mathbf{X}) \right)}}.$$

Solution. The right side is Wilks' quantity $Q(\mathbf{X} \mid \lambda)$ of (10.4.1), whose intervals (10.4.2) are asymptotically shortest; so it suffices to compute it. For $X_1, \ldots, X_n$ iid Poisson(λ),

$$\log L(\lambda \mid \mathbf{x}) = -n\lambda + \Big(\sum_{i=1}^{n} x_i \Big) \log \lambda - \sum_{i=1}^{n} \log(x_i!),$$

so

$$\frac{\partial}{\partial\lambda}\log L = -n + \frac{\sum_i x_i}{\lambda} = \frac{n(\bar{x}-\lambda)}{\lambda}, \qquad \frac{\partial^2}{\partial\lambda^2}\log L = -\frac{\sum_i x_i}{\lambda^2}.$$

Since $\mathrm{E}_\lambda \sum_i X_i = n\lambda$,

$$-\mathrm{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\log L\right) = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda},$$

and therefore

$$Q(\mathbf{X} \mid \lambda) = \frac{n(\bar{X}-\lambda)/\lambda}{\sqrt{n/\lambda}} = \frac{\sqrt{n}\,(\bar{X}-\lambda)}{\sqrt{\lambda}} = \frac{\bar{X}-\lambda}{\sqrt{\lambda/n}}.$$

The interval of Example 10.4.5 is thus exactly the set (10.4.2), which by Wilks (1938) is asymptotically the shortest in its class.

Problem 10.41. Let $X_1, \ldots, X_n$ be iid negative binomial(r, p), with pmf

$$f(x \mid r, p) = \binom{r+x-1}{x} p^r (1-p)^x, \qquad x = 0, 1, 2, \ldots,$$

mean $\mathrm{E}X = r(1-p)/p$ and variance $\operatorname{Var} X = r(1-p)/p^2$. We want to construct some approximate confidence intervals for the negative binomial parameters.
(a) Calculate Wilks' approximation (10.4.1) and show how to form confidence intervals with this expression. (Recall that $Q(\mathbf{X} \mid \theta) = \frac{\partial}{\partial\theta}\log L(\theta \mid \mathbf{X}) \big/ \sqrt{-\mathrm{E}_\theta \frac{\partial^2}{\partial\theta^2}\log L(\theta \mid \mathbf{X})}$ is asymptotically n$(0,1)$, and that $\{\theta\colon |Q(\mathbf{x} \mid \theta)| \le z_{\alpha/2}\}$ is an approximate $1-\alpha$ confidence set.)
(b) Find an approximate $1-\alpha$ confidence interval for the *mean* of the negative binomial distribution. Show how to incorporate the continuity correction into your interval.
(c) The aphid data of Exercise 9.23 can also be modeled using the negative binomial distribution. Construct an approximate 90% confidence interval for the aphid data using the results of part (b). Compare the interval to the Poisson-based intervals of Exercise 9.23. The data, the numbers of aphids per row in nine rows of a potato field, are

155 104 66 50 36 40 30 35 42

Solution. (a) With r known, $Q(\mathbf{X} \mid p) = \sqrt{n}\,[r(1-p) - p\bar{X}]/\sqrt{r(1-p)}$, and (10.4.2) is the interval between the roots of a quadratic in p. Indeed,

$$\log L(p \mid \mathbf{x}) = nr\log p + \Big(\sum_i x_i\Big)\log(1-p) + c(\mathbf{x}),$$

so

$$\frac{\partial}{\partial p}\log L = \frac{nr}{p} - \frac{\sum_i x_i}{1-p}, \qquad \frac{\partial^2}{\partial p^2}\log L = -\frac{nr}{p^2} - \frac{\sum_i x_i}{(1-p)^2}.$$

Since $\mathrm{E}_p \sum_i X_i = nr(1-p)/p$,

$$\begin{aligned} -\mathrm{E}_p\left(\frac{\partial^2}{\partial p^2}\log L\right) &= \frac{nr}{p^2} + \frac{nr(1-p)}{p(1-p)^2} \\ &= \frac{nr(1-p)+nrp}{p^2(1-p)} = \frac{nr}{p^2(1-p)}. \end{aligned}$$

Hence, writing the numerator over the common denominator $p(1-p)$,

$$\begin{aligned} Q(\mathbf{X} \mid p) &= \frac{n\big[r(1-p) - p\bar{X}\big]/[p(1-p)]}{\sqrt{nr/[p^2(1-p)]}} \\ &= \frac{\sqrt{n}\,\big[r(1-p) - p\bar{X}\big]}{\sqrt{r(1-p)}} = -\frac{\bar{X}-\mu}{\sqrt{r(1-p)/(np^2)}}, \end{aligned}$$

where $\mu = r(1-p)/p$: apart from sign, Q is just $\bar{X}$ standardized by its own mean and standard deviation. The approximate $1-\alpha$ set (10.4.2) is

$$\left\{p\colon \frac{n\big[r-p(r+\bar{x})\big]^2}{r(1-p)} \le z_{\alpha/2}^2\right\},$$

and clearing the denominator shows this is $\{p\colon Ap^2 - Bp + C \le 0\}$ with

$$\begin{aligned} A &= n(r+\bar{x})^2, \\ B &= r\big[2n(r+\bar{x}) - z_{\alpha/2}^2\big], \\ C &= r\big(nr - z_{\alpha/2}^2\big). \end{aligned}$$

Since $A > 0$, the set is the interval between the roots,

$$\frac{B - \sqrt{B^2-4AC}}{2A} \;\le\; p \;\le\; \frac{B + \sqrt{B^2-4AC}}{2A},$$

intersected with $(0,1)$.
(b) The mean $\mu = r(1-p)/p$ is a strictly decreasing function of p, so inverting the interval of part (a) endpoint-for-endpoint is legitimate; equivalently, substitute $p = r/(r+\mu)$ directly into Q. Because

$$\operatorname{Var} X = \frac{r(1-p)}{p^2} = \frac{\mu}{p} = \mu + \frac{\mu^2}{r},$$

the set $\{|Q| \le z_{\alpha/2}\}$ reads, in terms of μ,

$$\left\{\mu\colon \frac{|\bar{x}-\mu|}{\sqrt{(\mu+\mu^2/r)/n}} \le z_{\alpha/2}\right\}.$$

Squaring gives the quadratic $a\mu^2 - b\mu + c \le 0$ with

$$a = 1 - \frac{z_{\alpha/2}^2}{nr}, \qquad b = 2\bar{x} + \frac{z_{\alpha/2}^2}{n}, \qquad c = \bar{x}^2,$$

so that (when $a > 0$) the approximate $1-\alpha$ interval for μ is

$$\frac{b - \sqrt{b^2-4ac}}{2a} \;\le\; \mu \;\le\; \frac{b + \sqrt{b^2-4ac}}{2a}.$$

The data are integer-valued and $n\bar{X} = \sum_i X_i$ is an integer, so the continuity correction of Example 3.3.2 replaces $|\bar{x}-\mu|$ by $\big(|\bar{x}-\mu| - \frac{1}{2n}\big)^+$, i.e. it moves $\bar{x}$ half a unit of $\sum x_i$ away from μ on each side. Operationally: compute the lower endpoint from the quadratic with $\bar{x}$ replaced by $\bar{x} - \frac{1}{2n}$ and the upper endpoint from the quadratic with $\bar{x}$ replaced by $\bar{x} + \frac{1}{2n}$. This can only lengthen the interval.
(c) For the aphid data $n = 9$, $\sum x_i = 558$, $\bar{x} = 62$, and $s^2 = 1735.75$. The interval of part (b) needs r; matching moments in $\operatorname{Var} X = \mu + \mu^2/r$ gives

$$\hat{r} = \frac{\bar{x}^2}{s^2 - \bar{x}} = \frac{62^2}{1735.75 - 62} = 2.297.$$

With $z_{.05} = 1.645$, $a = 1 - 1.645^2/(9 \cdot 2.297) = .8691$, $b = 2(62) + 1.645^2/9 = 124.301$, $c = 3844$, so $b^2 - 4ac = 2087.6$ and the 90% interval is

$$45.23 \;\le\; \mu \;\le\; 97.80,$$

becoming (45.19, 97.89) with the continuity correction. (The simpler Wald form $\bar{x} \pm z_{.05}\sqrt{(\bar{x}+\bar{x}^2/\hat{r})/n}$ is $\bar{x} \pm z_{.05}s/\sqrt{n} = (39.2,\ 84.8)$, since $\bar{x} + \bar{x}^2/\hat{r} = s^2$ by construction of $\hat{r}$; the score form above is preferable, being asymmetric in the direction the skewness demands.)
The Poisson-based intervals of Exercise 9.23 are (57.78, 66.42) from the LRT and (57.75, 66.49) from the method of Example 9.2.15 — about one-sixth the length of the negative binomial interval. The Poisson model forces $\operatorname{Var} X = \mu$, i.e. $s^2 \approx 62$, whereas the observed $s^2 = 1735.75$ is 28 times $\bar{x}$; the negative binomial accommodates that overdispersion through r.

Problem 10.42. Show that (10.4.5) is equivalent to the highest likelihood region (9.2.7) in that for any fixed α level, they will produce the same confidence set. Here (10.4.5) is the approximate $1-\alpha$ set based on the asymptotic chi squared distribution of $-2\log\lambda(\mathbf{x})$,

$$\left\{\theta\colon\ -2\log\left(\frac{L(\theta\mid\mathbf{x})}{L(\hat\theta\mid\mathbf{x})}\right)\le\chi^2_{1,\alpha}\right\},$$

$\hat\theta$ being the MLE, and (9.2.7) is the region obtained by inverting the LRT of $H_0\colon\theta=\theta_0$ versus $H_1\colon\theta\neq\theta_0$, which accepts H_0 when $L(\theta_0\mid\mathbf{x})/L(\hat\theta\mid\mathbf{x})\ge k(\theta_0)$ and so has the form

$$\{\theta\colon L(\theta\mid\mathbf{x})\ge k'(\mathbf{x},\theta)\},$$

for a function k' giving $1-\alpha$ confidence.

Solution. The two sets coincide with $k'(\mathbf{x},\theta)=e^{-\chi^2_{1,\alpha}/2}L(\hat\theta\mid\mathbf{x})$, since $t\mapsto e^{-t/2}$ is a strictly decreasing bijection of $[0,\infty)$ onto $(0,1]$:

$$\begin{aligned}-2\log\frac{L(\theta\mid\mathbf{x})}{L(\hat\theta\mid\mathbf{x})}\le\chi^2_{1,\alpha}\ &\iff\ \log\frac{L(\theta\mid\mathbf{x})}{L(\hat\theta\mid\mathbf{x})}\ge-\tfrac12\chi^2_{1,\alpha}\\ &\iff\ \frac{L(\theta\mid\mathbf{x})}{L(\hat\theta\mid\mathbf{x})}\ge e^{-\chi^2_{1,\alpha}/2}\\ &\iff\ L(\theta\mid\mathbf{x})\ge e^{-\chi^2_{1,\alpha}/2}L(\hat\theta\mid\mathbf{x}),\end{aligned}$$

every step being an equivalence because log and exp are strictly increasing and $L(\hat\theta\mid\mathbf{x})>0$. So (10.4.5) is exactly the region (9.2.7) of Section 9.2.3 produced by the LRT acceptance region with the constant cutoff $k(\theta_0)\equiv e^{-\chi^2_{1,\alpha}/2}$.
Reading the chain upward gives the converse: a region (9.2.7) with cutoff $k(\theta_0)\equiv k\in(0,1]$ constant in θ_0 is the set (10.4.5) with $\chi^2_{1,\alpha}$ replaced by $-2\log k$, and $\alpha\mapsto\chi^2_{1,\alpha}$ sweeps out all of $[0,\infty)$, so every such k is matched by some α.

10.7 Exercises 10.43–10.48

Problem 10.43. In Example 10.4.7, two modifications were made to the Wald interval. (There $Y=\sum_{i=1}^n X_i$ with $X_1,\dots,X_n$ iid Bernoulli(p), $\hat p=Y/n$, and the Wald interval (10.4.8) is

$$\hat p-z_{\alpha/2}\sqrt{\tfrac{\hat p(1-\hat p)}{n}}\ \le\ p\ \le\ \hat p+z_{\alpha/2}\sqrt{\tfrac{\hat p(1-\hat p)}{n}}.)$$

(a) At $y=0$ the upper interval endpoint was changed to $1-(\alpha/2)^{1/n}$, and at $y=n$ the lower interval endpoint was changed to $(\alpha/2)^{1/n}$. Justify the choice of these endpoints. (Hint: see Section 9.2.3.)
(b) The second modification was to truncate all intervals to be within $[0,1]$. Show that this change, together with the one in part (a), results in an improvement over the original Wald interval.

Solution. (a) They are the Clopper-Pearson endpoints of Theorem 9.2.14 at the two boundary counts. $Y\sim$ binomial(n,p) has $F_Y(y\mid p)$ decreasing in p for each y, so case (i) of that theorem defines $\theta_L(y),\theta_U(y)$ by

$$P(Y\le y\mid p_U(y))=\tfrac{\alpha}{2},\qquad P(Y\ge y\mid p_L(y))=\tfrac{\alpha}{2}.$$

At $y=0$ the second equation has no solution ($P(Y\ge0\mid p)\equiv1$), so $p_L(0)=0$, while the first reads

$$(1-p_U)^n=\tfrac{\alpha}{2}\quad\Longrightarrow\quad p_U(0)=1-(\alpha/2)^{1/n}.$$

At $y=n$ the first equation has no solution ($P(Y\le n\mid p)\equiv1$), so $p_U(n)=1$, and the second reads

$$p_L^n=\tfrac{\alpha}{2}\quad\Longrightarrow\quad p_L(n)=(\alpha/2)^{1/n}.$$

Some such repair is forced: at $y = 0$ one has $\hat{p}(1-\hat{p}) = 0$, so the Wald interval collapses to the single point $\{0\}$ (likewise $\{1\}$ at $y = n$), covering no $p \in (0,1)$.
(b) Write $C(y)$ for the Wald interval and $C^*(y)$ for the modified one, so that

$$C^*(y) = \begin{cases} [0,\, 1-(\alpha/2)^{1/n}], & y = 0, \\ C(y) \cap [0,1], & 1 \le y \le n-1, \\ [(\alpha/2)^{1/n},\, 1], & y = n. \end{cases}$$

For $1 \le y \le n-1$ truncation deletes only points outside the parameter space, so

$$\{p \in C^*(y)\} = \{p \in C(y)\} \quad \text{for every } p \in [0,1],$$

i.e. the coverage contribution of those y is unchanged while the length is weakly (and, whenever an endpoint spills out of $[0,1]$, strictly) reduced. Only $y \in \{0, n\}$ changes the coverage, and there $C(0) = \{0\}$ and $C(n) = \{1\}$ cover no $p \in (0,1)$ at all, whence for every $p \in (0,1)$

$$\begin{aligned} P_p\big(p \in C^*(Y)\big) - P_p\big(p \in C(Y)\big) &= (1-p)^n\, I\Big(p \le 1-(\alpha/2)^{1/n}\Big) \\ &\quad + p^n\, I\Big(p \ge (\alpha/2)^{1/n}\Big) \;\ge\; 0, \end{aligned}$$

strictly positive on $\big(0, 1-(\alpha/2)^{1/n}\big] \cup \big[(\alpha/2)^{1/n}, 1\big)$, the two neighbourhoods of the endpoints where the Wald interval was worst. So C^* has coverage nowhere smaller, strictly larger near 0 and 1, and length no larger for any $1 \le y \le n-1$: an improvement in both criteria. For $n = 12$, $\alpha = .1$ the Wald interval has infimum coverage 0 and average .757, against .749 and .905 for C^*.

Problem 10.44. Agresti and Coull (1998) "strongly recommend" the score interval for a binomial parameter but are concerned that a formula such as (10.4.7) might be a bit formidable for an elementary course in statistics. To produce a reasonable binomial interval with an easier formula, they suggest the following modification to the Wald interval: Add 2 successes and 2 failures; then use the original Wald formula (10.4.8). That is, use $\tilde{p} = (y+2)/(n+4)$ instead of $\hat{p} = y/n$. Using both length and coverage probability, compare this interval to the binomial score interval. Do you agree that it is a reasonable alternative to the score interval?
(Samuels and Lu 1992 suggest another modification to the Wald interval based on sample sizes. Agresti and Caffo 2000 extend these improved approximate intervals to the two sample problem.)
Here (10.4.7) gives the endpoints of the score interval as the two roots

$$\frac{2\hat{p} + z^2_{\alpha/2}/n \pm \sqrt{(2\hat{p} + z^2_{\alpha/2}/n)^2 - 4\hat{p}^2(1 + z^2_{\alpha/2}/n)}}{2(1 + z^2_{\alpha/2}/n)},$$

and (10.4.8) is the Wald interval $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$.

Solution. Yes: the Agresti-Coull interval has **exactly** the score interval's midpoint and a half-length that is never smaller, so it contains the score interval and therefore has uniformly at least its coverage. Write $z = z_{\alpha/2}$ and $\tilde{n} = n + z^2$ (so "add 2 and 2" is the case $z^2 = 4$, i.e. $\alpha \approx .05$, where $z^2_{.025} = 3.84$), and use $\tilde{p} = (y + z^2/2)/\tilde{n}$ in the Wald formula with $\tilde{n}$ in place of n:

$$\tilde{p} \pm z\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}.$$

The midpoint: the two roots in (10.4.7) average to

$$\frac{2\hat{p} + z^2/n}{2(1 + z^2/n)} = \frac{\hat{p} + z^2/(2n)}{1 + z^2/n} = \frac{y + z^2/2}{n + z^2} = \tilde{p}.$$

So the score interval is centred at $\tilde{p}$, not at $\hat{p}$; Agresti and Coull simply take that centre.

Half-lengths: simplifying the discriminant in (10.4.7),

$$\left(2\hat{p}+\tfrac{z^2}{n}\right)^2-4\hat{p}^2\left(1+\tfrac{z^2}{n}\right)=\frac{4z^2}{n}\hat{p}(1-\hat{p})+\frac{z^4}{n^2},$$

so with $n\hat{p}(1-\hat{p})=y(n-y)/n$ the score half-length is

$$h_S=\frac{z}{\tilde{n}}\sqrt{\frac{y(n-y)}{n}+\frac{z^2}{4}},\qquad h_{AC}=\frac{z}{\tilde{n}}\sqrt{\frac{(y+\frac{z^2}{2})(n-y+\frac{z^2}{2})}{\tilde{n}}}.$$

Expanding the numerator of h_{AC}^2 gives $y(n-y)+\frac{z^2n}{2}+\frac{z^4}{4}$, hence

$$\begin{aligned}\frac{\tilde{n}^2}{z^2}\left(h_{AC}^2-h_S^2\right)&=\frac{y(n-y)+\frac{z^2n}{2}+\frac{z^4}{4}}{n+z^2}-\frac{y(n-y)+\frac{z^2n}{4}}{n}\\&=\frac{z^2\left[\frac{n^2}{4}-y(n-y)\right]}{n(n+z^2)}=\frac{z^2\left(\frac{n}{2}-y\right)^2}{n(n+z^2)}\;\geq\;0,\end{aligned}$$

with equality only at $y=n/2$. Thus $h_{AC}\geq h_S$ always, and since the centres agree,

$$\text{score interval}\;\subseteq\;\text{Agresti-Coull interval}\quad\Longrightarrow\quad P_p(\text{AC covers})\;\geq\;P_p(\text{score covers})\;\;\forall p.$$

The inflation is worst at the boundary counts, where $h_{AC}/h_S=\sqrt{(2n+z^2)/(n+z^2)}\to\sqrt{2}$, and decreases monotonically to 1 at $y=n/2$ (the numerator of $h_{AC}^2-h_S^2$ falls and h_S rises as $y\to n/2$); at $n=20$, $\alpha=.05$ it is 1.36 at $y=0$ but already under 1.06 by $y=3$.
Coverage (nominal .95, exact binomial computation, infimum and average over $p\in(0,1)$):

n	score inf	score avg	AC inf	AC avg
12	.836	.954	.922	.964
20	.837	.953	.929	.962
50	.838	.952	.935	.958

So yes: the price of the simpler formula is a mild conservatism – average coverage about one point high, bought back as a much better-behaved infimum, paid at the extreme y where the score interval was the one dipping below nominal. The one caveat is that, like the Wald interval, its endpoints can leave $[0,1]$ (at $n=12$, $y=12$, $\alpha=.05$ the upper endpoint is 1.04) and must be truncated.

Problem 10.45. Solve for the endpoints of the approximate binomial confidence interval, with continuity correction, given in Example 10.4.6. Show that this interval is wider than the corresponding interval without continuity correction, and that the continuity corrected interval has a uniformly higher coverage probability. (In fact, the coverage probability of the uncorrected interval does not maintain $1-\alpha$; it dips below this level for some parameter values. The corrected interval does maintain a coverage probability greater than $1-\alpha$ for all parameter values.)
In Example 10.4.6, with $\hat{p}=\sum x_i/n$, the continuity-corrected score interval is obtained by solving the two separate quadratics

$$\left|\frac{\hat{p}+\frac{1}{2n}-p}{\sqrt{p(1-p)/n}}\right|\leq z_{\alpha/2}\quad\text{(larger root = upper endpoint)},$$

$$\left|\frac{\hat{p}-\frac{1}{2n}-p}{\sqrt{p(1-p)/n}}\right|\leq z_{\alpha/2}\quad\text{(smaller root = lower endpoint)},$$

with the conventions that the lower endpoint is 0 if $\sum x_i=0$ and the upper endpoint is 1 if $\sum x_i=n$.

Solution. Write $z = z_{\alpha/2}$ and, for a real number t, let $g(t) \le h(t)$ be the two roots of

$$Q_t(p) = (t-p)^2 - \frac{z^2}{n}p(1-p) = \Big(1+\tfrac{z^2}{n}\Big)p^2 - \Big(2t+\tfrac{z^2}{n}\Big)p + t^2.$$

The algebra of Example 10.4.6 (with $\hat p$ replaced by t), together with the discriminant simplification $\big(2t+\frac{z^2}{n}\big)^2 - 4t^2\big(1+\frac{z^2}{n}\big) = \frac{4z^2}{n}t(1-t) + \frac{z^4}{n^2}$, gives

$$g(t), h(t) = \frac{t + \frac{z^2}{2n} \mp \frac{z}{\sqrt{n}}\sqrt{t(1-t)+\frac{z^2}{4n}}}{1+\frac{z^2}{n}}.$$

Hence the corrected endpoints are

$$L_{cc} = g\Big(\hat p - \tfrac{1}{2n}\Big), \qquad U_{cc} = h\Big(\hat p + \tfrac{1}{2n}\Big),$$

that is, explicitly, with $\hat p_\mp = \hat p \mp \frac{1}{2n}$,

$$L_{cc} = \frac{\hat p_- + \frac{z^2}{2n} - \frac{z}{\sqrt{n}}\sqrt{\hat p_-(1-\hat p_-)+\frac{z^2}{4n}}}{1+\frac{z^2}{n}},$$

$$U_{cc} = \frac{\hat p_+ + \frac{z^2}{2n} + \frac{z}{\sqrt{n}}\sqrt{\hat p_+(1-\hat p_+)+\frac{z^2}{4n}}}{1+\frac{z^2}{n}},$$

while the uncorrected interval of (10.4.7) is $[L,U] = [g(\hat p), h(\hat p)]$, and at $y=0$, $y=n$ one takes $L_{cc}=0$, $U_{cc}=1$ as in Example 10.4.6.

Both roots are strictly increasing on $t \in [0,1]$. Let $0 \le t_1 < t_2 \le 1$. Since $Q_t(t) = -\frac{z^2}{n}t(1-t) \le 0$ there and Q_t opens upward, $g(t) \le t \le h(t)$. For any $p \ge t_2$ we have $|p-t_1| > |p-t_2|$, so $Q_{t_1}(p) > Q_{t_2}(p)$; taking $p = h(t_2)$ $(\ge t_2)$ gives $Q_{t_1}(h(t_2)) > 0$. As $h(t_2) \ge t_2 > t_1 \ge g(t_1)$, the point $h(t_2)$ lies above the smaller root of the upward parabola Q_{t_1}, and a point where $Q_{t_1} > 0$ above its smaller root must lie above its larger root: $h(t_1) < h(t_2)$. Symmetrically, for $p \le t_1$ we have $Q_{t_2}(p) > Q_{t_1}(p)$; taking $p = g(t_1)$ gives $Q_{t_2}(g(t_1)) > 0$ with $g(t_1) \le t_1 < t_2 \le h(t_2)$, so $g(t_1) < g(t_2)$.

Wider, and containing: for $1 \le y \le n-1$ all three of $\hat p_- < \hat p < \hat p_+$ lie in $[0,1]$, so

$$L_{cc} = g(\hat p_-) < g(\hat p) = L \le U = h(\hat p) < h(\hat p_+) = U_{cc}.$$

At $y=0$, $g(0)=0=L_{cc}$ and only the upper endpoint moves, $U = h(0) < h(1/(2n)) = U_{cc}$; at $y=n$, symmetrically. So $[L,U] \subsetneq [L_{cc}, U_{cc}]$ for every y: the corrected interval is strictly wider.

Uniformly higher coverage: containment for every y means $\{p \in [L(Y), U(Y)]\} \subseteq \{p \in [L_{cc}(Y), U_{cc}(Y)]\}$ as events, hence for every p

$$P_p\big(L_{cc}(Y) \le p \le U_{cc}(Y)\big) \;\ge\; P_p\big(L(Y) \le p \le U(Y)\big).$$

For $n=12$, $\alpha = .1$ the two coverage functions bear out the parenthetical remark: over $p \in (0,1)$ the uncorrected interval has infimum coverage .796 (average .910), dipping below the nominal .9, whereas the corrected interval has infimum .913 > .9 (average .959). The realized intervals are

y	$[L,U]$	$[L_{cc},U_{cc}]$
0	[.000, .184]	[.000, .246]
3	[.105, .487]	[.080, .529]
6	[.286, .714]	[.252, .748]
9	[.513, .895]	[.471, .920]
12	[.816, 1.000]	[.754, 1.000]

Problem 10.46. Expand the comparisons made in Example 10.4.8.
(a) Produce a table similar to Table 10.4.2 that examines the robustness of intervals for a location parameter based on the median. (Intervals based on the mean are done in Table 10.4.1.)
(b) Another interval based on Huber's M-estimator would be one that used a variance from a bootstrap calculation. Examine the robustness of such an interval.
In Example 10.4.8, $X_1, \dots, X_n$ are iid from $f(x-\theta)$ with f symmetric about 0, $\hat\theta_M$ is the Huber M-estimator (the ρ of (10.2.2) and $\psi(x) = \max\{-k, \min(k,x)\}$ of (10.2.7), with $k = 1.5$), and the interval is

$$\hat\theta_M \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat\theta_M)/n},$$

with the asymptotic variance (10.3.7) replaced by the estimate

$$\widehat{\mathrm{Var}}_2(\hat\theta_M) = \frac{\frac{1}{n}\sum_i (x_i-\hat\theta_M)^2 I(|x_i-\hat\theta_M|<k) + k^2\big(\frac{1}{n}\sum_i I(|x_i-\hat\theta_M|>k)\big)}{\big(1-\frac{1}{n}\sum_i I(|x_i-\hat\theta_M|<k)\big)^2}$$

of (10.3.8), or by the naive estimate $\widehat{\mathrm{Var}}_3(\hat\theta_M) = \frac{1}{n}\sum_i (x_i-\hat\theta_M)^2$ of (10.3.9). Table 10.4.2 reports the resulting confidence coefficients for nominal $1-\alpha = .9$, $n = 15$, based on 10,000 simulations:

	Normal	t_5	Logistic	Double exponential
Variance estimate (10.3.8)	.844	.856	.855	.889
Variance estimate (10.3.9)	.837	.867	.855	.910

Solution. (a) Median intervals are noticeably better behaved than the Huber intervals of Table 10.4.2, provided the variance estimate respects the median's asymptotic variance $1/[2f(0)]^2$ (Example 10.2.3) rather than borrowing a spread estimate from the mean.

Take $M_n \pm z_{\alpha/2}\sqrt{\widehat V/n}$ with three choices of $\widehat V$, plus the exact distribution-free competitor:

1. $\widehat V_f = 1/[2\hat f_n(M_n)]^2$, with $\hat f_n$ the Gaussian kernel density estimate at M_n using the Silverman bandwidth $h = 0.9\min\{s,\ \mathrm{IQR}/1.34\}n^{-1/5}$;
2. $\widehat V_{\text{naive}} = \frac{1}{n}\sum_i (x_i - M_n)^2$ – the exact analogue of (10.3.9);
3. $\widehat V_{\text{boot}} = n\cdot \widehat{\mathrm{Var}}^*(M_n^*)$ from $B = 200$ nonparametric bootstrap resamples;
4. the order-statistic (sign) interval $[X_{(4)}, X_{(12)}]$, whose coverage is **exactly** $\sum_{j=4}^{11}\binom{15}{j}2^{-15} = .9648$ for every continuous symmetric f.

Simulating 10,000 samples of size $n = 15$ at nominal $1-\alpha = .9$ (standard versions of the four pdfs; all four intervals are scale-equivariant, so the answers do not depend on the scale chosen):

	Normal	t_5	Logistic	Double exponential
$\widehat V_f$ (kernel)	.891	.899	.899	.930
$\widehat V_{\text{naive}}$ (10.3.9 analogue)	.797	.847	.838	.912
$\widehat V_{\text{boot}}$	.891	.894	.892	.921
$[X_{(4)}, X_{(12)}]$ (exact)	.967	.965	.965	.964

Two things stand out against Table 10.4.2. First, once $\widehat V$ estimates the right quantity, rows 1 and 3 sit within a point of .9 at the normal, t_5 and logistic – better than either row of Table 10.4.2 – and run conservative at the double exponential, where $f(0) = 1/2$ is large and the median is the MLE. Second, row 2 estimates the wrong quantity – $\frac{1}{n}\sum(x_i - M_n)^2 \to \sigma^2$, not $1/[2f(0)]^2$ – and the sign of the error predicts the row exactly:

$$\text{normal:}\quad \sigma^2 = 1 \ <\ \tfrac{\pi}{2} = \tfrac{1}{[2f(0)]^2},$$
$$\text{double exponential:}\quad \sigma^2 = 2 \ >\ 1 = \tfrac{1}{[2f(0)]^2},$$

so the interval is too short at the normal (.797) and too long at the double exponential (.912). The sign interval is the only genuinely distribution-free row, at the price of a confidence coefficient stuck at .9648.
(b) The bootstrap interval $\hat\theta_M \pm z_{.05}\,\widehat{\mathrm{se}}^*(\hat\theta_M)$ is as robust as the one built from (10.3.8) and markedly more stable than the naive (10.3.9), but it does not cure the shortfall of Table 10.4.2.

(As printed, the denominator of (10.3.8) reads $\big(1-\frac{1}{n}\sum I(|x_i-\hat\theta_M|<k)\big)^2$; (10.3.6) and (10.3.7) show the intended denominator is $\big(\frac{1}{n}\sum I(|x_i-\hat\theta_M|<k)\big)^2$, which is what is used below.) With $k=1.5$, $n=15$, $B=200$ resamples and 10,000 simulations:

	Normal	t_5	Logistic	Double exponential
Bootstrap standard error	.864	.878	.877	.887
Variance estimate (10.3.8)	.865	.876	.868	.884
Variance estimate (10.3.9)	.859	.886	.876	.904

The bootstrap row tracks (10.3.8) to within one simulation standard error at every pdf, and both vary far less across pdfs (ranges .023 and .019) than the naive estimate (range .045, running from .859 at the normal to .904 at the double exponential). All three stay two to four points below nominal, for the reason given in Example 10.3.7: the $z_{\alpha/2}$ cutoff ignores the variability of the estimated standard error, and at $n=15$ the bootstrap adds variability of its own rather than removing it.

Problem 10.47. Let $X_1,\dots,X_n$ be iid negative binomial(r,p).
(a) Complete the details of Example 10.4.9; that is, show that for small p, the interval

$$\left\{p:\ \frac{\chi^2_{2nr,1-\alpha/2}}{2\sum x}\ \le\ p\ \le\ \frac{\chi^2_{2nr,\alpha/2}}{2\sum x}\right\}$$

is an approximate $1-\alpha$ confidence interval.
(b) Show how to choose the endpoints in order to obtain a minimum length $1-\alpha$ interval.
(In Example 10.4.9, r is known, $Y=\sum X_i\sim$ negative binomial(nr,p), and Exercise 2.38 establishes that $2pY\to\chi^2_{2nr}$ in distribution as $p\to0$, so that $2pY$ is a pivot for small p.)

Solution. (a) $2pY$ has a limiting distribution free of p, so it is an (approximate) pivot. With $Y=\sum X_i\sim$ negative binomial(nr,p), whose mgf is $M_Y(t)=\big[p/(1-(1-p)e^t)\big]^{nr}$ for $t<-\log(1-p)$,

$$\begin{aligned}M_{2pY}(t)=M_Y(2pt)&=\left(\frac{p}{1-(1-p)e^{2pt}}\right)^{nr},\\ 1-(1-p)e^{2pt}&=1-(1-p)\big(1+2pt+O(p^2)\big)\\&=p(1-2t)+O(p^2),\end{aligned}$$

so that for each fixed $t<1/2$ (which also keeps $2pt<p\le-\log(1-p)$, so the mgf exists)

$$M_{2pY}(t)\ \longrightarrow\ \left(\frac{1}{1-2t}\right)^{nr}\qquad\text{as } p\to0,$$

the mgf of a χ^2_{2nr}. Convergence of mgfs on a neighbourhood of 0 gives $2pY\to\chi^2_{2nr}$ in distribution (Theorem 2.3.12), which is Exercise 2.38. Hence for small p, with $\chi^2_{\nu,\gamma}$ the upper-γ cutoff,

$$P_p\left(\chi^2_{2nr,1-\alpha/2}\le 2pY\le\chi^2_{2nr,\alpha/2}\right)\approx1-\alpha,$$

and since $Y>0$ the event may be solved for p without reversing the inequalities:

$$\chi^2_{2nr,1-\alpha/2}\le 2pY\le\chi^2_{2nr,\alpha/2}\iff\frac{\chi^2_{2nr,1-\alpha/2}}{2Y}\le p\le\frac{\chi^2_{2nr,\alpha/2}}{2Y}.$$

(If $Y=0$ the endpoints are undefined; take $[0,1]$, an event of probability $p^{nr}\to0$.)
(b) Replace the equal-tail cutoffs by any $a<b$ with $P(a\le\chi^2_{2nr}\le b)=1-\alpha$; the interval is $[a/(2\sum x),\ b/(2\sum x)]$, of length $(b-a)/(2\sum x)$. Minimizing the length therefore means minimizing $b-a$ subject to $\int_a^b f(t)\,dt=1-\alpha$, where f is the χ^2_{2nr} pdf. That pdf is unimodal with mode $2nr-2$, so Theorem 9.3.2 applies: the shortest such interval is the one with

$$f(a)=f(b)>0,\qquad a\le2nr-2\le b.$$

Since $f(t) \propto t^{nr-1}e^{-t/2}$, the pair (a,b) is the solution of

$$(nr-1)\log\frac{b}{a} = \frac{b-a}{2}, \qquad F_{2nr}(b) - F_{2nr}(a) = 1-\alpha,$$

solved numerically. For $nr = 5$, $\alpha = .1$: $[a,b] = [3.017, 16.711]$, length 13.693, against the equal-tail [3.940, 18.307] of length 14.367 – a 4.7% reduction; for $nr = 10$: [9.786, 29.876] (length 20.090) against [10.851, 31.410] (length 20.560). (For $nr = 1$ the χ^2_2 pdf is decreasing, so the mode condition of Theorem 9.3.2 cannot be met and the shortest interval is $a = 0$, $b = \chi^2_{2,\alpha}$.)

Problem 10.48. For the case of Fieller's confidence set (see Miscellanea 9.5.3), that is, given a random sample $(X_1,Y_1),\dots,(X_n,Y_n)$ from a bivariate normal distribution with parameters $(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$, find an approximate confidence interval for $\theta = \mu_Y/\mu_X$. Use the approximate moment calculations in Example 5.5.27 and apply the Central Limit Theorem.
Example 5.5.27 gives, for $g(\mu_X,\mu_Y) = \mu_X/\mu_Y$, the partials $\partial g/\partial\mu_X = 1/\mu_Y$ and $\partial g/\partial\mu_Y = -\mu_X/\mu_Y^2$, and hence

$$\operatorname{Var}\left(\frac{X}{Y}\right) \approx \left(\frac{\mu_X}{\mu_Y}\right)^2 \left(\frac{\operatorname{Var}X}{\mu_X^2} + \frac{\operatorname{Var}Y}{\mu_Y^2} - 2\frac{\operatorname{Cov}(X,Y)}{\mu_X\mu_Y}\right).$$

Solution. The interval is

$$\frac{\bar Y}{\bar X} \;\pm\; \frac{z_{\alpha/2}}{\sqrt{n}}\,\frac{\sqrt{S_Y^2 - 2\hat\theta S_{XY} + \hat\theta^{\,2} S_X^2}}{|\bar X|}, \qquad \hat\theta = \frac{\bar Y}{\bar X},$$

with S_X^2, S_Y^2, S_{XY} the sample variances and covariance.
Apply the Multivariate Delta Method (Theorem 5.5.28) to $g(\mu_X,\mu_Y) = \mu_Y/\mu_X$, whose partials are those of Example 5.5.27 with the roles of X and Y interchanged:

$$\frac{\partial g}{\partial\mu_Y} = \frac{1}{\mu_X}, \qquad \frac{\partial g}{\partial\mu_X} = -\frac{\mu_Y}{\mu_X^2}.$$

Its hypotheses hold: $(\bar X,\bar Y)$ is the mean of an iid bivariate sample with finite covariances, g has continuous partials at (μ_X,μ_Y) provided $\mu_X \neq 0$, and

$$\begin{aligned}\tau^2 &= \sum_i\sum_j \sigma_{ij}\frac{\partial g}{\partial\mu_i}\frac{\partial g}{\partial\mu_j}\\ &= \left(\frac{\mu_Y}{\mu_X}\right)^2\left(\frac{\sigma_Y^2}{\mu_Y^2} + \frac{\sigma_X^2}{\mu_X^2} - 2\frac{\rho\sigma_X\sigma_Y}{\mu_X\mu_Y}\right)\\ &= \frac{1}{\mu_X^2}\left(\sigma_Y^2 - 2\theta\rho\sigma_X\sigma_Y + \theta^2\sigma_X^2\right) > 0\end{aligned}$$

unless the bivariate normal is degenerate. Hence

$$\sqrt{n}\left(\frac{\bar Y}{\bar X} - \theta\right) \;\longrightarrow\; \mathrm{n}(0,\tau^2) \quad \text{in distribution,}$$

which is Fieller's $\bar Z_\theta = \bar Y - \theta\bar X$ of Miscellanea 9.5.3 divided by $\bar X$: $\tau^2 = n\operatorname{Var}(\bar Z_\theta)/\mu_X^2$.
By the WLLN $S_X^2 \to \sigma_X^2$, $S_Y^2 \to \sigma_Y^2$, $S_{XY} \to \rho\sigma_X\sigma_Y$, $\bar X \to \mu_X$ in probability, and $\hat\theta \to \theta$ by the Continuous Mapping Theorem (Theorem 5.5.4), so

$$\hat\tau^2 = \frac{S_Y^2 - 2\hat\theta S_{XY} + \hat\theta^{\,2}S_X^2}{\bar X^2} \;\longrightarrow\; \tau^2 \quad \text{in probability.}$$

Slutsky's Theorem (Theorem 5.5.17) then gives

$$\frac{\hat\theta - \theta}{\hat\tau/\sqrt{n}} \;\longrightarrow\; \mathrm{n}(0,1),$$

so $\hat\theta \pm z_{\alpha/2}\hat\tau/\sqrt{n}$ is an approximate $1-\alpha$ interval, which is the display above.

11 Analysis of Variance and Regression

Contents

11.1 Exercises 11.1–11.7 426
11.2 Exercises 11.8–11.14 430
11.3 Exercises 11.15–11.21 436
11.4 Exercises 11.22–11.28 443
11.5 Exercises 11.29–11.35 448
11.6 Exercises 11.36–11.41 454

11.1 Exercises 11.1–11.7

Problem 11.1. An ANOVA variance-stabilizing transformation stabilizes variances in the following approximate way. Let Y have mean θ and variance $v(\theta)$.
(a) Use arguments as in Section 10.1.3 to show that a one-term Taylor series approximation of the variance of $g(Y)$ is given by

$$\operatorname{Var}\big(g(Y)\big) = \Big[\tfrac{d}{d\theta} g(\theta)\Big]^2 v(\theta).$$

(b) Show that the approximate variance of $g^*(Y)$ is independent of θ, where

$$g^*(y) = \int \Big[1/\sqrt{v(y)}\Big]\, dy.$$

Solution. (a) Expand g about the mean and keep one term:

$$g(Y) \;=\; g(\theta) + g'(\theta)(Y-\theta) + R,$$

and drop R. The surviving expression is an affine function of Y, so

$$\operatorname{Var}\big(g(Y)\big) \;\approx\; \big[g'(\theta)\big]^2 \operatorname{Var}(Y) \;=\; \big[g'(\theta)\big]^2 v(\theta),$$

the Delta Method variance of Theorem 5.5.24 as it is used in Section 10.1.3 (its hypothesis is that g is differentiable at θ with $g'(\theta) \neq 0$).
(b) By the Fundamental Theorem of Calculus, $g^{*\prime}(\theta) = 1/\sqrt{v(\theta)}$ ($v > 0$ and continuous makes the integrand continuous). Substituting into (a),

$$\operatorname{Var}\big(g^*(Y)\big) \;\approx\; \left[\frac{1}{\sqrt{v(\theta)}}\right]^2 v(\theta) \;=\; 1$$

for every θ.

Problem 11.2. Verify that the following transformations are approximately variance-stabilizing in the sense of Exercise 11.1.
(a) $Y \sim$ Poisson, $g^*(y) = \sqrt{y}$.
(b) $Y \sim$ binomial(n, p), $g^*(y) = \sin^{-1}\big(\sqrt{y/n}\big)$.
(c) Y has variance $v(\theta) = K\theta^2$ for some constant K, $g^*(y) = \log(y)$.
(Conditions for the existence of variance-stabilizing transformations go back at least to Curtiss 1943, with refinements given by Bar-Lev and Enis 1988, 1990.)

Solution. The three approximate variances are $1/4$, $1/(4n)$, and K – all free of the parameter. Each g^* is a constant multiple of the integral in Exercise 11.1(b), which changes the approximate variance only by that constant squared, so Exercise 11.1(a) is all that is needed.
(a) $\theta = \mathrm{E}Y$ and $v(\theta) = \theta$. With $g^*(y) = \sqrt{y}$, $g^{*\prime}(\theta) = 1/(2\sqrt{\theta})$, so

$$\operatorname{Var}\big(\sqrt{Y}\big) \;\approx\; \frac{1}{4\theta} \cdot \theta \;=\; \frac{1}{4}.$$

(b) $\theta = \mathrm{E}Y = np$ and $v(\theta) = np(1-p) = \theta(1-\theta/n)$. Differentiating $g^*(y) = \sin^{-1}\sqrt{y/n}$,

$$g^{*\prime}(y) = \frac{1}{\sqrt{1-y/n}} \cdot \frac{1}{2\sqrt{y/n}} \cdot \frac{1}{n} = \frac{1}{2\sqrt{ny-y^2}},$$

so at $y = \theta = np$ we get $g^{*\prime}(np) = 1/\big(2n\sqrt{p(1-p)}\big)$ and

$$\operatorname{Var}\Big(\sin^{-1}\sqrt{Y/n}\Big) \;\approx\; \frac{np(1-p)}{4n^2p(1-p)} \;=\; \frac{1}{4n},$$

free of p.
(c) $g^{*\prime}(\theta) = 1/\theta$, hence

$$\operatorname{Var}\bigl(\log Y\bigr) \;\approx\; \frac{1}{\theta^2}\cdot K\theta^2 \;=\; K.$$

Problem 11.3. The Box-Cox family of power transformations (Box and Cox 1964) is defined by

$$g^*_\lambda(y) = \begin{cases} (y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\ \log y & \text{if } \lambda = 0, \end{cases}$$

where λ is a free parameter.
(a) Show that, for each y, $g^*_\lambda(y)$ is continuous in λ. In particular, show that

$$\lim_{\lambda\to 0} (y^\lambda - 1)/\lambda = \log y.$$

(b) Find the function $v(\theta)$, the approximate variance of Y, that $g^*_\lambda(y)$ stabilizes. (Note that $v(\theta)$ will most likely also depend on λ.)
Analysis of transformed data in general and the Box-Cox power transformation in particular has been the topic of some controversy in the statistical literature. See Bickel and Doksum (1981), Box and Cox (1982), and Hinkley and Runger (1984).

Solution. (a) Fix $y > 0$ and write $y^\lambda = e^{\lambda \log y}$. For $\lambda \neq 0$ the map $\lambda \mapsto (e^{\lambda \log y} - 1)/\lambda$ is a quotient of entire functions with nonvanishing denominator, hence continuous there; continuity at $\lambda = 0$ is exactly the stated limit. Expanding the exponential,

$$\frac{y^\lambda - 1}{\lambda} = \frac{1}{\lambda}\sum_{j\geq 1} \frac{(\lambda \log y)^j}{j!} = \log y + \frac{\lambda(\log y)^2}{2} + O(\lambda^2),$$

so the limit as $\lambda \to 0$ is $\log y = g^*_0(y)$. (Equivalently, l'Hopital's rule in λ: the derivative of the numerator is $y^\lambda \log y \to \log y$ and that of the denominator is 1.)
(b) $v(\theta) = K\theta^{2(1-\lambda)}$ for an arbitrary constant $K > 0$. Indeed $g^{*\prime}_\lambda(y) = y^{\lambda-1}$ for every λ, including $\lambda = 0$, so by Exercise 11.1(a)

$$\operatorname{Var}\bigl(g^*_\lambda(Y)\bigr) \;\approx\; \theta^{2(\lambda-1)} v(\theta),$$

and this is free of θ if and only if $v(\theta) = K\theta^{2(1-\lambda)}$, in which case the approximate variance equals K.

Problem 11.4. A most famous (and useful) variance-stabilizing transformation is Fisher's z-transformation, which we have already encountered in Exercise 10.17. Here we will look at a few more details. Suppose that (X, Y) are bivariate normal with correlation coefficient ϱ and sample correlation r.
(a) Starting from Exercise 10.17, part (d) – which states that for a sample of size n from a bivariate normal population, $\sqrt{n}\,(r - \varrho) \to \mathrm{n}\bigl(0, (1-\varrho^2)^2\bigr)$ in distribution – use the Delta Method to show that

$$\frac{1}{2}\left[\log\left(\frac{1+r}{1-r}\right) - \log\left(\frac{1+\varrho}{1-\varrho}\right)\right]$$

is approximately normal with mean 0 and variance $1/n$.
(b) Fisher actually used a somewhat more accurate expansion (Stuart and Ord 1987, Section 16.33) and established that the quantity in part (a) is approximately normal with

$$\text{mean} = \frac{\varrho}{2(n-1)} \qquad \text{and} \qquad \text{variance} = \frac{1}{n-1} + \frac{4-\varrho^2}{2(n-1)^2}.$$

Show that for small ϱ and moderate n, we can approximate this mean and variance by 0 and $1/(n-3)$, which is the most popular form of Fisher's z-transformation.

Solution. (a) The transformation is $g(t) = \frac{1}{2}\log\big((1+t)/(1-t)\big) = \tanh^{-1} t$, whose derivative on $(-1,1)$ is

$$g'(t) = \frac{1}{2}\left(\frac{1}{1+t} + \frac{1}{1-t}\right) = \frac{1}{1-t^2},$$

continuous and nonzero at $t = \varrho$ for $|\varrho| < 1$, which is the hypothesis needed in the Delta Method (Theorem 5.5.24). Applying it to the limit of Exercise 10.17(d),

$$\sqrt{n}\,\big(g(r) - g(\varrho)\big) \;\to\; \mathrm{n}\left(0,\; \frac{(1-\varrho^2)^2}{(1-\varrho^2)^2}\right) \;=\; \mathrm{n}(0,1)$$

in distribution, that is, $g(r) - g(\varrho)$ is approximately $\mathrm{n}(0, 1/n)$.
(b) The mean $\varrho/(2(n-1))$ is $O(\varrho/n)$, while the standard deviation is of order $n^{-1/2}$; their ratio is $O(\varrho n^{-1/2})$, negligible for small ϱ and n not tiny. So set the mean to 0.
For the variance, drop ϱ^2 against 4:

$$\frac{1}{n-1} + \frac{4-\varrho^2}{2(n-1)^2} \;\approx\; \frac{1}{n-1} + \frac{2}{(n-1)^2} \;=\; \frac{n+1}{(n-1)^2}.$$

Now compare with $1/(n-3)$ by forming the ratio, using $(n+1)(n-3) = (n-1)^2 - 4$:

$$\frac{(n+1)/(n-1)^2}{1/(n-3)} = \frac{(n+1)(n-3)}{(n-1)^2} = 1 - \frac{4}{(n-1)^2}.$$

So the two agree to relative error $4/(n-1)^2$: about 4.9% at $n = 10$, 1.1% at $n = 20$, 0.5% at $n = 30$. Hence for moderate n the variance may be taken to be $1/(n-3)$.

Problem 11.5. Suppose that random variables Y_{ij} are observed according to the overparameterized oneway ANOVA model in (11.2.2), namely

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \qquad i = 1, \ldots, k, \quad j = 1, \ldots, n_i,$$

with $\mathrm{E}\epsilon_{ij} = 0$. Show that, without some restriction on the parameters, this model is not identifiable by exhibiting two distinct collections of parameters that lead to exactly the same distribution of the Y_{ij}s.

Solution. Take $(\mu, \tau_1, \ldots, \tau_k)$ and $(\mu + c,\ \tau_1 - c,\ \ldots,\ \tau_k - c)$ for any $c \neq 0$. These are distinct points of the parameter space, yet

$$(\mu + c) + (\tau_i - c) = \mu + \tau_i \qquad \text{for every } i,$$

so the two parameter vectors give the same $\mathrm{E}Y_{ij}$ for all i, j. Since the distribution of (Y_{ij}) depends on $(\mu, \tau_1, \ldots, \tau_k)$ only through the k numbers $\mu + \tau_i$ (the Y_{ij} are $\mu + \tau_i$ plus errors whose distribution does not involve μ or the τ_i), the two collections yield identical joint pdfs. Distinct parameter values thus fail to give distinct pdfs, so by Definition 11.2.2 the model is not identifiable.

Problem 11.6. Under the oneway ANOVA assumptions – $Y_{ij} = \theta_i + \epsilon_{ij}$, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, with the ϵ_{ij} independent $\mathrm{n}(0, \sigma^2)$ – and with $N = \sum_i n_i$:
(a) Show that the set of statistics $(\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{k\cdot}, S_p^2)$ is sufficient for $(\theta_1, \theta_2, \ldots, \theta_k, \sigma^2)$.
(b) Show that $S_p^2 = \frac{1}{N-k}\sum_{i=1}^k (n_i - 1)S_i^2$ is independent of each $\bar{Y}_{i\cdot}$, $i = 1, \ldots, k$. (See Lemma 5.3.3.)
(c) If σ^2 is known, explain how the ANOVA data are equivalent to their canonical version in Miscellanea 11.5.6, where one observes independent $X_i \sim \mathrm{n}(\xi_i, 1)$, $i = 1, \ldots, p$.

Solution. (a) Factor the joint pdf. Within group i the usual decomposition $\sum_j (y_{ij} - \theta_i)^2 = \sum_j (y_{ij} -$

$\bar{y}_{i\cdot})^2 + n_i(\bar{y}_{i\cdot} - \theta_i)^2$ (the cross term vanishes) summed over i gives, by (11.2.5),

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \theta_i)^2 = (N-k)s_p^2 + \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \theta_i)^2.$$

Hence

$$f(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \times \exp\left\{-\frac{(N-k)s_p^2 + \sum_i n_i(\bar{y}_{i\cdot} - \theta_i)^2}{2\sigma^2}\right\},$$

which depends on $\mathbf{y}$ only through $(\bar{y}_{1\cdot}, \ldots, \bar{y}_{k\cdot}, s_p^2)$; take $h(\mathbf{y}) \equiv 1$ in the Factorization Theorem (Theorem 6.2.6).
(b) The whole data vector is jointly normal, so Lemma 5.3.3 applies: zero covariance gives independence. For the residuals in group i,

$$\operatorname{Cov}(\bar{Y}_{i\cdot},\, Y_{ij} - \bar{Y}_{i\cdot}) = \frac{\sigma^2}{n_i} - \frac{\sigma^2}{n_i} = 0,$$

and for $l \neq i$, $\operatorname{Cov}(\bar{Y}_{i\cdot},\, Y_{lj} - \bar{Y}_{l\cdot}) = 0$ because groups i and l are independent. Thus by Lemma 5.3.3(b) the single linear combination $\bar{Y}_{i\cdot}$ is independent of the entire residual vector $\{Y_{lj} - \bar{Y}_{l\cdot} : 1 \le l \le k,\ 1 \le j \le n_l\}$. Since

$$S_p^2 = \frac{1}{N-k}\sum_{l=1}^{k}\sum_{j=1}^{n_l}(Y_{lj} - \bar{Y}_{l\cdot})^2$$

is a function of that residual vector, S_p^2 is independent of $\bar{Y}_{i\cdot}$ for each i.
(c) With σ^2 known the sufficient statistic of part (a) reduces to the cell means alone: in the factorization of (a) the factor $(2\pi\sigma^2)^{-N/2}e^{-(N-k)s_p^2/(2\sigma^2)}$ no longer involves an unknown parameter, so it may be taken as $h(\mathbf{y})$ and Theorem 6.2.6 makes $(\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{k\cdot})$ sufficient for $\boldsymbol{\theta}$. These satisfy

$$\bar{Y}_{i\cdot} \sim \mathrm{n}\left(\theta_i, \frac{\sigma^2}{n_i}\right), \qquad i = 1, \ldots, k, \quad \text{independent.}$$

Now standardize: put

$$X_i = \frac{\sqrt{n_i}}{\sigma}\,\bar{Y}_{i\cdot}, \qquad \xi_i = \frac{\sqrt{n_i}}{\sigma}\,\theta_i.$$

Then $X_1, \ldots, X_k$ are independent with $X_i \sim \mathrm{n}(\xi_i, 1)$, which is exactly the canonical version of Miscellanea 11.5.6 with $p = k$. Since σ and the n_i are known, $\theta_i \mapsto \xi_i$ is a known one-to-one rescaling, so the ANOVA estimation problem and the canonical one are the same problem.

Problem 11.7. Complete the proof of Theorem 11.2.8 by showing that

$$\frac{1}{\sigma^2}\sum_{i=1}^{k} n_i\Big((\bar{Y}_{i\cdot} - \bar{\bar{Y}}) - (\theta_i - \bar{\theta})\Big)^2 \;\sim\; \chi^2_{k-1},$$

where $\bar{\bar{Y}} = \sum_i n_i\bar{Y}_{i\cdot}/\sum_i n_i$ and $\bar{\theta} = \sum_i n_i\theta_i/\sum_i n_i$.
(Hint: Define $\bar{U}_i = \bar{Y}_{i\cdot} - \theta_i$, $i = 1, \ldots, k$. Show that the $\bar{U}_i$ are independent $\mathrm{n}(0, \sigma^2/n_i)$. Then adapt the induction argument of Lemma 5.3.2 to show that $\sum n_i(\bar{U}_i - \bar{\bar{U}})^2/\sigma^2 \sim \chi^2_{k-1}$, where $\bar{\bar{U}} = \sum n_i\bar{U}_i/\sum n_i$.)

Solution. Put $\bar{U}_i = \bar{Y}_{i\cdot} - \theta_i$ and $N = \sum_i n_i$. Under the oneway ANOVA assumptions $\bar{Y}_{i\cdot} \sim \mathrm{n}(\theta_i, \sigma^2/n_i)$ independent, so the $\bar{U}_i$ are independent $\mathrm{n}(0, \sigma^2/n_i)$, and since $\bar{\bar{U}} = \sum_i n_i\bar{U}_i/N = \bar{\bar{Y}} - \bar{\theta}$,

$$(\bar{Y}_{i\cdot} - \bar{\bar{Y}}) - (\theta_i - \bar{\theta}) = \bar{U}_i - \bar{\bar{U}}.$$

So it suffices to show $Q := \sum_i n_i(\bar{U}_i - \bar{\bar{U}})^2/\sigma^2 \sim \chi^2_{k-1}$.
Standardize by $V_i = \sqrt{n_i}\,\bar{U}_i/\sigma$, so $V_1, \ldots, V_k$ are iid $\mathrm{n}(0,1)$, and set $a_i = \sqrt{n_i/N}$, so $\sum_i a_i^2 = 1$. Expanding the square,

$$\begin{aligned} Q &= \frac{1}{\sigma^2}\sum_{i=1}^{k} n_i\bar{U}_i^2 \;-\; \frac{1}{N\sigma^2}\Big(\sum_{i=1}^{k} n_i\bar{U}_i\Big)^2 \\ &= \sum_{i=1}^{k} V_i^2 \;-\; \frac{1}{N}\Big(\sum_{i=1}^{k} \sqrt{n_i}\,V_i\Big)^2 \\ &= \sum_{i=1}^{k} V_i^2 \;-\; \Big(\sum_{i=1}^{k} a_i V_i\Big)^2. \end{aligned}$$

Let A be a $k \times k$ orthogonal matrix whose first row is $\mathbf{a} = (a_1, \ldots, a_k)$ (Gram-Schmidt supplies the remaining rows, since $|\mathbf{a}| = 1$), and set $\mathbf{Z} = A\mathbf{V}$. Each $Z_i = \sum_j A_{ij}V_j$ is normal with mean 0 and variance $\sum_j A_{ij}^2 = 1$, and $\mathrm{Cov}(Z_i, Z_l) = \sum_j A_{ij}A_{lj} = 0$ for $i \neq l$; by Lemma 5.3.3 the Z_i are therefore mutually independent $\mathrm{n}(0,1)$. Orthogonality also gives $\sum_i Z_i^2 = |A\mathbf{V}|^2 = |\mathbf{V}|^2 = \sum_i V_i^2$, while $Z_1 = \sum_i a_i V_i$. Hence

$$Q = \sum_{i=1}^{k} Z_i^2 - Z_1^2 = \sum_{i=2}^{k} Z_i^2 \sim \chi^2_{k-1}$$

by Lemma 5.3.2, parts (a) and (b).
Method (2) (the induction the hint asks for). For $1 \le m \le k$ write $N_m = \sum_{i\le m} n_i$, $\bar{\bar{U}}_m = \sum_{i\le m} n_i\bar{U}_i/N_m$, and $T_m = \sum_{i\le m} n_i(\bar{U}_i - \bar{\bar{U}}_m)^2$. The analogue of (5.3.1) is

$$T_{m+1} = T_m + \frac{N_m n_{m+1}}{N_{m+1}}\big(\bar{U}_{m+1} - \bar{\bar{U}}_m\big)^2 \qquad \text{(Check!)}$$

Since $\mathrm{Cov}(\bar{\bar{U}}_m, \bar{U}_i) = \sigma^2/N_m = \mathrm{Var}(\bar{\bar{U}}_m)$ for $i \le m$, Lemma 5.3.3(b) makes $\bar{\bar{U}}_m$ independent of $(\bar{U}_1 - \bar{\bar{U}}_m, \ldots, \bar{U}_m - \bar{\bar{U}}_m)$ and hence of T_m; and $\bar{U}_{m+1}$ is independent of everything indexed by $i \le m$. So the added term is independent of T_m, and because

$$\bar{U}_{m+1} - \bar{\bar{U}}_m \sim \mathrm{n}\left(0,\ \sigma^2\frac{N_{m+1}}{n_{m+1}N_m}\right),$$

that term divided by σ^2 is χ^2_1 by Lemma 5.3.2(a). At $m = 1$ we have $T_1 = 0$, so the identity gives $T_2/\sigma^2 \sim \chi^2_1$, the base case; applying Lemma 5.3.2(b) at each subsequent step gives $T_m/\sigma^2 \sim \chi^2_{m-1}$, and at $m = k$ this is $Q \sim \chi^2_{k-1}$.

11.2 Exercises 11.8–11.14

Problem 11.8. Show that under the oneway ANOVA assumptions, for any set of constants $\mathbf{a} = (a_1, \ldots, a_k)$, the quantity $\sum a_i\bar{Y}_{i\cdot}$ is normally distributed with mean $\sum a_i\theta_i$ and variance $\sigma^2\sum a_i^2/n_i$. (See Corollary 4.6.10.)

Solution. Corollary 4.6.10 applied to the k independent normals $\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{k\cdot}$.
Under the assumptions the $Y_{ij} \sim \mathrm{n}(\theta_i, \sigma^2)$ are all independent, so Corollary 4.6.10 within cell i (constants $a_j = 1/n_i$, $b_j = 0$) gives

$$\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \sim \mathrm{n}\left(\theta_i,\ \frac{\sigma^2}{n_i}\right),$$

and $\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{k\cdot}$ are mutually independent, being functions of the disjoint independent blocks $\{Y_{ij} : j \le n_i\}$ (Theorem 4.6.12).

Now apply Corollary 4.6.10 a second time, to $X_i = \bar{Y}_{i\cdot}$ with the constants a_i and $b_i = 0$:

$$\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot} \sim \mathrm{n}\left(\sum_{i=1}^{k} a_i \theta_i,\ \sum_{i=1}^{k} a_i^2 \frac{\sigma^2}{n_i} \right) = \mathrm{n}\left(\sum_{i=1}^{k} a_i \theta_i,\ \sigma^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i} \right).$$

The corollary's hypotheses – mutual independence and marginal normality of the summands – are exactly what the preceding paragraph supplies.

Problem 11.9. Using an argument similar to that which led to the t test in (11.2.7), show how to construct a t test for
(a) $H_0 :\ \sum a_i \theta_i = \delta$ versus $H_1 :\ \sum a_i \theta_i \neq \delta$.
(b) $H_0 :\ \sum a_i \theta_i \leq \delta$ versus $H_1 :\ \sum a_i \theta_i > \delta$, where δ is a specified constant.

Solution. Recentre the pivot (11.2.6) at δ: with $\mathbf{a} \neq \mathbf{0}$ put

$$T \ = \ \frac{\sum_{i=1}^{k} a_i \bar{Y}_{i\cdot} - \delta}{\sqrt{S_p^2 \sum_{i=1}^{k} a_i^2 / n_i}},$$

and reject H_0 if (a) $|T| > t_{N-k,\alpha/2}$, (b) $T > t_{N-k,\alpha}$.
(a) By (11.2.6), $\big(\sum a_i \bar{Y}_{i\cdot} - \sum a_i \theta_i\big) / \sqrt{S_p^2 \sum a_i^2 / n_i} \sim t_{N-k}$ – its hypotheses hold here because $\sum a_i \bar{Y}_{i\cdot}$ is normal with the mean and variance of Exercise 11.8, $(N-k)S_p^2/\sigma^2 \sim \chi^2_{N-k}$, and S_p^2 is independent of every $\bar{Y}_{i\cdot}$ (Exercise 11.6(b)). Under H_0 the numerator recentring is exact, so $T \sim t_{N-k}$ and

$$P_{H_0}\big(|T| > t_{N-k,\alpha/2}\big) = \alpha$$

for every θ with $\sum a_i \theta_i = \delta$ and every σ^2: the test has size α. This is (11.2.7) with $\sum a_i \bar{Y}_{i\cdot}$ replaced by $\sum a_i \bar{Y}_{i\cdot} - \delta$.
(b) Write $\eta = \sum a_i \theta_i$ and $D = \sqrt{S_p^2 \sum a_i^2 / n_i} > 0$, so that

$$T \ = \ T_0 + \frac{\eta - \delta}{D}, \qquad T_0 = \frac{\sum a_i \bar{Y}_{i\cdot} - \eta}{D} \sim t_{N-k}.$$

Since $D > 0$ pointwise, $\eta \leq \delta$ gives $T \leq T_0$ pointwise, hence

$$\beta(\theta, \sigma^2) = P\big(T > t_{N-k,\alpha}\big) \ \leq \ P\big(T_0 > t_{N-k,\alpha}\big) = \alpha,$$

with equality at $\eta = \delta$. Thus $\sup_{H_0} \beta = \alpha$ and the one-sided test has size α.

Problem 11.10. Suppose we have a oneway ANOVA with five treatments. Denote the treatment means by $\theta_1, \ldots, \theta_5$, where θ_1 is a control and $\theta_2, \ldots, \theta_5$ are alternative new treatments, and assume that an equal number of observations per treatment is taken. Consider the four contrasts $\sum a_i \theta_i$ defined by

$$\begin{aligned}
\mathbf{a}_1 &= \big(1, -\tfrac{1}{4}, -\tfrac{1}{4}, -\tfrac{1}{4}, -\tfrac{1}{4}\big), \\
\mathbf{a}_2 &= \big(0, 1, -\tfrac{1}{3}, -\tfrac{1}{3}, -\tfrac{1}{3}\big), \\
\mathbf{a}_3 &= \big(0, 0, 1, -\tfrac{1}{2}, -\tfrac{1}{2}\big), \\
\mathbf{a}_4 &= (0, 0, 0, 1, -1).
\end{aligned}$$

(a) Argue that the results of the four t tests using these contrasts can lead to conclusions about the ordering of $\theta_1, \ldots, \theta_5$. What conclusions might be made?
(b) Show that any two contrasts $\sum a_i \bar{Y}_{i\cdot}$ formed from the four $\mathbf{a}_i$s in part (a) are uncorrelated. (Recall that these are called orthogonal contrasts.)

(c) For the fertilizer experiment of Example 11.2.3, the following contrasts were planned:

$$\begin{aligned}\mathbf{a}_1 &= (-1, 1, 0, 0, 0),\\ \mathbf{a}_2 &= \left(0, -1, \tfrac{1}{2}, \tfrac{1}{2}, 0\right),\\ \mathbf{a}_3 &= (0, 0, 1, -1, 0),\\ \mathbf{a}_4 &= (0, -1, 0, 0, 1).\end{aligned}$$

Show that these contrasts are not orthogonal. Interpret these contrasts in the context of the fertilizer experiment, and argue that they are a sensible set of contrasts.
(In Example 11.2.3 the response is the zinc content of spinach plants, and the five fertilizer treatments, in pounds per acre, are:)

Treatment	Magnesium	Potassium	Zinc
1	0	0	0
2	0	200	0
3	50	200	0
4	200	200	0
5	0	200	15

Solution. (a) These are the Helmert contrasts: $\mathbf{a}_j$ compares θ_j with the average of all treatments that follow it,

$$\sum_i a_{ji}\theta_i = \theta_j - \frac{1}{5-j}\sum_{i>j}\theta_i, \qquad j = 1, 2, 3, 4.$$

Run each as the two-sided t test (11.2.7), or as the one-sided test of Exercise 11.9(b) with $\delta = 0$. Suppose all four are rejected in the positive direction, i.e. each $\sum_i a_{ji}\bar{Y}_{i\cdot}$ is significantly > 0. Then we conclude

$$\begin{aligned}&\theta_4 > \theta_5, && \theta_3 > \tfrac{1}{2}(\theta_4 + \theta_5),\\ &\theta_2 > \tfrac{1}{3}(\theta_3 + \theta_4 + \theta_5), && \theta_1 > \tfrac{1}{4}(\theta_2 + \theta_3 + \theta_4 + \theta_5).\end{aligned}$$

These chain downward: $\theta_4 > \theta_5$ makes $\frac{1}{2}(\theta_4+\theta_5) > \theta_5$, so $\theta_3 > \theta_5$; then the average of $\theta_3, \theta_4, \theta_5$ exceeds θ_5, so $\theta_2 > \theta_5$; and likewise $\theta_1 > \theta_5$. The conclusion available is that θ_5 is the smallest of the five means, each treatment in turn beating the average of those below it.
No complete ordering follows, though. From $\theta_3 > \frac{1}{2}(\theta_4 + \theta_5)$ and $\theta_4 > \theta_5$ one cannot get $\theta_3 > \theta_4$: $(\theta_3, \theta_4, \theta_5) = (6, 10, 0)$ satisfies both with $\theta_4 > \theta_3$. This is the caution on page 530 – rejecting $\theta_1 = \frac{1}{2}(\theta_2 + \theta_3)$ and $\theta_2 = \theta_3$ orders θ_3 against θ_1 and θ_2 but leaves θ_1 versus θ_2 undecided. Weaker rejection patterns say less: rejecting only $\mathbf{a}_4$ says only $\theta_4 \neq \theta_5$, and rejecting only $\mathbf{a}_1$ says only that the control differs from the average of the four new treatments.
(b) With $n_i = n$ for all i, Exercise 11.11 gives

$$\operatorname{Cov}\left(\sum_i a_i\bar{Y}_{i\cdot},\ \sum_i b_i\bar{Y}_{i\cdot}\right) = \sigma^2\sum_i \frac{a_i b_i}{n} = \frac{\sigma^2}{n}\,\mathbf{a}\cdot\mathbf{b},$$

so uncorrelatedness is exactly $\mathbf{a}\cdot\mathbf{b} = 0$. For $j < j'$ the vector $\mathbf{a}_{j'}$ is supported on coordinates $> j$, where $\mathbf{a}_j$ is constant equal to $-1/(5-j)$; hence

$$\mathbf{a}_j \cdot \mathbf{a}_{j'} = -\frac{1}{5-j}\sum_{i>j} a_{j'i} = -\frac{1}{5-j}\cdot 0 = 0,$$

because each $\mathbf{a}_{j'}$ is a contrast whose coordinates sum to 0 and which vanishes on $i \le j$. (Check: $\mathbf{a}_1 \cdot \mathbf{a}_2 = -\frac{1}{4} + 3\cdot\frac{1}{12} = 0$.) All six pairs are orthogonal.
(c) Orthogonality fails on three of the six pairs:

$$\begin{aligned}\mathbf{a}_1 \cdot \mathbf{a}_2 &= (1)(-1) = -1,\\ \mathbf{a}_1 \cdot \mathbf{a}_4 &= (1)(-1) = -1,\\ \mathbf{a}_2 \cdot \mathbf{a}_4 &= (-1)(-1) = 1,\\ \mathbf{a}_1 \cdot \mathbf{a}_3 &= \mathbf{a}_2 \cdot \mathbf{a}_3 = \mathbf{a}_3 \cdot \mathbf{a}_4 = 0.\end{aligned}$$

Reading the contrasts against the treatment table:

$\mathbf{a}_1 : \theta_2 - \theta_1$	effect of adding 200 lb potassium to the control;
$\mathbf{a}_2 : \frac{1}{2}(\theta_3 + \theta_4) - \theta_2$	average effect of adding magnesium, given potassium;
$\mathbf{a}_3 : \theta_3 - \theta_4$	50 vs. 200 lb magnesium, given potassium;
$\mathbf{a}_4 : \theta_5 - \theta_2$	effect of adding 15 lb zinc, given potassium.

Each contrast varies exactly one fertilizer component against its proper baseline – treatment 1 for potassium, treatment 2 for magnesium and for zinc – so together they are the four questions the experiment was built to answer: does potassium help, does magnesium help on top of potassium, how much magnesium, does zinc help on top of potassium. Sharing treatment 2 as that baseline is what forces the nonzero inner products, and orthogonality is a convenience, not a requirement, for the t tests of (11.2.7).

Problem 11.11. For any sets of constants $\mathbf{a} = (a_1, \ldots, a_k)$ and $\mathbf{b} = (b_1, \ldots, b_k)$, show that under the oneway ANOVA assumptions,

$$\operatorname{Cov}\Big(\sum a_i \bar{Y}_{i\cdot},\ \sum b_i \bar{Y}_{i\cdot}\Big) = \sigma^2 \sum \frac{a_i b_i}{n_i}.$$

Hence, in the oneway ANOVA, contrasts are uncorrelated (orthogonal) if $\sum a_i b_i / n_i = 0$.

Solution. Bilinearity of covariance plus independence of the cell means.
The $\bar{Y}_{i\cdot}$ are functions of disjoint independent blocks, hence independent (Theorem 4.6.12), with $\operatorname{Var} \bar{Y}_{i\cdot} = \sigma^2/n_i$; so $\operatorname{Cov}(\bar{Y}_{i\cdot}, \bar{Y}_{j\cdot}) = 0$ for $i \neq j$ and

$$\begin{aligned}\operatorname{Cov}\left(\sum_{i=1}^k a_i \bar{Y}_{i\cdot},\ \sum_{j=1}^k b_j \bar{Y}_{j\cdot}\right) &= \sum_{i=1}^k \sum_{j=1}^k a_i b_j \operatorname{Cov}(\bar{Y}_{i\cdot}, \bar{Y}_{j\cdot}) \\ &= \sum_{i=1}^k a_i b_i \operatorname{Var} \bar{Y}_{i\cdot} = \sigma^2 \sum_{i=1}^k \frac{a_i b_i}{n_i}.\end{aligned}$$

Since $\sigma^2 > 0$, the covariance vanishes exactly when $\sum a_i b_i / n_i = 0$, which for two contrasts is the stated orthogonality condition; in the balanced case $n_i \equiv n$ it reduces to $\mathbf{a} \cdot \mathbf{b} = 0$.

Problem 11.12. Suppose that we have a oneway ANOVA with equal numbers of observations on each treatment, that is, $n_i = n$, $i = 1, \ldots, k$. In this case the F test can be considered an average t test.
(a) Show that a t test of $H_0 : \theta_i = \theta_{i'}$ versus $H_1 : \theta_i \neq \theta_{i'}$ can be based on the statistic

$$t_{ii'}^2 = \frac{(\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot})^2}{S_p^2(2/n)}.$$

(b) Show that

$$\frac{1}{k(k-1)} \sum_{i,i'} t_{ii'}^2 = F,$$

where F is the usual ANOVA F statistic. (Hint: See Exercise 5.8(a).) (Communicated by George McCabe, who learned it from John Tukey.)

Solution. (a) Take $\mathbf{a}$ to be the pairwise contrast with $a_i = 1$, $a_{i'} = -1$ and all other coordinates 0.

Then $\sum a_j\theta_j = \theta_i - \theta_{i'}$, and $\sum a_j^2/n_j = 2/n$, so the statistic of (11.2.6)/(11.2.7) is

$$t_{ii'} = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}}{\sqrt{S_p^2(2/n)}} \sim t_{N-k} \quad \text{under } H_0,$$

exactly the two-treatment comparison of Example 11.2.6 with $n_1 = n_2 = n$. Rejecting when $|t_{ii'}| > t_{N-k,\alpha/2}$, i.e. when $t_{ii'}^2 > t_{N-k,\alpha/2}^2$, is a size-$\alpha$ test of $H_0 : \theta_i = \theta_{i'}$.
(b) The sum runs over all $k(k-1)$ ordered pairs $i \neq i'$ (the diagonal terms vanish), and balance makes $\bar{\bar{Y}} = \frac{1}{k}\sum_i \bar{Y}_{i\cdot}$ the grand mean. Exercise 5.8(a), applied to the k numbers $\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{k\cdot}$, gives

$$\sum_{i,i'} (\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot})^2 = 2k \sum_{i=1}^{k} \big(\bar{Y}_{i\cdot} - \bar{\bar{Y}}\big)^2.$$

Therefore

$$\begin{aligned}
\frac{1}{k(k-1)} \sum_{i,i'} t_{ii'}^2 &= \frac{1}{k(k-1)} \cdot \frac{n}{2S_p^2} \sum_{i,i'} (\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot})^2 \\
&= \frac{1}{k(k-1)} \cdot \frac{n}{2S_p^2} \cdot 2k \sum_{i=1}^{k} (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 \\
&= \frac{n \sum_{i=1}^{k} (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2}{(k-1)S_p^2} = \frac{\sum_{i=1}^{k} n_i (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2}{(k-1)S_p^2} = F,
\end{aligned}$$

the last expression being the left side of (11.2.14) divided by $k-1$, that is, the ANOVA F statistic referred to $F_{k-1,N-k}$.

Problem 11.13. Under the oneway ANOVA assumptions, show that the likelihood ratio test of $H_0 :\ \theta_1 = \theta_2 = \cdots = \theta_k$ is given by the F test of (11.2.14).

Solution. The LRT statistic is a strictly decreasing function of F, so the rejection region $\lambda(\mathbf{y}) \le c$ is a rejection region $F \ge c'$.
The likelihood for $(\theta_1, \ldots, \theta_k, \sigma^2)$ is

$$L = \big(2\pi\sigma^2\big)^{-N/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \theta_i)^2 \right\}, \qquad N = \sum_i n_i.$$

Unrestricted maximization: for fixed σ^2 each inner sum of squares is minimized at $\hat{\theta}_i = \bar{y}_{i\cdot}$, and then the usual normal calculation gives

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i,j} (y_{ij} - \bar{y}_{i\cdot})^2 = \frac{N-k}{N}\, s_p^2, \qquad \sup_{\Theta} L = \big(2\pi\hat{\sigma}^2\big)^{-N/2} e^{-N/2}.$$

Under H_0 all θ_i equal a common θ, and $\sum_{i,j}(y_{ij}-\theta)^2$ is minimized at the grand mean $\hat{\hat{\theta}} = \bar{\bar{y}} = \frac{1}{N}\sum_{i,j} y_{ij}$, whence

$$\hat{\hat{\sigma}}^2 = \frac{1}{N} \sum_{i,j} (y_{ij} - \bar{\bar{y}})^2, \qquad \sup_{\Theta_0} L = \big(2\pi\hat{\hat{\sigma}}^2\big)^{-N/2} e^{-N/2}.$$

Hence

$$\lambda(\mathbf{y}) = \frac{\sup_{\Theta_0} L}{\sup_{\Theta} L} = \left(\frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2} \right)^{N/2} = \left(\frac{\sum_{i,j} (y_{ij} - \bar{y}_{i\cdot})^2}{\sum_{i,j} (y_{ij} - \bar{\bar{y}})^2} \right)^{N/2}.$$

By the sum-of-squares partition of Theorem 11.2.11, equation (11.2.15),

$$\sum_{i,j} (y_{ij} - \bar{\bar{y}})^2 = \sum_{i,j} (y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{\bar{y}})^2,$$

so, dividing through by $\sum_{i,j}(y_{ij}-\bar y_{i\cdot})^2=(N-k)s_p^2$,

$$\lambda(\mathbf{y})^{-2/N}=1+\frac{\sum_i n_i(\bar y_{i\cdot}-\bar{\bar y})^2}{(N-k)s_p^2}=1+\frac{k-1}{N-k}\,F,\qquad F=\frac{\sum_i n_i(\bar y_{i\cdot}-\bar{\bar y})^2}{(k-1)s_p^2}.$$

The map $F\mapsto\left(1+\frac{k-1}{N-k}F\right)^{-N/2}$ is strictly decreasing on $F\ge 0$, so

$$\lambda(\mathbf{y})\le c\iff F\ge c'$$

for the corresponding constant c'. The LRT rejection region is therefore $\{F\ge c'\}$, and taking $c'=F_{k-1,N-k,\alpha}$ makes the size exactly α, since under H_0 the statistic F has the $F_{k-1,N-k}$ distribution of (11.2.13). Written out, reject H_0 if

$$\frac{\sum_{i=1}^k n_i(\bar Y_{i\cdot}-\bar{\bar Y})^2}{S_p^2}>(k-1)F_{k-1,N-k,\alpha},$$

which is exactly (11.2.14).

Problem 11.14. The Scheffe simultaneous interval procedure actually works for all linear combinations, not just contrasts. Show that under the oneway ANOVA assumptions, if $\mathbf{M}=\sqrt{kF_{k,N-k,\alpha}}$ (note the change in the numerator degrees of freedom), then the probability is $1-\alpha$ that

$$\begin{aligned}\sum_{i=1}^k a_i\bar Y_{i\cdot}-\mathbf{M}\sqrt{S_p^2\sum_{i=1}^k\frac{a_i^2}{n_i}}\;&\le\;\sum_{i=1}^k a_i\theta_i\\&\le\;\sum_{i=1}^k a_i\bar Y_{i\cdot}+\mathbf{M}\sqrt{S_p^2\sum_{i=1}^k\frac{a_i^2}{n_i}}\end{aligned}$$

simultaneously for all $\mathbf{a}=(a_1,\dots,a_k)$. It is probably easiest to proceed by first establishing, in the spirit of Lemma 11.2.7, that if $v_1,\dots,v_k$ are constants and $c_1,\dots,c_k$ are positive constants, then

$$\max_{\mathbf{a}}\left\{\frac{\left(\sum_{i=1}^k a_iv_i\right)^2}{\sum_{i=1}^k a_i^2/c_i}\right\}=\sum_{i=1}^k c_iv_i^2.$$

The proof of Theorem 11.2.10 can then be adapted to establish the result.

Solution. Dropping the constraint $\sum a_i=0$ turns Lemma 11.2.7 into plain Cauchy–Schwarz, and the χ^2_{k-1} of (11.2.13) into a χ^2_k.

The maximization. For $\mathbf{a}\neq\mathbf{0}$ write $a_iv_i=(a_i/\sqrt{c_i})(\sqrt{c_i}\,v_i)$; the Cauchy–Schwarz Inequality (Theorem 4.7.3, in the summation form used to prove Lemma 11.2.7) gives

$$\left(\sum_{i=1}^k a_iv_i\right)^2\le\left(\sum_{i=1}^k\frac{a_i^2}{c_i}\right)\left(\sum_{i=1}^k c_iv_i^2\right),$$

that is, the ratio is $\le\sum c_iv_i^2$. Equality holds iff $(a_i/\sqrt{c_i})$ and $(\sqrt{c_i}\,v_i)$ are proportional, i.e. $a_i=Kc_iv_i$, and such an $\mathbf{a}$ is admissible here because there is no constraint $\sum a_i=0$ to satisfy; so the maximum is attained and equals $\sum_i c_iv_i^2$. That constraint is the whole difference from Lemma 11.2.7, which replaces v_i by $v_i-\bar v_c$ and lowers the maximum to $\sum c_i(v_i-\bar v_c)^2$ – and costs one degree of freedom below.

The probability statement. Put $U_i=\bar Y_{i\cdot}-\theta_i$ and, as in (11.2.9),

$$T_{\mathbf{a}}^2=\frac{\left(\sum_i a_i\bar Y_{i\cdot}-\sum_i a_i\theta_i\right)^2}{S_p^2\sum_i a_i^2/n_i}=\frac{\left(\sum_i a_iU_i\right)^2}{S_p^2\sum_i a_i^2/n_i}.$$

Since S_p^2 does not enter the maximization, applying the display above with $v_i = U_i$ and $c_i = n_i$ gives

$$\sup_{\mathbf{a}} T_{\mathbf{a}}^2 = \frac{\sum_{i=1}^k n_i U_i^2}{S_p^2} = \frac{\sum_{i=1}^k n_i (\bar{Y}_{i\cdot} - \theta_i)^2}{S_p^2}.$$

Now $\bar{Y}_{i\cdot} \sim \mathrm{n}(\theta_i, \sigma^2/n_i)$ (Exercise 11.8 with $\mathbf{a} = \mathbf{e}_i$), independent across i by Theorem 4.6.12, so the $\sqrt{n_i}\, U_i/\sigma$ are iid $\mathrm{n}(0,1)$ and

$$\frac{1}{\sigma^2} \sum_{i=1}^k n_i U_i^2 \sim \chi_k^2$$

by Lemma 5.3.2. Also $(N-k)S_p^2/\sigma^2 \sim \chi_{N-k}^2$, independent of all the $\bar{Y}_{i\cdot}$ (Exercise 11.6(b)), so numerator and denominator are independent and

$$\frac{\sup_{\mathbf{a}} T_{\mathbf{a}}^2}{k} = \frac{\left(\sum_i n_i U_i^2/\sigma^2\right)/k}{\left(S_p^2/\sigma^2\right)} \sim F_{k,N-k}.$$

Finally the simultaneous statement is one event about that supremum:

$$\begin{aligned} \left| \sum_i a_i \bar{Y}_{i\cdot} - \sum_i a_i \theta_i \right| \le \mathbf{M} \sqrt{S_p^2 \sum_i \frac{a_i^2}{n_i}} \quad &\text{for all } \mathbf{a} \\ \iff \quad T_{\mathbf{a}}^2 \le \mathbf{M}^2 \quad &\text{for all } \mathbf{a} \neq \mathbf{0} \\ \iff \quad \sup_{\mathbf{a}} T_{\mathbf{a}}^2 \le \mathbf{M}^2 & \end{aligned}$$

($\mathbf{a} = \mathbf{0}$ gives $0 \le 0$, always true). Hence

$$P\left(\sup_{\mathbf{a}} T_{\mathbf{a}}^2 \le \mathbf{M}^2\right) = P\left(F_{k,N-k} \le \frac{\mathbf{M}^2}{k}\right) = 1 - \alpha$$

precisely when $\mathbf{M}^2/k = F_{k,N-k,\alpha}$, i.e. $\mathbf{M} = \sqrt{kF_{k,N-k,\alpha}}$, which is the assertion.

11.3 Exercises 11.15–11.21

Problem 11.15. (a) Show that for the t and F distributions, for any ν, α, and k,

$$t_{\nu,\alpha/2} \le \sqrt{(k-1)F_{k-1,\nu,\alpha}}.$$

(Recall the relationship between the t and the F. This inequality is a consequence of the fact that the distributions $kF_{k,\nu}$ are stochastically increasing in k for fixed ν but is actually a weaker statement. See Exercise 5.19.)
(b) Explain how the above inequality shows that the simultaneous Scheffe intervals are always wider than the single-contrast intervals.
(c) Show that it also follows from the above inequality that Scheffe tests are less powerful than t tests.

Solution. (a) Squared, the two sides are the upper-α cutoffs of $1 \cdot F_{1,\nu}$ and of $(k-1)F_{k-1,\nu}$, which Exercise 5.19(b) orders stochastically.
By Theorem 5.3.8(b), $T_\nu^2 \sim F_{1,\nu}$, and the two-sided α tail of T_ν is the upper α tail of T_ν^2, so

$$t_{\nu,\alpha/2}^2 \;=\; F_{1,\nu,\alpha}.$$

Exercise 5.19(b) (whose hypothesis is only that ν be held fixed) says $mF_{m,\nu}$ is stochastically increasing in m; taking $m = 1$ and $m = k-1 \ge 1$,

$$P\big((k-1)F_{k-1,\nu} > c\big) \;\ge\; P\big(F_{1,\nu} > c\big) \quad \text{for all } c.$$

Evaluating at $c = F_{1,\nu,\alpha}$ gives $P((k-1)F_{k-1,\nu} > F_{1,\nu,\alpha}) \geq \alpha$, so the upper-$\alpha$ cutoff of $(k-1)F_{k-1,\nu}$ is at least $F_{1,\nu,\alpha}$:

$$(k-1)F_{k-1,\nu,\alpha} \;\geq\; F_{1,\nu,\alpha} \;=\; t^2_{\nu,\alpha/2}.$$

Take square roots. (Iterating Exercise 5.19(c) down from $k-1$ to 1 gives the same chain.)

(b) Both intervals for $\sum_i a_i\theta_i$ have the same standard-error factor $\sqrt{S_p^2 \sum_i a_i^2/n_i}$ and differ only in the multiplier: the single-contrast interval (11.2.8) uses $t_{N-k,\alpha/2}$, while Theorem 11.2.10 uses $\mathbf{M} = \sqrt{(k-1)F_{k-1,N-k,\alpha}}$. By part (a) with $\nu = N-k$,

$$\frac{\text{Scheffe half-length}}{t \text{ half-length}} = \frac{\sqrt{(k-1)F_{k-1,N-k,\alpha}}}{t_{N-k,\alpha/2}} \;\geq\; 1,$$

for every $\mathbf{a}$ and every realization of S_p^2.

(c) The Scheffe test of $H_{0\mathbf{a}}\colon \sum_i a_i\theta_i = 0$ rejects when $T_\mathbf{a} > \mathbf{M}$, the t test (11.2.7) when $T_\mathbf{a} > t_{N-k,\alpha/2}$. Since $t_{N-k,\alpha/2} \leq \mathbf{M}$, the Scheffe rejection region is contained in the t rejection region, so for every θ

$$P_\theta(T_\mathbf{a} > \mathbf{M}) \;\leq\; P_\theta(T_\mathbf{a} > t_{N-k,\alpha/2}).$$

Problem 11.16. In Theorem 11.2.5 we saw that the ANOVA null is equivalent to all contrasts being 0. We can also write the ANOVA null as the intersection over another set of hypotheses.

(a) Show that the hypotheses

$$H_0\colon\ \theta_1 = \theta_2 = \cdots = \theta_k \qquad \text{versus} \qquad H_1\colon\ \theta_i \neq \theta_j \text{ for some } i,j$$

and the hypotheses

$$H_0\colon\ \theta_i - \theta_j = 0 \text{ for all } i,j \qquad \text{versus} \qquad H_1\colon\ \theta_i - \theta_j \neq 0 \text{ for some } i,j$$

are equivalent.

(b) Express H_0 and H_1 of the ANOVA test as unions and intersections of the sets

$$\Theta_{ij} = \{\theta = (\theta_1, \ldots, \theta_k)\colon \theta_i - \theta_j = 0\}.$$

Describe how these expressions can be used to construct another (different) union-intersection test of the ANOVA null hypothesis. (See Miscellanea 11.5.2.)

Solution. (a) The two nulls are the same subset of $\mathbb{R}^k$, hence so are the two alternatives.

If $\theta_1 = \cdots = \theta_k$ then $\theta_i - \theta_j = 0$ for every pair; conversely if $\theta_i - \theta_j = 0$ for all i,j, take $j = 1$ to get $\theta_i = \theta_1$ for every i. So

$$\{\theta\colon \theta_1 = \cdots = \theta_k\} = \{\theta\colon \theta_i - \theta_j = 0 \text{ for all } i,j\},$$

and each alternative is the complement of this common set in $\mathbb{R}^k$.

(b) In terms of the Θ_{ij},

$$H_0\colon\ \theta \in \bigcap_{1\leq i<j\leq k} \Theta_{ij}, \qquad H_1\colon\ \theta \in \bigcup_{1\leq i<j\leq k} \Theta_{ij}^c.$$

(Pairs with $i = j$ contribute $\Theta_{ii} = \mathbb{R}^k$ and may be dropped; $\Theta_{ij} = \Theta_{ji}$, so $i < j$ suffices.)

This is a union-intersection decomposition in the sense of Section 8.2.3, with the family indexed by pairs rather than by all $\mathbf{a} \in \mathcal{A}$. Each component null $H_{0ij}\colon \theta \in \Theta_{ij}$ is a single-contrast hypothesis, tested by the t statistic of (11.2.6) with $\mathbf{a}$ the (i,j) difference vector,

$$T_{ij} \;=\; \frac{|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}|}{\sqrt{S_p^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}},$$

so the union-intersection test rejects the ANOVA null when

$$\max_{1\le i<j\le k} T_{ij} \;>\; c,$$

with c chosen so that $P_{H_0}(\max_{i<j} T_{ij} > c) = \alpha$. When $n_i \equiv n$ this maximum is the studentized range $\max_{i,j} |\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}|/\sqrt{S_p^2(2/n)}$ and c is the upper-α point of the Q distribution: this is Tukey's Q method of Miscellanea 11.5.2.
It differs from the ANOVA F test, which is the union-intersection test over the larger family $\{\Theta_{\mathbf{a}} : \mathbf{a} \in \mathcal{A}\}$ of all contrasts; since $\{\Theta_{ij}\} \subset \{\Theta_{\mathbf{a}}\}$, the maximum is over fewer statistics and $c < \mathbf{M}$.

Problem 11.17. A multiple comparison procedure called the *Protected LSD* (Protected Least Significant Difference) is performed as follows. If the ANOVA F test rejects H_0 at level α, then for each pair of means θ_i and $\theta_{i'}$, declare the means different if

$$\frac{|\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}|}{\sqrt{S_p^2\big(\frac{1}{n_i} + \frac{1}{n_{i'}}\big)}} \;>\; t_{\alpha/2,\,N-k}.$$

Note that each t test is done at the same α level as the ANOVA F test. Here we are using an *experimentwise* α level, where

$$\text{experimentwise } \alpha = P\big(\text{at least one false assertion of difference} \mid \text{all the means are equal}\big).$$

(a) Prove that no matter how many means are in the experiment, simultaneous inference from the Protected LSD is made at level α.
(b) The ordinary (or *unprotected*) LSD simply does the individual t tests, at level α, no matter what the outcome of the ANOVA F test. Show that the ordinary LSD can have an experimentwise error rate greater than α. (The unprotected LSD does maintain a *comparisonwise* error rate of α.)
(c) Perform the LSD procedure on the fish toxin data of Example 11.2.1. What are the conclusions? The data are the amounts of deterioration (in standard units) of the liver of each sacrificed fish:

Toxin 1	Toxin 2	Toxin 3	Control
28	33	18	11
23	36	21	14
14	34	20	11
27	29	22	16
	31	24	
	34		

Solution. (a) The Protected LSD asserts nothing at all unless the F test rejects, so its error event is contained in the F test's rejection event.
Let $R = \{\text{ANOVA } F > F_{k-1,N-k,\alpha}\}$ and let E be the event that at least one pair is declared different. By construction $E \subset R$, and under the complete null $\theta_1 = \cdots = \theta_k$ every assertion of difference is false, so

$$\text{experimentwise } \alpha = P_{H_0}(E) \;\le\; P_{H_0}(R) \;=\; \alpha,$$

the last equality because (11.2.14) is an exact α level test. The bound involves no union over pairs, so it does not degrade as k grows.
(b) With $A_{ii'} = \{T_{ii'} > t_{\alpha/2,N-k}\}$ the unprotected LSD commits an experimentwise error on $\bigcup_{i<i'} A_{ii'}$, and for $k \ge 3$

$$\begin{aligned} P_{H_0}\Big(\bigcup_{i<i'} A_{ii'}\Big) &\ge P_{H_0}(A_{12} \cup A_{13}) \\ &= P_{H_0}(A_{12}) + P_{H_0}(A_{13} \cap A_{12}^c) \\ &= \alpha + P_{H_0}(A_{13} \cap A_{12}^c). \end{aligned}$$

The last probability is strictly positive: $(\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{k\cdot}, S_p^2)$ has a joint density that is strictly positive on $\mathbb{R}^k \times (0,\infty)$, and $A_{13} \cap A_{12}^c$ contains a nonempty open set there (take $\bar{y}_{1\cdot} = \bar{y}_{2\cdot}$ and $\bar{y}_{3\cdot}$ far from both, with s_p^2 moderate). Hence the experimentwise rate exceeds α strictly; at $\alpha = 0.05$ with $n_i \equiv 5$ it is 0.115 for $k = 3$ and 0.337 for $k = 6$. Each single test is nevertheless an exact level α test of its own $H_{0ii'}$, which is the comparisonwise statement.
(c) Every pair differs except Toxin 1 versus Toxin 3.
From Example 11.2.12, $F = 26.09 > F_{3,15,0.05} = 3.29$, so the protection clears and we proceed. Here $k = 4$, $N = 19$, $S_p^2 = 190.83/15 = 12.72$, $t_{0.025,15} = 2.131$, and

$$\bar{y}_{1\cdot} = 23.00, \quad \bar{y}_{2\cdot} = 32.83, \quad \bar{y}_{3\cdot} = 21.00, \quad \bar{y}_{4\cdot} = 13.00.$$

With difference $= |\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}|$ and $\mathrm{LSD}_{ii'} = t_{0.025,15}\sqrt{S_p^2(1/n_i + 1/n_{i'})}$:

pair	difference	$\mathrm{LSD}_{ii'}$	$T_{ii'}$	decision
1 vs 2	9.83	4.91	4.27	different
1 vs 3	2.00	5.10	0.84	–
1 vs control	10.00	5.38	3.96	different
2 vs 3	11.83	4.60	5.48	different
2 vs control	19.83	4.91	8.61	different
3 vs control	8.00	5.10	3.34	different

The ordering is control < (toxin 3, toxin 1) < toxin 2: all three toxins deteriorate the liver more than the control, toxin 2 more than either of the other two, and toxins 1 and 3 cannot be separated.

Problem 11.18. Demonstrate that "data snooping," that is, testing hypotheses that are suggested by the data, is generally not a good practice.
(a) Show that, for any random variable Y and constants a and b with $a > b$ and $P(Y > b) < 1$,

$$P(Y > a \mid Y > b) \;>\; P(Y > a).$$

(b) Apply the inequality in part (a) to the size of a data-suggested hypothesis test by letting Y be a test statistic and a be a cutoff point.

Solution. (a) Since $a > b$, the event $\{Y > a\}$ is contained in $\{Y > b\}$, so the conditioning does nothing to the numerator and only shrinks the denominator:

$$P(Y > a \mid Y > b) = \frac{P(Y > a,\, Y > b)}{P(Y > b)} = \frac{P(Y > a)}{P(Y > b)} \;\geq\; P(Y > a),$$

with strict inequality because $P(Y > b) < 1$. (As printed the inequality also needs $P(Y > a) > 0$, and $P(Y > b) > 0$ for the conditional probability to be defined; if $P(Y > a) = 0$ both sides are 0.)
(b) Let Y be the test statistic and let a be the level-α cutoff, so that the nominal size is

$$P_{H_0}(Y > a) = \alpha.$$

Data snooping means the hypothesis is formulated only after the data have been inspected, and only when the statistic already looks extreme – that is, the test is carried out only on the event $\{Y > b\}$ for some $b < a$. The relevant Type I error rate is then the conditional one, and by part (a)

$$P_{H_0}(Y > a \mid Y > b) = \frac{\alpha}{P_{H_0}(Y > b)} \;>\; \alpha.$$

The more selective the snooping (the larger b, hence the smaller $P_{H_0}(Y > b)$), the larger the true size. Concretely, in a oneway ANOVA one inspects the cell means, notices that $\bar{Y}_{1\cdot}$ and $\bar{Y}_{2\cdot}$ are far apart – that is, that $Y = T_{12}$ already exceeds the informal threshold b at which a discrepancy catches the eye – and only then tests $H_0\colon \theta_1 = \theta_2$ at nominal level α. Its true size is $\alpha / P_{H_0}(T_{12} > b) > \alpha$.

Problem 11.19. Let $X_i \sim \text{gamma}(\lambda_i, 1)$ independently for $i = 1, \ldots, n$. Define

$$Y_i = \frac{X_{i+1}}{\sum_{j=1}^{i} X_j}, \quad i = 1, \ldots, n-1, \qquad \text{and} \qquad Y_n = \sum_{j=1}^{n} X_j.$$

(a) Find the joint and marginal distributions of Y_i, $i = 1, \ldots, n$.
(b) Connect your results to any distributions that are commonly employed in the ANOVA.

Solution. (a) Write $S_i = \sum_{j \le i} X_j$ and $\alpha_i = \sum_{j \le i} \lambda_j$. Then $Y_1, \ldots, Y_n$ are mutually independent, $Y_n \sim \text{gamma}(\alpha_n, 1)$, and for $i < n$

$$f_{Y_i}(y) = \frac{\Gamma(\alpha_{i+1})}{\Gamma(\alpha_i)\Gamma(\lambda_{i+1})} \frac{y^{\lambda_{i+1}-1}}{(1+y)^{\alpha_{i+1}}}, \qquad y > 0,$$

so the joint density is the product of these n factors.
Marginals: $S_i \sim \text{gamma}(\alpha_i, 1)$ (sums of independent gammas with common scale) and S_i is independent of $X_{i+1} \sim \text{gamma}(\lambda_{i+1}, 1)$. By Exercise 4.24 applied to S_i and X_{i+1},

$$B_i \;:=\; \frac{S_i}{S_{i+1}} \sim \text{beta}(\alpha_i, \lambda_{i+1}) \quad \text{independent of} \quad S_{i+1} \sim \text{gamma}(\alpha_{i+1}, 1).$$

Since $Y_i = X_{i+1}/S_i = (1-B_i)/B_i$, substituting $b = 1/(1+y)$, $|db/dy| = (1+y)^{-2}$, into the beta density gives

$$f_{Y_i}(y) = \frac{\Gamma(\alpha_{i+1})}{\Gamma(\alpha_i)\Gamma(\lambda_{i+1})} (1+y)^{-(\alpha_i - 1)} \Big(\tfrac{y}{1+y}\Big)^{\lambda_{i+1}-1} (1+y)^{-2},$$

which is the displayed density. And $Y_n = S_n \sim \text{gamma}(\alpha_n, 1)$.
Independence, by induction on n: assume $B_1, \ldots, B_{n-2}, S_{n-1}$ are mutually independent. They are functions of $(X_1, \ldots, X_{n-1})$, which is independent of X_n, so $(B_1, \ldots, B_{n-2})$, S_{n-1}, X_n are mutually independent. Exercise 4.24 on the independent pair (S_{n-1}, X_n) makes B_{n-1} and S_n independent of each other, and both are functions of (S_{n-1}, X_n) alone, hence independent of $(B_1, \ldots, B_{n-2})$. So $B_1, \ldots, B_{n-1}, S_n$ are mutually independent, and since $Y_i = (1 - B_i)/B_i$ depends only on B_i and $Y_n = S_n$, Theorem 4.6.12 gives the claim.
(b) Writing $\lambda_i = \nu_i/2$ makes $2X_i \sim \chi^2_{\nu_i}$, which is exactly the scaled sums of squares in an ANOVA table: independent chi squareds whose degrees of freedom add. Then $2S_i \sim \chi^2_{\nu_1 + \cdots + \nu_i}$, and

$$\frac{\alpha_i}{\lambda_{i+1}} Y_i = \frac{2X_{i+1}/(2\lambda_{i+1})}{2S_i/(2\alpha_i)} \sim F_{2\lambda_{i+1}, 2\alpha_i} = F_{\nu_{i+1}, \nu_1 + \cdots + \nu_i}$$

by Definition 5.3.6 (the numerator and denominator are independent, as noted above). Thus Y_i is a multiple of the F statistic testing the $(i+1)$st source of variation against the pooled earlier sources, Y_n is half the total sum of squares, and part (a) says these successive F statistics and the total are mutually independent.

Problem 11.20. Assume the oneway ANOVA null hypothesis is true.
(a) Show that $\sum_i n_i (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2/(k-1)$ gives an unbiased estimate of σ^2.
(b) Show how to use the method of Example 5.3.5 to derive the ANOVA F test.

Solution. (a) $\text{E} \sum_i n_i (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 = (k-1)\sigma^2$, by the shortcut expansion.
Under the ANOVA assumptions $\bar{Y}_{i\cdot} \sim \text{n}(\theta_i, \sigma^2/n_i)$ independently, and $\bar{\bar{Y}} = \sum_i n_i \bar{Y}_{i\cdot}/N$ has $\text{Var}\, \bar{\bar{Y}} =$

$\sum_i n_i^2(\sigma^2/n_i)/N^2 = \sigma^2/N$. Under H_0 all $\theta_i = \theta$, so $\mathrm{E}\bar{Y}_{i\cdot} = \mathrm{E}\bar{\bar{Y}} = \theta$ and

$$\begin{aligned}\mathrm{E}\sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 &= \mathrm{E}\Big[\sum_i n_i \bar{Y}_{i\cdot}^2 - N\bar{\bar{Y}}^2\Big] \\ &= \sum_i n_i\Big(\frac{\sigma^2}{n_i} + \theta^2\Big) - N\Big(\frac{\sigma^2}{N} + \theta^2\Big) \\ &= k\sigma^2 - \sigma^2 = (k-1)\sigma^2.\end{aligned}$$

Dividing by $k-1$ gives an unbiased estimator, $\mathrm{E}\,\mathrm{MSB} = \sigma^2$. (Without H_0 the same computation gives $\mathrm{E}\,\mathrm{MSB} = \sigma^2 + \sum_i n_i(\theta_i - \bar{\theta})^2/(k-1)$, $\bar{\theta} = \sum n_i\theta_i/N$.)
(b) Example 5.3.5 forms the ratio of two independent estimators of the same variance, each a scaled chi squared, and Definition 5.3.6 names the result F. Here the two estimators are MSB (unbiased for σ^2 by part (a)) and $\mathrm{MSW} = S_p^2$ of (11.2.5), also unbiased for σ^2.
The two required hypotheses hold. Chi squaredness: by (11.2.16) $(N-k)S_p^2/\sigma^2 \sim \chi^2_{N-k}$ always, and under H_0 by (11.2.17) $(k-1)\mathrm{MSB}/\sigma^2 \sim \chi^2_{k-1}$. Independence: within treatment i, $\bar{Y}_{i\cdot}$ and S_i^2 are independent by Theorem 5.3.1(a), and the k treatment samples are independent, so $(\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{k\cdot})$ is independent of $(S_1^2, \ldots, S_k^2)$, hence MSB is independent of MSW. Therefore, exactly as in (5.3.8),

$$\begin{aligned}F \;=\; \frac{\mathrm{MSB}}{\mathrm{MSW}} &= \frac{\big[(k-1)\mathrm{MSB}/\sigma^2\big]/(k-1)}{\big[(N-k)\mathrm{MSW}/\sigma^2\big]/(N-k)} \\ &\sim \frac{\chi^2_{k-1}/(k-1)}{\chi^2_{N-k}/(N-k)} \;\sim\; F_{k-1,\,N-k}\end{aligned}$$

under H_0, the unknown σ^2 cancelling. When H_0 fails, the parenthetical in (a) gives $\mathrm{E}\,\mathrm{MSB} > \sigma^2$ while $\mathrm{E}\,\mathrm{MSW} = \sigma^2$ still, so the ratio is inflated and we reject for large values:

$$\frac{\sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2/(k-1)}{S_p^2} \;>\; F_{k-1,N-k,\alpha},$$

which is the test (11.2.14).

Problem 11.21. (a) Illustrate the partitioning of the sums of squares in the ANOVA by calculating the complete ANOVA table for the following data. To determine diet quality, male weanling rats were fed diets with various protein levels. Each of 15 rats was randomly assigned to one of three diets, and their weight gain in grams was recorded.
Diet protein level:

Low	Medium	High
3.89	8.54	20.39
3.87	9.32	24.22
3.26	8.76	30.91
2.70	9.30	22.78
3.82	10.45	26.33

(b) Analytically verify the partitioning of the ANOVA sums of squares by completing the proof of Theorem 11.2.11, which states that for any numbers y_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{\bar{y}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2,$$

where $\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_j y_{ij}$ and $\bar{\bar{y}} = \sum_i n_i\bar{y}_{i\cdot}/\sum_i n_i$.
(c) Illustrate the relationship between the t and F statistics, given in Exercise 11.12(b), using the

data of part (a). That exercise states that with $n_i \equiv n$,

$$\frac{1}{k(k-1)}\sum_{i,i'} t_{ii'}^2 = F, \qquad t_{ii'}^2 = \frac{(\bar Y_{i\cdot} - \bar Y_{i'\cdot})^2}{S_p^2(2/n)}.$$

Solution. (a) $F = 110.43$ on 2 and 12 degrees of freedom: protein level matters enormously. Here $k = 3$, $n_i \equiv n = 5$, $N = 15$. Group totals 17.54, 46.37, 124.63 give

$$\bar y_{1\cdot} = 3.508, \quad \bar y_{2\cdot} = 9.274, \quad \bar y_{3\cdot} = 24.926, \quad \bar{\bar y} = \frac{188.54}{15} = 12.5693.$$

Then

$$\mathrm{SSB} = 5\big[(-9.0613)^2 + (-3.2953)^2 + (12.3567)^2\big] = 1228.2710,$$
$$\mathrm{SSW} = 1.0887 + 2.1887 + 63.4585 = 66.7359,$$

the three within-group terms being $\sum_j (y_{ij} - \bar y_{i\cdot})^2$ for low, medium, high. The direct total $\mathrm{SST} = \sum\sum (y_{ij} - \bar{\bar y})^2 = 1295.0069 = \mathrm{SSB} + \mathrm{SSW}$, illustrating (11.2.15).

Source of variation	df	Sum of squares	Mean square	F
Between diets	2	1228.2710	614.1355	110.43
Within diets	12	66.7359	5.5613	
Total	14	1295.0069		

Since $110.43 > F_{2,12,0.05} = 3.89$ (indeed $p \approx 1.9 \times 10^{-8}$), reject $H_0\colon \theta_1 = \theta_2 = \theta_3$.
(b) Expand the suggested decomposition $y_{ij} - \bar{\bar y} = (y_{ij} - \bar y_{i\cdot}) + (\bar y_{i\cdot} - \bar{\bar y})$:

$$\sum_i\sum_j (y_{ij} - \bar{\bar y})^2 = \sum_i\sum_j (y_{ij} - \bar y_{i\cdot})^2 + \sum_i\sum_j (\bar y_{i\cdot} - \bar{\bar y})^2 + 2\sum_i\sum_j (y_{ij} - \bar y_{i\cdot})(\bar y_{i\cdot} - \bar{\bar y}).$$

The middle term has summand free of j, so it equals $\sum_i n_i(\bar y_{i\cdot} - \bar{\bar y})^2$. In the cross term the second factor is free of j and the inner sum of the first factor vanishes by the definition of $\bar y_{i\cdot}$:

$$2\sum_{i=1}^{k} (\bar y_{i\cdot} - \bar{\bar y}) \sum_{j=1}^{n_i} (y_{ij} - \bar y_{i\cdot}) = 2\sum_{i=1}^{k} (\bar y_{i\cdot} - \bar{\bar y}) \cdot 0 = 0.$$

Regrouping gives (11.2.15).
(c) With $S_p^2 = \mathrm{MSW} = 5.5613$ and $S_p^2(2/n) = 2(5.5613)/5 = 2.22453$, the three distinct pairwise statistics are

$$t_{12}^2 = \frac{(3.508 - 9.274)^2}{2.22453} = \frac{33.2468}{2.22453} = 14.9455, \qquad t_{12} = 3.8659,$$
$$t_{13}^2 = \frac{(3.508 - 24.926)^2}{2.22453} = \frac{458.7307}{2.22453} = 206.2146, \quad t_{13} = 14.3602,$$
$$t_{23}^2 = \frac{(9.274 - 24.926)^2}{2.22453} = \frac{244.9851}{2.22453} = 110.1289, \quad t_{23} = 10.4942.$$

The sum in Exercise 11.12(b) runs over all $k(k-1) = 6$ ordered pairs, and $t_{ii'}^2 = t_{i'i}^2$, so

$$\frac{1}{k(k-1)}\sum_{i,i'} t_{ii'}^2 = \frac{2(14.9455 + 206.2146 + 110.1289)}{6} = \frac{662.5780}{6} = 110.4297 = F,$$

matching the F of the table in part (a).

11.4 Exercises 11.22–11.28

Problem 11.22. Calculate the expected values of MSB and MSW given in the oneway ANOVA table. (Such expectations are formally known as *expected mean squares* and can be used to help identify F tests in complicated ANOVAs. An algorithm exists for calculating expected mean squares. See, for example, Kirk 1982 for details about the algorithm.)
Recall Table 11.2.1 (ANOVA table for oneway classification), under the oneway ANOVA assumptions $Y_{ij} \sim n(\theta_i, \sigma^2)$ independent, $i = 1, \dots, k$, $j = 1, \dots, n_i$, with $N = \sum_i n_i$:

Source	df	Sum of squares	Mean square	F
Between	$k-1$	$\text{SSB} = \sum_i n_i(\bar{y}_{i\cdot} - \bar{\bar{y}})^2$	$\text{MSB} = \text{SSB}/(k-1)$	MSB/MSW
Within	$N-k$	$\text{SSW} = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2$	$\text{MSW} = \text{SSW}/(N-k)$	
Total	$N-1$	$\text{SST} = \sum_i \sum_j (y_{ij} - \bar{\bar{y}})^2$		

Solution.

$$\mathrm{E}\,\text{MSW} = \sigma^2, \qquad \mathrm{E}\,\text{MSB} = \sigma^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i(\theta_i - \bar{\theta})^2,$$

where $\bar{\theta} = \sum_i n_i\theta_i/N$ is the weighted grand mean.
For MSW, each within-group sum of squares is $(n_i - 1)S_i^2$ with $\mathrm{E}\,S_i^2 = \sigma^2$ (Theorem 5.2.6), so

$$\mathrm{E}\,\text{SSW} = \sum_{i=1}^{k}(n_i - 1)\sigma^2 = (N-k)\sigma^2.$$

For MSB, expand the square; the cross term collapses because $\sum_i n_i\bar{Y}_{i\cdot} = N\bar{\bar{Y}}$ (the weighted analogue of Theorem 5.2.4(b)):

$$\sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 = \sum_{i=1}^{k} n_i\bar{Y}_{i\cdot}^2 - N\bar{\bar{Y}}^2.$$

Since $\bar{Y}_{i\cdot} \sim n(\theta_i, \sigma^2/n_i)$ and $\bar{\bar{Y}} \sim n(\bar{\theta}, \sigma^2/N)$,

$$\begin{aligned}
\mathrm{E}\,\text{SSB} &= \sum_{i=1}^{k} n_i\left(\frac{\sigma^2}{n_i} + \theta_i^2\right) - N\left(\frac{\sigma^2}{N} + \bar{\theta}^2\right) \\
&= (k-1)\sigma^2 + \sum_{i=1}^{k} n_i\theta_i^2 - N\bar{\theta}^2 \\
&= (k-1)\sigma^2 + \sum_{i=1}^{k} n_i(\theta_i - \bar{\theta})^2,
\end{aligned}$$

the last line because $\sum_i n_i\theta_i = N\bar{\theta}$. Dividing by the degrees of freedom gives the two expected mean squares.

Problem 11.23. Use the model in Miscellanea 11.5.3.
(a) Show that the mean and variance of Y_{ij} are $\mathrm{E}\,Y_{ij} = \mu + \tau_i$ and $\mathrm{Var}\,Y_{ij} = \sigma_B^2 + \sigma^2$.
(b) If $\sum a_i = 0$, show that the unconditional variance of $\sum a_i\bar{Y}_{i\cdot}$ is

$$\mathrm{Var}\Big(\sum_i a_i\bar{Y}_{i\cdot}\Big) = \frac{1}{r}(\sigma^2 + \sigma_B^2)(1-\rho)\sum_i a_i^2,$$

where $\rho =$ intraclass correlation.
(Randomized Complete Block ANOVA assumptions, Miscellanea 11.5.3: random variables Y_{ij} are observed according to

$$Y_{ij} \mid \mathbf{b} = \mu + \tau_i + b_j + \epsilon_{ij}, \qquad i = 1, \dots, k, \quad j = 1, \dots, r,$$

where (i) $\epsilon_{ij} \sim$ iid $n(0, \sigma^2)$, and (ii) $B_1, \ldots, B_r$, whose realized but unobserved values are the blocks $b_1, \ldots, b_r$, are iid $n(0, \sigma_B^2)$ and independent of ϵ_{ij} for all i, j. The intraclass correlation is $\rho = \operatorname{Corr}(Y_{ij}, Y_{i'j}) = \sigma_B^2/(\sigma_B^2 + \sigma^2)$ for $i \neq i'$.)

Solution. (a) Unconditionally $Y_{ij} = \mu + \tau_i + B_j + \epsilon_{ij}$ with B_j and ϵ_{ij} independent, mean 0:

$$\mathrm{E}\, Y_{ij} = \mu + \tau_i + \mathrm{E}B_j + \mathrm{E}\epsilon_{ij} = \mu + \tau_i,$$

$$\operatorname{Var} Y_{ij} = \operatorname{Var} B_j + \operatorname{Var} \epsilon_{ij} = \sigma_B^2 + \sigma^2.$$

(Equivalently by conditioning: $\mathrm{E}(Y_{ij} \mid \mathbf{b}) = \mu + \tau_i + b_j$ and $\operatorname{Var}(Y_{ij} \mid \mathbf{b}) = \sigma^2$, so Theorems 4.4.3 and 4.4.7 give $\mu + \tau_i$ and $\sigma^2 + \operatorname{Var} B_j$.)
(b) Averaging over the blocks, $\bar{Y}_{i\cdot} = \mu + \tau_i + \bar{B} + \bar{\epsilon}_{i\cdot}$, where $\bar{B} = r^{-1} \sum_j B_j$ and $\bar{\epsilon}_{i\cdot} = r^{-1} \sum_j \epsilon_{ij}$ are independent, with $\operatorname{Var} \bar{B} = \sigma_B^2/r$ and $\operatorname{Var} \bar{\epsilon}_{i\cdot} = \sigma^2/r$, and $\bar{\epsilon}_{1\cdot}, \ldots, \bar{\epsilon}_{k\cdot}$ mutually independent. Hence

$$\operatorname{Var} \bar{Y}_{i\cdot} = \frac{\sigma_B^2 + \sigma^2}{r}, \qquad \operatorname{Cov}(\bar{Y}_{i\cdot}, \bar{Y}_{i'\cdot}) = \frac{\sigma_B^2}{r} \quad (i \neq i').$$

Since $\sum_i a_i = 0$, the common term $\bar{B}$ drops out of the contrast:

$$\begin{aligned} \sum_{i=1}^k a_i \bar{Y}_{i\cdot} &= \sum_{i=1}^k a_i(\mu + \tau_i) + \Big(\sum_{i=1}^k a_i\Big)\bar{B} + \sum_{i=1}^k a_i \bar{\epsilon}_{i\cdot} \\ &= \sum_{i=1}^k a_i \tau_i + \sum_{i=1}^k a_i \bar{\epsilon}_{i\cdot}, \end{aligned}$$

so by independence of the $\bar{\epsilon}_{i\cdot}$,

$$\operatorname{Var}\Big(\sum_i a_i \bar{Y}_{i\cdot}\Big) = \frac{\sigma^2}{r} \sum_{i=1}^k a_i^2.$$

This is the stated form, because with $\rho = \sigma_B^2/(\sigma_B^2 + \sigma^2)$,

$$(\sigma^2 + \sigma_B^2)(1 - \rho) = (\sigma^2 + \sigma_B^2) \cdot \frac{\sigma^2}{\sigma_B^2 + \sigma^2} = \sigma^2.$$

Problem 11.24. The form of the Stein estimator of Miscellanea 11.5.6 can be justified somewhat by an *empirical Bayes* argument given in Efron and Morris (1972), which can be quite useful in data analysis. Such an argument may have been known by Stein (1956), although he makes no mention of it. Let $X_i \sim n(\theta_i, 1)$, $i = 1, \ldots, p$, and θ_i be iid $n(0, \tau^2)$.
(a) Show that the X_is, marginally, are iid $n(0, \tau^2 + 1)$, and, hence, $\sum X_i^2/(\tau^2 + 1) \sim \chi_p^2$.
(b) Using the marginal distribution, show that $\mathrm{E}\big(1 - \big((p-2)/\sum_{j=1}^p X_j^2\big)\big) = \tau^2/(\tau^2 + 1)$ if $p \geq 3$. Thus, the Stein estimator of Miscellanea 11.5.6 is an empirical Bayes version of the Bayes estimator $\delta_i^\pi(\mathbf{X}) = [\tau^2/(\tau^2 + 1)]X_i$.
(c) Show that the argument fails if $p < 3$ by showing that $\mathrm{E}(1/Y) = \infty$ if $Y \sim \chi_p^2$ with $p < 3$.
The Stein estimator of Miscellanea 11.5.6 is

$$\delta_i^S(\mathbf{X}) = \left(1 - \frac{p-2}{\sum_{i=1}^p X_i^2}\right) X_i.$$

Solution. (a) Write $X_i = \theta_i + Z_i$ with $Z_i \sim n(0, 1)$ independent of $\theta_i \sim n(0, \tau^2)$. By Corollary 4.6.10 the sum of independent normals is normal, so $X_i \sim n(0, \tau^2 + 1)$; independence across i holds because

the pairs (θ_i, Z_i) are independent across i. Hence $X_i/\sqrt{\tau^2+1}$ are iid $n(0,1)$ and, by Lemma 5.3.2,

$$\frac{\sum_{i=1}^p X_i^2}{\tau^2+1} = \sum_{i=1}^p \left(\frac{X_i}{\sqrt{\tau^2+1}}\right)^2 \sim \chi_p^2.$$

(b) Put $W = \sum_j X_j^2/(\tau^2+1) \sim \chi_p^2$. For $p \geq 3$,

$$\mathrm{E}\frac{1}{W} = \frac{1}{\Gamma(p/2)2^{p/2}} \int_0^\infty y^{p/2-2} e^{-y/2}\, dy = \frac{\Gamma(p/2-1)2^{p/2-1}}{\Gamma(p/2)2^{p/2}} = \frac{1}{p-2},$$

using $\Gamma(p/2) = (p/2-1)\Gamma(p/2-1)$. Therefore

$$\begin{aligned}\mathrm{E}\left(1 - \frac{p-2}{\sum_j X_j^2}\right) &= 1 - \frac{p-2}{\tau^2+1}\,\mathrm{E}\frac{1}{W} \\ &= 1 - \frac{1}{\tau^2+1} = \frac{\tau^2}{\tau^2+1},\end{aligned}$$

so the random shrinkage factor of δ_i^S is an unbiased estimate of the Bayes shrinkage factor $\tau^2/(\tau^2+1)$ appearing in δ_i^π.
(c) For $Y \sim \chi_p^2$ with $p < 3$ (that is, $p = 1$ or $p = 2$), bound the integrand below on $(0,1)$ by $e^{-y/2} > e^{-1/2}$:

$$\mathrm{E}\frac{1}{Y} \geq \frac{e^{-1/2}}{\Gamma(p/2)2^{p/2}} \int_0^1 y^{p/2-2}\, dy = \infty,$$

since $p/2 - 2 \leq -1$ makes $\int_0^1 y^{p/2-2} dy$ divergent. Thus no multiple of $1/\sum_j X_j^2$ has finite expectation and the calculation of part (b) is unavailable.

Problem 11.25. In Section 11.3.1, we found the least squares estimators of α and β by a two-stage minimization. This minimization can also be done using partial derivatives.
(a) Compute $\frac{\partial \mathrm{RSS}}{\partial c}$ and $\frac{\partial \mathrm{RSS}}{\partial d}$ and set them equal to 0. Show that the resulting two equations can be written as

$$nc + \left(\sum_{i=1}^n x_i\right) d = \sum_{i=1}^n y_i \quad \text{and} \quad \left(\sum_{i=1}^n x_i\right) c + \left(\sum_{i=1}^n x_i^2\right) d = \sum_{i=1}^n x_i y_i.$$

(These equations are called the *normal equations* for this minimization problem.)
(b) Show that $c = a$ and $d = b$ are the solutions to the normal equations.
(c) Check the second partial derivative condition to verify that the point $c = a$ and $d = b$ is indeed the minimum of RSS.
(Here $\mathrm{RSS} = \sum_{i=1}^n (y_i - (c + dx_i))^2$ and, from (11.3.8), $b = S_{xy}/S_{xx}$, $a = \bar{y} - b\bar{x}$, with $S_{xx} = \sum_i (x_i - \bar{x})^2$ and $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$.)

Solution. (a) Differentiating $\mathrm{RSS} = \sum_i (y_i - c - dx_i)^2$,

$$\frac{\partial \mathrm{RSS}}{\partial c} = -2\sum_{i=1}^n (y_i - c - dx_i), \qquad \frac{\partial \mathrm{RSS}}{\partial d} = -2\sum_{i=1}^n x_i(y_i - c - dx_i).$$

Setting each to 0 and collecting terms gives exactly

$$nc + \Big(\sum_i x_i\Big) d = \sum_i y_i, \qquad \Big(\sum_i x_i\Big) c + \Big(\sum_i x_i^2\Big) d = \sum_i x_i y_i.$$

(b) Substitute $c = a = \bar{y} - b\bar{x}$ and $d = b$. The first equation:

$$n(\bar{y} - b\bar{x}) + n\bar{x}\, b = n\bar{y}. \quad \checkmark$$

The second, using $\sum_i x_i^2 = S_{xx} + n\bar{x}^2$ and $bS_{xx} = S_{xy}$:

$$n\bar{x}(\bar{y} - b\bar{x}) + (S_{xx} + n\bar{x}^2)b = n\bar{x}\bar{y} + bS_{xx}$$
$$= n\bar{x}\bar{y} + S_{xy} = \sum_{i=1}^{n} x_i y_i. \quad \checkmark$$

(The coefficient matrix has determinant $n\sum_i x_i^2 - (\sum_i x_i)^2 = nS_{xx} \neq 0$ whenever the x_i are not all equal, so this is the unique solution.)
(c) The Hessian of RSS is constant:

$$H = 2\begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}, \qquad \frac{\partial^2 \mathrm{RSS}}{\partial c^2} = 2n > 0,$$

$$\det H = 4\Big(n\sum_i x_i^2 - \big(\textstyle\sum_i x_i\big)^2\Big) = 4nS_{xx} > 0$$

(again provided the x_i are not all equal). So H is positive definite, RSS is a strictly convex function of (c, d), and the unique stationary point (a, b) is the global minimum.

Problem 11.26. Suppose n is an even number. The values of the predictor variable, $x_1, \ldots, x_n$, all must be chosen to be in the interval $[e, f]$. Show that the choice that maximizes S_{xx} is for half of the x_i equal to e and the other half equal to f. (This was the choice mentioned in Section 11.3.2 that minimizes $\operatorname{Var} b$.)

Solution. The maximum value is $n(f-e)^2/4$, attained only at the half-and-half choice. For the upper bound, centre at the midpoint $m = (e+f)/2$ rather than at $\bar{x}$; since $\bar{x}$ minimizes $\sum_i (x_i - c)^2$ over c (Theorem 5.2.4(a)),

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \le \sum_{i=1}^{n}(x_i - m)^2 \le n\left(\frac{f-e}{2}\right)^2,$$

the last step because $x_i \in [e, f]$ forces $|x_i - m| \le (f-e)/2$.
The half-and-half choice attains the bound: with $n/2$ of the x_i at e and $n/2$ at f we get $\bar{x} = m$ and every $|x_i - \bar{x}| = (f-e)/2$, so

$$S_{xx} = n\left(\frac{f-e}{2}\right)^2 = \frac{n(f-e)^2}{4},$$

which is therefore the maximum, and by (11.3.20) it minimizes $\operatorname{Var} b = \sigma^2/S_{xx}$.
It is the only maximizing choice when $e < f$: equality in the second inequality requires $x_i \in \{e, f\}$ for every i, and equality in the first requires $\bar{x} = m$; if q of the x_i equal e then $\bar{x} = m$ reads $[qe + (n-q)f]/n = (e+f)/2$, i.e. $(n - 2q)(f - e) = 0$, so $q = n/2$.

Problem 11.27. Observations (x_i, Y_i), $i = 1, \ldots, n$, follow the model $Y_i = \alpha + \beta x_i + \epsilon_i$, where $\mathrm{E}\,\epsilon_i = 0$, $\operatorname{Var}\epsilon_i = \sigma^2$, and $\operatorname{Cov}(\epsilon_i, \epsilon_j) = 0$ if $i \neq j$. Find the best linear unbiased estimator of α.

Solution. The BLUE is the least squares intercept

$$\hat{\alpha} = \bar{Y} - b\bar{x} = \sum_{i=1}^{n} d_i^* Y_i, \qquad d_i^* = \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}},$$

where $b = S_{xY}/S_{xx}$ is the BLUE of β from Section 11.3.2 and $S_{xx} > 0$ (the x_i are not all equal). Indeed $\sum_i d_i^* Y_i = \bar{Y} - (\bar{x}/S_{xx})\sum_i (x_i - \bar{x})Y_i = \bar{Y} - b\bar{x}$, the least squares a of (11.3.8).

Unbiasedness. As in (11.3.16), $\mathrm{E}\sum_i d_i Y_i = \alpha\sum_i d_i + \beta\sum_i d_i x_i$ equals α for all α, β if and only if

$$\sum_{i=1}^{n} d_i = 1 \qquad \text{and} \qquad \sum_{i=1}^{n} d_i x_i = 0,$$

and d^* qualifies: $\sum_i d_i^* = 1 - 0 = 1$ and $\sum_i d_i^* x_i = \bar{x} - \bar{x}S_{xx}/S_{xx} = 0$.
Minimum variance. The Y_i are uncorrelated with common variance σ^2, so $\mathrm{Var}\sum_i d_i Y_i = \sigma^2\sum_i d_i^2$ and it suffices to minimize $\sum_i d_i^2$ under those two constraints. For any admissible d put $\eta_i = d_i - d_i^*$; then $\sum_i \eta_i = 0$ and $\sum_i \eta_i x_i = 0$, hence $\sum_i \eta_i(x_i - \bar{x}) = 0$ and

$$\sum_{i=1}^{n} d_i^* \eta_i = \frac{1}{n}\sum_i \eta_i - \frac{\bar{x}}{S_{xx}}\sum_i \eta_i(x_i - \bar{x}) = 0,$$

so that

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} d_i^{*2} + \sum_{i=1}^{n} \eta_i^2 \;\geq\; \sum_{i=1}^{n} d_i^{*2},$$

with equality only at $\eta \equiv 0$; so $\hat{\alpha}$ is the unique BLUE. Its variance is

$$\begin{aligned}\sigma^2 \sum_{i=1}^{n} d_i^{*2} &= \sigma^2 \left(\frac{1}{n} - \frac{2\bar{x}}{nS_{xx}} \sum_i (x_i - \bar{x}) \right. \\ &\qquad \left. + \frac{\bar{x}^2}{S_{xx}^2} \sum_i (x_i - \bar{x})^2 \right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right).\end{aligned}$$

Problem 11.28. Show that in the conditional normal model for simple linear regression, the MLE of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

(The conditional normal model (11.3.22): $Y_i \sim n(\alpha + \beta x_i, \sigma^2)$, independent, $i = 1, \ldots, n$, with $x_1, \ldots, x_n$ known fixed constants; $\hat{\alpha}$ and $\hat{\beta}$ are the MLEs of α and β, which coincide with the least squares a and b.)

Solution. From the joint pdf (11.3.24), the log likelihood is

$$\log L(\alpha, \beta, \sigma^2 \mid \mathbf{x}, \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

For every fixed $\sigma^2 > 0$ the only (α, β)-dependence is through $-\sum_i (y_i - \alpha - \beta x_i)^2/(2\sigma^2)$, so $\log L$ is maximized over (α, β) at the RSS minimizers, $\hat{\alpha} = a$ and $\hat{\beta} = b$ of Section 11.3.1 (Exercise 11.25), and these do not depend on σ^2. Writing $\mathrm{RSS} = \sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$, the profile log likelihood is

$$\ell(\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\mathrm{RSS}}{2\sigma^2}.$$

Differentiating in σ^2,

$$\frac{d\ell}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{\mathrm{RSS}}{2(\sigma^2)^2} = \frac{n}{2(\sigma^2)^2}\left(\frac{\mathrm{RSS}}{n} - \sigma^2 \right),$$

which (on $\mathrm{RSS} > 0$, an event of probability 1) is positive for $\sigma^2 < \mathrm{RSS}/n$ and negative for $\sigma^2 > \mathrm{RSS}/n$. Hence ℓ increases then decreases and the unique global maximum is at

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i)^2,$$

exactly as in the ordinary normal sampling computation of Example 7.2.11.

11.5 Exercises 11.29–11.35

Problem 11.29. Consider the residuals $\hat\epsilon_1,\dots,\hat\epsilon_n$ defined in Section 11.3.4 by $\hat\epsilon_i = Y_i - \hat\alpha - \hat\beta x_i$.
(a) Show that $\mathrm{E}\hat\epsilon_i = 0$.
(b) Verify that

$$\begin{aligned}\operatorname{Var}\hat\epsilon_i = \operatorname{Var} Y_i + \operatorname{Var}\hat\alpha + x_i^2 \operatorname{Var}\hat\beta \\ - 2\operatorname{Cov}(Y_i,\hat\alpha) - 2x_i\operatorname{Cov}(Y_i,\hat\beta) \\ + 2x_i\operatorname{Cov}(\hat\alpha,\hat\beta).\end{aligned}$$

(c) Use Lemma 11.3.2 to show that

$$\operatorname{Cov}(Y_i,\hat\alpha) = \sigma^2\left(\frac{1}{n} + \frac{(x_i-\bar x)\bar x}{S_{xx}}\right), \qquad \operatorname{Cov}(Y_i,\hat\beta) = \sigma^2\frac{x_i-\bar x}{S_{xx}},$$

and use these to verify (11.3.28), namely

$$\begin{aligned}\operatorname{Var}\hat\epsilon_i = \left(\frac{n-2}{n} + \frac{1}{S_{xx}}\Big(\frac{1}{n}\sum_{j=1}^n x_j^2 + x_i^2 \right. \\ \left. - 2(x_i-\bar x)^2 - 2x_i\bar x\Big)\right)\sigma^2.\end{aligned}$$

Solution. (a) $\mathrm{E}\hat\epsilon_i = \mathrm{E}Y_i - \mathrm{E}\hat\alpha - x_i\mathrm{E}\hat\beta = (\alpha+\beta x_i) - \alpha - \beta x_i = 0$, since $\hat\alpha,\hat\beta$ are unbiased (Theorem 11.3.3).
(b) $\hat\epsilon_i$ is the linear combination $(1)Y_i + (-1)\hat\alpha + (-x_i)\hat\beta$, and for any three random variables $\operatorname{Var}(a_1U_1 + a_2U_2 + a_3U_3) = \sum_k a_k^2 \operatorname{Var}U_k + 2\sum_{k<l} a_k a_l \operatorname{Cov}(U_k,U_l)$. With $(a_1,a_2,a_3) = (1,-1,-x_i)$ the three cross terms are $-2\operatorname{Cov}(Y_i,\hat\alpha)$, $-2x_i\operatorname{Cov}(Y_i,\hat\beta)$ and $+2x_i\operatorname{Cov}(\hat\alpha,\hat\beta)$, which is the display.
(c) The printed sign is a misprint: the correct formula carries a minus, $\operatorname{Cov}(Y_i,\hat\alpha) = \sigma^2\big(\frac{1}{n} - (x_i - \bar x)\bar x/S_{xx}\big)$, as the derivation below shows (and as is forced by $\hat\alpha = \sum c_i Y_i$ with the c_i of Theorem 11.3.3).
Write all three variables as linear combinations of the uncorrelated $Y_1,\dots,Y_n$ (each with variance σ^2): $Y_i = \sum_j \delta_{ij} Y_j$, and by Theorem 11.3.3 and (11.3.30),

$$\hat\alpha = \sum_j c_j Y_j, \quad \hat\beta = \sum_j d_j Y_j, \quad c_j = \frac{1}{n} - \frac{(x_j-\bar x)\bar x}{S_{xx}}, \quad d_j = \frac{x_j-\bar x}{S_{xx}}.$$

Lemma 11.3.2 applies (no normality or independence needed, only $\operatorname{Var}Y_i = \sigma^2$ and uncorrelatedness) and gives

$$\operatorname{Cov}(Y_i,\hat\alpha) = \sigma^2\sum_j \delta_{ij}c_j = \sigma^2 c_i = \sigma^2\left(\frac{1}{n} - \frac{(x_i-\bar x)\bar x}{S_{xx}}\right),$$

$$\operatorname{Cov}(Y_i,\hat\beta) = \sigma^2\sum_j \delta_{ij}d_j = \sigma^2 d_i = \sigma^2\frac{x_i-\bar x}{S_{xx}}.$$

Now substitute into (b), using $\operatorname{Var}\hat\alpha = \sigma^2\sum_j x_j^2/(nS_{xx})$, $\operatorname{Var}\hat\beta = \sigma^2/S_{xx}$ and $\operatorname{Cov}(\hat\alpha,\hat\beta) = -\sigma^2\bar x/S_{xx}$

from Theorem 11.3.3. Dividing by σ^2 and writing $\frac{1}{n}\sum_j x_j^2 = \frac{S_{xx}}{n} + \bar{x}^2$,

$$\begin{aligned}\frac{\operatorname{Var}\hat{\epsilon}_i}{\sigma^2} &= 1 + \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) + \frac{x_i^2}{S_{xx}} - \frac{2}{n} + \frac{2(x_i-\bar{x})\bar{x}}{S_{xx}} \\ &\quad - \frac{2x_i(x_i-\bar{x})}{S_{xx}} - \frac{2x_i\bar{x}}{S_{xx}} \\ &= 1 - \frac{1}{n} + \frac{(x_i-\bar{x})^2 - 2(x_i-\bar{x})^2}{S_{xx}} \\ &= 1 - \frac{1}{n} - \frac{(x_i-\bar{x})^2}{S_{xx}},\end{aligned}$$

where the middle step collects $\bar{x}^2 + x_i^2 - 2x_i\bar{x} = (x_i-\bar{x})^2$ and $2(x_i-\bar{x})\bar{x} - 2x_i(x_i-\bar{x}) = -2(x_i-\bar{x})^2$. Finally this is exactly (11.3.28): in that display $\frac{n-2}{n} = 1 - \frac{2}{n}$, and

$$\begin{aligned}\frac{1}{n}\sum_j x_j^2 + x_i^2 - 2(x_i-\bar{x})^2 - 2x_i\bar{x} &= \frac{1}{n}\sum_j x_j^2 - \bar{x}^2 - (x_i-\bar{x})^2 \\ &= \frac{S_{xx}}{n} - (x_i-\bar{x})^2,\end{aligned}$$

so the bracket of (11.3.28) equals $1 - \frac{2}{n} + \frac{1}{n} - \frac{(x_i-\bar{x})^2}{S_{xx}}$, as required.

Problem 11.30. Fill in the details about the distribution of $\hat{\alpha}$ left out of the proof of Theorem 11.3.3.
(a) Show that the estimator $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ can be expressed as $\hat{\alpha} = \sum_{i=1}^n c_i Y_i$, where

$$c_i = \frac{1}{n} - \frac{(x_i-\bar{x})\bar{x}}{S_{xx}}.$$

(b) Verify that

$$\mathrm{E}\hat{\alpha} = \alpha \qquad \text{and} \qquad \operatorname{Var}\hat{\alpha} = \sigma^2\left[\frac{1}{nS_{xx}}\sum_{i=1}^n x_i^2\right].$$

(c) Verify that

$$\operatorname{Cov}(\hat{\alpha},\hat{\beta}) = -\frac{\sigma^2\bar{x}}{S_{xx}}.$$

Solution. (a) Substitute $\hat{\beta} = \sum_i d_i Y_i$, $d_i = (x_i-\bar{x})/S_{xx}$ (Section 11.3.2, (11.3.19)):

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} = \sum_{i=1}^n \frac{Y_i}{n} - \bar{x}\sum_{i=1}^n \frac{(x_i-\bar{x})}{S_{xx}}Y_i = \sum_{i=1}^n \left[\frac{1}{n} - \frac{(x_i-\bar{x})\bar{x}}{S_{xx}}\right]Y_i.$$

(b) The two identities $\sum_i(x_i-\bar{x}) = 0$ and $\sum_i(x_i-\bar{x})x_i = \sum_i(x_i-\bar{x})^2 = S_{xx}$ do all the work here and in (c). They give $\sum_i c_i = 1 - \frac{\bar{x}}{S_{xx}}\sum_i(x_i-\bar{x}) = 1$ and $\sum_i c_i x_i = \bar{x} - \frac{\bar{x}}{S_{xx}}S_{xx} = 0$, so with $\mathrm{E}Y_i = \alpha + \beta x_i$,

$$\mathrm{E}\hat{\alpha} = \sum_i c_i(\alpha+\beta x_i) = \alpha\sum_i c_i + \beta\sum_i c_i x_i = \alpha.$$

By Lemma 11.3.2 (with d_i replaced by c_i), $\operatorname{Var}\hat{\alpha} = \sigma^2\sum_i c_i^2$, and

$$\begin{aligned}\sum_i c_i^2 &= \sum_i\left[\frac{1}{n^2} - \frac{2(x_i-\bar{x})\bar{x}}{nS_{xx}} + \frac{(x_i-\bar{x})^2\bar{x}^2}{S_{xx}^2}\right] \\ &= \frac{1}{n} - 0 + \frac{\bar{x}^2}{S_{xx}} = \frac{S_{xx}+n\bar{x}^2}{nS_{xx}} = \frac{1}{nS_{xx}}\sum_i x_i^2,\end{aligned}$$

the last step because $S_{xx} = \sum_i x_i^2 - n\bar{x}^2$.
(c) By Lemma 11.3.2,

$$\begin{aligned}\operatorname{Cov}(\hat{\alpha}, \hat{\beta}) = \sigma^2 \sum_i c_i d_i &= \sigma^2 \sum_i \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right] \frac{x_i - \bar{x}}{S_{xx}} \\ &= \frac{\sigma^2}{nS_{xx}} \sum_i (x_i - \bar{x}) - \frac{\sigma^2 \bar{x}}{S_{xx}^2} \sum_i (x_i - \bar{x})^2 \\ &= 0 - \frac{\sigma^2 \bar{x}}{S_{xx}} = -\frac{\sigma^2 \bar{x}}{S_{xx}}.\end{aligned}$$

Problem 11.31. Verify the claim in Theorem 11.3.3, that $\hat{\epsilon}_i$ is uncorrelated with $\hat{\alpha}$ and $\hat{\beta}$. (Show that $\hat{\epsilon}_i = \sum_j e_j Y_j$, where the e_js are given by (11.3.30), namely

$$\hat{\epsilon}_i = \sum_{j=1}^{n} \big[\delta_{ij} - (c_j + d_j x_i)\big] Y_j,$$

with

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}, \quad c_j = \frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{S_{xx}}, \quad d_j = \frac{x_j - \bar{x}}{S_{xx}}.$$

Then, using the facts that we can write $\hat{\alpha} = \sum_j c_j Y_j$ and $\hat{\beta} = \sum_j d_j Y_j$, verify that $\sum_j e_j c_j = \sum_j e_j d_j = 0$ and apply Lemma 11.3.2.)

Solution. $\operatorname{Cov}(\hat{\epsilon}_i, \hat{\alpha}) = \operatorname{Cov}(\hat{\epsilon}_i, \hat{\beta}) = 0$ for every i.
Substituting $\hat{\alpha} = \sum_j c_j Y_j$ and $\hat{\beta} = \sum_j d_j Y_j$ into $\hat{\epsilon}_i = Y_i - \hat{\alpha} - \hat{\beta} x_i$ and writing $Y_i = \sum_j \delta_{ij} Y_j$ gives (11.3.30), i.e. $\hat{\epsilon}_i = \sum_j e_j Y_j$ with $e_j = \delta_{ij} - c_j - d_j x_i$.
The sums needed are all in Exercise 11.30 (or follow from $\sum_j (x_j - \bar{x}) = 0$, $\sum_j (x_j - \bar{x}) x_j = S_{xx}$):

$$\sum_j c_j^2 = \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}, \quad \sum_j d_j^2 = \frac{1}{S_{xx}}, \quad \sum_j c_j d_j = -\frac{\bar{x}}{S_{xx}},$$

together with $\sum_j \delta_{ij} c_j = c_i$ and $\sum_j \delta_{ij} d_j = d_i$. Hence

$$\begin{aligned}\sum_j e_j c_j &= c_i - \sum_j c_j^2 - x_i \sum_j c_j d_j \\ &= \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} - \frac{1}{n} - \frac{\bar{x}^2}{S_{xx}} + \frac{x_i \bar{x}}{S_{xx}} \\ &= \frac{-x_i \bar{x} + \bar{x}^2 - \bar{x}^2 + x_i \bar{x}}{S_{xx}} = 0,\end{aligned}$$

and

$$\begin{aligned}\sum_j e_j d_j &= d_i - \sum_j c_j d_j - x_i \sum_j d_j^2 \\ &= \frac{x_i - \bar{x}}{S_{xx}} + \frac{\bar{x}}{S_{xx}} - \frac{x_i}{S_{xx}} = 0.\end{aligned}$$

The Y_j are uncorrelated with common variance σ^2, so Lemma 11.3.2 applies to each pair of linear forms and yields

$$\operatorname{Cov}(\hat{\epsilon}_i, \hat{\alpha}) = \sigma^2 \sum_j e_j c_j = 0, \qquad \operatorname{Cov}(\hat{\epsilon}_i, \hat{\beta}) = \sigma^2 \sum_j e_j d_j = 0.$$

Problem 11.32. Observations (x_i, Y_i), $i = 1, \ldots, n$, are made according to the model

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

where $x_1, \ldots, x_n$ are fixed constants and $\epsilon_1, \ldots, \epsilon_n$ are iid $\mathrm{n}(0, \sigma^2)$. The model is then reparameterized as

$$Y_i = \alpha' + \beta'(x_i - \bar{x}) + \epsilon_i.$$

Let $\hat{\alpha}$ and $\hat{\beta}$ denote the MLEs of α and β, respectively, and $\hat{\alpha}'$ and $\hat{\beta}'$ denote the MLEs of α' and β', respectively.
(a) Show that $\hat{\beta}' = \hat{\beta}$.
(b) Show that $\hat{\alpha}' \neq \hat{\alpha}$. In fact, show that $\hat{\alpha}' = \bar{Y}$. Find the distribution of $\hat{\alpha}'$.
(c) Show that $\hat{\alpha}'$ and $\hat{\beta}'$ are uncorrelated and, hence, independent under normality.

Solution. $\hat{\beta}' = \hat{\beta}$ and $\hat{\alpha}' = \bar{Y}$. The second model is the conditional normal model (11.3.22) run on the centered predictor $z_i = x_i - \bar{x}$, so Section 11.3.4 gives $\hat{\beta}' = S_{zy}/S_{zz}$ and $\hat{\alpha}' = \bar{Y} - \hat{\beta}'\bar{z}$, and $\bar{z} = 0$ makes

$$S_{zz} = \sum_i (z_i - \bar{z})^2 = \sum_i (x_i - \bar{x})^2 = S_{xx}, \qquad S_{zy} = \sum_i (x_i - \bar{x})(Y_i - \bar{Y}) = S_{xy}.$$

(a) $\hat{\beta}' = S_{zy}/S_{zz} = S_{xy}/S_{xx} = \hat{\beta}$.
(b) $\hat{\alpha}' = \bar{Y} - \hat{\beta}' \cdot 0 = \bar{Y}$, whereas $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$; these differ unless $\bar{x} = 0$ or $\hat{\beta} = 0$, and so do the parameters, since $\alpha + \beta x_i = (\alpha + \beta\bar{x}) + \beta(x_i - \bar{x})$ gives $\alpha' = \alpha + \beta\bar{x}$. As $\bar{Y}$ is an average of independent $\mathrm{n}(\alpha + \beta x_i, \sigma^2)$ variables,

$$\hat{\alpha}' = \bar{Y} \sim \mathrm{n}\left(\alpha + \beta\bar{x},\ \frac{\sigma^2}{n}\right) = \mathrm{n}\left(\alpha',\ \frac{\sigma^2}{n}\right).$$

(c) Write $\hat{\alpha}' = \sum_i \frac{1}{n} Y_i$ and $\hat{\beta}' = \sum_i d_i Y_i$ with $d_i = (x_i - \bar{x})/S_{xx}$. The Y_i are uncorrelated with variance σ^2, so Lemma 11.3.2 gives

$$\mathrm{Cov}(\hat{\alpha}', \hat{\beta}') = \sigma^2 \sum_{i=1}^{n} \frac{1}{n} \cdot \frac{x_i - \bar{x}}{S_{xx}} = \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

Both are linear functions of the independent normal Y_i, hence jointly bivariate normal, and for the bivariate normal zero correlation implies independence.

Problem 11.33. Observations (X_i, Y_i), $i = 1, \ldots, n$, are made from a bivariate normal population with parameters $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, and the model $Y_i = \alpha + \beta x_i + \epsilon_i$ is going to be fit.
(a) Argue that the hypothesis $H_0\colon \beta = 0$ is true if and only if the hypothesis $H_0'\colon \rho = 0$ is true. (See (11.3.25), which gives $\mathrm{E}(Y \mid x) = \left[\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X\right] + \left[\rho\frac{\sigma_Y}{\sigma_X}\right] x$.)
(b) Show algebraically that

$$\frac{\hat{\beta}}{S/\sqrt{S_{xx}}} = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}},$$

where r is the sample correlation coefficient, the MLE of ρ.
(c) Show how to test $H_0'\colon \rho = 0$, given only r^2 and n, using Student's t with $n-2$ degrees of freedom (see (11.3.33)). (Fisher derived an approximate confidence interval for ρ, using a variance-stabilizing transformation. See Exercise 11.4.)

Solution. (a) Matching (11.3.25) against $\mathrm{E}(Y \mid x) = \alpha + \beta x$ identifies $\beta = \rho\,\sigma_Y/\sigma_X$. A bivariate normal distribution has $\sigma_X, \sigma_Y > 0$, so σ_Y/σ_X is a strictly positive constant and $\beta = 0 \iff \rho = 0$.
(b) Write $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$ and $\hat{\beta} = S_{xy}/S_{xx}$. By (11.3.36) together with $\sum_i (\hat{y}_i - \bar{y})^2 = S_{xy}^2/S_{xx}$

(Exercise 11.34(c)),

$$\mathrm{RSS} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = S_{yy}\left(1 - r^2\right),$$

so by (11.3.29) $S^2 = \mathrm{RSS}/(n-2) = S_{yy}(1-r^2)/(n-2)$. Therefore

$$\frac{\hat\beta}{S/\sqrt{S_{xx}}} = \frac{S_{xy}}{S_{xx}} \cdot \frac{\sqrt{S_{xx}}}{S} = \frac{S_{xy}}{\sqrt{S_{xx}}\,S} = \frac{S_{xy}\sqrt{n-2}}{\sqrt{S_{xx}}\sqrt{S_{yy}(1-r^2)}} = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}.$$

(Here $r^2 < 1$, i.e. $\mathrm{RSS} > 0$, with probability 1.)
(c) By part (a), testing $H_0'\colon \rho = 0$ is testing $H_0\colon \beta = 0$, and by (11.3.33) the statistic $\hat\beta/(S/\sqrt{S_{xx}})$ has a t_{n-2} distribution when $\beta = 0$. By (b) that statistic is a function of r and n alone, so

$$T = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} \sim t_{n-2} \quad \text{under } H_0'.$$

Reject $H_0'\colon \rho = 0$ at level α against $H_1'\colon \rho \neq 0$ if

$$|T| = \sqrt{\frac{(n-2)r^2}{1-r^2}} > t_{n-2,\alpha/2},$$

which requires only r^2 and n (the sign of r is irrelevant for the two-sided test).

Problem 11.34. (a) Illustrate the partitioning of the sum of squares for simple linear regression by calculating the regression ANOVA table for the following data. Parents are often interested in predicting the eventual heights of their children. The following is a portion of the data taken from a study that might have been suggested by Galton's analysis.

Height (inches) at age 2 (x)	39	30	32	34	35	36	36	30
Height (inches) as an adult (y)	71	63	63	67	68	68	70	64

(b) Analytically establish the partitioning of the sum of squares for simple linear regression by verifying (11.3.36), namely

$$\sum_{i=1}^n (y_i - \bar y)^2 = \sum_{i=1}^n (\hat y_i - \bar y)^2 + \sum_{i=1}^n (y_i - \hat y_i)^2, \qquad \hat y_i = \hat\alpha + \hat\beta x_i.$$

(c) Prove that the two expressions for the regression sum of squares are, in fact, equal; that is, show that

$$\sum_{i=1}^n (\hat y_i - \bar y)^2 = \frac{S_{xy}^2}{S_{xx}}.$$

(d) Show that the coefficient of determination, r^2, given by

$$r^2 = \frac{\sum_{i=1}^n (\hat y_i - \bar y)^2}{\sum_{i=1}^n (y_i - \bar y)^2},$$

can be derived as the square of the sample correlation coefficient either of the n pairs $(y_1, x_1), \ldots, (y_n, x_n)$ or of the n pairs $(y_1, \hat y_1), \ldots, (y_n, \hat y_n)$.

Solution. (a) $F = 50.7$ on $(1, 6)$ degrees of freedom: the fit is highly significant. From $n = 8$, $\sum x_i = 272$, $\sum y_i = 534$, $\sum x_i^2 = 9318$, $\sum y_i^2 = 35712$, $\sum x_i y_i = 18221$,

$$\bar x = 34, \quad \bar y = 66.75, \quad S_{xx} = 70, \quad S_{yy} = 67.5, \quad S_{xy} = 65,$$

so $\hat\beta = 65/70 = 13/14 = 0.9286$ and $\hat\alpha = 66.75 - \frac{13}{14}(34) = 985/28 = 35.179$. The two components are

$$\text{Reg. SS} = \frac{S_{xy}^2}{S_{xx}} = \frac{65^2}{70} = \frac{845}{14} = 60.357,$$

$$\text{RSS} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 67.5 - \frac{845}{14} = \frac{50}{7} = 7.143,$$

and they add to $\text{SST} = S_{yy} = 67.5$, illustrating (11.3.36). In the format of Table 11.3.2:

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic
Regression (slope)	1	60.357	60.357	50.70
Residual	6	7.143	1.190	
Total	7	67.500		

Since $F = 50.70 > F_{1,6,.05} = 5.99$ (indeed $p = .00039$), $H_0\colon \beta = 0$ is rejected. Also $r^2 = 60.357/67.5 = .894$.

(b) Add and subtract $\hat y_i$ and expand:

$$\sum_i (y_i - \bar y)^2 = \sum_i (\hat y_i - \bar y)^2 + \sum_i (y_i - \hat y_i)^2 + 2\sum_i (\hat y_i - \bar y)(y_i - \hat y_i).$$

The cross-term vanishes. Indeed $\hat y_i - \bar y = \hat\beta(x_i - \bar x)$ because $\hat\alpha = \bar y - \hat\beta\bar x$, and $y_i - \hat y_i = (y_i - \bar y) - \hat\beta(x_i - \bar x)$, so

$$\begin{aligned}\sum_i (\hat y_i - \bar y)(y_i - \hat y_i) &= \hat\beta \sum_i (x_i - \bar x)\big[(y_i - \bar y) - \hat\beta(x_i - \bar x)\big] \\ &= \hat\beta\left(S_{xy} - \hat\beta S_{xx}\right) = 0,\end{aligned}$$

since $\hat\beta = S_{xy}/S_{xx}$.

(c)

$$\sum_i (\hat y_i - \bar y)^2 = \hat\beta^2 \sum_i (x_i - \bar x)^2 = \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}.$$

(d) By (c), $r^2 = \dfrac{S_{xy}^2/S_{xx}}{S_{yy}} = \dfrac{S_{xy}^2}{S_{xx}S_{yy}}$, which is the square of the sample correlation coefficient of the pairs (y_i, x_i).

For the pairs $(y_i, \hat y_i)$, use $\hat y_i - \bar{\hat y} = \hat y_i - \bar y = \hat\beta(x_i - \bar x)$, so

$$S_{y\hat y} = \hat\beta S_{xy} = \frac{S_{xy}^2}{S_{xx}}, \qquad S_{\hat y\hat y} = \hat\beta^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}},$$

and the squared sample correlation of $(y_i, \hat y_i)$ is

$$\frac{S_{y\hat y}^2}{S_{yy}S_{\hat y\hat y}} = \frac{(S_{xy}^2/S_{xx})^2}{S_{yy}\,(S_{xy}^2/S_{xx})} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2.$$

Problem 11.35. Observations $Y_1, \ldots, Y_n$ are described by the relationship $Y_i = \theta x_i^2 + \epsilon_i$, where $x_1, \ldots, x_n$ are fixed constants and $\epsilon_1, \ldots, \epsilon_n$ are iid $\text{n}(0, \sigma^2)$.
(a) Find the least squares estimator of θ.
(b) Find the MLE of θ.
(c) Find the best unbiased estimator of θ.

Solution. All three are

$$\hat\theta = \frac{\sum_{i=1}^n x_i^2 Y_i}{\sum_{i=1}^n x_i^4} \qquad \text{(assuming not all } x_i = 0\text{)}.$$

(a) $\mathrm{RSS}(\theta) = \sum_i (y_i - \theta x_i^2)^2$ is a strictly convex quadratic in θ (leading coefficient $\sum_i x_i^4 > 0$), so its unique minimizer is the root of

$$\frac{d}{d\theta}\mathrm{RSS}(\theta) = -2\sum_i x_i^2\big(y_i - \theta x_i^2\big) = 0, \qquad \text{i.e.} \qquad \hat\theta = \frac{\sum_i x_i^2 y_i}{\sum_i x_i^4}.$$

(b) The Y_i are independent $\mathrm{n}(\theta x_i^2, \sigma^2)$, so

$$L(\theta, \sigma^2 \mid \mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_i (y_i - \theta x_i^2)^2\right],$$

which for each fixed σ^2 is maximized over θ by minimizing $\mathrm{RSS}(\theta)$. The maximizing θ therefore does not depend on σ^2 (as in Section 11.3.4), so the MLE is the least squares estimator of (a).
(c) Write the joint pdf in exponential family form:

$$f(\mathbf{y} \mid \theta, \sigma^2) = \frac{e^{-\theta^2 \sum_i x_i^4/(2\sigma^2)}}{(2\pi\sigma^2)^{n/2}} \exp\left[\frac{\theta}{\sigma^2}\sum_i x_i^2 y_i - \frac{1}{2\sigma^2}\sum_i y_i^2\right].$$

The natural parameter is $\eta = \big(\theta/\sigma^2,\ -1/(2\sigma^2)\big)$, and as (θ, σ^2) ranges over $\mathbb{R} \times (0, \infty)$ the point η ranges over the open half-plane $\{(\eta_1, \eta_2) : \eta_2 < 0\}$, which contains an open subset of $\mathbb{R}^2$. So the family is full rank and, by Theorem 6.2.25,

$$T(\mathbf{Y}) = \left(\sum_i x_i^2 Y_i,\ \sum_i Y_i^2\right)$$

is a complete sufficient statistic. Now $\hat\theta$ is a function of T and is unbiased:

$$\mathrm{E}\hat\theta = \frac{\sum_i x_i^2\, \mathrm{E}Y_i}{\sum_i x_i^4} = \frac{\theta \sum_i x_i^4}{\sum_i x_i^4} = \theta.$$

By Theorem 7.3.23 it is therefore the best unbiased estimator of θ, with

$$\mathrm{Var}\,\hat\theta = \frac{\sigma^2 \sum_i x_i^4}{\big(\sum_i x_i^4\big)^2} = \frac{\sigma^2}{\sum_i x_i^4}.$$

11.6 Exercises 11.36–11.41

Problem 11.36. Observations $Y_1, \ldots, Y_n$ are made according to the model $Y_i = \alpha + \beta x_i + \epsilon_i$, where $x_1, \ldots, x_n$ are fixed constants and $\epsilon_1, \ldots, \epsilon_n$ are iid $\mathrm{n}(0, \sigma^2)$. Let $\hat\alpha$ and $\hat\beta$ denote MLEs of α and β.
(a) Assume that $x_1, \ldots, x_n$ are observed values of iid random variables $X_1, \ldots, X_n$ with distribution $\mathrm{n}(\mu_X, \sigma_X^2)$. Prove that when we take expectations over the joint distribution of X and Y, we still get $\mathrm{E}\hat\alpha = \alpha$ and $\mathrm{E}\hat\beta = \beta$.
(b) The phenomenon of part (a) does not carry over to the covariance. Calculate the unconditional covariance of $\hat\alpha$ and $\hat\beta$ (using the joint distribution of X and Y).

Solution. (a) Condition on $X = (X_1, \ldots, X_n)$ and use the tower property; the conditional estimators are unbiased for each realization of X, so averaging changes nothing. With $\hat\beta = S_{XY}/S_{XX}$ and $\hat\alpha = \bar Y - \hat\beta \bar X$, Theorem 11.3.3 gives $\mathrm{E}[\hat\beta \mid X] = \beta$ and $\mathrm{E}[\hat\alpha \mid X] = \alpha$ on the event $\{S_{XX} > 0\}$, which has probability 1 for $n \geq 2$ since the X_i are continuous. Hence

$$\mathrm{E}\hat{\beta} = \mathrm{E}\big[\mathrm{E}[\hat{\beta} \mid X]\big] = \mathrm{E}[\beta] = \beta,$$
$$\mathrm{E}\hat{\alpha} = \mathrm{E}\big[\mathrm{E}[\hat{\alpha} \mid X]\big] = \mathrm{E}[\alpha] = \alpha.$$

(b) The unconditional covariance is

$$\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \mu_X}{(n-3)\,\sigma_X^2}, \qquad n > 3,$$

and it fails to exist for $n \le 3$, since then $\mathrm{E}[1/S_{XX}] = \infty$.
By the conditional covariance identity (the bivariate form of Theorem 4.4.7),

$$\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = \mathrm{E}\big[\mathrm{Cov}(\hat{\alpha}, \hat{\beta} \mid X)\big] + \mathrm{Cov}\big(\mathrm{E}[\hat{\alpha} \mid X], \mathrm{E}[\hat{\beta} \mid X]\big).$$

The second term vanishes because by (a) the two conditional means are the constants α and β. For the first, Theorem 11.3.3 gives $\mathrm{Cov}(\hat{\alpha}, \hat{\beta} \mid X) = -\sigma^2 \bar{X}/S_{XX}$, so

$$\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = -\sigma^2\,\mathrm{E}\left[\frac{\bar{X}}{S_{XX}}\right] = -\sigma^2\,\mathrm{E}[\bar{X}]\,\mathrm{E}\left[\frac{1}{S_{XX}}\right],$$

the factorization being legitimate because $\bar{X}$ and S_{XX} are independent for a normal sample (Theorem 5.3.1). Finally $S_{XX}/\sigma_X^2 \sim \chi^2_{n-1}$, and for $W \sim \chi^2_k$ with $k > 2$,

$$\mathrm{E}\,W^{-1} = \int_0^\infty \frac{w^{k/2-2} e^{-w/2}}{\Gamma(k/2)2^{k/2}}\,dw = \frac{\Gamma(\frac{k}{2}-1)2^{k/2-1}}{\Gamma(\frac{k}{2})2^{k/2}} = \frac{1}{k-2}.$$

With $k = n-1$ this gives $\mathrm{E}[1/S_{XX}] = 1/\big((n-3)\sigma_X^2\big)$, and $\mathrm{E}\bar{X} = \mu_X$, yielding the stated formula.

Problem 11.37. We observe random variables $Y_1, \ldots, Y_n$ that are mutually independent, each with a normal distribution with variance σ^2. Furthermore, $\mathrm{E}Y_i = \beta x_i$, where β is an unknown parameter and $x_1, \ldots, x_n$ are fixed constants not all equal to 0.
(a) Find the MLE of β. Compute its mean and variance.
(b) Compute the Cramer–Rao Lower Bound for the variance of an unbiased estimator of β.
(c) Find a best unbiased estimator of β.
(d) If you could place the values $x_1, \ldots, x_n$ anywhere within a given nondegenerate closed interval $[A, B]$, where would you place these values? Justify your answer.
(e) For a given positive value r, the maximum probability estimator of β with respect to r is the value of D that maximizes the integral

$$\int_{D-r}^{D+r} f(y_1, \ldots, y_n \mid \beta)\,d\beta,$$

where $f(y_1, \ldots, y_n \mid \beta)$ is the joint pdf of $Y_1, \ldots, Y_n$. Find this estimator.

Solution. (a) $\hat{\beta} = \dfrac{\sum_i x_i Y_i}{\sum_i x_i^2}$, with $\mathrm{E}\hat{\beta} = \beta$ and $\mathrm{Var}\,\hat{\beta} = \sigma^2/\sum_i x_i^2$.
The log likelihood is $\log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (y_i - \beta x_i)^2$, and

$$\frac{\partial}{\partial\beta}\log L = \frac{1}{\sigma^2}\sum_i x_i(y_i - \beta x_i) = 0 \iff \beta = \frac{\sum_i x_i y_i}{\sum_i x_i^2},$$

a maximum since $\partial^2 \log L/\partial\beta^2 = -\sum_i x_i^2/\sigma^2 < 0$ (not all x_i are 0). Being a linear function of the Y_i,

$$\mathrm{E}\hat{\beta} = \frac{\sum_i x_i(\beta x_i)}{\sum_i x_i^2} = \beta, \qquad \mathrm{Var}\,\hat{\beta} = \frac{\sum_i x_i^2 \sigma^2}{(\sum_i x_i^2)^2} = \frac{\sigma^2}{\sum_i x_i^2}.$$

(b) CRLB $= \dfrac{\sigma^2}{\sum_i x_i^2}$. The family satisfies the regularity conditions of Theorem 7.3.9 (exponential family in β), and from (a) the second derivative $-\sum_i x_i^2/\sigma^2$ is constant, so the information number is $\mathrm{E}\big[-\partial^2 \log L/\partial\beta^2\big] = \sum_i x_i^2/\sigma^2$; the bound for an unbiased estimator is its reciprocal.
(c) $\hat\beta$ is the best unbiased estimator: it is unbiased and its variance equals the CRLB of (b), so by Theorem 7.3.9 no unbiased estimator of β has smaller variance.
(d) Put every x_i at whichever endpoint is larger in absolute value, so that $\sum_i x_i^2 = n\max(A^2, B^2)$. Since $\operatorname{Var}\hat\beta = \sigma^2/\sum_i x_i^2$ is decreasing in $\sum_i x_i^2$, we want to maximize $\sum_i x_i^2$ over $[A, B]^n$; $x \mapsto x^2$ is convex, so its maximum over $[A, B]$ occurs at an endpoint, giving $x_i^2 \le \max(A^2, B^2)$ termwise. The resulting variance is $\sigma^2/\big(n\max(A^2, B^2)\big)$.
(e) $D = \hat\beta$: the maximum probability estimator is the MLE. Completing the square in β,

$$\sum_i (y_i - \beta x_i)^2 = \Big(\sum_i x_i^2\Big)(\beta - \hat\beta)^2 + \sum_i y_i^2 - \frac{(\sum_i x_i y_i)^2}{\sum_i x_i^2},$$

so as a function of β the joint pdf is $f(y \mid \beta) = c(y)\exp\big\{-\frac{1}{2\sigma^2}(\sum_i x_i^2)(\beta - \hat\beta)^2\big\}$, a constant multiple of an $\mathrm{n}\big(\hat\beta, \sigma^2/\sum_i x_i^2\big)$ density in β. Differentiating the integral,

$$\frac{d}{dD}\int_{D-r}^{D+r} f(y \mid \beta)\, d\beta = f(y \mid D + r) - f(y \mid D - r) = 0$$

forces $(D+r-\hat\beta)^2 = (D-r-\hat\beta)^2$, i.e. $r(D-\hat\beta) = 0$, so $D = \hat\beta$. The derivative is positive for $D < \hat\beta$ and negative for $D > \hat\beta$ (the exponent is decreasing in $|\beta - \hat\beta|$), so this stationary point is the maximum.

Problem 11.38. An ecologist takes data (x_i, Y_i), $i = 1, \dots, n$, where x_i is the size of an area and Y_i is the number of moss plants in the area. We model the data by $Y_i \sim \text{Poisson}(\theta x_i)$, the Y_i being independent.
(a) Show that the least squares estimator of θ is $\sum_i x_i Y_i/\sum_i x_i^2$. Show that this estimator has variance $\theta\sum_i x_i^3/(\sum_i x_i^2)^2$. Also, compute its bias.
(b) Show that the MLE of θ is $\sum_i Y_i/\sum_i x_i$ and has variance $\theta/\sum_i x_i$. Compute its bias.
(c) Find a best unbiased estimator of θ and show that its variance attains the Cramer–Rao Lower Bound.

Solution. (a) $\hat\theta_{LS} = \sum_i x_i Y_i/\sum_i x_i^2$, unbiased, so its bias is 0. Least squares minimizes $\sum_i \big(Y_i - \mathrm{E}Y_i\big)^2 = \sum_i (Y_i - \theta x_i)^2$, and

$$\frac{d}{d\theta}\sum_i (Y_i - \theta x_i)^2 = -2\sum_i x_i(Y_i - \theta x_i) = 0 \iff \theta = \frac{\sum_i x_i Y_i}{\sum_i x_i^2},$$

the second derivative $2\sum_i x_i^2$ being positive. Since $\mathrm{E}Y_i = \operatorname{Var} Y_i = \theta x_i$,

$$\mathrm{E}\hat\theta_{LS} = \frac{\sum_i x_i(\theta x_i)}{\sum_i x_i^2} = \theta,$$
$$\operatorname{Var}\hat\theta_{LS} = \frac{\sum_i x_i^2 \operatorname{Var} Y_i}{(\sum_i x_i^2)^2} = \frac{\theta\sum_i x_i^3}{(\sum_i x_i^2)^2}.$$

(b) $\hat\theta = \sum_i Y_i/\sum_i x_i$, also unbiased, so its bias is 0. The log likelihood is

$$\log L(\theta \mid y) = -\theta\sum_i x_i + \Big(\sum_i y_i\Big)\log\theta + \sum_i \big(y_i \log x_i - \log y_i!\big),$$

so $\partial \log L/\partial\theta = -\sum_i x_i + \theta^{-1}\sum_i y_i$, which vanishes only at $\theta = \sum_i y_i/\sum_i x_i$; since $\partial^2 \log L/\partial\theta^2 = -\theta^{-2}\sum_i y_i < 0$ this is the maximum. As $\sum_i Y_i \sim \text{Poisson}(\theta\sum_i x_i)$,

$$\mathrm{E}\hat\theta = \frac{\theta\sum_i x_i}{\sum_i x_i} = \theta, \qquad \operatorname{Var}\hat\theta = \frac{\theta\sum_i x_i}{(\sum_i x_i)^2} = \frac{\theta}{\sum_i x_i}.$$

(c) $\hat\theta = \sum_i Y_i / \sum_i x_i$ is the best unbiased estimator. Writing the joint pmf as

$$f(y \mid \theta) = \Big(\prod_i \frac{x_i^{y_i}}{y_i!}\Big) \exp\Big\{ (\log\theta) \sum_i y_i - \theta \sum_i x_i \Big\},$$

the joint pmf is a full one-parameter exponential family in which the natural parameter $\log\theta$ ranges over all of $\mathbb{R}$, an open set; the argument of Theorem 6.2.25 (which uses only that the natural parameter space contains an open set, not that the sample is iid) therefore makes $T = \sum_i Y_i$ complete, and it is sufficient by the Factorization Theorem 6.2.6. Since $\hat\theta = T / \sum_i x_i$ is an unbiased function of T, Theorem 7.3.23 makes it the unique best unbiased estimator of θ.
For the bound, the family is regular (Theorem 7.3.9 applies to exponential families), and from (b)

$$\mathrm{E}\Big[-\frac{\partial^2}{\partial\theta^2}\log L\Big] = \mathrm{E}\Big[\frac{\sum_i Y_i}{\theta^2}\Big] = \frac{\theta \sum_i x_i}{\theta^2} = \frac{\sum_i x_i}{\theta},$$

so $\mathrm{CRLB} = \theta / \sum_i x_i = \operatorname{Var}\hat\theta$.

Problem 11.39. Verify that the simultaneous confidence intervals in (11.3.42) have the claimed coverage probability. That is, working under the conditional normal regression model (11.3.22), with $\hat\alpha$, $\hat\beta$ the least squares estimators, $S^2 = \sum_i (Y_i - \hat\alpha - \hat\beta x_i)^2/(n-2)$, and $x_{01}, \ldots, x_{0m}$ any m fixed values, show that the probability is at least $1-\alpha$ that

$$\begin{aligned} \hat\alpha + \hat\beta x_{0i} - t_{n-2,\alpha/(2m)}\, S\sqrt{\frac{1}{n} + \frac{(x_{0i}-\bar x)^2}{S_{xx}}} \;&< \alpha + \beta x_{0i} \\ &< \hat\alpha + \hat\beta x_{0i} + t_{n-2,\alpha/(2m)}\, S\sqrt{\frac{1}{n} + \frac{(x_{0i}-\bar x)^2}{S_{xx}}} \end{aligned}$$

simultaneously for $i = 1, \ldots, m$.

Solution. Each interval covers with probability exactly $1 - \alpha/m$, and Boole's Inequality (Theorem 1.2.11) then bounds the chance that any one of them misses by $m \cdot (\alpha/m) = \alpha$.
Fix x_0 and set $v(x_0) = \frac{1}{n} + \frac{(x_0-\bar x)^2}{S_{xx}}$. By Theorem 11.3.3 the estimator $\hat\alpha + \hat\beta x_0$ is normal (a linear function of the Y_i) with mean $\alpha + \beta x_0$ and, using $\sum_i x_i^2 = S_{xx} + n\bar x^2$,

$$\begin{aligned} \operatorname{Var}(\hat\alpha + \hat\beta x_0) &= \frac{\sigma^2 \sum_i x_i^2}{n S_{xx}} + x_0^2 \frac{\sigma^2}{S_{xx}} - 2x_0 \frac{\sigma^2 \bar x}{S_{xx}} \\ &= \sigma^2 \Big[\frac{1}{n} + \frac{\bar x^2 - 2x_0\bar x + x_0^2}{S_{xx}}\Big] = \sigma^2 v(x_0). \end{aligned}$$

Theorem 11.3.3 also gives $(n-2)S^2/\sigma^2 \sim \chi^2_{n-2}$ independent of $(\hat\alpha, \hat\beta)$, so by Definition 5.3.4

$$\begin{aligned} T(x_0) &= \frac{(\hat\alpha + \hat\beta x_0) - (\alpha + \beta x_0)}{S\sqrt{v(x_0)}} \\ &= \frac{\big[(\hat\alpha + \hat\beta x_0) - (\alpha + \beta x_0)\big] / \sigma\sqrt{v(x_0)}}{\sqrt{\big[(n-2)S^2/\sigma^2\big]/(n-2)}} \;\sim\; t_{n-2}. \end{aligned}$$

Let A_i be the event that the ith displayed interval contains $\alpha + \beta x_{0i}$, i.e. $|T(x_{0i})| < t_{n-2,\alpha/(2m)}$. By symmetry of the t_{n-2} distribution, $P(A_i^c) = 2 \cdot \alpha/(2m) = \alpha/m$ for each i. Hence

$$P\Big(\bigcap_{i=1}^m A_i\Big) = 1 - P\Big(\bigcup_{i=1}^m A_i^c\Big) \ge 1 - \sum_{i=1}^m P(A_i^c) = 1 - m \cdot \frac{\alpha}{m} = 1 - \alpha,$$

which is the claim.

Problem 11.40. (a) Prove that if a, b, c, and d are constants, with $c > 0$ and $d > 0$, then

$$\max_t \frac{(a+bt)^2}{c+dt^2} = \frac{a^2}{c} + \frac{b^2}{d}.$$

(b) Use part (a) to verify equation (11.3.44), namely

$$\max_t \frac{\big((\bar{Y}-\mu_{\bar{Y}}) + (\hat{\beta}-\beta)t\big)^2}{S^2\big[\frac{1}{n} + \frac{t^2}{S_{xx}}\big]} = \frac{n(\bar{Y}-\mu_{\bar{Y}})^2 + S_{xx}(\hat{\beta}-\beta)^2}{S^2}$$
$$= \frac{\frac{(\bar{Y}-\mu_{\bar{Y}})^2}{\sigma^2/n} + \frac{(\hat{\beta}-\beta)^2}{\sigma^2/S_{xx}}}{S^2/\sigma^2},$$

and hence fill in the gap in Theorem 11.3.6 (the Scheffe confidence band $\hat{\alpha}+\hat{\beta}x \pm M_\alpha S\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$ with $M_\alpha = \sqrt{2F_{2,n-2,\alpha}}$).
(c) Use part (a) to find a Scheffe-type simultaneous band using the prediction intervals of (11.3.41). That is, rewriting the prediction intervals as was done in Theorem 11.3.6, show that

$$\max_t \frac{\big((\bar{Y}-\mu_{\bar{Y}}) + (\hat{\beta}-\beta)t\big)^2}{S^2\big[1 + \frac{1}{n} + \frac{t^2}{S_{xx}}\big]} = \frac{\frac{n}{n+1}(\bar{Y}-\mu_{\bar{Y}})^2 + S_{xx}(\hat{\beta}-\beta)^2}{S^2}.$$

(d) The distribution of the maximum is not easy to write down, but we could approximate it. Approximate the statistic by using moment matching, as done in Example 7.2.3.
Here $t = x - \bar{x}$, $\mu_{\bar{Y}} = \mathrm{E}\bar{Y} = \alpha + \beta\bar{x}$, and the parameterization $\alpha + \beta x = \mu_{\bar{Y}} + \beta(x-\bar{x})$ of Exercise 11.32 makes $\bar{Y}$ and $\hat{\beta}$ independent.

Solution. (a) Cauchy–Schwarz. Write $a + bt = \frac{a}{\sqrt{c}}\sqrt{c} + \frac{b}{\sqrt{d}}\big(\sqrt{d}\,t\big)$; then

$$(a+bt)^2 \le \Big(\frac{a^2}{c} + \frac{b^2}{d}\Big)\big(c + dt^2\big),$$

so the ratio never exceeds $a^2/c + b^2/d$. Equality in Cauchy–Schwarz requires $(\sqrt{c}, \sqrt{d}\,t) \propto (a/\sqrt{c}, b/\sqrt{d})$, i.e. $t = bc/(ad)$, which is attained whenever $a \neq 0$. (If $a = 0$ the value $b^2/d = a^2/c + b^2/d$ is approached as $t \to \infty$, so the maximum should be read as a supremum in that one case.)
(b) Apply (a) with

$$a = \bar{Y} - \mu_{\bar{Y}}, \quad b = \hat{\beta} - \beta, \quad c = \frac{S^2}{n}, \quad d = \frac{S^2}{S_{xx}},$$

both $c, d > 0$ since $S^2 > 0$ with probability 1. Part (a) gives at once

$$\max_t \frac{\big((\bar{Y}-\mu_{\bar{Y}}) + (\hat{\beta}-\beta)t\big)^2}{S^2\big[\frac{1}{n} + \frac{t^2}{S_{xx}}\big]} = \frac{n(\bar{Y}-\mu_{\bar{Y}})^2}{S^2} + \frac{S_{xx}(\hat{\beta}-\beta)^2}{S^2},$$

which is the first line of (11.3.44); dividing numerator and denominator by σ^2 gives the second. Now $\bar{Y} \sim \mathrm{n}(\mu_{\bar{Y}}, \sigma^2/n)$ and $\hat{\beta} \sim \mathrm{n}(\beta, \sigma^2/S_{xx})$ are independent (Theorem 11.3.3 and Exercise 11.32), so the two numerator terms are independent χ^2_1 variables and their sum is χ^2_2; by Theorem 11.3.3 the denominator is $S^2/\sigma^2 = \big[(n-2)S^2/\sigma^2\big]/(n-2)$, a χ^2_{n-2} divided by its degrees of freedom and independent of $(\bar{Y}, \hat{\beta})$. Hence

$$\max_t \frac{\big((\bar{Y}-\mu_{\bar{Y}}) + (\hat{\beta}-\beta)t\big)^2}{S^2\big[\frac{1}{n} + \frac{t^2}{S_{xx}}\big]} = 2 \cdot \frac{\chi^2_2/2}{\chi^2_{n-2}/(n-2)} \sim 2F_{2,n-2},$$

so $P(\max_t \cdot \le M_\alpha^2) = 1 - \alpha$ exactly when $M_\alpha = \sqrt{2F_{2,n-2,\alpha}}$, completing the proof of Theorem 11.3.6.

(c) The prediction interval (11.3.41) has the same centre but the wider standard error $S\sqrt{1+\frac{1}{n}+\frac{t^2}{S_{xx}}}$, so apply (a) with the same a, b and

$$c = S^2\Big(1+\frac{1}{n}\Big) = S^2\,\frac{n+1}{n}, \qquad d = \frac{S^2}{S_{xx}},$$

giving

$$\max_t \frac{\big((\bar Y-\mu_{\bar Y})+(\hat\beta-\beta)t\big)^2}{S^2\big[1+\frac{1}{n}+\frac{t^2}{S_{xx}}\big]} = \frac{\frac{n}{n+1}(\bar Y-\mu_{\bar Y})^2 + S_{xx}(\hat\beta-\beta)^2}{S^2},$$

as claimed. Writing W for this maximum and dividing through by σ^2,

$$W = \frac{\frac{1}{n+1}\cdot\frac{(\bar Y-\mu_{\bar Y})^2}{\sigma^2/n} + \frac{(\hat\beta-\beta)^2}{\sigma^2/S_{xx}}}{S^2/\sigma^2} \;\sim\; \frac{\frac{1}{n+1}W_1 + W_2}{\chi^2_{n-2}/(n-2)},$$

with W_1, W_2 independent χ^2_1 variables independent of the χ^2_{n-2}. The band is

$$\hat\alpha+\hat\beta x \pm M_\alpha S\sqrt{1+\frac{1}{n}+\frac{(x-\bar x)^2}{S_{xx}}}, \qquad \text{all } x,$$

with M_α^2 the $1-\alpha$ quantile of W; the unequal weights keep the numerator from being chi squared, so M_α has no closed form.

(d) Moment matching (Example 7.2.3) on the numerator $U = \frac{1}{n+1}W_1 + W_2$: match U to $c\chi^2_\nu$ by equating the first two moments. Since $\mathrm{E}\chi^2_1 = 1$ and $\operatorname{Var}\chi^2_1 = 2$,

$$\mathrm{E}U = \frac{n+2}{n+1}, \qquad \operatorname{Var}U = \frac{2}{(n+1)^2} + 2 = \frac{2\big[(n+1)^2+1\big]}{(n+1)^2}.$$

Solving $c\nu = \mathrm{E}U$ and $2c^2\nu = \operatorname{Var}U$ gives

$$\nu = \frac{2(\mathrm{E}U)^2}{\operatorname{Var}U} = \frac{(n+2)^2}{(n+1)^2+1}, \qquad c = \frac{\operatorname{Var}U}{2\,\mathrm{E}U} = \frac{(n+1)^2+1}{(n+1)(n+2)}.$$

Therefore $U \approx c\chi^2_\nu$, and since $c\nu = \mathrm{E}U = (n+2)/(n+1)$,

$$W = \frac{U}{\chi^2_{n-2}/(n-2)} \approx c\nu\,\frac{\chi^2_\nu/\nu}{\chi^2_{n-2}/(n-2)} \sim \frac{n+2}{n+1}\,F_{\nu,\,n-2},$$

so the band of (c) may be used with

$$M_\alpha \approx \sqrt{\frac{n+2}{n+1}\,F_{\nu,\,n-2,\,\alpha}}, \qquad \nu = \frac{(n+2)^2}{(n+1)^2+1}.$$

Problem 11.41. In the discussion in Example 12.4.2, note that there was one observation from the potoroo data that had a missing value. Suppose that on the 24th animal it was observed that $O_2 = 16.3$.

Example 12.4.2 describes an experiment (McPherson 1990) in which the levels of carbon dioxide (CO_2) and oxygen (O_2) were measured in the pouches of 24 potoroos; interest is in the regression of CO_2 on O_2, so $x_i = \%O_2$ and $Y_i = \%CO_2$, with $Y_i \sim \mathrm{n}(\alpha+\beta x_i, \sigma^2)$ independent. The data for the 23 complete animals (Table 12.4.1) are

Animal	% O2	% CO2
1	20.0	1.0
2	19.6	1.2
3	19.6	1.1
4	19.4	1.4
5	18.4	2.3
6	19.0	1.7
7	19.0	1.7
8	18.3	2.4
9	18.2	2.1
10	18.6	2.1
11	19.2	1.2
12	18.2	2.3
13	18.7	1.9
14	18.5	2.4
15	18.0	2.6
16	17.4	2.9
17	16.5	4.0
18	17.2	3.3
19	17.3	3.0
20	17.8	3.4
21	17.3	2.9
22	18.4	1.9
23	16.9	3.9

(a) Write down the observed data and expected complete data log likelihood functions.
(b) Describe the E step and the M step of an EM algorithm to find the MLEs.
(c) Find the MLEs using all 24 observations.
(d) Actually, the O_2 reading on the 24th animal was not observed, but rather the CO_2 was observed to be 4.2 (and the O_2 was missing). Set up the EM algorithm in this case and find the MLEs. (This is a much harder problem, as you now have to take expectations over the xs. This means you have to formulate the regression problem using the bivariate normal distribution.)

Solution. (a) In the conditional model the value $x_{24} = 16.3$ alone carries no information about $(\alpha, \beta, \sigma^2)$, so the observed data log likelihood is just the 23-point one:

$$L_{\text{obs}}(\alpha, \beta, \sigma^2) = -\frac{23}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{23}(y_i - \alpha - \beta x_i)^2.$$

The complete data log likelihood is the same sum run to $i = 24$. Writing $\theta' = (\alpha', \beta', \sigma'^2)$ for the current parameter value and $\hat{y}'_{24} = \alpha' + \beta' x_{24}$, the missing Y_{24} has distribution $\mathrm{n}(\hat{y}'_{24}, \sigma'^2)$, so

$$\mathrm{E}_{\theta'}\left[(Y_{24} - \alpha - \beta x_{24})^2\right] = (\hat{y}'_{24} - \alpha - \beta x_{24})^2 + \sigma'^2,$$

and the expected complete data log likelihood is

$$\begin{aligned} Q(\theta \mid \theta') = &-\frac{24}{2}\log(2\pi\sigma^2) \\ &- \frac{1}{2\sigma^2}\Big[\sum_{i=1}^{23}(y_i - \alpha - \beta x_i)^2 + (\hat{y}'_{24} - \alpha - \beta x_{24})^2 + \sigma'^2\Big]. \end{aligned}$$

(b) E step: at iteration r impute $\hat{y}_{24}^{(r)} = \alpha^{(r)} + \beta^{(r)}(16.3)$ for the missing response and record the extra variance term $\sigma^{2(r)}$, i.e. form $Q(\theta \mid \theta^{(r)})$ above. M step: maximize Q. In (α, β) this is ordinary least squares on the 24 pairs $(x_1, y_1), \ldots, (x_{23}, y_{23}), (16.3, \hat{y}_{24}^{(r)})$, so with a bar and S denoting means and

sums of squares over those 24 points,

$$\beta^{(r+1)} = \frac{S_{xy}^{(24)}}{S_{xx}^{(24)}}, \qquad \alpha^{(r+1)} = \bar{y}^{(24)} - \beta^{(r+1)}\bar{x}^{(24)},$$

and then

$$\sigma^{2(r+1)} = \frac{1}{24}\Big[\mathrm{RSS}_{24}^{(r+1)} + \sigma^{2(r)}\Big],$$

where $\mathrm{RSS}_{24}^{(r+1)}$ is the residual sum of squares of that 24-point fit.
(c) $\hat{\alpha} = 18.675$, $\hat{\beta} = -0.894$, $\hat{\sigma}^2 = 0.0503$ — exactly the 23-observation MLEs; the extra animal changes nothing. At a fixed point the imputed value is $\hat{y}_{24} = \hat{\alpha} + \hat{\beta}x_{24}$, so the 24th term of each normal equation,

$$\sum_{i=1}^{24}(y_i - \alpha - \beta x_i) = 0, \qquad \sum_{i=1}^{24} x_i(y_i - \alpha - \beta x_i) = 0,$$

vanishes at $(\alpha, \beta) = (\hat{\alpha}, \hat{\beta})$; the equations collapse to the 23-point normal equations, whose solution is

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{-17.1148}{19.1443} = -0.8940, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 18.6746,$$

using $\bar{x} = 18.3261$, $\bar{y} = 2.2913$, $S_{xx} = 19.1443$, $S_{xy} = -17.1148$, so the imputed value is $\hat{y}_{24} = 18.6746 - 0.8940(16.3) = 4.103$. Its residual contribution being $0 + \hat{\sigma}^2$, the variance equation reads $24\hat{\sigma}^2 = \mathrm{RSS}_{23} + \hat{\sigma}^2$, so

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}_{23}}{23} = \frac{S_{yy} - \hat{\beta}S_{xy}}{23} = \frac{1.15788}{23} = 0.05034,$$

which is the maximizer of L_{obs} in (a), as EM must deliver.
(d) Now model (X_i, Y_i), $i = 1, \ldots, 24$, as iid bivariate normal with parameters $\theta = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY})$, the complete data log likelihood being (Definition 4.5.10)

$$-24\log(2\pi) - 12\log\big(\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2\big) - \tfrac{1}{2}\sum_{i=1}^{24}(z_i - \mu)^{\mathsf{T}}\Sigma^{-1}(z_i - \mu),$$

$z_i = (x_i, y_i)^{\mathsf{T}}$; it depends on the data only through $T = \big(\sum x_i, \sum y_i, \sum x_i^2, \sum y_i^2, \sum x_i y_i\big)$.
E step: with X_{24} missing and $y_{24} = 4.2$ observed, the conditional distributions listed after Definition 4.5.10 give

$$X_{24} \mid Y_{24} = 4.2 \;\sim\; \mathrm{n}\Big(\mu_X' + \frac{\sigma_{XY}'}{{\sigma_Y'}^2}(4.2 - \mu_Y'),\; {\sigma_X'}^2 - \frac{{\sigma_{XY}'}^2}{{\sigma_Y'}^2}\Big) \equiv \mathrm{n}(m', v'),$$

so replace the missing entries of T by $\mathrm{E}X_{24} = m'$, $\mathrm{E}X_{24}^2 = m'^2 + v'$, and $\mathrm{E}[X_{24}y_{24}] = 4.2\,m'$. M step: the complete data MLEs are the sample moments, so set $\mu_X = \sum x_i/24$, $\mu_Y = \sum y_i/24$, $\sigma_X^2 = \sum x_i^2/24 - \mu_X^2$, and likewise for σ_Y^2 and σ_{XY}, evaluated at the completed T.
The limit can be identified exactly, since the missing pattern is monotone: factor the bivariate density as $f_Y(y)f_{X|Y}(x \mid y)$, writing $X \mid Y = y \sim \mathrm{n}(\gamma_0 + \gamma_1 y, \tau^2)$. The observed data likelihood is then

$$\Big[\prod_{i=1}^{24} f_Y(y_i)\Big]\Big[\prod_{i=1}^{23} f_{X|Y}(x_i \mid y_i)\Big],$$

a product of two pieces with no shared parameters. Maximizing the first over (μ_Y, σ_Y^2) uses all 24 responses, and maximizing the second over $(\gamma_0, \gamma_1, \tau^2)$ is the least squares regression of x on y over the 23 complete pairs:

$$\hat{\mu}_Y = \frac{56.9}{24} = 2.3708, \quad \hat{\sigma}_Y^2 = 0.8312,$$
$$\hat{\gamma}_1 = -1.0399, \qquad \hat{\gamma}_0 = 20.7088, \qquad \hat{\tau}^2 = 0.05856.$$

Transforming back by $\mu_X = \gamma_0 + \gamma_1 \mu_Y$, $\sigma_{XY} = \gamma_1 \sigma_Y^2$, $\sigma_X^2 = \tau^2 + \gamma_1^2 \sigma_Y^2$,

$$\hat{\mu}_X = 18.2434, \quad \hat{\sigma}_X^2 = 0.9574, \quad \hat{\sigma}_{XY} = -0.8644, \quad \hat{\rho} = -0.9689,$$

and the EM iteration above indeed converges to these values. The implied regression of CO_2 on O_2 is

$$\hat{\beta} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2} = -0.9028, \qquad \hat{\alpha} = \hat{\mu}_Y - \hat{\beta}\hat{\mu}_X = 18.8414,$$

with residual variance $\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}^2/\hat{\sigma}_X^2 = 0.0508$; the imputed oxygen reading is $\hat{\gamma}_0 + \hat{\gamma}_1(4.2) = 16.341$.

12 Regression Models

Contents

12.1 Exercises 12.1–12.7 464
12.2 Exercises 12.8–12.14 470
12.3 Exercises 12.15–12.21 477
12.4 Exercises 12.22–12.28 484
12.5 Exercises 12.29–12.31 490

12.1 Exercises 12.1–12.7

Problem 12.1. Verify the expressions in (12.2.7). *(Hint: Use the Pythagorean Theorem.)*
That is, for a data point (x', y') and a line $y = a + bx$, show that the point $(\hat{x}', \hat{y}')$ on the line that is closest to (x', y') when distance is measured orthogonally (perpendicular to the line, as in Figure 12.2.1) is

$$\hat{x}' = \frac{by' + x' - ab}{1 + b^2}, \qquad \hat{y}' = a + \frac{b}{1 + b^2}\big(by' + x' - ab\big).$$

Solution. The foot of the perpendicular is the unique point of $y = a + bx$ whose displacement from (x', y') is orthogonal to the direction vector $(1, b)$ of the line; solving the two linear conditions

$$\begin{aligned} \hat{y}' &= a + b\,\hat{x}' && \text{(on the line)},\\ (x' - \hat{x}') + b\,(y' - \hat{y}') &= 0 && \text{(perpendicular)} \end{aligned}$$

gives, on substituting the first into the second,

$$(x' - \hat{x}') + b\,(y' - a - b\hat{x}') = 0 \iff (1 + b^2)\,\hat{x}' = x' + by' - ab,$$

which is the stated $\hat{x}'$, and then $\hat{y}' = a + b\hat{x}'$ is the stated $\hat{y}'$.
That this stationary point is the *minimizing* point is the Pythagorean Theorem: for any other point P on the line, the triangle with vertices (x', y'), $(\hat{x}', \hat{y}')$, P has its right angle at $(\hat{x}', \hat{y}')$, so

$$\big\|(x', y') - P\big\|^2 = \big\|(x', y') - (\hat{x}', \hat{y}')\big\|^2 + \big\|(\hat{x}', \hat{y}') - P\big\|^2 > \big\|(x', y') - (\hat{x}', \hat{y}')\big\|^2.$$

Problem 12.2. Show that the extrema of

$$f(b) = \frac{1}{1 + b^2}\big[S_{yy} - 2bS_{xy} + b^2 S_{xx}\big]$$

are given by

$$b = \frac{-(S_{xx} - S_{yy}) \pm \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}}{2S_{xy}}.$$

Show that the "+" solution gives the minimum of $f(b)$.

Solution. Differentiating $f = N/D$ with $N = S_{yy} - 2bS_{xy} + b^2 S_{xx}$ and $D = 1 + b^2$,

$$\begin{aligned} D^2 f'(b) &= (2bS_{xx} - 2S_{xy})(1 + b^2) - 2b\big(S_{yy} - 2bS_{xy} + b^2 S_{xx}\big)\\ &= 2\big[\, S_{xy}b^2 + (S_{xx} - S_{yy})b - S_{xy}\,\big], \end{aligned}$$

so (assuming $S_{xy} \neq 0$, else f is monotone on each half-line and the formula is vacuous) $f'(b) = 0$ exactly at the two roots of $S_{xy}b^2 + (S_{xx} - S_{yy})b - S_{xy} = 0$, namely

$$b = \frac{-(S_{xx} - S_{yy}) \pm \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}}{2S_{xy}},$$

the discriminant being strictly positive, so the roots are real and distinct.
To identify the minimum, determine the range of f directly. For $c \in \mathbb{R}$, the equation $f(b) = c$ is

$$(S_{xx} - c)b^2 - 2S_{xy}b + (S_{yy} - c) = 0,$$

which has a real root if and only if $S_{xy}^2 - (S_{xx} - c)(S_{yy} - c) \geq 0$, i.e.

$$c^2 - (S_{xx} + S_{yy})c + \big(S_{xx}S_{yy} - S_{xy}^2\big) \leq 0 \iff c_- \leq c \leq c_+,$$

where, writing $R = \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}$,

$$c_\pm = \frac{(S_{xx} + S_{yy}) \pm R}{2}.$$

Hence $\min_b f = c_-$, attained at the (double) root $b = S_{xy}/(S_{xx} - c_-)$ of the displayed quadratic. Since $S_{xx} - c_- = \big[(S_{xx} - S_{yy}) + R\big]/2$,

$$b_{\min} = \frac{2S_{xy}}{(S_{xx} - S_{yy}) + R} = \frac{2S_{xy}\big[R - (S_{xx} - S_{yy})\big]}{R^2 - (S_{xx} - S_{yy})^2} = \frac{-(S_{xx} - S_{yy}) + R}{2S_{xy}},$$

using $R^2 - (S_{xx} - S_{yy})^2 = 4S_{xy}^2$. This is the "+" root, and it is therefore the minimum; the "−" root is the maximum, with value c_+.

Problem 12.3. In maximizing the likelihood (12.2.13), we first minimized, for each value of α, β, and σ_δ^2, the function

$$f(\xi_1, \dots, \xi_n) = \sum_{i=1}^{n} \Big((x_i - \xi_i)^2 + \lambda \big(y_i - (\alpha + \beta\xi_i)\big)^2 \Big)$$

with respect to $\xi_1, \dots, \xi_n$.
(a) Prove that this function is minimized at

$$\xi_i^* = \frac{x_i + \lambda\beta(y_i - \alpha)}{1 + \lambda\beta^2}.$$

(b) Show that the function

$$D_\lambda\big((x, y), (\xi, \alpha + \beta\xi)\big) = (x - \xi)^2 + \lambda\big(y - (\alpha + \beta\xi)\big)^2$$

defines a *metric* between the points (x, y) and $(\xi, \alpha + \beta\xi)$. A metric is a distance measure, a function D that measures the distance between two points A and B. A metric satisfies the following four properties:
i. $D(A, A) = 0$.
ii. $D(A, B) > 0$ if $A \neq B$.
iii. $D(A, B) = D(B, A)$ (reflexive).
iv. $D(A, B) \leq D(A, C) + D(C, B)$ (triangle inequality).

Solution. (a) f separates: $f = \sum_i f_i(\xi_i)$ with $f_i(t) = (x_i - t)^2 + \lambda(y_i - \alpha - \beta t)^2$, a quadratic in t with leading coefficient $1 + \lambda\beta^2 > 0$, hence strictly convex with a unique global minimum at its stationary point:

$$\tfrac{1}{2} f_i'(t) = -(x_i - t) - \lambda\beta\,(y_i - \alpha - \beta t) = 0 \iff (1 + \lambda\beta^2)\, t = x_i + \lambda\beta(y_i - \alpha),$$

so $\xi_i^* = \big[x_i + \lambda\beta(y_i - \alpha)\big]/(1 + \lambda\beta^2)$, and $f_i'' \equiv 2(1 + \lambda\beta^2) > 0$. Minimizing each summand separately minimizes the sum.
(b) As printed, (iv) fails for the *squared* distance D_λ (take $\lambda = 1$, $A = (0, 0)$, $C = (1, 0)$, $B = (2, 0)$: $D_\lambda(A, B) = 4 > 1 + 1$), so the intended claim is that $d_\lambda = \sqrt{D_\lambda}$ is a metric; (i)–(iii) hold verbatim for D_λ as well.

Write a generic point of the plane as $A = (a_1, a_2)$ and let $T = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{\lambda} \end{pmatrix}$, which is invertible because $\lambda > 0$. Then

$$d_\lambda(A, B) = \sqrt{(a_1 - b_1)^2 + \lambda(a_2 - b_2)^2} = \big\|T(A - B)\big\|,$$

the ordinary Euclidean norm of $T(A - B)$. The four properties follow in one clause each:
i. $d_\lambda(A, A) = \|T\mathbf{0}\| = 0$.
ii. If $A \neq B$ then $T(A - B) \neq \mathbf{0}$ (T is invertible), so $d_\lambda(A, B) > 0$.

iii. $\|T(A-B)\| = \|-T(B-A)\| = \|T(B-A)\|$.
iv. By linearity of T and the triangle inequality for $\|\cdot\|$,

$$\|T(A-B)\| = \|T(A-C)+T(C-B)\| \le \|T(A-C)\| + \|T(C-B)\|.$$

Specializing $A=(x,y)$ and $B=(\xi,\alpha+\beta\xi)$ gives the assertion.

Problem 12.4. Consider the MLE of the slope in the EIV model

$$\hat\beta(\lambda) = \frac{-(S_{xx}-\lambda S_{yy}) + \sqrt{(S_{xx}-\lambda S_{yy})^2 + 4\lambda S_{xy}^2}}{2\lambda S_{xy}},$$

where $\lambda = \sigma_\delta^2/\sigma_\epsilon^2$ is assumed known.
(a) Show that $\lim_{\lambda\to 0}\hat\beta(\lambda) = S_{xy}/S_{xx}$, the slope of the ordinary regression of y on x.
(b) Show that $\lim_{\lambda\to\infty}\hat\beta(\lambda) = S_{yy}/S_{xy}$, the reciprocal of the slope of the ordinary regression of x on y.
(c) Show that $\hat\beta(\lambda)$ is, in fact, monotone in λ and is increasing if $S_{xy}>0$ and decreasing if $S_{xy}<0$.
(d) Show that the orthogonal least squares line ($\lambda=1$) is always between the lines given by the ordinary regressions of y on x and of x on y.
(e) The following data were collected in a study to examine the relationship between brain weight and body weight in a number of animal species.

Species	Body weight (kg) (x)	Brain weight (g) (y)
Arctic fox	3.385	44.50
Owl monkey	.480	15.50
Mountain beaver	1.350	8.10
Guinea pig	1.040	5.50
Chinchilla	.425	6.40
Ground squirrel	.101	4.00
Tree hyrax	2.000	12.30
Big brown bat	.023	.30

Calculate the MLE of the slope assuming the EIV model. Also, calculate the least squares slopes of the regressions of y on x and of x on y, and show how these quantities bound the MLE.

Solution. Everything follows from the rationalized form of $\hat\beta(\lambda)$. Writing $u(\lambda) = S_{xx}-\lambda S_{yy}$ and $R(\lambda) = \sqrt{u(\lambda)^2+4\lambda S_{xy}^2}$, multiplying numerator and denominator by $R+u$ and using $R^2-u^2 = 4\lambda S_{xy}^2$ gives

$$\hat\beta(\lambda) = \frac{R-u}{2\lambda S_{xy}} = \frac{2S_{xy}}{R(\lambda)+u(\lambda)} \equiv \frac{2S_{xy}}{g(\lambda)}, \qquad g(\lambda) := R(\lambda)+u(\lambda). \qquad (*)$$

Note $g>0$ for $\lambda>0$ (as $R>|u|$ when $S_{xy}\neq 0$), so $(*)$ never divides by zero. Throughout, $S_{xx}>0$, $S_{yy}>0$ and $S_{xy}\neq 0$.
(a) Let $\lambda\to 0$ in $(*)$: $u\to S_{xx}$ and $R\to |S_{xx}| = S_{xx}$, so

$$\lim_{\lambda\to 0}\hat\beta(\lambda) = \frac{2S_{xy}}{2S_{xx}} = \frac{S_{xy}}{S_{xx}},$$

the ordinary least squares slope of y on x from (11.3.7).
(b) Divide numerator and denominator of the original expression by λ:

$$\hat\beta(\lambda) = \frac{-\left(\frac{S_{xx}}{\lambda}-S_{yy}\right) + \sqrt{\left(\frac{S_{xx}}{\lambda}-S_{yy}\right)^2 + \frac{4S_{xy}^2}{\lambda}}}{2S_{xy}}$$
$$\xrightarrow[\lambda\to\infty]{} \frac{S_{yy}+|S_{yy}|}{2S_{xy}} = \frac{S_{yy}}{S_{xy}},$$

the reciprocal of the slope S_{xy}/S_{yy} of the ordinary regression of x on y.

(c) By $(*)$ it suffices to show g is strictly decreasing. With $u' = -S_{yy}$,

$$g'(\lambda) = \frac{2uu' + 4S_{xy}^2}{2R} + u' = \frac{2S_{xy}^2 - S_{yy}(u+R)}{R},$$

so $g' < 0$ iff $S_{yy}(u+R) > 2S_{xy}^2$. If $2S_{xy}^2 - S_{yy}u \le 0$ this is immediate since $S_{yy}R > 0$; otherwise both sides are positive and we may square:

$$\begin{aligned} S_{yy}^2R^2 - \big(2S_{xy}^2 - S_{yy}u\big)^2 &= S_{yy}^2\big(u^2 + 4\lambda S_{xy}^2\big) - 4S_{xy}^4 + 4S_{xy}^2S_{yy}u - S_{yy}^2u^2 \\ &= 4S_{xy}^2\big[\lambda S_{yy}^2 - S_{xy}^2 + S_{yy}u\big] \\ &= 4S_{xy}^2\big[S_{xx}S_{yy} - S_{xy}^2\big] > 0, \end{aligned}$$

the last step using $S_{yy}u = S_{xx}S_{yy} - \lambda S_{yy}^2$ and, for positivity, the Cauchy–Schwarz inequality $S_{xy}^2 \le S_{xx}S_{yy}$, strict unless the points are exactly collinear. Hence $g' < 0$, and by $(*)$ $\hat\beta$ is strictly increasing in λ when $S_{xy} > 0$ and strictly decreasing when $S_{xy} < 0$.

(d) By (c) the map $\lambda \mapsto \hat\beta(\lambda)$ is strictly monotone on $(0,\infty)$, and by (a) and (b) its two endpoint limits are S_{xy}/S_{xx} and S_{yy}/S_{xy}. A continuous strictly monotone function maps $(0,\infty)$ onto the open interval determined by its endpoint limits, so for *every* $\lambda \in (0,\infty)$, in particular $\lambda = 1$,

$$\hat\beta(\lambda) \text{ lies strictly between } \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \frac{S_{yy}}{S_{xy}}.$$

Since all three lines pass through $(\bar x, \bar y)$ (each has intercept $\bar y - b\bar x$), the orthogonal least squares line lies between the two ordinary regression lines.

(e) With $n = 8$,

$$\begin{aligned} &\sum x_i = 8.804, \quad \sum y_i = 96.6, \\ &\sum x_i^2 = 18.78408, \quad \sum y_i^2 = 2524.70, \quad \sum x_iy_i = 202.4584, \end{aligned}$$

so $\bar x = 1.1005$, $\bar y = 12.075$ and

$$\begin{aligned} S_{xx} &= 18.78408 - 8(1.1005)^2 = 9.09528, \\ S_{yy} &= 2524.70 - 8(12.075)^2 = 1358.255, \\ S_{xy} &= 202.4584 - 8(1.1005)(12.075) = 96.1501. \end{aligned}$$

The two ordinary regression slopes are

$$\frac{S_{xy}}{S_{xx}} = \frac{96.1501}{9.09528} = 10.5714 \qquad \text{and} \qquad \frac{S_{yy}}{S_{xy}} = \frac{1358.255}{96.1501} = 14.1264.$$

Taking $\lambda = 1$ (equal error variances, so the MLE is the orthogonal least squares line), $S_{xx} - S_{yy} = -1349.1597$ and

$$\sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2} = \sqrt{1857211.3} = 1362.7954,$$

giving

$$\begin{aligned} \hat\beta(1) &= \frac{1349.1597 + 1362.7954}{2(96.1501)} = \frac{2711.9551}{192.3002} = 14.1027, \\ \hat\alpha &= \bar y - \hat\beta\bar x = -3.4450. \end{aligned}$$

As part (d) predicts,

$$\frac{S_{xy}}{S_{xx}} = 10.5714 \;<\; \hat\beta(1) = 14.1027 \;<\; \frac{S_{yy}}{S_{xy}} = 14.1264.$$

Problem 12.5. In the EIV functional relationship model, where $\lambda = \sigma_\delta^2/\sigma_\epsilon^2$ is assumed known, show that the MLE of σ_δ^2 is given by (12.2.18), that is,

$$\hat\sigma_\delta^2 = \frac{1}{2n}\,\frac{\lambda}{1+\lambda\hat\beta^2}\sum_{i=1}^{n}\Big(y_i - (\hat\alpha+\hat\beta x_i)\Big)^2,$$

where $\hat\alpha$ and $\hat\beta$ are the MLEs of (12.2.16) and the partially maximized likelihood is, by (12.2.17),

$$\begin{aligned}&\max_{\alpha,\beta,\xi_1,\dots,\xi_n} L\big(\alpha,\beta,\xi_1,\dots,\xi_n,\sigma_\delta^2 \mid \mathbf{x},\mathbf{y}\big)\\ &\qquad = \frac{1}{(2\pi)^n}\frac{\lambda^{n/2}}{(\sigma_\delta^2)^n}\exp\left[-\frac{1}{2\sigma_\delta^2}\,\frac{\lambda}{1+\lambda\hat\beta^2}\sum_{i=1}^{n}\Big(y_i-(\hat\alpha+\hat\beta x_i)\Big)^2\right].\end{aligned}$$

Solution. Write $t=\sigma_\delta^2$ and

$$K = \frac{\lambda}{1+\lambda\hat\beta^2}\sum_{i=1}^{n}\big(y_i-(\hat\alpha+\hat\beta x_i)\big)^2,$$

a constant free of t because $\hat\alpha,\hat\beta$ of (12.2.16) do not involve σ_δ^2 (this is what licenses the staged maximization). From (12.2.17),

$$\log L^*(t) = -n\log t - \frac{K}{2t} + \text{const}, \qquad \frac{d}{dt}\log L^*(t) = -\frac{n}{t}+\frac{K}{2t^2} = \frac{K-2nt}{2t^2}.$$

The derivative is positive for $t < K/(2n)$ and negative for $t > K/(2n)$, so the unique global maximum over $t>0$ is at

$$\hat\sigma_\delta^2 = \frac{K}{2n} = \frac{1}{2n}\,\frac{\lambda}{1+\lambda\hat\beta^2}\sum_{i=1}^{n}\Big(y_i-(\hat\alpha+\hat\beta x_i)\Big)^2,$$

which is (12.2.18).

Problem 12.6. Show that in the linear structural relationship model (12.2.6), if we integrate out ξ_i, the marginal distribution of (X_i,Y_i) is given by (12.2.19).
Here (12.2.6) is the hierarchy

$$\begin{aligned}Y_i &= \alpha+\beta\xi_i+\epsilon_i, \quad \epsilon_i\sim \mathrm{n}(0,\sigma_\epsilon^2),\\ X_i &= \xi_i+\delta_i, \qquad\qquad \delta_i\sim \mathrm{n}(0,\sigma_\delta^2),\end{aligned}$$

with $\xi_i\sim$ iid $\mathrm{n}(\xi,\sigma_\xi^2)$, the ϵ_is and δ_is independent of each other and of the ξ_is; and (12.2.19) asserts

$$(X_i,Y_i)\sim \text{bivariate normal}\big(\xi,\ \alpha+\beta\xi,\ \sigma_\delta^2+\sigma_\xi^2,\ \sigma_\epsilon^2+\beta^2\sigma_\xi^2,\ \beta\sigma_\xi^2\big),$$

the last three entries being $\operatorname{Var}X_i$, $\operatorname{Var}Y_i$, and $\operatorname{Cov}(X_i,Y_i)$.

Solution. Integrating out ξ_i is done cleanly on the mgf. Condition on ξ_i and use the independence of $\delta_i,\epsilon_i,\xi_i$ together with the normal mgf $\mathrm{E}e^{sZ}=e^{\mu s+\sigma^2s^2/2}$ for $Z\sim\mathrm{n}(\mu,\sigma^2)$ (Section 3.3):

$$\begin{aligned}M_{X_i,Y_i}(s,t) &= \mathrm{E}\,\exp\big[s(\xi_i+\delta_i)+t(\alpha+\beta\xi_i+\epsilon_i)\big]\\ &= e^{t\alpha}\,\mathrm{E}e^{s\delta_i}\,\mathrm{E}e^{t\epsilon_i}\,\mathrm{E}\,e^{(s+t\beta)\xi_i}\\ &= e^{t\alpha}\,\exp\Big[\tfrac12 s^2\sigma_\delta^2+\tfrac12 t^2\sigma_\epsilon^2\Big]\\ &\qquad\times\exp\Big[(s+t\beta)\xi+\tfrac12(s+t\beta)^2\sigma_\xi^2\Big].\end{aligned}$$

Collecting the exponent by powers of s and t,

$$\log M_{X_i,Y_i}(s,t) = s\,\xi + t\,(\alpha+\beta\xi) + \tfrac{1}{2}\Big[s^2\big(\sigma_\delta^2+\sigma_\xi^2\big) + 2st\,\beta\sigma_\xi^2 + t^2\big(\sigma_\epsilon^2+\beta^2\sigma_\xi^2\big)\Big],$$

which is precisely the mgf of the bivariate normal of Definition 4.5.10 with

$$\mu_X = \xi, \quad \mu_Y = \alpha+\beta\xi, \quad \sigma_X^2 = \sigma_\delta^2+\sigma_\xi^2, \quad \sigma_Y^2 = \sigma_\epsilon^2+\beta^2\sigma_\xi^2, \quad \rho\sigma_X\sigma_Y = \beta\sigma_\xi^2.$$

Since the mgf is finite for all (s,t) in a neighborhood of the origin, it determines the joint distribution (Theorem 2.3.11 and its bivariate version, Section 4.6), so (X_i, Y_i) has the distribution claimed in (12.2.19), with correlation

$$\rho = \frac{\beta\sigma_\xi^2}{\sqrt{\big(\sigma_\delta^2+\sigma_\xi^2\big)\big(\sigma_\epsilon^2+\beta^2\sigma_\xi^2\big)}}.$$

Problem 12.7. Consider a linear structural relationship model where we assume that ξ_i has an improper distribution, $\xi_i \sim \text{uniform}(-\infty,\infty)$.
(a) Show that for each i,

$$\begin{aligned}\int_{-\infty}^{\infty} &\frac{1}{2\pi}\frac{1}{\sigma_\delta\sigma_\epsilon}\exp\left[-\frac{(x_i-\xi_i)^2}{2\sigma_\delta^2}\right] \\ &\qquad\times\exp\left[-\frac{\big(y_i-(\alpha+\beta\xi_i)\big)^2}{2\sigma_\epsilon^2}\right]d\xi_i \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\beta^2\sigma_\delta^2+\sigma_\epsilon^2}}\exp\left[-\frac{1}{2}\frac{\big(y_i-(\alpha+\beta x_i)\big)^2}{\beta^2\sigma_\delta^2+\sigma_\epsilon^2}\right].\end{aligned}$$

(Completing the square in the exponential makes the integration easy.)
(b) The result of the integration in part (a) looks like a pdf, and if we consider it a pdf of Y conditional on X, then we seem to have a linear relationship between X and Y. Thus, it is sometimes said that this "limiting case" of the structural relationship leads to simple linear regression and ordinary least squares. Explain why this interpretation of the above function is wrong.

Solution. (a) Split the constant as $\frac{1}{2\pi\sigma_\delta\sigma_\epsilon} = \frac{1}{\sqrt{2\pi}\sigma_\delta}\cdot\frac{1}{\sqrt{2\pi}\sigma_\epsilon}$ and read the integrand as a mixture: the left factor is the $\text{n}(x_i,\sigma_\delta^2)$ pdf in the variable ξ_i, and the right factor is the $\text{n}(\alpha+\beta\xi_i,\sigma_\epsilon^2)$ pdf evaluated at y_i. Hence the integral is the pdf, evaluated at y_i, of

$$W = \alpha+\beta U+\epsilon, \qquad U\sim\text{n}(x_i,\sigma_\delta^2), \quad \epsilon\sim\text{n}(0,\sigma_\epsilon^2)\text{ independent},$$

and by Corollary 4.6.10 (a linear combination of independent normals is normal)

$$W\sim\text{n}\big(\alpha+\beta x_i,\ \beta^2\sigma_\delta^2+\sigma_\epsilon^2\big),$$

which is exactly the right-hand side.
Method (2) (completing the square). The exponent is $-\frac{1}{2}Q(\xi_i)$ with

$$\begin{aligned}Q(\xi_i) &= \frac{\xi_i^2}{\tau^2} - 2c\,\xi_i + \text{const},\\ \frac{1}{\tau^2} &= \frac{1}{\sigma_\delta^2}+\frac{\beta^2}{\sigma_\epsilon^2} = \frac{\beta^2\sigma_\delta^2+\sigma_\epsilon^2}{\sigma_\delta^2\sigma_\epsilon^2},\end{aligned}$$

so $\int e^{-Q/2}d\xi_i = \sqrt{2\pi}\,\tau\,e^{-Q(\tau^2c)/2}$, and

$$\frac{1}{2\pi\sigma_\delta\sigma_\epsilon}\cdot\sqrt{2\pi}\,\tau = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\beta^2\sigma_\delta^2+\sigma_\epsilon^2}},$$

the leftover exponent being

$$-\tfrac{1}{2}\,\frac{\big(y_i-(\alpha+\beta x_i)\big)^2}{\beta^2\sigma_\delta^2+\sigma_\epsilon^2}.$$

(Check!)
(b) There is no joint distribution, so there is nothing to condition on, and the function is not a regression of Y on X even read formally.
(i) Since uniform$(-\infty,\infty)$ is not a probability distribution, the function $h(x,y)$ on the right of (a) is obtained by integrating against Lebesgue measure and is the joint pdf of nothing: $\int h(x,y)\,dy=1$ for each x, whence $\iint h=\infty$. A conditional pdf is by Definition 4.2.1 the ratio $f_{X,Y}/f_X$, and here neither term exists.
(ii) h depends on (x,y) only through $y-\beta x$: it is the n$(\alpha,\ \beta^2\sigma_\delta^2+\sigma_\epsilon^2)$ density of the single scalar $Y_i-\beta X_i=\alpha+\epsilon_i-\beta\delta_i$, whose "error variance" has absorbed the measurement error δ_i. So the regressor is correlated with the error,

$$\operatorname{Cov}\big(X_i,\ Y_i-\alpha-\beta X_i\big)=\operatorname{Cov}\big(\xi_i+\delta_i,\ \epsilon_i-\beta\delta_i\big)=-\beta\sigma_\delta^2\neq 0,$$

which the model of Section 11.3 forbids. Indeed, at every finite σ_ξ^2 the genuine structural model of Exercise 12.6 has, by the bivariate normal conditional mean applied to (12.2.19),

$$\mathrm{E}(Y_i\mid X_i=x)=\alpha+\beta\xi+\frac{\beta\sigma_\xi^2}{\sigma_\xi^2+\sigma_\delta^2}(x-\xi),$$

an *attenuated* slope, and that is what least squares estimates – never β.

12.2 Exercises 12.8–12.14

Problem 12.8. Verify the nonidentifiability problems in the structural relationship model in the following ways.
(a) Produce two different sets of parameters that give the same marginal distribution to (X_i,Y_i).
(b) Show that there are at least two distinct parameter vectors that yield the same solution to the equations given in (12.2.20).

Solution. (a) The parameter vectors $(\alpha,\beta,\xi,\sigma_\epsilon^2,\sigma_\delta^2,\sigma_\xi^2)=(0,1,0,1,1,1)$ and $(0,\frac{3}{2},0,\frac{1}{2},\frac{4}{3},\frac{2}{3})$ both give (X_i,Y_i) the bivariate normal law with means $(0,0)$, variances $(2,2)$, and covariance 1.
By (12.2.19) the marginal distribution in the structural model (12.2.6) is

$$(X_i,Y_i)\sim\text{bivariate normal}\big(\xi,\ \alpha+\beta\xi,\ \sigma_\delta^2+\sigma_\xi^2,\ \sigma_\epsilon^2+\beta^2\sigma_\xi^2,\ \beta\sigma_\xi^2\big),$$

so the six parameters enter only through these five functions. Fixing $\xi=0$ and $\alpha=0$ and targeting means $(0,0)$, variances $(2,2)$, covariance 1, the whole one-parameter family

$$\sigma_\xi^2=\frac{1}{\beta},\qquad \sigma_\delta^2=2-\frac{1}{\beta},\qquad \sigma_\epsilon^2=2-\beta,\qquad \tfrac{1}{2}<\beta<2,$$

works: $\sigma_\delta^2+\sigma_\xi^2=2$, $\beta\sigma_\xi^2=1$, and $\sigma_\epsilon^2+\beta^2\sigma_\xi^2=2-\beta+\beta=2$, with all three variances strictly positive on $(\frac{1}{2},2)$. Taking $\beta=1$ and $\beta=\frac{3}{2}$ gives the two displayed vectors, so β—and with it every variance component—is unidentifiable.
(b) The same degree of freedom appears in the sample equations. In (12.2.20) there are five equations in the six unknowns $(\alpha,\beta,\xi,\sigma_\epsilon^2,\sigma_\delta^2,\sigma_\xi^2)$, and for every $\hat\beta\neq 0$ the assignment

$$\begin{aligned}
\hat\xi&=\bar x, & \hat\alpha&=\bar y-\hat\beta\bar x,\\
\hat\sigma_\xi^2&=\frac{S_{xy}}{n\hat\beta}, & \hat\sigma_\delta^2&=\frac{1}{n}\Big(S_{xx}-\frac{S_{xy}}{\hat\beta}\Big),\\
\hat\sigma_\epsilon^2&=\frac{1}{n}\big(S_{yy}-\hat\beta S_{xy}\big)
\end{aligned}$$

satisfies all five equations identically (substitute and check). It is an admissible parameter vector exactly when the three variances are nonnegative, which by Exercise 12.9 holds iff $\hat{\beta}$ has the sign of S_{xy} and

$$\frac{|S_{xy}|}{S_{xx}} \;\le\; |\hat{\beta}| \;\le\; \frac{S_{yy}}{|S_{xy}|}.$$

This interval has nonempty interior whenever $S_{xy}^2 < S_{xx}S_{yy}$, i.e. whenever the data are not exactly collinear, so infinitely many—in particular at least two—distinct parameter vectors solve (12.2.20) and hence maximize the likelihood.

Problem 12.9. In the structural relationship model, the solution to the equations in (12.2.20) implies a restriction on $\hat{\beta}$, the same restriction seen in the functional relationship case (see Exercise 12.4).
(a) Show that in (12.2.20), the MLE of σ_δ^2 is nonnegative only if $S_{xx} \ge (1/\hat{\beta})S_{xy}$. Also, the MLE of σ_ϵ^2 is nonnegative only if $S_{yy} \ge \hat{\beta}S_{xy}$.
(b) Show that the restrictions in part (a), together with the rest of the equations in (12.2.20), imply that

$$\frac{|S_{xy}|}{S_{xx}} \;\le\; |\hat{\beta}| \;\le\; \frac{S_{yy}}{|S_{xy}|}.$$

Solution. (a) Eliminate $\hat{\sigma}_\xi^2$ between the last two equations of (12.2.20) using the last one. From $\frac{1}{n}S_{xy} = \hat{\beta}\hat{\sigma}_\xi^2$ we get $\hat{\sigma}_\xi^2 = S_{xy}/(n\hat{\beta})$, and substituting into $\frac{1}{n}S_{xx} = \hat{\sigma}_\delta^2 + \hat{\sigma}_\xi^2$ and $\frac{1}{n}S_{yy} = \hat{\sigma}_\epsilon^2 + \hat{\beta}^2\hat{\sigma}_\xi^2$ gives

$$\hat{\sigma}_\delta^2 = \frac{1}{n}\Big(S_{xx} - \frac{1}{\hat{\beta}}S_{xy}\Big), \qquad \hat{\sigma}_\epsilon^2 = \frac{1}{n}\big(S_{yy} - \hat{\beta}S_{xy}\big).$$

So $\hat{\sigma}_\delta^2 \ge 0$ iff $S_{xx} \ge (1/\hat{\beta})S_{xy}$ and $\hat{\sigma}_\epsilon^2 \ge 0$ iff $S_{yy} \ge \hat{\beta}S_{xy}$.
(b) The remaining equation $\hat{\sigma}_\xi^2 = S_{xy}/(n\hat{\beta}) \ge 0$ forces $\hat{\beta}$ and S_{xy} to have the same sign, so

$$\hat{\beta}S_{xy} = |\hat{\beta}|\,|S_{xy}|, \qquad \frac{1}{\hat{\beta}}S_{xy} = \frac{|S_{xy}|}{|\hat{\beta}|}.$$

Part (a) therefore reads $S_{xx} \ge |S_{xy}|/|\hat{\beta}|$ and $S_{yy} \ge |\hat{\beta}|\,|S_{xy}|$, that is,

$$\frac{|S_{xy}|}{S_{xx}} \;\le\; |\hat{\beta}| \;\le\; \frac{S_{yy}}{|S_{xy}|},$$

the two bounds being the moduli of the $\lambda \to 0$ and $\lambda \to \infty$ limits of $\hat{\beta}(\lambda)$ in Exercise 12.4(a),(b).

Problem 12.10. (a) Derive the MLEs for $(\alpha, \beta, \sigma_\epsilon^2, \sigma_\delta^2, \sigma_\xi^2)$ in the structural relationship model by solving the equations (12.2.20) under the assumption that $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$.
(b) Calculate the MLEs for $(\alpha, \beta, \sigma_\epsilon^2, \sigma_\delta^2, \sigma_\xi^2)$ for the data of Exercise 12.4 by assuming the structural relationship model holds and that $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$. (The data of Exercise 12.4(e) are body weight x in kg and brain weight y in g for eight animal species.)

Species	Body weight (x)	Brain weight (y)
Arctic fox	3.385	44.50
Owl monkey	.480	15.50
Mountain beaver	1.350	8.10
Guinea pig	1.040	5.50
Chinchilla	.425	6.40
Ground squirrel	.101	4.00
Tree hyrax	2.000	12.30
Big brown bat	.023	.30

(c) Verify the relationship between variance estimates in the functional and structural relationship models. In particular, show that

$$\widehat{\operatorname{Var}} X_i(\text{structural}) = 2\,\widehat{\operatorname{Var}} X_i(\text{functional}).$$

That is, verify

$$\Big(S_{xx} - \frac{S_{xy}}{\hat\beta}\Big) = \frac{\lambda}{1+\lambda\hat\beta^2}\sum_{i=1}^{n}\big(y_i - (\hat\alpha + \hat\beta x_i)\big)^2.$$

(d) Verify the following equality, which is implicit in the MLE variance estimates given in (12.2.21). Show that

$$S_{xx} - \frac{S_{xy}}{\hat\beta} = \lambda\big(S_{yy} - \hat\beta S_{xy}\big).$$

Solution. (a) The constraint makes the system square, and it reproduces (12.2.16) and (12.2.21):

$$\hat\xi = \bar x, \qquad \hat\alpha = \bar y - \hat\beta\bar x,$$

$$\hat\beta = \frac{-(S_{xx} - \lambda S_{yy}) + \sqrt{(S_{xx} - \lambda S_{yy})^2 + 4\lambda S_{xy}^2}}{2\lambda S_{xy}},$$

$$\hat\sigma_\xi^2 = \frac{1}{n}\frac{S_{xy}}{\hat\beta}, \qquad \hat\sigma_\epsilon^2 = \frac{1}{n}\big(S_{yy} - \hat\beta S_{xy}\big),$$

$$\hat\sigma_\delta^2 = \lambda\hat\sigma_\epsilon^2 = \frac{1}{n}\Big(S_{xx} - \frac{S_{xy}}{\hat\beta}\Big).$$

Indeed, the first two equations of (12.2.20) give $\hat\xi$ and $\hat\alpha$ at once, and the last three give the three variance formulas exactly as in Exercise 12.9(a). The sixth relation $\hat\sigma_\delta^2 = \lambda\hat\sigma_\epsilon^2$ then determines $\hat\beta$:

$$S_{xx} - \frac{S_{xy}}{\hat\beta} = \lambda\big(S_{yy} - \hat\beta S_{xy}\big)$$

$$\iff \quad \lambda S_{xy}\hat\beta^2 + (S_{xx} - \lambda S_{yy})\hat\beta - S_{xy} = 0,$$

after multiplying by $\hat\beta$ and collecting terms. The two roots have product $-S_{xy}/(\lambda S_{xy}) = -1/\lambda < 0$, so they are of opposite sign, and only the root sharing the sign of S_{xy} makes $\hat\sigma_\xi^2 \ge 0$ (Exercise 12.9(b)); that is the "+" root displayed above, which is exactly $\hat\beta$ of (12.2.16). Thus the structural MLEs of α and β coincide with the functional ones, while the variance estimates are those of (12.2.21).
(b) With $\lambda = 1$ (the case treated in Exercise 12.4(e), where the EIV slope is compared with the two least squares slopes):

$$n = 8, \quad \bar x = 1.1005, \quad \bar y = 12.075,$$

$$S_{xx} = 9.0953, \qquad S_{yy} = 1358.2550, \qquad S_{xy} = 96.1501.$$

Then $S_{xx} - \lambda S_{yy} = -1349.1597$ and $\sqrt{(S_{xx} - \lambda S_{yy})^2 + 4\lambda S_{xy}^2} = 1362.7954$, so

$$\hat\beta = \frac{1349.1597 + 1362.7954}{2(96.1501)} = 14.1027, \qquad \hat\alpha = 12.075 - 14.1027(1.1005) = -3.4450.$$

The variance estimates are

$$\hat\sigma_\xi^2 = \frac{96.1501}{8(14.1027)} = .8522, \qquad \hat\sigma_\epsilon^2 = \frac{1358.2550 - 14.1027(96.1501)}{8} = .2847,$$

and $\hat\sigma_\delta^2 = \lambda\hat\sigma_\epsilon^2 = .2847$; here $\hat\sigma_\delta^2 + \hat\sigma_\xi^2 = 1.1369 = S_{xx}/n$ and $\hat\beta\hat\sigma_\xi^2 = 12.0188 = S_{xy}/n$. (Check!)
(c) Since $\hat\alpha = \bar y - \hat\beta\bar x$,

$$\sum_{i=1}^n \big(y_i - (\hat\alpha + \hat\beta x_i)\big)^2 = \sum_{i=1}^n \big((y_i - \bar y) - \hat\beta(x_i - \bar x)\big)^2 = S_{yy} - 2\hat\beta S_{xy} + \hat\beta^2 S_{xx},$$

so the asserted identity is equivalent, after multiplying by $1 + \lambda\hat\beta^2$, to

$$\begin{aligned} &\Big(S_{xx} - \frac{S_{xy}}{\hat\beta}\Big)(1 + \lambda\hat\beta^2) = \lambda\big(S_{yy} - 2\hat\beta S_{xy} + \hat\beta^2 S_{xx}\big) \\ \iff\quad & S_{xx} - \frac{S_{xy}}{\hat\beta} - \lambda\hat\beta S_{xy} = \lambda S_{yy} - 2\lambda\hat\beta S_{xy} \\ \iff\quad & \lambda S_{xy}\hat\beta^2 + (S_{xx} - \lambda S_{yy})\hat\beta - S_{xy} = 0, \end{aligned}$$

where the last step multiplied by $\hat\beta$. That is the defining quadratic for $\hat\beta$ from part (a), so the identity holds. Since the structural estimate is $\hat\sigma_\delta^2 = \frac{1}{n}(S_{xx} - S_{xy}/\hat\beta)$ while the functional estimate (12.2.18) is $\frac{1}{2n}\lambda(1 + \lambda\hat\beta^2)^{-1}\sum_i (y_i - (\hat\alpha + \hat\beta x_i))^2$, the identity says precisely that the structural estimate is twice the functional one.
(d) Divide the quadratic $\lambda S_{xy}\hat\beta^2 + (S_{xx} - \lambda S_{yy})\hat\beta - S_{xy} = 0$ by $\hat\beta$:

$$\lambda\hat\beta S_{xy} + S_{xx} - \lambda S_{yy} - \frac{S_{xy}}{\hat\beta} = 0 \quad\iff\quad S_{xx} - \frac{S_{xy}}{\hat\beta} = \lambda\big(S_{yy} - \hat\beta S_{xy}\big),$$

which is the equality $\hat\sigma_\delta^2 = \lambda\hat\sigma_\epsilon^2$ implicit in (12.2.21).

Problem 12.11. (a) Show that for random variables X and Y and constants a, b, c, d,

$$\operatorname{Cov}(aY + bX,\, cY + dX) = ac\operatorname{Var} Y + (bc + ad)\operatorname{Cov}(X, Y) + bd\operatorname{Var} X.$$

(b) Use the result in part (a) to verify that in the structural relationship model with $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$,

$$\operatorname{Cov}(\beta\lambda Y_i + X_i,\, Y_i - \beta X_i) = 0,$$

the identity on which the Creasy–Williams confidence set is based.
(c) Use the results of part (b) to show that

$$\frac{\sqrt{n-2}\, r_\lambda(\beta)}{\sqrt{1 - r_\lambda^2(\beta)}} \sim t_{n-2}$$

for any value of β, where $r_\lambda(\beta)$ is given in (12.2.23). Also, show that the confidence set defined in (12.2.24) has constant coverage probability equal to $1 - \alpha$.

Solution. (a) Expand by bilinearity of covariance:

$$\begin{aligned} \operatorname{Cov}(aY + bX,\, cY + dX) &= ac\operatorname{Cov}(Y, Y) + ad\operatorname{Cov}(Y, X) \\ &\quad + bc\operatorname{Cov}(X, Y) + bd\operatorname{Cov}(X, X) \\ &= ac\operatorname{Var} Y + (bc + ad)\operatorname{Cov}(X, Y) + bd\operatorname{Var} X. \end{aligned}$$

(b) By (12.2.19), $\operatorname{Var} X_i = \sigma_\delta^2 + \sigma_\xi^2$, $\operatorname{Var} Y_i = \sigma_\epsilon^2 + \beta^2\sigma_\xi^2$, and $\operatorname{Cov}(X_i, Y_i) = \beta\sigma_\xi^2$. Apply part (a) with

$a=\beta\lambda$, $b=1$, $c=1$, $d=-\beta$:

$$\begin{aligned}\operatorname{Cov}(\beta\lambda Y_i+X_i,\,Y_i-\beta X_i) &= \beta\lambda\big(\sigma_\epsilon^2+\beta^2\sigma_\xi^2\big)+\big(1-\lambda\beta^2\big)\beta\sigma_\xi^2-\beta\big(\sigma_\delta^2+\sigma_\xi^2\big)\\ &= \beta\lambda\sigma_\epsilon^2+\lambda\beta^3\sigma_\xi^2+\beta\sigma_\xi^2-\lambda\beta^3\sigma_\xi^2-\beta\sigma_\delta^2-\beta\sigma_\xi^2\\ &= \beta\big(\lambda\sigma_\epsilon^2-\sigma_\delta^2\big)=0,\end{aligned}$$

the last step by the assumption $\sigma_\delta^2=\lambda\sigma_\epsilon^2$.
(c) Put $U_i=\beta\lambda Y_i+X_i$ and $V_i=Y_i-\beta X_i$ at the true β. The map $(x,y)\mapsto(\beta\lambda y+x,\,y-\beta x)$ is linear with determinant $1+\lambda\beta^2\neq 0$, so (U_i,V_i), $i=1,\dots,n$, are iid nondegenerate bivariate normal, and by part (b) their correlation is 0. Now $r_\lambda(\beta)$ of (12.2.23) is by definition the sample correlation coefficient of the pairs (U_i,V_i)—the second expression in (12.2.23) is the first one with the sums expanded, since

$$\begin{aligned}\textstyle\sum_i(U_i-\bar U)(V_i-\bar V) &= \beta\lambda S_{yy}+(1-\beta^2\lambda)S_{xy}-\beta S_{xx},\\ \textstyle\sum_i(U_i-\bar U)^2 &= \beta^2\lambda^2 S_{yy}+2\beta\lambda S_{xy}+S_{xx},\\ \textstyle\sum_i(V_i-\bar V)^2 &= S_{yy}-2\beta S_{xy}+\beta^2 S_{xx}.\end{aligned}$$

Exercise 11.33 gives the null distribution of the sample correlation coefficient of an iid bivariate normal sample with zero correlation:

$$T(\beta)=\frac{\sqrt{n-2}\,r_\lambda(\beta)}{\sqrt{1-r_\lambda^2(\beta)}}\sim t_{n-2}.$$

Finally $T^2(\beta)\sim F_{1,n-2}$, the square of a t_{n-2} variable being $F_{1,n-2}$, and the set in (12.2.24) contains β exactly when $T^2(\beta)\le F_{1,n-2,\alpha}$, so

$$P\big(\beta\in C_\lambda(\hat\beta)\big)=P\big(T^2(\beta)\le F_{1,n-2,\alpha}\big)=1-\alpha$$

for every $(\alpha,\beta,\sigma_\epsilon^2,\sigma_\xi^2)$, since the law of $T(\beta)$ is free of all parameters.

Problem 12.12. Verify the following facts about $\hat\beta$ (the MLE of β when we assume $\sigma_\delta^2=\lambda\sigma_\epsilon^2$), $r_\lambda(\beta)$ of (12.2.23), and $C_\lambda(\hat\beta)$, the Creasy–Williams confidence set of (12.2.24).
(a) $\hat\beta$ and $-1/(\lambda\hat\beta)$ are the two roots of the quadratic equation defining the zeros of the first derivative of the likelihood function (12.2.14).
(b) $r_\lambda(\beta)=-r_\lambda(-1/(\lambda\beta))$ for every β.
(c) If $\beta\in C_\lambda(\hat\beta)$, then $-1/(\lambda\beta)\in C_\lambda(\hat\beta)$.

Solution. (a) The quadratic is

$$q(\beta)\;=\;\lambda S_{xy}\beta^2+(S_{xx}-\lambda S_{yy})\beta-S_{xy}=0,$$

and its two roots multiply to $-S_{xy}/(\lambda S_{xy})=-1/\lambda$, so if one root is $\hat\beta$ the other is $(-1/\lambda)/\hat\beta=-1/(\lambda\hat\beta)$. To see that q is the derivative equation: in (12.2.14) the parameter β enters only through

$$h(\beta)=\frac{\lambda}{1+\lambda\beta^2}\sum_{i=1}^n\big(y_i-(\alpha+\beta x_i)\big)^2,$$

which appears with a negative sign in the exponent, so L is maximized by minimizing h. For fixed β the inner sum is minimized at $\alpha=\bar y-\beta\bar x$, giving $h(\beta)=\lambda\big(S_{yy}-2\beta S_{xy}+\beta^2S_{xx}\big)/(1+\lambda\beta^2)$, and

$$\begin{aligned}\frac{(1+\lambda\beta^2)^2}{2\lambda}h'(\beta) &= (\beta S_{xx}-S_{xy})(1+\lambda\beta^2)-\lambda\beta\big(S_{yy}-2\beta S_{xy}+\beta^2S_{xx}\big)\\ &= \lambda S_{xy}\beta^2+(S_{xx}-\lambda S_{yy})\beta-S_{xy}\;=\;q(\beta),\end{aligned}$$

the cubic terms $\lambda\beta^3S_{xx}$ cancelling. The root with the sign of S_{xy} is the minimizer of h, hence the MLE (12.2.16) (Exercise 12.2 checks this for $\lambda=1$, and the reparametrization (12.2.15) carries it to general λ); the other root, $-1/(\lambda\hat\beta)$, maximizes h.

(b) Write $\beta' = -1/(\lambda\beta)$ and let N, A, B denote the numerator and the two sums of squares in (12.2.23),

$$N(\beta) = \lambda\beta S_{yy} + (1-\lambda\beta^2)S_{xy} - \beta S_{xx} = -q(\beta),$$
$$A(\beta) = \lambda^2\beta^2 S_{yy} + 2\lambda\beta S_{xy} + S_{xx}, \qquad B(\beta) = S_{yy} - 2\beta S_{xy} + \beta^2 S_{xx},$$

so that $r_\lambda(\beta) = N(\beta)/\sqrt{A(\beta)B(\beta)}$. Substituting and clearing denominators,

$$N(\beta') = -\frac{N(\beta)}{\lambda\beta^2}, \qquad A(\beta') = \frac{B(\beta)}{\beta^2}, \qquad B(\beta') = \frac{A(\beta)}{\lambda^2\beta^2}.$$

(For the first: $-\lambda\beta^2 N(\beta') = \lambda\beta S_{yy} - \lambda\beta^2 S_{xy} + S_{xy} - \beta S_{xx} = N(\beta)$.) Hence $A(\beta')B(\beta') = A(\beta)B(\beta)/(\lambda\beta^2)^2$ and, taking positive square roots,

$$r_\lambda(\beta') = \frac{-N(\beta)/(\lambda\beta^2)}{\sqrt{A(\beta)B(\beta)}/(\lambda\beta^2)} = -r_\lambda(\beta).$$

(c) Membership in (12.2.24) is a condition on r_λ^2 alone, and part (b) gives $r_\lambda^2(\beta') = (-r_\lambda(\beta))^2 = r_\lambda^2(\beta)$ for $\beta' = -1/(\lambda\beta)$, so

$$\frac{(n-2)r_\lambda^2(\beta')}{1-r_\lambda^2(\beta')} = \frac{(n-2)r_\lambda^2(\beta)}{1-r_\lambda^2(\beta)} \le F_{1,n-2,\alpha}$$

whenever $\beta \in C_\lambda(\hat\beta)$; hence $\beta' \in C_\lambda(\hat\beta)$. (In particular $\hat\beta$ and $-1/(\lambda\hat\beta)$ both lie in the set, the numerator $-q$ of r_λ vanishing at each by part (a).)

Problem 12.13. There is an interesting connection between the Creasy–Williams confidence set of (12.2.24) and the interval $C_{\mathrm{G}}(\hat\beta)$ of (12.2.22).
(a) Show that

$$C_{\mathrm{G}}(\hat\beta) = \left\{ \beta : \frac{(\beta-\hat\beta)^2}{\hat\sigma_\beta^2/(n-2)} \le F_{1,n-2,\alpha} \right\},$$

where $\hat\beta$ is the MLE of β and $\hat\sigma_\beta^2$ is the previously defined consistent estimator of σ_β^2.
(b) Show that the Creasy–Williams set can be written in the form

$$\left\{ \beta : \frac{(\beta-\hat\beta)^2}{\hat\sigma_\beta^2/(n-2)} \left[\frac{(1+\lambda\beta\hat\beta)^2}{(1+\lambda\beta^2)^2}\right] \le F_{1,n-2,\alpha} \right\}.$$

Hence $C_{\mathrm{G}}(\hat\beta)$ can be derived by replacing the term in square brackets with 1, its probability limit. (In deriving this representation, the fact that $\hat\beta$ and $-1/(\lambda\hat\beta)$ are roots of the numerator of $r_\lambda(\beta)$ is of great help. In particular, the fact that

$$\frac{r_\lambda^2(\beta)}{1-r_\lambda^2(\beta)} = \frac{\lambda^2 S_{xy}^2(\beta-\hat\beta)^2\big(\beta+(1/(\lambda\hat\beta))\big)^2}{(1+\lambda\beta^2)^2\big(S_{xx}S_{yy}-S_{xy}^2\big)}$$

is straightforward to establish.)

Solution. (a) Square the two-sided bound. The interval (12.2.22) is

$$\hat\beta - \frac{t_{n-2,\alpha/2}\hat\sigma_\beta}{\sqrt{n-2}} \le \beta \le \hat\beta + \frac{t_{n-2,\alpha/2}\hat\sigma_\beta}{\sqrt{n-2}} \iff \frac{|\beta-\hat\beta|}{\hat\sigma_\beta/\sqrt{n-2}} \le t_{n-2,\alpha/2},$$

and squaring, together with $t_{n-2,\alpha/2}^2 = F_{1,n-2,\alpha}$, gives the stated form.
(b) Two facts do all the work. First, by Exercise 12.12(a) the numerator of $r_\lambda(\beta)$ in (12.2.23) is $-q(\beta)$ with $q(\beta) = \lambda S_{xy}\beta^2 + (S_{xx} - \lambda S_{yy})\beta - S_{xy}$, whose roots are $\hat\beta$ and $-1/(\lambda\hat\beta)$ and whose leading coefficient

is λS_{xy}; hence

$$N(\beta) = -\lambda S_{xy}(\beta - \hat\beta)\Big(\beta + \frac{1}{\lambda\hat\beta}\Big).$$

Second, with A, B as in Exercise 12.12(b), a direct expansion gives

$$A(\beta)B(\beta) - N(\beta)^2 = (1+\lambda\beta^2)^2\big(S_{xx}S_{yy} - S_{xy}^2\big). \qquad \text{(Check!)}$$

Since $r_\lambda^2/(1-r_\lambda^2) = N^2/(AB-N^2)$, the two displays give exactly the identity quoted in the hint. Now simplify $\hat\sigma_\beta^2$. Writing $D = \sqrt{(S_{xx}-\lambda S_{yy})^2 + 4\lambda S_{xy}^2}$, the two roots of q differ by $D/(\lambda S_{xy})$, so

$$\Big(\hat\beta + \frac{1}{\lambda\hat\beta}\Big)^2 = \frac{D^2}{\lambda^2 S_{xy}^2}, \qquad \text{i.e.} \qquad D^2 = \lambda^2 S_{xy}^2 \cdot \frac{(1+\lambda\hat\beta^2)^2}{\lambda^2\hat\beta^2} = \frac{S_{xy}^2(1+\lambda\hat\beta^2)^2}{\hat\beta^2},$$

and therefore

$$\hat\sigma_\beta^2 = \frac{(1+\lambda\hat\beta^2)^2(S_{xx}S_{yy}-S_{xy}^2)}{D^2} = \frac{\hat\beta^2\big(S_{xx}S_{yy}-S_{xy}^2\big)}{S_{xy}^2}.$$

Substituting $S_{xx}S_{yy} - S_{xy}^2 = \hat\sigma_\beta^2 S_{xy}^2/\hat\beta^2$ into the hint identity and using $\lambda\hat\beta\big(\beta + 1/(\lambda\hat\beta)\big) = 1 + \lambda\beta\hat\beta$,

$$\begin{aligned}\frac{(n-2)r_\lambda^2(\beta)}{1-r_\lambda^2(\beta)} &= \frac{(n-2)\lambda^2 S_{xy}^2\hat\beta^2(\beta-\hat\beta)^2\big(\beta+\frac{1}{\lambda\hat\beta}\big)^2}{(1+\lambda\beta^2)^2\,\hat\sigma_\beta^2\,S_{xy}^2} \\ &= \frac{(\beta-\hat\beta)^2}{\hat\sigma_\beta^2/(n-2)} \cdot \frac{(1+\lambda\beta\hat\beta)^2}{(1+\lambda\beta^2)^2}.\end{aligned}$$

Imposing $\le F_{1,n-2,\alpha}$ gives the stated representation of (12.2.24).
Finally, in the structural model $\hat\beta \to \beta$ in probability (the estimators of (12.2.16) and (12.2.21) are consistent when $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$), and $b \mapsto (1+\lambda\beta b)^2/(1+\lambda\beta^2)^2$ is continuous, so by Theorem 5.5.4 the bracketed factor satisfies

$$\frac{(1+\lambda\beta\hat\beta)^2}{(1+\lambda\beta^2)^2} \longrightarrow \frac{(1+\lambda\beta^2)^2}{(1+\lambda\beta^2)^2} = 1$$

in probability, and replacing it by 1 turns the Creasy–Williams set into $C_{\mathrm{G}}(\hat\beta)$ of part (a).

Problem 12.14. Graph the logistic regression function $\pi(x)$ from (12.3.2),

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}},$$

for these three cases: $\alpha = \beta = 1$, $\alpha = \beta = 2$, and $\alpha = \beta = 3$.

Solution. All three are the same S-shaped curve rescaled about the point $(-1, \frac12)$: with $\alpha = \beta = c$,

$$\pi_c(x) = \frac{e^{c(x+1)}}{1+e^{c(x+1)}} = \frac{1}{1+e^{-c(x+1)}}.$$

Each is strictly increasing from 0 (as $x \to -\infty$) to 1 (as $x \to \infty$), each passes through $\pi(-\alpha/\beta) = \pi(-1) = \frac12$, and each is symmetric about that point, $\pi_c(-1+t) = 1 - \pi_c(-1-t)$. By (12.3.3), the slope is $c\,\pi_c(x)(1-\pi_c(x))$, largest at $x=-1$ where it equals $c/4$: the curves have slopes .25, .50, .75 at the common crossing point, so the $c=2$ curve is the $c=1$ curve horizontally compressed by a factor 2 toward $x=-1$, and the $c=3$ curve by a factor 3.
Plotting the values below and joining them smoothly between the asymptotes $\pi=0$ and $\pi=1$ gives the three curves.

x	-4	-3	-2	-1.5	-1	-0.5	0	1	2
$\pi_1(x)$	.047	.119	.269	.378	.500	.622	.731	.881	.953
$\pi_2(x)$	.002	.018	.119	.269	.500	.731	.881	.982	.998
$\pi_3(x)$	.000	.002	.047	.182	.500	.818	.953	.998	1.000

12.3 Exercises 12.15–12.21

Problem 12.15. For the logistic regression function in (12.3.2),

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}},$$

verify these relationships.
(a) $\pi(-\alpha/\beta) = 1/2$
(b) $\pi((-\alpha/\beta)+c) = 1-\pi((-\alpha/\beta)-c)$ for any c
(c) (12.3.3) for $d\pi(x)/dx$, namely $d\pi(x)/dx = \beta\pi(x)(1-\pi(x))$
(d) (12.3.4) about the odds ratio, namely

$$e^{\beta} = \frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x))}$$

(e) (12.3.5) about the multiplicative change in odds, namely

$$\frac{\pi(x+1)}{1-\pi(x+1)} = e^{\beta}\,\frac{\pi(x)}{1-\pi(x)}$$

(f) (12.3.6) and (12.3.8) regarding the likelihood equations for a Bernoulli GLM: with $Y_i \sim$ Bernoulli(π_i), $\pi(x) = F(\alpha+\beta x)$, $F_i = F(\alpha+\beta x_i)$, $dF(w)/dw = f(w)$ and $f_i = f(\alpha+\beta x_i)$,

$$\frac{\partial}{\partial\alpha}\log\!\left(\frac{F_i}{1-F_i}\right) = \frac{f_i}{F_i(1-F_i)},$$

$$\frac{\partial}{\partial\beta}\log L(\alpha,\beta|\mathbf{y}) = \sum_{i=1}^{n}(y_i-F_i)\frac{f_i}{F_i(1-F_i)}\,x_i.$$

(g) For logistic regression, $f_i/(F_i(1-F_i)) = \beta$ in (12.3.7) and (12.3.8).

Solution. All seven read off $\pi(x) = e^u/(1+e^u)$, $1-\pi(x) = 1/(1+e^u)$ with $u = \alpha+\beta x$; equivalently (12.3.1), $\log(\pi(x)/(1-\pi(x))) = \alpha+\beta x$.
(a) At $x = -\alpha/\beta$ we have $u = 0$, so $\pi(-\alpha/\beta) = e^0/(1+e^0) = 1/2$.
(b) At $x = (-\alpha/\beta)\pm c$ we have $u = \pm\beta c$, so

$$\begin{aligned}\pi\!\left(\tfrac{-\alpha}{\beta}+c\right)+\pi\!\left(\tfrac{-\alpha}{\beta}-c\right) &= \frac{e^{\beta c}}{1+e^{\beta c}}+\frac{e^{-\beta c}}{1+e^{-\beta c}}\\ &= \frac{e^{\beta c}}{1+e^{\beta c}}+\frac{1}{1+e^{\beta c}} = 1.\end{aligned}$$

(c) Differentiating the quotient and recognizing the factors,

$$\frac{d\pi(x)}{dx} = \frac{\beta e^u(1+e^u)-e^u\beta e^u}{(1+e^u)^2} = \beta\,\frac{e^u}{1+e^u}\cdot\frac{1}{1+e^u} = \beta\pi(x)(1-\pi(x)).$$

(d) By (12.3.1) the odds at x are $\pi(x)/(1-\pi(x)) = e^{\alpha+\beta x}$, so

$$\frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x))} = \frac{e^{\alpha+\beta(x+1)}}{e^{\alpha+\beta x}} = e^{\beta}.$$

(e) Multiply (d) through by $\pi(x)/(1-\pi(x))$.

(f) For a Bernoulli GLM the log likelihood is

$$\log L(\alpha, \beta|\mathbf{y}) = \sum_{i=1}^{n} \left\{ \log(1 - F_i) + y_i \log\left(\frac{F_i}{1 - F_i}\right) \right\}.$$

Since $\partial F_i/\partial\alpha = f_i$,

$$\frac{\partial}{\partial\alpha} \log\left(\frac{F_i}{1 - F_i}\right) = \frac{f_i}{F_i} + \frac{f_i}{1 - F_i} = \frac{f_i}{F_i(1 - F_i)},$$

which is (12.3.6), and $\partial \log(1 - F_i)/\partial\alpha = -f_i/(1 - F_i) = -F_i f_i/(F_i(1 - F_i))$. Adding,

$$\frac{\partial}{\partial\alpha} \log L = \sum_{i=1}^{n} (y_i - F_i) \frac{f_i}{F_i(1 - F_i)},$$

which is (12.3.7). Since $\partial F_i/\partial\beta = x_i f_i$, every derivative above merely acquires a factor x_i, giving (12.3.8).
(g) Errata: the printed value β should be 1, as the text itself records on p. 594; β is the value of the ratio only if f_i is read as $d\pi(x)/dx$ at x_i rather than as the book's $f = dF(w)/dw$. With the book's definition, $F(w) = e^w/(1 + e^w)$ gives

$$f(w) = \frac{e^w}{(1 + e^w)^2} = F(w)(1 - F(w)),$$

so $f_i/(F_i(1 - F_i)) = 1$ and (12.3.7), (12.3.8) collapse to $\sum_i (y_i - F_i) = 0$, $\sum_i (y_i - F_i)x_i = 0$. Under the other reading, part (c) gives $d\pi(x_i)/dx = \beta F_i(1 - F_i)$ and the ratio is β.

Problem 12.16. Consider this logistic regression data. Only two values, $x = 0$ and 1, are observed. For $x = 0$ there are 10 successes in 10 trials. For $x = 1$ there are 5 successes in 10 trials. Show that the logistic regression MLEs $\hat{\alpha}$ and $\hat{\beta}$ do not exist for these data by verifying the following.
(a) The MLEs for $\pi(0)$ and $\pi(1)$, not restricted by (12.3.2), are given by $\hat{\pi}(0) = 1$ and $\hat{\pi}(1) = .5$.
(b) The overall maximum of the likelihood function given by the estimates in part (a) can not be achieved at any finite values of the logistic regression parameters α and β, but can be achieved in the limit as $\beta \to -\infty$ and $\alpha = -\beta$.

Solution. (a) Write $p = \pi(0)$, $q = \pi(1)$. With $Y_1^* \sim$ binomial$(10, p)$ and $Y_2^* \sim$ binomial$(10, q)$ independent, and $y_1^* = 10$, $y_2^* = 5$, the likelihood factors as

$$L(p, q) = \binom{10}{10} p^{10} \cdot \binom{10}{5} q^5 (1 - q)^5 = 252\, p^{10} q^5 (1 - q)^5,$$

a product of a factor in p alone and a factor in q alone, so each may be maximized separately over $[0, 1]$. The first is strictly increasing, giving $\hat{\pi}(0) = 1$; the second is the binomial likelihood with MLE the sample proportion (Example 7.2.7), $\hat{\pi}(1) = 5/10 = .5$. The unrestricted maximum value is therefore

$$L(1, .5) = 252 \cdot 1 \cdot (.5)^5 (.5)^5 = 252 \cdot 2^{-10}.$$

(b) Under (12.3.2) the parameters are $p = e^\alpha/(1 + e^\alpha)$ and $q = e^{\alpha+\beta}/(1 + e^{\alpha+\beta})$, and for every finite (α, β) we have $0 < p < 1$ strictly. Hence

$$L(p, q) = 252\, p^{10}\, q^5 (1 - q)^5 < 252 \cdot 1 \cdot 2^{-10},$$

the strict inequality coming from $p^{10} < 1$ and the bound $q^5(1 - q)^5 \le 2^{-10}$ valid for all q. So no finite (α, β) attains the bound $252 \cdot 2^{-10}$, and a maximizer of the logistic likelihood does not exist.
The bound is however the supremum. Take $\alpha = -\beta$ with $\beta \to -\infty$. Then $\alpha + \beta = 0$, so $q = e^0/(1 + e^0) = .5$ exactly, while $\alpha = -\beta \to +\infty$ gives

$$p = \frac{e^\alpha}{1 + e^\alpha} \longrightarrow 1.$$

Therefore

$$L \longrightarrow 252 \cdot 1 \cdot (.5)^5 (.5)^5 = 252 \cdot 2^{-10},$$

so the supremum $252 \cdot 2^{-10}$ is attained only in the limit $\beta \to -\infty$, $\alpha = -\beta \to +\infty$.

Problem 12.17. In *probit regression*, the link function is the standard normal cdf $\Phi(x) = P(Z \le x)$, where $Z \sim \mathrm{n}(0,1)$. Thus, in this model we observe $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$, where $Y_i \sim$ Bernoulli(π_i) and $\pi_i = \Phi(\alpha + \beta x_i)$.
(a) Write out the likelihood function and show how to solve for the MLEs of α and β.
(b) Fit the probit model to the data of Table 12.3.1. Comment on any differences from the logistic fit.

Table 12.3.1. Temperature at flight time (deg F) and failure of O-rings (1 = failure, 0 = success):

Flight no.	14	9	23	10	1	5	13	15	4	3	8	17
Failure	1	1	1	1	0	0	0	0	0	0	0	0
Temp.	53	57	58	63	66	67	67	67	68	69	70	70

Flight no.	2	11	6	7	16	21	19	22	12	20	18
Failure	1	1	0	0	0	1	0	0	0	0	0
Temp.	70	70	72	73	75	75	76	76	78	79	81

For reference, the logistic fit of Example 12.3.1 to these data is $\hat\alpha = 15.043$, $\hat\beta = -.232$, with $\hat\pi(31^\circ) = .9996$.

Solution. (a) Writing $\Phi_i = \Phi(\alpha + \beta x_i)$ and $\phi_i = \phi(\alpha + \beta x_i)$ with $\phi = \Phi'$ the standard normal pdf,

$$L(\alpha, \beta | \mathbf{y}) = \prod_{i=1}^{n} \Phi_i^{y_i} \, (1 - \Phi_i)^{1 - y_i},$$

$$\log L(\alpha, \beta | \mathbf{y}) = \sum_{i=1}^{n} \left\{ \log(1 - \Phi_i) + y_i \log\!\left(\frac{\Phi_i}{1 - \Phi_i} \right) \right\}.$$

This is the Bernoulli GLM of Section 12.3.2 with $F = \Phi$, $f = \phi$, so (12.3.7) and (12.3.8) give the likelihood equations directly:

$$0 = \sum_{i=1}^{n} (y_i - \Phi_i) \, \frac{\phi_i}{\Phi_i (1 - \Phi_i)},$$

$$0 = \sum_{i=1}^{n} (y_i - \Phi_i) \, \frac{\phi_i}{\Phi_i (1 - \Phi_i)} \, x_i.$$

Unlike the logistic case the factor $\phi_i / (\Phi_i (1 - \Phi_i))$ does not cancel (Exercise 12.15(g)), so these are genuinely nonlinear in (α, β) and must be solved numerically. For probit regression $\log L$ is strictly concave (p. 594), so any solution of the pair above is the unique MLE. Solve by Fisher scoring, the form of Newton–Raphson that replaces the observed second derivatives (12.3.9) by their expectations: with $\boldsymbol{\theta} = (\alpha, \beta)$, iterate

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + I(\boldsymbol{\theta}^{(k)})^{-1} \, \nabla \log L(\boldsymbol{\theta}^{(k)}),$$

where, $\partial \log L / \partial \alpha$ being a sum of independent mean-zero terms,

$$I(\alpha, \beta) = \sum_{i=1}^{n} \frac{\phi_i^2}{\Phi_i (1 - \Phi_i)} \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix},$$

which is positive definite whenever the x_i are not all equal, so the iteration is well posed.
(b) Running that iteration on Table 12.3.1 ($n = 23$, $x =$ temperature, $y =$ failure) gives

$$\hat\alpha = 8.775, \qquad \hat\beta = -.1351, \qquad \log L = -10.189,$$

with $\mathrm{se}(\hat\alpha) = 3.872$, $\mathrm{se}(\hat\beta) = .0565$ from the inverse information matrix. The Wald statistic for $H_0 : \beta = 0$ is $-.1351/.0565 = -2.39$, so $\beta < 0$ is supported, as in Example 12.3.2.

Differences from the logistic fit are slight over the observed range and pronounced in the extrapolated tail. The coefficients differ by the familiar scale factor

$$\frac{\hat\beta_{\text{logit}}}{\hat\beta_{\text{probit}}} = \frac{-.2322}{-.1351} = 1.72,$$

close to the 1.81-to-1 ratio of the standard deviations of the logistic$(0,1)$ and n$(0,1)$ link distributions, so the two fitted curves are near-reparametrizations of each other. The fitted failure probabilities agree to about one percentage point throughout:

Temp.	53	60	65	70	75	81
logistic	.9392	.7527	.4881	.2300	.0855	.0227
probit	.9468	.7483	.4975	.2477	.0873	.0151

The maximized log likelihoods, -10.158 (logistic) versus -10.189 (probit), are likewise indistinguishable: these data cannot discriminate between the two links. Where they part company is the extrapolation to the Challenger launch temperature 31°F, far outside the data:

$$\hat\pi_{\text{probit}}(31) = \Phi(8.775 - .1351(31)) = \Phi(4.587) = .999998,$$

against .9996 for the logistic. The normal tail decays like $e^{-w^2/2}$ while the logistic tail decays like e^{-w}, so the probit model is the more emphatic of the two about a 31° failure.

Problem 12.18. Brown and Rothery (1993, Chapter 4) discuss a generalization of the linear logistic model to the quadratic model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i + \gamma x_i^2.$$

(a) Write out the likelihood function and show how to solve for the MLEs of α, β, and γ.
(b) Using the log LRT, show how to test the hypothesis $H_0 : \gamma = 0$, that is, that the model is really linear logistic.
(c) Fit the quadratic logistic model to the data in the table on survival of sparrowhawks of different ages.

Age	1	2	3	4	5	6	7	8	9
No. of birds	77	149	182	118	78	46	27	10	4
No. surviving	35	89	130	79	52	28	14	3	1

(d) Decide whether the better model for the sparrowhawks is linear or quadratic, that is, test $H_0 : \gamma = 0$.

Solution. (a) The data are grouped as in the last display of p. 594: at age x_j there are n_j birds of which $Y_j^* \sim$ binomial$(n_j, \pi(x_j))$ survive, $j = 1, \ldots, J$. With

$$\eta_j = \alpha + \beta x_j + \gamma x_j^2, \qquad F_j = \frac{e^{\eta_j}}{1+e^{\eta_j}},$$

$$L(\alpha,\beta,\gamma|\mathbf{y}^*) = \prod_{j=1}^{J} \binom{n_j}{y_j^*} F_j^{\,y_j^*}(1-F_j)^{n_j - y_j^*},$$

$$\log L = \text{const} + \sum_{j=1}^{J} \left\{ n_j \log(1-F_j) + y_j^* \eta_j \right\}.$$

Since $\partial\eta_j/\partial(\alpha,\beta,\gamma) = (1, x_j, x_j^2)$ and $\partial F_j/\partial\eta_j = F_j(1-F_j)$, the logistic cancellation of Exercise

12.15(g) applies in each coordinate and the three likelihood equations are

$$0 = \sum_{j=1}^{J} (y_j^* - n_j F_j),$$
$$0 = \sum_{j=1}^{J} (y_j^* - n_j F_j)\, x_j,$$
$$0 = \sum_{j=1}^{J} (y_j^* - n_j F_j)\, x_j^2.$$

These are nonlinear in (α, β, γ) and are solved numerically by the Newton–Raphson iteration of Exercise 12.17(a), now in three parameters; for a logistic link the observed and expected information agree (the remark below (12.3.10)), and extending (12.3.10) gives

$$I(\alpha, \beta, \gamma) = \sum_{j=1}^{J} n_j F_j (1 - F_j) \begin{pmatrix} 1 & x_j & x_j^2 \\ x_j & x_j^2 & x_j^3 \\ x_j^2 & x_j^3 & x_j^4 \end{pmatrix}.$$

This is positive definite as soon as at least three distinct x_j carry data (the summands are rank-one outer products of $(1, x_j, x_j^2)$, whose Vandermonde structure makes any three of them independent), so $\log L$ is strictly concave and the solution is the unique MLE.
(b) The restricted space is $\Theta_0 = \{(\alpha, \beta, 0)\}$, so by Definition 8.2.1 the LRT statistic is

$$-2 \log \lambda(\mathbf{y}^*) = 2 \left[\log L(\hat\alpha, \hat\beta, \hat\gamma | \mathbf{y}^*) - \log L(\hat\alpha_0, \hat\beta_0, 0 | \mathbf{y}^*) \right],$$

where $(\hat\alpha, \hat\beta, \hat\gamma)$ solves the three equations of part (a) and $(\hat\alpha_0, \hat\beta_0)$ solves the first two with γ held at 0 — that is, the ordinary linear logistic fit. The full space has three free parameters and Θ_0 has two, so the degrees of freedom prescribed by Theorem 10.3.3 is $3 - 2 = 1$ (the theorem is stated for an iid sample, but the book applies it verbatim to this independent-but-not-identically-distributed logistic setting on p. 596, the regularity conditions of Miscellanea 10.6.2 holding because each Y_j^* comes from a regular exponential family with θ interior to Θ). Under H_0,

$$-2 \log \lambda(\mathbf{Y}^*) \xrightarrow{d} \chi^2_{3-2} = \chi^2_1.$$

For a level-.05 test, reject $H_0 : \gamma = 0$ when $-2 \log \lambda \geq \chi^2_{1,.05} = 3.84$.
(c) With $J = 9$, $x_j = j$, $(n_j) = (77, 149, 182, 118, 78, 46, 27, 10, 4)$ and $(y_j^*) = (35, 89, 130, 79, 52, 28, 14, 3, 1)$, the Newton–Raphson iteration of part (a) converges to

$$\hat\alpha = -.8450, \qquad \hat\beta = .8372, \qquad \hat\gamma = -.1030,$$

with $\log L(\hat\alpha, \hat\beta, \hat\gamma) = -445.890$ (constant term omitted throughout) and standard errors .328, .182, .0221 from the inverse of $I(\hat\alpha, \hat\beta, \hat\gamma)$. The fitted survival probabilities track the observed proportions closely:

Age	1	2	3	4	5	6	7	8	9
observed	.4545	.5973	.7143	.6695	.6667	.6087	.5185	.3000	.2500
quadratic	.4723	.6029	.6770	.7019	.6829	.6157	.4926	.3237	.1611
linear	.6195	.6213	.6230	.6247	.6265	.6282	.6300	.6317	.6334

Since $\hat\gamma < 0$ the fitted log odds is concave with maximum at $x = -\hat\beta/(2\hat\gamma) = 4.07$ years: survival rises through the birds' fourth year and falls thereafter.
(d) The linear logistic fit to the same data gives $\hat\alpha_0 = .4800$, $\hat\beta_0 = .00743$ with $\log L = -457.572$ — essentially a flat curve, as the third row of the table above shows. Hence

$$-2 \log \lambda = 2 \left[-445.890 - (-457.572) \right] = 23.36.$$

Since $23.36 \geq \chi^2_{1,.05} = 3.84$ (indeed $P(\chi^2_1 \geq 23.36) = 1.3 \times 10^{-6}$), reject $H_0 : \gamma = 0$: the quadratic logistic model is the better model for the sparrowhawks.

Problem 12.19. For the logistic regression model:
(a) Show that $\left(\sum_{j=1}^{J} Y_j^*,\ \sum_{j=1}^{J} Y_j^* x_j\right)$ is a sufficient statistic for (α, β).
(b) Verify the formula for the logistic regression information matrix in (12.3.10), namely

$$I(\alpha,\beta) = \begin{pmatrix} \sum_{j=1}^{J} n_j F_j(1-F_j) & \sum_{j=1}^{J} x_j n_j F_j(1-F_j) \\ \sum_{j=1}^{J} x_j n_j F_j(1-F_j) & \sum_{j=1}^{J} x_j^2 n_j F_j(1-F_j) \end{pmatrix},$$

where $F_j = e^{\alpha+\beta x_j}/(1+e^{\alpha+\beta x_j})$.

Solution. (a) The likelihood admits $\mathbf{y}^*$ only through the two sums. With $Y_j^* \sim \text{binomial}(n_j, F_j)$ independently and n_j, x_j known, (12.3.1) gives $\log(F_j/(1-F_j)) = \alpha+\beta x_j$ and $1-F_j = (1+e^{\alpha+\beta x_j})^{-1}$, so

$$\begin{aligned} L(\alpha,\beta|\mathbf{y}^*) &= \prod_{j=1}^{J} \binom{n_j}{y_j^*} \left(\frac{F_j}{1-F_j}\right)^{y_j^*} (1-F_j)^{n_j} \\ &= \left[\prod_{j=1}^{J} \binom{n_j}{y_j^*}\right] \exp\left\{\sum_{j=1}^{J} y_j^*(\alpha+\beta x_j)\right\} \prod_{j=1}^{J} \left(1+e^{\alpha+\beta x_j}\right)^{-n_j} \\ &= h(\mathbf{y}^*)\, g\Big(\textstyle\sum_j y_j^*,\ \sum_j y_j^* x_j \mid \alpha,\beta\Big), \end{aligned}$$

with

$$g(t_1,t_2|\alpha,\beta) = e^{\alpha t_1+\beta t_2} \prod_{j=1}^{J} \left(1+e^{\alpha+\beta x_j}\right)^{-n_j}$$

depending on $\mathbf{y}^*$ only through (t_1,t_2), and $h(\mathbf{y}^*) = \prod_j \binom{n_j}{y_j^*}$ free of (α,β). By the Factorization Theorem (Theorem 6.2.6), $\left(\sum_j Y_j^*, \sum_j Y_j^* x_j\right)$ is sufficient for (α,β).
(b) From the factorization above,

$$\log L = \text{const} + \alpha \sum_j y_j^* + \beta \sum_j y_j^* x_j - \sum_j n_j \log\left(1+e^{\alpha+\beta x_j}\right).$$

Since $\partial \log(1+e^{\alpha+\beta x_j})/\partial\alpha = F_j$ and, by the computation of (12.3.3) with the chain factor 1 in place of β, $\partial F_j/\partial\alpha = F_j(1-F_j)$,

$$\frac{\partial}{\partial\alpha}\log L = \sum_j \left(y_j^* - n_j F_j\right), \qquad -\frac{\partial^2}{\partial\alpha^2}\log L = \sum_j n_j F_j(1-F_j).$$

Each further differentiation with respect to β inserts one factor x_j, so

$$-\frac{\partial^2}{\partial\alpha\,\partial\beta}\log L = \sum_j x_j n_j F_j(1-F_j),$$
$$-\frac{\partial^2}{\partial\beta^2}\log L = \sum_j x_j^2 n_j F_j(1-F_j).$$

Substituting these four entries into the definition (12.3.9) of $I(\theta_1,\theta_2)$ gives (12.3.10). No expectation step is needed: the second derivatives contain no y_j^*, so observed and expected information coincide, as the book remarks below (12.3.10).

Problem 12.20. Consider a logistic regression model and assume $\beta = 0$.
(a) If $0 < \sum_{i=1}^{n} y_i < n$, show that the MLE of $\pi(x)$ (which does not depend on x in this case) is

$\hat{\pi} = \sum_{i=1}^n y_i/n$.
(b) If $0 < \sum_{i=1}^n y_i < n$, show that the MLE of α is

$$\hat{\alpha}_0 = \log\left(\frac{\sum_{i=1}^n y_i}{n - \sum_{i=1}^n y_i}\right).$$

(c) Show that if $\sum_{i=1}^n y_i = 0$ or n, $\hat{\alpha}_0$ does not exist, but the LRT statistic for testing $H_0 : \beta = 0$ is still well defined.

Solution. With $\beta = 0$, (12.3.2) makes $\pi(x) = e^\alpha/(1+e^\alpha) = \pi$ free of x, so the Y_i are iid Bernoulli(π) and, with $s = \sum_{i=1}^n y_i$,

$$L(\alpha, 0|\mathbf{y}) = \prod_{i=1}^n \pi^{y_i}(1-\pi)^{1-y_i} = \pi^s(1-\pi)^{n-s}.$$

(a) As α ranges over $\mathbb{R}$, $\pi = e^\alpha/(1+e^\alpha)$ ranges over all of $(0,1)$, so maximizing L in α is the same as maximizing $\pi^s(1-\pi)^{n-s}$ over $\pi \in (0,1)$. Setting the derivative of the log likelihood to zero,

$$\frac{d}{d\pi}\left[s\log\pi + (n-s)\log(1-\pi)\right] = \frac{s}{\pi} - \frac{n-s}{1-\pi} = 0 \quad \Longrightarrow \quad \pi = \frac{s}{n},$$

and when $0 < s < n$ this root lies in $(0,1)$ and is a maximum, since the second derivative $-s/\pi^2 - (n-s)/(1-\pi)^2$ is negative throughout. Hence $\hat{\pi} = s/n = \sum_i y_i/n$ (Example 7.2.7).
(b) The map $\alpha \mapsto \pi = e^\alpha/(1+e^\alpha)$ is a bijection of $\mathbb{R}$ onto $(0,1)$ with inverse $\alpha = \log(\pi/(1-\pi))$. By the invariance property of MLEs (Theorem 7.2.10),

$$\hat{\alpha}_0 = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \log\left(\frac{s/n}{1-s/n}\right) = \log\left(\frac{\sum_{i=1}^n y_i}{n - \sum_{i=1}^n y_i}\right),$$

which is finite exactly because $0 < s < n$.
(c) Nonexistence. If $s = 0$ then $L(\alpha, 0|\mathbf{y}) = (1-\pi)^n$, strictly decreasing in π and hence strictly decreasing in α; its supremum 1 is approached only as $\alpha \to -\infty$ and is attained at no finite α. If $s = n$ then $L = \pi^n$ is strictly increasing in α, supremum 1 approached only as $\alpha \to +\infty$. Either way $\hat{\alpha}_0$ does not exist.
The LRT statistic survives because Definition 8.2.1 uses suprema, not maxima:

$$\lambda(\mathbf{y}) = \frac{\sup_\alpha L(\alpha, 0|\mathbf{y})}{\sup_{\alpha,\beta} L(\alpha, \beta|\mathbf{y})},$$

and both suprema are finite, being suprema of a likelihood bounded above by 1. Explicitly, take $s = 0$ (the case $s = n$ is symmetric under $y_i \mapsto 1 - y_i$). The numerator is 1 by the paragraph above. For the denominator, $L(\alpha, \beta|\mathbf{y}) = \prod_i (1 - F_i) \le 1$, while the choice $\beta = 0$, $\alpha \to -\infty$ already drives it to 1; hence the denominator is 1 too, and

$$\lambda(\mathbf{y}) = 1, \qquad -2\log\lambda(\mathbf{y}) = 0.$$

So the statistic is well defined, and equal to 0, even though neither the restricted nor the unrestricted maximizer exists.

Problem 12.21. Let $Y \sim$ binomial(n, π), and let $\hat{\pi} = Y/n$ denote the MLE of π. Let $W = \log(\hat{\pi}/(1-\hat{\pi}))$ denote the sample logit, the MLE of $\log(\pi/(1-\pi))$. Use the Delta Method to show that $1/(n\hat{\pi}(1-\hat{\pi}))$ is a reasonable estimate of $\operatorname{Var} W$.

Solution. $\operatorname{Var} W \approx 1/(n\pi(1-\pi))$, by the Delta Method (Theorem 5.5.24) applied to $g(p) =$

$\log(p/(1-p))$, whose derivative

$$g'(p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

exists and is nonzero at every $p \in (0,1)$, so the theorem's hypothesis holds at $p = \pi$ (W itself is defined on the event $0 < Y < n$, whose probability tends to 1).
Writing $\hat\pi = \bar X_n$ for $X_1, \dots, X_n$ iid Bernoulli(π), whose mgf exists on all of $\mathbb{R}$, Theorem 5.5.14 gives

$$\sqrt{n}\,(\hat\pi - \pi) \;\longrightarrow\; \mathrm{n}(0, \pi(1-\pi)) \quad \text{in distribution,}$$

so (5.5.10) yields

$$\begin{aligned}\sqrt{n}\left[W - \log\!\left(\tfrac{\pi}{1-\pi}\right)\right] &\longrightarrow \mathrm{n}\Big(0,\ \pi(1-\pi)\left[g'(\pi)\right]^2\Big)\\ &= \mathrm{n}\left(0,\ \frac{\pi(1-\pi)}{\pi^2(1-\pi)^2}\right) = \mathrm{n}\left(0,\ \frac{1}{\pi(1-\pi)}\right).\end{aligned}$$

In the variance-approximation form (5.5.9) with $k = 1$, using $\operatorname{Var}\hat\pi = \pi(1-\pi)/n$,

$$\operatorname{Var} W \;\approx\; \left[g'(\pi)\right]^2 \operatorname{Var}\hat\pi = \frac{1}{\pi^2(1-\pi)^2} \cdot \frac{\pi(1-\pi)}{n} = \frac{1}{n\pi(1-\pi)}.$$

Since $\hat\pi \to \pi$ in probability (WLLN, Theorem 5.5.2; the X_i are bounded, so $\operatorname{Var} X_i < \infty$) and $p \mapsto 1/(p(1-p))$ is continuous on $(0,1)$, Theorem 5.5.4 gives

$$n\,\hat\pi(1-\hat\pi) \big/ n\pi(1-\pi) \;\longrightarrow\; 1 \quad \text{in probability,}$$

so substituting $\hat\pi$ for π changes the approximation by a factor tending to 1 and

$$\widehat{\operatorname{Var}}\, W = \frac{1}{n\hat\pi(1-\hat\pi)}$$

is a reasonable (consistent) estimate of $\operatorname{Var} W$.

12.4 Exercises 12.22–12.28

Problem 12.22. In Example 12.4.1 we examined how small perturbations affected the least squares estimate of slope. Perform the analogous calculation and assess the robustness (to small perturbations) of the least squares estimate of intercept.
(Recall the setting of Example 12.4.1: $Y_i = \alpha + \beta x_i + \varepsilon_i$ with the ε_i uncorrelated, $\mathrm{E}\,\varepsilon_i = 0$, and the contaminated variance

$$\operatorname{Var}(\varepsilon_i) = \begin{cases} \sigma^2 & \text{with probability } 1-\delta,\\ \tau^2 & \text{with probability } \delta,\end{cases}$$

for which the least squares slope $b = \sum d_i Y_i$, $d_i = (x_i - \bar x)/\sum (x_i - \bar x)^2$, was shown to satisfy $\operatorname{Var}(b) = [(1-\delta)\sigma^2 + \delta\tau^2]/\sum_{i=1}^n (x_i - \bar x)^2$.)

Solution. The intercept is exactly as robust as the slope:

$$\operatorname{Var}(a) = \left[(1-\delta)\sigma^2 + \delta\tau^2\right]\left[\frac{1}{n} + \frac{\bar x^2}{\sum_{i=1}^n (x_i - \bar x)^2}\right],$$

the same contamination factor $(1-\delta)\sigma^2 + \delta\tau^2$ multiplying the uncontaminated variance.

Write $a = \bar{Y} - b\bar{x} = \sum_{i=1}^n c_i Y_i$ with $c_i = \frac{1}{n} - \bar{x}\, d_i$. Since $\sum d_i = 0$ and $\sum d_i^2 = 1/\sum (x_i - \bar{x})^2$,

$$\begin{aligned}\sum_{i=1}^n c_i^2 &= \sum_{i=1}^n \frac{1}{n^2} - \frac{2\bar{x}}{n}\sum_{i=1}^n d_i + \bar{x}^2 \sum_{i=1}^n d_i^2 \\ &= \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.\end{aligned}$$

The ε_i are uncorrelated with common variance $\operatorname{Var}(\varepsilon_i) = (1-\delta)\sigma^2 + \delta\tau^2$ (the mixture has mean 0 in both components), so $\operatorname{Var}(a) = \sum_i c_i^2 \operatorname{Var}(\varepsilon_i)$ gives the display above.
Comparing to the uncontaminated case $\delta = 0$,

$$\frac{\operatorname{Var}_\delta(a)}{\operatorname{Var}_0(a)} = (1-\delta) + \delta\frac{\tau^2}{\sigma^2} = 1 + \delta\left(\frac{\tau^2}{\sigma^2} - 1\right),$$

exactly the inflation factor found for b in Example 12.4.1. The distortion is $O(\delta)$ for fixed τ^2, so the intercept is no less robust to small perturbations than the slope.

Problem 12.23. In Example 12.4.1, in contrast to Example 10.2.1, when we introduced the contaminated distribution for ε_i, we did not introduce a bias. Show that if we had, it would not have mattered. That is, if we assume

$$(\mathrm{E}\,\varepsilon_i, \operatorname{Var}\varepsilon_i) = \begin{cases}(0, \sigma^2) & \text{with probability } 1-\delta, \\ (\mu, \tau^2) & \text{with probability } \delta,\end{cases}$$

then:
(a) the least squares estimator b would still be an unbiased estimator of β.
(b) the least squares estimator a has expectation $\alpha + \delta\mu$, so the model may just as well be assumed to be $Y_i = \alpha + \delta\mu + \beta x_i + \varepsilon_i$.

Solution. The contamination shifts every error by the same amount, and a common shift is invisible to a fitted intercept.
Under the stated mixture $\mathrm{E}\,\varepsilon_i = (1-\delta)\cdot 0 + \delta\mu = \delta\mu$ for every i, so $\mathrm{E}\,Y_i = \alpha + \delta\mu + \beta x_i$.
(a) With $d_i = (x_i - \bar{x})/\sum_j (x_j - \bar{x})^2$ we have $b = \sum_i d_i Y_i$, and $\sum_i d_i = 0$, $\sum_i d_i x_i = 1$. Hence

$$\begin{aligned}\mathrm{E}\,b &= \sum_{i=1}^n d_i(\alpha + \delta\mu + \beta x_i) \\ &= (\alpha + \delta\mu)\sum_{i=1}^n d_i + \beta\sum_{i=1}^n d_i x_i = \beta.\end{aligned}$$

The bias $\delta\mu$ is a constant and is annihilated by $\sum_i d_i = 0$; b remains unbiased.
(b) $a = \bar{Y} - b\bar{x}$, so by (a)

$$\mathrm{E}\,a = (\alpha + \delta\mu + \beta\bar{x}) - \beta\bar{x} = \alpha + \delta\mu.$$

Thus a estimates $\alpha + \delta\mu$, not α. Setting $e_i = \varepsilon_i - \delta\mu$, which has $\mathrm{E}\,e_i = 0$ and $\operatorname{Var}(e_i) = \operatorname{Var}(\varepsilon_i)$, the model is

$$Y_i = (\alpha + \delta\mu) + \beta x_i + e_i,$$

a simple linear regression with mean-zero errors, unchanged slope and unchanged error variances, whose intercept is the unknown constant $\alpha + \delta\mu$.

Problem 12.24. For the model $Y_i = \beta x_i + \varepsilon_i$, show that the LAD estimator is given by $t_{(k^*+1)}$,

where $t_i = y_i/x_i$, $t_{(1)} \le \cdots \le t_{(n)}$ and, if $x_{(i)}$ is the x value paired with $t_{(i)}$, k^* satisfies

$$\sum_{i=1}^{k^*} |x_{(i)}| \le \sum_{i=k^*+1}^{n} |x_{(i)}| \quad \text{and} \quad \sum_{i=1}^{k^*+1} |x_{(i)}| > \sum_{i=k^*+2}^{n} |x_{(i)}|.$$

Solution. The LAD problem for $Y_i = \beta x_i + \varepsilon_i$ is a **weighted median** problem: minimizing $\sum_i |y_i - \beta x_i|$ is minimizing $\sum_i |x_i|\,|t_i - \beta|$, and $t_{(k^*+1)}$ is the weighted median of the t_i with weights $|x_i|$.
Assume $x_i \neq 0$ for all i (an observation with $x_i = 0$ contributes the constant $|y_i|$ and may be discarded). Since $y_i - \beta x_i = x_i(t_i - \beta)$,

$$L(\beta) = \sum_{i=1}^{n} |y_i - \beta x_i| = \sum_{i=1}^{n} |x_i|\,|t_i - \beta|,$$

which is convex and piecewise linear in β with kinks exactly at the points $t_{(1)} \le \cdots \le t_{(n)}$. On the open interval $(t_{(k)}, t_{(k+1)})$ its slope is constant and equal to

$$L'(\beta) = \sum_{i=1}^{k} |x_{(i)}| - \sum_{i=k+1}^{n} |x_{(i)}| \;=:\; S_k,$$

with the conventions $S_0 = -\sum_{i=1}^{n} |x_{(i)}|$ on $(-\infty, t_{(1)})$ and $S_n = +\sum_{i=1}^{n} |x_{(i)}|$ on $(t_{(n)}, \infty)$. Because

$$S_k - S_{k-1} = 2|x_{(k)}| \ge 0,$$

the slope sequence $S_0 \le S_1 \le \cdots \le S_n$ is nondecreasing, so L is decreasing up to the first index at which the slope turns positive and increasing thereafter. Hence L is minimized at the kink $t_{(k+1)}$ where the slope changes sign, i.e. at the index k^* characterized by

$$S_{k^*} \le 0 < S_{k^*+1},$$

and by definition of S_k these two inequalities read

$$\sum_{i=1}^{k^*} |x_{(i)}| \le \sum_{i=k^*+1}^{n} |x_{(i)}|, \qquad \sum_{i=1}^{k^*+1} |x_{(i)}| > \sum_{i=k^*+2}^{n} |x_{(i)}|.$$

(Such a k^* exists because $S_0 < 0 < S_n$, provided the $|x_{(i)}|$ are not all 0.) Therefore $\hat{\beta}_L = t_{(k^*+1)}$; when $S_{k^*} = 0$ the whole interval $[t_{(k^*)}, t_{(k^*+1)}]$ minimizes L and $t_{(k^*+1)}$ is the stated choice (Exercise 12.25).

Problem 12.25. A problem with the LAD regression line is that it is not always uniquely defined.
(a) Show that, for a data set with three observations, (x_1, y_1), (x_1, y_2), and (x_3, y_3) (note the first two xs are the same), any line that goes through (x_3, y_3) and lies between (x_1, y_1) and (x_1, y_2) is a least absolute deviation line.
(b) For three individuals, measurements are taken on heart rate (x, in beats per minute) and oxygen consumption (y, in ml/kg). The (x, y) pairs are $(127, 14.4)$, $(127, 11.9)$, and $(136, 17.9)$. Calculate the slope and intercept of the least squares line and the range of the least absolute deviation lines.

Solution. (a) Every such line attains the value $|y_2 - y_1|$, which is a lower bound for the LAD criterion over all lines.
Write a candidate line $y = a + bx$ by the two values it takes at the two distinct abscissas, $v = a + bx_1$ and $u = a + bx_3$; since $x_1 \neq x_3$ the map $(a, b) \mapsto (v, u)$ is a bijection onto $\mathbb{R}^2$. The criterion is

$$\begin{aligned} D(a, b) &= |y_1 - v| + |y_2 - v| + |y_3 - u| \\ &\ge |y_1 - v| + |y_2 - v| \;\ge\; |y_2 - y_1|, \end{aligned}$$

the last step by the triangle inequality. Equality holds in the second inequality precisely when $u = y_3$, and in the third precisely when v lies between y_1 and y_2. Hence $\min D = |y_2 - y_1|$ and the minimizers are exactly the lines through (x_3, y_3) whose value at $x = x_1$ lies in the interval $[\,\min(y_1, y_2), \max(y_1, y_2)\,]$ – that is, the lines through (x_3, y_3) lying between (x_1, y_1) and (x_1, y_2). Whenever $y_1 \neq y_2$ this is a one-parameter family, so the LAD line is not unique.
(b) Here $x_1 = 127$ with $y_1 = 14.4$, $y_2 = 11.9$, and $(x_3, y_3) = (136, 17.9)$.
Least squares: with $\bar{x} = 130$, $\bar{y} = 221/15 = 14.7\overline{3}$,

$$S_{xx} = (-3)^2 + (-3)^2 + 6^2 = 54,$$
$$S_{xy} = (-3)(-\tfrac{1}{3}) + (-3)(-\tfrac{17}{6}) + 6(\tfrac{19}{6}) = \tfrac{57}{2},$$

so

$$\hat{\beta} = \frac{57/2}{54} = \frac{19}{36} \approx 0.5278, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = -\frac{4849}{90} \approx -53.878.$$

LAD range: by part (a) the LAD lines are those joining $(136, 17.9)$ to $(127, v)$ for $11.9 \le v \le 14.4$, with

$$b(v) = \frac{17.9 - v}{136 - 127} = \frac{17.9 - v}{9}, \qquad a(v) = v - 127\, b(v).$$

The endpoints are

$$v = 14.4: \quad b = \tfrac{7}{18}, \quad a = -\tfrac{3149}{90} \approx -34.99,$$
$$v = 11.9: \quad b = \tfrac{2}{3}, \quad a = -\tfrac{2183}{30} \approx -72.77.$$

Since $b(v)$ is decreasing in v and $a(v)$ is increasing in v, the LAD lines are exactly those with

$$\text{slope} \in \left[\tfrac{7}{18}, \tfrac{2}{3}\right] \approx [0.389, 0.667],$$
$$\text{intercept} \in \left[-\tfrac{2183}{30}, -\tfrac{3149}{90}\right] \approx [-72.77, -34.99],$$

the two ranges being traversed together as v runs over $[11.9, 14.4]$.

Problem 12.26. (a) Throughout Example 12.4.3 we assumed that $\frac{1}{n}\sum_{i=1}^n x_i^2 \to \sigma_x^2 < \infty$. Show that this condition is satisfied by (i) $x_i = 1$ (the case of the ordinary median) and (ii) $|x_i| \le 1$ (the case of bounded x_i).
(b) Show that, under the conditions on x_i in part (a), $\frac{1}{n}\sum_{i=1}^n \psi(y_i - \hat{\beta}_L x_i) \to 0$ in probability, where, as in Example 12.4.3,

$$\psi(y_i - \beta x_i) = x_i I(y_i > \beta x_i) - x_i I(y_i < \beta x_i)$$

and $\hat{\beta}_L$ is the LAD estimator in the model $Y_i = \beta x_i + \varepsilon_i$.

Solution. (a) (i) $\frac{1}{n}\sum_{i=1}^n 1^2 = 1$ for every n, so $\sigma_x^2 = 1 < \infty$.
(ii) If $|x_i| \le 1$ then $0 \le \frac{1}{n}\sum_{i=1}^n x_i^2 \le 1$ for every n, so the averages stay in $[0, 1]$ and the limit σ_x^2, whenever it exists, satisfies $\sigma_x^2 \le 1 < \infty$. (Boundedness supplies the finiteness; existence of the limit is what the design assumption itself contributes, and along a subsequence it is automatic by Bolzano-Weierstrass.)
(b) The bound is almost sure and of order $1/n$. Write $t_i = y_i/x_i$ (discard any $x_i = 0$, which contributes nothing to ψ). Since $y_i - \beta x_i = x_i(t_i - \beta)$,

$$\psi(y_i - \beta x_i) = x_i \operatorname{sgn}\big(x_i(t_i - \beta)\big) = |x_i| \operatorname{sgn}(t_i - \beta),$$

so with $L(\beta) = \sum_i |y_i - \beta x_i| = \sum_i |x_i||t_i - \beta|$ as in Exercise 12.24,

$$\sum_{i=1}^n \psi(y_i - \beta x_i) = \sum_{t_i > \beta} |x_i| - \sum_{t_i < \beta} |x_i|$$

is exactly $-L'(\beta)$ at every β that is not one of the t_i. Let $J = \{i : t_i = \hat{\beta}_L\}$ and $W = \sum_{i \in J} |x_i|$. The

one-sided derivatives of the convex function L at its minimizer satisfy $L'(\hat\beta_L-) \le 0 \le L'(\hat\beta_L+)$, and

$$L'(\hat\beta_L\mp) = -\sum_{i=1}^n \psi(y_i - \hat\beta_L x_i) \mp W,$$

so combining the two inequalities gives the deterministic bound

$$\left|\sum_{i=1}^n \psi(y_i - \hat\beta_L x_i)\right| \le W = \sum_{i\in J} |x_i| \le |J| \max_i |x_i| \le |J|,$$

using $|x_i| \le 1$ (case (i) is the special case $|x_i| = 1$). Because the ε_i have a continuous distribution, the $t_i = \beta + \varepsilon_i/x_i$ are distinct with probability 1, so $|J| = 1$ almost surely and

$$\left|\frac{1}{n}\sum_{i=1}^n \psi(y_i - \hat\beta_L x_i)\right| \le \frac{1}{n} \longrightarrow 0.$$

Convergence to 0 with probability 1 implies convergence in probability (Section 5.5.2).

Problem 12.27. In the setting of Example 12.4.3, with $Y_i = \beta x_i + \varepsilon_i$, the ε_i iid with cdf F, density f and median 0 (so $F(0) = 1/2$), $\psi(y_i - \beta x_i) = x_i I(y_i > \beta x_i) - x_i I(y_i < \beta x_i)$, and $\frac{1}{n}\sum_{i=1}^n x_i^2 \to \sigma_x^2 < \infty$ with $|x_i| \le 1$:

(a) Verify that $\dfrac{-1}{\sqrt{n}}\displaystyle\sum_{i=1}^n \psi(Y_i - \hat\beta x_i) \to \mathrm{n}(0, \sigma_x^2)$.

(b) Verify that

$$\frac{1}{n}\sum_{i=1}^n \frac{d}{d\beta_0}\mathrm{E}_\beta\big[\psi(Y_i - \beta_0 x_i)\big]\Big|_{\beta_0=\beta} = 2f(0)\,\frac{1}{n}\sum_{i=1}^n x_i^2,$$

and, with part (a), conclude that

$$\sqrt{n}\,(\hat\beta_L - \beta) \to \mathrm{n}\left(0, \frac{1}{4f(0)^2\sigma_x^2}\right).$$

Solution. (a) The summands are independent, bounded and mean zero with $\operatorname{Var}\psi(Y_i - \beta x_i) = x_i^2$, so Lindeberg-Feller applies. (As in (12.4.3), the $\hat\beta$ printed in the statement is the true β of the numerator of (12.4.2); the leading minus sign is immaterial since the limit law is symmetric.)
Since $Y_i - \beta x_i = \varepsilon_i$, the identity of Exercise 12.26 gives

$$\psi(Y_i - \beta x_i) = x_i \operatorname{sgn}(\varepsilon_i),$$

and $\mathrm{E}\operatorname{sgn}(\varepsilon_i) = P(\varepsilon_i > 0) - P(\varepsilon_i < 0) = 0$ because $F(0) = 1/2$, while $\mathrm{E}\operatorname{sgn}^2(\varepsilon_i) = 1$ (ε_i is continuous, so $P(\varepsilon_i = 0) = 0$). Hence

$$\mathrm{E}\,\psi(Y_i - \beta x_i) = 0, \qquad \operatorname{Var}\psi(Y_i - \beta x_i) = x_i^2,$$

and $s_n^2 := \sum_{i=1}^n x_i^2 = n\big(\sigma_x^2 + o(1)\big) \to \infty$ (assuming $\sigma_x^2 > 0$). The summands obey $|x_i \operatorname{sgn}(\varepsilon_i)| \le 1$, so for any $\eta > 0$ the event $\{|x_i \operatorname{sgn}(\varepsilon_i)| > \eta s_n\}$ is empty once $s_n > 1/\eta$, and the Lindeberg condition holds trivially. By the Lindeberg-Feller central limit theorem (named in Miscellanea 5.8.1),

$$\frac{1}{s_n}\sum_{i=1}^n \psi(Y_i - \beta x_i) \to \mathrm{n}(0, 1),$$

and since $s_n/\sqrt{n} \to \sigma_x$, Slutsky's Theorem (Theorem 5.5.17) gives

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(Y_i - \beta x_i) = \frac{s_n}{\sqrt{n}} \cdot \frac{1}{s_n}\sum_{i=1}^n \psi(Y_i - \beta x_i) \to \mathrm{n}(0, \sigma_x^2).$$

(b) For any β_0, using $Y_i = \beta x_i + \varepsilon_i$,

$$\begin{aligned}\mathrm{E}_\beta\big[\psi(Y_i - \beta_0 x_i)\big] &= x_i\big[P_\beta(Y_i > \beta_0 x_i) - P_\beta(Y_i < \beta_0 x_i)\big] \\ &= x_i\big[1 - F\big((\beta_0 - \beta)x_i\big)\big] - x_i F\big((\beta_0 - \beta)x_i\big) \\ &= x_i\big[1 - 2F\big((\beta_0 - \beta)x_i\big)\big],\end{aligned}$$

since $P_\beta(\varepsilon_i > (\beta_0 - \beta)x_i) = 1 - F((\beta_0 - \beta)x_i)$ whatever the sign of x_i. Differentiating,

$$\frac{d}{d\beta_0}\mathrm{E}_\beta\big[\psi(Y_i - \beta_0 x_i)\big] = -2x_i^2 f\big((\beta_0 - \beta)x_i\big),$$

so at $\beta_0 = \beta$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \frac{d}{d\beta_0}\mathrm{E}_\beta\big[\psi(Y_i - \beta_0 x_i)\big]\Big|_{\beta_0=\beta} &= -2f(0)\,\frac{1}{n}\sum_{i=1}^n x_i^2 \\ &\longrightarrow -2f(0)\sigma_x^2.\end{aligned}$$

(The printed statement, like the book's own display (12.4.4), drops the minus sign; it cancels below, so the stated conclusion is unaffected.)
Now combine. Equation (12.4.2) reads

$$\sqrt{n}(\hat\beta_L - \beta) = \frac{-\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(Y_i - \beta x_i)}{\frac{1}{n}\frac{d}{d\hat\beta_L}\sum_{i=1}^n \psi(Y_i - \hat\beta_L x_i)\big|_{\hat\beta_L=\beta}},$$

its left-hand side having been set to 0 on the strength of Exercise 12.26(b). Replacing the denominator by its expectation as in (12.4.4) (the law of large numbers, Theorem 5.5.2, applied before differentiating) makes it $-2f(0)\sigma_x^2 \neq 0$, so by part (a) and Slutsky,

$$\sqrt{n}(\hat\beta_L - \beta) \to \frac{\mathrm{n}(0,\sigma_x^2)}{2f(0)\sigma_x^2} = \mathrm{n}\left(0, \frac{\sigma_x^2}{4f(0)^2\sigma_x^4}\right) = \mathrm{n}\left(0, \frac{1}{4f(0)^2\sigma_x^2}\right),$$

which is (12.4.5).

Problem 12.28. In the setting of Example 12.4.3 ($Y_i = \beta x_i + \varepsilon_i$, the ε_i iid with mean 0 and variance σ^2, $|x_i| \le 1$ and $\frac{1}{n}\sum_{i=1}^n x_i^2 \to \sigma_x^2 < \infty$), show that the least squares estimator is given by

$$\hat\beta = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}, \qquad \text{and} \qquad \sqrt{n}\,(\hat\beta - \beta) \to \mathrm{n}\left(0, \frac{1}{\sigma_x^2}\right).$$

Solution. $\hat\beta = \sum_i x_i y_i / \sum_i x_i^2$ is the unique stationary point of a strictly convex quadratic.
Minimizing $Q(\beta) = \sum_{i=1}^n (y_i - \beta x_i)^2$,

$$Q'(\beta) = -2\sum_{i=1}^n x_i(y_i - \beta x_i) = 0 \iff \hat\beta = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2},$$

and $Q''(\beta) = 2\sum_i x_i^2 > 0$, so this is the minimum.
Substituting $Y_i = \beta x_i + \varepsilon_i$ gives the error representation

$$\sqrt{n}(\hat\beta - \beta) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i\varepsilon_i}{\frac{1}{n}\sum_{i=1}^n x_i^2}.$$

The numerator has independent, mean-zero summands with $\mathrm{Var}(x_i\varepsilon_i) = \sigma^2 x_i^2$, so $s_n^2 = \sigma^2\sum_{i=1}^n x_i^2 =$

$n\sigma^2(\sigma_x^2+o(1))\to\infty$. Because $|x_i|\le 1$, the event $\{|x_i\varepsilon_i|>\eta s_n\}$ is contained in $\{|\varepsilon_i|>\eta s_n\}$, whence

$$\frac{1}{s_n^2}\sum_{i=1}^{n}\mathrm{E}\big[x_i^2\varepsilon_i^2 I(|x_i\varepsilon_i|>\eta s_n)\big]\le\frac{\sum_{i=1}^n x_i^2}{s_n^2}\,\mathrm{E}\big[\varepsilon^2 I(|\varepsilon|>\eta s_n)\big]$$
$$=\frac{1}{\sigma^2}\,\mathrm{E}\big[\varepsilon^2 I(|\varepsilon|>\eta s_n)\big]\to 0$$

by dominated convergence, since $\mathrm{E}\,\varepsilon^2=\sigma^2<\infty$ and $s_n\to\infty$. The Lindeberg condition holds, so by the Lindeberg-Feller central limit theorem (named in Miscellanea 5.8.1) and $s_n/\sqrt{n}\to\sigma\sigma_x$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i\varepsilon_i\to\mathrm{n}(0,\sigma^2\sigma_x^2).$$

The denominator is the constant sequence $\frac{1}{n}\sum_i x_i^2\to\sigma_x^2\neq 0$, so by Slutsky's Theorem (Theorem 5.5.17)

$$\sqrt{n}(\hat\beta-\beta)\to\mathrm{n}\left(0,\frac{\sigma^2\sigma_x^2}{\sigma_x^4}\right)=\mathrm{n}\left(0,\frac{\sigma^2}{\sigma_x^2}\right),$$

which is the stated $\mathrm{n}(0,1/\sigma_x^2)$ under the normalization $\sigma^2=\mathrm{Var}\,\varepsilon_i=1$ left implicit in Example 12.4.3 (and in Table 10.2.1, whose AREs it reproduces).

12.5 Exercises 12.29–12.31

Problem 12.29. Using a Taylor series argument as in Example 12.4.3, derive the asymptotic distribution of the median in iid sampling.

Solution. If $X_1,\dots,X_n$ are iid with cdf F and density f, and θ is the unique median ($F(\theta)=1/2$, $f(\theta)>0$), then the sample median M_n satisfies

$$\sqrt{n}\,(M_n-\theta)\;\to\;\mathrm{n}\left(0,\;\frac{1}{4f(\theta)^2}\right).$$

This is Example 12.4.3 run with $x_i\equiv 1$, case (i) of Exercise 12.26(a): $Y_i=\beta x_i+\varepsilon_i$ becomes $X_i=\theta+\varepsilon_i$, the criterion $\sum_i|y_i-\beta x_i|$ becomes $\sum_i|X_i-\theta|$ with minimizer the median, and $\sigma_x^2=\lim_n\frac{1}{n}\sum_i x_i^2=1$, so $f(0)$ there is the data density at the median. In detail, $\rho(t)=|t|$ gives $\psi=\rho'$,

$$\psi(X_i-\theta)=I(X_i>\theta)-I(X_i<\theta),$$

and M_n solves $\sum_i\psi(X_i-\hat\theta)=0$ up to a term that vanishes in probability (Exercise 12.26(b) with $x_i\equiv 1$), which is (12.4.1). Taylor expanding about θ and rearranging as in (12.4.2),

$$\sqrt{n}\,(\hat\theta-\theta)=\frac{-\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi(X_i-\theta)}{\frac{1}{n}\frac{d}{dt}\sum_{i=1}^n\psi(X_i-t)\big|_{t=\theta}}.$$

Here $\mathrm{E}_\theta\,\psi(X_i-\theta)=P(X_i>\theta)-P(X_i<\theta)=0$ and $\psi^2\equiv 1$, so $\mathrm{Var}\,\psi(X_i-\theta)=1$ and the Central Limit Theorem gives

$$\frac{-1}{\sqrt{n}}\sum_{i=1}^{n}\psi(X_i-\theta)\;\to\;\mathrm{n}(0,1),$$

which is (12.4.3) with $\sigma_x^2=1$. Since ψ is not differentiable at 0, the denominator is handled as in (12.4.4), averaging first and then differentiating:

$$\frac{1}{n}\frac{d}{dt}\sum_{i=1}^{n}\psi(X_i-t)\approx\frac{d}{dt}\,\mathrm{E}_\theta\big[I(X_1>t)-I(X_1<t)\big]$$
$$=\frac{d}{dt}\big[1-2F(t)\big]=-2f(t),$$

so the denominator tends to the nonrandom nonzero limit $-2f(\theta)$. ((12.4.4) prints these two signs positive; they cancel the -1 in the numerator, so the symmetric limit is unchanged.) Slutsky's Theorem (Theorem 5.5.17) then gives

$$\sqrt{n}\,(M_n - \theta) \;\to\; \frac{\mathrm{n}(0,1)}{2f(\theta)} = \mathrm{n}\!\left(0, \frac{1}{4f(\theta)^2}\right),$$

which is (12.4.5) at $\sigma_x^2 = 1$.

Problem 12.30. For the data of Table 12.4.1, use the parametric bootstrap to assess the standard error from the LAD and M-estimator fit. In particular:
(a) Fit the line $y = \alpha + \beta x$ to get estimates $\hat\alpha$ and $\hat\beta$.
(b) Calculate the residual mean squared error

$$\hat\sigma^2 = \frac{1}{n-2}\sum_{i=1}^{n}\big[y_i - (\hat\alpha + \hat\beta x_i)\big]^2.$$

(c) Generate new residuals from $\mathrm{n}(0, \hat\sigma^2)$ and re-estimate α and β.
(d) Do part (c) B times and calculate the standard deviation of $\hat\alpha$ and $\hat\beta$.
(e) Repeat parts (a)–(d) using both the double exponential and Laplace distributions for the errors. Compare your answers to the normal.
Table 12.4.1 gives the values of CO_2 and O_2 in the pouches of 23 potoroos (McPherson 1990); the regression of interest is CO_2 (y) on O_2 (x).

Animal	1	2	3	4	5	6	7	8
% O_2	20	19.6	19.6	19.4	18.4	19	19	18.3
% CO_2	1	1.2	1.1	1.4	2.3	1.7	1.7	2.4
Animal	9	10	11	12	13	14	15	16
% O_2	18.2	18.6	19.2	18.2	18.7	18.5	18	17.4
% CO_2	2.1	2.1	1.2	2.3	1.9	2.4	2.6	2.9
Animal	17	18	19	20	21	22	23	
% O_2	16.5	17.2	17.3	17.8	17.3	18.4	16.9	
% CO_2	4.0	3.3	3.0	3.4	2.9	1.9	3.9	

Solution. Part (e) is an erratum: the double exponential and the Laplace are the same distribution, and the third error law intended is the logistic, completing the trio of Example 12.4.5; that is what is fitted below.
(a) With $n = 23$ the three fits of Examples 12.4.2 and 12.4.4 are reproduced (least squares in closed form, LAD by minimizing $\sum_i |y_i - \alpha - \beta x_i|$, and the M-estimator by minimizing $\sum_i \rho_i(\alpha,\beta)$ of (12.4.6) with $k = 1.5\hat\sigma$):

$$\begin{aligned} \text{least squares:}\quad & y = 18.675 - 0.894x, \\ \text{LAD:}\quad & y = 18.589 - 0.889x, \\ \text{M-estimate } (k = 1.5\hat\sigma):\quad & y = 18.497 - 0.885x. \end{aligned}$$

(b) From the least squares fit,

$$\hat\sigma^2 = \frac{1}{21}\sum_{i=1}^{23}\big[y_i - (18.675 - 0.894x_i)\big]^2 = 0.05514, \qquad \hat\sigma = 0.2348,$$

the value .23 quoted in Example 12.4.4.
(c)–(d) The parametric bootstrap holds the design $x_1, \ldots, x_{23}$ and the fitted line fixed and resamples the errors: for $b = 1, \ldots, B$ draw $\varepsilon_1^*, \ldots, \varepsilon_n^*$ iid $\mathrm{n}(0, \hat\sigma^2)$, set

$$y_i^* = \hat\alpha + \hat\beta x_i + \varepsilon_i^*,$$

and refit by each of the three methods to get $(\hat\alpha^{*b}, \hat\beta^{*b})$. The bootstrap standard error is the square root of (10.1.11), the sample variance of the B values, with $\mathrm{n}(0, \hat\sigma^2)$ as the plug-in distribution of Example 10.1.22. With $B = 1000$:

errors	estimator	se($\hat\alpha$)	se($\hat\beta$)
normal	LS	0.995	0.0542
normal	LAD	1.227	0.0669
normal	M	1.001	0.0546
double exponential	LS	0.988	0.0538
double exponential	LAD	0.926	0.0505
double exponential	M	0.917	0.0499
logistic	LS	1.002	0.0546
logistic	LAD	1.107	0.0602
logistic	M	0.961	0.0523

(All three error laws are scaled to have standard deviation $\hat\sigma = 0.2348$, so the columns are comparable.) The normal row for least squares checks the simulation: the exact standard errors from Section 11.3 are

$$\hat\sigma\sqrt{\tfrac{1}{n} + \tfrac{\bar x^2}{S_{xx}}} = 0.985, \qquad \frac{\hat\sigma}{\sqrt{S_{xx}}} = 0.0537,$$

matching the bootstrap values 0.995 and 0.0542 to simulation error.

(e) Comparison through the ratios of squared standard errors for the slope, which estimate the reciprocal AREs of Definition 10.1.16:

ratio of variances	normal	double exponential	logistic
LAD vs. LS	1.52	0.88	1.22
M vs. LS	1.01	0.86	0.92
M vs. LAD	0.67	0.98	0.75

Under normal errors LAD pays the price predicted by Example 12.4.3, whose $\mathrm{ARE}(\hat\beta_L, \hat\beta) = 4f(0)^2 = 2/\pi \approx .64$ is a variance ratio $\pi/2 = 1.57$. Under double exponential errors LAD is the maximum likelihood procedure and beats least squares; under the logistic it sits between. The M-estimator costs about 1% against least squares under normal errors, improves on it under both heavier-tailed laws, and is never worse than LAD – the conclusion of Example 12.4.5, whose AREs are 0.98, 1.03, 1.07 versus least squares and 1.39, 1.27, 1.14 versus LAD.

Problem 12.31. For the data of Table 12.4.1 (the O_2/CO_2 readings on 23 potoroos, reproduced in Exercise 12.30), we could also use the nonparametric bootstrap to assess the standard error from the LAD and M-estimator fit.
(a) Fit the line $y = \alpha + \beta x$ to get estimates $\hat\alpha$ and $\hat\beta$.
(b) Generate new residuals by resampling from the fitted residuals and re-estimate α and β.
(c) Do part (b) B times and calculate the standard deviation of $\hat\alpha$ and $\hat\beta$.

Solution. The nonparametric (residual) bootstrap standard errors, $B = 4000$:

estimator	$\hat\alpha$	se($\hat\alpha$)	$\hat\beta$	se($\hat\beta$)
LS	18.675	0.993	-0.894	0.0541
LAD	18.589	0.908	-0.889	0.0494
M	18.497	0.936	-0.885	0.0510

(a) The three fits are those of Exercise 12.30(a), i.e. Examples 12.4.2 and 12.4.4: $y = 18.675 - 0.894x$ (least squares), $y = 18.589 - 0.889x$ (LAD), and $y = 18.497 - 0.885x$ (M-estimate with $k = 1.5\hat\sigma$, $\hat\sigma = 0.2348$).

(b) Only the error-generating step of Exercise 12.30 changes: instead of drawing ε_i^* from a named density, draw them with replacement from the $n = 23$ fitted residuals of the method being assessed. Writing $\hat r_i = y_i - \hat\alpha - \hat\beta x_i$ and $\bar{\hat r}$ for their mean, put

$$\varepsilon_1^*, \dots, \varepsilon_n^* \ \text{ iid from the uniform distribution on } \left\{\sqrt{\tfrac{n}{n-2}}\,(\hat r_i - \bar{\hat r})\right\}_{i=1}^{n},$$

set $y_i^* = \hat\alpha + \hat\beta x_i + \varepsilon_i^*$, and refit. This is the plug-in of Section 10.1.4 with F replaced by the empirical

cdf of the residuals; the recentering makes that cdf have mean 0 (needed for $\hat{\alpha}^*$ to be centered at $\hat{\alpha}$), and the factor $\sqrt{n/(n-2)}$ undoes the shrinkage $\mathrm{E}\sum \hat{r}_i^2 = (n-2)\sigma^2$ so the resampling distribution has variance $\hat{\sigma}^2$.
(c) Repeating $B = 4000$ times and taking sample standard deviations of $\{\hat{\alpha}^{*b}\}$ and $\{\hat{\beta}^{*b}\}$ gives the table above, whose slope variance ratios

$$\left(\frac{0.0494}{0.0541}\right)^2 = 0.83 \;\; \text{(LAD vs. LS)}, \qquad \left(\frac{0.0510}{0.0541}\right)^2 = 0.89 \;\; \text{(M vs. LS)},$$

reverse the normal-error verdict of Exercise 12.30 because the 23 fitted residuals are heavier-tailed than normal (excess kurtosis .91, largest residual $2.7\hat{\sigma}$).