Solutions to Gelman et al.'s Bayesian Data Analysis

Aayush Bajaj

2026-09-21T10:00:00+10:00

Contents

Solutions to every exercise in *Bayesian Data Analysis* (3rd edition, CRC Press, 2013) by Gelman, Carlin, Stern, Dunson, Vehtari and Rubin — 215 exercises across chapters 1–23. The book is filed at Bayesian Data Analysis.

1 Probability and Inference

Contents

1.1 Exercises 1.1–1.7

Problem 1.1. Conditional probability: suppose that if $\omega = 1$, then y has a normal distribution with mean 1 and standard deviation σ, and if $\omega = 2$, then y has a normal distribution with mean 2 and standard deviation σ. Also, suppose $\Pr(\omega = 1) = 0.5$ and $\Pr(\omega = 2) = 0.5$.
(a) For $\sigma = 2$, write the formula for the marginal probability density for y and sketch it.
(b) What is $\Pr(\omega = 1 \mid y = 1)$, again supposing $\sigma = 2$?
(c) Describe how the posterior density of ω changes in shape as σ is increased and as it is decreased.

Solution. (a) Averaging the conditional density over the two values of ω,

$$p(y) = \tfrac{1}{2}\mathrm{N}(y \mid 1, \sigma^2) + \tfrac{1}{2}\mathrm{N}(y \mid 2, \sigma^2),$$

so at $\sigma = 2$,

$$p(y) = \frac{1}{4\sqrt{2\pi}}\left[e^{-(y-1)^2/8} + e^{-(y-2)^2/8}\right].$$

The sketch is a single symmetric bump centred at $y = 1.5$, with $p(1.5) = 0.193$, visually indistinguishable from a normal curve: the component means are separated by $1 = \sigma/2$, far short of the 2σ separation needed for an equally weighted two-component normal mixture to be bimodal. (Check! $p''(1.5) < 0$.)
(b) $\Pr(\omega = 1 \mid y = 1) = 0.531$. By Bayes' rule the prior odds are 1, so the posterior odds are the likelihood ratio, and in general

$$\begin{aligned}\frac{\Pr(\omega = 1 \mid y)}{\Pr(\omega = 2 \mid y)} &= \frac{\exp\bigl(-(y-1)^2/(2\sigma^2)\bigr)}{\exp(-(y-2)^2/(2\sigma^2))} \\ &= \exp\left(\frac{3-2y}{2\sigma^2}\right),\end{aligned}$$

whence

$$\Pr(\omega = 1 \mid y) = \mathrm{logit}^{-1}\left(\frac{3-2y}{2\sigma^2}\right).$$

At $y = 1$, $\sigma = 2$ the exponent is $1/8$ and

$$\Pr(\omega = 1 \mid y = 1) = \frac{1}{1 + e^{-1/8}} = 0.5312.$$

(c) The posterior is the logistic curve above, evaluated at the argument $(3 - 2y)/(2\sigma^2)$, i.e. a logistic function of y with slope $-1/\sigma^2$ centred at the midpoint $y = 1.5$. As σ increases the slope flattens and $\Pr(\omega = 1 \mid y) \to \frac{1}{2}$ for every fixed y: the data become uninformative and the posterior reverts to the prior. As σ decreases the curve steepens to a step at $y = 1.5$, $\Pr(\omega = 1 \mid y) \to 1$ for $y < 1.5$ and $\to 0$ for $y > 1.5$: ω is identified by which mean y falls nearer.

Problem 1.2. Conditional means and variances: show that (1.8) and (1.9) hold if u is a vector. Here (1.8) and (1.9) are the identities of Section 1.8,

$$E(u) = E(E(u \mid v)),$$

$$\mathrm{var}(u) = E(\mathrm{var}(u \mid v)) + \mathrm{var}(E(u \mid v)),$$

in which the inner expectation averages over u conditional on v and the outer expectation averages over v; for a vector u the variance is the covariance matrix

$$\mathrm{var}(u) = \int (u - E(u))(u - E(u))^T p(u)\, du.$$

Solution. Both derivations in Section 1.8 go through verbatim, the scalar square u^2 being replaced by the outer product uu^T; assume $E(u^T u) < \infty$ so every entry below is finite.
For (1.8) the integrand is a vector and integration is entrywise, so the book's factorization applies to each component at once:

$$\begin{aligned} E(u) &= \iint u\, p(u,v)\, du\, dv \\ &= \int \left[\int u\, p(u \mid v)\, du \right] p(v)\, dv \\ &= \int E(u \mid v)\, p(v)\, dv = E(E(u \mid v)). \end{aligned}$$

Nothing above used the dimension of v either.
For (1.9), write both variances in the form $\operatorname{var}(w) = E(ww^T) - E(w)E(w)^T$ and expand:

$$\begin{aligned} &E(\operatorname{var}(u \mid v)) + \operatorname{var}(E(u \mid v)) \\ &\quad = E\big[E(uu^T \mid v) - E(u \mid v)E(u \mid v)^T\big] \\ &\qquad + E\big[E(u \mid v)E(u \mid v)^T\big] - E(u)E(u)^T \\ &\quad = E(uu^T) - E(u)E(u)^T = \operatorname{var}(u). \end{aligned}$$

The second-last equality is (1.8) applied entrywise to the matrix-valued uu^T, and (1.8) applied to u is what makes the outer variance's mean term $E(u)E(u)^T$.

Problem 1.3. Probability calculation for genetics (from Lindley, 1965): suppose that in each individual of a large population there is a pair of genes, each of which can be either x or X, that controls eye color: those with xx have blue eyes, while heterozygotes (those with Xx or xX) and those with XX have brown eyes. The proportion of blue-eyed individuals is p^2 and of heterozygotes is $2p(1-p)$, where $0 < p < 1$. Each parent transmits one of its own genes to the child; if a parent is a heterozygote, the probability that it transmits the gene of type X is $\frac{1}{2}$. Assuming random mating, show that among brown-eyed children of brown-eyed parents, the expected proportion of heterozygotes is $2p/(1+2p)$. Suppose Judy, a brown-eyed child of brown-eyed parents, marries a heterozygote, and they have n children, all brown-eyed. Find the posterior probability that Judy is a heterozygote and the probability that her first grandchild has blue eyes.

Solution. Everything follows from $q = p/(1+p)$, the probability that a brown-eyed parent transmits an x gene: a brown-eyed individual is Xx or XX with the population proportions renormalized,

$$\Pr(Xx \mid \text{brown}) = \frac{2p(1-p)}{1-p^2} = \frac{2p}{1+p},$$

and only a heterozygote can transmit x, doing so with probability $\frac{1}{2}$, so

$$q = \tfrac{1}{2} \cdot \frac{2p}{1+p} = \frac{p}{1+p}.$$

Under random mating the two brown-eyed parents transmit independently, so the child is xx with probability q^2 and Xx with probability $2q(1-q)$; conditioning on the child being brown,

$$\begin{aligned} \Pr(Xx \mid \text{brown child of brown parents}) &= \frac{2q(1-q)}{1-q^2} = \frac{2q}{1+q} \\ &= \frac{2p/(1+p)}{(1+2p)/(1+p)} = \frac{2p}{1+2p}, \end{aligned}$$

which is the required proportion. (The last step used $1 + q = (1+2p)/(1+p)$.)
Judy's prior is therefore $\pi = \Pr(Xx) = 2p/(1+2p)$, with $\Pr(XX) = 1/(1+2p)$. Her husband is Xx. If Judy is Xx each child is xx with probability $\frac{1}{4}$, so n brown-eyed children have probability $(3/4)^n$;

if Judy is Xx every child is brown, probability 1. Bayes' rule, with the common factor $1/(1+2p)$ cancelling, gives

$$\pi_n := \Pr(Xx \mid n \text{ brown children}) = \frac{2p\,(3/4)^n}{2p\,(3/4)^n + 1}.$$

For the grandchild, first update Judy's eldest child, who is brown-eyed. (i) If Judy is Xx, the cross $Xx \times Xx$ gives xx, Xx, XX with probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$, so the brown child is Xx with probability $(1/2)/(3/4) = 2/3$. (ii) If Judy is XX, the cross $XX \times Xx$ gives Xx or XX each with probability $\frac{1}{2}$, all brown, so the child is Xx with probability $\frac{1}{2}$. Averaging over the posterior,

$$\Pr(\text{child } Xx \mid \text{data}) = \tfrac{2}{3}\pi_n + \tfrac{1}{2}(1-\pi_n) = \tfrac{1}{2} + \tfrac{1}{6}\pi_n.$$

Take the child's mate to be a random member of the population, which transmits x with probability $p^2 + \frac{1}{2}\cdot 2p(1-p) = p$. The grandchild is blue-eyed only if it receives x from both, and the child transmits x with probability $\frac{1}{2}$ when Xx and never when XX:

$$\begin{aligned}\Pr(\text{grandchild } xx \mid \text{data}) &= \tfrac{1}{2}\left(\tfrac{1}{2} + \tfrac{1}{6}\pi_n\right) p \\ &= \frac{p}{4} + \frac{p\,\pi_n}{12}.\end{aligned}$$

Problem 1.4. Probability assignment: we will use the football dataset to estimate some conditional probabilities about professional football games. There were twelve games with point spreads of 8 points; the outcomes in those games were -7, -5, -3, -3, 1, 6, 7, 13, 15, 16, 20, and 21, with positive values indicating wins by the favorite and negative values indicating wins by the underdog. Consider the following conditional probabilities:

$$\Pr(\text{favorite wins} \mid \text{point spread} = 8),$$

$$\Pr(\text{favorite wins by at least 8} \mid \text{point spread} = 8),$$

$$\Pr(\text{favorite wins by at least 8} \mid \text{point spread} = 8 \text{ and favorite wins}).$$

(a) Estimate each of these using the relative frequencies of games with a point spread of 8.
(b) Estimate each using the normal approximation for the distribution of (outcome − point spread).

Solution. (a) Counting the twelve outcomes: eight are positive and five are at least 8 (namely 13, 15, 16, 20, 21), and those five are among the eight wins, so

$$\begin{aligned}\Pr(\text{fav. wins}) &= \tfrac{8}{12} = 0.67,\\ \Pr(\text{fav. wins by} \geq 8) &= \tfrac{5}{12} = 0.42,\\ \Pr(\text{fav. wins by} \geq 8 \mid \text{fav. wins}) &= \tfrac{5}{8} = 0.63,\end{aligned}$$

the third being the ratio of the second to the first, as the event "wins by at least 8" is contained in "wins."
(b) Section 1.6 fits $d = (\text{outcome}) - (\text{point spread})$ as $d \mid x \sim \mathrm{N}(0, 14^2)$, independent of the spread x (sample mean 0.07, sample s.d. 13.86 over the 672 games). At $x = 8$, the favorite wins when outcome > 0, i.e. $d > -8$, and wins by at least 8 when $d \geq 0$:

$$\begin{aligned}\Pr(\text{fav. wins}) &= \Phi\left(\tfrac{8}{14}\right) = \Phi(0.571) = 0.72,\\ \Pr(\text{fav. wins by} \geq 8) &= \Phi(0) = 0.50,\\ \Pr(\text{fav. wins by} \geq 8 \mid \text{fav. wins}) &= \frac{0.50}{0.7161} = 0.70.\end{aligned}$$

This is the convention of Section 1.6 itself, which assigns $\Pr_{\text{norm}}(y > 0 \mid x) = \Phi(x/14)$ with no correction for the integer-valued scores; restoring a continuity correction, so that "wins" is $y \geq 1$ and "wins by at least 8" is $y \geq 8$, gives instead $\Phi(7.5/13.86) = 0.71$, $\Phi(0.5/13.86) = 0.51$ and 0.73.

Problem 1.5. Probability assignment: the 435 U.S. Congressmembers are elected to two-year terms; the number of voters in an individual congressional election varies from about 50,000 to 350,000. We will use various sources of information to estimate roughly the probability that at least one congressional election is tied in the next national election.
(a) Use any knowledge you have about U.S. politics. Specify clearly what information you are using to construct this conditional probability, even if your answer is just a guess.
(b) Use the following information: in the period 1900-1992, there were 20,597 congressional elections, out of which 6 were decided by fewer than 10 votes and 49 decided by fewer than 100 votes.
See Gelman, King, and Boscardin (1998), Mulligan and Hunter (2001), and Gelman, Katz, and Tuerlinckx (2002) for more on this topic.

Solution. Roughly one chance in 200: about 0.004 from general knowledge, about 0.006 from the historical near-tie counts.
(a) The information used: of the 435 seats, the great majority are safe, and in a typical year only some 40 races are competitive enough that the winning margin could plausibly have landed within a few thousand votes; a competitive race draws on the order of 200,000 votes; and for such a race the margin D near zero is spread out over a window of some $\pm$5,000 votes with no reason to favour one value in that window over another, so its density near zero is about 1/10,000 per vote. Hence for a competitive race

$$\Pr(D = 0) \approx 10^{-4},$$

and with the safe races contributing essentially nothing,

$$\Pr(\text{some tie}) \approx 40 \times 10^{-4} = 0.004.$$

(b) Let f denote the probability density (per vote) of the margin D at zero, taken as constant over a window of $\pm$100 votes, which the data below justify. One wrinkle must be handled: with turnout N fixed, $D \equiv N \pmod 2$, so D moves on a lattice of spacing 2, each attainable value carrying probability $2f$, and a tie is outright impossible when N is odd. The nonzero values in the window $|D| < 10$ number 8 if N is even $(\pm2, \pm4, \pm6, \pm8)$ and 10 if N is odd $(\pm1, \pm3, \pm5, \pm7, \pm9)$, so averaging over the parity of N,

$$\Pr(0 < |D| < 10) \approx 9 \cdot 2f = 18f = \frac{6}{20597},$$

while an exact tie requires N even and then $D = 0$:

$$\Pr(D = 0) = \tfrac{1}{2} \cdot 2f = f = \frac{6/20597}{18} = 1.6 \times 10^{-5}.$$

The parity correction has cancelled: 18 is also the naive count of nonzero integers in the window. The same argument applied to the wider window gives the companion estimate

$$\Pr(D = 0) = \frac{49/20597}{198} = 1.2 \times 10^{-5},$$

the two agreeing to within 30%, which is as much as the flat-density assumption deserves. Taking $f \approx 1.4 \times 10^{-5}$ per race, or about 1 in 70,000, and treating the 435 races as independent for this purpose (a national swing shifts all margins together but does not change a first-order count of near-ties),

$$\begin{aligned}\Pr(\text{some tie}) &= 1 - (1 - 1.4 \times 10^{-5})^{435} \\ &\approx 435 \times 1.4 \times 10^{-5} = 0.006,\end{aligned}$$

one chance in about 165, the two window-based estimates bracketing this at 0.005 and 0.007.

Problem 1.6. Conditional probability: approximately 1/125 of all births are fraternal twins and 1/300 of births are identical twins. Elvis Presley had a twin brother (who died at birth). What is the probability that Elvis was an identical twin? (You may approximate the probability of a boy or girl birth as $\frac{1}{2}$.)

Solution. $5/11 = 0.45$.
Write I and F for the events that the birth was an identical-twin birth and a fraternal-twin birth (reading the quoted rates as fractions of birth events), and let E be the datum that Elvis, a boy, had a twin brother. The prior odds are

$$\frac{\Pr(I)}{\Pr(F)} = \frac{1/300}{1/125} = \frac{5}{12},$$

and, given that Elvis is a boy, an identical co-twin is a boy with certainty while a fraternal co-twin is a boy with probability $\frac{1}{2}$:

$$\frac{\Pr(E \mid I)}{\Pr(E \mid F)} = \frac{1}{1/2} = 2.$$

Multiplying, the posterior odds are $\frac{5}{12} \cdot 2 = \frac{5}{6}$, so

$$\Pr(I \mid E) = \frac{5}{5+6} = \frac{5}{11} = 0.4545.$$

Problem 1.7. Conditional probability: the following problem is loosely based on the television game show Let's Make a Deal. At the end of the show, a contestant is asked to choose one of three large boxes, where one box contains a fabulous prize and the other two boxes contain lesser prizes. After the contestant chooses a box, Monty Hall, the host of the show, opens one of the two boxes containing smaller prizes. (In order to keep the conclusion suspenseful, Monty does not open the box selected by the contestant.) Monty offers the contestant the opportunity to switch from the chosen box to the remaining unopened box. Should the contestant switch or stay with the original choice? Calculate the probability that the contestant wins under each strategy. This is an exercise in being clear about the information that should be conditioned on when constructing a probability judgment. See Selvin (1975) and Morgan et al. (1991) for further discussion of this problem.

Solution. Switch: switching wins with probability 2/3, staying with probability 1/3.
Label the boxes so that the contestant picks box 1, let $\theta \in \{1, 2, 3\}$ be the box holding the prize, with $\Pr(\theta = j) = 1/3$, and let M be the box Monty opens. Monty's behaviour is the model: he always opens a box that is neither box 1 nor the prize box, and when $\theta = 1$ he is free to choose, opening box 3 with probability q. Then

$$\Pr(M = 3 \mid \theta = 1) = q, \quad \Pr(M = 3 \mid \theta = 2) = 1, \quad \Pr(M = 3 \mid \theta = 3) = 0,$$

and Bayes' rule with the uniform prior gives

$$\Pr(\theta = 1 \mid M = 3) = \frac{\frac{1}{3}q}{\frac{1}{3}q + \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 0} = \frac{q}{1+q},$$
$$\Pr(\theta = 2 \mid M = 3) = \frac{1}{1+q},$$

the first being the winning probability for staying and the second for switching. The contestant has no information distinguishing boxes 2 and 3, so $q = \frac{1}{2}$ and these are 1/3 and 2/3.
The conditioning is the whole content. The datum is not "box 3 holds a lesser prize," which would leave boxes 1 and 2 symmetric at 1/2 each; it is "Monty, who is barred from opening box 1 and from revealing the prize, opened box 3." That rule makes $M = 3$ certain when the prize is in box 2 but only a coin flip when it is in box 1, which is exactly the likelihood ratio 2 in favour of switching.

1.2 Exercises 1.8–1.9

Problem 1.8. Subjective probability: discuss the following statement. 'The probability of event E is considered "subjective" if two rational persons A and B can assign unequal probabilities to E, $P_A(E)$ and $P_B(E)$. These probabilities can also be interpreted as "conditional": $P_A(E) = P(E \mid I_A)$ and $P_B(E) = P(E \mid I_B)$, where I_A and I_B represent the knowledge available to persons A and B, respectively.' Apply this idea to the following examples.
(a) The probability that a '6' appears when a fair die is rolled, where A observes the outcome of the die roll and B does not.
(b) The probability that Brazil wins the next World Cup, where A is ignorant of soccer and B is a knowledgeable sports fan.

Solution. The conditional reading is exact in (a) and formal only in (b): in (a) the two persons share one probability model and differ by a conditioning event, in (b) they differ by the model itself.
(a) With $E = \{\text{die shows } 6\}$, I_B the knowledge that a fair die was rolled and $I_A = I_B$ together with the observed face,

$$P_B(E) = P(E \mid I_B) = \tfrac{1}{6},$$
$$P_A(E) = P(E \mid I_A) = \mathbf{1}\{\text{the face was } 6\} \in \{0, 1\},$$

unequal with neither person irrational. They are coherent, since $E[P_A(E) \mid I_B] = P(E \mid I_B) = 1/6$: B's probability is B's expectation of A's, and B would adopt A's value on being told I_A. The physical symmetry of the die makes $1/6$ the uniquely reasonable value given I_B (Section 1.5), so the subjectivity is about information only, not judgment.
(b) There is no single model P of which P_A and P_B are two conditionings: no symmetry argument and no replication of "the next World Cup". A and B differ not only in data (squad quality, injuries, the draw, betting markets) but in the model linking that data to a winner, and two equally informed fans may still disagree, which cannot happen in (a). What survives is the betting definition of Section 1.5: $P_A(E)$ is the p at which A is indifferent between the two bets, and nothing forces A and B to name the same p. Such probabilities are judged by calibration over many statements, as in the football point-spread example of Section 1.6, with B's extra knowledge showing up as sharper but still calibrated forecasts; and A, believing $I_A \subset I_B$, may rationally defer and adopt $P_B(E)$.

Problem 1.9. Simulation of a queuing problem: a clinic has three doctors. Patients come into the clinic at random, starting at 9 a.m., according to a Poisson process with time parameter 10 minutes: that is, the time after opening at which the first patient appears follows an exponential distribution with expectation 10 minutes and then, after each patient arrives, the waiting time until the next patient is independently exponentially distributed, also with expectation 10 minutes. When a patient arrives, he or she waits until a doctor is available. The amount of time spent by each doctor with each patient is a random variable, uniformly distributed between 5 and 20 minutes. The office stops admitting new patients at 4 p.m. and closes when the last patient is through with the doctor.
(a) Simulate this process once. How many patients came to the office? How many had to wait for a doctor? What was their average wait? When did the office close?
(b) Simulate the process 100 times and estimate the median and 50% interval for each of the summaries in (a).

Solution. In one simulated day the office saw 31 patients, of whom 2 had to wait, with mean wait 1.3 minutes among those two, and it closed at 4:05 p.m.; over 100 independent days the medians are 42 patients, 6 waiters, a 3.7-minute mean wait among waiters, and a 4:05 p.m. closing time.
The model is an $M/G/3$ queue on the fixed admission window $[0, T]$, $T = 420$ minutes, served first-come-first-served. Arrival times are the points of a Poisson process of rate $\lambda = 1/10$ per minute,

generated as the partial sums

$$t_k = \sum_{i=1}^{k} X_i, \qquad X_i \overset{\text{iid}}{\sim} \text{Expon}(1/10),$$

kept while $t_k \le T$; service times are $S_k \overset{\text{iid}}{\sim} \text{U}(5, 20)$, independent of the arrivals. With $c = 3$ doctors and $E[S] = 12.5$ the utilization is $\rho = \lambda E[S]/c = 5/12$, so the queue is lightly loaded and most patients are seen at once.

The recursion is the only thing needed. Let f_j denote the time doctor j next becomes free, initialized at $f_j = 0$. Patient k arriving at t_k goes to the doctor who is free earliest, $j_k = \arg\min_j f_j$, and starts at

$$b_k = \max(t_k, f_{j_k}), \qquad W_k = b_k - t_k, \qquad f_{j_k} \leftarrow b_k + S_k.$$

The number of patients is $N = \max\{k : t_k \le T\}$, the number who waited is $\#\{k : W_k > 0\}$, and the closing time is $\max_k(b_k + S_k)$. Since a patient can only be delayed when all three doctors are busy, $W_k > 0$ exactly when $\min_j f_j > t_k$, which is what the recursion tests.

(a) The first of the 100 replicate days of part (b):

summary	value
patients seen	31
patients who had to wait	2
mean wait among those who waited	1.30 minutes
mean wait over all patients	0.08 minutes
closing time	4:05 p.m. (425.4 min after 9 a.m.)

(b) Repeating the day 100 times independently, the estimands are the median and the central 50% interval (25th and 75th percentiles) of each summary over days; the last column is the bootstrap standard error of the median estimate across the 100 replicates, i.e. the Monte Carlo error from using 100 days rather than infinitely many.

summary	median	50% interval	MC se of median
patients seen	42	[39, 46]	0.9
patients who had to wait	6	[3, 9]	0.6
mean wait among those who waited (min)	3.67	[2.52, 5.35]	0.27
mean wait over all patients (min)	0.54	[0.22, 0.94]	0.11
closing time	4:05 p.m.	[4:00 p.m., 4:10 p.m.]	0.6 min

The closing-time interval is $[420.4, 429.8]$ minutes after 9 a.m.; at this Monte Carlo precision the counts are good to about one patient and the mean waits to a few seconds, so 100 replications support the medians above but not a third digit.

2 Single-Parameter Models

Contents

2.1 Exercises 2.1–2.7

Problem 2.1. Posterior inference: suppose you have a Beta$(4,4)$ prior distribution on the probability θ that a coin will yield a 'head' when spun in a specified manner. The coin is independently spun ten times, and 'heads' appear fewer than 3 times. You are not told how many heads were seen, only that the number is less than 3. Calculate your exact posterior density (up to a proportionality constant) for θ and sketch it.

Solution.

$$p(\theta \mid y<3) \;\propto\; \theta^3(1-\theta)^{11}\big(36\theta^2+8\theta+1\big), \qquad 0\le\theta\le 1.$$

The observation is the **event** $\{y<3\}$, so the likelihood is its probability under $y \mid \theta \sim \mathrm{Bin}(10,\theta)$:

$$\begin{aligned}\Pr(y<3\mid\theta) &= \sum_{k=0}^{2}\binom{10}{k}\theta^k(1-\theta)^{10-k}\\ &= (1-\theta)^{10}+10\theta(1-\theta)^9+45\theta^2(1-\theta)^8\\ &= (1-\theta)^8\big[(1-\theta)^2+10\theta(1-\theta)+45\theta^2\big]\\ &= (1-\theta)^8\big(36\theta^2+8\theta+1\big).\end{aligned}$$

Multiplying by the Beta$(4,4)$ density $p(\theta)\propto\theta^3(1-\theta)^3$ gives the display above, whose normalizing constant is

$$\Big[\int_0^1 \theta^3(1-\theta)^{11}\big(36\theta^2+8\theta+1\big)d\theta\Big]^{-1} = \tfrac{7735}{8} = 966.875,$$

by $\int_0^1\theta^{a-1}(1-\theta)^{b-1}d\theta = B(a,b)$ applied to the three monomials.
Term by term the posterior is the beta mixture

$$\tfrac{13}{128}\,\mathrm{Beta}(\theta\mid 4,14)+\tfrac{5}{16}\,\mathrm{Beta}(\theta\mid 5,13)+\tfrac{75}{128}\,\mathrm{Beta}(\theta\mid 6,12),$$

with weights proportional to $\binom{10}{k}B(4+k,14-k)$, $k=0,1,2$ (Check!), so $E(\theta\mid y<3)=\frac{39}{128}=0.305$ and the mode is $\theta=0.281$.
Sketch: a single hump on $[0,1]$, zero at both endpoints, rising steeply to its peak at $\theta\approx 0.28$ and decaying with a mild right tail that is negligible beyond $\theta=0.7$.

Problem 2.2. Predictive distributions: consider two coins, C_1 and C_2, with the following characteristics: $\Pr(\text{heads}\mid C_1)=0.6$ and $\Pr(\text{heads}\mid C_2)=0.4$. Choose one of the coins at random and imagine spinning it repeatedly. Given that the first two spins from the chosen coin are tails, what is the expectation of the number of additional spins until a head shows up?

Solution.

$$E(N\mid TT)=\tfrac{175}{78}=2.244.$$

Two tails is evidence for the tail-heavy coin C_2. With $\Pr(C_1)=\Pr(C_2)=\frac12$, Bayes' rule gives

$$\begin{aligned}\Pr(C_1\mid TT) &= \frac{(0.4)^2}{(0.4)^2+(0.6)^2}=\frac{0.16}{0.52}=\tfrac{4}{13},\\ \Pr(C_2\mid TT) &= \tfrac{9}{13}.\end{aligned}$$

Given the coin, the spins are independent, so the number N of additional spins up to and including the first head is geometric with mean $1/\Pr(\text{heads}\mid C)$; by (1.8),

$$\begin{aligned}E(N\mid TT) &= \tfrac{4}{13}\cdot\tfrac{1}{0.6}+\tfrac{9}{13}\cdot\tfrac{1}{0.4}\\ &= \tfrac{4}{13}\cdot\tfrac{5}{3}+\tfrac{9}{13}\cdot\tfrac{5}{2}=\tfrac{20}{39}+\tfrac{45}{26}=\tfrac{175}{78}.\end{aligned}$$

Problem 2.3. Predictive distributions: let y be the number of 6's in 1000 rolls of a fair die.
(a) Sketch the approximate distribution of y, based on the normal approximation.
(b) Using the normal distribution table, give approximate 5%, 25%, 50%, 75%, and 95% points for the distribution of y.

Solution. (a) $y \sim \text{Bin}(1000, \frac{1}{6})$, so

$$E(y) = 1000 \cdot \tfrac{1}{6} = 166.7,$$
$$\text{sd}(y) = \sqrt{1000 \cdot \tfrac{1}{6} \cdot \tfrac{5}{6}} = \sqrt{138.9} = 11.8,$$

and y is approximately $\text{N}(166.7,\, 11.8^2)$: a symmetric bell centred at 167, with essentially all its mass on $[130, 205]$ and inflection points at 155 and 178.
(b) The quantiles are $166.7 + 11.8\, z_p$ with $z_p = -1.645,\ -0.674,\ 0,\ 0.674,\ 1.645$, rounded to integers since y is discrete:

p	5%	25%	50%	75%	95%
y_p	147	159	167	175	186

Problem 2.4. Predictive distributions: let y be the number of 6's in 1000 independent rolls of a particular real die, which may be unfair. Let θ be the probability that the die lands on '6.' Suppose your prior distribution for θ is as follows:

$$\Pr(\theta = 1/12) = 0.25,$$
$$\Pr(\theta = 1/6) = 0.5,$$
$$\Pr(\theta = 1/4) = 0.25.$$

(a) Using the normal approximation for the conditional distributions, $p(y \mid \theta)$, sketch your approximate prior predictive distribution for y.
(b) Give approximate 5%, 25%, 50%, 75%, and 95% points for the distribution of y. (Be careful here: y does not have a normal distribution, but you can still use the normal distribution as part of your analysis.)

Solution. (a) The prior predictive is the three-component normal mixture

$$p(y) \approx 0.25\, \text{N}(y \mid 83.3,\, 8.74^2) + 0.5\, \text{N}(y \mid 166.7,\, 11.79^2) + 0.25\, \text{N}(y \mid 250,\, 13.69^2),$$

since $y \mid \theta \sim \text{Bin}(1000, \theta)$ has $E(y \mid \theta) = 1000\theta$ and $\text{sd}(y \mid \theta) = \sqrt{1000\,\theta(1-\theta)}$, evaluated at $\theta = 1/12,\ 1/6,\ 1/4$.
The sketch is three well-separated bumps centred at 83, 167 and 250: adjacent centres are more than six standard deviations apart ($83.4/11.79 = 7.1$, $83.3/13.69 = 6.1$), so the density sinks to 7×10^{-6} near $y = 119$ and 1.0×10^{-4} near $y = 207$. The middle bump carries twice the area of each outer one and is the tallest, with peak height 0.0169 against 0.0114 and 0.0073.
(b) Let Φ denote the standard normal cdf and write the mixture cdf

$$F(y) = 0.25\, \Phi\Big(\tfrac{y-83.3}{8.74}\Big) + 0.5\, \Phi\Big(\tfrac{y-166.7}{11.79}\Big) + 0.25\, \Phi\Big(\tfrac{y-250}{13.69}\Big).$$

Because the components barely overlap, each quantile is located by inverting a single component, the other two terms of F equalling 0 or 1 to within 10^{-9} there.
(i) $p = 0.05$ lies inside the first component: $0.25\, \Phi(z) = 0.05$, so $\Phi(z) = 0.2$, $z = -0.842$, and

$$y_{0.05} = 83.3 - 0.842(8.74) = 76.0.$$

(ii) $p = 0.50$ lies at the centre of the second: $0.25 + 0.5\, \Phi(z) = 0.50$ gives $z = 0$ and $y_{0.50} = 166.7$.
(iii) $p = 0.95$ lies inside the third: $0.75 + 0.25\, \Phi(z) = 0.95$, so $\Phi(z) = 0.8$, $z = 0.842$, and

$$y_{0.95} = 250 + 0.842(13.69) = 261.5.$$

(iv) $p = 0.25$ and $p = 0.75$ are exactly the cumulative masses **between** components, so they fall in the empty valleys and are essentially undetermined: F crosses 0.25 at $y = 118$ but $|F(y) - 0.25| < 0.0005$ for every y in $[110, 130]$, and F crosses 0.75 at $y = 206$ with $|F(y) - 0.75| < 0.0015$ throughout $[200, 215]$.

p	5%	25%	50%	75%	95%
y_p	76	118	167	206	262

Problem 2.5. Posterior distribution as a compromise between prior information and data: let y be the number of heads in n spins of a coin, whose probability of heads is θ.
(a) If your prior distribution for θ is uniform on the range $[0, 1]$, derive your prior predictive distribution for y,

$$\Pr(y = k) = \int_0^1 \Pr(y = k \mid \theta)\, d\theta,$$

for each $k = 0, 1, \ldots, n$.
(b) Suppose you assign a Beta(α, β) prior distribution for θ, and then you observe y heads out of n spins. Show algebraically that your posterior mean of θ always lies between your prior mean, $\frac{\alpha}{\alpha+\beta}$, and the observed relative frequency of heads, $\frac{y}{n}$.
(c) Show that, if the prior distribution on θ is uniform, the posterior variance of θ is always less than the prior variance.
(d) Give an example of a Beta(α, β) prior distribution and data y, n, in which the posterior variance of θ is higher than the prior variance.

Solution. (a) $\Pr(y = k) = \dfrac{1}{n+1}$ for every $k = 0, 1, \ldots, n$ – the discrete uniform distribution:

$$\begin{aligned}\int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k}\, d\theta &= \binom{n}{k} B(k+1,\, n-k+1) \\ &= \frac{n!}{k!\,(n-k)!} \cdot \frac{k!\,(n-k)!}{(n+1)!} = \frac{1}{n+1}.\end{aligned}$$

(b) The posterior is $\theta \mid y \sim \text{Beta}(\alpha + y, \beta + n - y)$, so with $\lambda = \dfrac{\alpha+\beta}{\alpha+\beta+n} \in (0, 1)$,

$$\begin{aligned}E(\theta \mid y) &= \frac{\alpha + y}{\alpha + \beta + n} \\ &= \frac{\alpha+\beta}{\alpha+\beta+n} \cdot \frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n} \cdot \frac{y}{n} \\ &= \lambda \frac{\alpha}{\alpha+\beta} + (1-\lambda)\frac{y}{n}.\end{aligned}$$

A convex combination of two numbers with strictly positive weights lies between them (and equals them only when they coincide).
(c) Uniform means Beta$(1, 1)$, whose variance is $\frac{1\cdot 1}{2^2 \cdot 3} = \frac{1}{12}$. Writing $m = E(\theta \mid y) = \frac{y+1}{n+2}$, the posterior Beta$(1 + y,\, 1 + n - y)$ has

$$\text{var}(\theta \mid y) = \frac{m(1-m)}{n+3} \le \frac{1/4}{n+3} \le \frac{1}{16} < \frac{1}{12}$$

for every $n \ge 1$, since $m(1-m) \le \frac{1}{4}$.
(d) Prior Beta$(1, 3)$ with $n = 1$, $y = 1$:

$$\begin{aligned}\text{var}(\theta) &= \frac{1 \cdot 3}{4^2 \cdot 5} = \frac{3}{80} = 0.0375, \\ \text{var}(\theta \mid y) &= \frac{2 \cdot 3}{5^2 \cdot 6} = \frac{1}{25} = 0.0400.\end{aligned}$$

Problem 2.6. Predictive distributions: Derive the mean and variance (2.17) of the negative binomial predictive distribution for the cancer rate example, using the mean and variance formulas (1.8) and (1.9).
For reference: the cancer-rate model (2.16) is $y_j \mid \theta_j \sim \text{Poisson}(10n_j\theta_j)$ with $\theta_j \sim \text{Gamma}(\alpha, \beta)$, so that $y_j \sim \text{Neg-bin}\big(\alpha, \frac{\beta}{10n_j}\big)$; formulas (1.8) and (1.9) are $E(u) = E(E(u \mid v))$ and $\text{var}(u) = E(\text{var}(u \mid v)) + \text{var}(E(u \mid v))$; and (2.17) reads

$$E(y_j) = 10n_j \frac{\alpha}{\beta},$$
$$\text{var}(y_j) = 10n_j \frac{\alpha}{\beta} + (10n_j)^2 \frac{\alpha}{\beta^2}.$$

Solution. By (1.8), conditioning on θ_j,

$$E(y_j) = E\big(E(y_j \mid \theta_j)\big) = E(10n_j\theta_j) = 10n_j \frac{\alpha}{\beta},$$

using $E(\theta_j) = \alpha/\beta$ and $\text{var}(\theta_j) = \alpha/\beta^2$ for $\theta_j \sim \text{Gamma}(\alpha, \beta)$ (Appendix A) and the Poisson identity $E(y_j \mid \theta_j) = \text{var}(y_j \mid \theta_j) = 10n_j\theta_j$.
By (1.9),

$$\begin{aligned}\text{var}(y_j) &= E\big(\text{var}(y_j \mid \theta_j)\big) + \text{var}\big(E(y_j \mid \theta_j)\big) \\ &= E(10n_j\theta_j) + \text{var}(10n_j\theta_j) \\ &= 10n_j \frac{\alpha}{\beta} + (10n_j)^2 \frac{\alpha}{\beta^2},\end{aligned}$$

which is (2.17).

Problem 2.7. Noninformative prior densities:
(a) For the binomial likelihood, $y \sim \text{Bin}(n, \theta)$, show that $p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$ is the uniform prior distribution for the natural parameter of the exponential family.
(b) Show that if $y = 0$ or n, the resulting posterior distribution is improper.

Solution. (a) The natural parameter is $\phi = \text{logit}(\theta) = \log \frac{\theta}{1-\theta}$, because

$$p(y \mid \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \binom{n}{y}(1-\theta)^n \exp\Big\{y \log \tfrac{\theta}{1-\theta}\Big\},$$

which is of the exponential-family form $f(y)g(\theta)e^{\phi(\theta)^T u(y)}$ of Section 2.4 with $f(y) = \binom{n}{y}$, $g(\theta) = (1-\theta)^n$, $u(y) = y$ and $\phi(\theta) = \text{logit}(\theta)$.
Now put $p(\phi) \propto 1$ on $\phi \in (-\infty, \infty)$ and transform back by (2.19), using

$$\frac{d\phi}{d\theta} = \frac{d}{d\theta}\big[\log\theta - \log(1-\theta)\big] = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)} :$$

$$p(\theta) = p(\phi)\left|\frac{d\phi}{d\theta}\right| \propto \theta^{-1}(1-\theta)^{-1}.$$

(b) The unnormalized posterior is

$$p(\theta \mid y) \propto \theta^{-1}(1-\theta)^{-1}\theta^y(1-\theta)^{n-y} = \theta^{y-1}(1-\theta)^{n-y-1},$$

the kernel of Beta(y, $n-y$), normalizable exactly when $y > 0$ and $n - y > 0$; the prior is itself improper ($\int_0^1 \theta^{-1}(1-\theta)^{-1}d\theta = \infty$), so properness must be checked for each y rather than assumed.
(i) If $y = 0$ the integrand is $\theta^{-1}(1-\theta)^{n-1}$, and near 0 it is bounded below by $\frac{1}{2}\theta^{-1}$ for θ small, so $\int_0^1 \theta^{-1}(1-\theta)^{n-1}d\theta = \infty$.
(ii) If $y = n$ the integrand is $\theta^{n-1}(1-\theta)^{-1}$ and the same argument applies at 1.

2.2 Exercises 2.8–2.14

Problem 2.8. Normal distribution with unknown mean: a random sample of n students is drawn from a large population, and their weights are measured. The average weight of the n sampled students is $\bar{y} = 150$ pounds. Assume the weights in the population are normally distributed with unknown mean θ and known standard deviation 20 pounds. Suppose your prior distribution for θ is normal with mean 180 and standard deviation 40.
(a) Give your posterior distribution for θ. (Your answer will be a function of n.)
(b) A new student is sampled at random from the same population and has a weight of $\tilde{y}$ pounds. Give a posterior predictive distribution for $\tilde{y}$. (Your answer will still be a function of n.)
(c) For $n = 10$, give a 95% posterior interval for θ and a 95% posterior predictive interval for $\tilde{y}$.
(d) Do the same for $n = 100$.

Solution.

$$\theta \mid y \;\sim\; \mathrm{N}\!\left(\frac{180 + 600n}{1 + 4n},\; \frac{1600}{1 + 4n}\right).$$

(a) Apply (2.11)–(2.12) with $\mu_0 = 180$, $\tau_0^2 = 1600$, $\sigma^2 = 400$, $\bar{y} = 150$:

$$\frac{1}{\tau_n^2} = \frac{1}{1600} + \frac{n}{400} = \frac{1 + 4n}{1600},$$
$$\mu_n = \tau_n^2 \left(\frac{180}{1600} + \frac{150\, n}{400}\right) = \frac{180 + 600n}{1 + 4n}.$$

(b) Section 2.5 (p. 42) computes $\mathrm{E}(\tilde{y} \mid y) = \mu_n$ and, by identities (2.7)–(2.8), $\mathrm{var}(\tilde{y} \mid y) = \mathrm{E}(\sigma^2 \mid y) + \mathrm{var}(\theta \mid y) = \sigma^2 + \tau_n^2$; the pair $(\tilde{y}, \theta)$ is jointly normal, so

$$\tilde{y} \mid y \;\sim\; \mathrm{N}\!\left(\frac{180 + 600n}{1 + 4n},\; 400 + \frac{1600}{1 + 4n}\right).$$

(c) $n = 10$: $\mu_{10} = 6180/41 = 150.73$ and $\tau_{10}^2 = 1600/41 = 39.02$, so $\tau_{10} = 6.247$. The central 95% intervals are $\mu_{10} \pm 1.96\,\tau_{10}$ and $\mu_{10} \pm 1.96\sqrt{400 + \tau_{10}^2}$:

$$\theta:\; [138.5,\; 163.0],$$
$$\tilde{y}:\; 150.73 \pm 1.96\sqrt{439.02} = [109.7,\; 191.8].$$

(d) $n = 100$: $\mu_{100} = 60180/401 = 150.07$ and $\tau_{100}^2 = 1600/401 = 3.990$, so $\tau_{100} = 1.998$:

$$\theta:\; [146.2,\; 154.0],$$
$$\tilde{y}:\; 150.07 \pm 1.96\sqrt{403.99} = [110.7,\; 189.5].$$

Problem 2.9. Setting parameters for a beta prior distribution: suppose your prior distribution for θ, the proportion of Californians who support the death penalty, is beta with mean 0.6 and standard deviation 0.3.
(a) Determine the parameters α and β of your prior distribution. Sketch the prior density function.
(b) A random sample of 1000 Californians is taken, and 65% support the death penalty. What are your posterior mean and variance for θ? Draw the posterior density function.
(c) Examine the sensitivity of the posterior distribution to different prior means and widths including a non-informative prior.

Solution.

$$\alpha = 1, \qquad \beta = \tfrac{2}{3}.$$

(a) This is the method of moments (A.3), p. 585: with $\mu = \alpha/(\alpha + \beta)$ and $\mathrm{var} = \mu(1 - \mu)/(\alpha + \beta + 1)$

from Table A.1,

$$0.3^2 = \frac{(0.6)(0.4)}{\alpha+\beta+1} \implies \alpha+\beta+1 = \tfrac{0.24}{0.09} = \tfrac{8}{3},$$
$$\alpha = 0.6(\alpha+\beta) = 0.6 \cdot \tfrac{5}{3} = 1, \qquad \beta = \tfrac{5}{3} - 1 = \tfrac{2}{3}.$$

The density is $p(\theta) = \frac{2}{3}(1-\theta)^{-1/3}$ on $(0,1)$: strictly increasing from $\frac{2}{3}$ at $\theta = 0$, blowing up at $\theta = 1$ but with finite integral. (A standard deviation of 0.3 is close to the maximum $\sqrt{0.24} = 0.49$ attainable on $[0,1]$ with mean 0.6, which is why so little mass sits near the mean.)
(b) With $y = 650$ successes out of $n = 1000$, beta-binomial conjugacy (Section 2.4, p. 35, $\theta \mid y \sim \text{Beta}(\alpha+y, \beta+n-y)$) gives $\theta \mid y \sim \text{Beta}(1+650,\ \frac{2}{3}+350) = \text{Beta}(651,\ 350.67)$, so

$$\text{E}(\theta \mid y) = \frac{651}{1001.67} = 0.6499,$$
$$\text{var}(\theta \mid y) = \frac{(0.6499)(0.3501)}{1002.67} = 2.269 \times 10^{-4},$$

i.e. posterior standard deviation 0.01506. The density is essentially $\text{N}(0.650, 0.0151^2)$ to the eye: a sharp bell on $[0.60, 0.70]$, invisible on the scale of the prior sketch.
(c) Reparameterize a prior by its mean μ and prior sample size $\kappa = \alpha+\beta$, so $\alpha = \kappa\mu$, $\beta = \kappa(1-\mu)$, and

$$\text{E}(\theta \mid y) = \frac{\kappa\mu + 650}{\kappa + 1000},$$

a weighted average of μ and 0.650 with weights κ and 1000. Numerically:

prior	α	β	post. mean	post. sd
mean 0.6, sd 0.3	1	0.667	0.6499	0.01506
uniform Beta(1, 1)	1	1	0.6497	0.01506
Jeffreys Beta($\frac{1}{2}, \frac{1}{2}$)	0.5	0.5	0.6499	0.01507
improper Beta(0, 0)	0	0	0.6500	0.01508
mean 0.8, sd 0.15	4.89	1.22	0.6509	0.01502
mean 0.4, sd 0.10	9.2	13.8	0.6444	0.01496
mean 0.4, sd 0.05	38	57	0.6283	0.01460

Every prior with $\kappa \ll 1000$, the three noninformative choices included, shifts the posterior mean by less than 0.006 and leaves the posterior sd unchanged in the third decimal. Sensitivity appears only at $\kappa \approx 100$ (last row), a claim of prior information worth a tenth of the sample.

Problem 2.10. Discrete sample spaces: suppose there are N cable cars in San Francisco, numbered sequentially from 1 to N. You see a cable car at random; it is numbered 203. You wish to estimate N. (See Goodman, 1952, for a discussion and references to several versions of this problem, and Jeffreys, 1961, Lee, 1989, and Jaynes, 2003, for Bayesian treatments.)
(a) Assume your prior distribution on N is geometric with mean 100; that is,

$$p(N) = (1/100)(99/100)^{N-1}, \qquad \text{for } N = 1, 2, \ldots.$$

What is your posterior distribution for N?
(b) What are the posterior mean and standard deviation of N? (Sum the infinite series analytically or approximate them on the computer.)
(c) Choose a reasonable 'noninformative' prior distribution for N and give the resulting posterior distribution, mean, and standard deviation for N.

Solution.

$$p(N \mid y = 203) \;=\; \frac{1}{C}\,\frac{(0.99)^{N-1}}{N}, \qquad N = 203, 204, \ldots,$$

with $C = \sum_{N \geq 203} (0.99)^{N-1}/N = 0.047051$.
(a) The sampling model is $\Pr(y = 203 \mid N) = 1/N$ for $N \geq 203$ and 0 otherwise, so Bayes' rule multiplies the geometric prior by $1/N$ and truncates below at 203; the constant 1/100 is absorbed into

C. In closed form, using $\sum_{N\ge1} x^N/N = -\log(1-x)$,

$$C \;=\; \frac{1}{x}\left(-\log(1-x) - \sum_{N=1}^{202}\frac{x^N}{N}\right), \quad x = 0.99.$$

(b) The $1/N$ in the posterior cancels the N in the first moment, so the mean is a bare geometric tail sum:

$$\begin{aligned}
\mathrm{E}(N \mid y) &= \frac{1}{C}\sum_{N\ge203}(0.99)^{N-1} = \frac{(0.99)^{202}}{C\,(1-0.99)} = 279.09,\\
\mathrm{E}(N^2 \mid y) &= \frac{1}{C}\sum_{N\ge203} N(0.99)^{N-1}\\
&= \frac{1}{C}\cdot\frac{203(0.99)^{202}(0.01) + (0.99)^{203}}{(0.01)^2} = 84285,
\end{aligned}$$

so $\mathrm{sd}(N \mid y) = \sqrt{84285 - 279.09^2} = 79.96$.
(c) Take the scale-invariant prior $p(N) \propto 1/N$, the natural noninformative choice for a positive quantity whose order of magnitude is unknown (and the discrete analogue of Jeffreys' $p(\sigma) \propto \sigma^{-1}$; the flat prior $p(N) \propto 1$ will not do, since it leaves $p(N \mid y) \propto 1/N$, whose sum diverges). Then

$$p(N \mid y) = \frac{1}{C_2}\frac{1}{N^2}, \quad N \ge 203, \quad C_2 = \sum_{N\ge203} N^{-2} = \psi'(203) = 0.0049383,$$

which is proper. But

$$\mathrm{E}(N \mid y) \;=\; \frac{1}{C_2}\sum_{N\ge203}\frac{1}{N} \;=\; \infty,$$

and a fortiori the standard deviation is infinite: one draw from a discrete uniform bounds N below but says nothing about its scale, so a scale-free prior leaves a tail too heavy for moments. Quantiles are the usable summaries; from $\Pr(N > M \mid y) = \psi'(M+1)/\psi'(203) \approx 202.5/(M+\frac{1}{2})$,

$$\text{median } 405, \qquad \text{central 95\% interval } [208,\ 8100].$$

Problem 2.11. Computing with a nonconjugate single-parameter model: suppose $y_1, \ldots, y_5$ are independent samples from a Cauchy distribution with unknown center θ and known scale 1: $p(y_i \mid \theta) \propto 1/(1+(y_i-\theta)^2)$. Assume, for simplicity, that the prior distribution for θ is uniform on $[0, 100]$. Given the observations $(y_1, \ldots, y_5) = (43, 44, 45, 46.5, 47.5)$:
(a) Compute the unnormalized posterior density function, $p(\theta)p(y \mid \theta)$, on a grid of points $\theta = 0, \frac{1}{m}, \frac{2}{m}, \ldots, 100$, for some large integer m. Using the grid approximation, compute and plot the normalized posterior density function, $p(\theta \mid y)$, as a function of θ.
(b) Sample 1000 draws of θ from the posterior density and plot a histogram of the draws.
(c) Use the 1000 samples of θ to obtain 1000 samples from the predictive distribution of a future observation, y_6, and plot a histogram of the predictive draws.

Solution.

$$p(\theta \mid y) \;\propto\; \prod_{i=1}^{5}\frac{1}{1+(y_i-\theta)^2}, \qquad 0 \le \theta \le 100,$$

the uniform prior contributing only the truncation.
(a) Evaluate this on the grid $\theta_j = j/m$, $j = 0, \ldots, 100m$, with $m = 1000$; compute it as exp of the log-likelihood $-\sum_i \log(1+(y_i-\theta)^2)$ recentred at its maximum to avoid underflow, and normalize by $\sum_j$. The result is a single sharp mode on the data,

$$\begin{aligned}
\text{mode } &= 44.86, \quad \text{median } = 45.02, \quad \text{mean } = 45.08,\\
&\text{sd} = 0.986, \quad \text{central 95\%: } [43.28,\ 47.08],
\end{aligned}$$

and posterior mass 1.5×10^{-5} outside $[40, 50]$; the grid spacing 0.001 resolves a mode of width ≈ 2, so the discretization is harmless. The five likelihood factors merge into one bump, mildly right-skewed (skewness 0.20, and mode < median < mean above).
(b) Draw j from the normalized grid probabilities with replacement and jitter by $\mathrm{U}(-\frac{1}{2m}, \frac{1}{2m})$; 1000 such draws gave sample mean 45.04, sd 0.970, and quantiles

$$2.5\% : 43.30, \quad 25\% : 44.34, \quad 50\% : 44.98,$$
$$75\% : 45.71, \quad 97.5\% : 46.90,$$

matching the grid summaries in (a) to within Monte Carlo error $\mathrm{sd}/\sqrt{1000} \approx 0.03$. (Check!)
(c) For each posterior draw $\theta^{(s)}$ set $y_6^{(s)} = \theta^{(s)} + Z^{(s)}$ with $Z^{(s)}$ an independent standard Cauchy, which is exactly a draw from $p(y_6 \mid y) = \int p(y_6 \mid \theta) p(\theta \mid y)\, d\theta$ by (1.4). The predictive draws had

$$\text{median } 44.97, \quad \text{quartiles } 43.60,\ 46.31,$$
$$\text{central } 95\%\text{: } [35.0,\ 57.6],$$

with 87.5% of draws inside $[40, 50]$ and extreme values -1879 and 755: the predictive distribution has no mean, its spread being set by the unit Cauchy noise, not by the ± 1 uncertainty in θ. The histogram must therefore be plotted on a truncated axis (here $[35, 55]$) to show anything at all.

Problem 2.12. Jeffreys' prior distributions: suppose $y \mid \theta \sim \mathrm{Poisson}(\theta)$. Find Jeffreys' prior density for θ, and then find α and β for which the $\mathrm{Gamma}(\alpha, \beta)$ density is a close match to Jeffreys' density.

Solution.

$$p_{\text{Jeffreys}}(\theta) \propto \theta^{-1/2}, \qquad \text{i.e.} \qquad \alpha = \tfrac{1}{2}, \quad \beta = 0.$$

By the definition in Section 2.8, $p(\theta) \propto [J(\theta)]^{1/2}$ with J the Fisher information. For $\log p(y \mid \theta) = -\theta + y \log \theta - \log(y!)$,

$$\frac{d^2}{d\theta^2} \log p(y \mid \theta) = -\frac{y}{\theta^2},$$
$$J(\theta) = \mathrm{E}\left(-\frac{d^2}{d\theta^2} \log p(y \mid \theta) \;\middle|\; \theta \right) = \frac{\mathrm{E}(y \mid \theta)}{\theta^2} = \frac{1}{\theta},$$

so $p(\theta) \propto \theta^{-1/2}$.
The $\mathrm{Gamma}(\alpha, \beta)$ density is $\propto \theta^{\alpha-1} e^{-\beta\theta}$; matching the power gives $\alpha - 1 = -\frac{1}{2}$ and matching the exponential factor (which must be absent) gives $\beta = 0$. Thus Jeffreys' prior is the improper limit $\mathrm{Gamma}(\frac{1}{2}, 0)$, and for any $n \geq 1$ observations the conjugate updating (2.15) still yields a proper posterior $\mathrm{Gamma}\big(\frac{1}{2} + \sum_{i=1}^n y_i,\ n\big)$, since its shape $\frac{1}{2} + \sum y_i > 0$ and rate $n > 0$.

Problem 2.13. Discrete data: Table 2.2 gives the number of fatal accidents and deaths on scheduled airline flights per year over a ten-year period. We use these data as a numerical example for fitting discrete data models. Table 2.2 (worldwide airline fatalities, 1976–1985; death rate is passenger deaths per 100 million passenger miles; source: Statistical Abstract of the United States):

Year	Fatal accidents	Passenger deaths	Death rate
1976	24	734	0.19
1977	25	516	0.12
1978	31	754	0.15
1979	31	877	0.16
1980	22	814	0.14
1981	21	362	0.06
1982	26	764	0.13
1983	20	809	0.13
1984	16	223	0.03
1985	22	1066	0.15

(a) Assume that the numbers of fatal accidents in each year are independent with a Poisson(θ) distribution. Set a prior distribution for θ and determine the posterior distribution based on the data from 1976 through 1985. Under this model, give a 95% predictive interval for the number of fatal accidents in 1986. You can use the normal approximation to the gamma and Poisson or compute using simulation.
(b) Assume that the numbers of fatal accidents in each year follow independent Poisson distributions with a constant rate and an exposure in each year proportional to the number of passenger miles flown. Set a prior distribution for θ and determine the posterior distribution based on the data for 1976–1985. (Estimate the number of passenger miles flown in each year by dividing the appropriate columns of Table 2.2 and ignoring round-off errors.) Give a 95% predictive interval for the number of fatal accidents in 1986 under the assumption that 8×10^{11} passenger miles are flown that year.
(c) Repeat (a) above, replacing 'fatal accidents' with 'passenger deaths.'
(d) Repeat (b) above, replacing 'fatal accidents' with 'passenger deaths.'
(e) In which of the cases (a)–(d) above does the Poisson model seem more or less reasonable? Why? Discuss based on general principles, without specific reference to the numbers in Table 2.2.
Incidentally, in 1986, there were 22 fatal accidents, 546 passenger deaths, and a death rate of 0.06 per 100 million miles flown. We return to this example in Exercises 3.12, 6.2, 6.3, and 8.14.

Solution. All four parts are the single template: with $y_+ = \sum_i y_i$, $x_+ = \sum_i x_i$ and $y_i \mid \theta \sim \text{Poisson}(x_i\theta)$ as in (2.14),

$$\theta \mid y \sim \text{Gamma}\big(\tfrac{1}{2} + y_+,\ x_+\big), \quad \tilde{y} \mid y \sim \text{Neg-bin}\big(\tfrac{1}{2} + y_+,\ \tfrac{x_+}{\tilde{x}}\big),$$

by (2.14)–(2.15) and the negative binomial mixture identity of Section 2.6 (p. 44), $\int \text{Poisson}(\tilde{y} \mid \tilde{x}\theta)\,\text{Gamma}(\theta \mid \alpha, \beta)\,d\theta = \text{Neg-bin}(\tilde{y} \mid \alpha, \beta/\tilde{x})$, with

$$\text{E}(\tilde{y} \mid y) = \frac{(\frac{1}{2} + y_+)\tilde{x}}{x_+}, \quad \text{var}(\tilde{y} \mid y) = \text{E}(\tilde{y} \mid y)\left(1 + \frac{\tilde{x}}{x_+}\right).$$

Parts (a) and (c) are the case $x_i \equiv 1$, $\tilde{x} = 1$. The prior taken throughout is Jeffreys' Gamma$(\frac{1}{2}, 0)$ from Exercise 2.12; it is improper, but the posterior is proper for every part since its shape $\frac{1}{2} + y_+ > 0$ and rate $x_+ > 0$, and with ten years of data a prior of this weight is inconsequential. Given θ, the ten years are exchangeable in (a) and (c); in (b) and (d) model (2.14) is not exchangeable in the y_i, only in the pairs (x_i, y_i).
(a) $y_+ = 238$, $n = 10$, so $\theta \mid y \sim \text{Gamma}(238.5, 10)$: posterior mean 23.85, sd 1.54. The 1986 predictive has mean 23.85 and variance $23.85(1.1) = 26.24$, sd 5.12; the exact negative-binomial central 95% interval is

$$[14,\ 34]$$

(the normal approximation $23.85 \pm 1.96(5.12)$ gives $[13.8, 33.9]$, the same after rounding to integers).
(b) Exposures x_i = (passenger deaths)/(death rate), in units of 10^{11} passenger miles:

Year	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
x_i	3.863	4.300	5.027	5.481	5.814	6.033	5.877	6.223	7.433	7.107

with $x_+ = 57.159$. Then $\theta \mid y \sim \text{Gamma}(238.5,\ 57.159)$: posterior mean 4.173 fatal accidents per 10^{11} passenger miles, sd 0.270. With $\tilde{x} = 8$ the predictive mean is 33.38 and the variance $33.38(1 +$

8/57.159) = 38.05, sd 6.17, giving

$$[22,\ 46].$$

(c) $y_+ = 6919$, so $\theta \mid y \sim \text{Gamma}(6919.5,\ 10)$: posterior mean 691.95, sd 8.32. Predictive mean 691.95, variance 691.95(1.1) = 761.1, sd 27.59, and

$$[638,\ 747].$$

(d) $\theta \mid y \sim \text{Gamma}(6919.5,\ 57.159)$: posterior mean 121.06 deaths per 10^{11} passenger miles, sd 1.46. With $\tilde{x} = 8$, predictive mean 968.5, variance 968.5(1.13996) = 1104, sd 33.2, and

$$[904,\ 1034].$$

(e) The Poisson model is most defensible in (b) and least in (c) and (d).
(i) *Accidents versus deaths.* A Poisson count arises from many independent opportunities each with tiny probability. Fatal accidents plausibly satisfy this. Passenger deaths do not: deaths arrive in clusters, one per accident, of size equal to the number aboard. Deaths are therefore a compound Poisson sum $\sum_k m_k$ over accidents k, with var = E(accidents) E(m^2), which exceeds the mean by the factor E(m^2)/E(m) – of order the typical aircraft capacity. A Poisson likelihood for deaths therefore understates predictive uncertainty by roughly an order of magnitude in the standard deviation.
(ii) *Exposure versus no exposure.* Conditional on the count being Poisson, modelling the rate as constant per passenger mile is more credible than as constant per year, since traffic volume grows steadily and is the obvious driver of the number of opportunities for an accident. Model (a) forces the 1986 expectation to be the ten-year average of a series generated under systematically smaller exposure than 1986's; model (b) does not.
(iii) *Constant θ across a decade.* If safety improves over time, all four models are misspecified, overdispersed relative to Poisson through between-year variation in the rate; a time trend is the repair (Exercise 3.12).

Problem 2.14. Algebra of the normal model:
(a) Fill in the steps to derive (2.9)–(2.10), and (2.11)–(2.12).
(b) Derive (2.11) and (2.12) by starting with a $\text{N}(\mu_0, \tau_0^2)$ prior distribution and adding data points one at a time, using the posterior distribution at each step as the prior distribution for the next.
Here (2.9)–(2.10) state that for a single observation $y \mid \theta \sim \text{N}(\theta, \sigma^2)$ with σ^2 known and prior $\theta \sim \text{N}(\mu_0, \tau_0^2)$,

$$p(\theta \mid y) \propto \exp\left(-\frac{1}{2\tau_1^2}(\theta - \mu_1)^2\right),$$

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}, \qquad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2};$$

and (2.11)–(2.12) state that for $y = (y_1, \ldots, y_n)$ i.i.d. $\text{N}(\theta, \sigma^2)$, $p(\theta \mid y) = \text{N}(\theta \mid \mu_n, \tau_n^2)$ with

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \qquad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}.$$

Solution. Both derivations are the identity: a sum of two quadratics in θ is a single quadratic whose curvature is the sum of the curvatures. (The text on pp. 41–42 cites parts 2.14(b) and 2.14(c), which the printed exercise numbers differently: its 2.14(b) is the second half of the printed (a), and its 2.14(c) is the printed (b).)
(a) *Single observation.* With τ_1^2 defined by $\tau_1^{-2} = \tau_0^{-2} + \sigma^{-2}$ and $\mu_1 = \tau_1^2(\mu_0\tau_0^{-2} + y\sigma^{-2})$, collect powers

of θ:

$$\begin{aligned}\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2} &= \theta^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_0^2}\right)-2\theta\left(\frac{y}{\sigma^2}+\frac{\mu_0}{\tau_0^2}\right)+c_1\\ &= \frac{1}{\tau_1^2}\left(\theta^2-2\theta\mu_1\right)+c_1\\ &= \frac{1}{\tau_1^2}(\theta-\mu_1)^2+c_2,\end{aligned}$$

where $c_1 = y^2/\sigma^2+\mu_0^2/\tau_0^2$ and $c_2 = c_1-\mu_1^2/\tau_1^2$ are free of θ. Substituting into $p(\theta \mid y) \propto \exp(-\frac{1}{2}[\cdots])$ and absorbing $e^{-c_2/2}$ into the normalizing constant gives (2.9) with (2.10).
n observations. Apply the sum-of-squares decomposition $\sum_i(y_i-\theta)^2 = \sum_i(y_i-\bar y)^2+n(\bar y-\theta)^2$, whose first term is free of θ:

$$\begin{aligned}&\frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\\ &\qquad= \frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{n}{\sigma^2}(\theta-\bar y)^2+c_3.\end{aligned}$$

This is the single-observation exponent with y replaced by $\bar y$ and σ^2 by σ^2/n – equivalently, $\bar y \mid \theta \sim \mathrm{N}(\theta, \sigma^2/n)$ is sufficient – so (2.10) with those substitutions is exactly (2.12), and (2.11) follows.
(b) Let (μ_k, τ_k^2) denote the parameters of $p(\theta \mid y_1, \ldots, y_k)$. Because the y_i are conditionally independent given θ,

$$p(\theta \mid y_1, \ldots, y_k) \propto p(\theta \mid y_1, \ldots, y_{k-1})\, p(y_k \mid \theta),$$

so each step is an instance of (2.9)–(2.10) with prior $\mathrm{N}(\mu_{k-1}, \tau_{k-1}^2)$ and single observation y_k of variance σ^2. Claim, by induction on k:

$$\frac{1}{\tau_k^2} = \frac{1}{\tau_0^2}+\frac{k}{\sigma^2}, \qquad \frac{\mu_k}{\tau_k^2} = \frac{\mu_0}{\tau_0^2}+\frac{1}{\sigma^2}\sum_{i=1}^{k} y_i.$$

The case $k = 0$ is the prior. For the step, (2.10) gives

$$\begin{aligned}\frac{1}{\tau_k^2} &= \frac{1}{\tau_{k-1}^2}+\frac{1}{\sigma^2} = \frac{1}{\tau_0^2}+\frac{k-1}{\sigma^2}+\frac{1}{\sigma^2},\\ \frac{\mu_k}{\tau_k^2} &= \frac{\mu_{k-1}}{\tau_{k-1}^2}+\frac{y_k}{\sigma^2} = \frac{\mu_0}{\tau_0^2}+\frac{1}{\sigma^2}\sum_{i=1}^{k-1} y_i+\frac{y_k}{\sigma^2},\end{aligned}$$

which is the claim at k. Setting $k = n$ and using $\sum_{i=1}^{n} y_i = n\bar y$ yields precisely (2.12), and hence (2.11).

2.3 Exercises 2.15–2.21

Problem 2.15. Beta distribution: assume the result, from standard advanced calculus, that

$$\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

If Z has a beta distribution with parameters α and β, find $E[Z^m(1-Z)^n]$ for any nonnegative integers m and n. Hence derive the mean and variance of Z.

Solution.

$$E[Z^m(1-Z)^n] = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(\alpha+m)\Gamma(\beta+n)}{\Gamma(\alpha+\beta+m+n)}.$$

Indeed, with the Beta(α,β) density of Appendix A,

$$E[Z^m(1-Z)^n] = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 z^{\alpha+m-1}(1-z)^{\beta+n-1}\,dz$$
$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(\alpha+m)\Gamma(\beta+n)}{\Gamma(\alpha+\beta+m+n)},$$

the quoted integral applying because $\alpha+m>0$ and $\beta+n>0$.
Taking $(m,n)=(1,0)$ and $(2,0)$ and using $\Gamma(x+1)=x\Gamma(x)$,

$$E(Z) = \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+1)}{\Gamma(\alpha)\Gamma(\alpha+\beta+1)} = \frac{\alpha}{\alpha+\beta},$$
$$E(Z^2) = \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+2)}{\Gamma(\alpha)\Gamma(\alpha+\beta+2)} = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}.$$

Hence

$$\operatorname{var}(Z) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2}$$
$$= \frac{\alpha(\alpha+1)(\alpha+\beta)-\alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Problem 2.16. Beta-binomial distribution and Bayes' prior distribution: suppose y has a binomial distribution for given n and unknown parameter θ, where the prior distribution of θ is Beta(α,β).
(a) Find $p(y)$, the *marginal distribution* of y, for $y=0,\dots,n$ (unconditional on θ). This discrete distribution is known as the *beta-binomial*, for obvious reasons.
(b) Show that if the beta-binomial probability is constant in y, then the prior distribution has to have $\alpha=\beta=1$.

Solution. (a)

$$p(y) = \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(\alpha+y)\,\Gamma(\beta+n-y)}{\Gamma(\alpha+\beta+n)}, \qquad y=0,\dots,n,$$

by integrating the joint density over θ and reading off the value of the integral from Exercise 2.15 with exponents y and $n-y$:

$$p(y) = \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}\,\mathrm{Beta}(\theta\mid\alpha,\beta)\,d\theta$$
$$= \binom{n}{y}E\left[\theta^y(1-\theta)^{n-y}\right].$$

(b) Assume $n\ge 2$ (for $n=1$ constancy only forces $\alpha=\beta$, which is the one gap in the printed statement). Constancy is equivalent to $p(y+1)/p(y)=1$ for $y=0,\dots,n-1$, and

$$\frac{p(y+1)}{p(y)} = \frac{n-y}{y+1}\cdot\frac{\alpha+y}{\beta+n-y-1},$$

since $\Gamma(\alpha+y+1)=(\alpha+y)\Gamma(\alpha+y)$ and $\Gamma(\beta+n-y)=(\beta+n-y-1)\Gamma(\beta+n-y-1)$. Taking $y=0$ and $y=n-1$ gives the two equations

$$n\alpha = \beta+n-1, \qquad \alpha+n-1 = n\beta.$$

Adding them gives $(n+1)\alpha=(n+1)\beta$, so $\alpha=\beta$; the first equation then reads $n\alpha=\alpha+n-1$, i.e. $(n-1)\alpha=n-1$, so $\alpha=\beta=1$.

Problem 2.17. Posterior intervals: unlike the central posterior interval, the highest posterior interval is not invariant to transformation. For example, suppose that, given σ^2, the quantity nv/σ^2 is distributed as χ^2_n, and that σ has the (improper) noninformative prior density $p(\sigma) \propto \sigma^{-1}$, $\sigma > 0$.
(a) Prove that the corresponding prior density for σ^2 is $p(\sigma^2) \propto \sigma^{-2}$.
(b) Show that the 95% highest posterior density region for σ^2 is not the same as the region obtained by squaring the endpoints of a posterior interval for σ.

Solution. (a) Write $\psi = \sigma^2$, so $\sigma = \psi^{1/2}$ and $d\sigma/d\psi = \frac{1}{2}\psi^{-1/2}$; by the change-of-variables formula (1.4),

$$p(\psi) = p(\sigma)\left|\frac{d\sigma}{d\psi}\right| \propto \psi^{-1/2} \cdot \tfrac{1}{2}\psi^{-1/2} \propto \psi^{-1} = \sigma^{-2}.$$

(b) The transformation $\psi \mapsto \sigma = \psi^{1/2}$ is nonlinear, so it multiplies the density by the non-constant Jacobian 2σ; a level set of $p(\psi \mid v)$ is therefore *not* a level set of $p(\sigma \mid v)$, and an HPD region in one parameterization maps to a non-HPD region in the other.
Concretely, v has density $p(v \mid \psi) = (n/\psi)\,\chi^2_n(nv/\psi) \propto \psi^{-n/2}e^{-nv/(2\psi)}$, so with the prior of part (a)

$$p(\psi \mid v) \propto \psi^{-n/2-1}e^{-nv/(2\psi)}, \qquad \psi \mid v \sim \text{Inv-gamma}\left(\tfrac{n}{2}, \tfrac{nv}{2}\right),$$

which is the scaled inverse-χ^2 posterior Inv-$\chi^2(n, v)$ of Section 2.6 – proper for $n \geq 1, v > 0$, despite the improper prior. That density is unimodal (its log-derivative $-(n/2+1)/\psi + nv/(2\psi^2)$ vanishes only at $\psi = nv/(n+2)$), and likewise for σ, so each 95% HPD region is a single interval whose endpoints carry equal density. Thus $[\psi_L, \psi_U]$ of content 0.95 is HPD for ψ iff $p(\psi_L \mid v) = p(\psi_U \mid v)$, while $[\psi_L^{1/2}, \psi_U^{1/2}]$ (same content) is HPD for σ iff

$$2\psi_L^{1/2}\, p(\psi_L \mid v) = 2\psi_U^{1/2}\, p(\psi_U \mid v).$$

Both can hold only if $\psi_L = \psi_U$, impossible for a nondegenerate 95% region. Hence squaring the endpoints of the 95% posterior interval of highest density for σ never returns the 95% HPD interval for σ^2.

Problem 2.18. Poisson model: derive the gamma posterior distribution (2.15) for the Poisson model parameterized in terms of rate and exposure with conjugate prior distribution. That is, for the extended Poisson model (2.14),

$$y_i \sim \text{Poisson}(x_i\theta), \qquad i = 1, \ldots, n,$$

with known positive exposures x_i and prior $\theta \sim \text{Gamma}(\alpha, \beta)$, show that

$$\theta \mid y \sim \text{Gamma}\left(\alpha + \sum_{i=1}^n y_i,\ \beta + \sum_{i=1}^n x_i\right).$$

Solution. The likelihood is gamma-shaped in θ (the y_i are independent given θ, the model being exchangeable in the pairs (x_i, y_i) rather than in the y_i alone):

$$p(y \mid \theta) = \prod_{i=1}^n \frac{(x_i\theta)^{y_i}e^{-x_i\theta}}{y_i!} \ \propto\ \theta^{\sum_i y_i}\, e^{-(\sum_i x_i)\theta},$$

the discarded factor $\prod_i x_i^{y_i}/y_i!$ being free of θ. Multiplying by the Gamma(α, β) prior density $p(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta}$,

$$p(\theta \mid y) \ \propto\ \theta^{\,\alpha+\sum_i y_i-1}\, e^{-(\beta+\sum_i x_i)\theta}, \qquad \theta > 0,$$

which is the unnormalized Gamma$(\alpha + \sum_i y_i,\ \beta + \sum_i x_i)$ density; both parameters are positive (the x_i are known positive exposures), so the density is proper, the constant of proportionality is determined, and (2.15) follows.

Problem 2.19. Exponential model with conjugate prior distribution:
(a) Show that if $y \mid \theta$ is exponentially distributed with rate θ, then the gamma prior distribution is conjugate for inferences about θ given an independent and identically distributed sample of y values.
(b) Show that the equivalent prior specification for the mean, $\phi = 1/\theta$, is inverse-gamma. (That is, derive the latter density function.)
(c) The length of life of a light bulb manufactured by a certain process has an exponential distribution with unknown rate θ. Suppose the prior distribution for θ is a gamma distribution with coefficient of variation 0.5. (The *coefficient of variation* is defined as the standard deviation divided by the mean.) A random sample of light bulbs is to be tested and the lifetime of each obtained. If the coefficient of variation of the distribution of θ is to be reduced to 0.1, how many light bulbs need to be tested?
(d) In part (c), if the coefficient of variation refers to ϕ instead of θ, how would your answer be changed?

Solution. (a) For $y = (y_1, \ldots, y_n)$ i.i.d. Expon(θ),

$$p(y \mid \theta) = \prod_{i=1}^{n} \theta e^{-\theta y_i} = \theta^n e^{-\theta \sum_i y_i},$$

so with $\theta \sim \text{Gamma}(\alpha, \beta)$,

$$p(\theta \mid y) \propto \theta^{\alpha+n-1} e^{-(\beta+\sum_i y_i)\theta}, \qquad \theta \mid y \sim \text{Gamma}(\alpha + n,\ \beta + \textstyle\sum_i y_i)\,,$$

again a gamma distribution, so the family is conjugate.
(b) With $\phi = 1/\theta$, so $\theta = 1/\phi$ and $|d\theta/d\phi| = \phi^{-2}$, formula (1.4) gives

$$\begin{aligned} p(\phi) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\phi}\right)^{\alpha-1} e^{-\beta/\phi} \cdot \frac{1}{\phi^2} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \phi^{-(\alpha+1)} e^{-\beta/\phi}, \qquad \phi > 0, \end{aligned}$$

which is the Inv-gamma(α, β) density of Appendix A.
(c) $n = 96$. The Gamma(a, b) distribution has mean a/b and variance a/b^2, hence coefficient of variation

$$\text{CV} = \frac{\sqrt{a}/b}{a/b} = a^{-1/2},$$

free of b. The prior requirement $\alpha^{-1/2} = 0.5$ gives $\alpha = 4$; by (a) the posterior shape after n bulbs is $\alpha + n$, whatever the observed lifetimes, so

$$(\alpha + n)^{-1/2} = 0.1 \implies \alpha + n = 100 \implies n = 100 - 4 = 96.$$

(d) $n = 96$ again. By (b), $\phi \sim$ Inv-gamma(a, b) has mean $b/(a-1)$ and variance $b^2/\{(a-1)^2(a-2)\}$ for $a > 2$, so

$$\text{CV} = \frac{b/\{(a-1)\sqrt{a-2}\}}{b/(a-1)} = (a-2)^{-1/2}.$$

Now the prior condition $(\alpha-2)^{-1/2} = 0.5$ gives $\alpha = 6$, and the posterior condition $(\alpha+n-2)^{-1/2} = 0.1$ gives $\alpha + n = 102$, so

$$n = 102 - 6 = 96.$$

The prior is a different one, but the answer to (c) is unchanged.

Problem 2.20. Censored and uncensored data in the exponential model:
(a) Suppose $y \mid \theta$ is exponentially distributed with rate θ, and the marginal (prior) distribution of θ is Gamma(α, β). Suppose we observe that $y \geq 100$, but do not observe the exact value of y. What is the posterior distribution, $p(\theta \mid y \geq 100)$, as a function of α and β? Write down the posterior mean and variance of θ.
(b) In the above problem, suppose that we are now told that y is exactly 100. Now what are the posterior mean and variance of θ?
(c) Explain why the posterior variance of θ is higher in part (b) even though more information has been observed. Why does this not contradict identity (2.8) on page 32?

Solution. (a) $\theta \mid y \geq 100 \sim \text{Gamma}(\alpha,\ \beta + 100)$. The censored observation contributes the survival function $\Pr(y \geq 100 \mid \theta) = e^{-100\theta}$ as its likelihood, so

$$p(\theta \mid y \geq 100) \propto \theta^{\alpha-1} e^{-\beta\theta} \cdot e^{-100\theta} = \theta^{\alpha-1} e^{-(\beta+100)\theta},$$

whence

$$E(\theta \mid y \geq 100) = \frac{\alpha}{\beta + 100}, \qquad \text{var}(\theta \mid y \geq 100) = \frac{\alpha}{(\beta + 100)^2}.$$

(b) $\theta \mid y = 100 \sim \text{Gamma}(\alpha + 1,\ \beta + 100)$, by Exercise 2.19(a) with $n = 1$ and $\sum y_i = 100$, so

$$E(\theta \mid y = 100) = \frac{\alpha + 1}{\beta + 100}, \qquad \text{var}(\theta \mid y = 100) = \frac{\alpha + 1}{(\beta + 100)^2}.$$

(c) The exact observation adds the density factor $\theta e^{-100\theta}$ where the censored one added only $e^{-100\theta}$; the extra θ raises the shape from α to $\alpha + 1$ while leaving the rate at $\beta + 100$, and a gamma variance is proportional to its shape. So

$$\text{var}(\theta \mid y = 100) - \text{var}(\theta \mid y \geq 100) = \frac{1}{(\beta + 100)^2} > 0.$$

There is no contradiction with (2.8), for two reasons.
(i) Identity (2.8) is a statement about an *average* over the data, not about one dataset: it says $\text{var}(\theta) = E[\text{var}(\theta \mid y)] + \text{var}(E(\theta \mid y))$, so $E[\text{var}(\theta \mid y)] \leq \text{var}(\theta)$, while individual values of y may well raise the posterior variance.
(ii) Averaged in the right place, (2.8) holds exactly here. Take $p(\theta \mid y \geq 100)$ as the prior and the exact y as the new datum; writing $r = \beta + 100$, the predictive density is $p(y \mid y \geq 100) = \alpha r^\alpha / (r + y - 100)^{\alpha+1}$ for $y \geq 100$, and since $\text{var}(\theta \mid y) = (\alpha + 1)/(\beta + y)^2$ and $E(\theta \mid y) = (\alpha + 1)/(\beta + y)$ are both decreasing in y, $y = 100$ is the least favourable point of that range. Integrating,

$$E[\text{var}(\theta \mid y) \mid y \geq 100] = \frac{\alpha(\alpha + 1)}{(\alpha + 2) r^2},$$

$$\text{var}(E(\theta \mid y) \mid y \geq 100) = \frac{\alpha(\alpha + 1)^2}{(\alpha + 2) r^2} - \frac{\alpha^2}{r^2} = \frac{\alpha}{(\alpha + 2) r^2},$$

whose sum is $\alpha / r^2 = \text{var}(\theta \mid y \geq 100)$, exactly as (2.8) requires.

Problem 2.21. Simple hierarchical modeling:
The file `pew_research_center_june_elect_wknd_data.dta` has data from Pew Research Center polls taken during the 2008 election campaign (31,201 respondents; the variable `ideo` records self-placed ideology as one of *very liberal*, *liberal*, *moderate*, *conservative*, *very conservative*, or `dk/refused`, and `state` records the respondent's state). You can read these data into R using the `read.dta()` function (after first loading the `foreign` package into R).
Your task is to estimate the percentage of the (adult) population in each state (excluding Alaska, Hawaii, and the District of Columbia) who label themselves as 'very liberal,' following the general

procedure that was used in Section 2.7 to estimate cancer rates, but using the binomial and beta rather than Poisson and gamma distributions. But you do not need to make maps; it will be enough to make scatterplots, plotting the estimate vs. Barack Obama's vote share in 2008 (data available at `2008ElectionResult.csv`, readable in R using `read.csv()`).
Make the following four graphs on a single page:

- Graph proportion very liberal among the survey respondents in each state vs. Obama vote share – that is, a scatterplot using the two-letter state abbreviations (see `state.abb()` in R).
- Graph the Bayes posterior mean in each state vs. Obama vote share.
- Repeat graphs (a) and (b) using the number of respondents in the state on the x-axis.

This exercise has four challenges: first, manipulating the data in order to get the totals by state; second, estimating the parameters of the prior distribution; third, doing the Bayesian analysis by state; and fourth, making the graphs.

Solution. The model is $y_j \mid \theta_j \sim \text{Bin}(n_j, \theta_j)$ with the θ_j exchangeable across states, $\theta_j \sim \text{Beta}(\alpha, \beta)$ independently, giving the shrinkage estimate

$$E(\theta_j \mid y_j) = \frac{\alpha + y_j}{\alpha + \beta + n_j},$$

the beta-binomial analogue of the kidney-cancer formula $E(\theta_j \mid y_j) = (20 + y_j)/(430{,}000 + 10n_j)$ of Section 2.7.
Totals by state. Dropping the 1,787 respondents with missing or 'dk-refused' ideology leaves n_j respondents in state j, of whom y_j answered 'very liberal.' Alaska does not appear in the file at all; deleting Hawaii and the District of Columbia leaves $J = 48$ states, $\sum_j n_j = 29{,}353$ and $\sum_j y_j = 1{,}463$, a pooled proportion of 0.0498. The state sample sizes are very unequal – from $n = 29$ (Wyoming) through a median of 472 to $n = 2{,}688$ (California) – which is exactly the situation in which shrinkage matters.
Prior. Estimate (α, β) by matching the first two moments of the raw proportions $p_j = y_j/n_j$, correcting the second for binomial noise, as in Section 2.7. Write $\mu = \alpha/(\alpha+\beta)$ and $K = \alpha + \beta$. Then $E(p_j) = \mu$ and

$$\text{var}(p_j) = \text{var}(\theta_j) + E\left[\frac{\theta_j(1-\theta_j)}{n_j}\right],$$

by (2.8). The observed values are $\widehat{\text{var}}(p_j) = 3.493 \times 10^{-4}$ and a mean within-state sampling variance $\overline{p_j(1-p_j)/n_j} = 1.700 \times 10^{-4}$, leaving $\text{var}(\theta_j) \approx 1.794 \times 10^{-4}$. With $\mu = 0.0498$ and $\text{var}(\theta_j) = \mu(1-\mu)/(K+1)$ from Exercise 2.15,

$$K = \frac{\mu(1-\mu)}{\text{var}(\theta_j)} - 1 = 263.0, \qquad \alpha = \mu K = 13.1, \qquad \beta = (1-\mu)K = 249.9.$$

So the prior carries about as much information as a poll of 263 people. (Maximizing the beta-binomial marginal likelihood $\prod_j p(y_j \mid \alpha, \beta)$ from Exercise 2.16(a) instead gives $\hat\alpha = 16.29$, $\hat\beta = 316.0$, i.e. $K = 332$ and $\mu = 0.0490$; no posterior mean below moves by as much as 0.004, so the conclusions do not turn on the fitting method.)
Estimates. Shrinkage is severe for the small states and negligible for the large ones:

state	y_j	n_j	p_j	$E(\theta_j \mid y_j)$	Obama share
Wyoming	1	29	0.0345	0.0483	0.327
Idaho	14	138	0.1014	0.0676	0.361
New Hampshire	1	158	0.0063	0.0335	0.543
Utah	15	261	0.0575	0.0536	0.342
Oregon	40	442	0.0905	0.0753	0.571
New York	104	1,568	0.0663	0.0640	0.622
California	179	2,688	0.0666	0.0651	0.609

The raw proportions span [0.0063, 0.1014] with standard deviation 0.0187; the posterior means span only [0.0330, 0.0753] with standard deviation 0.0099. Both extremes of the raw scale are small-sample artifacts – New Hampshire's 1/158 and Idaho's 14/138 – and both are pulled most of the way back to the national 5%.

Graphs. The four panels (Figure `bda3-ch02-very-liberal-by-state`) plot each state as its two-letter abbreviation. Panel (a), raw proportion against Obama's 2008 share, is a shapeless cloud of correlation 0.18; panel (b), posterior mean against Obama share, tightens to correlation 0.21 on a vertical scale less than half as tall. Panels (c) and (d) put n_j (log scale) on the horizontal axis and make the point of the exercise visible: in (c) the spread of p_j narrows like $n_j^{-1/2}$, the outlying rates all belonging to states with a few hundred respondents, exactly as in Figure 2.8 for the kidney-cancer rates; in (d) the posterior means show no such funnel, because the small-n_j states are the ones nearly replaced by the prior mean $\alpha/(\alpha+\beta) = 0.0498$, precisely the flattening seen in Figure 2.9a.

2.4 Exercises 2.22–2.22

Problem 2.22. Prior distributions:
A (hypothetical) study is performed to estimate the effect of a simple training program on basketball free-throw shooting. A random sample of 100 college students is recruited into the study. Each student first shoots 100 free-throws to establish a baseline success probability. Each student then takes 50 practice shots each day for a month. At the end of that time, he or she takes 100 shots for a final measurement. Let θ be the average improvement in success probability.
Give three prior distributions for θ (explaining each in a sentence):
(a) A noninformative prior,
(b) A subjective prior based on your best knowledge, and
(c) A weakly informative prior.

Solution. Take, respectively, $\theta \sim \mathrm{U}(-1, 1)$, $\theta \sim \mathrm{N}(0.03,\, 0.02^2)$, and $\theta \sim \mathrm{N}(0,\, 0.2^2)$.
Each student contributes $\hat{\theta}_i = \hat{p}_i^{\text{final}} - \hat{p}_i^{\text{base}}$, a difference of two independent binomial proportions out of 100 shots, so at a baseline near $p = 0.65$

$$\mathrm{sd}(\hat{\theta}_i) = \sqrt{\tfrac{2p(1-p)}{100}} \approx 0.067,$$
$$\mathrm{sd}(\bar{\hat{\theta}}) = \mathrm{sd}(\hat{\theta}_i)/\sqrt{100} \approx 0.0067,$$

making the likelihood for θ nearly normal with scale about 0.007 (a floor: between-student variation in true improvement only widens it). Each prior below is judged against that number.
(a) Noninformative: $p(\theta) \propto 1$ on $[-1, 1]$, flat over the whole logically possible range for a difference of two probabilities; since the likelihood is confined to a window of width ≈ 0.03 deep inside that interval, the improper version $p(\theta) \propto 1$ on $\mathbb{R}$ gives the same posterior to several decimal places.
(b) Subjective: $\theta \sim \mathrm{N}(0.03,\, 0.02^2)$, a 95% prior interval of $(-0.009,\, 0.069)$. About 1500 practice shots over a month is genuine but modest training, so I expect a gain of roughly three percentage points, would be astonished by ten (that moves a typical 65% shooter to 75%), and leave a little mass below zero for fatigue or a baseline inflated by chance. Prior scale 0.02 is three times the likelihood scale, so this is mildly informative: the posterior standard deviation drops to

$$\left(0.02^{-2} + 0.0067^{-2}\right)^{-1/2} \approx 0.0064,$$

with weight $0.02^{-2}/(0.02^{-2} + 0.0067^{-2}) \approx 0.10$ on the prior mean.
(c) Weakly informative: $\theta \sim \mathrm{N}(0,\, 0.2^2)$ (or t_4 at 0 with scale 0.1 for heavier tails), a 95% prior interval of $(-0.39,\, 0.39)$. Centered at the null value $\theta = 0$ and wide enough that any reasonable dataset dominates it — the posterior standard deviation moves by 0.06% (Check!) — yet proper, and still ruling out the absurd, since a student starting near 0.65 cannot improve by 0.8.

3 Introduction to Multiparameter Models

Contents

3.1 Exercises 3.1–3.7

Problem 3.1. Binomial and multinomial models: suppose data $(y_1, \ldots, y_J)$ follow a multinomial distribution with parameters $(\theta_1, \ldots, \theta_J)$. Also suppose that $\theta = (\theta_1, \ldots, \theta_J)$ has a Dirichlet prior distribution. Let $\alpha = \frac{\theta_1}{\theta_1+\theta_2}$.
(a) Write the marginal posterior distribution for α.
(b) Show that this distribution is identical to the posterior distribution for α obtained by treating y_1 as an observation from the binomial distribution with probability α and sample size $y_1 + y_2$, ignoring the data $y_3, \ldots, y_J$.
This result justifies the application of the binomial distribution to multinomial problems when we are only interested in two of the categories; for example, see the next problem.

Solution. (a) $\alpha \mid y \sim \text{Beta}(a_1 + y_1,\, a_2 + y_2)$, writing $\text{Dirichlet}(a_1, \ldots, a_J)$ for the prior.
By Section 3.4 the Dirichlet is conjugate to the multinomial, so $\theta \mid y \sim \text{Dirichlet}(a_1 + y_1, \ldots, a_J + y_J)$. Use the gamma construction of the Dirichlet (Appendix A): let $x_1, \ldots, x_J$ be independent with $x_j \sim \text{Gamma}(a_j + y_j, 1)$ (all shape parameters positive, as required), and set $\theta_j = x_j / \sum_k x_k$. Then

$$\alpha = \frac{\theta_1}{\theta_1 + \theta_2} = \frac{x_1}{x_1 + x_2},$$

and the ratio of a $\text{Gamma}(a_1 + y_1, 1)$ variate to its sum with an independent $\text{Gamma}(a_2 + y_2, 1)$ variate is $\text{Beta}(a_1 + y_1, a_2 + y_2)$ (Appendix A). Since α is a function of x_1, x_2 alone, that is its marginal posterior – free of $y_3, \ldots, y_J$ and of $a_3, \ldots, a_J$.
(b) The same computation with $y = 0$ gives the induced prior $\alpha \sim \text{Beta}(a_1, a_2)$. Treating $y_1 \mid \alpha \sim \text{Bin}(y_1 + y_2, \alpha)$ contributes the likelihood factor $\alpha^{y_1}(1-\alpha)^{y_2}$, so

$$\begin{aligned} p(\alpha \mid y_1, y_2) &\propto \alpha^{a_1 - 1}(1-\alpha)^{a_2 - 1} \cdot \alpha^{y_1}(1-\alpha)^{y_2} \\ &= \alpha^{a_1 + y_1 - 1}(1-\alpha)^{a_2 + y_2 - 1}, \end{aligned}$$

that is, $\text{Beta}(a_1 + y_1, a_2 + y_2)$ – the distribution found in (a).

Problem 3.2. Comparison of two multinomial observations: on September 25, 1988, the evening of a presidential campaign debate, ABC News conducted a survey of registered voters in the United States; 639 persons were polled before the debate, and 639 different persons were polled after. The results are displayed in Table 3.2:

Survey	Bush	Dukakis	No opinion/other	Total
pre-debate	294	307	38	639
post-debate	288	332	19	639

Assume the surveys are independent simple random samples from the population of registered voters. Model the data with two different multinomial distributions. For $j = 1, 2$, let α_j be the proportion of voters who preferred Bush, out of those who had a preference for either Bush or Dukakis at the time of survey j. Plot a histogram of the posterior density for $\alpha_2 - \alpha_1$. What is the posterior probability that there was a shift toward Bush?

Solution. $\Pr(\alpha_2 > \alpha_1 \mid y) = 0.19$: the data give no evidence of a shift toward Bush, and in fact mildly favour a shift away.
Put independent uniform (that is, $\text{Dirichlet}(1, 1, 1)$) priors on the two survey probability vectors. By Section 3.4 the posteriors are independent $\text{Dirichlet}(295, 308, 39)$ and $\text{Dirichlet}(289, 333, 20)$, and by Exercise 3.1 the induced marginals of the two Bush-versus-Dukakis proportions are independent with

$$\alpha_1 \mid y \sim \text{Beta}(295, 308), \qquad \alpha_2 \mid y \sim \text{Beta}(289, 333).$$

Their posterior means are $295/603 = 0.489$ and $289/622 = 0.465$.

Drawing 2×10^5 independent pairs from these two beta distributions and forming $\alpha_2 - \alpha_1$ gives

$$\begin{aligned} \mathrm{E}(\alpha_2 - \alpha_1 \mid y) &= -0.025, \\ \mathrm{sd}(\alpha_2 - \alpha_1 \mid y) &= 0.028, \end{aligned}$$

(Monte Carlo standard error 0.0001 on each), with central 95% posterior interval $[-0.080,\ 0.031]$ and

$$\Pr(\alpha_2 - \alpha_1 > 0 \mid y) = 0.19.$$

The histogram is essentially normal, centred just below zero, with nearly a fifth of its mass to the right of 0.

Problem 3.3. Estimation from two independent experiments: an experiment was performed on the effects of magnetic fields on the flow of calcium out of chicken brains. Two groups of chickens were involved: a control group of 32 chickens and an exposed group of 36 chickens. One measurement was taken on each chicken, and the purpose of the experiment was to measure the average flow μ_c in untreated (control) chickens and the average flow μ_t in treated chickens. The 32 measurements on the control group had a sample mean of 1.013 and a sample standard deviation of 0.24. The 36 measurements on the treatment group had a sample mean of 1.173 and a sample standard deviation of 0.20.
(a) Assuming the control measurements were taken at random from a normal distribution with mean μ_c and variance σ_c^2, what is the posterior distribution of μ_c? Similarly, use the treatment group measurements to determine the marginal posterior distribution of μ_t. Assume a uniform prior distribution on $(\mu_c, \mu_t, \log \sigma_c, \log \sigma_t)$.
(b) What is the posterior distribution for the difference, $\mu_t - \mu_c$? To get this, you may sample from the independent t distributions you obtained in part (a) above. Plot a histogram of your samples and give an approximate 95% posterior interval for $\mu_t - \mu_c$.
The problem of estimating two normal means with unknown ratio of variances is called the Behrens–Fisher problem.

Solution. (a) $\mu_c \mid y \sim t_{31}(1.013,\ 0.24^2/32)$ and $\mu_t \mid y \sim t_{35}(1.173,\ 0.20^2/36)$, independently.
The uniform prior on $(\mu_c, \mu_t, \log \sigma_c, \log \sigma_t)$ factors across the two groups and is the noninformative prior $p(\mu, \sigma^2) \propto \sigma^{-2}$ of Section 3.2 within each group; the two likelihoods also factor, so the two posteriors are independent and each is the standard one derived there,

$$\mu \mid y \sim t_{n-1}\big(\bar{y},\ s^2/n\big)\,.$$

Numerically the scales are $s_c/\sqrt{n_c} = 0.24/\sqrt{32} = 0.04243$ and $s_t/\sqrt{n_t} = 0.20/6 = 0.03333$.
(b) $\mu_t - \mu_c$ is the difference of these two independent t variables; it has no closed form (this is the Behrens–Fisher problem), so simulate. Its mean and variance are available exactly, since t_ν has variance $\nu/(\nu-2)$ times its squared scale:

$$\begin{aligned} \mathrm{E}(\mu_t - \mu_c \mid y) &= 1.173 - 1.013 = 0.160, \\ \mathrm{var}(\mu_t - \mu_c \mid y) &= \tfrac{0.20^2}{36} \cdot \tfrac{35}{33} + \tfrac{0.24^2}{32} \cdot \tfrac{31}{29}, \end{aligned}$$

giving $\mathrm{sd}(\mu_t - \mu_c \mid y) = 0.0557$. Drawing 2×10^5 independent pairs from the two t distributions reproduces these (0.1602 and 0.0557) and gives the central 95% posterior interval

$$\mu_t - \mu_c \in [0.050,\ 0.269],$$

with $\Pr(\mu_t > \mu_c \mid y) = 0.997$. The histogram is unimodal and symmetric about 0.16, with slightly heavier tails than a normal and essentially all of its mass positive.

Problem 3.4. Inference for a 2×2 table: an experiment was performed to estimate the effect of beta-blockers on mortality of cardiac patients. A group of patients were randomly assigned to treatment and control groups: out of 674 patients receiving the control, 39 died, and out of 680 receiving the treatment, 22 died. Assume that the outcomes are independent and binomially distributed, with probabilities of death of p_0 and p_1 under the control and treatment, respectively. We return to this example in Section 5.6.
(a) Set up a noninformative prior distribution on (p_0, p_1) and obtain posterior simulations.
(b) Summarize the posterior distribution for the *odds ratio*, $(p_1/(1-p_1))/(p_0/(1-p_0))$.
(c) Discuss the sensitivity of your inference to your choice of noninformative prior density.

Solution. (a) Take independent uniform priors, $p_0, p_1 \sim \text{Beta}(1,1)$. By Section 2.1 the posteriors are independent,

$$p_0 \mid y \sim \text{Beta}(40, 636), \qquad p_1 \mid y \sim \text{Beta}(23, 659),$$

and 2×10^5 independent draws from each beta distribution are posterior simulations of (p_0, p_1).
(b) Transform each draw to $\rho = \dfrac{p_1/(1-p_1)}{p_0/(1-p_0)}$. The simulations give

$$\begin{aligned} \text{median}(\rho \mid y) &= 0.55, \qquad \text{E}(\rho \mid y) = 0.57, \\ 95\%\ \text{interval} &= [0.32,\ 0.93], \\ \Pr(\rho < 1 \mid y) &= 0.988. \end{aligned}$$

On the log scale the posterior is close to normal, $\text{E}(\log\rho \mid y) = -0.60$ with sd $= 0.27$. Beta-blockers reduce the odds of death by roughly 45%, and the posterior gives odds of about 80 to 1 that the effect is beneficial.
(c) Negligible. Repeating (a)–(b) with the Jeffreys prior $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ and with the Haldane prior $\text{Beta}(0,0)$ (proper posterior here, since no cell count is 0) gives

$$\begin{aligned} \text{Beta}(\tfrac{1}{2},\tfrac{1}{2}): &\quad \text{median } 0.55,\ [0.32,\ 0.92],\ \Pr(\rho<1 \mid y) = 0.989, \\ \text{Beta}(0,0): &\quad \text{median } 0.54,\ [0.31,\ 0.92],\ \Pr(\rho<1 \mid y) = 0.989. \end{aligned}$$

Each prior contributes at most one pseudo-observation per cell against 39 and 22 deaths, so it moves the posterior median of ρ by under 2% and the interval endpoints by about 0.01. The choice would matter only if a cell count were near zero.

Problem 3.5. Rounded data: it is a common problem for measurements to be observed in rounded form (for a review, see Heitjan, 1989). For a simple example, suppose we weigh an object five times and measure weights, rounded to the nearest pound, of 10, 10, 12, 11, 9. Assume the unrounded measurements are normally distributed with a noninformative prior distribution on the mean μ and variance σ^2.
(a) Give the posterior distribution for (μ, σ^2) obtained by pretending that the observations are exact unrounded measurements.
(b) Give the correct posterior distribution for (μ, σ^2) treating the measurements as rounded.
(c) How do the incorrect and correct posterior distributions differ? Compare means, variances, and contour plots.
(d) Let $z = (z_1, \ldots, z_5)$ be the original, unrounded measurements corresponding to the five observations above. Draw simulations from the posterior distribution of z. Compute the posterior mean of $(z_1 - z_2)^2$.

Solution. (a) With $n = 5$, $\bar{y} = 10.4$ and $s^2 = 1.3$, the noninformative-prior results (3.5) and (3.3) of Section 3.2 give

$$\sigma^2 \mid y \sim \text{Inv-}\chi^2(4,\ 1.3), \qquad \mu \mid \sigma^2, y \sim \text{N}(10.4,\ \sigma^2/5),$$

so that $\mu \mid y \sim t_4(10.4,\ 0.26)$.

(b) The data are $y_i = \text{round}(z_i)$, so observing y_i is exactly the event $z_i \in (y_i - \frac{1}{2},\, y_i + \frac{1}{2})$. Hence the correct likelihood is the probability of that event, not a density:

$$p(y \mid \mu, \sigma^2) = \prod_{i=1}^{5} \left[\Phi\left(\frac{y_i + \frac{1}{2} - \mu}{\sigma} \right) - \Phi\left(\frac{y_i - \frac{1}{2} - \mu}{\sigma} \right) \right],$$

and with $p(\mu, \sigma^2) \propto \sigma^{-2}$,

$$p(\mu, \sigma^2 \mid y) \propto \sigma^{-2} \prod_{i=1}^{5} \left[\Phi\left(\frac{y_i + \frac{1}{2} - \mu}{\sigma} \right) - \Phi\left(\frac{y_i - \frac{1}{2} - \mu}{\sigma} \right) \right].$$

This has no conjugate form; evaluate it on a grid in $(\mu, \log \sigma)$.
(c) They agree about μ and disagree about σ: rounding inflates the naive variance estimate. Normalizing both densities on a grid covering $\mu \in [-20, 41]$, $\sigma \in [0.05, 400]$:

quantity	ignoring rounding	rounded likelihood
$\mathrm{E}(\mu)$	10.400	10.401
$\mathrm{sd}(\mu)$	0.721	0.708
$\mathrm{E}(\sigma)$	1.429	1.370
$\mathrm{E}(\sigma^2)$	2.599	2.429

The posterior of μ is unchanged because rounding is symmetric about each y_i; the posterior of σ^2 shrinks by $2.429/2.599 = 0.935$, almost exactly Sheppard's correction factor $(s^2 - \frac{1}{12})/s^2 = 1.217/1.3 = 0.936$, the naive s^2 having absorbed the variance $\frac{1}{12}$ of the rounding error. Contour plots of the two densities in (μ, σ) are nested teardrop-shaped curves, symmetric about $\mu = 10.4$ and flaring outwards in μ as σ grows; the two sets very nearly coincide, with the rounded-likelihood contours displaced downwards in σ by a few percent.
(d) $\mathrm{E}\big[(z_1 - z_2)^2 \mid y\big] = 0.159$.
Given (μ, σ) and y, the z_i are independent, z_i being $\mathrm{N}(\mu, \sigma^2)$ truncated to $(y_i - \frac{1}{2},\, y_i + \frac{1}{2})$; simulate by the inverse-cdf method. Since $y_1 = y_2 = 10$, z_1 and z_2 are conditionally i.i.d.\ on $(9.5, 10.5)$, so

$$\mathrm{E}\big[(z_1 - z_2)^2 \mid \mu, \sigma\big] = 2\,\mathrm{var}(z_1 \mid \mu, \sigma),$$

and averaging this truncated-normal variance over the grid posterior of (b) gives 0.1592 without simulation error; 2×10^5 raw draws of $(z_1 - z_2)^2$ reproduce it as 0.159, Monte Carlo standard error 0.0003.

Problem 3.6. Binomial with unknown probability and sample size: some of the difficulties with setting prior distributions in multiparameter models can be illustrated with the simple binomial distribution. Consider data $y_1, \ldots, y_n$ modeled as independent $\mathrm{Bin}(N, \theta)$, with both N and θ unknown. Defining a convenient family of prior distributions on (N, θ) is difficult, partly because of the discreteness of N.
Raftery (1988) considers a hierarchical approach based on assigning the parameter N a Poisson distribution with *unknown* mean μ. To define a prior distribution on (θ, N), Raftery defines $\lambda = \mu\theta$ and specifies a prior distribution on (λ, θ). The prior distribution is specified in terms of λ rather than μ because 'it would seem easier to formulate prior information about λ, the unconditional expectation of the observations, than about μ, the mean of the unobserved quantity N.'
(a) A suggested noninformative prior distribution is $p(\lambda, \theta) \propto \lambda^{-1}$. What is a motivation for this noninformative distribution? Is the distribution improper? Transform to determine $p(N, \theta)$.
(b) The Bayesian method is illustrated on counts of waterbuck obtained by remote photography on five separate days in Kruger Park in South Africa. The counts were 53, 57, 66, 67, and 72. Perform the Bayesian analysis on these data and display a scatterplot of posterior simulations of (N, θ). What is the posterior probability that $N > 100$?
(c) Why not simply use a Poisson with fixed μ as a prior distribution for N?

Solution. (a) It is uniform in $(\log\lambda, \theta)$: $\lambda = \mathrm{E}(y_i)$ is a positive scale parameter, for which Section 2.8 prescribes the flat prior on the log scale, and θ is a probability on the bounded range $[0, 1]$, for which the flat prior is the natural noninformative choice. The distribution is improper, since $\int_0^\infty \lambda^{-1}\, d\lambda$ diverges at both ends.

For the transformation, first change variables from λ to $\mu = \lambda/\theta$ at fixed θ; the Jacobian is $|\partial\lambda/\partial\mu| = \theta$, so

$$p(\mu, \theta) \propto (\mu\theta)^{-1} \cdot \theta = \mu^{-1}.$$

Now average the Poisson prior for N over μ:

$$\begin{aligned} p(N, \theta) &\propto \int_0^\infty \frac{e^{-\mu}\mu^N}{N!}\, \mu^{-1}\, d\mu \\ &= \frac{\Gamma(N)}{N!} = \frac{1}{N}, \qquad N = 1, 2, \ldots \end{aligned}$$

So $p(N, \theta) \propto N^{-1}$ on $N \geq 1$, uniform in θ, and again improper since $\sum_{N\geq 1} N^{-1} = \infty$.

(b) $\Pr(N > 100 \mid y) = 0.96$.

With $n = 5$, $\sum y_i = 315$ and $\max y_i = 72$, the posterior is

$$p(N, \theta \mid y) \propto \frac{1}{N} \prod_{i=1}^{5} \binom{N}{y_i}\, \theta^{315}(1-\theta)^{5N-315},$$

supported on $N \geq 72$. Integrating θ out with the beta integral gives the exact marginal

$$p(N \mid y) \propto \frac{1}{N} \left[\prod_{i=1}^{5} \binom{N}{y_i}\right] B(316,\, 5N - 314),$$

and $\theta \mid N, y \sim \mathrm{Beta}(316,\, 5N - 314)$, so simulation is exact: draw N from the tabulated marginal, then θ from its beta conditional.

This marginal is proper but extremely heavy-tailed. As $N \to \infty$, $\prod_i \binom{N}{y_i} \sim N^{315}/\prod_i y_i!$ and $B(316, 5N - 314) \sim \Gamma(316)(5N)^{-316}$, so

$$p(N \mid y) \sim c\, N^{-2},$$

with $c = 119.8$ after normalizing. Hence $\mathrm{E}(N \mid y) = \infty$ and only quantiles are meaningful:

quantity	2.5%	25%	median	75%	97.5%
N	95	149	235	479	4795

with posterior mode $N = 122$ and $\Pr(N > 100 \mid y) = 0.958$. The corresponding draws of θ have median 0.27 and 95% interval [0.013, 0.66].

The scatterplot of (N, θ) on a log-N axis is a narrow hyperbolic ridge running from $(N, \theta) \approx (80, 0.8)$ down to $(N, \theta) \approx (3000, 0.02)$: $\log N$ and $\log\theta$ have posterior correlation -0.999. What the data determine is the product $\lambda = N\theta$ – posterior median 63.1, 95% interval [57.3, 69.2] – and almost nothing about where along the ridge the truth lies.

(c) Because a fixed μ would *determine* the answer: Poisson(μ) has relative standard deviation $\mu^{-1/2}$, so it pins N to within a few percent, while the data themselves say almost nothing about N given λ – that is the ridge in (b). The posterior would be an artifact of the chosen μ. And μ is the mean of the unobserved N, precisely the quantity Raftery says one has no basis for describing.

Problem 3.7. Poisson and binomial distributions: a student sits on a street corner for an hour and records the number of bicycles b and the number of other vehicles v that go by. Two models are considered:

- The outcomes b and v have independent Poisson distributions, with unknown means θ_b and θ_v.
- The outcome b has a binomial distribution, with unknown probability p and sample size $b + v$.

Show that the two models have the same likelihood if we define $p = \frac{\theta_b}{\theta_b + \theta_v}$.

Solution. The Poisson likelihood factors as (a Poisson likelihood for the total $b + v$, involving only $\tau = \theta_b + \theta_v$) times (exactly the binomial likelihood of the second model). Reparametrize by (p, τ), a bijection of $(0, \infty)^2$ onto $(0, 1) \times (0, \infty)$ with $\theta_b = p\tau$ and $\theta_v = (1 - p)\tau$:

$$\begin{aligned} p(b, v \mid \theta_b, \theta_v) &= \frac{e^{-\theta_b}\theta_b^b}{b!} \cdot \frac{e^{-\theta_v}\theta_v^v}{v!} \\ &= \frac{e^{-\tau}\tau^{b+v}}{b!\, v!}\, p^b(1-p)^v \\ &= \underbrace{\frac{e^{-\tau}\tau^{b+v}}{(b+v)!}}_{\text{Poisson}(b+v|\tau)} \cdot \underbrace{\binom{b+v}{b} p^b(1-p)^v}_{\text{Bin}(b|b+v,p)}, \end{aligned}$$

using $\binom{b+v}{b}/(b+v)! = 1/(b!\, v!)$ on the last line. The first factor is free of p, so as functions of p at any fixed τ – equivalently, conditional on the observed total $b + v$, which is ancillary for p – the two likelihoods are proportional.

3.2 Exercises 3.8–3.14

Problem 3.8. Analysis of proportions: a survey was done of bicycle and other vehicular traffic in the neighborhood of the campus of the University of California, Berkeley, in the spring of 1993. Sixty city blocks were selected at random; each block was observed for one hour, and the numbers of bicycles and other vehicles traveling along that block were recorded. The sampling was stratified into six types of city blocks: busy, fairly busy, and residential streets, with and without bike routes, with ten blocks measured in each stratum. Table 3.3 displays the number of bicycles and other vehicles recorded in the study. For this problem, restrict your attention to the first four rows of the table: the data on residential streets, reproduced here as counts of bicycles/other vehicles. (The data for two of the residential blocks were lost.)

Type of street	Bike route?	Counts of bicycles/other vehicles
Residential	yes	16/58, 9/90, 10/48, 13/57, 19/103, 20/57, 18/86, 17/112, 35/273, 55/64
Residential	no	12/113, 1/18, 2/14, 4/44, 9/208, 7/67, 9/29, 8/154

(a) Let $y_1, \ldots, y_{10}$ and $z_1, \ldots, z_8$ be the observed proportion of traffic that was on bicycles in the residential streets with bike lanes and with no bike lanes, respectively (so $y_1 = 16/(16 + 58)$ and $z_1 = 12/(12 + 113)$, for example). Set up a model so that the y_i's are independent and identically distributed given parameters θ_y and the z_i's are independent and identically distributed given parameters θ_z.
(b) Set up a prior distribution that is independent in θ_y and θ_z.
(c) Determine the posterior distribution for the parameters in your model and draw 1000 simulations from the posterior distribution. (Hint: θ_y and θ_z are independent in the posterior distribution, so they can be simulated independently.)
(d) Let $\mu_y = E(y_i \mid \theta_y)$ be the mean of the distribution of the y_i's; μ_y will be a function of θ_y. Similarly, define μ_z. Using your posterior simulations from (c), plot a histogram of the posterior simulations of $\mu_y - \mu_z$, the expected difference in proportions in bicycle traffic on residential streets with and without bike lanes.
We return to this example in Exercise 5.13.

Solution. (a) Take the two normal models

$$y_i \mid \mu_y, \sigma_y^2 \sim \text{N}(\mu_y, \sigma_y^2), \qquad z_i \mid \mu_z, \sigma_z^2 \sim \text{N}(\mu_z, \sigma_z^2),$$

independently, so $\theta_y = (\mu_y, \sigma_y^2)$ and $\theta_z = (\mu_z, \sigma_z^2)$: the proportions themselves are the data, and the parameters describe block-to-block variation. A single binomial for the counts is untenable here, since it would force all the spread in y_i to come from the block totals, and the ten bike-route proportions range from 0.091 to 0.462 on totals in the hundreds. (Exercise 5.13 repairs this with a hierarchical beta-binomial.)
(b) Use the standard noninformative prior of Section 3.2 separately in each stratum,

$$p(\mu_y, \sigma_y^2, \mu_z, \sigma_z^2) \;\propto\; \sigma_y^{-2}\,\sigma_z^{-2},$$

which is uniform on $(\mu_y, \log\sigma_y, \mu_z, \log\sigma_z)$ and factors into independent pieces.
(c) The two strata are a priori independent and involve disjoint data, so the posterior factors and each factor is given by (3.5) and (3.3):

$$\begin{aligned}\sigma_y^2 \mid y &\sim \text{Inv-}\chi^2(n_y - 1,\; s_y^2), \quad \mu_y \mid \sigma_y^2, y \sim \text{N}(\bar{y},\; \sigma_y^2/n_y),\\ \sigma_z^2 \mid z &\sim \text{Inv-}\chi^2(n_z - 1,\; s_z^2), \quad \mu_z \mid \sigma_z^2, z \sim \text{N}(\bar{z},\; \sigma_z^2/n_z),\end{aligned}$$

with $n_y = 10$, $n_z = 8$. The observed proportions are

$$\begin{aligned}y = (&0.2162,\ 0.0909,\ 0.1724,\ 0.1857,\ 0.1557,\\ &0.2597,\ 0.1731,\ 0.1318,\ 0.1136,\ 0.4622),\\ z = (&0.0960,\ 0.0526,\ 0.1250,\ 0.0833,\ 0.0415,\\ &0.0946,\ 0.2368,\ 0.0494),\end{aligned}$$

giving the sufficient statistics

$$\bar{y} = 0.1961, \quad s_y = 0.1055, \qquad \bar{z} = 0.0974, \quad s_z = 0.0631.$$

Marginally $\mu_y \mid y \sim t_9(\bar{y}, s_y^2/10)$ and $\mu_z \mid z \sim t_7(\bar{z}, s_z^2/8)$, with 95% central intervals [0.121, 0.272] and [0.045, 0.150]. Drawing σ^2 from its scaled inverse-χ^2 and then μ from its normal gives 1000 independent draws of (μ_y, μ_z).
(d) Differencing the draws estimates the posterior of the estimand $\mu_y - \mu_z$:

$$\begin{aligned}E(\mu_y - \mu_z \mid y, z) &= 0.099, &&\text{sd} = 0.046,\\ 95\%\text{ interval} &= [\,0.007,\ 0.190\,], &&\Pr(\mu_y > \mu_z \mid y, z) = 0.98.\end{aligned}$$

These are computed from 10^5 draws; with the 1000 the exercise asks for, the Monte Carlo error on the mean is $0.046/\sqrt{1000} = 0.0015$ and the interval endpoints and tail probability are good only to about ± 0.01. The histogram (figure `bda3-ch03-bike-lane-difference`) is unimodal and slightly right-skewed about 0.10, the skew inherited from the outlying block 55/64, which inflates s_y.

Problem 3.9. Conjugate normal model: suppose y is an independent and identically distributed sample of size n from the distribution $\text{N}(\mu, \sigma^2)$, where the prior distribution for (μ, σ^2) is $\text{N-Inv-}\chi^2(\mu, \sigma^2 \mid \mu_0, \sigma_0^2/\kappa_0;\ \nu_0, \sigma_0^2)$; that is, $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$ and $\mu \mid \sigma^2 \sim \text{N}(\mu_0, \sigma^2/\kappa_0)$. The posterior distribution, $p(\mu, \sigma^2 \mid y)$, is also normal-inverse-χ^2; derive explicitly its parameters in terms of the prior parameters and the sufficient statistics of the data.

Solution. The parameters are

$$\begin{aligned}\mu_n &= \frac{\kappa_0}{\kappa_0 + n}\,\mu_0 + \frac{n}{\kappa_0 + n}\,\bar{y},\\ \kappa_n &= \kappa_0 + n, \qquad \nu_n = \nu_0 + n,\\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2.\end{aligned}$$

Multiplying the prior density (3.6) by the likelihood (3.2) gives, as in (3.7),

$$p(\mu, \sigma^2 \mid y) \propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)}(\sigma^2)^{-n/2}\, e^{-Q/(2\sigma^2)},$$

$$Q = \nu_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2 + (n-1)s^2 + n(\bar y - \mu)^2,$$

where $\bar y$ and s^2 are the sufficient statistics of (3.2). Everything rests on the identity

$$\begin{aligned}\kappa_0(\mu-\mu_0)^2 + n(\mu-\bar y)^2 &= (\kappa_0+n)\mu^2 - 2(\kappa_0\mu_0 + n\bar y)\mu \\ &\qquad + \kappa_0\mu_0^2 + n\bar y^2 \\ &= \kappa_n(\mu-\mu_n)^2 \\ &\qquad + \kappa_0\mu_0^2 + n\bar y^2 - \kappa_n\mu_n^2,\end{aligned}$$

with $\kappa_n = \kappa_0 + n$ and $\mu_n = (\kappa_0\mu_0 + n\bar y)/\kappa_n$ read off as the coefficient ratio. The leftover constant simplifies:

$$\begin{aligned}\kappa_0\mu_0^2 + n\bar y^2 - \frac{(\kappa_0\mu_0 + n\bar y)^2}{\kappa_0+n} &= \frac{\kappa_0 n(\mu_0^2 - 2\mu_0\bar y + \bar y^2)}{\kappa_0+n} \\ &= \frac{\kappa_0 n}{\kappa_0+n}(\bar y - \mu_0)^2.\end{aligned}$$

(Check! Expand the numerator of the first line over the common denominator $\kappa_0 + n$.) Hence

$$Q = \nu_n\sigma_n^2 + \kappa_n(\mu-\mu_n)^2$$

with $\nu_n\sigma_n^2$ as displayed above, and

$$p(\mu, \sigma^2 \mid y) \propto \sigma^{-1}(\sigma^2)^{-(\nu_n/2+1)} \exp\Big[-\tfrac{1}{2\sigma^2}\big(\nu_n\sigma_n^2 + \kappa_n(\mu-\mu_n)^2\big)\Big],$$

since $\nu_0/2 + 1 + n/2 = \nu_n/2 + 1$. Comparing with (3.6) identifies this as N-Inv-$\chi^2(\mu_n, \sigma_n^2/\kappa_n;\ \nu_n, \sigma_n^2)$.

Problem 3.10. Comparison of normal variances: for $j = 1, 2$, suppose that

$$y_{j1}, \ldots, y_{jn_j} \mid \mu_j, \sigma_j^2 \sim \text{iid } \mathrm{N}(\mu_j, \sigma_j^2),$$
$$p(\mu_j, \sigma_j^2) \propto \sigma_j^{-2},$$

and (μ_1, σ_1^2) are independent of (μ_2, σ_2^2) in the prior distribution. Show that the posterior distribution of $(s_1^2/s_2^2)/(\sigma_1^2/\sigma_2^2)$ is F with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom. (Hint: to show the required form of the posterior density, you do not need to carry along all the normalizing constants.)

Solution. The quantity is a ratio of two independent χ^2 variables, each over its own degrees of freedom. The prior factors across j and the two samples are independent given the parameters, so the posterior factors too:

$$p(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2 \mid y) = \prod_{j=1}^{2} p(\mu_j, \sigma_j^2 \mid y_j),$$

each factor being the noninformative-prior normal posterior of Section 3.2. Its marginal for the variance is (3.5):

$$\sigma_j^2 \mid y \sim \text{Inv-}\chi^2(n_j - 1,\ s_j^2), \qquad j = 1, 2,$$

independently, where s_j^2 is the sample variance of group j (and $n_j \geq 2$, so the degrees of freedom are positive). By the definition of the scaled inverse-χ^2 distribution (Appendix A), $X \sim \text{Inv-}\chi^2(\nu, s^2)$ means exactly $\nu s^2/X \sim \chi^2_\nu$; hence

$$u_j := \frac{(n_j-1)s_j^2}{\sigma_j^2} \;\Big|\; y \;\sim\; \chi^2_{n_j-1},$$

and u_1, u_2 are posterior independent. Now simply rearrange the quantity of interest:

$$\frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{u_1/(n_1-1)}{u_2/(n_2-1)}.$$

The right-hand side is a ratio of two independent χ^2 variables each divided by its own degrees of freedom, which is the definition of the F distribution (Appendix A). Therefore

$$\frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} \;\Big|\; y \;\sim\; F_{n_1-1,\,n_2-1}.$$

Problem 3.11. Computation: in the bioassay example, replace the uniform prior density by a joint normal prior distribution on (α, β), with $\alpha \sim \mathrm{N}(0, 2^2)$, $\beta \sim \mathrm{N}(10, 10^2)$, and $\mathrm{corr}(\alpha, \beta) = 0.5$. The bioassay data of Table 3.1 (Racine et al., 1986) are

Dose, x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

with model $y_i \mid \theta_i \sim \mathrm{Bin}(n_i, \theta_i)$ and $\mathrm{logit}(\theta_i) = \alpha + \beta x_i$, as in (3.14) and (3.15). Figure 3.3 of the text displays, for the uniform prior $p(\alpha, \beta) \propto 1$, (a) a contour plot of the posterior density on $[-5, 10] \times [-10, 40]$, with contours at $0.05, 0.15, \ldots, 0.95$ times the density at the mode, and (b) a scatterplot of 1000 posterior draws; both show a banana-shaped, strongly positively correlated ridge running up and to the right from about $(\alpha, \beta) = (0, 0)$.
(a) Repeat all the computations and plots of Section 3.7 with this new prior distribution.
(b) Check that your contour plot and scatterplot look like a compromise between the prior distribution and the likelihood (as displayed in Figure 3.3).
(c) Discuss the effect of this hypothetical prior information on the conclusions in the applied context.

Solution. (a) The prior is the bivariate normal

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}0\\ 10\end{pmatrix},\; \begin{pmatrix}4 & 10\\ 10 & 100\end{pmatrix}\right),$$

since $\mathrm{cov}(\alpha, \beta) = 0.5 \cdot 2 \cdot 10 = 10$, so by (3.15) the unnormalized log posterior is

$$\begin{aligned}\log p(\alpha, \beta \mid y) = \sum_{i=1}^{4} \Big[& y_i(\alpha + \beta x_i) \\ & - n_i \log\!\big(1 + e^{\alpha + \beta x_i}\big)\Big] \\ & - \tfrac{1}{2} \begin{pmatrix}\alpha\\ \beta - 10\end{pmatrix}^{T} \begin{pmatrix}4 & 10\\ 10 & 100\end{pmatrix}^{-1} \begin{pmatrix}\alpha\\ \beta - 10\end{pmatrix} + \text{ const.}\end{aligned}$$

Evaluating this on a 600×600 grid over $[-5, 10] \times [-10, 40]$ (the same range as Figure 3.3), subtracting the maximum before exponentiating, and normalizing the grid to total probability 1 gives a posterior mode at $(\hat\alpha, \hat\beta) = (0.71,\ 7.95)$ and moments

$$\begin{aligned} E(\alpha \mid y) &= 0.98, \quad \mathrm{sd}(\alpha \mid y) = 0.90, \\ E(\beta \mid y) &= 10.48, \quad \mathrm{sd}(\beta \mid y) = 4.60, \end{aligned} \qquad \mathrm{corr}(\alpha, \beta \mid y) = 0.60.$$

Drawing 1000 values by the marginal-then-conditional grid method of Section 3.7 (draw α^s from the numerically summed $p(\alpha \mid y)$, then β^s from $p(\beta \mid \alpha^s, y)$, then jitter by a uniform of grid width) gives

$$\begin{aligned} \alpha: &\quad \text{mean } 0.94, \;\; 95\%\ [-0.56,\ 2.81], \\ \beta: &\quad \text{mean } 10.27, \;\; 95\%\ [3.53,\ 20.42], \end{aligned}$$

so the Monte Carlo error on those means is $0.90/\sqrt{1000} = 0.03$ and $4.60/\sqrt{1000} = 0.15$. All 1000 draws have $\beta > 0$; summing the grid gives $\Pr(\beta > 0 \mid y) = 0.99999$. The LD50 $= -\alpha/\beta$, conditional on $\beta > 0$ as in Section 3.7, has posterior median -0.093 and 95% interval $[-0.246,\ 0.119]$ on the log-dose scale (grid values; the 1000 draws reproduce them to ± 0.01).
(b) Compare the three sets of moments. The prior has $(E\alpha, E\beta) = (0, 10)$ with sds $(2, 10)$; the likelihood alone, that is the uniform-prior posterior of Section 3.7, has

$$E(\alpha \mid y) = 1.31, \quad \text{sd} = 1.10,$$
$$E(\beta \mid y) = 11.61, \quad \text{sd} = 5.71,$$

and the new posterior sits strictly between the two in location, $0 < 0.98 < 1.31$ and $10 < 10.48 < 11.61$, and is tighter than either in scale, $0.90 < \min(2, 1.10)$ and $4.60 < \min(10, 5.71)$. The contour plot keeps the banana shape of Figure 3.3a — the prior is too diffuse to impose its elliptical contours — but is pulled down and to the left and is visibly narrower, and the scatterplot of the 1000 draws is correspondingly a slightly compressed version of Figure 3.3b. (Figure `bda3-ch03-bioassay-normal-prior`.)
(c) The prior is weak relative to the data for β (prior sd 10 against a likelihood sd of 5.7) and only comparable for α (prior sd 2 against 1.1), so it changes nothing qualitative: the compound is still judged harmful with posterior probability above 0.999, and the LD50 is still about -0.09 on the log-dose scale, a dose of roughly $e^{-0.09} = 0.92$ g/ml. What the prior buys is precision: the posterior sds fall by 18% for α and 20% for β, and the LD50 interval narrows from $[-0.276, 0.104]$ to $[-0.246, 0.119]$. Since the prior encodes a genuine expectation from related bioassays — a steep positive slope near 10 and an intercept near 0, meaning a compound roughly 50-50 lethal at unit dose — the gain is legitimate; but with only 20 animals the analyst should say plainly that the shift in the LD50 median from -0.112 to -0.093 is prior-driven, not data-driven.

Problem 3.12. Poisson regression model: expand the model of Exercise 2.13(a) — worldwide airline fatal accidents modeled as independent Poisson counts — by assuming that the number of fatal accidents in year t follows a Poisson distribution with mean $\alpha + \beta t$. The data are the first two columns of Table 2.2:

Year	Fatal accidents	Passenger deaths	Death rate
1976	24	734	0.19
1977	25	516	0.12
1978	31	754	0.15
1979	31	877	0.16
1980	22	814	0.14
1981	21	362	0.06
1982	26	764	0.13
1983	20	809	0.13
1984	16	223	0.03
1985	22	1066	0.15

(Death rate is passenger deaths per 100 million passenger miles.) You will estimate α and β, following the example of the analysis in Section 3.7.
(a) Discuss various choices for a 'noninformative' prior for (α, β). Choose one.
(b) Discuss what would be a realistic informative prior distribution for (α, β). Sketch its contours and then put it aside. Do parts (c)-(h) of this problem using your noninformative prior distribution from (a).
(c) Write the posterior density for (α, β). What are the sufficient statistics?
(d) Check that the posterior density is proper.
(e) Calculate crude estimates and uncertainties for (α, β) using linear regression.
(f) Plot the contours and take 1000 draws from the joint posterior density of (α, β).
(g) Using your samples of (α, β), plot a histogram of the posterior density for the expected number of fatal accidents in 1986, $\alpha + 1986\beta$.

(h) Create simulation draws and obtain a 95% predictive interval for the number of fatal accidents in 1986.
(i) How does your hypothetical informative prior distribution in (b) differ from the posterior distribution in (f) and (g), obtained from the noninformative prior distribution and the data? If they disagree, discuss.

Solution. (a) Take $p(\alpha,\beta)\propto 1$ on the region

$$R=\{(\alpha,\beta):\ \alpha+\beta t>0 \text{ for } t=1976,\dots,1985\},$$

which is forced on us, since the Poisson mean must be positive at every observed year. Three candidates are natural. The flat density on R is invariant under re-origining time, $t\mapsto t-c$, which sends $(\alpha,\beta)\mapsto(\alpha+c\beta,\beta)$ with unit Jacobian, and this is the only reparametrization the problem really invites; it is improper but gives a proper posterior by (d). A second candidate is the Jeffreys prior

$$p(\alpha,\beta)\propto\Big|\textstyle\sum_t \frac{1}{\alpha+\beta t}\binom{1}{t}\binom{1}{t}^{T}\Big|^{1/2},$$

which is invariant but depends on the design and pulls mass toward the boundary of R. A third is flat in $(\log\alpha,\log\beta)$, but β is expected to be negative here, so a log scale is inadmissible. We use the flat prior on R.
(b) A realistic informative prior would say: worldwide fatal accidents run at a few dozen per year and are slowly declining. In the re-origined parameters $\gamma=\alpha+1980.5\beta$ (the expected count at the center of the data) and β, take

$$\gamma\sim \mathrm{N}(25,\,5^2),\qquad \beta\sim \mathrm{N}(-0.5,\,1^2),$$

independent, whose contours are axis-aligned ellipses centered at $(25,-0.5)$ in (γ,β), five times as wide in γ as tall in β. Put aside.
(c) By (3.15) with the flat prior,

$$\begin{aligned}p(\alpha,\beta\mid y)&\propto\prod_{t=1976}^{1985}(\alpha+\beta t)^{y_t}e^{-(\alpha+\beta t)}\\&=\exp\Big[-\big(10\alpha+\beta\textstyle\sum_t t\big)\Big]\prod_t(\alpha+\beta t)^{y_t},\end{aligned}$$

on R and zero off it, with $\sum_t t=19805$. There is no reduction: the minimal sufficient statistic is the full vector $y=(y_{1976},\dots,y_{1985})$ of ten counts. The terms $y_t\log(\alpha+\beta t)$ cannot be collected into finitely many data-dependent functions because the identity link makes this a curved, non-exponential family in (α,β). (Under a log link, $\log\mu_t=\alpha+\beta t$, the sufficient statistics would collapse to $\sum_t y_t$ and $\sum_t t\,y_t$.)
(d) Change variables to the endpoint means

$$u=\alpha+1976\beta,\qquad v=\alpha+1985\beta,$$

so $d\alpha\,d\beta=du\,dv/9$ and, writing $t_j=1975+j$ and $w_j=(j-1)/9$ for $j=1,\dots,10$,

$$\mu_j:=\alpha+\beta t_j=(1-w_j)u+w_jv.$$

Each μ_j is a convex combination of u and v, so $R=\{u>0,\ v>0\}$ and $\mu_j\le u+v$. Also $\sum_j w_j=45/9=5$, hence $\sum_j\mu_j=5(u+v)$. With $Y=\sum_t y_t=238$,

$$\begin{aligned}\int_R\prod_j\mu_j^{y_j}e^{-\sum_j\mu_j}\,d\alpha\,d\beta&\le\frac19\int_0^\infty\!\!\int_0^\infty(u+v)^Ye^{-5(u+v)}du\,dv\\&=\frac19\int_0^\infty s^{Y+1}e^{-5s}\,ds\\&=\frac{\Gamma(240)}{9\cdot5^{240}}<\infty,\end{aligned}$$

the middle step by substituting $s = u + v$ and integrating over $u \in (0, s)$, which contributes the extra factor s. The posterior is therefore proper.
(e) Least squares of y_t on t, computed with the centered predictor $x_t = t - 1980.5$ (so $\sum x_t^2 = 82.5$), gives

$$\hat{\gamma} = \bar{y} = 23.80 \quad (\text{se } 1.27),$$
$$\hat{\beta} = -0.921 \qquad (\text{se } 0.443),$$

with residual standard deviation $s = 4.02$ on 8 degrees of freedom. Transforming back, $\hat{\alpha} = \hat{\gamma} - 1980.5\hat{\beta} = 1848$ with standard error 877 — the huge value and its huge uncertainty are artifacts of extrapolating the line back to $t = 0$, and $(\hat{\alpha}, \hat{\beta})$ have correlation -1.000 to three decimals.
(f) Evaluate the log posterior of (c) on a 900×900 grid in $(\gamma, \beta) \in [10, 40] \times [-3, 2]$, zeroing the cells outside R, subtract the maximum, exponentiate and normalize. The mode is at $(\gamma, \beta) = (23.82, -0.948)$ and

$$E(\gamma \mid y) = 24.00, \qquad \text{sd}(\gamma \mid y) = 1.55,$$
$$E(\beta \mid y) = -0.940, \quad \text{sd}(\beta \mid y) = 0.540.$$

The contours are close to elliptical and nearly axis-aligned in (γ, β), which is exactly why the centered parameterization is the right one to draw. Sampling 1000 draws by the marginal-then-conditional grid method of Section 3.7 and mapping back by $\alpha = \gamma - 1980.5\beta$ gives

$$\beta: \ \text{mean} \ -0.94, \quad 95\% \ [-1.99, \ 0.12],$$

matching the grid to within the Monte Carlo error $0.54/\sqrt{1000} = 0.017$ on the mean. Summing the grid, $\Pr(\beta < 0 \mid y) = 0.96$: the evidence for a declining trend is suggestive but not decisive. (Figure `bda3-ch03-poisson-regression`.)
(g) For each draw form $\mu_{1986} = \alpha + 1986\beta = \gamma + 5.5\beta$. The histogram is unimodal and nearly symmetric with

$$E(\mu_{1986} \mid y) = 18.8, \quad \text{sd} = 3.19, \quad 95\% \ [12.9, \ 25.4].$$

(h) Draw $\tilde{y}^s \sim \text{Poisson}(\mu_{1986}^s)$ for each draw of μ_{1986}. The predictive distribution has mean 18.8 and standard deviation 5.4, and its 2.5% and 97.5% quantiles give

$$\tilde{y}_{1986} \in [9, \ 30].$$

Discreteness, and the Monte Carlo error of 1000 draws, leave each endpoint uncertain by about one count.
(i) They agree. The informative prior of (b) put γ at 25 ± 5 and β at -0.5 ± 1; the data-driven posterior of (f) puts them at 24.0 ± 1.55 and -0.940 ± 0.540. Both prior means lie well inside the corresponding posterior intervals, and the prior is three to four times wider on each coordinate, so it would have contributed almost nothing had it been used. The one place the analyses part company is emphasis: the prior asserted a declining trend outright, whereas the data leave a 4% posterior probability that the trend is flat or rising.

Problem 3.13. Multivariate normal model: derive equations (3.12) by completing the square in vector-matrix notation. That is, for $y_1, \ldots, y_n$ independent and identically distributed $\text{N}(\mu, \Sigma)$ with Σ known and prior $\mu \sim \text{N}(\mu_0, \Lambda_0)$, show that $p(\mu \mid y, \Sigma) = \text{N}(\mu \mid \mu_n, \Lambda_n)$ with

$$\mu_n = \left(\Lambda_0^{-1} + n\Sigma^{-1}\right)^{-1}\left(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}\right),$$
$$\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}.$$

Solution. Complete the square on the quadratic form in μ. From (3.11) and the normal prior,

$$\begin{aligned} &p(\mu \mid y, \Sigma) \\ &\propto \exp\Big[-\tfrac{1}{2}\Big((\mu-\mu_0)^T \Lambda_0^{-1} (\mu-\mu_0) \\ &\qquad + \textstyle\sum_{i=1}^n (y_i-\mu)^T \Sigma^{-1} (y_i-\mu) \Big)\Big]. \end{aligned}$$

Split the data term about $\bar y = \frac{1}{n}\sum_i y_i$, the cross terms vanishing because $\sum_i (y_i - \bar y) = 0$:

$$\begin{aligned} \sum_{i=1}^n (y_i-\mu)^T \Sigma^{-1} (y_i-\mu) = &\sum_{i=1}^n (y_i-\bar y)^T \Sigma^{-1} (y_i-\bar y) \\ &+ n(\mu-\bar y)^T \Sigma^{-1} (\mu-\bar y), \end{aligned}$$

and the first sum is free of μ, so it goes into the proportionality constant. What remains is

$$\begin{aligned} Q(\mu) &= (\mu-\mu_0)^T \Lambda_0^{-1} (\mu-\mu_0) \\ &\quad + n(\mu-\bar y)^T \Sigma^{-1} (\mu-\bar y) \\ &= \mu^T \big(\Lambda_0^{-1} + n\Sigma^{-1}\big)\mu \\ &\quad - 2\mu^T \big(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar y\big) + c, \end{aligned}$$

where $c = \mu_0^T \Lambda_0^{-1} \mu_0 + n \bar y^T \Sigma^{-1} \bar y$ is constant in μ; the two cross terms combined into one because Λ_0^{-1} and Σ^{-1} are symmetric, so $\mu^T A \mu_0 = \mu_0^T A \mu$ for A symmetric. Set

$$\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}, \qquad \mu_n = \Lambda_n\big(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar y\big).$$

This is legitimate: Λ_0 and Σ are positive definite variance matrices, hence so are their inverses and the sum Λ_n^{-1}, which is therefore invertible. Then

$$\begin{aligned} (\mu-\mu_n)^T \Lambda_n^{-1} (\mu-\mu_n) = &\mu^T \Lambda_n^{-1} \mu - 2\mu^T \Lambda_n^{-1} \mu_n \\ &+ \mu_n^T \Lambda_n^{-1} \mu_n, \end{aligned}$$

and $\Lambda_n^{-1}\mu_n = \Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar y$ by construction, so $Q(\mu)$ and this expression agree up to a constant. Hence

$$p(\mu \mid y, \Sigma) \propto \exp\Big[-\tfrac{1}{2}(\mu-\mu_n)^T \Lambda_n^{-1} (\mu-\mu_n)\Big] = \mathrm{N}(\mu \mid \mu_n, \Lambda_n),$$

which is (3.12).

Problem 3.14. Improper prior and proper posterior distributions: prove that the posterior density (3.15) for the bioassay example has a finite integral over the range $(\alpha, \beta) \in (-\infty, \infty) \times (-\infty, \infty)$. Recall that with the uniform prior $p(\alpha, \beta) \propto 1$ and the logistic model (3.14), the unnormalized posterior is

$$p(\alpha, \beta \mid y) \propto \prod_{i=1}^4 \big[\mathrm{logit}^{-1}(\alpha + \beta x_i)\big]^{y_i} \big[1 - \mathrm{logit}^{-1}(\alpha + \beta x_i)\big]^{n_i - y_i},$$

with the data of Table 3.1:

i	Dose, x_i (log g/ml)	n_i	y_i
1	-0.86	5	0
2	-0.30	5	1
3	-0.05	5	3
4	0.73	5	5

Solution. The likelihood is bounded by $e^{-\varphi}$ for a positively homogeneous convex φ that vanishes only at the origin, and such an exponential is integrable on the plane.

Write $u_i = \alpha + \beta x_i$ and $g(u) = \text{logit}^{-1}(u) = (1+e^{-u})^{-1}$. Since $\log(1+e^{-u}) \geq \max(0,-u)$ and $\log(1+e^{u}) \geq \max(0,u)$,

$$\log g(u) = -\log(1+e^{-u}) \leq -\max(0,-u),$$
$$\log\big(1-g(u)\big) = -\log(1+e^{u}) \leq -\max(0,u).$$

Multiplying by the nonnegative weights y_i and $n_i - y_i$ and summing,

$$L(\alpha,\beta) := \prod_{i=1}^{4} g(u_i)^{y_i}\big(1-g(u_i)\big)^{n_i-y_i} \leq e^{-\varphi(\alpha,\beta)},$$

$$\varphi(\alpha,\beta) = \sum_{i=1}^{4}\Big[y_i\,(u_i)^- + (n_i-y_i)\,(u_i)^+\Big],$$

where $(u)^+ = \max(0,u)$ and $(u)^- = \max(0,-u)$.
Two properties of φ. It is nonnegative, and since each $u_i = \alpha + \beta x_i$ is linear (no intercept) while $(\cdot)^{\pm}$ are positively homogeneous, $\varphi(c\alpha, c\beta) = c\,\varphi(\alpha,\beta)$ for every $c > 0$. Second, $\varphi(\alpha,\beta) = 0$ forces, term by term:

$$\begin{aligned}
&\text{(i) } i=1 \ (y_1=0,\ n_1-y_1=5): && u_1 \leq 0,\\
&\text{(ii) } i=2 \ (y_2=1,\ n_2-y_2=4): && u_2 = 0,\\
&\text{(iii) } i=3 \ (y_3=3,\ n_3-y_3=2): && u_3 = 0,\\
&\text{(iv) } i=4 \ (y_4=5,\ n_4-y_4=0): && u_4 \geq 0.
\end{aligned}$$

Leaves (ii) and (iii) alone give $\alpha - 0.30\beta = 0$ and $\alpha - 0.05\beta = 0$; the coefficient matrix has determinant $-0.05 + 0.30 = 0.25 \neq 0$, so $(\alpha,\beta) = (0,0)$. Hence $\varphi > 0$ everywhere on the unit circle S, and φ is continuous, so by compactness

$$c := \min_{(\alpha,\beta)\in S} \varphi(\alpha,\beta) > 0$$

(numerically $c = 0.2497$). Homogeneity then upgrades this to a global bound, $\varphi(v) \geq c\,\|v\|$ for all $v = (\alpha,\beta) \in \mathbb{R}^2$. Therefore, in polar coordinates,

$$\begin{aligned}
\int_{\mathbb{R}^2} L(\alpha,\beta)\,d\alpha\,d\beta &\leq \int_{\mathbb{R}^2} e^{-c\|v\|}\,dv\\
&= 2\pi\int_0^{\infty} r e^{-cr}\,dr = \frac{2\pi}{c^2} < \infty,
\end{aligned}$$

so the posterior density (3.15) is proper.

3.3 Exercises 3.15–3.15

Problem 3.15. Joint distributions: The autoregressive time-series model $y_1, y_2, \ldots$ with mean level 0, autocorrelation 0.8, residual standard deviation 1, and normal errors can be written as

$$(y_t \mid y_{t-1}, y_{t-2}, \ldots) \sim \text{N}(0.8y_{t-1}, 1)$$

for all t.
(a) Prove that the distribution of y_t, given the observations at all other integer time points t, depends only on y_{t-1} and y_{t+1}.
(b) What is the distribution of y_t given y_{t-1} and y_{t+1}?

Solution. (a) Only two factors of the joint density contain y_t. For any $n > t+1$ the model's defining conditionals telescope into

$$p(y_1,\ldots,y_n) \;=\; p(y_1)\prod_{s=2}^{n} p(y_s \mid y_{s-1}),$$

so, treating every y_s with $s \neq t$ as fixed and dropping all factors free of y_t,

$$\begin{aligned} p(y_t \mid y_{-t}) &\propto p(y_t \mid y_{t-1})\, p(y_{t+1} \mid y_t) \\ &\propto \exp\Big(-\tfrac{1}{2}(y_t - 0.8 y_{t-1})^2 \Big) \\ &\quad \times \exp\Big(-\tfrac{1}{2}(y_{t+1} - 0.8 y_t)^2 \Big). \end{aligned}$$

The right-hand side involves the conditioning variables only through y_{t-1} and y_{t+1}, and the normalizing constant is its integral over y_t; being free of n, the result holds for every finite block and hence for the whole series, so $p(y_t \mid y_{-t}) = p(y_t \mid y_{t-1}, y_{t+1})$.
(b) Complete the square in y_t. The exponent above is

$$-\tfrac{1}{2}\Big[(1 + 0.8^2)\, y_t^2 - 2(0.8)(y_{t-1} + y_{t+1})\, y_t\Big] + \text{const},$$

a quadratic with precision $1 + 0.64 = 1.64$ and linear coefficient $0.8(y_{t-1} + y_{t+1})$, so

$$(y_t \mid y_{t-1}, y_{t+1}) \sim \mathrm{N}\!\left(\frac{0.8}{1.64}\,(y_{t-1} + y_{t+1}),\ \frac{1}{1.64} \right),$$

that is, mean $\frac{20}{41}(y_{t-1} + y_{t+1}) = 0.4878(y_{t-1} + y_{t+1})$ and variance $\frac{25}{41} = 0.6098$.

4 Asymptotics and Connections to Non-Bayesian Approaches

Contents

4.1 Exercises 4.1–4.7

Problem 4.1. Normal approximation: suppose that $y_1, \dots, y_5$ are independent samples from a Cauchy distribution with unknown center θ and known scale 1: $p(y_i \mid \theta) \propto 1/(1 + (y_i - \theta)^2)$. Assume that the prior distribution for θ is uniform on $[0, 1]$. Given the observations $(y_1, \dots, y_5) = (-2, -1, 0, 1.5, 2.5)$:
(a) Determine the derivative and the second derivative of the log posterior density.
(b) Find the posterior mode of θ by iteratively solving the equation determined by setting the derivative of the log-likelihood to zero.
(c) Construct the normal approximation based on the second derivative of the log posterior density at the mode. Plot the approximate normal density and compare to the exact density as computed using the approach described in Exercise 2.11.

Solution. (a) On $0 \le \theta \le 1$ the uniform prior is a constant, so $\log p(\theta \mid y) = \text{const} - \sum_{i=1}^{5} \log\bigl(1 + (y_i - \theta)^2\bigr)$ and

$$\frac{d}{d\theta} \log p(\theta \mid y) = \sum_{i=1}^{5} \frac{2(y_i - \theta)}{1 + (y_i - \theta)^2},$$

$$\frac{d^2}{d\theta^2} \log p(\theta \mid y) = \sum_{i=1}^{5} \frac{2\left[(y_i - \theta)^2 - 1\right]}{\left[1 + (y_i - \theta)^2\right]^2},$$

the second line by differentiating $2u/(1 + u^2)$ in $u = y_i - \theta$ and multiplying by $du/d\theta = -1$.
(b) Newton-Raphson, $\theta^{t+1} = \theta^t - L'(\theta^t)/L''(\theta^t)$ with $L = \log p(\cdot \mid y)$, started at the sample median $\theta^0 = 0$:

t	θ^t	$L'(\theta^t)$	$L''(\theta^t)$
0	0.000000	-0.187268	-1.323551
1	-0.141489	0.005278	-1.374539
2	-0.137649	-7.0×10^{-7}	-1.374890
3	-0.137649	$< 10^{-11}$	-1.374890

so the likelihood equation has the unique root $\hat\theta = -0.137649$ (a scan of L' over $[-6, 6]$ shows only this one sign change, so the Cauchy likelihood is unimodal for these data).
This root lies outside the prior support $[0, 1]$, on which $L' < 0$ throughout, so the posterior mode as printed is the boundary point $\theta = 0$; both values are carried into (c).
(c) With $I(\theta) = -L''(\theta)$,

$$I(\hat\theta) = 1.374890, \qquad p(\theta \mid y) \approx \mathrm{N}(-0.1376,\ 0.7273), \quad \text{sd} = 0.8528,$$

and at the boundary mode, $I(0) = 1.323551$, giving N(0, 0.7555), sd = 0.8692. Both are renormalized to $[0, 1]$ for comparison. The exact density is obtained as in Exercise 2.11: evaluate the unnormalized $\exp\{L(\theta)\}$ on a grid of 20001 points in $[0, 1]$ and normalize by the trapezoidal sum. Densities (all normalized on $[0, 1]$):

θ	exact	$\mathrm{N}(-0.1376, 0.7273)$	$\mathrm{N}(0, 0.7555)$
0.0	1.2515	1.3392	1.2238
0.5	1.0012	1.0259	1.0372
1.0	0.7259	0.5573	0.6314

Exact posterior mean and sd on $[0, 1]$: 0.4534 and 0.2845; the truncated normal at $\hat\theta$ gives 0.4310 and 0.2772, the one at the boundary mode 0.4476 and 0.2795. Against the exact density the maximum absolute errors are 0.169 and 0.094 (about 13% and 8% of the peak 1.252): with $n = 5$ both reproduce the gentle downward slope but exaggerate it.

Problem 4.2. Normal approximation: derive the analytic form of the information matrix and the normal approximation variance for the bioassay example.
(The bioassay example of Section 3.7: $y_i \mid \alpha, \beta \sim \mathrm{Bin}(n_i, \theta_i)$ independently, with $\mathrm{logit}(\theta_i) = \alpha + \beta x_i$ as in (3.14), uniform prior $p(\alpha, \beta) \propto 1$, and the data of Table 3.1 from Racine et al. (1986):)

Dose, x_i (log g/ml)	Animals, n_i	Deaths, y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

Solution.

$$I(\alpha,\beta) = \begin{pmatrix} \sum_i n_i\theta_i(1-\theta_i) & \sum_i n_i x_i\theta_i(1-\theta_i) \\ \sum_i n_i x_i\theta_i(1-\theta_i) & \sum_i n_i x_i^2\theta_i(1-\theta_i) \end{pmatrix} = X^T W X,$$

with X the 4×2 matrix of rows $(1, x_i)$ and $W = \mathrm{Diag}\{n_i\theta_i(1-\theta_i)\}$, where $\theta_i = \mathrm{logit}^{-1}(\alpha + \beta x_i)$. Writing $\eta_i = \alpha + \beta x_i$, the log posterior under the uniform prior (3.15) is (the binomial form presumes the n_i animals within a dose group are exchangeable, and the groups independent)

$$\log p(\alpha,\beta \mid y) = \text{const} + \sum_{i=1}^{k} \Big[y_i\eta_i - n_i \log\big(1 + e^{\eta_i}\big) \Big],$$

and since $d\theta_i/d\eta_i = \theta_i(1-\theta_i)$,

$$\begin{aligned} \frac{\partial}{\partial \alpha} \log p &= \sum_i (y_i - n_i\theta_i), \\ \frac{\partial}{\partial \beta} \log p &= \sum_i x_i(y_i - n_i\theta_i), \\ \frac{\partial^2}{\partial \alpha^2} \log p &= -\sum_i n_i\theta_i(1-\theta_i), \\ \frac{\partial^2}{\partial \alpha\, \partial \beta} \log p &= -\sum_i n_i x_i\theta_i(1-\theta_i), \end{aligned}$$

and $\partial^2 \log p/\partial\beta^2 = -\sum_i n_i x_i^2 \theta_i(1-\theta_i)$. The y_i have dropped out (the logit is the canonical link), so the observed information equals the expected information, and I is positive definite whenever the x_i are not all equal. The normal approximation (4.2) is therefore

$$p(\alpha,\beta \mid y) \approx \mathrm{N}\left(\begin{pmatrix} \hat\alpha \\ \hat\beta \end{pmatrix}, \left(X^T \hat W X \right)^{-1} \right), \qquad \hat W = W\big|_{(\hat\alpha,\hat\beta)}.$$

Numerically, the mode (equivalently the maximum likelihood estimate) is $(\hat\alpha, \hat\beta) = (0.8466,\ 7.7487)$, giving

x_i	$\hat\eta_i$	$\hat\theta_i$	$\hat w_i = n_i\hat\theta_i(1-\hat\theta_i)$
-0.86	-5.8173	0.0030	0.0148
-0.30	-1.4780	0.1857	0.7561
-0.05	0.4591	0.6128	1.1864
0.73	6.5031	0.9985	0.0075

so that

$$I(\hat\alpha,\hat\beta) = \begin{pmatrix} 1.9648 & -0.2934 \\ -0.2934 & 0.08594 \end{pmatrix}, \quad I^{-1} = \begin{pmatrix} 1.0385 & 3.5459 \\ 3.5459 & 23.743 \end{pmatrix},$$

that is, posterior standard deviations 1.02 for α and 4.87 for β with correlation 0.714 – the standard errors 1.0 and 4.9 quoted in Section 3.7.

Problem 4.3. Normal approximation to the marginal posterior distribution of an estimand: in the bioassay example, the normal approximation to the joint posterior distribution of (α, β) is obtained. The posterior distribution of any estimand, such as the LD50, can be approximated by a normal distribution fit to its marginal posterior mode and the curvature of the marginal posterior density about the mode. This is sometimes called the 'delta method.' Expand the posterior distribution of the LD50, $-\alpha/\beta$, as a Taylor series around the posterior mode and thereby derive the asymptotic posterior median and standard deviation. Compare to the histogram in Figure 4.2.
(Figure 4.2 is a histogram of $-\alpha/\beta$ computed from 1000 draws of the normal approximation $p(\alpha, \beta \mid y) \approx \mathrm{N}(\hat{\theta}, I(\hat{\theta})^{-1})$ of the bioassay example, restricted to the 950 draws with $\beta > 0$; those 950 values of the LD50 ranged over $[-12.4, 5.4]$, and panel (b) shows the central 95% of them.)

Solution.

$$\mathrm{LD50} \mid y \;\approx\; \mathrm{N}\!\left(-\frac{\hat{\alpha}}{\hat{\beta}},\; \nabla g^T I(\hat{\alpha}, \hat{\beta})^{-1} \nabla g\right) = \mathrm{N}(-0.1093,\; 0.09546^2),$$

so the asymptotic posterior median is -0.1093 and the asymptotic posterior standard deviation is 0.0955.
Write $g(\alpha, \beta) = -\alpha/\beta$ and expand about the joint posterior mode $(\hat{\alpha}, \hat{\beta}) = (0.8466,\ 7.7487)$ of Exercise 4.2:

$$g(\alpha, \beta) = -\frac{\hat{\alpha}}{\hat{\beta}} - \frac{1}{\hat{\beta}}(\alpha - \hat{\alpha}) + \frac{\hat{\alpha}}{\hat{\beta}^2}(\beta - \hat{\beta}) + O\big(\|\cdot\|^2\big)\,,$$

a linear function of (α, β) with gradient $\nabla g = (-1/\hat{\beta},\ \hat{\alpha}/\hat{\beta}^2) = (-0.129053,\ 0.014099)$. Under the normal approximation (4.2), a linear function of a normal vector is exactly normal, whence the display above; its median equals its mean $g(\hat{\alpha}, \hat{\beta}) = -0.10925$, and with I^{-1} from Exercise 4.2,

$$\begin{aligned}\nabla g^T I^{-1} \nabla g &= 1.0385(0.129053)^2 - 2(3.5459)(0.129053)(0.014099) \\ &\qquad + 23.743(0.014099)^2 \\ &= 0.017296 - 0.012903 + 0.004720 = 0.009113,\end{aligned}$$

so $\mathrm{sd} = \sqrt{0.009113} = 0.09546$ and the asymptotic central 95% interval is $-0.1093 \pm 1.96(0.0955) = (-0.296,\ 0.078)$.
Against Figure 4.2: drawing 2×10^5 values from the same $\mathrm{N}(\hat{\theta}, I(\hat{\theta})^{-1})$ and forming $-\alpha/\beta$ on the 94.5% with $\beta > 0$ gives, averaged over twelve replicate runs, median -0.1118 and central 95% interval $(-0.395,\ 0.453)$, with between-run standard errors 0.0002, 0.002 and 0.009 respectively; so the delta-method centre agrees with the histogram but its interval, of width 0.374, is less than half the histogram's 0.848. The histogram's spread is not comparable to 0.0955 at all: since $\hat{\beta}/\mathrm{sd}(\beta) = 7.75/4.87 = 1.59$, the normal approximation puts mass arbitrarily close to $\beta = 0$, the simulated LD50 ranged over several thousand and its sample standard deviation does not converge. The delta method removes that tail by truncating at first order, and its interval is the closer of the two to the exact posterior of Figure 3.4 (grid computation: median -0.112, central 95% interval $(-0.276,\ 0.103)$, $\Pr(\beta > 0) > 0.999$).

Problem 4.4. Asymptotic normality: assuming the regularity conditions hold, we know that $p(\theta \mid y)$ approaches normality as $n \to \infty$. In addition, if $\phi = f(\theta)$ is any one-to-one continuous transformation of θ, we can express the Bayesian inference in terms of ϕ and find that $p(\phi \mid y)$ also approaches normality. But a nonlinear transformation of a normal distribution is no longer normal. How can both limiting normal distributions be valid?

Solution. Because the posterior concentrates: all but a vanishing fraction of the posterior mass sits in a neighborhood of θ_0 of width $O(n^{-1/2})$, and on such a shrinking neighborhood a smooth f is linear to within $o(n^{-1/2})$ – and a linear transformation of a normal distribution is exactly normal.
Quantitatively, let f be twice differentiable with $f'(\theta_0) \neq 0$ – exactly the condition under which the transformed problem still meets the regularity conditions of Section 4.2, since $J_\phi = J/(f')^2$ must stay

finite and positive; a one-to-one f with vanishing derivative, such as $\theta \mapsto \theta^3$ at $\theta_0 = 0$, is excluded. Taylor expansion about θ_0 gives, for some $\tilde{\theta}$ between θ and θ_0,

$$\begin{aligned}\sqrt{n}\big(\phi - f(\theta_0)\big) &= f'(\theta_0)\,\sqrt{n}(\theta - \theta_0) \\ &\quad + \tfrac{1}{2} f''(\tilde{\theta})\, n^{-1/2}\big[\sqrt{n}(\theta - \theta_0)\big]^2.\end{aligned}$$

By Section 4.2, $\sqrt{n}(\theta - \theta_0) \mid y \to \mathrm{N}\big(0, [J(\theta_0)]^{-1}\big)$ in distribution, so the bracket is $O_p(1)$ and the whole second term is $O_p(n^{-1/2}) \to 0$; Slutsky's theorem then gives

$$\sqrt{n}\big(\phi - f(\theta_0)\big) \mid y \;\to\; \mathrm{N}\!\left(0,\; \frac{[f'(\theta_0)]^2}{J(\theta_0)}\right),$$

which is exactly the normal limit predicted by the information for ϕ, $J_\phi(\phi_0) = J(\theta_0)/[f'(\theta_0)]^2$.
Neither posterior is ever exactly normal at finite n; the curvature of f distorts the θ-normal only on scales of order 1, to which the posterior assigns probability tending to 0.

Problem 4.5. Approximate mean and variance:
(a) Suppose x and y are independent normally distributed random variables, where x has mean 4 and standard deviation 1, and y has mean 3 and standard deviation 2. What are the mean and standard deviation of y/x? Compute this using simulation.
(b) Suppose x and y are independent random variables, where x has mean 4 and standard deviation 1, and y has mean 3 and standard deviation 2. What are the approximate mean and standard deviation of y/x? Determine this without using simulation.
(c) What assumptions are required for the approximation in (b) to be reasonable?

Solution. (a) They do not exist. Since the $\mathrm{N}(4,1)$ density is strictly positive at 0, $\mathrm{E}\,|1/x| = \infty$, so y/x has no finite mean and no finite variance; the simulation exhibits exactly this. Drawing 10^5, 10^6 and 10^7 pairs under three seeds gave sample standard deviations

$$0.75,\ 4.65,\ 8.63; \qquad 5.20,\ 3.43,\ 2.73; \qquad 0.73,\ 1.14,\ 34.2,$$

which refuse to settle, with $\max|y/x|$ growing from 90 to 10^5 as n grows – the signature of an undefined second moment. What is stable is the centre of the distribution: across all runs the sample median was 0.750, the central 95% interval was $(-0.238,\ 2.21)$, and the standard deviation of the central 95% of the draws was 0.501. (The sample mean hovered near 0.81, but this too is not a convergent quantity.) So the honest simulation answer is: median 0.750, spread about 0.5 in the bulk, moments infinite.
(b) With only means and standard deviations specified, linearize $g(x,y) = y/x$ about $(\mu_x, \mu_y) = (4, 3)$:

$$\nabla g = \left(-\frac{\mu_y}{\mu_x^2},\ \frac{1}{\mu_x}\right) = \left(-\tfrac{3}{16},\ \tfrac{1}{4}\right),$$

so, using independence,

$$\begin{aligned}\mathrm{E}\Big(\frac{y}{x}\Big) &\approx \frac{\mu_y}{\mu_x} = \frac{3}{4} = 0.75, \\ \mathrm{var}\Big(\frac{y}{x}\Big) &\approx \left(\frac{\mu_y}{\mu_x^2}\right)^2 \sigma_x^2 + \left(\frac{1}{\mu_x}\right)^2 \sigma_y^2 \\ &= \left(\tfrac{3}{16}\right)^2 (1) + \left(\tfrac{1}{4}\right)^2 (4) \\ &= 0.035156 + 0.25 = 0.285156,\end{aligned}$$

giving $\mathrm{sd} = \sqrt{0.285156} = 0.534$. Equivalently $\mathrm{var}(y/x) \approx (\mu_y/\mu_x)^2\left[\sigma_y^2/\mu_y^2 + \sigma_x^2/\mu_x^2\right]$: relative variances add. Both agree with the stable features of (a), median 0.750 and bulk spread 0.501.
(c) That σ_x/μ_x be small – here 0.25 – so that x is effectively never near 0 and $1/x$ is well approximated by its tangent line over the effective range of x; that x and y be independent, or else a $2\,\mathrm{cov}(x,y)\,\partial_x g\,\partial_y g$ term must be added; and that the third and higher moments of x be small enough that the neglected

quadratic term $\mu_y(x-\mu_x)^2/\mu_x^3$ is negligible. None of this makes the exact moments finite when x is normal, as (a) shows; the approximation describes the bulk of the distribution, not its tails.

Problem 4.6. Statistical decision theory: a decision-theoretic approach to the estimation of an unknown parameter θ introduces the loss function $L(\theta, a)$ which, loosely speaking, gives the cost of deciding that the parameter has the value a, when it is in fact equal to θ. The estimate a can be chosen to minimize the posterior expected loss,

$$\mathrm{E}(L(a \mid y)) = \int L(\theta, a)\, p(\theta \mid y)\, d\theta.$$

This optimal choice of a is called a Bayes estimate for the loss function L. Show that:
(a) If $L(\theta, a) = (\theta - a)^2$ (squared error loss), then the posterior mean, $\mathrm{E}(\theta \mid y)$, if it exists, is the unique Bayes estimate of θ.
(b) If $L(\theta, a) = |\theta - a|$, then any posterior median of θ is a Bayes estimate of θ.
(c) If k_0 and k_1 are nonnegative numbers, not both zero, and

$$L(\theta, a) = \begin{cases} k_0(\theta - a) & \text{if } \theta \geq a \\ k_1(a - \theta) & \text{if } \theta < a, \end{cases}$$

then any $\dfrac{k_0}{k_0 + k_1}$ quantile of the posterior distribution $p(\theta \mid y)$ is a Bayes estimate of θ.

Solution. (a) Complete the square, writing $\mu = \mathrm{E}(\theta \mid y)$:

$$\mathrm{E}\big[(\theta - a)^2 \mid y\big] = \mathrm{var}(\theta \mid y) + \big(\mu - a\big)^2,$$

a strictly convex function of a whose unique minimum is at $a = \mu$. (Assume $\mathrm{E}(\theta^2 \mid y) < \infty$; otherwise the expected loss is $+\infty$ for every a and no comparison is possible.)
(b) Let m be any posterior median, so $\Pr(\theta \leq m) \geq \frac{1}{2}$ and $\Pr(\theta \geq m) \geq \frac{1}{2}$ (here and in (c), Pr is posterior probability given y), and assume $\mathrm{E}(|\theta| \mid y) < \infty$. For $a > m$ the pointwise inequality

$$|\theta - a| - |\theta - m| \;\geq\; (a - m)\big[\mathbf{1}\{\theta \leq m\} - \mathbf{1}\{\theta > m\}\big]$$

holds – with equality when $\theta \leq m$, and by the triangle inequality $|\theta - a| \geq |\theta - m| - (a - m)$ otherwise. Taking posterior expectations,

$$\mathrm{E}\big[|\theta - a| \mid y\big] - \mathrm{E}\big[|\theta - m| \mid y\big] \geq (a - m)\big[\Pr(\theta \leq m) - \Pr(\theta > m)\big] \geq 0.$$

The case $a < m$ is the mirror image, using $\Pr(\theta \geq m) \geq \frac{1}{2}$. Hence m minimizes the posterior expected loss.
(c) Put $q = k_0/(k_0 + k_1) \in [0, 1]$ and let m be any q-quantile, that is

$$\Pr(\theta < m) \;\leq\; q \;\leq\; \Pr(\theta \leq m).$$

Fix a and set $D(\theta) = L(\theta, a) - L(\theta, m)$.
(i) $a > m$. The three regions give

$$D(\theta) = \begin{cases} k_1(a - m), & \theta \leq m, \\ k_1(a - m) - (k_0 + k_1)(\theta - m), & m < \theta < a, \\ -k_0(a - m), & \theta \geq a, \end{cases}$$

(the middle line by writing $k_1(a - \theta) - k_0(\theta - m)$ as $k_1(a - m) - (k_0 + k_1)(\theta - m)$), and on $m < \theta < a$ that line is decreasing in θ, hence $\geq -k_0(a - m)$. So $D(\theta) \geq (a - m)\big[k_1\mathbf{1}\{\theta \leq m\} - k_0\mathbf{1}\{\theta > m\}\big]$, and

$$\begin{aligned} \mathrm{E}\big[D(\theta) \mid y\big] &\geq (a - m)\big[k_1 \Pr(\theta \leq m) - k_0 \Pr(\theta > m)\big] \\ &= (a - m)\big[(k_0 + k_1) \Pr(\theta \leq m) - k_0\big] \;\geq\; 0, \end{aligned}$$

the last step because $\Pr(\theta \le m) \ge q$.
(ii) $a < m$. Symmetrically $D(\theta) = k_0(m-a)$ for $\theta \ge m$, $D(\theta) = -k_1(m-a)$ for $\theta < a$, and $D(\theta) = k_0(\theta - a) - k_1(m-\theta)$ is increasing on $a \le \theta < m$, hence $\ge -k_1(m-a)$ there. Therefore

$$\begin{aligned}\mathrm{E}\big[D(\theta) \mid y\big] &\ge (m-a)\big[k_0 \Pr(\theta \ge m) - k_1 \Pr(\theta < m)\big] \\ &= (m-a)\big[k_0 - (k_0+k_1)\Pr(\theta < m)\big] \;\ge\; 0,\end{aligned}$$

now because $\Pr(\theta < m) \le q$.

Problem 4.7. Unbiasedness: prove that the Bayesian posterior mean, based on a proper prior distribution, cannot be an unbiased estimator except in degenerate problems (see Bickel and Blackwell, 1967, and Lehmann, 1983, p. 244).

Solution. If $\hat\theta(y) = \mathrm{E}(\theta \mid y)$ were unbiased then $\mathrm{E}\big[(\theta - \hat\theta(y))^2\big] = 0$, so $\theta = \hat\theta(y)$ with probability 1 and the data determine the parameter exactly.
Because the prior is proper, $p(\theta, y) = p(\theta)p(y \mid \theta)$ is a genuine joint probability distribution and every expectation below is taken over it; assume $\mathrm{E}(\theta^2) < \infty$ so that all the second moments exist. Suppose $\mathrm{E}\big(\hat\theta(y) \mid \theta\big) = \theta$ for all θ. Evaluate the cross-moment $\mathrm{E}\big[\theta\,\hat\theta(y)\big]$ by the tower property in the two possible orders.
Conditioning on θ (so that θ is fixed and unbiasedness applies):

$$\mathrm{E}\big[\theta\,\hat\theta(y)\big] = \mathrm{E}\Big[\theta\,\mathrm{E}\big(\hat\theta(y) \mid \theta\big)\Big] = \mathrm{E}\big[\theta \cdot \theta\big] = \mathrm{E}(\theta^2).$$

Conditioning on y (so that $\hat\theta(y)$ is fixed, being $\sigma(y)$-measurable, and the definition of the posterior mean applies):

$$\mathrm{E}\big[\theta\,\hat\theta(y)\big] = \mathrm{E}\Big[\hat\theta(y)\,\mathrm{E}\big(\theta \mid y\big)\Big] = \mathrm{E}\big[\hat\theta(y)^2\big].$$

Hence $\mathrm{E}(\theta^2) = \mathrm{E}(\hat\theta^2) = \mathrm{E}(\theta\hat\theta)$, and therefore

$$\mathrm{E}\big[(\theta - \hat\theta(y))^2\big] = \mathrm{E}(\theta^2) - 2\,\mathrm{E}(\theta\hat\theta) + \mathrm{E}(\hat\theta^2) = 0.$$

A nonnegative random variable with zero expectation is zero almost surely, so $\theta = \mathrm{E}(\theta \mid y)$ for almost every (θ, y); equivalently $\mathrm{E}\big[\mathrm{var}(\theta \mid y)\big] = 0$, so the posterior variance vanishes for almost every y. That is the degenerate case: the data reveal θ with certainty (for instance $p(y \mid \theta)$ a point mass at an invertible function of θ), and only then can a posterior mean be unbiased.

4.2 Exercises 4.8–4.14

Problem 4.8. Regression to the mean: work through the details of the example of mother's and daughter's heights on page 94, illustrating with a sketch of the joint distribution and relevant conditional distributions.
(The example on page 94, restated so this problem is self-contained. Let θ be the height of an adult daughter and y the height of her mother. Assume (y, θ) are jointly normal with known equal means of 160 centimeters, equal variances, and a known correlation of 0.5. Conditioning on the observed y, the posterior mean of θ is

$$E(\theta \mid y) = 160 + 0.5\,(y - 160),$$

which is the book's equation (4.5). This is not unbiased in the sense of repeated sampling of y given fixed θ: since $E(y \mid \theta) = 160 + 0.5(\theta - 160)$, the estimate (4.5) has expectation $160 + 0.25(\theta - 160)$ and is biased towards the grand mean. The estimate $\hat\theta = 160 + 2(y - 160)$ is unbiased under repeated sampling of y given θ, but makes no sense: a mother 10 centimeters taller than average is credited with a daughter 20 centimeters taller than average.)

Solution. The two conditionals are mirror images, each with slope $\frac{1}{2}$ in its own conditioning variable, and that symmetry is the whole content of regression to the mean. With $\mu = 160$ and σ^2 the common variance, the joint distribution is normal with correlation $\rho = \frac{1}{2}$, and the bivariate normal conditioning formula of Appendix A gives

$$\theta \mid y \sim \mathrm{N}\big(\mu + \tfrac{1}{2}(y-\mu),\ \tfrac{3}{4}\sigma^2\big)\,,$$
$$y \mid \theta \sim \mathrm{N}\big(\mu + \tfrac{1}{2}(\theta-\mu),\ \tfrac{3}{4}\sigma^2\big)\,.$$

The first line is (4.5). Nothing shrinks in the population: $\mathrm{Var}(\theta) = \mathrm{Var}(y) = \sigma^2$, and the joint density is symmetric about the 45° line.

Sampling bias of the posterior mean. Taking expectations in (4.5) over $p(y \mid \theta)$,

$$E\big(E(\theta \mid y) \,\big|\, \theta\big) = \mu + \tfrac{1}{2}\big(E(y \mid \theta) - \mu\big) = \mu + \tfrac{1}{4}(\theta - \mu),$$

so that

$$\mathrm{bias}(\theta) = \mu + \tfrac{1}{4}(\theta-\mu) - \theta = -\tfrac{3}{4}(\theta-\mu),$$

shrinkage towards the grand mean by three quarters of the distance.

The unbiased alternative. Since $E(y \mid \theta) = \mu + \frac{1}{2}(\theta-\mu)$, inverting the map gives $\hat\theta = \mu + 2(y-\mu)$, and indeed

$$E(\hat\theta \mid \theta) = \mu + 2\big(\tfrac{1}{2}(\theta-\mu)\big) = \theta, \qquad \mathrm{Var}(\hat\theta \mid \theta) = 4 \cdot \tfrac{3}{4}\sigma^2 = 3\sigma^2.$$

Mean squared error, conditional on θ. For the posterior mean, $\mathrm{Var}(E(\theta \mid y) \mid \theta) = \frac{1}{4} \cdot \frac{3}{4}\sigma^2$, so

$$\mathrm{MSE}_{(4.5)}(\theta) = \tfrac{3}{16}\sigma^2 + \tfrac{9}{16}(\theta-\mu)^2,$$
$$\mathrm{MSE}_{\hat\theta}(\theta) = 3\sigma^2.$$

The unbiased estimate has the smaller mean squared error only when

$$\tfrac{9}{16}(\theta-\mu)^2 > 3\sigma^2 - \tfrac{3}{16}\sigma^2 = \tfrac{45}{16}\sigma^2, \quad \text{i.e.} \quad |\theta-\mu| > \sqrt{5}\,\sigma.$$

So the unbiased estimate wins only for $|\theta-\mu| > \sqrt{5}\,\sigma$, that is on about 2.5% of the population; averaged over $p(\theta)$ the comparison is $\frac{3}{4}\sigma^2$ against $3\sigma^2$, a factor of exactly 4. Concretely, with $\sigma = 7.5$ cm and a mother of height $y = 175$, $\theta \mid y \sim \mathrm{N}(167.5, 6.50^2)$ gives the 95% interval $[154.8, 180.2]$, while $\hat\theta = 190$ cm lies outside it and above the 99.99th percentile of $p(\theta)$.

Sketch. The joint density is concentric ellipses centred at $(160, 160)$ with major axis along the 45° line. Three lines pass through the centre: $E(\theta \mid y) = 160 + 0.5(y-160)$, of slope $\frac{1}{2}$, which bisects every vertical chord of every ellipse; the mirror line $E(y \mid \theta)$, which on these axes reads $\theta = 160 + 2(y-160)$, of slope 2 — this is $\hat\theta$, and it bisects every horizontal chord; and the 45° line between them. The vertical slices at $y = 150, 160, 175$ are, after normalising, three normal curves of common standard deviation 0.866σ centred at $155, 160, 167.5$: each centre is pulled half way back to 160.

Problem 4.9. Point estimation: suppose a measurement y is recorded with a $\mathrm{N}(\theta, \sigma^2)$ sampling distribution, with σ known exactly and θ known to lie in the interval $[0, 1]$. Consider two point estimates of θ: (1) the maximum likelihood estimate, restricted to the range $[0, 1]$, and (2) the posterior mean based on the assumption of a uniform prior distribution on θ. Show that if σ is large enough, estimate (1) has a higher mean squared error than (2) for any value of θ in $[0, 1]$. (The unrestricted maximum likelihood estimate has even higher mean squared error.)

Solution. For large σ estimate (1) degenerates into a coin flip between the two endpoints, and a coin flip between 0 and 1 costs exactly $\frac{1}{4}$ more than the constant $\frac{1}{2}$ at every θ; estimate (2) degenerates into that constant. The two estimates are

$$\hat\theta_1(y) = \mathrm{median}(0, y, 1), \qquad \hat\theta_2(y) = \frac{\int_0^1 u\, e^{-(y-u)^2/2\sigma^2}\, du}{\int_0^1 e^{-(y-u)^2/2\sigma^2}\, du}.$$

Lower bound for estimate (1). Discarding the event $\{0 \le y \le 1\}$, on which the loss is nonnegative, and

using $\Pr(y<0)=\Phi(-\theta/\sigma)\ge\Phi(-1/\sigma)$ and $\Pr(y>1)=\Phi((\theta-1)/\sigma)\ge\Phi(-1/\sigma)$ for $\theta\in[0,1]$, write $p_\sigma=\Phi(-1/\sigma)$; then

$$\begin{aligned}\mathrm{MSE}_1(\theta) &\ge \theta^2\Pr(y<0)+(1-\theta)^2\Pr(y>1)\\ &\ge p_\sigma\big[\theta^2+(1-\theta)^2\big] = p_\sigma\Big[2\big(\theta-\tfrac12\big)^2+\tfrac12\Big].\end{aligned}$$

Upper bound for estimate (2). Cancel the factor $e^{-y^2/2\sigma^2}$ from numerator and denominator: with $w(u)=\exp\big(yu/\sigma^2-u^2/2\sigma^2\big)$,

$$\hat\theta_2(y)=\frac{\int_0^1 u\,w(u)\,du}{\int_0^1 w(u)\,du},\qquad e^{-\lambda}\le w(u)\le e^{\lambda}\ \text{ on }[0,1],$$

where $\lambda=(|y|+\tfrac12)/\sigma^2$. Hence $\tfrac12 e^{-2\lambda}\le\hat\theta_2\le\tfrac12 e^{2\lambda}$, so, together with the trivial bound $|\hat\theta_2-\tfrac12|\le\tfrac12$,

$$\big|\hat\theta_2-\tfrac12\big| \le \tfrac12\min\big(1,\ e^{2\lambda}-1\big).$$

On $\{|y-\theta|\le\sigma^{3/2}\}$ we have $\lambda\le\lambda_\sigma:=\sigma^{-1/2}+\frac32\sigma^{-2}$, and the complement has probability $2\Phi(-\sigma^{1/2})$ whatever θ is, so

$$E\big|\hat\theta_2-\tfrac12\big| \le \tfrac12\big(e^{2\lambda_\sigma}-1\big)+\Phi(-\sigma^{1/2}) =: \tfrac23\,\varepsilon_\sigma,$$

uniformly in $\theta\in[0,1]$, and $\varepsilon_\sigma\to0$ as $\sigma\to\infty$. Expanding about $\frac12$ and using $|2(\frac12-\theta)|\le1$ and $(\hat\theta_2-\frac12)^2\le\frac12|\hat\theta_2-\frac12|$,

$$\mathrm{MSE}_2(\theta)=\big(\theta-\tfrac12\big)^2+2\big(\tfrac12-\theta\big)E\big[\hat\theta_2-\tfrac12\big]+E\big[(\hat\theta_2-\tfrac12)^2\big] \le \big(\theta-\tfrac12\big)^2+\varepsilon_\sigma.$$

Comparison. Subtracting, and using $0\le(\theta-\frac12)^2\le\frac14$ with $2p_\sigma-1\le0$,

$$\begin{aligned}\mathrm{MSE}_1(\theta)-\mathrm{MSE}_2(\theta) &\ge (2p_\sigma-1)\big(\theta-\tfrac12\big)^2+\tfrac{p_\sigma}{2}-\varepsilon_\sigma\\ &\ge \tfrac{2p_\sigma-1}{4}+\tfrac{p_\sigma}{2}-\varepsilon_\sigma = p_\sigma-\tfrac14-\varepsilon_\sigma.\end{aligned}$$

Since $p_\sigma\to\frac12$ and $\varepsilon_\sigma\to0$, the right side tends to $\frac14$ and in particular is positive for all σ large enough — with the bounds as written, already for $\sigma\ge100$, where $p_\sigma-\frac14=0.246$ and $\varepsilon_\sigma\le0.167$. The bound is uniform in θ, which is what was asked.
The limits are exact: $\hat\theta_1\to$ a fair coin flip on $\{0,1\}$ and $\hat\theta_2\to\frac12$, so

$$\mathrm{MSE}_1(\theta)\to\big(\theta-\tfrac12\big)^2+\tfrac14,\qquad \mathrm{MSE}_2(\theta)\to\big(\theta-\tfrac12\big)^2,$$

a gap of exactly $\frac14$ at every θ. The constant 100 is an artifact of the bounding: numerical quadrature of the two mean squared errors (the minimum over θ is attained at an endpoint of $[0,1]$) puts the crossover at $\sigma=0.753$,

σ	0.5	0.75	1	2	5
$\min_\theta(\mathrm{MSE}_1-\mathrm{MSE}_2)$	−0.040	−0.0006	0.039	0.129	0.198

Problem 4.10. Non-Bayesian inference: replicate the analysis of the bioassay example in Section 3.7 using non-Bayesian inference. This problem does not have a unique answer, so be clear on what methods you are using.
(a) Construct an "estimator" of (α,β); that is, a function whose input is a dataset, (x,n,y), and whose output is a point estimate $(\hat\alpha,\hat\beta)$. Compute the value of the estimate for the data given in Table 3.1.
(b) The bias and variance of this estimate are functions of the true values of the parameters (α,β) and also of the sampling distribution of the data, given α,β. Assuming the binomial model, estimate the bias and variance of your estimator.

(c) Create approximate 95% confidence intervals for α, β, and the LD50 based on asymptotic theory and the estimated bias and variance.
(d) Does the inaccuracy of the normal approximation for the posterior distribution (compare Figures 3.3 and 4.1) cast doubt on the coverage properties of your confidence intervals in (c)? If so, why?
(e) Create approximate 95% confidence intervals for α, β, and the LD50 using the jackknife or bootstrap (see Efron and Tibshirani, 1993).
(f) Compare your 95% intervals for the LD50 in (c) and (e) to the posterior distribution displayed in Figure 3.4 and the posterior distribution based on the normal approximation, displayed in 4.2b. Comment on the similarities and differences among the four intervals. Which do you prefer as an inferential summary about the LD50? Why?
Table 3.1, bioassay data from Racine et al. (1986):

Dose, x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

The model of Section 3.7 is $y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$ with $\text{logit}(\theta_i) = \alpha + \beta x_i$ (3.14), and LD50 is the dose at which $\theta = \frac{1}{2}$, namely $-\alpha/\beta$. Figure 3.3 shows the exact posterior under $p(\alpha, \beta) \propto 1$ (contours and 1000 draws), noticeably skewed towards large α, β; Figure 3.4 is the histogram of the exact posterior draws of LD50, conditional on $\beta > 0$; Figure 4.1 is the normal approximation at the mode, an ellipse with no such skew; Figure 4.2b is the central 95% of the LD50 draws from that normal approximation, conditional on $\beta > 0$.

Solution. (a) Take the estimator to be the maximum likelihood estimate, $(\hat\alpha, \hat\beta) = \arg\max \sum_i [\, y_i(\alpha + \beta x_i) - n_i \log(1 + e^{\alpha+\beta x_i}) \,]$, computed by Newton–Raphson. Because the prior in Section 3.7 is uniform, this is exactly the posterior mode quoted on page 86. For Table 3.1,

$$(\hat\alpha, \hat\beta) = (0.847,\ 7.749), \qquad \widehat{\text{LD50}} = -\hat\alpha/\hat\beta = -0.109,$$

with fitted death probabilities (0.003, 0.186, 0.613, 0.998) at the four doses.
The estimator needs a caveat that will dominate everything below: the maximum is attained at a finite point only if the data are not (quasi-)separated, i.e. only if there is no index j with $y_i = 0$ for all $i < j$ and $y_i = n_i$ for all $i > j$. Table 3.1 is not separated; many datasets from this design are.
(b) The observed information at $(\hat\alpha, \hat\beta)$, with $w_i = n_i \hat\theta_i (1 - \hat\theta_i)$, is

$$J = \begin{pmatrix} \sum w_i & \sum w_i x_i \\ \sum w_i x_i & \sum w_i x_i^2 \end{pmatrix} = \begin{pmatrix} 1.965 & -0.293 \\ -0.293 & 0.0859 \end{pmatrix},$$
$$J^{-1} = \begin{pmatrix} 1.039 & 3.546 \\ 3.546 & 23.74 \end{pmatrix},$$

so the asymptotic standard errors are $\text{se}(\hat\alpha) = 1.019$, $\text{se}(\hat\beta) = 4.873$, correlation 0.714, and the asymptotic bias is $O(n^{-1})$, nominally zero.
That is the textbook answer; it is badly wrong here, and the parametric bootstrap says why. Plugging $\hat\theta = (0.003, 0.186, 0.613, 0.998)$ into the binomial model and enumerating all $6^4 = 1296$ possible datasets exactly (no Monte Carlo error is needed at this size),

$$\Pr\big(\text{the MLE is infinite}\big) = 0.404.$$

Under the fitted model, two replications in five of this experiment produce a separated dataset and no finite estimate at all, so the bias and variance of $(\hat\alpha, \hat\beta)$ are undefined. Conditioning on the 59.6% of datasets with a finite MLE,

$$E(\hat\alpha) = 0.930, \qquad \text{bias} = +0.083, \qquad \text{sd} = 0.777,$$
$$E(\hat\beta) = 7.217, \qquad \text{bias} = -0.531, \qquad \text{sd} = 2.240.$$

The conditional standard deviations sit well below the asymptotic standard errors (0.78 against 1.02, and 2.24 against 4.87), because conditioning on a finite estimate removes exactly the replications that would have supplied the large values. For LD50 the delta method gives, with $g = (-1/\hat\beta,\ \hat\alpha/\hat\beta^2) = (-0.129,\ 0.0141)$,

$$\mathrm{se}(\widehat{\mathrm{LD50}}) = \sqrt{g^T J^{-1} g} = 0.0955,$$

in agreement with the 0.096 obtained by numerical differentiation of the exact marginal posterior of LD50. Its exact conditional sampling bias is $E(\widehat{\mathrm{LD50}}) - \widehat{\mathrm{LD50}} = -0.1214 + 0.1093 = -0.012$. (Its conditional *variance* does not exist for practical purposes — a few very improbable datasets give $\hat\beta$ near 0 and make the second moment arbitrarily large — but its quantiles, used in (e), are perfectly well behaved.)

(c) Using $\hat\theta - \widehat{\mathrm{bias}} \pm 1.96\,\mathrm{se}$ with the asymptotic standard errors and the bias estimates from (b):

$$\begin{aligned} \alpha: &\quad 0.764 \pm 1.96(1.019) = [-1.23,\ 2.76], \\ \beta: &\quad 8.280 \pm 1.96(4.873) = [-1.27,\ 17.83], \\ \mathrm{LD50}: &\quad -0.097 \pm 1.96(0.0955) = [-0.284,\ 0.090]. \end{aligned}$$

Without the bias correction these are $[-1.15, 2.84]$, $[-1.80, 17.30]$ and $[-0.296, 0.078]$; the correction is immaterial next to the standard errors.

(d) Yes, decisively, and for two related reasons. First, the interval for β contains 0, which is the same artifact visible in Figure 4.1: the normal approximation assigns probability $\Phi(-7.749/4.873) = 0.056$ to $\beta \le 0$, whereas the exact posterior of Figure 3.3 assigns 4×10^{-6} (the book reports $\Pr(\beta > 0) > 0.999$ on page 87). Since $\mathrm{LD50} = -\alpha/\beta$ has a pole at $\beta = 0$, no interval built from a normal approximation that straddles $\beta = 0$ can be trusted, and the delta-method variance — a first-order expansion of a function with a nearby singularity — is not a reliable scale. Second, and worse, the 40% separation probability found in (b) means the nominal 95% procedure fails outright on two replications in five. Enumerating the exact sampling distribution and counting, the actual coverage of the intervals in (c) is

$$\Pr(\text{covers } \alpha) = 0.587, \quad \Pr(\text{covers } \beta) = 0.573, \quad \Pr(\text{covers LD50}) = 0.593,$$

counting a separated replication (where no finite interval is produced) as a failure. Conditionally on a finite MLE the coverages are 0.986, 0.962 and 0.996 — the procedure over-covers when it works at all, and the 95% label describes neither situation. With $n = 20$ animals and four design points, the asymptotics of Section 4.2 simply have not engaged.

(e) Parametric bootstrap: resample $y_i^* \sim \mathrm{Bin}(5, \hat\theta_i)$ and refit. Again this is done by exact enumeration of the 1296 outcomes, so the reported percentiles are the exact bootstrap percentiles. Discarding the separated replicates,

$$\alpha \in [-0.48,\ 1.95], \quad \beta \in [3.57,\ 11.16], \quad \mathrm{LD50} \in [-0.312,\ 0.086].$$

The nonparametric bootstrap, resampling the 20 animals with replacement (20,000 replications, 43.8% separated and discarded, Monte Carlo standard error about 0.003 on the LD50 endpoints), gives the very similar $\alpha \in [-0.55, 2.31]$, $\beta \in [4.00, 12.04]$, $\mathrm{LD50} \in [-0.287, 0.104]$. All three intervals are conditional on the 56%–60% event that the replicate admits a finite estimate. (For a *quasi*-separated replicate, one whose middle group sits at the separating dose, $-\hat\alpha/\hat\beta$ still has a finite limit, namely that dose; for a completely separated one it does not.)

Jackknife, deleting one animal at a time from the 20: this breaks down for α and β, since deleting the single death at $x = -0.30$ yields $y = (0, 0, 3, 5)$ with $n = (5, 4, 5, 5)$, which is separated at $x = -0.05$; that pseudo-value is infinite, so the jackknife standard errors for α and β do not exist. For LD50 the pseudo-values stay finite (the separated one contributes -0.05) and the jackknife gives $\mathrm{se} = 0.113$ with bias estimate $+0.027$, hence

$$\mathrm{LD50} \in -0.109 \pm 1.96(0.113) = [-0.331,\ 0.113].$$

(f) Collecting the four intervals for LD50:

method	95% interval	width
(c) asymptotic, delta method	$[-0.284,\ 0.090]$	0.374
(e) parametric bootstrap percentile	$[-0.312,\ 0.086]$	0.398
(e) jackknife	$[-0.331,\ 0.113]$	0.444
exact posterior, Figure 3.4	$[-0.276,\ 0.103]$	0.379
normal-approximation posterior, Figure 4.2b	$[-0.395,\ 0.445]$	0.840

The first four agree remarkably well: all are centred near -0.11 and all have width close to 0.4. The odd one out is the interval derived from the normal approximation to $p(\alpha, \beta|y)$, which is more than twice as wide, because the 5.6% of its draws with β near or below zero send $-\alpha/\beta$ off towards $\pm\infty$; the book records a range of $[-12.4,\ 5.4]$ for the 950 draws behind Figure 4.2a, and Figure 4.2b only hides the tails by truncation.

I prefer the exact posterior interval of Figure 3.4. It requires no asymptotics, it is exactly the quantity asked about — a statement of where LD50 lies given these 20 animals — and it is the only one of the five that is well defined for *every* dataset this design can produce. The delta-method and bootstrap intervals happen to land in the right place for this dataset, but (d) shows their calibration is an accident of these particular data: on two replications in five they do not exist, and their unconditional coverage is 59%, not 95%. The jackknife is the weakest of the four, being both the widest and the most fragile.

Problem 4.11. Bayesian interpretation of non-Bayesian estimates: consider the following estimation procedure, which is based on classical hypothesis testing. A matched pairs experiment is done, and the differences $y_1, \ldots, y_n$ are recorded and modeled as independent draws from $\mathrm{N}(\theta, \sigma^2)$. For simplicity, assume σ^2 is known. The parameter θ is estimated as the average observed difference if it is "statistically significant" and zero otherwise:

$$\hat{\theta} = \begin{cases} \bar{y} & \text{if } \bar{y} \geq 1.96\,\sigma/\sqrt{n}, \\ 0 & \text{otherwise.} \end{cases}$$

Can this be interpreted, in some sense, as an approximate summary (for example, a posterior mean or mode) of a Bayesian inference under some prior distribution on θ?

Solution. Yes as a posterior *mode*, under a spike-and-slab prior: a lump of prior mass at $\theta = 0$ mixed with a flat density on $\theta > 0$. Never as a posterior *mean*, for any prior.

Write $\tau = \sigma/\sqrt{n}$, so that $\bar{y}$ is sufficient with $\bar{y} \mid \theta \sim \mathrm{N}(\theta, \tau^2)$, and take

$$p(\theta) = (1-\pi)\,\mathrm{N}(\theta \mid 0, \epsilon^2) + \pi\,\mathrm{U}(\theta \mid 0, M),$$

with ϵ tiny and M large. The unnormalised posterior is $q(\theta) = p(\theta)\,\exp\big(-(\bar{y}-\theta)^2/2\tau^2\big)$, and it has at most two local maxima, one contributed by each prior component.

(i) $\bar{y} < 0$. On $(0, M)$ the likelihood is strictly decreasing, so the slab branch has no interior maximum and the mode is the spike, $\hat{\theta} = 0$.

(ii) $\bar{y} > 0$. The slab branch peaks at $\theta = \bar{y}$ (for $\bar{y} < M$), where $q(\bar{y}) = \pi/M$; as $\epsilon \to 0$ the spike branch peaks at $\theta \approx 0$, where $q(0) \approx \big[(1-\pi)/(\sqrt{2\pi}\epsilon)\big]e^{-\bar{y}^2/2\tau^2}$. Comparing,

$$\frac{q(\bar{y})}{q(0)} = \frac{\pi}{M} \cdot \frac{\sqrt{2\pi}\,\epsilon}{1-\pi}\, e^{\bar{y}^2/2\tau^2} \;>\; 1 \iff \frac{\bar{y}^2}{\tau^2} \;>\; 2\log K, \quad K := \frac{(1-\pi)M}{\sqrt{2\pi}\,\pi\,\epsilon}.$$

Choosing the prior constants so that $2\log K = 1.96^2 = 3.8416$, i.e. $K = e^{1.9208} = 6.83$, makes the posterior mode equal to $\bar{y}$ exactly when $\bar{y} \geq 1.96\tau = 1.96\sigma/\sqrt{n}$, and 0 otherwise. This is the stated estimator.

The threshold $\tau\sqrt{2\log K}$ carries the factor $\sigma/\sqrt{n}$ by itself, so a *single* fixed prior reproduces the procedure at every sample size; and the one-sidedness of the printed rule is exactly the one-sidedness of the slab, since a uniform component on $(-M, M)$ gives instead $\hat{\theta} = \bar{y}\,\mathbf{1}\{|\bar{y}| \geq 1.96\tau\}$.

It is not a posterior mean under any prior G, however. Bayes' rule (1.1) with the normal kernel gives

$$E(\theta \mid \bar{y}) = \frac{\int \theta\, e^{\theta\bar{y}/\tau^2 - \theta^2/2\tau^2}\, dG(\theta)}{\int e^{\theta\bar{y}/\tau^2 - \theta^2/2\tau^2}\, dG(\theta)},$$

a ratio of two-sided Laplace transforms of finite measures, hence real-analytic — in particular continuous — in $\bar{y}$ on the interior of its domain of convergence. The estimator jumps from 0 to 1.96τ at $\bar{y} = 1.96\tau$, so no prior makes it a posterior mean.

Problem 4.12. Bayesian interpretation of non-Bayesian estimates: repeat the above problem but with σ replaced by s, the sample standard deviation of $y_1, \dots, y_n$. That is, $y_1, \dots, y_n$ are modeled as independent $\mathrm{N}(\theta, \sigma^2)$ draws with σ^2 now unknown, and

$$\hat{\theta} = \begin{cases} \bar{y} & \text{if } \bar{y} \geq 1.96\, s/\sqrt{n}, \\ 0 & \text{otherwise,} \end{cases} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Can this be interpreted, in some sense, as an approximate summary (for example, a posterior mean or mode) of a Bayesian inference under some prior distribution on θ?

Solution. Yes, as a posterior mode under the same spike-and-slab prior on θ as in 4.11, now multiplied by the standard noninformative prior $p(\sigma^2) \propto \sigma^{-2}$; the implied cutoff is not exactly 1.96 for finite n, but converges to it.

Keep the prior

$$p(\theta, \sigma^2) \;\propto\; \Big[(1-\pi)\,\mathrm{N}(\theta \mid 0, \epsilon^2) + \pi\,\mathrm{U}(\theta \mid 0, M)\Big]\,\sigma^{-2},$$

ϵ tiny, M large. Marginalising σ^2 is exactly the gamma-integral calculation of Section 3.2 that produces the marginal posterior of μ under $p(\mu, \log\sigma) \propto 1$, and it is unaffected by the factor $p(\theta)$, which does not involve σ^2:

$$p(\theta \mid y) \;\propto\; p(\theta)\left[1 + \frac{n(\bar{y}-\theta)^2}{(n-1)s^2}\right]^{-n/2},$$

i.e. the prior times a $t_{n-1}(\bar{y}, s^2/n)$ kernel. As in 4.11 there are two candidate modes, and for $\bar{y} < 0$ the slab branch is monotone decreasing on $(0, M)$ so the mode is 0. For $\bar{y} > 0$, writing $t = \sqrt{n}\,\bar{y}/s$ and $K = (1-\pi)M/(\sqrt{2\pi}\,\pi\,\epsilon)$ as before,

$$\frac{p(\bar{y} \mid y)}{p(0 \mid y)} = \frac{1}{K}\left[1 + \frac{t^2}{n-1}\right]^{n/2} > 1 \iff t^2 > c_n^2 := (n-1)\big(K^{2/n} - 1\big).$$

So the estimator is again a hard threshold on the usual t statistic, exactly the form of the stated rule. Calibrating with the same $K = e^{1.9208} = 6.83$ that worked in 4.11,

$$c_n^2 = (n-1)\big(e^{3.8416/n} - 1\big) \longrightarrow 3.8416 = 1.96^2 \quad (n \to \infty),$$

since $n(e^{a/n} - 1) \to a$. The finite-n cutoffs, alongside the classical $t_{n-1,0.975}$ they are imitating:

n	5	10	20	50	100	1000
c_n	2.150	2.053	2.006	1.978	1.969	1.961
$t_{n-1,0.975}$	2.776	2.262	2.093	2.010	1.984	1.962

The implied cutoff exceeds 1.96 and decreases to it, tracking $t_{n-1,0.975}$ from below. So the single n-free prior with $K = 6.83$ reproduces the stated rule only in the limit, but to within the same order of accuracy with which 1.96 itself approximates $t_{n-1,0.975}$; an exact match at a given n costs an n-dependent prior, $K = (1 + 3.8416/(n-1))^{n/2}$.

It is still not a posterior mean. For fixed s the t kernel is bounded by 1 and continuous in $\bar{y}$, so whenever $\int |\theta|\, dG(\theta) < \infty$ (which is what makes the posterior mean exist at all) dominated convergence makes $E(\theta \mid \bar{y}, s)$ continuous in $\bar{y}$, while $\hat{\theta}$ jumps from 0 to $1.96\, s/\sqrt{n}$.

Problem 4.13. Objections to Bayesian inference: discuss the criticism, "Bayesianism assumes: (a) Either a weak or uniform prior [distribution], in which case why bother?, (b) Or a strong prior [distribution], in which case why collect new data?, (c) Or more realistically, something in between,

in which case Bayesianism always seems to duck the issue" (Ehrenberg, 1986). Feel free to use any of the examples covered so far to illustrate your points.

Solution. The trilemma is false in each of its three arms, because it assumes that the prior distribution is the only thing Bayesian inference contributes and that the likelihood is the only thing it shares with everyone else. The Bayesian contribution is the posterior *distribution*, and that is worth having no matter how weak the prior is.

(a) *Weak or uniform prior, so why bother?* Because the answer is a distribution rather than an estimate and a standard error, and in small or awkward problems those are not the same thing. The bioassay analysis of Section 3.7 uses $p(\alpha, \beta) \propto 1$ — the prior contributes literally nothing — and yet the Bayesian and non-Bayesian analyses of the same likelihood differ sharply. As computed in 4.10, under the fitted model the maximum likelihood estimate is infinite with probability 0.40, the nominal 95% asymptotic confidence intervals have actual coverage 59%, and LD50 $= -\alpha/\beta$ has a pole at $\beta = 0$ that no delta-method standard error can represent. The posterior histogram of Figure 3.4 has none of these problems and required no asymptotics. Similarly in Section 2.4 the posterior for the placenta previa sex ratio is reported as an interval on the probability scale directly, with no transformation or continuity correction. The prior did no work in either case; Bayes did.

Two further points about arm (a). A uniform prior is a substantive choice, not an abstention: it is not invariant to reparameterisation, which is exactly the motivation for Jeffreys' prior in Section 2.8, and in Chapter 5 the flat hyperprior $p(\log \tau) \propto 1$ yields an improper posterior. And "weak" is relative to the likelihood: the same prior is noninformative for $n = 1000$ and influential for $n = 5$.

(b) *Strong prior, so why collect data?* Because a strong prior is not an infinitely strong one, and precisions add. For the normal model, (2.12) gives

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2},$$

so the data contribution n/σ^2 grows without bound while the prior contribution $1/\tau_0^2$ stays fixed; this is the statement on page 88 that the likelihood dominates the prior. The placenta previa example is the illustration: with 980 births, every informative prior in Table 2.1 — prior sample sizes $\alpha + \beta$ from 2 up to 200 — yields the same posterior median to within 0.001 and essentially the same 95% interval. If a prior really is so strong that n new observations cannot move it, that is a discovery worth making, and it is made by writing the prior down and doing the arithmetic. Furthermore, data do something a prior cannot: they check the model. The posterior predictive checks of Chapter 6 can reject the sampling model outright, whatever the prior on its parameters.

(c) *Something in between, so Bayesianism ducks the issue.* This inverts the situation. The "issue" is how much weight partial prior information should get, and the Bayesian answer is a number, stated in advance, with its influence on the conclusion measurable by the sensitivity analysis described in Section 6.1 and carried out in Section 6.5. The alternative is not an analysis free of prior information; it is one in which the same information enters through the choice of model, the choice of test statistic, which covariates are retained, which outliers are excluded, and when sampling stopped — none of which appear in the reported standard error. Section 4.5 makes the sharper version of this point: insisting on unbiasedness does not avoid prior information, it merely discards it, as in the mother-and-daughter example worked in 4.8, where the unbiased estimate assigns a daughter a height above the 99.99th percentile of its own population distribution.

The arm (c) objection also ignores the case that occupies most of this book. In a hierarchical model (Chapter 5) the "in between" prior is not a subjective compromise at all: its parameters are estimated from the same data, by pooling information across units. The answer to "where did the prior come from?" is then "from the other 70 experiments in the table."

What survives of Ehrenberg's criticism is a reporting standard, not an objection: state the prior, state the likelihood, and show how much the conclusion would change under reasonable alternatives to each.

Problem 4.14. Objectivity and subjectivity: discuss the statement, "People tend to believe results that support their preconceptions and disbelieve results that surprise them. Bayesian methods encourage this undisciplined mode of thinking."

Solution. Exactly backwards: the premise is right and the conclusion reverses it. Bayesian methods are the only ones that require the preconception to be written down as a probability distribution, published with the analysis, and combined with the data by a rule fixed in advance. That is the definition of disciplining a preconception, not of encouraging one.

The prior is auditable; the alternatives are not. An analyst who dismisses a surprising result in a Bayesian report must exhibit the prior under which it is unsurprising, and must have committed to that prior before seeing the data for the report to be honest. A non-Bayesian analysis of the same data offers a much larger and wholly unreported set of levers for the same purpose: which model to fit, which test statistic to use, which covariates to keep, which observations are "outliers," when to stop sampling, how many comparisons were actually made. None of these appear in the standard error, and all of them respond to preconceptions. The prior is the one place where the subjectivity is labelled.

The influence of a preconception is bounded, and the bound is computable. For the normal model with prior $\mathrm{N}(\mu_0, \tau_0^2)$, (2.12) gives the posterior mean as the precision-weighted average

$$E(\theta \mid y) = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}},$$

whose weight on the prior is $O(1/n)$. To sustain a preconception against accumulating data an analyst would need τ_0^{-2} to grow with n, which a stated prior cannot do; and since the asymptotic results of Section 4.2 make the posterior converge to θ_0 for *any* prior that is positive in a neighbourhood of θ_0, the only way to be immovable is to assign a surprising conclusion prior probability exactly zero — which is visible on the page.

Bayesian analysis comes with its own anti-self-deception machinery. The posterior predictive checks of Chapter 6 confront the fitted model with the data and can discredit it regardless of what the prior says, and the sensitivity analysis of Sections 6.1 and 6.5 reports how much of the conclusion is prior and how much is likelihood. The classical apparatus has nothing corresponding to the second of these.

Finally, the "undisciplined" behaviour is sometimes the correct behaviour. Discounting a surprising estimate is right when the surprise is more probably noise than signal. That is the content of the regression example of 4.8, where the estimate that refuses to shrink towards the population mean predicts a 190 cm daughter, and of the hierarchical shrinkage of Chapter 5, where an extreme rate from a small sample is pulled towards the population of similar experiments and thereby made more accurate. A method that takes every surprising point estimate at face value is not more objective; it is just less calibrated.

The legitimate residue of the complaint is that a strong prior *can* be used to bury an inconvenient likelihood, and that reporting only a posterior mean conceals this. The remedy is the reporting standard — prior, likelihood, and sensitivity, shown separately — not the abandonment of the only framework in which the preconception has to be declared.

4.3 Exercises 4.15–4.15

Problem 4.15. Coverage of posterior intervals:

(a) Consider a model with scalar parameter ω. Prove that, if you draw ω from the prior, draw $y \mid \omega$ from the data model, then perform Bayesian inference for ω given y, that there is a 50 percent probability that your 50 percent interval for ω contains the true value.

(b) Suppose $\omega \sim \mathrm{N}(0, 2^2)$ and $y \mid \omega \sim \mathrm{N}(\omega, 1)$. Suppose the true value of ω is 1. What is the coverage of the posterior 50 percent interval for ω? (You have to work this one out; it's not 50 percent or any other number you could just guess.)

(c) Suppose $\omega \sim \mathrm{N}(0, 2^2)$ and $y \mid \omega \sim \mathrm{N}(\omega, 1)$. Suppose the true value of ω is ω_0. Make a plot

showing the coverage of the posterior 50 percent interval for ω, as a function of ω_0.

Solution. (a) Average the conditional coverage over the marginal of y. Let $S(y)$ be any set built from y with $\Pr(\omega \in S(y) \mid y) = 1/2$ for every y, the defining property of a 50 percent posterior interval. Since the probability is taken under the joint density $p(\omega, y) = p(\omega)p(y \mid \omega)$, the tower property over the sigma-algebra generated by y gives

$$\begin{aligned}\Pr\big(\omega \in S(y)\big) &= \mathrm{E}\Big[\mathrm{E}\big[1_{\{\omega \in S(y)\}} \,\big|\, y\big]\Big] \\ &= \mathrm{E}\big[\Pr(\omega \in S(y) \mid y)\big] = \mathrm{E}\big[\tfrac{1}{2}\big] = \tfrac{1}{2}.\end{aligned}$$

(b) 0.535. By (2.10), with prior $\mathrm{N}(\mu_0, \tau_0^2) = \mathrm{N}(0, 4)$ and a single observation of variance $\sigma^2 = 1$, the posterior precision is $\tau_0^{-2} + \sigma^{-2} = \frac{1}{4} + 1 = \frac{5}{4}$ and the posterior mean is $(\frac{1}{4} \cdot 0 + 1 \cdot y)/\frac{5}{4}$, so

$$\omega \mid y \sim \mathrm{N}\big(\tfrac{4}{5}y,\ \tfrac{4}{5}\big),$$

and the central 50 percent posterior interval is $\frac{4}{5}y \pm c$ with

$$c = z_{0.75}\sqrt{4/5} = 0.67449 \times 0.89443 = 0.60328.$$

Now fix the true value ω_0 and let $y \sim \mathrm{N}(\omega_0, 1)$. The miss $\omega_0 - \frac{4}{5}y$ is normal with mean $\omega_0 - \frac{4}{5}\omega_0 = \omega_0/5$ and standard deviation $4/5$, so the coverage is

$$\begin{aligned}C(\omega_0) &= \Pr\big(\big|\omega_0 - \tfrac{4}{5}y\big| \le c\big) \\ &= \Phi\left(\frac{c - \omega_0/5}{4/5}\right) - \Phi\left(\frac{-c - \omega_0/5}{4/5}\right).\end{aligned}$$

At $\omega_0 = 1$ the two arguments are $(0.60328 - 0.2)/0.8 = 0.50410$ and -1.00410, giving

$$C(1) = 0.69291 - 0.15766 = 0.53524.$$

(c) The displayed $C(\omega_0)$ is the curve to plot: even in ω_0, peaking at the prior mean with $C(0) = 2\Phi(0.75410) - 1 = 0.549$ and decaying to 0 in both tails. Values:

ω_0	0	0.5	1	2	3	4	5	6	8	10
$C(\omega_0)$	0.549	0.546	0.535	0.495	0.435	0.363	0.288	0.216	0.104	0.040

Coverage exceeds the nominal 1/2 for $|\omega_0| < 1.908$, where shrinkage toward 0 pulls the interval toward the truth, crosses 1/2 at $|\omega_0| = 1.908$, and falls below it thereafter.

5 Hierarchical Models

Contents

5.1 Exercises 5.1–5.7

Problem 5.1. Exchangeability with known model parameters: For each of the following three examples, answer: (i) Are observations y_1 and y_2 exchangeable? (ii) Are observations y_1 and y_2 independent? (iii) Can we act as if the two observations are independent?
(a) A box has one black ball and one white ball. We pick a ball y_1 at random, put it back, and pick another ball y_2 at random.
(b) A box has one black ball and one white ball. We pick a ball y_1 at random, we do not put it back, then we pick ball y_2.
(c) A box has a million black balls and a million white balls. We pick a ball y_1 at random, we do not put it back, then we pick ball y_2 at random.

Solution. Exchangeable in all three; independent only in (a); safe to treat as independent in (a) and (c). Code black as 1, white as 0.
(a) The joint distribution is $p(y_1, y_2) = \frac{1}{4}$ on $\{0,1\}^2$, symmetric in its arguments and a product of two Bernoulli($\frac{1}{2}$) marginals. So (i) yes, (ii) yes, (iii) yes.
(b) The second draw is the ball the first draw left behind, so

$$p(y_1, y_2) = \tfrac{1}{2} \quad \text{on } (0,1) \text{ and } (1,0), \qquad p(y_1, y_2) = 0 \text{ otherwise.}$$

This is symmetric under swapping y_1, y_2, so (i) yes. But $y_2 = 1 - y_1$ with probability one, whereas $p(y_1)p(y_2) = \frac{1}{4}$ everywhere; equivalently

$$\operatorname{cov}(y_1, y_2) = E(y_1 y_2) - E(y_1)E(y_2) = 0 - \tfrac{1}{4} = -\tfrac{1}{4},$$

the extreme negative value for Bernoulli($\frac{1}{2}$) variables. So (ii) no, and (iii) emphatically no: the observations are as dependent as two binary variables can be.
(c) With $N = 10^6$ of each colour, every ordered pair of draws is equally likely, so $p(y_1, y_2)$ depends only on the multiset $\{y_1, y_2\}$ and (i) holds. For (ii),

$$p(y_2 = 1 \mid y_1 = 1) = \frac{N-1}{2N-1} = \frac{999{,}999}{1{,}999{,}999} \neq \tfrac{1}{2},$$

so no. But the dependence is numerically nil:

$$\operatorname{corr}(y_1, y_2) = -\frac{1}{2N-1} = -5.0 \times 10^{-7},$$

so for any practical inference (iii) yes – acting as if the draws were independent Bernoulli($\frac{1}{2}$) perturbs no probability by more than about 10^{-6}.

Problem 5.2. Exchangeability with unknown model parameters: For each of the following three examples, answer: (i) Are observations y_1 and y_2 exchangeable? (ii) Are observations y_1 and y_2 independent? (iii) Can we act as if the two observations are independent?
(a) A box has n black and white balls but we do not know how many of each color. We pick a ball y_1 at random, put it back, and pick another ball y_2 at random.
(b) A box has n black and white balls but we do not know how many of each color. We pick a ball y_1 at random, we do not put it back, then we pick ball y_2 at random.
(c) Same as (b) but we know that there are many balls of each color in the box.

Solution. Exchangeable in all three; independent in none; and in none of the three may we act as if they were independent – though in (a) and (c) they are independent **conditional** on the unknown composition. Write θ for the unknown proportion of black balls, with prior $p(\theta)$, and code black as 1.

(a) Given θ the draws are i.i.d. Bernoulli(θ), so

$$p(y_1, y_2) = \int \theta^{y_1+y_2}(1-\theta)^{2-y_1-y_2}\, p(\theta)\, d\theta,$$

which depends on (y_1, y_2) only through $y_1 + y_2$: exchangeable, (i) yes. This is exactly the mixture-of-i.i.d. form (5.2), and by Exercise 5.5,

$$\operatorname{cov}(y_1, y_2) = \operatorname{var}\big(E(y_1 \mid \theta)\big) = \operatorname{var}(\theta) > 0$$

unless $p(\theta)$ is a point mass. So (ii) no. For (iii): y_1 is informative about θ and hence about y_2, by an amount $\operatorname{var}(\theta)$ that is not small when we are genuinely ignorant of the composition (for $\theta \sim \mathrm{U}(0,1)$, cov $= 1/12$ and corr $= 1/3$). We may act as if they are **conditionally** independent given θ, which is what the hierarchical model does, but not as if they are marginally independent.
(b) Given the number $n\theta$ of black balls the two draws are an exchangeable sample without replacement, so $p(y_1, y_2 \mid \theta)$ is symmetric, and averaging over $p(\theta)$ preserves symmetry: (i) yes. Now two dependencies act at once,

$$\operatorname{cov}(y_1, y_2) = \underbrace{E\big[\operatorname{cov}(y_1, y_2 \mid \theta)\big]}_{= -E[\theta(1-\theta)]/(n-1)} + \underbrace{\operatorname{var}\big(E(y_1 \mid \theta)\big)}_{= \operatorname{var}(\theta)},$$

the negative finite-population term and the positive learning-about-θ term. Both are generally nonzero, so (ii) no and (iii) no.
(c) Exchangeable for the same reason, (i) yes. "Many balls of each color" kills the first term above – it is $O(1/n)$ – but leaves $\operatorname{var}(\theta)$ untouched, so (ii) no, and (iii) no: unlike Exercise 5.1(c), where $\theta = \frac{1}{2}$ was known, here the residual dependence is the whole of our uncertainty about the composition. What we may do is treat the draws as i.i.d. Bernoulli(θ) given θ, ignoring the without-replacement correction.

Problem 5.3. Hierarchical models and multiple comparisons:
(a) Reproduce the computations in Section 5.5 for the educational testing example. Use the posterior simulations to estimate (i) for each school j, the probability that its coaching program is the best of the eight; and (ii) for each pair of schools, j and k, the probability that the coaching program in school j is better than that in school k.
(b) Repeat (a), but for the simpler model with τ set to ∞ (that is, separate estimation for the eight schools). In this case, the probabilities (ii) can be computed analytically.
(c) Discuss how the answers in (a) and (b) differ.
(d) In the model with τ set to 0, the probabilities (i) and (ii) have degenerate values; what are they?
The data of Table 5.2 (observed effects of special preparation on SAT-V scores in eight randomized experiments) are:

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28	15
B	8	10
C	−3	16
D	7	11
E	−1	9
F	1	11
G	18	10
H	12	18

The model of Section 5.5 is $y_j \mid \theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$ with σ_j known, $\theta_j \mid \mu, \tau \sim \mathrm{N}(\mu, \tau^2)$, and $p(\mu, \tau) \propto 1$.

Solution. Under the hierarchical model no pairwise comparison is sharper than about 0.73 and no school is best with probability above 0.25; under separate estimation school A is best with probability 0.55 and beats school E with probability 0.95.

(a) Simulation follows Section 5.4. The hyperprior $p(\mu, \tau) \propto 1$ is improper, but Section 5.4 records that uniform τ (unlike uniform $\log \tau$) leaves $p(\tau \mid y)$ integrable, so the posterior below is proper. Evaluate the marginal posterior (5.21) for τ on a grid, $\tau \in (0, 40]$ in steps of 0.01, where with $V_j = \sigma_j^2 + \tau^2$,

$$\hat{\mu}(\tau) = \frac{\sum_j y_j/V_j}{\sum_j 1/V_j}, \qquad V_\mu(\tau) = \Big(\sum_j 1/V_j\Big)^{-1},$$

$$p(\tau \mid y) \propto V_\mu(\tau)^{1/2} \prod_{j=1}^{8} V_j^{-1/2} \exp\Big(-\frac{(y_j - \hat{\mu})^2}{2V_j}\Big),$$

then draw τ from the grid, $\mu \mid \tau, y \sim \mathrm{N}(\hat{\mu}, V_\mu)$ by (5.20), and finally, independently for each j, by (5.17),

$$\theta_j \mid \mu, \tau, y \sim \mathrm{N}(\hat{\theta}_j, V_j^*),$$
$$\hat{\theta}_j = \frac{y_j/\sigma_j^2 + \mu/\tau^2}{1/\sigma_j^2 + 1/\tau^2}, \qquad V_j^* = \Big(\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\Big)^{-1}.$$

With $S = 200{,}000$ draws the posterior quantiles reproduce Table 5.3 (posterior means 11.4, 7.9, 6.2, 7.7, 5.1, 6.2, 10.7, 8.5; posterior median of τ equal to 5.2). Counting draws gives (i), with Monte Carlo standard error at most 0.001:

School	A	B	C	D	E	F	G	H
$\Pr(\theta_j$ largest)	0.25	0.10	0.08	0.10	0.05	0.07	0.21	0.14

and for (ii), $\Pr(\theta_j > \theta_k \mid y)$ with j indexing rows:

	A	B	C	D	E	F	G	H
A		0.63	0.68	0.64	0.73	0.69	0.52	0.61
B	0.37		0.57	0.51	0.63	0.58	0.38	0.48
C	0.32	0.43		0.44	0.55	0.51	0.33	0.42
D	0.36	0.49	0.56		0.62	0.57	0.37	0.47
E	0.27	0.37	0.45	0.38		0.45	0.27	0.36
F	0.31	0.42	0.49	0.43	0.55		0.31	0.41
G	0.48	0.62	0.67	0.63	0.73	0.69		0.59
H	0.39	0.52	0.58	0.53	0.64	0.59	0.41	

Every entry lies in $[0.27, 0.73]$: the data do not permit a confident ranking of any pair.

(b) With $\tau = \infty$ the prior on θ is flat and the eight analyses separate: $\theta_j \mid y \sim \mathrm{N}(y_j, \sigma_j^2)$ independently. Then $\theta_j - \theta_k \sim \mathrm{N}(y_j - y_k, \sigma_j^2 + \sigma_k^2)$, so (ii) is available in closed form,

$$\Pr(\theta_j > \theta_k \mid y) = \Phi\left(\frac{y_j - y_k}{\sqrt{\sigma_j^2 + \sigma_k^2}}\right),$$

giving (rows j):

	A	B	C	D	E	F	G	H
A		0.87	0.92	0.87	0.95	0.93	0.71	0.75
B	0.13		0.72	0.53	0.75	0.68	0.24	0.42
C	0.08	0.28		0.30	0.46	0.42	0.13	0.27
D	0.13	0.47	0.70		0.71	0.65	0.23	0.41
E	0.05	0.25	0.54	0.29		0.44	0.08	0.26
F	0.07	0.32	0.58	0.35	0.56		0.13	0.30
G	0.29	0.76	0.87	0.77	0.92	0.87		0.61
H	0.25	0.58	0.73	0.59	0.74	0.70	0.39	

The "best of eight" probabilities (i) are an eight-dimensional orthant probability; simulating 200,000 independent draws of θ gives (Monte Carlo standard error at most 0.001)

School	A	B	C	D	E	F	G	H
$\Pr(\theta_j$ largest)	0.549	0.035	0.026	0.036	0.003	0.012	0.170	0.168

(c) The $\tau = \infty$ probabilities are far more extreme. Shrinkage towards a common μ (posterior median $\tau = 5.2$, small relative to the σ_j) pulls the differences $\theta_j - \theta_k$ towards zero while their posterior spread shrinks much less, so under (a) no pairwise probability leaves $[0.27, 0.73]$ and every school's chance of being best lies between 0.05 and 0.25, near the 1/8 of complete ignorance; under (b) school A is best with probability 0.55 and beats school E with probability 0.95. A pair can even reverse: $\Pr(\theta_C > \theta_E \mid y) = 0.55$ in (a) but 0.46 in (b), because the σ_j differ and shrinkage is stronger for the noisier school C. The hierarchical model, by letting the data report that τ is small, supplies the multiplicity correction that separate estimation would need to add by hand.
(d) With $\tau = 0$ all eight parameters coincide, $\theta_1 = \cdots = \theta_8 = \mu$, so $\Pr(\theta_j > \theta_k \mid y) = 0$ for every $j \neq k$ and the probability that any given school is strictly the best is 0. Breaking the ties at random – the usual convention – these become $\Pr(\theta_j > \theta_k) = \frac{1}{2}$ for all $j \neq k$ and $\Pr(j \text{ best}) = \frac{1}{8}$ for all j: the data are given no weight at all in ranking the schools.

Problem 5.4. Exchangeable prior distributions: suppose it is known a priori that the $2J$ parameters $\theta_1, \ldots, \theta_{2J}$ are clustered into two groups, with exactly half being drawn from a $\mathrm{N}(1, 1)$ distribution, and the other half being drawn from a $\mathrm{N}(-1, 1)$ distribution, but we have not observed which parameters come from which distribution.
(a) Are $\theta_1, \ldots, \theta_{2J}$ exchangeable under this prior distribution?
(b) Show that this distribution cannot be written as a mixture of independent and identically distributed components.
(c) Why can we not simply take the limit as $J \to \infty$ and get a counterexample to de Finetti's theorem?
See Exercise 8.10 for a related problem.

Solution. Yes to (a); (b) follows from Exercise 5.5 because the covariances are negative; (c) because the family is not the set of marginals of any one infinite exchangeable sequence, and its limit is i.i.d. anyway.
(a) Yes. Let S be the set of indices drawn from $\mathrm{N}(1, 1)$, uniform over the $\binom{2J}{J}$ subsets of size J. Then

$$p(\theta) = \binom{2J}{J}^{-1} \sum_{|S|=J} \prod_{j \in S} \mathrm{N}(\theta_j \mid 1, 1) \prod_{j \notin S} \mathrm{N}(\theta_j \mid -1, 1),$$

and a permutation π of $(1, \ldots, 2J)$ merely relabels the summand S as $\pi^{-1}(S)$, permuting the terms of a sum over **all** subsets of size J. So p is invariant to permutations: exchangeable.
(b) Let $z_j = 1$ if $j \in S$ and $z_j = 0$ otherwise, so $E(\theta_j \mid z) = 2z_j - 1$ and $E(\theta_j) = 0$ by symmetry. The z_j are a simple random sample of J ones among $2J$ positions, so for $i \neq j$,

$$\Pr(z_i = z_j) = 2 \cdot \frac{J(J-1)}{2J(2J-1)} = \frac{J-1}{2J-1},$$

and, using $\mathrm{cov}(\theta_i, \theta_j) = E[\mathrm{cov}(\theta_i, \theta_j \mid z)] + \mathrm{cov}(E(\theta_i \mid z), E(\theta_j \mid z))$ with the first term zero (θ are independent given z),

$$\begin{aligned} \mathrm{cov}(\theta_i, \theta_j) &= E\big[(2z_i - 1)(2z_j - 1)\big] \\ &= \Pr(z_i = z_j) - \Pr(z_i \neq z_j) \\ &= \frac{J-1}{2J-1} - \frac{J}{2J-1} = -\frac{1}{2J-1} < 0. \end{aligned}$$

By Exercise 5.5 every mixture of i.i.d. components has nonnegative covariances, so this distribution is not of that form.
(c) Because there is no single infinite exchangeable sequence here to which de Finetti's theorem could be applied. The construction gives, for each J, a distribution on $2J$ variables, and these are not consistent: the $2J$-dimensional law is not the marginal of the $2(J+1)$-dimensional one (the first has covariance $-1/(2J-1)$, the second $-1/(2J+1)$). De Finetti's theorem requires a single distribution on $(\theta_1, \theta_2, \ldots)$ whose finite-dimensional marginals are exchangeable.

Nor does the limit produce a counterexample: for fixed k, the count of the first k indices in S is Hypergeometric$(2J, J, k) \to \text{Bin}(k, \frac{1}{2})$, so the k-dimensional marginal converges to

$$\prod_{j=1}^{k} \Big[\tfrac{1}{2}\mathrm{N}(\theta_j \mid 1, 1) + \tfrac{1}{2}\mathrm{N}(\theta_j \mid -1, 1)\Big],$$

i.i.d. as de Finetti requires, with $\text{cov}(\theta_i, \theta_j) = -1/(2J-1) \to 0$.

Problem 5.5. Mixtures of independent distributions: suppose the distribution of $\theta = (\theta_1, \ldots, \theta_J)$ can be written as a mixture of independent and identically distributed components:

$$p(\theta) = \int \prod_{j=1}^{J} p(\theta_j \mid \phi)\, p(\phi)\, d\phi.$$

Prove that the covariances $\text{cov}(\theta_i, \theta_j)$ are all nonnegative.

Solution. $\text{cov}(\theta_i, \theta_j) = \text{var}\big(E(\theta_i \mid \phi)\big) \geq 0$.
Assume $E(\theta_j^2) < \infty$, so that all the quantities below exist. Apply (1.9) to $\theta_i + \theta_j$ and to $\theta_i - \theta_j$ and subtract; the resulting bilinear form of that identity reads

$$\text{cov}(\theta_i, \theta_j) = E\big[\,\text{cov}(\theta_i, \theta_j \mid \phi)\big] + \text{cov}\,\big(E(\theta_i \mid \phi),\, E(\theta_j \mid \phi)\big).$$

For $i \neq j$ the first term vanishes: given ϕ the components are independent by hypothesis, so $\text{cov}(\theta_i, \theta_j \mid \phi) = 0$. In the second term the components are identically distributed given ϕ, so $E(\theta_i \mid \phi) = E(\theta_j \mid \phi) =: m(\phi)$, and the covariance of a random variable with itself is its variance:

$$\text{cov}(\theta_i, \theta_j) = 0 + \text{cov}\,\big(m(\phi), m(\phi)\big) = \text{var}\,\big(m(\phi)\big) \geq 0.$$

For $i = j$ the statement is $\text{var}(\theta_i) \geq 0$.

Problem 5.6. Exchangeable models:
(a) In the divorce rate example of Section 5.2, set up a prior distribution for the values $y_1, \ldots, y_8$ that allows for one low value (Utah) and one high value (Nevada), with independent and identical distributions for the other six values. This prior distribution should be exchangeable, because it is not known which of the eight states correspond to Utah and Nevada.
(b) Determine the posterior distribution for y_8 under this model given the observed values of $y_1, \ldots, y_7$ given in the example. This posterior distribution should probably have two or three modes, corresponding to the possibilities that the missing state is Utah, Nevada, or one of the other six.
(c) Now consider the entire set of eight data points, including the value for y_8 given at the end of the example. Are these data consistent with the prior distribution you gave in part (a) above? In particular, did your prior distribution allow for the possibility that the actual data have an outlier (Nevada) at the high end, but no outlier at the low end?
In the example of Section 5.2, $y_1, \ldots, y_8$ are the 1981 divorce rates per 1000 population of eight Mountain states (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming), presented in a random order so that the index does not reveal the state. Seven of the eight are observed:

$$5.8, \quad 6.6, \quad 7.8, \quad 5.6, \quad 7.0, \quad 7.1, \quad 5.4.$$

At the end of the example we are told that the unobserved state is Nevada, whose 1981 divorce rate was 13.9.

Solution. The posterior for y_8 is a three-component mixture, putting weight 0.22 on Utah, 0.44 on Nevada and 0.34 on an ordinary state; and the eight actual data points are not consistent with the

prior, which insists on a low outlier that is not there.
(a) Introduce a latent labelling: let (L, H) be an ordered pair of distinct indices drawn uniformly from the $8 \cdot 7 = 56$ possibilities, L marking Utah and H Nevada. Given the labels,

$$y_L \sim \mathrm{N}(4, 1^2), \qquad y_H \sim \mathrm{N}(12, 3^2),$$
$$y_j \overset{\text{iid}}{\sim} \mathrm{N}(6.5, 1.5^2) \quad (j \neq L, H),$$

all eight independent. (The means are a pre-data guess: a typical Mountain state near 6.5 per 1000, Utah well below, Nevada far above, with a generous spread on Nevada.) Marginally

$$p(y_1, \ldots, y_8) = \frac{1}{56} \sum_{L \neq H} \mathrm{N}(y_L \mid 4, 1)\, \mathrm{N}(y_H \mid 12, 9) \times \prod_{j \neq L, H} \mathrm{N}(y_j \mid 6.5, 2.25),$$

and permuting the indices permutes the 56 summands among themselves, so the prior is exchangeable.
(b) Condition on the seven observed values y_{obs} and partition by the role of index 8:

- (i) $8 = L$: then one of the seven observed is Nevada and six are ordinary;
- (ii) $8 = H$: then one of the seven observed is Utah and six are ordinary;
- (iii) $8 \notin \{L, H\}$: both Utah and Nevada are among the seven observed.

Each case contributes its likelihood summed over the admissible label placements; with f_L, f_H, f_0 the three densities and $P_0 = \prod_{i=1}^{7} f_0(y_i)$,

$$w_{(\text{i})} \propto P_0 \sum_{i=1}^{7} \frac{f_H(y_i)}{f_0(y_i)},$$
$$w_{(\text{ii})} \propto P_0 \sum_{i=1}^{7} \frac{f_L(y_i)}{f_0(y_i)},$$
$$w_{(\text{iii})} \propto P_0 \sum_{i \neq i'} \frac{f_L(y_i)}{f_0(y_i)} \frac{f_H(y_{i'})}{f_0(y_{i'})}.$$

Evaluating at $y_{\text{obs}} = (5.8, 6.6, 7.8, 5.6, 7.0, 7.1, 5.4)$ and normalizing gives

$$w_{(\text{i})} = 0.222, \qquad w_{(\text{ii})} = 0.440, \qquad w_{(\text{iii})} = 0.338.$$

Since y_8 is independent of the rest given the labels, the posterior is the corresponding mixture of the three prior components:

$$p(y_8 \mid y_{\text{obs}}) = 0.222\, \mathrm{N}(4, 1^2) + 0.440\, \mathrm{N}(12, 3^2) + 0.338\, \mathrm{N}(6.5, 1.5^2),$$

trimodal with modes at 4, 6.5 and 12, posterior mean 8.4 and standard deviation 4.0, and $\Pr(y_8 > 10 \mid y_{\text{obs}}) = 0.33$. The largest weight falls on "the missing state is Nevada" because the seven observed values are all tightly clustered and one of them (5.4, with conditional probability 0.45 of being Utah, or 5.6 with 0.30) can play Utah comfortably, whereas nothing in the range 5.4 to 7.8 plays Nevada well.
(c) No. With $y_8 = 13.9$ the complete data are 5.4, 5.6, 5.8, 6.6, 7.0, 7.1, 7.8, 13.9: one clear high outlier and no low one. But the prior of part (a) forces exactly one draw from $\mathrm{N}(4, 1)$, and the smallest observation is 5.4. Taking $T(y) = \min_j y_j$ as a check statistic, the prior predictive probability is

$$\Pr\left(T(y^{\text{rep}}) \geq 5.4\right) = 0.016,$$

whereas the high end is entirely comfortable, $\Pr(\max_j y_j^{\text{rep}} \geq 13.9) = 0.26$. So the prior did not allow a high outlier without a matching low one: Utah's actual 1981 rate sits inside the cluster of ordinary states. A prior that would accommodate these data must let the number of outliers at each end be random – an exchangeable i.i.d. mixture, or simply a heavy-tailed i.i.d. distribution centered at 6.5.

Problem 5.7. Continuous mixture models:
(a) If $y \mid \theta \sim \text{Poisson}(\theta)$, and $\theta \sim \text{Gamma}(\alpha, \beta)$, then the marginal (prior predictive) distribution of y is negative binomial with parameters α and β (or $p = \beta/(1+\beta)$). Use the formulas (2.7) and (2.8) to derive the mean and variance of the negative binomial.
(b) In the normal model with unknown location and scale (μ, σ^2), the noninformative prior density, $p(\mu, \sigma^2) \propto 1/\sigma^2$, results in a normal-inverse-χ^2 posterior distribution for (μ, σ^2). Marginally then $\sqrt{n}(\mu - \bar{y})/s$ has a posterior distribution that is t_{n-1}. Use (2.7) and (2.8) to derive the first two moments of the latter distribution, stating the appropriate condition on n for existence of both moments.

Solution. Mean α/β and variance $(\alpha/\beta)(1 + 1/\beta)$ in (a); mean 0 and variance $(n-1)/(n-3)$ in (b), requiring $n > 2$ and $n > 3$ respectively. Both parts apply (2.7) and (2.8) with the conditioning variable named below.
(a) Condition on θ, so $E(y \mid \theta) = \text{var}(y \mid \theta) = \theta$, and use $E(\theta) = \alpha/\beta$, $\text{var}(\theta) = \alpha/\beta^2$ for the $\text{Gamma}(\alpha, \beta)$ prior (Appendix A):

$$\begin{aligned} E(y) &= E\big(E(y \mid \theta)\big) = E(\theta) = \frac{\alpha}{\beta}, \\ \text{var}(y) &= E\big(\text{var}(y \mid \theta)\big) + \text{var}\big(E(y \mid \theta)\big) \\ &= E(\theta) + \text{var}(\theta) = \frac{\alpha}{\beta} + \frac{\alpha}{\beta^2} = \frac{\alpha}{\beta}\Big(1 + \frac{1}{\beta}\Big). \end{aligned}$$

In terms of $p = \beta/(1+\beta)$, so that $1 - p = 1/(1+\beta)$ and $\beta = p/(1-p)$, these read

$$E(y) = \frac{\alpha(1-p)}{p}, \qquad \text{var}(y) = \frac{\alpha(1-p)}{p^2} = \frac{E(y)}{p},$$

the standard negative binomial moments.
(b) Write $t = \sqrt{n}(\mu - \bar{y})/s$ and condition on σ^2, treating $\bar{y}$ and s as fixed data. The posterior factors as $\mu \mid \sigma^2, y \sim \text{N}(\bar{y}, \sigma^2/n)$ and $\sigma^2 \mid y \sim \text{Inv-}\chi^2(n-1, s^2)$ (equations (3.3) and (3.5)), so

$$t \mid \sigma^2, y \sim \text{N}\Big(0,\ \frac{\sigma^2}{s^2}\Big).$$

Hence $E(t \mid \sigma^2) = 0$ and $\text{var}(t \mid \sigma^2) = \sigma^2/s^2$, and by (2.7) and (2.8),

$$\begin{aligned} E(t) &= E\big(E(t \mid \sigma^2)\big) = 0, \\ \text{var}(t) &= E\Big(\frac{\sigma^2}{s^2}\Big) + \text{var}(0) = \frac{1}{s^2} \cdot \frac{(n-1)s^2}{(n-1)-2} = \frac{n-1}{n-3}, \end{aligned}$$

using $E(\sigma^2 \mid y) = \nu s^2/(\nu - 2)$ for $\sigma^2 \sim \text{Inv-}\chi^2(\nu, s^2)$ with $\nu = n - 1$.
The conditions are the hypotheses of those two identities, that the moments being averaged exist: $E|t| < \infty$ needs $E(\sigma \mid y) < \infty$, finite for $\nu > 1$, that is $n > 2$; and $E(\sigma^2 \mid y) < \infty$ needs $\nu > 2$, that is $n > 3$.

5.2 Exercises 5.8–5.14

Problem 5.8. Discrete mixture models: if $p_m(\omega)$, for $m = 1, \ldots, M$, are conjugate prior densities for the sampling model $y \mid \omega$, show that the class of finite mixture prior densities given by

$$p(\omega) = \sum_{m=1}^{M} \lambda_m \, p_m(\omega)$$

is also a conjugate class, where the λ_m's are nonnegative weights that sum to 1. This can provide a useful extension of the natural conjugate prior family to more flexible distributional forms. As an

example, use the mixture form to create a bimodal prior density for a normal mean, that is thought to be near 1, with a standard deviation of 0.5, but has a small probability of being near -1, with the same standard deviation. If the variance of each observation $y_1, \dots, y_{10}$ is known to be 1, and their observed mean is $\bar{y} = -0.25$, derive your posterior distribution for the mean, making a sketch of both prior and posterior densities. Be careful: the prior and posterior mixture proportions are different.

Solution. The posterior is the mixture of the component posteriors with weights reweighted by the component prior predictive densities:

$$p(\omega \mid y) = \sum_{m=1}^{M} \lambda'_m \, p_m(\omega \mid y), \qquad \lambda'_m = \frac{\lambda_m \, q_m(y)}{\sum_l \lambda_l \, q_l(y)},$$

where $q_m(y) = \int p(y \mid \omega) p_m(\omega) \, d\omega$ is the marginal likelihood under component m. Indeed,

$$\begin{aligned} p(\omega \mid y) &\propto p(y \mid \omega) \sum_m \lambda_m p_m(\omega) \\ &= \sum_m \lambda_m \, p(y \mid \omega) p_m(\omega) \\ &= \sum_m \lambda_m \, q_m(y) \, p_m(\omega \mid y), \end{aligned}$$

and each $p_m(\omega \mid y)$ lies in the same parametric family as p_m by conjugacy. Hence the posterior is again a finite mixture of members of that family, i.e. the mixture class is closed under sampling. (Each $q_m(y)$ is finite and positive, so the λ'_m are well defined weights summing to 1.)
For the example, take the two-component normal prior

$$p(\omega) = 0.9 \, \mathrm{N}(\omega \mid 1, 0.5^2) + 0.1 \, \mathrm{N}(\omega \mid -1, 0.5^2),$$

which is bimodal with modes at ± 1 (heights 0.718 and 0.080). The data enter only through $\bar{y} \mid \omega \sim \mathrm{N}(\omega, \sigma^2/n)$ with $\sigma^2/n = 1/10$. By (2.12) each component posterior has precision $0.5^{-2} + 10 = 14$, so

$$V = \tfrac{1}{14} = 0.07143, \qquad \sqrt{V} = 0.26726,$$

and component means

$$\begin{aligned} \mu'_1 &= \frac{4(1) + 10(-0.25)}{14} = \tfrac{3}{28} = 0.10714, \\ \mu'_2 &= \frac{4(-1) + 10(-0.25)}{14} = -\tfrac{13}{28} = -0.46429. \end{aligned}$$

The marginal likelihoods are $q_m = \mathrm{N}(\bar{y} \mid \mu_m, \, 0.25 + 0.1)$, so

$$\frac{q_1}{q_2} = \exp\!\left(-\frac{(-1.25)^2 - (0.75)^2}{2(0.35)} \right) = e^{-10/7} = 0.23965,$$

giving

$$\lambda'_1 = \frac{0.9(0.23965)}{0.9(0.23965) + 0.1} = 0.6832, \qquad \lambda'_2 = 0.3168.$$

Thus

$$p(\omega \mid y) = 0.683 \, \mathrm{N}(\omega \mid 0.107, \, 0.267^2) + 0.317 \, \mathrm{N}(\omega \mid -0.464, \, 0.267^2),$$

with mean -0.0739 and standard deviation 0.377: the prior odds 0.9:0.1 become posterior odds 0.683:0.317, since $\bar{y} = -0.25$ sits much closer to -1 than to 1. Sketch: the prior has two well-separated humps at -1 and 1, the right one about nine times taller; the posterior collapses onto $(-1, 1)$, its two component means only 0.571 apart against a common standard deviation 0.267, so it is unimodal with a single peak at $\omega = 0.074$ and a pronounced left shoulder near -0.46.

Problem 5.9. Noninformative hyperprior distributions: consider the hierarchical binomial model in Section 5.3. Improper posterior distributions are, in fact, a general problem with hierarchical models when a uniform prior distribution is specified for the logarithm of the population standard deviation of the exchangeable parameters. In the case of the beta population distribution, the prior variance is approximately $(\alpha+\beta)^{-1}$ (see Appendix A), and so a uniform distribution on $\log(\alpha+\beta)$ is approximately uniform on the log standard deviation. The resulting unnormalized posterior density (5.8),

$$p(\alpha,\beta \mid y) \propto p(\alpha,\beta)\prod_{j=1}^{J}\frac{B(\alpha+y_j,\ \beta+n_j-y_j)}{B(\alpha,\beta)},$$

has an infinite integral in the limit as the population standard deviation approaches 0. We encountered the problem again in Section 5.4 for the hierarchical normal model.
(a) Show that, with a uniform prior density on $(\log(\alpha/\beta), \log(\alpha+\beta))$, the unnormalized posterior density has an infinite integral.
(b) A simple way to avoid the impropriety is to assign a uniform prior distribution to the standard deviation parameter itself, rather than its logarithm. For the beta population distribution we are considering here, this is achieved approximately by assigning a uniform prior distribution to $(\alpha+\beta)^{-1/2}$. Show that combining this with an independent uniform prior distribution on $\frac{\alpha}{\alpha+\beta}$ yields the prior density (5.10),

$$p\Big(\log\Big(\tfrac{\alpha}{\beta}\Big), \log(\alpha+\beta)\Big) \propto \alpha\beta(\alpha+\beta)^{-5/2}.$$

(c) Show that the resulting posterior density (5.8) is proper as long as $0 < y_j < n_j$ for at least one experiment j.

Solution. Throughout write $s = \alpha+\beta$ and $p = \alpha/(\alpha+\beta)$, so that $u = \log(\alpha/\beta) = \operatorname{logit}(p)$ and $v = \log s$. The Jacobian of $(u,v) \mapsto (\alpha,\beta)$ is $\alpha\beta$, since

$$\left|\frac{\partial(u,v)}{\partial(\alpha,\beta)}\right| = \left|\frac{1}{\alpha}\cdot\frac{1}{s} + \frac{1}{\beta}\cdot\frac{1}{s}\right| = \frac{1}{\alpha\beta}.$$

Each factor of (5.8) is the beta-binomial weight

$$L_j(p,s) \;=\; \frac{B(\alpha+y_j,\ \beta+n_j-y_j)}{B(\alpha,\beta)} \;=\; \mathrm{E}_{\omega\sim\mathrm{Beta}(\alpha,\beta)}\big[\omega^{y_j}(1-\omega)^{n_j-y_j}\big] \le 1.$$

(a) The integrand does not decay as $v \to +\infty$. Population standard deviation $\approx s^{-1/2} \to 0$ means $s \to \infty$ at fixed p, and there $\mathrm{Beta}(\alpha,\beta)$ concentrates at p: writing the factor out,

$$L_j = \frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)}\cdot\frac{\Gamma(\beta+n_j-y_j)}{\Gamma(\beta)}\cdot\frac{\Gamma(s)}{\Gamma(s+n_j)},$$

with $\Gamma(\alpha+y_j)/\Gamma(\alpha) = \alpha(\alpha+1)\cdots(\alpha+y_j-1) \sim \alpha^{y_j}$, likewise $\sim \beta^{n_j-y_j}$, and $\Gamma(s)/\Gamma(s+n_j) \sim s^{-n_j}$, so

$$L_j(p,s) \longrightarrow p^{y_j}(1-p)^{n_j-y_j} \qquad (s\to\infty,\ p \text{ fixed}).$$

With $p(u,v) \propto 1$ the unnormalized posterior therefore tends to the strictly positive constant $c(p) = \prod_j p^{y_j}(1-p)^{n_j-y_j}$, and the convergence is uniform on any compact u-interval K. Choosing V so large that the integrand exceeds $\frac{1}{2}\min_{u\in K} c(p) > 0$ for $v > V$,

$$\int_K\int_V^\infty p(u,v \mid y)\,dv\,du \;\ge\; \tfrac{1}{2}|K|\min_{u\in K} c(p)\int_V^\infty dv = \infty.$$

(b) Put $t = s^{-1/2}$ and let (p,t) be independent uniform, i.e. $p(p,t) \propto 1$. Changing variables one coordinate at a time,

$$\frac{dp}{du} = p(1-p) = \frac{\alpha\beta}{s^2}, \qquad \left|\frac{dt}{dv}\right| = \tfrac{1}{2}e^{-v/2} \propto s^{-1/2},$$

so

$$p(u,v) \;\propto\; \frac{\alpha\beta}{s^2}\cdot s^{-1/2} \;=\; \alpha\beta\,(\alpha+\beta)^{-5/2},$$

which is (5.10). Dividing by the Jacobian $\alpha\beta$ recovers $p(\alpha,\beta) \propto (\alpha+\beta)^{-5/2}$, i.e. (5.9).
(c) Work in the coordinates $(p,t) \in (0,1)\times(0,\infty)$, where by (b) the prior is Lebesgue measure and the posterior mass is $\int_0^1\int_0^\infty \prod_j L_j\,dt\,dp$. Split at $t=1$.
(i) On $t \in (0,1)$: $\prod_j L_j \le 1$ and the region has measure 1, so this part contributes at most 1.
(ii) On $t>1$, i.e. $s<1$: fix an index j^* with $0 < y_{j^*} < n_{j^*}$ and bound all other factors by 1. For $s \le 1$ we have $\alpha,\beta \le 1$, hence

$$\begin{aligned}
\frac{\Gamma(\alpha+y)}{\Gamma(\alpha)} &= \alpha\prod_{i=1}^{y-1}(\alpha+i) \le \alpha\,y!,\\
\frac{\Gamma(\beta+n-y)}{\Gamma(\beta)} &\le \beta\,(n-y)!,\\
\frac{\Gamma(s)}{\Gamma(s+n)} &= \frac{1}{s(s+1)\cdots(s+n-1)} \le \frac{1}{s\,(n-1)!},
\end{aligned}$$

writing $y = y_{j^*}$, $n = n_{j^*}$ (the first two products are nonempty precisely because $0<y<n$). Multiplying, and using $\alpha\beta = s^2p(1-p) \le s^2$,

$$L_{j^*}(p,s) \;\le\; \frac{y!\,(n-y)!}{(n-1)!}\,\frac{\alpha\beta}{s} \;\le\; C\,s \;=\; C\,t^{-2},$$

uniformly in p. Hence this part contributes at most $\int_0^1\int_1^\infty Ct^{-2}\,dt\,dp = C < \infty$.
So the posterior is proper. (If instead every $y_j \in \{0, n_j\}$ then the same expansion gives $L_j \to (1-p)$ or p as $s \to 0$, the integrand does not decay in t, and the integral diverges — so the stated condition is sharp.)

Problem 5.10. Checking the integrability of the posterior distribution: consider the hierarchical normal model in Section 5.4, in which $\bar y_{\cdot j} \mid \omega_j \sim \mathrm{N}(\omega_j, \sigma_j^2)$ with known σ_j^2, $\omega_j \mid \mu,\tau \sim \mathrm{N}(\mu,\tau^2)$ for $j = 1,\dots,J$, and the marginal posterior density of τ is (5.21),

$$p(\tau \mid y) \propto p(\tau)\,V_\mu^{1/2}\prod_{j=1}^J(\sigma_j^2+\tau^2)^{-1/2}\exp\left(-\frac{(\bar y_{\cdot j}-\hat\mu)^2}{2(\sigma_j^2+\tau^2)}\right),$$

with $\hat\mu$ and V_μ given by (5.20),

$$\hat\mu = \frac{\sum_j \frac{1}{\sigma_j^2+\tau^2}\bar y_{\cdot j}}{\sum_j \frac{1}{\sigma_j^2+\tau^2}}, \qquad V_\mu^{-1} = \sum_{j=1}^J \frac{1}{\sigma_j^2+\tau^2}.$$

(a) If the hyperprior distribution is $p(\mu,\tau) \propto \tau^{-1}$ (that is, $p(\mu,\log\tau) \propto 1$), show that the posterior density is improper.
(b) If the hyperprior distribution is $p(\mu,\tau) \propto 1$, show that the posterior density is proper if $J > 2$.
(c) How would you analyze SAT coaching data if $J = 2$ (that is, data from only two schools)?

Solution. Since $p(\mu \mid \tau, y)$ is a proper normal density (5.20) and each $p(\omega_j \mid \mu,\tau,y)$ is a proper normal density (5.17), propriety of the joint posterior is exactly propriety of $\int_0^\infty p(\tau \mid y)\,d\tau$, so (5.21) decides both parts.
(a) Improper, by a nonintegrable singularity at $\tau = 0$. Write $g(\tau)$ for the factor of (5.21) following $p(\tau)$. Every ingredient of g is continuous on $[0,\infty)$ and positive at $\tau = 0$:

$$V_\mu^{1/2} \to \Big(\sum_j \sigma_j^{-2}\Big)^{-1/2}, \qquad \prod_j (\sigma_j^2+\tau^2)^{-1/2} \to \prod_j \sigma_j^{-1},$$
$$\exp(\cdot) \to \exp\Big(-\sum_j (\bar y_{\cdot j} - \hat\mu_0)^2/(2\sigma_j^2)\Big) > 0,$$

with $\hat\mu_0$ the precision-weighted mean at $\tau = 0$. So $g(0) = c > 0$ and $g(\tau) \geq c/2$ on some $(0, \varepsilon)$, whence

$$\int_0^\infty \tau^{-1} g(\tau)\, d\tau \;\geq\; \frac{c}{2}\int_0^\varepsilon \frac{d\tau}{\tau} \;=\; \infty.$$

(b) With $p(\tau) \propto 1$ the integrand is $g(\tau)$, which is continuous on $[0, \infty)$, so only the tail matters. Let $\sigma_{\max}^2 = \max_j \sigma_j^2$. From $V_\mu^{-1} = \sum_j (\sigma_j^2+\tau^2)^{-1} \geq J(\sigma_{\max}^2+\tau^2)^{-1}$ and $\exp(\cdot) \leq 1$ and $(\sigma_j^2+\tau^2)^{-1/2} \leq \tau^{-1}$,

$$g(\tau) \;\leq\; \frac{(\sigma_{\max}^2+\tau^2)^{1/2}}{\sqrt{J}} \cdot \tau^{-J} \;\leq\; \frac{\sigma_{\max}+\tau}{\sqrt{J}}\,\tau^{-J},$$

and $\int_1^\infty (\sigma_{\max}\tau^{-J} + \tau^{1-J})\, d\tau < \infty$ exactly when $J - 1 > 1$, i.e. $J > 2$. Combined with continuity on $[0, 1]$ this gives a finite integral, so the posterior is proper. The bound is tight: the same limits give $V_\mu^{1/2} \sim \tau/\sqrt{J}$, $\prod_j (\sigma_j^2+\tau^2)^{-1/2} \sim \tau^{-J}$ and $\exp(\cdot) \to 1$, so $g(\tau) \sim \tau^{1-J}/\sqrt{J}$ and the integral diverges for $J \leq 2$.
(c) Do not use the flat prior: by (b) it fails, and $g(\tau) \sim \tau^{-1}/\sqrt{2}$ means the improper posterior puts all its mass at $\tau = \infty$, i.e. no pooling at all — so that analysis is just the two schools fitted separately. Two group means give one contrast, hence almost no information with which to separate τ from the known σ_j^2, and whatever is assumed about τ is what gets reported. Two honest options: (i) a proper weakly informative prior, the half-Cauchy $\tau \sim \text{Cauchy}^+(0, A)$ of Section 5.7 with A at the high end of plausible school-to-school variation (order 25 points for the coaching data), whose τ^{-2} tail restores integrability while staying flat for small τ; (ii) a sensitivity analysis reporting $\text{E}(\omega_j \mid \tau, y)$ and $\text{var}(\omega_j \mid \tau, y)$ from Exercise 5.12 as functions of τ, bracketed by complete pooling at $\tau = 0$ and no pooling as $\tau \to \infty$. An inverse-gamma(ϵ, ϵ) prior will not do: by Section 5.7 its behavior as $\tau \to 0$ is arbitrarily sensitive to ϵ precisely when J is small.

Problem 5.11. Nonconjugate hierarchical models: suppose that in the rat tumor example (Section 5.3), where $y_j \mid \omega_j \sim \text{Bin}(n_j, \omega_j)$ independently for $j = 1, \ldots, J$, we wish to use a normal population distribution on the log-odds scale: $\text{logit}(\omega_j) \sim \text{N}(\mu, \tau^2)$, for $j = 1, \ldots, J$. As in Section 5.3, you will assign a noninformative prior distribution to the hyperparameters and perform a full Bayesian analysis.
(a) Write the joint posterior density, $p(\omega, \mu, \tau \mid y)$.
(b) Show that the integral (5.4), $p(\phi \mid y) = \int p(\omega, \phi \mid y)\, d\omega$ with $\phi = (\mu, \tau)$, has no closed-form expression.
(c) Why is expression (5.5), $p(\phi \mid y) = p(\omega, \phi \mid y)/p(\omega \mid \phi, y)$, no help for this problem?
In practice, we can solve this problem by normal approximation, importance sampling, and Markov chain simulation, as described in Part III.

Solution. (a) Take the hyperprior $p(\mu, \tau) \propto 1$, which is the uniform-on-τ choice of Section 5.4 (uniform on $\log \tau$ would be improper by Exercise 5.10(a)). The population density for ω_j carries the Jacobian $d\,\text{logit}(\omega_j)/d\omega_j = [\omega_j(1-\omega_j)]^{-1}$, so

$$p(\omega, \mu, \tau \mid y) \;\propto\; \tau^{-J} \prod_{j=1}^J \omega_j^{y_j-1}(1-\omega_j)^{n_j-y_j-1} \times \exp\left(-\frac{\big(\text{logit}(\omega_j) - \mu\big)^2}{2\tau^2}\right),$$

on $\omega \in (0, 1)^J$, $\mu \in \mathbb{R}$, $\tau > 0$.
(b) The integral factors over j, and each factor is a logistic-normal integral. Substituting $\theta_j = \text{logit}(\omega_j)$, so that $\omega_j^{y_j-1}(1-\omega_j)^{n_j-y_j-1}\, d\omega_j = e^{y_j\theta_j}(1+e^{\theta_j})^{-n_j}\, d\theta_j$,

$$p(\mu,\tau \mid y) \;\propto\; \tau^{-J}\prod_{j=1}^{J} I_j(\mu,\tau),$$

$$I_j(\mu,\tau) = \int_{-\infty}^{\infty} \frac{e^{y_j\theta}}{(1+e^{\theta})^{n_j}} \exp\!\left(-\frac{(\theta-\mu)^2}{2\tau^2}\right) d\theta.$$

What made (5.7)–(5.8) work in Section 5.3 was that the beta kernel $\omega^{\alpha-1}(1-\omega)^{\beta-1}$ has exactly the algebraic form of the binomial likelihood, so the integral was a beta function. Here the kernels are of incompatible type: I_j convolves a Gaussian with the binomial-logit likelihood, and already its smallest instance $n_j = 1$, $y_j = 0$ is the logistic-normal integral $\int (1+e^{\theta})^{-1}\mathrm{N}(\theta \mid \mu,\tau^2)\,d\theta$, for which no expression in elementary or standard special functions is known — which is why the logistic link is replaced by a probit, or approximated by $\mathrm{logit}^{-1}(\theta) \approx \Phi(\theta/1.702)$, whenever a closed form is wanted. Expanding $(1+e^{\theta})^{-n_j}$ in a series and integrating term by term yields only an infinite series in Φ. (Nonexistence in the Liouville sense is not proved here; the claim asserted is the book's, that no closed form is available.)
(c) Identity (5.5) reads $p(\phi \mid y) = p(\omega,\phi \mid y)/p(\omega \mid \phi, y)$, and it is useful only when the denominator is known in normalized form. From (a), the conditional posterior factors as

$$p(\omega_j \mid \mu,\tau,y) = \frac{\omega_j^{\,y_j-1}(1-\omega_j)^{\,n_j-y_j-1}\exp\!\big(-(\mathrm{logit}\,\omega_j-\mu)^2/(2\tau^2)\big)}{I_j(\mu,\tau)},$$

whose normalizing constant is precisely $I_j(\mu,\tau)$ — the very integral of part (b). So (5.5) merely relocates the unknown constant from numerator to denominator; it is circular here, whereas in Section 5.3 the denominator was the known beta density (5.7).

Problem 5.12. Conditional posterior means and variances: derive analytic expressions for $\mathrm{E}(\omega_j \mid \tau, y)$ and $\mathrm{var}(\omega_j \mid \tau, y)$ in the hierarchical normal model of Section 5.4 (and used in Figures 5.6 and 5.7). Recall from (5.17) that

$$\omega_j \mid \mu,\tau,y \sim \mathrm{N}(\hat\omega_j, V_j),$$
$$\hat\omega_j = \frac{\frac{1}{\sigma_j^2}\bar y_{\cdot j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}}, \qquad V_j = \frac{1}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}},$$

and from (5.20) that $\mu \mid \tau, y \sim \mathrm{N}(\hat\mu, V_\mu)$ with

$$\hat\mu = \frac{\sum_j \frac{\bar y_{\cdot j}}{\sigma_j^2+\tau^2}}{\sum_j \frac{1}{\sigma_j^2+\tau^2}}, \qquad V_\mu^{-1} = \sum_{j=1}^{J}\frac{1}{\sigma_j^2+\tau^2}.$$

(Hint: use (2.7) and (2.8), averaging over μ.)

Solution. Substitute $\hat\mu$ for μ in the mean, and add the shrinkage-weighted variance of $\hat\mu$:

$$\mathrm{E}(\omega_j \mid \tau, y) = \frac{\frac{1}{\sigma_j^2}\bar y_{\cdot j} + \frac{1}{\tau^2}\hat\mu}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}},$$
$$\mathrm{var}(\omega_j \mid \tau, y) = V_j + B_j^2\, V_\mu,$$

where $B_j = \dfrac{1/\tau^2}{1/\sigma_j^2+1/\tau^2} = \dfrac{\sigma_j^2}{\sigma_j^2+\tau^2}$ is the shrinkage factor.
For the mean, condition on μ and use (2.7) with everything also conditioned on (τ, y):

$$\mathrm{E}(\omega_j \mid \tau, y) = \mathrm{E}\big[\mathrm{E}(\omega_j \mid \mu,\tau,y) \mid \tau, y\big] = \mathrm{E}(\hat\omega_j \mid \tau, y),$$

and $\hat\omega_j = (1-B_j)\bar y_{\cdot j} + B_j\mu$ is affine in μ, so the expectation is obtained by replacing μ with $\mathrm{E}(\mu \mid \tau, y) = \hat\mu$.

For the variance, (2.8) in the same conditional form gives

$$\begin{aligned}\operatorname{var}(\omega_j \mid \tau, y) &= \mathrm{E}\big[\operatorname{var}(\omega_j \mid \mu, \tau, y) \mid \tau, y\big] \\ &\quad + \operatorname{var}\big[\mathrm{E}(\omega_j \mid \mu, \tau, y) \mid \tau, y\big] \\ &= \mathrm{E}(V_j \mid \tau, y) + \operatorname{var}(\hat{\omega}_j \mid \tau, y) \\ &= V_j + B_j^2 \operatorname{var}(\mu \mid \tau, y) = V_j + B_j^2 V_\mu,\end{aligned}$$

since V_j is a function of (σ_j, τ) only, hence constant given (τ, y), and $\hat{\omega}_j$ is affine in μ with slope B_j. Written out in terms of σ_j and τ alone,

$$\begin{aligned}\mathrm{E}(\omega_j \mid \tau, y) &= \frac{\tau^2}{\sigma_j^2 + \tau^2}\,\bar{y}_{\cdot j} + \frac{\sigma_j^2}{\sigma_j^2 + \tau^2}\,\hat{\mu}, \\ \operatorname{var}(\omega_j \mid \tau, y) &= \frac{\sigma_j^2 \tau^2}{\sigma_j^2 + \tau^2} + \left(\frac{\sigma_j^2}{\sigma_j^2 + \tau^2}\right)^2 \left(\sum_{k=1}^{J} \frac{1}{\sigma_k^2 + \tau^2}\right)^{-1}.\end{aligned}$$

Problem 5.13. Hierarchical binomial model: Exercise 3.8 described a survey of bicycle traffic in Berkeley, California, with data displayed in Table 3.3. For this problem, restrict your attention to the first two rows of the table: residential streets labeled as 'bike routes,' which we will use to illustrate this computational exercise. The relevant counts (bicycles y_j and other vehicles, in one hour, in each of 10 residential blocks with a bike route) are

j	1	2	3	4	5	6	7	8	9	10
bicycles y_j	16	9	10	13	19	20	18	17	35	55
other vehicles	58	90	48	57	103	57	86	112	273	64
total n_j	74	99	58	70	122	77	104	129	308	119

(a) Set up a model for the data in Table 3.3 so that, for $j = 1, \ldots, 10$, the observed number of bicycles at location j is binomial with unknown probability ω_j and sample size equal to the total number of vehicles (bicycles included) in that block. The parameter ω_j can be interpreted as the underlying or 'true' proportion of traffic at location j that is bicycles. (See Exercise 3.8.) Assign a beta population distribution for the parameters ω_j and a noninformative hyperprior distribution as in the rat tumor example of Section 5.3. Write down the joint posterior distribution.
(b) Compute the marginal posterior density of the hyperparameters and draw simulations from the joint posterior distribution of the parameters and hyperparameters, as in Section 5.3.
(c) Compare the posterior distributions of the parameters ω_j to the raw proportions, (number of bicycles / total number of vehicles) in location j. How do the inferences from the posterior distribution differ from the raw proportions?
(d) Give a 95% posterior interval for the average underlying proportion of traffic that is bicycles.
(e) A new city block is sampled at random and is a residential street with a bike route. In an hour of observation, 100 vehicles of all kinds go by. Give a 95% posterior interval for the number of those vehicles that are bicycles. Discuss how much you trust this interval in application.
(f) Was the beta distribution for the ω_j's reasonable?

Solution. (a) The rat-tumor model of Section 5.3 verbatim, with (y_j, n_j) from the table above:

$$\begin{aligned}y_j \mid \omega_j &\sim \mathrm{Bin}(n_j, \omega_j), \quad j = 1, \ldots, 10, \text{ independent}, \\ \omega_j \mid \alpha, \beta &\sim \mathrm{Beta}(\alpha, \beta), \\ p(\alpha, \beta) &\propto (\alpha + \beta)^{-5/2} \quad (5.9),\end{aligned}$$

so that, by (5.6),

$$p(\omega, \alpha, \beta \mid y) \;\propto\; (\alpha+\beta)^{-5/2} \prod_{j=1}^{10} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \omega_j^{\alpha-1}(1-\omega_j)^{\beta-1} \times \omega_j^{y_j}(1-\omega_j)^{n_j-y_j}.$$

(b) Integrating out ω gives (5.8),

$$p(\alpha, \beta \mid y) \;\propto\; (\alpha+\beta)^{-5/2} \prod_{j=1}^{10} \frac{B(\alpha+y_j,\ \beta+n_j-y_j)}{B(\alpha,\beta)},$$

and on the coordinates $(u, v) = (\log(\alpha/\beta), \log(\alpha+\beta))$ the density picks up the Jacobian $\alpha\beta$, giving (5.10) in the prior factor. Evaluating the log density on a 400×400 grid over $u \in [-3.0, -0.2]$, $v \in [0.5, 7.0]$ (grid-edge mass below 10^{-4} of the total), the contours are the familiar banana, elongated in v, with posterior mode at $(u, v) = (-1.37, 2.89)$, i.e.

$$(\hat\alpha, \hat\beta) = (3.66,\ 14.42).$$

Drawing 10{,}000 pairs (α, β) from the grid (with jitter) and then $\omega_j \mid \alpha, \beta, y \sim \text{Beta}(\alpha+y_j,\ \beta+n_j-y_j)$ by (5.7) gives posterior medians and 95% central intervals

$$\alpha:\ 3.19\ [1.04,\ 8.77], \qquad \beta:\ 12.5\ [3.46,\ 36.1],$$

with $\alpha+\beta$ having median 15.7 and interval $[4.63, 44.6]$. Repeating with independent seeds moves these hyperparameter quantiles by at most 0.3 and every quantity on the probability scale below by at most 0.001.

(c) The posterior means shrink the raw proportions toward the population mean ≈ 0.21, and the shrinkage is strongest where n_j is smallest:

j	1	2	3	4	5	6	7	8	9	10
raw y_j/n_j	.216	.091	.172	.186	.156	.260	.173	.132	.114	.462
$\text{E}(\omega_j \mid y)$	.214	.107	.179	.189	.162	.249	.177	.140	.118	.429
2.5%	.136	.056	.100	.113	.105	.167	.114	.088	.085	.341
97.5%	.305	.171	.274	.278	.227	.344	.250	.201	.156	.520

The shifts are modest — n_j runs from 58 to 308, so each ω_j is already well determined by its own block — but systematic: block 2 (raw .091, $n = 99$) moves up to .107, block 10 (raw .462, $n = 119$) moves down to .429, while block 9 ($n = 308$) barely moves at all. The intervals are also slightly wider than separate-block binomial intervals, because α and β are themselves uncertain.

(d) The average underlying proportion is the population mean $\alpha/(\alpha+\beta)$, whose posterior median is 0.205 with 95% central interval

$$\Pr\left(0.145 \le \tfrac{\alpha}{\alpha+\beta} \le 0.293 \;\middle|\; y\right) = 0.95.$$

(e) Draw $\tilde\omega \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$ for a genuinely new block and then $\tilde y \mid \tilde\omega \sim \text{Bin}(100, \tilde\omega)$. The posterior predictive interval is

$$\Pr(3 \le \tilde y \le 50 \mid y) = 0.95, \qquad \text{E}(\tilde y \mid y) = 20.9,$$

with median 19 (the upper endpoint carries Monte Carlo error of about ± 1 at 10{,}000 draws; the lower endpoint is stable at 3). It is far wider than the interval for any single observed ω_j because it must absorb the between-block variation as well: the fitted population distribution has posterior median standard deviation 0.098 against a mean of 0.205, so blocks genuinely differ a great deal. How much to trust it: the arithmetic is right conditional on the model, but the model's exchangeability assumption is the weak point. The ten blocks were not a simple random sample of residential bike-route blocks in any well-defined population, they were observed for a single hour on particular days, and traffic composition depends on time of day, weather and the specific street — none of which is in the model. Furthermore $\tilde y$ is conditioned on exactly 100 vehicles passing, which is itself informative about the kind of block (compare Exercise 5.14, where the totals are modeled rather than conditioned on). I would treat $[3, 50]$ as a lower bound on the real uncertainty.

(f) Yes. Posterior predictive replication of all ten blocks from the fitted population distribution reproduces the observed spread well: with T ranging over the extremes and the dispersion of the raw proportions,

$T(y)$	max	min	sd	range
observed	0.462	0.091	0.105	0.371
predictive mean	0.401	0.067	0.108	0.334
$\Pr(T(y^{\text{rep}}) \geq T(y) \mid y)$	0.26	0.28	0.45	0.32

No p-value is near 0 or 1, so the two-parameter beta accommodates even block 10's outlying 0.462 without strain — though only by being genuinely dispersed, which is what makes (e) so wide. With $J = 10$ there is in any case little power to distinguish a beta from other unimodal population distributions on $(0, 1)$.

Problem 5.14. Hierarchical Poisson model: consider the dataset in the previous problem, but suppose only the total amount of traffic at each location is observed. That is, for the ten residential blocks with a bike route in Table 3.3, we observe only

j	1	2	3	4	5	6	7	8	9	10
total vehicles n_j	74	99	58	70	122	77	104	129	308	119

(a) Set up a model in which the total number of vehicles observed at each location j follows a Poisson distribution with parameter ω_j, the 'true' rate of traffic per hour at that location. Assign a gamma population distribution for the parameters ω_j and a noninformative hyperprior distribution. Write down the joint posterior distribution.
(b) Compute the marginal posterior density of the hyperparameters and plot its contours. Simulate random draws from the posterior distribution of the hyperparameters and make a scatterplot of the simulation draws.
(c) Is the posterior density integrable? Answer analytically by examining the joint posterior density at the limits or empirically by examining the plots of the marginal posterior density above.
(d) If the posterior density is not integrable, alter it and repeat the previous two steps.
(e) Draw samples from the joint posterior distribution of the parameters and hyperparameters, by analogy to the method used in the hierarchical binomial model.

Solution. (a) With Gamma(α, β) in the rate parametrization (density $\propto \omega^{\alpha-1}e^{-\beta\omega}$), and the noninformative hyperprior $p(\alpha, \beta) \propto (\alpha\beta)^{-1}$, i.e. uniform on $(\log\alpha, \log\beta)$:

$$p(\omega, \alpha, \beta \mid n) \;\propto\; \frac{1}{\alpha\beta} \prod_{j=1}^{10} \frac{\beta^\alpha}{\Gamma(\alpha)}\, \omega_j^{\alpha-1} e^{-\beta\omega_j}\; \frac{\omega_j^{n_j} e^{-\omega_j}}{n_j!}.$$

(b) The ω_j integrate out to independent negative binomials,

$$p(n_j \mid \alpha, \beta) = \frac{\Gamma(\alpha + n_j)}{\Gamma(\alpha)\, n_j!} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^{n_j},$$

so that

$$p(\alpha, \beta \mid n) \;\propto\; \frac{1}{\alpha\beta} \prod_{j=1}^{10} p(n_j \mid \alpha, \beta),$$

which on the natural coordinates $(\log(\alpha/\beta), \log\alpha) = (\log\mu, \log\alpha)$ — $\mu = \alpha/\beta$ being the population mean rate — is flat in the prior factor. Evaluated on a grid the contours are closed and concentrated around $\log\mu \approx 4.75$ ($\mu \approx 116$, the observed mean of the n_j) and $\log\alpha \approx 1.2$; nothing visible happens at the edges.
(c) Not integrable. Take $\alpha \to \infty$ with $\mu = \alpha/\beta$ held fixed. The population distribution Gamma$(\alpha, \alpha/\mu)$ has mean μ and variance $\mu^2/\alpha \to 0$, so it degenerates to a point mass at μ and

$$\prod_{j=1}^{10} p(n_j \mid \alpha,\beta) \;\longrightarrow\; c(\mu) := \prod_{j=1}^{10} \frac{\mu^{n_j} e^{-\mu}}{n_j!} \;>\; 0.$$

Since the prior is uniform in $\log\alpha$ (the map $(\log\alpha,\log\beta)\mapsto(\log\mu,\log\alpha)$ has unit Jacobian), integrating the ray gives $\int_A^\infty c(\mu)\,d\log\alpha = \infty$ for every fixed μ. This is exactly the failure of Exercise 5.9(a): a uniform prior on the log of the population dispersion leaves an infinite ridge as the dispersion goes to zero. The empirical check of (b) misses it, and must: the ridge height is $\max_\mu c(\mu) = c(116) = e^{-188.6} \approx 1.2\times10^{-82}$, eighty orders of magnitude below the mode, because the data ($\bar n = 116$, sample variance 5142) are wildly overdispersed relative to Poisson.

(d) Follow the recipe of Exercise 5.9(b): put a uniform prior on the dispersion itself rather than its logarithm. The scale-free dispersion of Gamma(α,β) is its coefficient of variation $\alpha^{-1/2}$ (playing the role of $(\alpha+\beta)^{-1/2}$ there), so take

$$p\Big(\log\Big(\tfrac{\alpha}{\beta}\Big),\,\alpha^{-1/2}\Big) \propto 1, \qquad \text{i.e.} \qquad p(\alpha,\beta) \propto \alpha^{-3/2}\beta^{-1},$$

the Jacobian being $\big|\partial(\log(\alpha/\beta),\alpha^{-1/2})/\partial(\alpha,\beta)\big| = \tfrac12\alpha^{-3/2}\beta^{-1}$. The offending direction $\alpha\to\infty$ now occupies the finite interval $\alpha^{-1/2}\in(0,\varepsilon)$, so it contributes at most ε times a bounded integrand. For the opposite corner, $\alpha\to0$ and $\beta\to0$, expand each factor: $\Gamma(\alpha+n)/\Gamma(\alpha)\to\alpha\,(n-1)!$ and $(\beta/(1+\beta))^\alpha \le \beta^\alpha$, so

$$\prod_j p(n_j\mid\alpha,\beta) \;\le\; C\,\alpha^J\beta^{\alpha J} \qquad (\alpha\le1,\ \beta\le1),$$

and with $J=10$,

$$\int_0^1\!\!\int_0^1 \alpha^{-3/2}\beta^{-1}C\alpha^J\beta^{\alpha J}\,d\beta\,d\alpha = \frac{C}{J}\int_0^1 \alpha^{J-5/2}\,d\alpha < \infty,$$

finite whenever $J>3/2$. The remaining directions die fast: as $\mu\to\infty$ at fixed α each factor is $O(\mu^{-\alpha})$, and as $\mu\to0$ it is $O(\mu^{n_j})$. So the posterior is proper. (A uniform prior on the population standard deviation $\sqrt{\alpha}/\beta$ rather than on the coefficient of variation would not work: its Jacobian gives $p(\alpha,\beta)\propto\alpha^{-1/2}\beta^{-2}$, and the β integral above then diverges for $\alpha\le1/J$.)

Recomputing on a 400×400 grid over $\log(\alpha/\beta)\in[4.0,5.6]$ and $\alpha^{-1/2}\in(0,3]$ (edge mass below 10^{-4}), the contours are closed with posterior mode at

$$\big(\log(\alpha/\beta),\ \alpha^{-1/2}\big) = (4.754,\ 0.466), \qquad (\hat\alpha,\hat\beta) = (4.60,\ 0.0396).$$

From 10,000 grid draws, posterior medians and 95% central intervals:

quantity	2.5%	median	97.5%
α	1.16	3.45	8.17
β	0.0087	0.0297	0.0727
$\mu=\alpha/\beta$	82.4	117.1	174.3
$\sqrt{\alpha}/\beta$	37.9	62.3	133.0

Across independent seeds these entries move by less than 0.1 on the α scale and less than 2 on the μ and $\sqrt{\alpha}/\beta$ scales. The population standard deviation of the rates, median 62 against a mean of 117, dwarfs the Poisson noise $\sqrt{116}\approx11$; the totals are far from a common Poisson rate, which is the whole point of the hierarchy.

(e) Exactly as in Section 5.3: having drawn (α,β) from the grid, draw each ω_j from its conditional posterior, which is gamma by conjugacy,

$$\omega_j \mid \alpha,\beta,n \;\sim\; \text{Gamma}(\alpha+n_j,\ \beta+1),$$

independently across j. The resulting posterior medians and 95% intervals are

j	1	2	3	4	5	6	7	8	9	10
n_j	74	99	58	70	122	77	104	129	308	119
median ω_j	74.9	99.1	59.6	71.4	121.7	77.7	103.9	128.1	302.0	118.5
2.5%	59.7	81.1	45.8	56.1	101.2	61.5	85.7	107.3	269.2	98.3
97.5%	92.8	118.9	76.1	89.3	143.8	96.1	124.8	151.1	337.5	141.4

Shrinkage toward μ is almost nil — the posterior mean of ω_j is $(\alpha+n_j)/(\beta+1)$, and with $\hat{\beta} \approx 0.04$ and $\hat{\alpha} \approx 3.5$ against n_j of order 100, each block's own count dominates. That is the correct answer here: the population distribution is so dispersed that block j's rate is learned almost entirely from block j.

5.3 Exercises 5.15–5.17

Problem 5.15. Meta-analysis: perform the computations for the meta-analysis data of Table 5.4.
(a) Plot the posterior density of τ over an appropriate range that includes essentially all of the posterior density, analogous to Figure 5.5.
(b) Produce graphs analogous to Figures 5.6 and 5.7 to display how the posterior means and standard deviations of the θ_j's depend on τ.
(c) Produce a scatterplot of the crude effect estimates vs. the posterior median effect estimates of the 22 studies. Verify that the studies with smallest sample sizes are partially pooled the most toward the mean.
(d) Draw simulations from the posterior distribution of a new treatment effect, $\tilde{\theta}_j$. Plot a histogram of the simulations.
(e) Given the simulations just obtained, draw simulated outcomes from replications of a hypothetical new experiment with 100 persons in each of the treated and control groups. Plot a histogram of the simulations of the crude estimated treatment effect (5.23) in the new experiment.
Table 5.4 (raw data and derived summaries): results of 22 clinical trials of beta-blockers for reducing mortality after myocardial infarction. Entries are deaths/total; y_j is the empirical log-odds ratio (5.23) and σ_j the approximate sampling standard deviation, the square root of (5.24). Negative effects correspond to reduced probability of death under the treatment.

j	Control	Treated	y_j	σ_j
1	3/39	3/38	0.028	0.850
2	14/116	7/114	-0.741	0.483
3	11/93	5/69	-0.541	0.565
4	127/1520	102/1533	-0.246	0.138
5	27/365	28/355	0.069	0.281
6	6/52	4/59	-0.584	0.676
7	152/939	98/945	-0.512	0.139
8	48/471	60/632	-0.079	0.204
9	37/282	25/278	-0.424	0.274
10	188/1921	138/1916	-0.335	0.117
11	52/583	64/873	-0.213	0.195
12	47/266	45/263	-0.039	0.229
13	16/293	9/291	-0.593	0.425
14	45/883	57/858	0.282	0.205
15	31/147	25/154	-0.321	0.298
16	38/213	33/207	-0.135	0.261
17	12/122	28/251	0.141	0.364
18	6/154	8/151	0.322	0.553
19	3/134	6/174	0.444	0.717
20	40/218	32/209	-0.218	0.260
21	43/364	27/391	-0.591	0.257
22	39/674	22/680	-0.608	0.272

Here (5.23) is the empirical-logit estimate

$$y_j = \log\left(\frac{y_{1j}}{n_{1j} - y_{1j}}\right) - \log\left(\frac{y_{0j}}{n_{0j} - y_{0j}}\right),$$

and (5.24) its approximate sampling variance

$$\sigma_j^2 = \frac{1}{y_{1j}} + \frac{1}{n_{1j} - y_{1j}} + \frac{1}{y_{0j}} + \frac{1}{n_{0j} - y_{0j}}.$$

Solution. Everything is the eight-schools computation of Section 5.5 run on (y_j, σ_j), $J = 22$: the model is

$$\begin{aligned} y_j \mid \theta_j &\sim \mathrm{N}(\theta_j, \sigma_j^2), \\ \theta_j \mid \mu, \tau &\sim \mathrm{N}(\mu, \tau^2), \\ p(\mu, \tau) &\propto 1, \end{aligned}$$

with σ_j treated as known (justified in Section 5.6: every arm but a handful has $n > 50$).
(a) By (5.21), with $\hat\mu(\tau)$ and $V_\mu(\tau)$ from (5.20),

$$\begin{aligned} V_\mu(\tau)^{-1} &= \sum_j \frac{1}{\sigma_j^2 + \tau^2}, \\ \hat\mu(\tau) &= V_\mu(\tau) \sum_j \frac{y_j}{\sigma_j^2 + \tau^2}, \\ p(\tau \mid y) &\propto V_\mu(\tau)^{1/2} \prod_{j=1}^{22} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(y_j - \hat\mu(\tau))^2}{2(\sigma_j^2 + \tau^2)}\right). \end{aligned}$$

Evaluated on a grid of $\tau \in (0, 1)$ this density is unimodal with mode at $\tau = 0.12$ and is negligible beyond $\tau = 0.6$; $\tau = 0$ remains entirely plausible, the density there being 0.68 of the modal value (Section 5.6 describes it as about 25% lower). Quantiles, agreeing with Table 5.5 except in the extreme lower tail (Table 5.5 prints 0.02 for the 2.5% point; the grid gives 0.008):

quantile	2.5%	25%	50%	75%	97.5%
τ	0.01	0.07	0.13	0.18	0.31

(figure `bda3-ch05-meta-tau-posterior`).
(b) Take $\hat\theta_j$ and $V_j = (\sigma_j^{-2} + \tau^{-2})^{-1}$ from (5.17) and average over $\mu \mid \tau, y \sim \mathrm{N}(\hat\mu(\tau), V_\mu(\tau))$, using $\mathrm{var}(\theta_j \mid \tau, y) = \mathrm{E}\,\mathrm{var}(\theta_j \mid \mu, \tau, y) + \mathrm{var}\,\mathrm{E}(\theta_j \mid \mu, \tau, y)$:

$$\begin{aligned} \mathrm{E}(\theta_j \mid \tau, y) &= V_j \left(\frac{y_j}{\sigma_j^2} + \frac{\hat\mu(\tau)}{\tau^2}\right), \\ \mathrm{sd}(\theta_j \mid \tau, y)^2 &= V_j + \left(\frac{V_j}{\tau^2}\right)^2 V_\mu(\tau), \end{aligned}$$

At $\tau = 0$ all 22 curves start at the pooled estimate -0.26 with common sd 0.05, and they fan out as τ grows, each $\mathrm{E}(\theta_j \mid \tau, y) \to y_j$ and $\mathrm{sd}(\theta_j \mid \tau, y) \to \sigma_j$. Five representative studies:

j	σ_j	y_j	E at 0.1	sd at 0.1	E at 0.2	sd at 0.2	E at 0.4	sd at 0.4
1	0.850	0.028	-0.246	0.115	-0.227	0.207	-0.188	0.373
7	0.139	-0.512	-0.340	0.090	-0.424	0.116	-0.483	0.132
10	0.117	-0.335	-0.286	0.083	-0.311	0.103	-0.327	0.113
14	0.205	0.282	-0.148	0.101	0.013	0.148	0.173	0.184
18	0.553	0.322	-0.232	0.113	-0.176	0.199	-0.044	0.332

(figures `bda3-ch05-meta-cond-mean`, `bda3-ch05-meta-cond-sd`). The imprecise studies 1 and 18 are still pinned near the mean at $\tau = 0.4$, while the precise studies 7, 10, 14 have all but reached their own y_j by $\tau = 0.2$.
(c) Drawing τ from the grid, then $\mu \mid \tau, y$, then $\theta_j \mid \mu, \tau, y$ (20{,}000 draws, Monte Carlo standard error below 0.006 on every quantile below) reproduces the posterior quantiles printed in Table 5.4.

The posterior medians span only $[-0.35, -0.12]$ while the y_j span $[-0.74, 0.44]$, so the pooling is severe throughout. The amount of pooling is the shrinkage factor

$$B_j = \frac{\sigma_j^2}{\sigma_j^2 + \tau^2} \qquad \text{(the weight on } \mu \text{ in } \mathrm{E}(\theta_j \mid \mu, \tau, y)),$$

evaluated here at the posterior median $\tau = 0.13$; it is increasing in σ_j, and σ_j falls roughly like $n_j^{-1/2}$, which is the assertion to be verified (the ordering by n_j is not exact, since σ_j depends on the event rates as well as on n_j: study 18 has $n = 305$ but a larger σ_j than study 2 with $n = 230$):

j	total n	σ_j	y_j	median θ_j	B_j
1	77	0.850	0.028	-0.246	0.977
6	111	0.676	-0.584	-0.261	0.964
3	162	0.565	-0.541	-0.263	0.950
2	230	0.483	-0.741	-0.277	0.932
18	305	0.553	0.322	-0.227	0.948
14	1741	0.205	0.282	-0.124	0.714
7	1884	0.139	-0.512	-0.349	0.532
4	3053	0.138	-0.246	-0.250	0.530
10	3837	0.117	-0.335	-0.288	0.448

The four smallest trials carry y_j ranging over 0.8 on the log-odds scale and yet have posterior medians within 0.04 of each other and of μ: they are pooled 93%-98% of the way to the mean, while the two largest retain about half their own signal. In the scatterplot (`bda3-ch05-meta-shrinkage`, point size proportional to $\sqrt{n_j}$) the small points sit on the horizontal line $\theta = \text{median}(\mu \mid y) = -0.25$ and the large points lie nearest the 45-degree line.

(d) $\tilde{\theta}_j \mid \mu, \tau \sim \mathrm{N}(\mu, \tau^2)$, the new study being assumed exchangeable with the 22 observed ones, drawn alongside each posterior draw gives a symmetric, slightly long-tailed histogram (`bda3-ch05-meta-predictive`) centred at -0.25 with sd 0.17:

quantile	2.5%	25%	50%	75%	97.5%
$\tilde{\theta}_j$	-0.60	-0.33	-0.25	-0.17	0.12

matching Table 5.5, and $\Pr(\tilde{\theta}_j > 0 \mid y) = 0.064$. (Section 5.6 reports this last probability as just over 10%, which is inconsistent with its own Table 5.5: a median of -0.25 and a 97.5% point of 0.11 force a value near 0.06.)

(e) The control rate must be supplied; take the pooled control mortality of the 22 trials, $p_0 = 0.100$, so $\text{logit}(p_0) = -2.197$ and, for each draw,

$$p_1 = \text{logit}^{-1}\Big(\text{logit}(p_0) + \tilde{\theta}_j\Big), \qquad \tilde{y}_0 \sim \text{Bin}(100, p_0), \quad \tilde{y}_1 \sim \text{Bin}(100, p_1),$$

and $\tilde{y}$ is then (5.23) applied to $\tilde{y}_0, \tilde{y}_1$ (with 0.5 added to each of the four counts, as recommended in Section 5.6, since $n = 100$ at a 10% rate produces empty cells in 0.05% of replications). The histogram (`bda3-ch05-meta-newexp`) has mean -0.25 and sd 0.54:

quantile	2.5%	50%	97.5%
$\tilde{y}$	-1.36	-0.23	0.80

and $\Pr(\tilde{y} < 0) = 0.64$ against $\Pr(\tilde{\theta}_j < 0) = 0.94$. Decomposing that sd by conditioning on $\tilde{\theta}_j$, binomial sampling noise contributes 0.51 and the genuine spread of the effect only 0.19, and $\sqrt{0.51^2 + 0.19^2} = 0.54$.

Problem 5.16. Equivalent data: Suppose we wish to apply the inferences from the meta-analysis example in Section 5.6 to data on a new study with equal numbers of people in the control and treatment groups. How large would the study have to be so that the prior and data were weighted equally in the posterior inference for that study?

Solution. About 850 people in each group, that is 1700 in total.

The new study's own data supply $y \mid \theta \sim \mathrm{N}(\theta, \sigma^2)$ and the meta-analysis supplies the prior $p(\theta \mid y_{1:22})$,

which by Exercise 5.15(d) has standard deviation $s = 0.17$ (Table 5.5: $\tilde{\theta}_j$ has 95% interval $[-0.58, 0.11]$). Since $\mathrm{E}(\theta \mid y)$ weights y and the prior mean by σ^{-2} and s^{-2}, equal weighting is exactly

$$\sigma^2 = s^2 = 0.17^2 = 0.0289.$$

With n in each arm and death probabilities $p_0 \approx p_1 \approx p$, (5.24) gives

$$\sigma^2 = \frac{1}{np} + \frac{1}{n(1-p)} + \frac{1}{np} + \frac{1}{n(1-p)} = \frac{2}{n\,p(1-p)},$$

so $n = 2/\big(p(1-p)s^2\big)$. The 22 trials pool to a mortality rate of $p = 0.089$ (1811 deaths in 20{,}290 patients), whence $2/(p(1-p)) = 24.6$ and

$$n = \frac{24.6}{0.0289} = 851 \quad \text{per group.}$$

(Check: $\sigma = \sqrt{24.6/851} = 0.170$.) Only trials 4, 7, 10 and 14 of Table 5.4 are this large.
Method (2): if one takes the prior for a new study to be $\mathrm{N}(\mu, \tau^2)$ with μ and τ fixed at their posterior medians, ignoring hyperparameter uncertainty, the requirement becomes $\sigma = \tau = 0.13$ and $n = 24.6/0.0169 = 1455$ per group.

Problem 5.17. Informative prior distributions: Continuing the example from Exercise 2.22, consider a (hypothetical) study of a simple training program for basketball free-throw shooting. A random sample of 100 college students is recruited into the study. Each student first shoots 100 free-throws to establish a baseline success probability. Each student then takes 50 practice shots each day for a month. At the end of that time, he or she takes 100 shots for a final measurement.
Let θ_i be the improvement in success probability for person i. For simplicity, assume the θ_i's are normally distributed with mean μ and standard deviation σ.
Give three joint prior distributions for μ, σ:
(a) A noninformative prior distribution,
(b) A subjective prior distribution based on your best knowledge, and
(c) A weakly informative prior distribution.

Solution. (a) $p(\mu, \sigma) \propto 1$ on $\mu \in \mathbb{R}$, $\sigma > 0$: the uniform density on (μ, σ), the default of Sections 5.4 and 5.7 for the hierarchical normal model. The posterior is proper: the uniform density on σ has a finite integral near $\sigma = 0$, and its infinite mass at $\sigma \to \infty$ makes the posterior improper only for $J = 1$ or 2 groups (Section 5.7), whereas here $J = 100$. The tempting alternative $p(\mu, \log\sigma) \propto 1$, that is $p(\sigma) \propto \sigma^{-1}$, must be avoided, since it gives an improper posterior with an infinite spike at $\sigma = 0$ (Section 5.4).
(b) Independently,

$$\mu \sim \mathrm{N}(0.05,\, 0.04^2), \qquad \sigma \sim \mathrm{N}^{+}(0,\, 0.05^2),$$

the second being a half-normal. The mean: a month of 50 shots a day is 1500 practice shots, enough to move a typical non-varsity college student, who starts around 60%, up by a few percentage points, but the gain is bounded by the 40% of headroom remaining and some students will get worse, so a prior 95% interval $(-0.03, 0.13)$ for μ is honest. The scale: the person-to-person spread of a real training gain should be comparable to its mean, and it is in any case hard to learn about, since each θ_i is measured by a difference of two binomial proportions whose own standard deviation is $\sqrt{2(0.6)(0.4)/100} = 0.07$; the half-normal above has median 0.034 and 95th percentile 0.098, which covers that range and puts vanishing mass on values of σ the design could never resolve.
(c) Independently,

$$\mu \sim \mathrm{N}(0,\, 0.2^2), \qquad \sigma \sim \text{half-Cauchy}(0,\, 0.2),$$

the half-Cauchy family of Section 5.7 with scale A set large relative to the plausible range. These are proper and encode only what the measurement scale itself guarantees, that θ_i is a difference of two probabilities and so lies in $[-1, 1]$: the prior 95% interval for μ is $(-0.39, 0.39)$, essentially the whole

range physically available, and the prior for σ has median 0.2, so it neither favours $\sigma = 0$ nor, having a gentle tail rather than a half-normal's, prevents the likelihood from pulling σ upward if the data ask for it.

6 Model Checking

Contents

6.1 Exercises 6.1–6.7

Problem 6.1. Posterior predictive checking:
(a) On page 120, the data from the SAT coaching experiments were checked against the model that assumed identical effects in all eight schools: the expected order statistics of the effect sizes were $(26, 19, 14, 10, 6, 2, -3, -9)$, compared to observed data of $(28, 18, 12, 8, 7, 1, -1, -3)$. Express this comparison formally as a posterior predictive check comparing this model to the data. Does the model fit the aspect of the data tested here?
(b) Explain why, even though the identical-schools model fits under this test, it is still unacceptable for some practical purposes.
The data of Table 5.2 (observed effects of special preparation on SAT-V scores in eight randomized experiments) are:

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Solution. (a) Yes, it fits: the posterior predictive p-value for the largest observed effect is 0.39.
The identical-schools model is $y_j \mid \theta \sim \mathrm{N}(\theta, \sigma_j^2)$ independently, $j = 1, \ldots, 8$, with the σ_j of Table 5.2 known and $p(\theta) \propto 1$ (proper posterior, since $\sum_j \sigma_j^{-2} > 0$). By (2.11)-(2.12),

$$\theta \mid y \sim \mathrm{N}(\hat{\theta}, V), \quad \hat{\theta} = \frac{\sum_j y_j/\sigma_j^2}{\sum_j 1/\sigma_j^2} = 7.69, \quad V = \Big(\sum_j \sigma_j^{-2}\Big)^{-1} = 16.6,$$

the pooled estimate 7.7 with standard error 4.1 of page 119. Replications y^{rep} are drawn from (6.1), here $\int \prod_j \mathrm{N}(y_j^{\mathrm{rep}} \mid \theta, \sigma_j^2) p(\theta \mid y)\, d\theta$; the aspect of the data examined on page 120 is the order-statistic vector $T(y) = (y_{(1)} \geq \cdots \geq y_{(8)})$, with scalar summary $T_{\max}(y) = \max_j y_j = 28$ carrying the question actually asked there ("would it be possible to have one school's observed effect be 28 just by chance?"). With $S = 200{,}000$ draws $\theta^s \sim \mathrm{N}(7.69, 4.07^2)$, $y_j^{\mathrm{rep}\,s} \sim \mathrm{N}(\theta^s, \sigma_j^2)$:

Order k	1	2	3	4	5	6	7	8
$\mathrm{E}[y_{(k)}^{\mathrm{rep}}]$	26.2	18.1	13.4	9.5	5.9	2.0	-2.8	-10.8
sd	9.9	7.5	6.8	6.5	6.5	6.8	7.4	9.9
observed $y_{(k)}$	28	18	12	8	7	1	-1	-3

(the first row reproduces the book's $(26, 19, 14, 10, 6, 2, -3, -9)$, which was computed from the cruder reference distribution $y_j \sim \mathrm{N}(8, 13^2)$). Every observed order statistic lies well inside one predictive standard deviation of its expectation – their individual upper-tail probabilities $\Pr(y_{(k)}^{\mathrm{rep}} \geq y_{(k)} \mid y)$ run from 0.22 to 0.59 – and

$$p_B = \Pr\big(T_{\max}(y^{\mathrm{rep}}) \geq 28 \mid y\big) = 0.39.$$

Companion quantities agree: $\Pr(\min_j y_j^{\mathrm{rep}} \leq -3 \mid y) = 0.79$ and $\Pr(\mathrm{range}(y^{\mathrm{rep}}) \geq 31 \mid y) = 0.67$. (Monte Carlo standard error $\sqrt{p(1-p)/S} \leq 0.002$ throughout.) The identical-schools model fits this aspect of the data.
(b) Because the test has essentially no power against the alternatives that matter. The observed spread of the y_j is generated almost entirely by the known sampling errors $\sigma_j \in [9, 18]$, which swamp any plausible between-school variation. Simulating eight schools with $\theta_j \sim \mathrm{N}(8, \tau^2)$ and $y_j \mid \theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$ gives

τ	0	5	10	15
$\mathrm{E}[\max_j y_j]$	26.5	27.8	31.3	36.2
$\Pr(\max_j y_j \geq 28)$	0.38	0.44	0.59	0.74
$\mathrm{E}[\mathrm{sd}(y)]$	12.4	13.3	15.7	19.0

so the data would look almost the same under $\tau = 0$ and under $\tau = 10$; passing the check is no evidence for $\tau = 0$. But the practical questions all turn on τ: the pooled model forces $\theta_A = \theta_C$, hence asserts $\Pr(\theta_A > \theta_C \mid y) = 0$ against the hierarchical model's 0.70 (page 123), hands every school the identical interval $[-0.3, 15.7]$, and predicts a new school's effect with standard error 4.1 rather than $\sqrt{\tau^2 + 4.1^2}$. As page 120 notes, it implies "the probability is $\frac{1}{2}$ that the true effect in A is less than 7.7", which no one reading Table 5.2 believes.

Problem 6.2. Model checking: in Exercise 2.13, the counts of airline fatalities in 1976-1985 were fitted to four different Poisson models.
(a) For each of the models, set up posterior predictive test quantities to check the following assumptions: (1) independent Poisson distributions, (2) no trend over time.
(b) For each of the models, use simulations from the posterior predictive distributions to measure the discrepancies. Display the discrepancies graphically and give p-values.
(c) Do the results of the posterior predictive checks agree with your answers in Exercise 2.13(e)?
The data of Table 2.2 (worldwide airline fatalities, 1976-1985; death rate is passenger deaths per 100 million passenger miles) are:

Year	Fatal accidents	Passenger deaths	Death rate
1976	24	734	0.19
1977	25	516	0.12
1978	31	754	0.15
1979	31	877	0.16
1980	22	814	0.14
1981	21	362	0.06
1982	26	764	0.13
1983	20	809	0.13
1984	16	223	0.03
1985	22	1066	0.15

The four models of Exercise 2.13 are: (a) fatal accidents $y_t \sim \text{Poisson}(\theta)$; (b) fatal accidents $y_t \sim \text{Poisson}(\theta x_t)$ with x_t the passenger miles flown; (c) passenger deaths $y_t \sim \text{Poisson}(\theta)$; (d) passenger deaths $y_t \sim \text{Poisson}(\theta x_t)$.

Solution. (a) Two discrepancy measures, both functions of data and parameter. Writing $E_t = \theta$ (models (a), (c)) or $E_t = \theta x_t$ (models (b), (d)) for the fitted rate in year t, $t = 1, \ldots, 10$, and $\bar{t} = 5.5$, take

$$T_1(y, \theta) = \sum_{t=1}^{10} \frac{(y_t - E_t)^2}{E_t},$$
$$T_2(y, \theta) = \frac{\sum_t (t - \bar{t})\,(y_t - E_t)/\sqrt{E_t}}{\sum_t (t - \bar{t})^2}.$$

T_1 is the χ^2 discrepancy of (6.7); under independent Poisson sampling it has mean 10 regardless of θ, so it detects the overdispersion (or underdispersion) produced by any failure of the independent-Poisson assumption. T_2 is the least-squares slope of the standardized residuals on time, so its posterior predictive distribution is centred at 0 under the model and it detects a monotone drift in the rate.
The exposures, obtained as $x_t = (\text{deaths}_t)/(\text{death rate}_t)$ in units of 10^{11} passenger miles, are

Year	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
x_t (10^{11}mi)	3.863	4.300	5.027	5.481	5.814	6.033	5.877	6.223	7.433	7.107
accidents per x_t	6.21	5.81	6.17	5.66	3.78	3.48	4.42	3.21	2.15	3.10

(b) With the Jeffreys prior $p(\theta) \propto \theta^{-1/2}$ of Exercise 2.12, the posterior is $\theta \mid y \sim \text{Gamma}(\sum_t y_t + \frac{1}{2},\ \sum_t x_t)$, and $S = 40{,}000$ draws $(\theta^s, y^{\text{rep}\,s})$ give

Model	$\hat{\theta}$	$\mathrm{E}[T_1(y,\theta) \mid y]$	p_1	$\mathrm{E}[T_2(y,\theta) \mid y]$	p_2
(a) accidents, constant	23.85	9.4	0.50	-0.189	0.044
(b) accidents, exposure	4.17	27.0	0.004	-0.499	<0.001
(c) deaths, constant	691.9	829.7	<0.001	0.097	0.81
(d) deaths, exposure	121.1	1025.5	<0.001	-1.578	<0.001

Here $p_1 = \Pr(T_1(y^{\mathrm{rep}},\theta) \geq T_1(y,\theta) \mid y)$ and, since the substantive alternative is a declining rate, $p_2 = \Pr(T_2(y^{\mathrm{rep}},\theta) \leq T_2(y,\theta) \mid y)$; the Monte Carlo standard error is at most 0.003. (For model (b), $\hat{\theta} = 4.17$ accidents per 10^{11} miles; for (d), 121 deaths per 10^{11} miles.) The graphical display is the realized-versus-replicated scatterplot of Figure 6.4: plot the S pairs $(T(y,\theta^s), T(y^{\mathrm{rep}\,s},\theta^s))$ with the 45° line, one panel per model, T_1 on a $\log_{10}$ scale because for (c) and (d) the realized values lie two orders of magnitude to the right of the cloud.

Reading the table: the accident counts are consistent with a constant-rate Poisson process ($p_1 = 0.50$) but show a marginal downward drift ($p_2 = 0.044$); once exposure is included the drift becomes overwhelming ($p_2 < 0.001$) and drags the dispersion with it ($p_1 = 0.004$), because exposure grew by 84% while accidents fell. The death counts are grossly overdispersed under both models: $T_1 = 830$ and 1026 against a predictive mean of 10, a p-value indistinguishable from 0.

(c) Yes. Exercise 2.13(e) argues on general principles that the Poisson model is most plausible for fatal accidents, since accidents occur as rare independent events in a large exposure, and least plausible for passenger deaths, since deaths arrive in clumps – one accident kills from 0 to several hundred people at once, so deaths are a compound rather than a simple Poisson process. The checks confirm exactly this ordering: model (a) passes the dispersion check and the other three fail, with the two death models failing by a factor of 100. The checks add one thing the general argument did not: the accident rate per passenger mile is clearly declining over the decade, so no constant-rate model – with or without exposure – is adequate.

Problem 6.3. Model improvement:
(a) Use the solution to the previous problem and your substantive knowledge to construct an improved model for airline fatalities.
(b) Fit the new model to the airline fatality data.
(c) Use your new model to forecast the airline fatalities in 1986. How does this differ from the forecasts from the previous models?
(d) Check the new model using the same posterior predictive checks as you used in the previous models. Does the new model fit better?
(Data as in Exercise 6.2 / Table 2.2. In 1986 there were in fact 22 fatal accidents, 546 passenger deaths, and a death rate of 0.06 per 100 million miles; assume 8×10^{11} passenger miles were flown in 1986.)

Solution. (a) A Poisson log-linear trend in the rate per passenger mile,

$$y_t \mid \alpha,\beta \sim \mathrm{Poisson}\big(x_t \exp(\alpha + \beta(t - \bar{t}))\big),$$
$$t = 1,\ldots,10, \quad \bar{t} = 5.5,$$

with y_t the fatal accidents, x_t the exposure in 10^{11} passenger miles, and $p(\alpha,\beta) \propto 1$ (proper posterior: the log-likelihood is strictly concave in (α,β) with $\to -\infty$ in every direction, since the x_t are positive and the t are not all equal). This is the minimal repair of the two defects Exercise 6.2 exposed in model (b): exposure belongs in the model (traffic grew 84% over the decade), and the rate per mile is not constant – aviation safety improved steadily through the late 1970s and early 1980s, so the rate should decay smoothly rather than jump. A multiplicative trend keeps the rate positive and matches the form the substantive claim takes, "safety improves by a few percent a year". Passenger deaths are not modelled here: their overdispersion is compound, not a trend.

(b) The posterior was computed on a 401×401 grid over $(\alpha,\beta) \in [0.5, 2.5] \times [-0.35, 0.05]$ and 20,000 draws taken from it:

$$\alpha \mid y: \ \mathrm{E} = 1.432, \ \mathrm{sd} = 0.066,$$
$$\beta \mid y: \ \mathrm{E} = -0.1045, \ \mathrm{sd} = 0.0231,$$
$$95\% \text{ interval } [-0.150, -0.060].$$

So the baseline rate at mid-decade is $e^{1.432} = 4.19$ fatal accidents per 10^{11} passenger miles, falling by a factor $e^{\beta} = 0.901$ per year, 95% interval $[0.860, 0.942]$ – about a 10% annual improvement, with $\Pr(\beta < 0 \mid y) > 0.9999$.
(c) $\mathrm{E}[y_{1986} \mid y] = 17.1$, with 95% posterior predictive interval $[8, 28]$, obtained from $y_{1986} \sim \text{Poisson}(8.0 \cdot e^{\alpha + 6.5\beta})$. Compared with the models of Exercise 2.13:

Model	$\mathrm{E}[y_{1986} \mid y]$	95% interval
2.13(a) constant count	23.8	$[14, 34]$
2.13(b) constant rate × mileage	33.4	$[22, 46]$
trend model	17.1	$[8, 28]$

The exposure-only model (b) is the worst forecaster of the three: it projects the 1986 traffic increase forward while holding the rate fixed at its decade average, so it predicts 33 accidents. The trend model does the opposite, extrapolating the 10% annual decline one year past the data. The realized value was 22, for which the trend model gives $\Pr(y_{1986} \geq 22 \mid y) = 0.18$ – inside the interval, but in its upper tail.
(d) Yes, decisively. The same two discrepancies give

$$T_1: \ \mathrm{E}[T_1 \mid y] = 7.5 \ \text{(predictive mean 10.0)}, \quad p_1 = 0.68,$$
$$T_2: \ \mathrm{E}[T_2 \mid y] = 0.006, \quad p_2 = 0.52,$$

against $p_1 = 0.004$, $p_2 < 0.001$ for model (b). The trend has absorbed the systematic drift, and with it the excess dispersion that the drift had been creating: ten counts now scatter about their fitted rates exactly as independent Poisson variables should.

Problem 6.4. Model checking and sensitivity analysis: find a published Bayesian data analysis from the statistical literature.
(a) Compare the data to posterior predictive replications of the data.
(b) Perform a sensitivity analysis by computing posterior inferences under plausible alternative models.

Solution. The analysis chosen is Rubin (1981), "Estimation in parallel randomized experiments", Journal of Educational Statistics 6, 377-401 – the hierarchical normal analysis of the eight SAT-V coaching experiments reproduced in Section 5.5. The fitted model is

$$y_j \mid \theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2), \quad \theta_j \mid \mu, \tau \sim \mathrm{N}(\mu, \tau^2), \quad p(\mu, \tau) \propto 1,$$

with (y_j, σ_j) as in Table 5.2 of Exercise 6.1; the θ_j are taken exchangeable, which is what licenses the common population distribution and is defensible here because nothing beyond y_j distinguishes the eight programs (Section 5.5). The uniform prior on τ yields a proper posterior (Section 5.4). Draws are taken by the factorization of Section 5.4: a grid draw of τ from (5.21), then $\mu \mid \tau, y$ from (5.20), then $\theta_j \mid \mu, \tau, y$ from (5.17). Posterior median $\tau = 5.3$; $\mathrm{E}[\theta_A \mid y] = 11.4$ with sd 8.4.
(a) The model fits: no test quantity is in a tail. Replications $y_j^{\mathrm{rep}} \sim \mathrm{N}(\theta_j^s, \sigma_j^2)$ from 20,000 posterior draws give

T	$T(y)$	$\mathrm{E}[T(y^{\mathrm{rep}}) \mid y]$	p_B
$\max_j y_j$	28	28.6	0.48
$\min_j y_j$	-3	-12.6	0.17
$\mathrm{sd}(y)$	10.44	13.8	0.78

with $p_B = \Pr(T(y^{\mathrm{rep}}) \geq T(y) \mid y)$, Monte Carlo standard error 0.003. The only mild signal is that the observed data are slightly less dispersed than typical replications, which is what one expects when the posterior for τ has substantial mass away from 0 while the data themselves are consistent with $\tau = 0$. The check has little power here, exactly as in Exercise 6.1(b), because τ is a free parameter devoted

precisely to matching the dispersion of the y_j.
(b) The inferences that matter are moderately sensitive to the prior on τ and insensitive to the shape of the population distribution.

Alternative model	median τ	$\mathrm{E}[\theta_A \mid y]$	sd	$\mathrm{E}[\theta_C \mid y]$	$\Pr(\theta_A > \theta_C \mid y)$
$p(\tau) \propto 1$ (published)	5.3	11.4	8.4	6.1	0.68
$\tau \sim$ half-Cauchy$(0, 25)$	4.8	11.0	7.9	6.3	0.67
$\tau \sim$ half-Cauchy$(0, 5)$	2.8	9.4	6.2	7.0	0.62
$\tau^2 \sim$ Inv-gamma$(0.001, 0.001)$	0.6	8.4	5.3	7.3	0.56
$\theta_j \sim \mu + \tau t_4$	4.3	11.5	8.8	6.0	0.68

The first three rows are the plausible alternatives and they agree: $\mathrm{E}[\theta_A \mid y]$ moves only from 11.4 to 9.4, and $\Pr(\theta_A > 28.4 \mid y)$ – the separate-analysis point estimate – stays between 0.01 and 0.04, so Rubin's central conclusion (school A's effect is nothing like 28, and the eight schools cannot be reliably ranked) survives. Replacing the normal population distribution by a t_4, fitted by Gibbs sampling with the normal scale-mixture representation of the t, changes nothing to within Monte Carlo error, because with one observation per group there is no information with which to identify the tail behaviour of $p(\theta_j \mid \mu, \tau)$.
The fourth row is the cautionary one, and it is the row Figure 5.9 warns about: Inv-gamma$(0.001, 0.001)$ on τ^2 is proper, but it is so sharply peaked near zero that the posterior median falls to 0.6 and the analysis slides towards complete pooling – the answer Section 5.5 rejects on substantive grounds. (A uniform density on $\log \tau$ is worse still: Section 5.4 notes it gives an improper posterior, so it cannot be used at all.) With only $J = 8$ groups the data carry little information about τ, so any prior peaked at 0 dominates; the uniform prior on τ is the defensible default and the published analysis is a fair representative of the reasonable answers.

Problem 6.5. Hypothesis testing: discuss the statement, "Null hypotheses of no difference are usually known to be false before the data are collected; when they are, their rejection or acceptance simply reflects the size of the sample and the power of the test, and is not a contribution to science" (Savage, 1957, quoted in Kish, 1965). If you agree with this statement, what does this say about the model checking discussed in this chapter?

Solution. Agreed, and it says that a model check must report a magnitude, not a verdict.
Savage's premise is right: a sharp null such as $\theta_1 = \theta_2$ is a measure-zero subset of the parameter space and is essentially never true of two real treatments, two real populations, or two real schools. Its rejection is then a statement about n, not about nature – for any fixed $\theta_1 \neq \theta_2$ the power tends to 1, so rejection is guaranteed by enough data. The Bayesian response (Section 6.1, and Exercises 3.2-3.4) is not to test $\theta_1 = \theta_2$ at all but to report the posterior distribution of $\theta_1 - \theta_2$: the issue is not whether the difference is exactly zero but how large it is and how well it is determined. The same holds for models, which are never exactly true, so

$$\Pr\left(T(y^{\mathrm{rep}}) \geq T(y) \mid y\right) \longrightarrow 0 \quad \text{as } n \to \infty$$

for any test quantity T sensitive to a discrepancy the model genuinely has, however small. A posterior predictive p-value is not an estimate of anything; it is a statement about the resolution of the current dataset.
That does not condemn the model checking of this chapter, because the enterprise is not hypothesis testing. Three differences do the work.
(i) The checks are not conducted with an accept/reject decision at the end. The chapter's own instruction (Section 6.3) is that "we prefer to look at the magnitude of the discrepancy as well as its p-value", and the graphical displays – Figures 6.3-6.5, the realized-versus-replicated scatterplots – are the primary output, the p-value a numerical caption. The speed-of-light check reports that $\min(y_i) = -44$ is far outside its replicate distribution, and the interesting content is the size of the gap, not that it is significant.
(ii) The null being examined is the fitted model in its entirety, which is being used, not merely entertained. A discrepancy is worth acting on when it distorts the inferences one intends to draw, which is

a question of effect size relative to purpose. Exercise 6.1 is the complement of Savage's point: there the identical-schools model passes the check at $p_B = 0.39$ and is nevertheless unacceptable, because the check had no power against the alternatives that matter. A p-value near $\frac{1}{2}$ is therefore no more a contribution to science than a p-value near 0; the choice of T is what carries the scientific content.
(iii) Rejection here is constructive rather than terminal. There is no "accept the null"; there is only the loop of Section 6.1 – find where the replications fail to look like the data, expand the model in the direction indicated, refit, and check again, as in Exercises 6.2 and 6.3, where the failed checks on the constant-rate airline models name the missing ingredient (exposure and a declining rate per mile) rather than merely declaring the models false.

Problem 6.6. Variety of predictive reference sets: in the example of binary outcomes on page 147, it is assumed that the number of measurements, n, is fixed in advance, and so the hypothetical replications under the binomial model are performed with $n = 20$. Suppose instead that the protocol for measurement is to stop once 13 zeros have appeared.
(a) Explain why the posterior distribution of the parameter θ under the assumed model does not change.
(b) Perform a posterior predictive check, using the same test quantity, T = number of switches, but simulating the replications y^{rep} under the new measurement protocol. Display the predictive simulations, $T(y^{\text{rep}})$, and discuss how they differ from Figure 6.5.
In the example of page 147 the binary outcomes $y_1, \ldots, y_n$ are modelled as independent trials with common probability of success θ, given a uniform prior distribution. The observed data, in order, are

$$1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0,$$

so $n = 20$, $\sum_i y_i = 7$, and the posterior is Beta$(8, 14)$. The test quantity is $T =$ the number of switches between 0 and 1 in the sequence, with observed value $T(y) = 3$. Figure 6.5 shows the histogram of $T(y^{\text{rep}\,s})$ over 10,000 draws under the fixed-n protocol, from which $\Pr(T(y^{\text{rep}}) \le T(y) \mid y) = 0.028$.

Solution. (a) Because the stopping rule contributes a factor that does not involve θ. Under the new protocol the sample size n is random, the final observation is necessarily a 0, and the data are determined by the positions of the 7 ones among the first 19 places. The sampling density is

$$p(y \mid \theta) = \theta^7(1-\theta)^{13} \quad \text{for any admissible sequence } y,$$

identical to the fixed-n density, since in both cases the probability of one particular ordered sequence with 7 ones and 13 zeros is $\theta^7(1-\theta)^{13}$. The two protocols differ only in which sequences are admissible, that is, in the support of $p(y \mid \theta)$ – and that support does not depend on θ. Hence the likelihood function of θ is the same up to a constant, and by the likelihood principle (Section 8.1, and the definition of ignorability in Section 8.2) the posterior is Beta$(8, 14)$ under both. What changes is not $p(\theta \mid y)$ but the reference set $\{y^{\text{rep}}\}$ used for the check.
(b) $\Pr(T(y^{\text{rep}}) \le 3 \mid y) = 0.070$, against 0.028 under the fixed-n protocol. Replications were drawn 2×10^5 times by $\theta^s \sim$ Beta$(8, 14)$, then the number of ones $k^s \sim$ Neg-bin$(13, 1-\theta^s)$ preceding the thirteenth zero, then a uniformly random arrangement of those k^s ones among the first $12+k^s$ positions with a 0 in the last. Display the histogram of $T(y^{\text{rep}\,s})$ on integer bins with a vertical line at $T(y) = 3$, beside Figure 6.5 on the same axes; numerically,

	fixed $n = 20$	stop at 13th zero
$\mathrm{E}[T(y^{\text{rep}}) \mid y]$	8.41	9.08
sd	2.57	4.12
$\Pr(T \le 3 \mid y)$	0.028	0.070
$\Pr(T \ge 3 \mid y)$	0.983	0.945

with Monte Carlo standard error at most 0.001, and with the new predictive distribution of n having mean 21.0, standard deviation 5.3, and 95% interval $[14, 34]$ (only 8.6% of replicate datasets have n exactly 20).

Three differences from Figure 6.5. First, the distribution is much wider – standard deviation 4.1 against 2.6 – because n is now random and T grows roughly linearly in n; the extra variance in the reference set is precisely the information the fixed-n protocol was supplying for free. Second, it is visibly bimodal in parity: T must be even when the sequence begins with 0 and odd when it begins with 1, and since $\mathrm{E}[\theta \mid y] = 8/22 = 0.36$ the even values carry about two-thirds of the mass, producing the alternating comb

T	0	1	2	3	4	5	6	7	8
Pr	.011	.003	.041	.015	.082	.035	.114	.055	.123

whereas under the fixed-n protocol both endpoints are free and no parity constraint arises. Third, and as a consequence of the first two, the check is less decisive: the observed $T(y) = 3$, an odd value falling in a trough of the comb, now sits at the 7th percentile rather than the 3rd. The evidence against the independence assumption is in the same direction and of the same kind, but the wider reference set costs the test some power.

Problem 6.7. Prior vs. posterior predictive checks (from Gelman, Meng, and Stern, 1996): consider 100 observations, $y_1, \ldots, y_n$, modeled as independent samples from a $\mathrm{N}(\theta, 1)$ distribution with a diffuse prior distribution, say, $p(\theta) = \frac{1}{2A}$ for $\theta \in [-A, A]$ with some extremely large value of A, such as 10^5. We wish to check the model using, as a test statistic, $T(y) = \max_i |y_i|$: is the maximum absolute observed value consistent with the normal model? Consider a dataset in which $\bar{y} = 5.1$ and $T(y) = 8.1$.
(a) What is the posterior predictive distribution for y^{rep}? Make a histogram for the posterior predictive distribution of $T(y^{\mathrm{rep}})$ and give the posterior predictive p-value for the observation $T(y) = 8.1$.
(b) The prior predictive distribution is $p(y^{\mathrm{rep}}) = \int p(y^{\mathrm{rep}} \mid \theta) p(\theta)\, d\theta$. (Compare to equation (6.1).) What is the prior predictive distribution for y^{rep} in this example? Roughly sketch the prior predictive distribution of $T(y^{\mathrm{rep}})$ and give the approximate prior predictive p-value for the observation $T(y) = 8.1$.
(c) Your answers for (a) and (b) should show that the data are consistent with the posterior predictive but not the prior predictive distribution. Does this make sense? Explain.

Solution. (a) $p_B = 0.13$. Since $|\bar{y}| = 5.1 \ll A$, the truncation is irrelevant and the posterior is $\theta \mid y \sim \mathrm{N}(\bar{y}, 1/n) = \mathrm{N}(5.1, 0.1^2)$ by (2.11)-(2.12). The posterior predictive distribution is therefore the exchangeable normal

$$y^{\mathrm{rep}} \mid y \sim \mathrm{N}\big(5.1\,\mathbf{1},\ \Sigma\big),$$
$$\Sigma = I_{100} + \tfrac{1}{100}\mathbf{1}\mathbf{1}^T,$$

that is, $y_i^{\mathrm{rep}} = \theta + \varepsilon_i$ with $\theta \sim \mathrm{N}(5.1, 0.01)$ and ε_i i.i.d. $\mathrm{N}(0, 1)$: each margin is $\mathrm{N}(5.1, 1.01)$ with pairwise correlation $1/101$. Because θ is pinned near 5.1, all hundred draws are positive with overwhelming probability and $T = \max_i y_i^{\mathrm{rep}}$, so

$$\begin{aligned} p_B &= \Pr\big(T(y^{\mathrm{rep}}) \geq 8.1 \mid y\big) \\ &= \int \Big(1 - \Phi(8.1 - \theta)^{100}\Big)\, \mathrm{N}(\theta \mid 5.1, 0.01)\, d\theta \;=\; 0.131, \end{aligned}$$

confirmed by 2×10^5 simulations, which give 0.1309 (Monte Carlo standard error 0.0008). The predictive distribution of T has mean 7.61, standard deviation 0.44, and 95% interval $[6.87, 8.59]$; its histogram is a single mound over roughly $[6.4, 9.2]$ with the observed 8.1 marked by a vertical line well inside the right shoulder. The observed maximum is entirely ordinary.
(b) $p \approx 0.9999$. The prior predictive distribution is $y_i^{\mathrm{rep}} = \theta + \varepsilon_i$ with $\theta \sim \mathrm{U}[-A, A]$, $A = 10^5$, and ε_i i.i.d. $\mathrm{N}(0, 1)$ – marginally each y_i^{rep} is very nearly uniform on $[-A, A]$, and the hundred components are locked together, all lying within about ± 3 of the common θ. Hence

$$T(y^{\mathrm{rep}}) = \max_i |\theta + \varepsilon_i| \approx |\theta|,$$

so $T(y^{\text{rep}})$ is approximately uniform on $[0, A]$ – a flat slab stretching from 0 to 100,000, with only an 8×10^{-5} sliver of its mass below 8.1 – and

$$\Pr\left(T(y^{\text{rep}}) \geq 8.1\right) \approx 1 - \frac{8.1}{10^5} = 0.99992,$$

simulation giving 0.99991. By the convention of Section 6.3 the p-value is 0.9999, which is as extreme as 0.0001: the observed statistic sits in the far left tail of its prior predictive reference distribution.
(c) Yes, and it is the point of the example: the two checks are testing different things. The prior predictive check evaluates the joint model $p(\theta)p(y \mid \theta)$, prior included, and here it detects nothing about the normal sampling model whatever – it detects that the prior is absurd as a description of where θ actually lives. The reference distribution of T is spread over 10^5 units, so every dataset with a plausible mean lands in its extreme lower tail, and the p-value is driven towards 1 by increasing A alone, without touching the data: it equals $1 - T(y)/A$, a function of the prior's range and nothing else, and has no limit as a probability statement when $A \to \infty$, the prior predictive distribution then being improper.
The posterior predictive check, by contrast, conditions on y through θ, so the location of the reference distribution is set by the data rather than by A, and the comparison is between the observed $\max_i |y_i|$ and what the $\text{N}(\theta, 1)$ sampling model predicts for it – the question actually asked. Its verdict, $p_B = 0.13$, correctly reports that a maximum absolute value of 8.1 among 100 observations centred at 5.1 with unit variance is unremarkable, so prior predictive checks are informative only when the prior is itself a substantive part of the model being tested.

6.2 Exercises 6.8–6.10

Problem 6.8. Variety of posterior predictive distributions: for the educational testing example in Section 6.5, we considered a reference set for the posterior predictive simulations in which $\theta = (\theta_1, \ldots, \theta_8)$ was fixed. This corresponds to a replication of the study with the same eight coaching programs. (Recall the model of Section 5.5: $y_j \mid \theta_j \sim N(\theta_j, \sigma_j^2)$ with σ_j known, $\theta_j \mid \mu, \tau \sim N(\mu, \tau^2)$ i.i.d., and $p(\mu, \tau) \propto 1$; the data of Table 5.2 are

School	y_j	σ_j
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

and the four test statistics of Section 6.5 are $\max_j y_j$, $\min_j y_j$, $\text{mean}(y_j)$ and $\text{sd}(y_j)$, whose posterior predictive p-values under the fixed-θ reference set are given in Figure 6.12.)
(a) Consider an alternative reference set, in which (μ, τ) are fixed but θ is allowed to vary. Define a posterior predictive distribution for y^{rep} under this replication, by analogy to (6.1),

$$p(y^{\text{rep}} \mid y) = \int p(y^{\text{rep}} \mid \theta)\, p(\theta \mid y)\, d\theta.$$

What is the experimental replication that corresponds to this reference set?
(b) Consider switching from the analysis of Section 6.5 to an analysis using this alternative reference set. Would you expect the posterior predictive p-values to be less extreme, more extreme, or stay about the same? Why?
(c) Reproduce the model checks of Section 6.5 based on this posterior predictive distribution. Compare to your speculations in part (b).

Solution. (a) Marginalizing the fresh school effects out of the population distribution,

$$p(y^{\text{rep}} \mid y) = \int\int \prod_{j=1}^{8} N\left(y_j^{\text{rep}} \mid \mu, \tau^2 + \sigma_j^2\right) p(\mu, \tau \mid y)\, d\mu\, d\tau,$$

since the inner replication draws $\theta_j^{\text{rep}} \mid \mu, \tau \sim N(\mu, \tau^2)$ and then $y_j^{\text{rep}} \mid \theta_j^{\text{rep}} \sim N(\theta_j^{\text{rep}}, \sigma_j^2)$, which convolve to $N(\mu, \tau^2 + \sigma_j^2)$ independently across j — the same marginalization that produces (5.18).
The corresponding experiment is a study of eight **new** coaching programs, drawn from the same population of schools that produced the original eight, and evaluated with the same designs (hence the same known standard errors σ_j). Under the reference set of Section 6.5 the same eight programs are re-evaluated; here only the population (μ, τ) is held fixed.
(b) Less extreme, because the reference distribution is strictly more dispersed: conditional on the same (μ, τ), y_j^{rep} has variance $\tau^2 + \sigma_j^2$ here against $V_j + \sigma_j^2$ in Section 6.5, where $V_j = (\sigma_j^{-2} + \tau^{-2})^{-1} < \tau^2$ by (5.17). So every test statistic has a wider predictive distribution and the observed value sits nearer its middle. The effect should be slight, since $E(\tau \mid y) \approx 6.5$ is small beside the σ_j, which run from 9 to 18.
(c) Drawing 200 000 posterior samples of (θ, μ, τ) (grid on τ from (5.21), then μ from (5.20) and θ from (5.17), as in Section 5.4; the improper $p(\tau) \propto 1$ does give a proper posterior here, which is why Section 5.4 adopts it rather than $p(\log \tau) \propto 1$) and replicating under each reference set gives $p = \Pr(T(y^{\text{rep}}) \geq T(y) \mid y)$, with Monte Carlo standard error at most 0.002:

$T(y)$	observed	p, θ fixed	p, (μ, τ) fixed
$\max_j y_j$	28.00	0.48	0.50
$\min_j y_j$	-3.00	0.17	0.18
$\text{mean}(y_j)$	8.75	0.45	0.45
$\text{sd}(y_j)$	10.44	0.78	0.79

The third column reproduces Figure 6.12, whose p-values from only 200 draws are 0.54, 0.19, 0.50, 0.78. The predictive standard deviations do widen as predicted, for instance $\text{sd}(\max_j y_j^{\text{rep}})$ rises from 10.9 to 12.4 and $\text{sd}(\text{mean}(y_j^{\text{rep}}))$ from 6.2 to 7.5, and the p-values for max, min and mean move toward 1/2 as speculated. The sd p-value instead edges from 0.78 to 0.79: widening the reference set also shifts $E(\text{sd}(y^{\text{rep}}))$ up from 13.8 to 14.5, and here the shift beats the spread. In neither reference set does any of the four statistics indicate misfit.

Problem 6.9. Model checking: check the assumed model fitted to the rat tumor data in Section 5.3. Define some test quantities that might be of scientific interest, and compare them to their posterior predictive distributions.
(Section 5.3 fits $y_j \mid \theta_j \sim \text{Bin}(n_j, \theta_j)$ independently for $j = 1, \ldots, 71$, with $\theta_j \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$ i.i.d. and hyperprior $p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$. The data of Table 5.1, displayed as $y_j/n_j =$ (rats with tumors)/(rats in group), are the 70 historical control groups

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

together with the current experiment, 4/14.)

Solution. The model passes every check of the within-group binomial fit and of the shape of the population of rates, but fails decisively on one scientifically pointed quantity: the tumor rate rises with the size of the experiment, which an exchangeable $\text{Beta}(\alpha, \beta)$ prior on the θ_j forbids.
Fitting as in Section 5.3 (grid on $(\log(\alpha/\beta), \log(\alpha+\beta))$ over the region of Figure 5.3, then $\theta_j \mid \alpha, \beta, y \sim \text{Beta}(\alpha + y_j, \beta + n_j - y_j)$ by (5.7); the improper hyperprior (5.9) is the one choice in Section 5.3 shown

to leave the posterior integrable, so truncating to the grid loses nothing) gives 20 000 draws with

$$E(\alpha \mid y) = 2.14, \quad E(\beta \mid y) = 12.71, \quad E\Big(\tfrac{\alpha}{\alpha+\beta} \,\Big|\, y\Big) = 0.145,$$

matching the book's approximate population rate of 0.14 on page 128. Two reference sets are useful:

- replication (i), the standard one of (6.1): $y_j^{\text{rep}} \sim \text{Bin}(n_j, \theta_j)$ with the posterior θ_j held fixed, which checks the binomial sampling model;
- replication (ii), 71 **new** experiments: $\theta_j^{\text{rep}} \sim \text{Beta}(\alpha, \beta)$ redrawn, then $y_j^{\text{rep}} \sim \text{Bin}(n_j, \theta_j^{\text{rep}})$, which checks the population model and the exchangeability of the θ_j across groups.

With $p = \Pr(T(y^{\text{rep}}) \geq T(y) \mid y)$ and $p_j = y_j/n_j$ (Monte Carlo standard error at most 0.004):

T	observed	p, (i)	p, (ii)
$\#\{j : y_j = 0\}$	14	0.16	0.16
$\max_j p_j$	0.375	0.91	0.94
$\text{sd}(p_1, \ldots, p_{71})$	0.104	0.71	0.76
$\text{corr}(p_j, n_j)$	0.335	0.11	0.003
Spearman $\text{corr}(p_j, n_j)$	0.364	0.10	0.001

The first three are unremarkable. The count of tumor-free groups, the scientifically salient "how often does a control group show nothing" quantity, is 14 against a predictive mean of 10.3 and 95% interval $[4, 17]$ under (i): high, but within the reference distribution. The largest observed rate, $9/24 = 0.375$, is if anything **smaller** than the model expects ($E = 0.47$), so there is no outlying group. The chi-square discrepancy

$$T(y, \theta) = \sum_{j=1}^{71} \frac{(y_j - n_j\theta_j)^2}{n_j\theta_j(1-\theta_j)}$$

gives $\Pr(T(y^{\text{rep}}, \theta) \geq T(y, \theta) \mid y) = 0.50$: no overdispersion beyond what the beta population already supplies.

The last two rows are the failure. Under (ii) the replicated correlation between rate and group size is centered at 0.00 with 95% interval $[-0.20, 0.23]$, and only 3 replications in 1000 reach the observed 0.335; the rank correlation is worse, 1 in 1000. (Under (i) the check has little power: the θ_j are conditioned on the very data whose correlation is at issue, so the reference distribution inherits it, centering at 0.22.) Substantively, the larger experiments in Tarone's table are systematically the ones with higher tumor incidence, so θ_j and n_j are dependent and the groups are not exchangeable; n_j should enter the model, for example through a regression of $\text{logit}(\theta_j)$ on $\log n_j$ as in the hierarchical regression of Chapter 15.

Problem 6.10. Checking the assumption of equal variance: Figures 1.1 and 1.2 on pages 14 and 15 display data on point spreads x and score differentials y of a set of professional football games. (The data are available at http://www.stat.columbia.edu/~gelman/book/.) Figure 1.1 is a scatterplot of the actual outcome y against the point spread x for each of 672 games (coordinates jittered by $\text{U}(-0.1, 0.1)$ in x and $\text{U}(-0.2, 0.2)$ in y); Figure 1.2a is the same scatterplot for $y - x$ against x, and Figure 1.2b a histogram of $y - x$ with the $N(0, 14^2)$ density superimposed. In Section 1.6, a model is fit of the form $y \sim N(x, 14^2)$. However, Figure 1.2a seems to show a pattern of decreasing variance of $y - x$ as a function of x.

(a) Simulate several replicated datasets y^{rep} under the model and, for each, create graphs like Figures 1.1 and 1.2. Display several graphs per page, and compare these to the corresponding graphs of the actual data. This is a graphical posterior predictive check as described in Section 6.4.

(b) Create a numerical summary $T(x, y)$ to capture the apparent decrease in variance of $y - x$ as a function of x. Compare this to the distribution of simulated test statistics, $T(x, y^{\text{rep}})$, and compute the p-value for this posterior predictive check.

Solution. The apparent decrease is not real: the sharpest summary of it, the ratio of residual standard deviations for large and small spreads, has posterior predictive p-value 0.13, and the replicated scatterplots are indistinguishable from the data.

The model of Section 1.6 has no free parameters, so $p(y^{\text{rep}} \mid y) = \prod_{i=1}^{672} N(y_i^{\text{rep}} \mid x_i, 14^2)$ and the posterior predictive reference distribution coincides with the prior predictive one; the p-values below are exact tail probabilities of a fully specified null. The observed residuals satisfy $\overline{y - x} = 0.07$ and $\text{sd}(y - x) = 13.86$, confirming the $N(0, 14^2)$ calibration overall.

(a) Five replications $y_i^{\text{rep}} = x_i + 14 z_i$, $z_i \sim N(0, 1)$, plotted with the jitter of Figures 1.1 and 1.2a alongside the real data, show the same funnel: the apparent narrowing at large x is present in the replications too, because point spreads above 10 occur in only 34 of the 672 games and extreme residuals are simply rarer where the data are thin. (Figures `bda3-ch06-football-outcome` and `bda3-ch06-football-resid`.)

(b) Take

$$T(x, y) = \frac{\text{sd}\{y_i - x_i : x_i \geq 6.5\}}{\text{sd}\{y_i - x_i : x_i \leq 3.5\}},$$

which splits the games into 177 large-spread and 327 small-spread games and centers at 1 under constant variance. Observed, $T(x, y) = 0.926$; over 10 000 replications $T(x, y^{\text{rep}})$ has mean 1.001, standard deviation 0.067 and central 95% interval $[0.874, 1.136]$, so

$$p = \Pr(T(x, y^{\text{rep}}) \leq T(x, y) \,|\, y) = 0.13$$

with Monte Carlo standard error 0.003 (the lower tail, since the alternative of interest is variance decreasing in x). A second summary, the correlation $T_2(x, y) = \text{corr}(|y_i - x_i|, x_i) = -0.030$ against a reference distribution of mean 0.000 and standard deviation 0.039, gives $p = 0.22$. Binning confirms the picture, with observed and 95% predictive standard deviations of $y - x$:

spread range	n	observed sd	95% predictive
$x = 0$	17	15.85	$[9.2, 18.8]$
$0.5 \leq x \leq 3$	251	13.73	$[12.8, 15.2]$
$3.5 \leq x \leq 6$	227	14.81	$[12.7, 15.3]$
$6.5 \leq x \leq 10$	143	12.84	$[12.4, 15.6]$
$x \geq 10.5$	34	11.65	$[10.6, 17.3]$

Every observed value lies inside its interval.

7 Evaluating, Comparing, and Expanding Models

Contents

7.1 Exercises 7.1–7.7

Problem 7.1. Predictive accuracy and cross-validation: Compute AIC, DIC, WAIC, and cross-validation for the logistic regression fit to the bioassay example of Section 3.7.
The bioassay data of Table 3.1 are

Dose, x_i (log g/ml)	Animals, n_i	Deaths, y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

with model $y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$, $\text{logit}(\theta_i) = \alpha + \beta x_i$, and the uniform prior density $p(\alpha, \beta) \propto 1$ of Section 3.7.

Solution.

$$\text{AIC} = 7.96, \quad \text{DIC} = 7.91, \quad \text{WAIC} = 7.81,$$

and LOO cross-validation gives $\text{lppd}_{\text{loo-cv}} = -4.89$; the four data points are the $n = 4$ prediction units.
AIC. The maximum likelihood estimate is $(\hat\alpha, \hat\beta)_{\text{mle}} = (0.85, 7.75)$ with

$$\log p(y \mid \hat\theta_{\text{mle}}) = \sum_{i=1}^{4} \log \text{Bin}(y_i \mid n_i, \theta_i) = -1.98.$$

Two parameters are estimated, so by (7.6)

$$\widehat{\text{elpd}}_{\text{AIC}} = -1.98 - 2 = -3.98, \qquad \text{AIC} = 7.96.$$

DIC. Posterior summaries come from the Section 3.7 grid posterior on $[-5, 10] \times [-10, 40]$. The posterior mean is $\hat\theta_{\text{Bayes}} = (1.31, 11.61)$, and

$$\log p(y \mid \hat\theta_{\text{Bayes}}) = -2.23,$$
$$E_{\text{post}}\big(\log p(y \mid \theta)\big) = -3.09.$$

Hence by (7.8)–(7.9)

$$p_{\text{DIC}} = 2\big(-2.23 - (-3.09)\big) = 1.73,$$

$$\widehat{\text{elpd}}_{\text{DIC}} = -2.23 - 1.73 = -3.95, \qquad \text{DIC} = 7.91.$$

The alternative (7.10) gives $p_{\text{DIC alt}} = 2\,\text{var}_{\text{post}}(\log p(y \mid \theta)) = 2.41$; the posterior of (α, β) is skewed, so the mean sits off the mode and the two effective-parameter counts straddle $k = 2$.
WAIC. Pointwise, with $\text{lppd}_i = \log E_{\text{post}} p(y_i \mid \theta)$ and $V_i = \text{var}_{\text{post}} \log p(y_i \mid \theta)$,

i	x_i	lppd_i	V_i
1	-0.86	-0.034	0.017
2	-0.30	-1.266	0.608
3	-0.05	-1.406	0.516
4	0.73	-0.031	0.024

so by (7.5), (7.11), (7.13)

$$\text{lppd} = -2.74, \qquad p_{\text{WAIC 1}} = 0.71,$$
$$p_{\text{WAIC 2}} = 1.17, \qquad \widehat{\text{elppd}}_{\text{WAIC}} = -3.90,$$

and $\text{WAIC} = -2(-3.90) = 7.81$. The two extreme doses are predicted almost perfectly and contribute essentially no effective parameters; all the fitting happens at x_2, x_3.
Cross-validation. By (7.14), evaluating each $p_{\text{post}(-i)}(y_i)$ on the Section 3.7 grid,

i	$\log p_{\text{post}(-i)}(y_i)$
1	-0.054
2	-2.217
3	-2.553
4	-0.069

$$\mathrm{lppd}_{\mathrm{loo-cv}} = -4.89, \qquad p_{\mathrm{loo-cv}} = \mathrm{lppd} - \mathrm{lppd}_{\mathrm{loo-cv}} = 2.16.$$

With $n = 4$ the bias correction is not negligible: $\mathrm{lppd}_{-i} = -3.27$, so $b = \mathrm{lppd} - \mathrm{lppd}_{-i} = 0.53$ and $\mathrm{lppd}_{c\,\mathrm{loo-cv}} = -4.36$.

The grid bound is not cosmetic. Under the **untruncated** flat prior the leave-one-out posteriors omitting $i = 2$ or $i = 3$ are improper: deleting $x_2 = -0.30$ leaves responses that are separated except at the single dose x_3, and along the ray $\beta \to \infty$ with $\alpha + \beta x_3$ held fixed the likelihood tends to the positive constant $\binom{5}{3}\theta^3(1-\theta)^2$, so $\iint p(y_{-2} \mid \alpha, \beta)\, d\alpha\, d\beta$ diverges linearly in β. (Numerically, doubling the integration box doubles the mass.) The numbers above therefore use the bounded grid of Section 3.7, i.e. the uniform prior truncated to $[-5, 10] \times [-10, 40]$, under which all four leave-one-out posteriors are proper; $\mathrm{lppd}_{\mathrm{loo-cv}}$ for $i = 2, 3$ would drift downward without that truncation. AIC, DIC and WAIC, which condition on all four points, are unaffected.

Problem 7.2. Information criteria: show that DIC yields an estimate of elpd that is correct in expectation, in the case of normal models or in the asymptotic limit of large sample sizes (see Spiegelhalter et al., 2002, p. 604).
Here $\widehat{\mathrm{elpd}}_{\mathrm{DIC}} = \log p(y \mid \hat{\theta}_{\mathrm{Bayes}}) - p_{\mathrm{DIC}}$ as in (7.7), with $p_{\mathrm{DIC}} = 2\big(\log p(y \mid \hat{\theta}_{\mathrm{Bayes}}) - E_{\mathrm{post}} \log p(y \mid \theta)\big)$ as in (7.8), and the target is the plug-in expected log predictive density (7.3), $\mathrm{elpd}_{\hat{\theta}} = E_f\big(\log p(\tilde{y} \mid \hat{\theta}(y))\big)$, where $\tilde{y}$ is a replicate data set of the same size n.

Solution. Both halves reduce to the same two facts: $p_{\mathrm{DIC}} = k$, and the in-sample log predictive density overshoots $\mathrm{elpd}_{\hat{\theta}}$ by exactly k.

The exact normal case. Take $y \sim \mathrm{N}(X\beta, \sigma^2 I_n)$ with σ^2 known, X of full column rank k, and $p(\beta) \propto 1$; then $\beta \mid y \sim \mathrm{N}(\hat{\beta}, \sigma^2 (X^T X)^{-1})$ with $\hat{\beta} = (X^T X)^{-1} X^T y = \hat{\beta}_{\mathrm{Bayes}}$. The log-likelihood is exactly quadratic,

$$\log p(y \mid \beta) = \log p(y \mid \hat{\beta}) - \tfrac{1}{2\sigma^2}(\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}),$$

so, since $E_{\mathrm{post}}(\beta - \hat{\beta})(\beta - \hat{\beta})^T = \sigma^2 (X^T X)^{-1}$,

$$\begin{aligned} E_{\mathrm{post}} \log p(y \mid \beta) &= \log p(y \mid \hat{\beta}) - \tfrac{1}{2\sigma^2} \operatorname{tr}\big(X^T X \cdot \sigma^2 (X^T X)^{-1}\big) \\ &= \log p(y \mid \hat{\beta}) - \tfrac{k}{2}, \end{aligned}$$

whence $p_{\mathrm{DIC}} = k$ exactly and $\widehat{\mathrm{elpd}}_{\mathrm{DIC}} = \log p(y \mid \hat{\beta}) - k$.

Now take expectations over $y \sim \mathrm{N}(X\beta_0, \sigma^2 I)$. Since $\|y - X\hat{\beta}\|^2 / \sigma^2 \sim \chi^2_{n-k}$,

$$E_y \log p(y \mid \hat{\beta}) = -\tfrac{n}{2} \log(2\pi\sigma^2) - \tfrac{n-k}{2}.$$

For an independent replicate $\tilde{y} \sim \mathrm{N}(X\beta_0, \sigma^2 I)$, using $\tilde{y} \perp \hat{\beta}$ and $E\|X(\hat{\beta} - \beta_0)\|^2 = \sigma^2 k$,

$$\begin{aligned} \mathrm{elpd}_{\hat{\theta}} &= -\tfrac{n}{2} \log(2\pi\sigma^2) - \tfrac{1}{2\sigma^2} E\|\tilde{y} - X\hat{\beta}\|^2 \\ &= -\tfrac{n}{2} \log(2\pi\sigma^2) - \tfrac{n}{2} - \tfrac{k}{2}. \end{aligned}$$

Subtracting, $E_y \log p(y \mid \hat{\beta}) - \mathrm{elpd}_{\hat{\theta}} = k$, i.e.

$$E_y\big(\widehat{\mathrm{elpd}}_{\mathrm{DIC}}\big) = \mathrm{elpd}_{\hat{\theta}}.$$

The asymptotic case. Under the Chapter 4 regularity conditions the posterior concentrates, $\theta \mid y \approx \mathrm{N}(\hat{\theta}, J^{-1})$ with $J = -\nabla^2 \log p(y \mid \theta)|_{\hat{\theta}}$ the observed information, $\hat{\theta}_{\mathrm{Bayes}} = \hat{\theta}_{\mathrm{mle}} + O_p(n^{-1})$, and the log-likelihood is quadratic to the same order:

$$\log p(y \mid \theta) = \log p(y \mid \hat{\theta}) - \tfrac{1}{2}(\theta - \hat{\theta})^T J (\theta - \hat{\theta}) + o_p(1).$$

Averaging over the posterior gives $E_{\mathrm{post}} \log p(y \mid \theta) = \log p(y \mid \hat{\theta}) - \tfrac{1}{2}\mathrm{tr}(JJ^{-1}) + o_p(1)$, so

$$p_{\mathrm{DIC}} = k + o_p(1),$$

the χ^2_k-based drop of $k/2$ noted on p. 171.
For the bias, let θ_0 be the true (or Kullback–Leibler optimal) value and $I = E_y(J) = n\,I_1$. Expanding in θ about $\hat\theta$ and about θ_0,

$$\log p(y \mid \hat\theta) - \log p(y \mid \theta_0) = \tfrac{1}{2}(\hat\theta - \theta_0)^T J (\hat\theta - \theta_0) + o_p(1),$$
$$E_{\tilde y} \log p(\tilde y \mid \theta_0) - E_{\tilde y} \log p(\tilde y \mid \hat\theta) = \tfrac{1}{2}(\hat\theta - \theta_0)^T I (\hat\theta - \theta_0) + o_p(1),$$

the second line because θ_0 maximizes $\theta \mapsto E_{\tilde y} \log p(\tilde y \mid \theta)$ with curvature I there. Asymptotic normality gives $\hat\theta - \theta_0 \sim \mathrm{N}(0, I^{-1})$ and $J = I + o_p(n)$, so both quadratic forms are χ^2_k in the limit and each has expectation $k/2$. Adding,

$$\begin{aligned} E_y \log p(y \mid \hat\theta) - \mathrm{elpd}_{\hat\theta} &= \tfrac{k}{2} + \tfrac{k}{2} + o(1) \\ &= k + o(1), \end{aligned}$$

since $E_y \log p(y \mid \theta_0) = E_{\tilde y} \log p(\tilde y \mid \theta_0)$ (y and $\tilde y$ have the same distribution). Combining with $p_{\mathrm{DIC}} = k + o_p(1)$ and $\log p(y \mid \hat\theta_{\mathrm{Bayes}}) = \log p(y \mid \hat\theta_{\mathrm{mle}}) + o_p(1)$,

$$E_y\big(\widehat{\mathrm{elpd}}_{\mathrm{DIC}}\big) = \mathrm{elpd}_{\hat\theta} + o(1).$$

Problem 7.3. Predictive accuracy for hierarchical models: Compute AIC, DIC, WAIC, and cross-validation for the meta-analysis example of Section 5.6.
That example takes the 22 beta-blocker trials of Table 5.4, summarized by the empirical log-odds ratios y_j of (5.23) and their approximate sampling standard deviations σ_j of (5.24):

j	y_j	σ_j	j	y_j	σ_j
1	0.028	0.850	12	−0.039	0.229
2	−0.741	0.483	13	−0.593	0.425
3	−0.541	0.565	14	0.282	0.205
4	−0.246	0.138	15	−0.321	0.298
5	0.069	0.281	16	−0.135	0.261
6	−0.584	0.676	17	0.141	0.364
7	−0.512	0.139	18	0.322	0.553
8	−0.079	0.204	19	0.444	0.717
9	−0.424	0.274	20	−0.218	0.260
10	−0.335	0.117	21	−0.591	0.257
11	−0.213	0.195	22	−0.608	0.272

The model is $y_j \mid \theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$ with σ_j known, $\theta_j \mid \mu, \tau \sim \mathrm{N}(\mu, \tau^2)$, and $p(\mu, \tau) \propto 1$, alongside its no-pooling ($\tau = \infty$) and complete-pooling ($\tau = 0$) limits.

Solution. The analogue of Table 7.1 for these 22 studies, with y_j the $n = 22$ prediction units:

	No pooling	Complete pooling	Hierarchical
$-2\,$lpd (mle)	−11.9	11.4	—
k	22.0	1.0	—
AIC	32.1	13.4	—
$-2\,$lpd (Bayes)	−11.9	11.4	1.7
p_{DIC}	22.0	1.0	6.0
DIC	32.1	13.4	13.7
$-2\,$lppd	3.4	11.2	3.9
$p_{\mathrm{WAIC}\,1}$	6.8	1.2	3.8
$p_{\mathrm{WAIC}\,2}$	11.0	1.3	4.8
WAIC	25.4	13.8	13.5
$p_{\mathrm{loo-cv}}$	—	1.3	5.3
$-2\,\mathrm{lppd}_{\mathrm{loo-cv}}$	—	13.9	14.5

Blanks are undefined exactly as in Table 7.1: AIC is defined relative to a maximum likelihood estimate and a fixed parameter count k, neither of which is appropriate for a hierarchical model whose 22 θ_j

are constrained by an estimated τ; LOO-CV needs a predictive density for a held-out y_j, and under no pooling θ_j has a flat prior and no other data bear on it.
How the entries are obtained. Under no pooling $\hat{\theta}_j = y_j$ and $\theta_j \mid y \sim \mathrm{N}(y_j, \sigma_j^2)$, so

$$-2 \log p(y \mid \hat{\theta}) = \sum_{j=1}^{22} \log(2\pi\sigma_j^2) = -11.9,$$

$p_{\mathrm{DIC}} = 22.0$ (exactly k, as in Exercise 7.2), and

$$p_{\mathrm{WAIC}\,2} = \sum_j \mathrm{var}_{\mathrm{post}} \log \mathrm{N}(y_j \mid \theta_j, \sigma_j^2) = \tfrac{22}{2} = 11,$$

since $(y_j - \theta_j)^2/\sigma_j^2 \sim \chi_1^2$ has variance 2 and enters with a factor $-\frac{1}{2}$.
Under complete pooling $\theta_j \equiv \mu$ with $\mu \mid y \sim \mathrm{N}(-0.260, 0.0503^2)$, giving $-2\,\mathrm{lpd} = 11.4$, $p_{\mathrm{DIC}} = 1.0$, $p_{\mathrm{WAIC}\,2} = 1.32$. LOO is available in closed form: $y_j \mid y_{-j} \sim \mathrm{N}(\hat{\mu}_{-j}, \sigma_j^2 + V_{-j})$ with $\hat{\mu}_{-j}, V_{-j}$ the precision-weighted mean and variance from (5.20) applied to y_{-j}, and summing gives $-2\,\mathrm{lppd}_{\mathrm{loo-cv}} = 13.85$ against WAIC $= 13.83$ — for a one-parameter model with $n = 22$ the two agree to the second digit, as they should.
For the hierarchical model we simulate from the Section 5.4 factorization: $p(\tau \mid y)$ on a grid, then $\mu \mid \tau, y$ and $\theta_j \mid \mu, \tau, y$ normal. This reproduces the Section 5.6 inference ($E(\tau \mid y) = 0.13$, $E(\mu \mid y) = -0.25$). With $S = 4 \times 10^5$ draws (Monte Carlo error under 0.02 on each entry),

$$\begin{aligned} -2 \log p(y \mid \hat{\theta}_{\mathrm{Bayes}}) &= 1.68, & p_{\mathrm{DIC}} &= 6.01, \\ -2\,\mathrm{lppd} &= 3.94, & p_{\mathrm{WAIC}\,2} &= 4.76. \end{aligned}$$

Both effective sizes fall between 1 and 22 as they must. For LOO, the held-out study's effect is a fresh draw $\theta_j \mid \mu, \tau \sim \mathrm{N}(\mu, \tau^2)$, so θ_j and μ integrate out analytically and only τ needs the grid:

$$p(y_j \mid y_{-j}) = \int \mathrm{N}\big(y_j \mid \hat{\mu}_{-j}(\tau),\ \sigma_j^2 + \tau^2 + V_{-j}(\tau)\big) \times p(\tau \mid y_{-j})\, d\tau,$$

refitting $p(\tau \mid y_{-j})$ for each j. Summing the 22 terms gives $-2\,\mathrm{lppd}_{\mathrm{loo-cv}} = 14.5$ and $p_{\mathrm{loo-cv}} = 5.3$.
The no-pooling model has by far the best raw fit and by far the worst corrected score, while the hierarchical and complete-pooling models are indistinguishable on all three criteria, as they must be with $\hat{\tau} \approx 0.13$ small relative to the typical $\sigma_j \approx 0.3$.

Problem 7.4. Bayes factors when the prior distribution is improper: on page 183, we discuss Bayes factors for comparing two extreme models for the SAT coaching example. The two models are

$$H_1 :\ p(y \mid \theta_1, \ldots, \theta_J) = \prod_{j=1}^{J} \mathrm{N}(y_j \mid \theta_j, \sigma_j^2),$$

with the θ_j unrelated, and

$$H_2 :\ p(y \mid \theta) = \prod_{j=1}^{J} \mathrm{N}(y_j \mid \theta, \sigma_j^2), \quad \theta_1 = \cdots = \theta_J = \theta,$$

with the σ_j known. The flat prior densities of page 183 are replaced by independent $\mathrm{N}(0, A^2)$ prior distributions for the free parameters: $\theta_j \sim \mathrm{N}(0, A^2)$ independently under H_1, and $\theta \sim \mathrm{N}(0, A^2)$ under H_2.
(a) Derive the Bayes factor, $p(H_2 \mid y)/p(H_1 \mid y)$, as a function of $y_1, \ldots, y_J$, $\sigma_1, \ldots, \sigma_J$, and A, for the models with $\mathrm{N}(0, A^2)$ prior distributions.

(b) Evaluate the Bayes factor in the limit $A \to \infty$.
(c) For fixed A, evaluate the Bayes factor as the number of schools, J, increases. Assume for simplicity that $\sigma_1 = \cdots = \sigma_J = \sigma$, and that the sample mean and variance of the y_j's do not change.

Solution. (a) With $w_j = \sigma_j^{-2}$, $W = \sum_j w_j$, $\bar{y} = W^{-1}\sum_j w_j y_j$ and $S = \sum_j w_j (y_j - \bar{y})^2$,

$$\frac{p(H_2 \mid y)}{p(H_1 \mid y)} = \frac{p(H_2)}{p(H_1)}\, B,$$

$$B = \frac{p(y \mid H_2)}{p(y \mid H_1)} = \frac{\prod_j \big(1 + A^2/\sigma_j^2\big)^{1/2}}{(1 + WA^2)^{1/2}} \times \exp\Big\{ -\tfrac{S}{2} - \frac{\bar{y}^2}{2(A^2 + W^{-1})} + \sum_j \frac{y_j^2}{2(\sigma_j^2 + A^2)} \Big\}.$$

Both marginal likelihoods are elementary normal integrals. Under H_1 the θ_j are independent, so the y_j are marginally independent:

$$p(y \mid H_1) = \prod_{j=1}^{J} \mathrm{N}(y_j \mid 0, \sigma_j^2 + A^2).$$

Under H_2, complete the square in θ using $\sum_j w_j (y_j - \theta)^2 = S + W(\theta - \bar{y})^2$:

$$\begin{aligned} p(y \mid H_2) &= \prod_j (2\pi\sigma_j^2)^{-1/2} e^{-S/2} \int e^{-\frac{W}{2}(\theta - \bar{y})^2} \mathrm{N}(\theta \mid 0, A^2)\, d\theta \\ &= \prod_j (2\pi\sigma_j^2)^{-1/2} e^{-S/2} \Big(\tfrac{2\pi}{W}\Big)^{1/2} \mathrm{N}\big(\bar{y} \mid 0, W^{-1} + A^2\big), \end{aligned}$$

the last step because $\int \mathrm{N}(\theta \mid \bar{y}, W^{-1}) \mathrm{N}(\theta \mid 0, A^2) d\theta = \mathrm{N}(\bar{y} \mid 0, W^{-1} + A^2)$. Dividing, the factors $(2\pi/W)^{1/2} (2\pi(W^{-1} + A^2))^{-1/2} = (1 + WA^2)^{-1/2}$ and $\prod_j (2\pi\sigma_j^2)^{-1/2} \big/ \prod_j (2\pi(\sigma_j^2 + A^2))^{-1/2} = \prod_j (1 + A^2/\sigma_j^2)^{1/2}$ give the display. (Checked against numerical integration of $p(y \mid H_2)$ for the eight-schools data of Table 5.2: $\log B$ agrees to six decimals, giving $B = 63.0$ at $A = 30$ and $B = 8.24 \times 10^{11}$ at $A = 1000$.)
(b) $B \to \infty$ like A^{J-1}: as $A \to \infty$,

$$(1 + WA^2)^{-1/2} \sim W^{-1/2} A^{-1}, \qquad \prod_j \big(1 + A^2/\sigma_j^2\big)^{1/2} \sim \frac{A^J}{\prod_j \sigma_j},$$

and both terms in the exponent involving A vanish, so

$$B \;\sim\; \frac{A^{J-1}}{W^{1/2} \prod_j \sigma_j}\, e^{-S/2} \;\longrightarrow\; \infty \qquad (J \geq 2).$$

So with noninformative priors taken as the limit of proper ones, $p(H_2 \mid y) \to 1$ whatever the data: the complete-pooling model wins by fiat, because H_1 pays the A^{-1} price of a diffuse prior J times and H_2 pays it once. This is the sensitivity to A^2 asserted on p. 184.
(c) $\log B = Jc - \frac{1}{2}\log J + O(1)$, so $B \to \infty$ or 0 exponentially in J according to the sign of

$$c = \tfrac{1}{2}\log\Big(1 + \frac{A^2}{\sigma^2}\Big) - \frac{v}{2\sigma^2} + \frac{v + \bar{y}^2}{2(\sigma^2 + A^2)},$$

where $\bar{y} = \frac{1}{J}\sum_j y_j$ and $v = \frac{1}{J}\sum_j (y_j - \bar{y})^2$ are the fixed sample mean and variance. Indeed with $\sigma_j \equiv \sigma$ we have $W = J/\sigma^2$, $S = Jv/\sigma^2$ and $\sum_j y_j^2 = J(v + \bar{y}^2)$, so

$$\log B = -\tfrac{1}{2}\log\big(1 + JA^2/\sigma^2\big) + \tfrac{J}{2}\log\big(1 + A^2/\sigma^2\big) - \frac{Jv}{2\sigma^2} - \frac{\bar{y}^2}{2(A^2 + \sigma^2/J)} + \frac{J(v + \bar{y}^2)}{2(\sigma^2 + A^2)},$$

and collecting the terms proportional to J gives c, the rest being $-\frac{1}{2}\log J + O(1)$.
The sign is decided by whether the observed spread matches what H_1 predicts. If $v = \sigma^2 + A^2$ and $\bar{y} = 0$ — exactly the marginal variance under H_1 — then, writing $r = A^2/\sigma^2$,

$$c = \tfrac{1}{2}\log(1+r) - \tfrac{1+r}{2} + \tfrac{1}{2} = \tfrac{1}{2}\big(\log(1+r) - r\big) < 0,$$

so H_1 wins exponentially; if the y_j are less dispersed than that, $c > 0$ and H_2 wins exponentially. For the eight-schools numbers ($\bar{y} = 8.8$, $v^{1/2} = 9.8$, $\sigma \approx 12.6$) and $A = 20$ one gets $c = +0.48$, hence $B = 9.4$ at $J = 8$ but $B = 3.7 \times 10^{15}$ at $J = 80$: with $\sigma_j \equiv \sigma$ the Bayes factor depends on the data only through $(\bar{y}, v)$, which are being held fixed, so the whole movement is the J in the exponent.

Problem 7.5. Power-transformed normal models: A natural expansion of the family of normal distributions, for all-positive data, is through power transformations, which are used in various contexts, including regression models. For simplicity, consider univariate data $y = (y_1, \ldots, y_n)$, that we wish to model as independent and identically normally distributed after transformation.
Box and Cox (1964) propose the model, $y_i^{(\phi)} \sim \mathrm{N}(\mu, \sigma^2)$, where

$$y_i^{(\phi)} = \begin{cases} (y_i^\phi - 1)/\phi & \text{for } \phi \neq 0, \\ \log y_i & \text{for } \phi = 0. \end{cases} \tag{7.19}$$

The parameterization in terms of $y_i^{(\phi)}$ allows a continuous family of power transformations that includes the logarithm as a special case. To perform Bayesian inference, one must set up a prior distribution for the parameters, (μ, σ, ϕ).
(a) It seems natural to apply a prior distribution of the form $p(\mu, \log\sigma, \phi) \propto p(\phi)$, where $p(\phi)$ is a prior distribution (perhaps uniform) on ϕ alone. Unfortunately, this prior distribution leads to unreasonable results. Set up a numerical example to show why. (Hint: consider what happens when all the data points y_i are multiplied by a constant factor.)
(b) Box and Cox (1964) propose a prior distribution that has the form $p(\mu, \log\sigma, \phi) \propto \dot{y}^{\,1-\phi} p(\phi)$, where $\dot{y} = (\prod_{i=1}^n y_i)^{1/n}$. Show that this prior distribution eliminates the problem in (a).
(c) Write the marginal posterior density, $p(\phi \mid y)$, for the model in (b).
(d) Discuss the implications of the fact that the prior distribution in (b) depends on the data.
(e) The power transformation model is used with the understanding that negative values of $y_i^{(\phi)}$ are not possible. Discuss the effect of the implicit truncation on the model.
See Pericchi (1981) and Hinkley and Runger (1984) for further discussion of Bayesian analysis of power transformations.

Solution. (a) Changing the units of y tilts the posterior of ϕ: for any $c > 0$,

$$p(\phi \mid cy) \;\propto\; c^{\phi}\, p(\phi \mid y).$$

To see this, integrate out $(\mu, \log\sigma)$ first. Including the Jacobian $\prod_i \mathrm{d}y_i^{(\phi)}/\mathrm{d}y_i = \prod_i y_i^{\phi-1}$,

$$p(y \mid \mu, \sigma, \phi) = \prod_i \mathrm{N}\big(y_i^{(\phi)} \mid \mu, \sigma^2\big) \prod_i y_i^{\phi-1},$$

and with $p(\mu, \log\sigma) \propto 1$ the standard normal integral (as in Section 3.2) gives

$$p(\phi \mid y) \;\propto\; p(\phi) \Big(\prod_i y_i\Big)^{\phi-1} \Big[\sum_i \big(y_i^{(\phi)} - \overline{y^{(\phi)}}\big)^2\Big]^{-(n-1)/2}.$$

Now $(cy_i)^{(\phi)} = c^\phi y_i^{(\phi)} + (c^\phi - 1)/\phi$, an affine map, so the bracket is multiplied by $c^{2\phi}$ and the bracket to the power $-(n-1)/2$ by $c^{-(n-1)\phi}$, while the Jacobian factor is multiplied by $c^{n(\phi-1)}$. The product of the two is $c^{\phi-n}$, which up to the ϕ-free constant c^{-n} is the claimed tilt.
Numerical example. Take $n = 10$ observations

$$y = (1.2,\ 2.4,\ 3.1,\ 4.0,\ 5.5,\ 6.2,\ 8.0,\ 9.7,\ 12.4,\ 18.0),$$

with $p(\phi) \propto 1$. The posterior mode of ϕ is

c	mode of $p(\phi \mid cy)$
0.1	+0.10
1	+0.53
10	+1.07
100	+1.91

The same measurements, reported in different units, are said to need a log transformation, a square-root transformation, no transformation, or a square. Worse, the tilt destroys propriety: as $\phi \to \infty$ the log posterior is asymptotically linear with slope $\sum_i \log(cy_i) - (n-1)\log(c\,y_{\max}) = -9.15 + \log c$, so for $c > e^{9.15} \approx 9.4 \times 10^3$ — a change from grams to milligrams and a little — $p(\phi \mid cy)$ is not integrable at all. A prior that is flat in $(\mu, \log\sigma)$ cannot be right, because μ and σ live on the $y^{(\phi)}$ scale, whose units change with ϕ.

(b) The factor $\dot{y}^{\,1-\phi}$ is exactly the compensating Jacobian. With it,

$$p(\phi \mid y) \propto p(\phi)\,\dot{y}^{\,1-\phi}\,\dot{y}^{\,n(\phi-1)}\Big[\sum_i \big(y_i^{(\phi)} - \overline{y^{(\phi)}}\big)^2\Big]^{-(n-1)/2}$$
$$= p(\phi)\Big[\frac{\sum_i (y_i^{(\phi)} - \overline{y^{(\phi)}})^2}{\dot{y}^{\,2(\phi-1)}}\Big]^{-(n-1)/2},$$

using $\prod_i y_i^{\phi-1} = \dot{y}^{\,n(\phi-1)}$. Under $y \to cy$ the numerator of the bracket picks up $c^{2\phi}$ and the denominator $c^{2(\phi-1)}$, so the whole bracket picks up c^2, independent of ϕ; hence

$$p(\phi \mid cy) = p(\phi \mid y) \quad \text{for every } c > 0.$$

Equivalently, the model in (b) is the flat-prior model applied to the **normalized** transformation $z_i(\phi) = y_i^{(\phi)}/\dot{y}^{\,\phi-1}$, whose units are those of y for every ϕ. In the example of (a) the posterior mode is $\hat{\phi} = +0.21$ for $c = 0.1, 1, 10, 100$ alike.

(c) From the display in (b),

$$p(\phi \mid y) \;\propto\; p(\phi)\,\dot{y}^{\,(n-1)(\phi-1)}\Big[\sum_{i=1}^{n} \big(y_i^{(\phi)} - \overline{y^{(\phi)}}\big)^2\Big]^{-(n-1)/2},$$

the omitted constant $(2\pi)^{-(n-1)/2} n^{-1/2} 2^{(n-3)/2}\Gamma(\frac{n-1}{2})$ depending only on n. The conditional posterior completing the model is the usual one: $\sigma^2 \mid \phi, y \sim \text{Inv-}\chi^2(n-1, s^2_{(\phi)})$ and $\mu \mid \sigma, \phi, y \sim \mathrm{N}(\overline{y^{(\phi)}}, \sigma^2/n)$, with $s^2_{(\phi)}$ the sample variance of the $y_i^{(\phi)}$.

(d) The dependence is on $\dot{y}$ only, and it is a change of units, not a change of belief. Because μ and σ are parameters of the **transformed** data, a prior on them can only be stated once a scale for $y^{(\phi)}$ is fixed, and that scale varies with ϕ; $\dot{y}$ supplies one that is common to all ϕ. Formally the rule is not a Bayesian prior — it violates the likelihood principle, two data sets with proportional likelihoods in ϕ but different geometric means get different answers, and the "prior" cannot be stated before seeing y — but it is an excellent approximation to the honest procedure, which is to put a genuine prior on (μ, σ) in fixed physical units of y and let it be vague on the scale of the observed data. Since $\dot{y}$ is a consistent estimate of that scale, using it changes the answer by $O(n^{-1/2})$ relative to the honest prior, while restoring exact invariance.

(e) The transformation (7.19) has bounded range: for $\phi > 0$, $y_i^{(\phi)} > -1/\phi$, and for $\phi < 0$, $y_i^{(\phi)} < 1/|\phi|$; only $\phi = 0$ gives the whole line. So the normal model is at best an approximation and the correct likelihood carries a truncation factor, for $\phi > 0$,

$$p(y \mid \mu, \sigma, \phi) = \Big[\Phi\Big(\tfrac{\mu + 1/\phi}{\sigma}\Big)\Big]^{-n} \prod_i \mathrm{N}\big(y_i^{(\phi)} \mid \mu, \sigma^2\big) \prod_i y_i^{\phi-1},$$

which is ≥ 1 and equals 1 to high accuracy whenever $\mu + 1/\phi \gg \sigma$. Three consequences. First, the uncorrected analysis is safe exactly when the fitted transformation works well — a good normal fit means few standard deviations of room below the boundary — and is misleading precisely in the

regions of (μ, σ, ϕ) where the normal approximation is poor, which are the regions a model check should be flagging. Second, the uncorrected likelihood is not a density in y, so posterior predictive draws $\tilde{y}^{(\phi)} < -1/\phi$ occur and correspond to no real $\tilde{y}$; predictive inference must either truncate or admit that $\tilde{y}$ is undefined with some probability. Third, and the reason this is more than a technicality, no $\phi \neq 0$ model can be literally true for all-positive data with support down to 0, so the family should be read as a device for finding a good approximate scale, with ϕ reported as an interval rather than selected and then conditioned on.

Problem 7.6. Fitting a power-transformed normal model: Table 7.3 gives short-term radon measurements for a sample of houses in three counties in Minnesota (see Section 9.4 for more on this example). For this problem, ignore the first-floor measurements (those indicated with asterisks in the table).

Table 7.3, short-term measurements of radon concentration (in picoCuries/liter); all measurements were recorded on the basement level of the houses, except for those marked *, which were recorded on the first floor:

County	Radon measurements (pCi/L)
Blue Earth	5.0, 13.0, 7.2, 6.8, 12.8, 5.8*, 9.5, 6.0, 3.8, 14.3*, 1.8, 6.9, 4.7, 9.5
Clay	0.9*, 12.9, 2.6, 3.5*, 26.6, 1.5, 13.0, 8.8, 19.5, 2.5*, 9.0, 13.1, 3.6, 6.9*
Goodhue	14.3, 6.9*, 7.6, 9.8*, 2.6, 43.5, 4.9, 3.5, 4.8, 5.6, 3.5, 3.9, 6.7

(a) Fit the power-transformed normal model from Exercise 7.5(b) to the basement measurements in Blue Earth County.

(b) Fit the power-transformed normal model to the basement measurements in all three counties, holding the parameter ϕ equal for all three counties but allowing the mean and variance of the normal distribution to vary.

(c) Check the fit of the model using posterior predictive simulations.

(d) Discuss whether it would be appropriate to simply fit a lognormal model to these data.

Solution. Deleting the asterisked values leaves $n = 12$, 10, 11 basement measurements:

County	n	Basement measurements	median	max
Blue Earth	12	5.0, 13.0, 7.2, 6.8, 12.8, 9.5, 6.0, 3.8, 1.8, 6.9, 4.7, 9.5	6.9	13.0
Clay	10	12.9, 2.6, 26.6, 1.5, 13.0, 8.8, 19.5, 9.0, 13.1, 3.6	11.0	26.6
Goodhue	11	14.3, 7.6, 2.6, 43.5, 4.9, 3.5, 4.8, 5.6, 3.5, 3.9, 6.7	4.9	43.5

(a) Blue Earth alone gives posterior mode $\hat{\phi} = 0.59$ with 95% interval $[-0.37, 1.81]$ — the data cannot distinguish a log transformation from none at all. With $C = 1$ group, Exercise 7.5(c) gives

$$p(\phi \mid y) \propto \dot{y}^{\,(n-1)(\phi-1)} \Big[\textstyle\sum_i \big(y_i^{(\phi)} - \overline{y^{(\phi)}}\big)^2\Big]^{-(n-1)/2},$$

evaluated on a grid for ϕ with $p(\phi) \propto 1$; here $\dot{y} = 6.42$. Posterior summaries: median 0.63, quartiles $[0.28, 1.01]$, $\Pr(\phi < 0 \mid y) = 0.11$. The posterior density at $\phi = 0$ (lognormal) is 0.51 of its maximum and at $\phi = 1$ (normal) is 0.75 of its maximum, so with $n = 12$ mildly skewed observations neither special case is contradicted.

(b) With $C = 3$ counties, $N = 33$ observations, common ϕ and free $(\mu_c, \log \sigma_c)$, the invariance argument of Exercise 7.5(b) forces the prior $p(\{\mu_c\}, \{\log \sigma_c\}, \phi) \propto \dot{y}^{\,C(1-\phi)} p(\phi)$, with $\dot{y}$ the geometric mean of all N observations, and integrating out the six nuisance parameters gives

$$p(\phi \mid y) \;\propto\; p(\phi)\, \dot{y}^{\,(N-C)(\phi-1)} \prod_{c=1}^{3} \Big[\sum_{i \in c} \big(y_i^{(\phi)} - \overline{y_c^{(\phi)}}\big)^2\Big]^{-(n_c-1)/2}.$$

(Each group contributes $c^{-(n_c-1)\phi}$ to the scaling and the Jacobian contributes $c^{N(\phi-1)}$, so the compensating exponent is $C(1-\phi)$, not $1-\phi$; check the case $C = 1$.) The result is $\hat{\phi} = -0.03$, 95% interval $[-0.51, 0.44]$: pooling the three counties pins ϕ down to roughly a factor of two in the exponent and puts it essentially at the logarithm. Given ϕ, $\sigma_c^2 \mid \phi, y \sim \text{Inv-}\chi^2(n_c - 1, s^2_{c(\phi)})$ and $\mu_c \mid \sigma_c, \phi, y \sim \text{N}(\overline{y_c^{(\phi)}}, \sigma_c^2/n_c)$; back-transforming μ_c gives the posterior median radon level in each county,

County	median of y, posterior median	95% interval
Blue Earth	6.4	[4.4, 9.2]
Clay	8.1	[4.0, 16.2]
Goodhue	6.1	[3.6, 10.9]

(Posterior median σ_c: 0.55, 0.92, 0.78; since $\hat{\phi} \approx 0$ these read as standard deviations on the log scale.) The disagreement among counties fitted separately is worth recording, because it is what the common-ϕ assumption is buying: $\hat{\phi} = 0.59$ (Blue Earth), 0.42 (Clay), -0.84 (Goodhue). Goodhue is pulled down by its single value of 43.5.

(c) Simulate ϕ from the grid posterior, then (σ_c^2, μ_c) from the conditionals above, then y_c^{rep} of the same size as y_c, and back-transform. With $S = 20{,}000$ replications, $\Pr\big(T(y^{\text{rep}}) \geq T(y) \mid y\big)$ is

Test quantity	Blue Earth	Clay	Goodhue
$\max_i y_i$	0.71	0.64	0.18
$\min_i y_i$	0.78	0.64	0.25
skewness of $y_i^{(\phi)}$	0.91	0.85	0.02

Only Goodhue fails, and it fails on skewness ($p = 0.02$): after the common transformation its 11 values are more right-skewed than replications allow, because 43.5 sits far above the rest. The maximum alone does not flag it ($p = 0.18$), since the fitted σ_3 is inflated by that very point — the familiar effect of checking an outlier with a statistic that the outlier itself calibrates.

One further check comes free. In 5.8% of the simulated replicate data sets (0.6%, 3.8%, 2.2% by county) at least one value of $y^{\text{rep},(\phi)}$ fell outside the range of the transformation ($y^{(\phi)} > 1/|\phi|$ for the negative ϕ draws, which are 56% of the posterior), so it corresponds to no real measurement. That is the implicit truncation of Exercise 7.5(e) showing up at a rate too large to ignore.

(d) Yes, for these data. The lognormal is $\phi = 0$, which is the posterior median of the common-ϕ fit ($\hat{\phi} = -0.03$), lies well inside the 95% interval $[-0.51, 0.44]$, and has posterior density 0.99 of the maximum; nothing is lost by setting $\phi = 0$ and much is gained — the parameters (μ_c, σ_c) become interpretable as the mean and standard deviation of log radon and are comparable across counties and across data sets, the truncation problem of part (c) disappears entirely since the lognormal has support $(0, \infty)$, and the three-county comparison becomes a standard normal-theory problem. The estimated transformation should be reported as the justification for the lognormal, not suppressed: the width of the interval for ϕ is a statement that these 33 observations could not have detected a moderate departure from lognormality, and the Goodhue skewness check says the one real defect — a heavy upper tail in one county — is not repaired by any choice of ϕ and calls instead for a longer-tailed model or for a hierarchical treatment of the counties as in Section 9.4.

Problem 7.7. Model expansion: consider the t model, $y_i \mid \mu, \sigma^2, \nu \sim t_\nu(\mu, \sigma^2)$, as a generalization of the normal. Suppose that, conditional on ν, you are willing to assign a noninformative uniform prior density on $(\mu, \log\sigma)$. Construct what you consider a noninformative joint prior density on $(\mu, \log\sigma, \nu)$, for the range $\nu \in [1, \infty)$. Address the issues raised in setting up a prior distribution for the power-transformed normal model in Exercise 7.5.

Solution. Take

$$p(\mu, \log\sigma, \nu) \;\propto\; \nu^{-2}, \qquad \nu \in [1, \infty),$$

that is, uniform on $\nu^{-1} \in (0, 1]$ with $\mu, \log\sigma$ flat given ν. It is proper in ν (total mass $\int_1^\infty \nu^{-2} d\nu = 1$) and improper only in the location and scale, where impropriety is harmless. Each of the four difficulties of Exercise 7.5 is either absent or is what dictates the choice.

Invariance under change of units (7.5a,b). Absent. For every fixed ν, $t_\nu(\mu, \sigma^2)$ is a location–scale family, so under $y \to a + cy$ the likelihood is exactly reproduced by $(\mu, \sigma, \nu) \to (a + c\mu,\, c\sigma,\, \nu)$, and $p(\mu, \log\sigma) \propto 1$ is invariant under that map. There is no Jacobian mismatch between different values of the expansion parameter — μ and σ mean the same thing for every ν, whereas in (7.19) they refer to the $y^{(\phi)}$ scale, whose units depend on ϕ. Hence no data-dependent factor $\dot{y}^{\,1-\phi}$ is needed and none should be used.

Propriety (the real constraint). As $\nu \to \infty$, $t_\nu(\mu, \sigma^2) \to \mathrm{N}(\mu, \sigma^2)$, so the integrated likelihood

$$p(y \mid \nu) = \iint \prod_i t_\nu(y_i \mid \mu, \sigma^2)\, d\mu\, \frac{d\sigma}{\sigma}$$

converges to $p_{\mathrm{N}}(y) > 0$ rather than to 0 (proper for $n \geq 2$ with at least two distinct y_i, as for the normal). Therefore

$$\int_1^\infty p(y \mid \nu)\, d\nu = \infty,$$

and a flat prior on $\nu \in [1, \infty)$ gives an improper posterior — the exact analogue of the divergent tail found in Exercise 7.5(a), and the reason "uniform on ν" is not an option. Any admissible $p(\nu)$ must be integrable at ∞; ν^{-2} is.

Which "uniform" (7.5's parameterization problem). Choose the parameterization in which the likelihood is regular at the limit point. The t_ν density is smooth in ν^{-1} at $\nu^{-1} = 0$, with

$$p(y \mid \nu) = p_{\mathrm{N}}(y)\Big(1 + \frac{\kappa(y)}{\nu} + O(\nu^{-2})\Big),$$

so ν^{-1} is the parameter on whose scale the data carry approximately constant information near the normal, and to leading order it matches the excess kurtosis $6/(\nu - 4)$ that the expansion is actually buying. A numerical illustration on $n = 20$ observations

$$\begin{aligned} y = (&1.15, -0.48, 2.31, 0.02, -1.77, 0.63, -0.21, 4.90, \\ &0.35, -0.94, 0.77, -2.55, 1.02, 0.18, -0.33, 0.44, \\ &-1.21, 0.91, -0.07, 2.02) \end{aligned}$$

gives $\log p(y \mid \nu) - \log p_{\mathrm{N}}(y) = 1.47,\ 0.82,\ 0.47,\ 0.103,\ 0.0209,\ 0.00525$ at $\nu = 3, 10, 20, 100, 500, 2000$: exactly the κ/ν behaviour, with $\kappa = 10.3$. Under the proposed prior the posterior of ν^{-1} is well behaved, with median 0.45 and 95% interval $[0.06, 0.95]$, i.e. $\nu \in [1.1, 17.2]$; under a flat prior on ν no such summary exists.

Data dependence (7.5d). None: ν^{-2} involves no function of y. The device forced on Box and Cox by the units problem is unnecessary here precisely because of the invariance noted above, so the prior can be stated, and criticized, before the data arrive.

Truncation (7.5e). None: $t_\nu(\mu, \sigma^2)$ has support $\mathbb{R}$ for every $\nu \geq 1$, so unlike (7.19) the model never places probability on impossible values, the likelihood needs no normalizing correction, and every posterior predictive draw is a legitimate observation. If the data are all-positive, that is a reason to model $\log y$ with the t, not a reason to truncate.

To admit the normal itself, extend to $\nu^{-1} \in [0, 1]$, which this prior does continuously: in ν^{-1} the null model is a boundary point at finite distance rather than at $\nu = \infty$.

8 Modeling Accounting for Data Collection

Contents

8.1 Exercises 8.1–8.7

Problem 8.1. Definition of concepts: the concepts of randomization, exchangeability, and ignorability have often been confused in the statistical literature. For each of the following statements, explain why it is false but also explain why it has a kernel of truth. Illustrate with examples from this chapter or earlier chapters.
(a) Randomization implies exchangeability: that is, if a randomized design is used, an exchangeable model is appropriate for the observed data, $y_{\text{obs}\,1}, \ldots, y_{\text{obs}\,n}$.
(b) Randomization is required for exchangeability: that is, an exchangeable model for $y_{\text{obs}\,1}, \ldots, y_{\text{obs}\,n}$ is appropriate only for data that were collected in a randomized fashion.
(c) Randomization implies ignorability; that is, if a randomized design is used, then it is ignorable.
(d) Randomization is required for ignorability; that is, randomized designs are the only designs that are ignorable.
(e) Ignorability implies exchangeability; that is, if an ignorable design is used, then an exchangeable model is appropriate for the observed data, $y_{\text{obs}\,1}, \ldots, y_{\text{obs}\,n}$.
(f) Ignorability is required for exchangeability; that is, an exchangeable model for the vector $y_{\text{obs}\,1}, \ldots, y_{\text{obs}\,n}$ is appropriate only for data that were collected using an ignorable design.

Solution. All six are false; the three concepts live in three different places. Randomization is a property of $p(I \mid x, y, \phi)$, ignorability is a property of the **pair** $(p(y \mid x, \omega),\, p(I \mid x, y, \phi))$ through (8.2) versus (8.3), and exchangeability is a property of $p(y \mid x)$ alone. Nothing about one forces anything about another.
(a) False. Randomization makes $p(I)$ invariant under permutation of the unit labels; it says nothing about $p(y)$. In the Latin square of Table 8.4 the treatments are randomized, yet the 25 observed yields are not exchangeable: the plot coordinates x distinguish the units, and the whole point of the analysis on p. 216 is to model y given x. Worse, even with no covariates the observed values $y_{\text{obs}\,i}$ are draws under **different** treatments, so they are exchangeable only after conditioning on the treatment received.
Kernel of truth: under complete randomization with no recorded covariates, the units are exchangeable given the treatment indicator, and randomization is what licenses treating the treatment-group labels as carrying no information beyond the treatment itself.
(b) False. Exchangeability is a judgement about y, not a consequence of data collection. The 71 rat tumor experiments of Section 5.1 and the eight coaching studies of Section 5.5 are modelled exchangeably at the study level, and no randomization whatsoever selected those studies; Simon Newcomb's speed-of-light measurements (Section 3.2) are modelled as exchangeable with no randomization at all.
Kernel of truth: randomization is a **sufficient** device for exchangeability. When the units genuinely carry distinguishing information and we do not wish to model it, randomizing over the labels is the one construction that makes the symmetric model defensible (p. 218).
(c) False. Ignorability requires missing at random **and** distinct parameters, and randomization delivers only the first. The book's own counterexample is on p. 220: an experimenter who uses complete randomization when he believes the treatment effects are large and randomized blocks when he suspects they are small. The choice of design is then a function of his prior for ω, so $p(\phi \mid x, \omega) \neq p(\phi \mid x)$, the parameters are not distinct, and the design is not ignorable even though every assignment was randomized. Exercise 8.3 is the same phenomenon in the milk-production experiment.
Kernel of truth: randomization guarantees $p(I \mid x, y, \phi) = p(I \mid x, \phi)$, which is free of y_{mis}; missing at random is the hard condition, and distinct parameters holds in almost every practical randomized experiment.
(d) False. Any design whose inclusion pattern depends only on fully observed covariates is ignorable. Auditing every tax return with declared income above \$1 million (p. 202) is deterministic and ignorable, because declared income is a fully observed x. The systematic field layout ABABABBABABA on twelve plots (p. 219) is deterministic and ignorable given plot location. The milk-production design of Section 8.4 is neither randomized in any documented way nor known, and it is still ignorable given the three pre-treatment covariates.

Kernel of truth: the italicized claim on p. 218 is nearly this statement with the covariates removed. If there are no fully observed x to condition on, then apart from degenerate designs the only $p(I_1, \ldots, I_n \mid \phi)$ that is both ignorable and invariant to permutation of the indexes is a randomization. Randomization also buys ignorability without our having to know, record, or model which covariates drove the assignment.

(e) False. Stratified sampling is the standing counterexample. The CBS survey of Table 8.2 is ignorable given the sixteen stratum indicators, yet the 1447 observed responses are plainly not exchangeable: the response distributions differ by stratum, and the analysis of Section 8.3 models sixteen distinct parameter vectors $(\omega_{1j}, \omega_{2j}, \omega_{3j})$. Likewise the audited tax returns are systematically the largest ones in the population.

Kernel of truth: a strongly ignorable design is implicitly randomized over every variable not in x (p. 220), so an ignorable design given x does deliver exchangeability of the units **conditional on** x: units i and j with $x_i = x_j$ have $p(I_i \mid x) = p(I_j \mid x)$ and may be modelled exchangeably. Exchangeability returns within cells, not across them.

(f) False. Censoring (Section 8.7) is the clean case: the underlying complete data $y_1, \ldots, y_N$ may be modelled as independent and identically distributed given ω, hence fully exchangeable, while $p(I \mid y, \phi)$ depends on y_{mis} and the design is nonignorable. The exchangeable model for y is not merely appropriate there, it is the model one actually fits; what the nonignorability forbids is dropping the factor $p(I \mid y, \phi)$ from the likelihood.

Kernel of truth: ignorability is exactly what transfers an exchangeable model for the complete data y to a usable exchangeable model for the **observed** data y_{obs}. Under a nonignorable design the observed values are a selected, hence distorted, subset, and applying $p(y \mid \omega)$ directly to y_{obs} as in (8.3) is wrong.

Problem 8.2. Application of design issues: choose an example from earlier in this book and discuss the relevance of the material in the current chapter to the analysis. In what way would you change the analysis, if at all, given what you have learned from the current chapter?

Solution. Take the eight-schools SAT coaching analysis of Section 5.5. Two design issues are relevant, at the two levels of the hierarchy, and only the second changes the analysis.

Level one, within schools: the students were randomly assigned to coaching and control within each school, so the assignment is strongly ignorable and known, with $p(I \mid x, y, \phi) = p(I \mid x)$, x being the school indicator. This is exactly the completely randomized experiment of Section 8.4 carried out eight times in parallel, so the school-level estimates y_j and their standard errors σ_j may be taken at face value and the posterior computed from $p(\omega \mid x, y_{\text{obs}}) \propto p(\omega)\, p(y_{\text{obs}} \mid x, \omega)$. Chapter 8 also sharpens what θ_j means: it is a superpopulation parameter, whereas the finite-population causal effect $\bar{y}_j^A - \bar{y}_j^B$ for the students actually in school j is the estimand of (8.17), and for large n_j the two coincide with θ_j to order $n_j^{-1/2}$.

Level two, selection of the schools: here the design is **not** ignorable and Section 5.5 tacitly assumes it is. The eight schools are the ones that ran a coaching program, evaluated it, and reported the result; nothing randomized their entry into the dataset. If the probability of a study reaching the analyst rises with the estimated effect, then $p(I_j \mid \theta_j, y_j, \phi)$ depends on y_j, which is the censoring structure of Section 8.7 and is nonignorable and unknown. This is the one place where Chapter 8 says the Section 5.5 posterior is wrong rather than merely incomplete: an exchangeable $\theta_j \sim \text{N}(\mu, \tau^2)$ fitted to a selected upper tail overstates μ, and, because selection truncates from below, understates τ.

Three concrete changes.

(i) Use the available covariates rather than the minimally adequate summary (p. 217). The sample sizes n_j, equivalently the σ_j, are fully observed and were chosen by the investigators, so the principle that everything used in the design belongs in the model applies. Replace the second level by

$$\theta_j \mid \alpha, \beta, \tau \sim \text{N}\big(\alpha + \beta\,(\log n_j - \overline{\log n}),\ \tau^2\big)\,,$$

which is the exact analogue of including M_k in the school-level model (8.10) for the Melbourne cluster sample, and check whether β is distinguishable from zero. The model of Section 5.5 is the special case $\beta = 0$.

(ii) Run the selection mechanism as a sensitivity analysis rather than assuming it away: attach a weight $\Pr(I_j = 1 \mid y_j, \phi) = \text{logit}^{-1}(\phi_0 + \phi_1 y_j/\sigma_j)$ and trace the posterior for μ as ϕ_1 ranges over plausible values. With eight studies, τ already poorly determined, and a posterior mean effect of about 8 points, even mild selection moves the estimate appreciably, and reporting that range is more honest than reporting the ignorable answer alone.
(iii) Widen the posterior predictive checks of Section 6.5: because the within-school assignment was randomized, replications may redraw I^{rep} as well as y^{rep} (p. 219).

Problem 8.3. Distinct parameters and ignorability:
(a) For the milk production experiment in Section 8.4, give an argument for why the parameters ϕ and ω may not be distinct.
(b) If the parameters are not distinct in this example, the design is no longer ignorable. Discuss how posterior inferences would be affected by using an appropriate nonignorable model. (You need not set up the model; just discuss the direction and magnitude of the changes in the posterior inferences for the treatment effect.)
(The experiment of Section 8.4: fifty cows were assigned to four diets differing in the level of a feed additive, methionine hydroxy analog, with six post-treatment measures of milk fat recorded per cow. Three covariates were recorded before assignment: lactation number, age, and initial weight. Cows were first assigned completely at random; the balance of the three covariates across treatment groups was then inspected, several randomizations were tried, and the assignment giving the best balance was used. Thus x is 50×3, the complete data y are 50×24, and one subvector of length 6 is observed per cow.)

Solution. (a) Because the re-randomization rule is the experimenter's judgement, and that judgement was formed from his beliefs about ω. He balanced on lactation number, age, and initial weight, and on nothing else, precisely because he believed those three variables predict milk fat; had he believed the coefficients of x in $p(y \mid x, \omega)$ were zero, one randomization would have been enough and no balance criterion would have been applied. The unstated rule, that is, the implicit weighting of the three covariates in the balance criterion and the number of randomizations tried before stopping, therefore carries information about the magnitudes of those coefficients, so

$$p(\phi \mid x, \omega) \;\neq\; p(\phi \mid x),$$

which is the negation of the distinct-parameters condition on p. 202. The p. 220 mechanism is available too: an experimenter searches harder for balance when he expects the treatment effect to be small relative to covariate-induced variation, so the effort ϕ encodes his prior for the treatment effect itself. Note that missing at random is untouched: the assignment used only the fully observed x, never the unrecorded outcomes, so the failure of ignorability here is entirely a failure of distinctness.
(b) The effect is a mild extra prior factor on ω, and it is negligible here. Writing ϕ^{obs} for what the design reveals, the nonignorable posterior is proportional to

$$\begin{aligned} p(\omega \mid x, y_{\text{obs}}, I) \;\propto\; & p(\omega \mid x)\; p(y_{\text{obs}} \mid x, \omega) \\ & \times \int p(\phi \mid x, \omega)\, p(I \mid x, y_{\text{obs}}, \phi)\, d\phi, \end{aligned}$$

and the ignorable analysis simply drops the last line. That line is a smooth, diffuse function of ω: it says roughly that the coefficients on lactation number, age, and weight are not all negligible, and, under the p. 220 variant, that the treatment effect was not expected to be enormous.
Direction. The balance criterion is symmetric in the sign of the treatment effect, so nothing shifts the location of the posterior for the treatment contrasts in a determinate direction. What the factor does favour is larger $|\beta_x|$, which would slightly reduce the posterior for the residual variance σ^2 and hence **sharpen**, not shift, the treatment-effect posterior. Under the reading in which he searched harder for balance because he expected a small effect, the factor is a weak shrinkage of the treatment contrast toward zero.

Magnitude. Negligible, for two independent reasons. First, the extra factor is vague compared with a likelihood based on fifty cows; it is at most worth a few pseudo-observations, whereas the data determine the regression coefficients to within a standard error that the factor cannot meaningfully compete with. Second, and more decisively, the re-randomization achieved **balance**, which is the situation of Figure 8.2a: with the covariate distributions matched across the four diets, the estimated treatment effect is insensitive to the form of $p(y \mid x, \omega)$, and a fortiori to a perturbation of the prior on its coefficients. This is why the text can say (p. 218) that reasonable violations of distinctness should not have large effects on the inferences for this problem. The conclusion fails only in the case the same page warns about: if the assignment rule had depended on beliefs about the **treatment efficacy** strongly enough to make ϕ nearly determine ω, the extra factor would no longer be vague and would have to be modelled.

Problem 8.4. Interaction between units: consider a hypothetical agricultural experiment in which each of two fertilizers is assigned to 10 plots chosen completely at random from a linear array of 20 plots, and the outcome is the average yield of the crops in each plot. Suppose there is interference between units, because each fertilizer leaches somewhat onto the two neighboring plots.
(a) Set up a model of potential data y, observed data y_{obs}, and inclusion indicators I. The potential data structure will have to be larger than a 20×4 matrix in order to account for the interference.
(b) Is the treatment assignment ignorable under this notation?
(c) Suppose the estimand of interest is the average difference in yields under the two treatments. Define the finite-population estimand mathematically in terms of y.
(d) Set up a probability model for y.

Solution. (a) Index the potential outcomes of plot i by the treatments on the **triple** $(i-1, i, i+1)$, which restores the stability assumption of p. 200 by enlarging y until I no longer changes it. Let $T = (T_1, \dots, T_{20}) \in \{A, B\}^{20}$ be the assignment, with $\sum_i \mathbb{1}(T_i = B) = 10$, and set

$$y \;=\; \big(y_{i,(a,b,c)}\big), \qquad i = 1, \dots, 20,\ (a,b,c) \in \{A,B\}^3,$$

where $y_{i,(a,b,c)}$ is the yield of plot i when plot $i-1$ gets a, plot i gets b, and plot $i+1$ gets c. This is a 20×8 array, larger than 20×4 as required. The edge plots have no outside neighbour, so row 1 depends only on (b, c) and row 20 only on (a, b); those two rows carry 4 distinct values each, and y has $18 \cdot 8 + 2 \cdot 4 = 152$ distinct entries.
The indicator array is $I = (I_{i,(a,b,c)})$ of the same 20×8 shape,

$$I_{i,(a,b,c)} \;=\; \mathbb{1}\big\{(T_{i-1}, T_i, T_{i+1}) = (a,b,c)\big\},$$

with exactly one 1 per row, so y_{obs} consists of 20 numbers and y_{mis} of the remaining 132. Note I and T determine each other: the middle index of the observed cell in row i is T_i. The inclusion model is

$$p(I \mid y, \phi) \;=\; p(I) \;=\; \binom{20}{10}^{-1} = \frac{1}{184756}$$

on the 184756 arrays generated by a balanced T, and 0 elsewhere.
(b) Yes, and in the strongest sense: $p(I \mid y, \phi) = p(I)$ is exactly (8.4), so the design is known and strongly ignorable, with no covariates needed and constant propensity score $\frac{1}{2}$. What the interference cost us is not ignorability but the size of y_{mis}: only 20 of 152 potential values are seen, against 20 of 40 in the no-interference 20×2 layout. Had we kept the naive structure, the difficulty would not have been a nonignorable design; the notation itself would have been incoherent, since y would depend on I.
(c) With interference, a treatment label alone no longer names a state of the field, so the estimand must compare two whole assignment vectors, and the natural pair is the all-B field against the all-A field:

$$\bar{y}_B - \bar{y}_A \;=\; \frac{1}{20} \sum_{i=1}^{20} \Big(y_{i,(B,B,B)} - y_{i,(A,A,A)} \Big),$$

reading $y_{1,(A,A,A)}$ as $y_{1,(\cdot,A,A)}$ and $y_{20,(A,A,A)}$ as $y_{20,(A,A,\cdot)}$. Of the 40 quantities in this sum, only those few plots whose realized triple happened to be AAA or BBB are observed; the rest are imputed from the model, which is why (d) matters more here than in a stability-satisfying experiment.
(d) Write $z_a = \mathbb{1}(a = B)$ and model an own-plot effect β and a per-neighbour leaching effect γ additively on top of a plot fertility term:

$$y_{i,(a,b,c)} \;=\; \mu + \delta_i + \beta z_b + \gamma\,(z_a + z_c) + \epsilon_{i,(a,b,c)},$$

with the edge rows omitting the absent neighbour's γz term. For the spatial fertility of a linear array take a first-order autoregression,

$$\delta_1 \sim \mathrm{N}\Big(0, \tfrac{\tau^2}{1-\rho^2}\Big), \qquad \delta_i = \rho\,\delta_{i-1} + \eta_i, \quad \eta_i \sim \mathrm{N}(0, \tau^2),$$

and let the 8-vector $\epsilon_i = (\epsilon_{i,(a,b,c)})$ be multivariate normal with mean 0 and exchangeable covariance

$$\mathrm{Cov}(\epsilon_{i,u}, \epsilon_{i,v}) = \sigma^2\big(\lambda + (1-\lambda)\mathbb{1}_{u=v}\big),$$

independent across i. Complete with $p(\mu, \beta, \gamma, \rho, \log\sigma, \log\tau) \propto 1$ on $|\rho| < 1$ and $\lambda \sim \mathrm{U}(0,1)$.
Two features of this model earn it its place. The design supplies plots with zero, one, and two treated neighbours, so β and γ are separately identified from the 20 observations. And the estimand of (c) reduces under the model to

$$\bar{y}_B - \bar{y}_A \;=\; \beta + 1.9\,\gamma + \frac{1}{20}\sum_{i=1}^{20}\big(\epsilon_{i,(B,B,B)} - \epsilon_{i,(A,A,A)}\big),$$

the coefficient $1.9 = (18 \cdot 2 + 2 \cdot 1)/20$ coming from the two edge plots, with μ and every δ_i cancelling. The within-row correlation λ never enters the likelihood, exactly as ω_{AB} does not on p. 215; it must be given a prior, it affects only the finite-population estimand through the residual term above, and its influence vanishes as the number of plots grows.

Problem 8.5. Analyzing a designed experiment: Table 8.6 displays the results of a randomized blocks experiment on penicillin production. Four manufacturing processes (treatments A, B, C, D) were each applied in five different conditions (blocks); four runs were made within each block, with the treatments assigned to the runs at random. The data, from Box, Hunter, and Hunter (1978), who adjusted them so that the averages are integers, are:

Block	A	B	C	D
1	89	88	97	94
2	84	77	92	79
3	81	87	87	85
4	87	92	89	84
5	79	81	80	88

(a) Express this experiment in the general notation of this chapter, specifying x, y_{obs}, y_{mis}, N, and I. Sketch the table of units by measurements. How many observed measurements and how many unobserved measurements are there in this problem?
(b) Under the randomized blocks design, what is the distribution of I? Is it ignorable? Is it known? Is it strongly ignorable? Are the propensity scores an adequate summary?
(c) Set up a normal-based model of the data and all relevant parameters that is conditional on enough information for the design to be ignorable.
(d) Suppose one is interested in the (superpopulation) average yields of penicillin, averaging over the block conditions, under each of the four treatments. Express this estimand in terms of the parameters in your model.
We return to this example in Exercise 15.2.

Solution. (a) The units are the $N = 20$ runs, not the five blocks. Label a run by the pair $i = (j, r)$ with block $j = 1, \dots, 5$ and $r = 1, \dots, 4$, and let y_{ik} be the yield run i would give under treatment $k \in \{A, B, C, D\}$; then y is a 20×4 array of potential outcomes and I is a 20×4 array of zeros and ones with exactly one 1 in each row. The fully observed covariate is the block label, $x = (j(i))_{i=1}^{20}$, equivalently a 20×5 matrix of stratum indicators. Since Table 8.6 reports yields by treatment rather than by run, we may without loss number the runs within each block so that run (j, r) received treatment r, making I the identity pattern within every block:

Unit	A	B	C	D
(1,1)	89	.	.	.
(1,2)	.	88	.	.
(1,3)	.	.	97	.
(1,4)	.	.	.	94
(2,1)	84	.	.	.

and so on through unit $(5, 4)$. There are 20 observed measurements and $20 \cdot 4 - 20 = 60$ unobserved ones, so three quarters of the potential data are missing, the usual state of affairs in an experiment (Table 8.1).

(b) Within each block the four treatments are allotted to the four runs by a uniform random permutation, independently across blocks:

$$p(I \mid x, y, \phi) = p(I \mid x) = \left(\frac{1}{4!}\right)^5 = \frac{1}{7962624}$$

on the set of I that place exactly one 1 in each row and exactly one of each treatment in each block, and 0 elsewhere.

Ignorable: yes. The expression is free of y outright, so missing at random holds, and there is no parameter ϕ at all, so distinct parameters holds vacuously. Known: yes, the display above is the complete specification. Strongly ignorable: yes, $p(I \mid x, y, \phi) = p(I \mid x)$ with x fully observed.

Propensity scores: no, they are not an adequate summary. Here

$$\pi_{ik} = \Pr(I_{ik} = 1 \mid x) = \tfrac{1}{4} \quad \text{for every } i \text{ and } k,$$

a constant, so conditioning on π is conditioning on nothing and discards the block labels entirely. The design is **not** strongly ignorable given π alone: $p(I \mid x)$ above is supported on the block-balanced arrays and is not a function of π. These same propensity scores would arise from complete randomization and from independent assignment with probability $\frac{1}{4}$ per run, which is exactly the phenomenon noted on p. 204, and it is why π supports neither the precision of part (c) nor a correct posterior predictive replication.

(c) Condition on the block indicators and nothing less. On the yield scale,

$$y_{ik} \mid \omega \;\sim\; \mathrm{N}\big(\mu + \beta_{j(i)} + \tau_k,\; \sigma^2\big)\,,$$

independently across runs i, with the block effects exchangeable,

$$\beta_j \mid \tau_\beta \;\sim\; \mathrm{N}(0, \tau_\beta^2), \qquad j = 1, \dots, 5,$$

treatment effects constrained by $\sum_k \tau_k = 0$, and $p(\mu, \tau, \log\sigma, \tau_\beta) \propto 1$. Because x enters $p(y \mid x, \omega)$, part (b) makes the design ignorable and the posterior is simply $p(\omega \mid x, y_{\text{obs}}) \propto p(\omega)\, p(y_{\text{obs}} \mid x, \omega)$, with

$$p(y_{\text{obs}} \mid x, \omega) = \prod_{i=1}^{20} \prod_{k} \left[p(y_{ik} \mid x, \omega)\right]^{I_{ik}}.$$

For finite-population estimands one further needs the joint law of the four potential outcomes in a row; take $(y_{iA}, \dots, y_{iD})$ multivariate normal with the above means and exchangeable correlation λ, which does not appear in the likelihood and must be assigned a prior. This is the model taken up again in Exercise 15.2.

(d) Averaging the mean over the block distribution $\beta_j \sim \mathrm{N}(0, \tau_\beta^2)$,

$$\theta_k \;=\; \mathrm{E}\big(y_{ik} \mid \omega\big) \;=\; \mu + \tau_k, \qquad k = A, B, C, D,$$

since $\mathrm{E}(\beta_j) = 0$; the pairwise comparisons are $\theta_k - \theta_{k'} = \tau_k - \tau_{k'}$. Had the question asked instead for the average over the five block conditions actually run, the estimand would be $\mu + \bar{\beta} + \tau_k$ with $\bar{\beta} = \frac{1}{5}\sum_{j=1}^{5} \beta_j$.

Problem 8.6. Including additional information beyond the adequate summary:
(a) Suppose that the experiment in the previous exercise had been performed by complete randomization (with each treatment coincidentally appearing once in each block), not randomized blocks. Explain why the appropriate Bayesian modeling and posterior inference would not change.
(b) Describe how the posterior predictive check would change under the assumption of complete randomization.
(c) Why is the randomized blocks design preferable to complete randomization in this problem?
(d) Give an example illustrating why too much blocking can make modeling more difficult and sensitive to assumptions.

Solution. (a) Because the design enters the posterior only through a factor that is the same under both mechanisms. Under complete randomization with five runs per treatment,

$$p(I \mid x, y, \phi) = p(I) = \binom{20}{5, 5, 5, 5}^{-1} = \frac{1}{11732745024},$$

which, like the randomized-blocks $p(I \mid x)$ of Exercise 8.5(b), is free of y_{mis} and carries no parameter, so both designs are ignorable and in (8.2) the factor integrates out, leaving (8.3):

$$p(\omega \mid x, y_{\mathrm{obs}}, I) = p(\omega \mid x, y_{\mathrm{obs}}) \propto p(\omega \mid x)\, p(y_{\mathrm{obs}} \mid x, \omega).$$

The realized (x, y_{obs}, I) are by hypothesis identical, and the prior and complete-data model are unchanged, so the posterior is identical number for number. The block labels stay in the model regardless: they are a fully observed covariate, and the principle of p. 217 is to use all available information and not merely the minimally adequate summary, which under complete randomization would not have included x at all. This is precisely the Latin-square argument of p. 216.
(b) The check changes because the replication redraws I^{rep} from $p(I \mid x, \phi)$, and that distribution is design-specific even though the posterior is not. Under randomized blocks, every I^{rep} is again a within-block permutation, so each replicated dataset has each treatment exactly once per block, each treatment mean is again an average of five observations one per block, and any test statistic measuring treatment-by-block balance is degenerate at its observed value. Under complete randomization, I^{rep} scatters five runs per treatment over all twenty, and balance is destroyed almost surely: the probability that a completely randomized replicate reproduces the balanced pattern is

$$\frac{(4!)^5}{20!/(5!)^4} = \frac{7962624}{11732745024} = 6.8 \times 10^{-4},$$

and the probability that some treatment is missing from some block entirely is 0.999. Three consequences. Test statistics must be defined for unbalanced tables, since $T(y^{\mathrm{rep}}, I^{\mathrm{rep}})$ will almost never be computable on a balanced layout. The reference distribution for treatment contrasts is wider, because the variance factor for $\tau_k - \tau_{k'}$ in the additive model averages $0.551\,\sigma^2$ over completely randomized assignments against a fixed $0.400\,\sigma^2$ under blocking, so the check is less stringent. And a new check becomes available: taking T to be a measure of treatment-by-block imbalance, the observed value sits below essentially every replicate, at the 0.07th percentile, which would correctly signal that the data did not arise from complete randomization.
(c) Because it buys precision and robustness at no cost. Precision: balance makes the treatment contrasts orthogonal to the block effects, giving the smallest achievable variance factor $2/5 = 0.400$

for $\tau_k - \tau_{k'}$; over completely randomized assignments the same factor has median 0.507, mean 0.551, and exceeds 1.39 in the worst one percent (Monte Carlo over 10^6 assignments; the reported digits are stable). With the residual mean square from Table 8.6 equal to 18.8 on 12 degrees of freedom, so $\hat{\sigma} = 4.34$, this is a posterior standard deviation of 2.74 for a treatment contrast under blocking against a typical 3.09 and an occasional 5.11 under complete randomization. The gain is real here because the blocks are genuinely different: the block mean square is 66.0 against a residual mean square of 18.8, and a model that dropped the blocks would raise the residual standard deviation from 4.34 to 5.53.
Robustness: balance is what makes the answer insensitive to the form of the block model, exactly as in Figure 8.2a. Under an unbalanced draw the treatment means must be corrected for block effects through the assumed additive structure, and since with probability 0.999 some treatment is absent from some block, at least one cell of the correction would rest entirely on additivity rather than on data.
(d) Block the twenty runs into ten blocks of two, say consecutive pairs of runs sharing a fermenter charge. Each block now holds two runs, hence at most two of the four treatments, so a block supplies a direct within-block comparison for at most one of the six treatment pairs, and the remaining contrasts are recovered only by chaining blocks through the additive model. Three difficulties follow, each an instance of p. 220, item 2. The parameter count rises: one intercept, nine block contrasts, and three treatment contrasts, leaving $20 - 13 = 7$ residual degrees of freedom instead of 12, so σ is poorly determined and the posterior for the treatment contrasts becomes sensitive to the prior on τ_β, which is now doing the work of holding ten weakly identified block effects together. Additivity becomes uncheckable: with only 7 residual degrees of freedom and at most one treatment pair per block, no block-by-treatment interaction can be estimated or even diagnosed, yet every between-block comparison assumes it away. And identification can fail outright: if the realized pairing happens to put only $\{A, B\}$ and only $\{C, D\}$ in blocks, the design is disconnected and $\tau_A - \tau_C$ has a flat likelihood, the extreme case in which the observed-data likelihood provides no information to separate treatment from block. In the limit of blocks of size one, every treatment effect is confounded with its block effect and the experiment says nothing at all.

Problem 8.7. Simple random sampling:
(a) Derive the exact posterior distribution for $\bar{y}$ under simple random sampling with the normal model and noninformative prior distribution.
(b) Derive the asymptotic result (8.6), namely

$$\bar{y} \mid y_{\text{obs}} \;\approx\; \text{N}\big(\bar{y}_{\text{obs}},\; \big(\tfrac{1}{n} - \tfrac{1}{N}\big)\, s^2_{\text{obs}}\big)\,.$$

Solution. (a) The answer is

$$\bar{y} \mid y_{\text{obs}} \;\sim\; t_{n-1}\big(\bar{y}_{\text{obs}},\; \big(\tfrac{1}{n} - \tfrac{1}{N}\big)\, s^2_{\text{obs}}\big)\,,$$

the result quoted on p. 206. Take $y_i \mid \mu, \sigma^2 \sim \text{N}(\mu, \sigma^2)$ independently for $i = 1, \ldots, N$ with $p(\mu, \sigma^2) \propto \sigma^{-2}$. Simple random sampling satisfies $p(I \mid y, \phi) = p(I)$, so the design is ignorable and $p(\mu, \sigma^2 \mid y_{\text{obs}}, I) = p(\mu, \sigma^2 \mid y_{\text{obs}})$, which by Section 3.2 is

$$\sigma^2 \mid y_{\text{obs}} \sim \text{Inv-}\chi^2\big(n-1,\, s^2_{\text{obs}}\big)\,, \qquad \mu \mid \sigma^2, y_{\text{obs}} \sim \text{N}\Big(\bar{y}_{\text{obs}}, \tfrac{\sigma^2}{n}\Big)\,.$$

Conditional on (μ, σ^2) the $N - n$ unsampled values are independent of the sampled ones, so exactly

$$\bar{y}_{\text{mis}} \mid \mu, \sigma^2, y_{\text{obs}} \;\sim\; \text{N}\Big(\mu,\; \tfrac{\sigma^2}{N-n}\Big)\,,$$

and averaging this normal over the normal posterior for μ adds the variances:

$$\bar{y}_{\text{mis}} \mid \sigma^2, y_{\text{obs}} \;\sim\; \text{N}\Big(\bar{y}_{\text{obs}},\; \sigma^2\Big(\tfrac{1}{n} + \tfrac{1}{N-n}\Big)\Big)\,.$$

Now $\bar{y} = \frac{n}{N}\bar{y}_{\text{obs}} + \frac{N-n}{N}\bar{y}_{\text{mis}}$ by (8.5) is an affine function of $\bar{y}_{\text{mis}}$, hence normal given σ^2, with mean

$\frac{n}{N}\bar{y}_{\text{obs}} + \frac{N-n}{N}\bar{y}_{\text{obs}} = \bar{y}_{\text{obs}}$ and variance scaled by $\left(\frac{N-n}{N}\right)^2$:

$$\left(\frac{N-n}{N}\right)^2 \left(\frac{1}{n} + \frac{1}{N-n}\right) = \frac{(N-n)^2}{N^2} \cdot \frac{N}{n(N-n)} = \frac{N-n}{Nn} = \frac{1}{n} - \frac{1}{N},$$

so that

$$\bar{y} \mid \sigma^2, y_{\text{obs}} \;\sim\; \text{N}\big(\bar{y}_{\text{obs}},\; \left(\tfrac{1}{n} - \tfrac{1}{N}\right)\sigma^2\big)\,.$$

Finally mix over σ^2. A normal with variance proportional to σ^2, mixed over $\sigma^2 \sim \text{Inv-}\chi^2(n-1, s^2_{\text{obs}})$, is Student-$t$ with $n-1$ degrees of freedom and the same scale factor applied to s^2_{obs} (Appendix A; the integral is the one performed in Section 3.2 to obtain $\mu \mid y \sim t_{n-1}(\bar{y}, s^2/n)$, the special case $N \to \infty$ here). This gives the stated result.

(b) The asymptotic form does not need the normal model; it needs only two central limit theorems. Write $\mu = \mu(\omega) = \text{E}(y_i \mid \omega)$ and $\sigma^2 = \sigma^2(\omega) = \text{var}(y_i \mid \omega)$ for a general exchangeable model $p(y \mid \omega) = \int \prod_i p(y_i \mid \omega) p(\omega) d\omega$, and let $n \to \infty$ and $N - n \to \infty$ with N/n fixed.

Averaging $N - n$ conditionally independent draws,

$$p(\bar{y}_{\text{mis}} \mid \omega) \;\approx\; \text{N}\Big(\bar{y}_{\text{mis}} \,\Big|\, \mu,\; \tfrac{\sigma^2}{N-n}\Big)\,.$$

By the large-sample normality of Section 4.2 applied to the ignorable posterior $p(\omega \mid y_{\text{obs}})$,

$$\text{E}(\mu \mid y_{\text{obs}}) \approx \bar{y}_{\text{obs}}, \quad \text{var}(\mu \mid y_{\text{obs}}) \approx \tfrac{1}{n} s^2_{\text{obs}}, \quad \text{E}(\sigma^2 \mid y_{\text{obs}}) \approx s^2_{\text{obs}},$$

with $\text{var}(\sigma^2 \mid y_{\text{obs}}) = O(n^{-1})$, so the uncertainty in σ^2 contributes only at lower order. Hence

$$p(\bar{y}_{\text{mis}} \mid y_{\text{obs}}) \approx \int p(\bar{y}_{\text{mis}} \mid \mu, \sigma)\, p(\mu, \sigma \mid y_{\text{obs}})\, d\mu\, d\sigma$$

is a normal mixture of normals with approximately normal mixing distribution, hence itself approximately normal. Its mean and variance follow from the conditional variance identity:

$$\begin{aligned} \text{E}(\bar{y}_{\text{mis}} \mid y_{\text{obs}}) &= \text{E}(\mu \mid y_{\text{obs}}) \;\approx\; \bar{y}_{\text{obs}}, \\ \text{var}(\bar{y}_{\text{mis}} \mid y_{\text{obs}}) &= \text{var}(\mu \mid y_{\text{obs}}) + \text{E}\Big(\tfrac{\sigma^2}{N-n} \,\Big|\, y_{\text{obs}}\Big) \\ &\approx\; \frac{s^2_{\text{obs}}}{n} + \frac{s^2_{\text{obs}}}{N-n} \;=\; \frac{N}{n(N-n)}\, s^2_{\text{obs}}. \end{aligned}$$

Substituting into (8.5), $\bar{y}$ is affine in $\bar{y}_{\text{mis}}$ and therefore approximately normal, with mean $\bar{y}_{\text{obs}}$ and variance

$$\left(\frac{N-n}{N}\right)^2 \cdot \frac{N}{n(N-n)}\, s^2_{\text{obs}} = \frac{N-n}{Nn}\, s^2_{\text{obs}} = \left(\frac{1}{n} - \frac{1}{N}\right) s^2_{\text{obs}},$$

which is (8.6).

Method (2): under the normal model of part (a) the derivation is immediate, since t_{n-1} converges to the normal as $n \to \infty$ with the location and scale of (a) unchanged.

8.2 Exercises 8.8–8.14

Problem 8.8. Finite-population inference for completely randomized experiments: in the completely randomized experiment of Section 8.4, n units are drawn at random from a population of N units, $n/2$ of them are assigned treatment A and $n/2$ treatment B, and the observed outcome of unit i is y_i^A or y_i^B according to its assignment. The estimand is the finite-population causal effect $\bar{y}^A - \bar{y}^B$, where $\bar{y}^A = \frac{1}{N}\sum_{i=1}^{N} y_i^A$ and likewise for B. The text asserts (equation (8.17)) that for large n and

large N/n,

$$(\bar{y}^A - \bar{y}^B)\,|\,y_{\text{obs}} \approx \text{N}\big(\bar{y}^A_{\text{obs}} - \bar{y}^B_{\text{obs}},\ \tfrac{2}{n}\left(s^{2A}_{\text{obs}} + s^{2B}_{\text{obs}}\right)\big)\,,$$

where s^{2A}_{obs} and s^{2B}_{obs} are the sample variances of the observed outcomes under the two treatments.
(a) Derive the asymptotic result (8.17).
(b) Derive the (finite-population) inference for $\bar{y}^A - \bar{y}^B$ under a model in which the pairs (y^A_i, y^B_i) are drawn from a bivariate normal distribution with mean (μ_A, μ_B), standard deviations (σ_A, σ_B), and correlation ρ.
(c) Discuss how inference in (b) depends on ρ and the implications in practice. Why does the dependence on ρ disappear in the limit of large N/n?

Solution. (a) The finite-population contrast collapses to the superpopulation contrast $\mu_A - \mu_B$, whose posterior distribution is normal with the stated moments.
Write $m = n/2$ for the number of units per arm. Given θ and y_{obs}, the $N - m$ unobserved values of y^A_i are independent draws from their superpopulation distribution, so

$$\bar{y}^A\,|\,\theta, y_{\text{obs}} \sim \text{N}\big(\tfrac{m}{N}\bar{y}^A_{\text{obs}} + \tfrac{N-m}{N}\mu_A,\ \tfrac{N-m}{N^2}\sigma^2_A\big)\,,$$

and likewise for B. Measured against the $O(\sigma_A m^{-1/2})$ posterior uncertainty in μ_A, the displacement $\frac{m}{N}(\bar{y}^A_{\text{obs}} - \mu_A)$ is smaller by a factor m/N and the added standard deviation $\sigma_A\sqrt{N-m}/N$ by a factor $\sqrt{m/N}$; both vanish as $N/n \to \infty$, leaving $\bar{y}^A - \bar{y}^B = \mu_A - \mu_B$ to that order.
The treatment assignment is ignorable and the likelihood factors into a part involving θ_A and a part involving θ_B (page 215), so the noninformative-prior normal-theory result of Section 3.2 applies separately in each arm: with $p(\mu_A, \mu_B, \log\sigma_A, \log\sigma_B) \propto 1$,

$$\begin{aligned}\mu_A\,|\,y_{\text{obs}} &\sim t_{m-1}\big(\bar{y}^A_{\text{obs}},\ s^{2A}_{\text{obs}}/m\big)\,,\\ \mu_B\,|\,y_{\text{obs}} &\sim t_{m-1}\big(\bar{y}^B_{\text{obs}},\ s^{2B}_{\text{obs}}/m\big)\,,\end{aligned}$$

independently. For large m the t densities approach normals (Section 4.4), so

$$(\mu_A - \mu_B)\,|\,y_{\text{obs}} \approx \text{N}\left(\bar{y}^A_{\text{obs}} - \bar{y}^B_{\text{obs}},\ \frac{s^{2A}_{\text{obs}} + s^{2B}_{\text{obs}}}{m}\right),$$

and $1/m = 2/n$ gives (8.17).
(b) Only the missing half of each sampled pair and the $N - n$ unsampled pairs need to be imputed. Let $\mathcal{A}$ and $\mathcal{B}$ be the two treatment groups and $d_i = y^A_i - y^B_i$. Under the bivariate normal model the conditional imputations are, for $i \in \mathcal{A}$,

$$y^B_i\,|\,y^A_i, \theta \sim \text{N}\Big(\mu_B + \rho\tfrac{\sigma_B}{\sigma_A}(y^A_i - \mu_A),\ \sigma^2_B(1-\rho^2)\Big)\,,$$

and symmetrically for $i \in \mathcal{B}$. Therefore

$$\begin{aligned}\text{E}[d_i\,|\,\theta, y_{\text{obs}}] &= \mu_A - \mu_B + \left(1 - \rho\tfrac{\sigma_B}{\sigma_A}\right)(y^A_i - \mu_A), \quad i \in \mathcal{A},\\ \text{E}[d_i\,|\,\theta, y_{\text{obs}}] &= \mu_A - \mu_B + \left(\rho\tfrac{\sigma_A}{\sigma_B} - 1\right)(y^B_i - \mu_B), \quad i \in \mathcal{B},\end{aligned}$$

with conditional variances $\sigma^2_B(1-\rho^2)$ and $\sigma^2_A(1-\rho^2)$ respectively, while for the $N-n$ unsampled units d_i has mean $\mu_A - \mu_B$ and variance $\sigma^2_A + \sigma^2_B - 2\rho\sigma_A\sigma_B$. Averaging over i,

$$\begin{aligned}\text{E}\big[\bar{y}^A - \bar{y}^B\,\big|\,\theta, y_{\text{obs}}\big] &= \mu_A - \mu_B\\ &\quad + \tfrac{m}{N}\left(1 - \rho\tfrac{\sigma_B}{\sigma_A}\right)(\bar{y}^A_{\text{obs}} - \mu_A)\\ &\quad + \tfrac{m}{N}\left(\rho\tfrac{\sigma_A}{\sigma_B} - 1\right)(\bar{y}^B_{\text{obs}} - \mu_B),\end{aligned}$$

$$\begin{aligned}\text{Var}\big[\bar{y}^A - \bar{y}^B\,\big|\,\theta, y_{\text{obs}}\big] &= \frac{1}{N^2}\Big[m(1-\rho^2)(\sigma^2_A + \sigma^2_B)\\ &\qquad + (N-n)\big(\sigma^2_A + \sigma^2_B - 2\rho\sigma_A\sigma_B\big)\Big] \;=:\; V_1.\end{aligned}$$

Now average over θ. Taking σ_A, σ_B as estimated by $s^A_{\text{obs}}, s^B_{\text{obs}}$ and writing $\mu_A = \bar{y}^A_{\text{obs}} + e_A$, $\mu_B = \bar{y}^B_{\text{obs}} + e_B$ with e_A, e_B independent $\text{N}(0, \sigma_A^2/m)$ and $\text{N}(0, \sigma_B^2/m)$, the conditional mean above equals $(\bar{y}^A_{\text{obs}} - \bar{y}^B_{\text{obs}}) + a\, e_A - b\, e_B$ with

$$a = 1 - \frac{m}{N}\left(1 - \rho\frac{\sigma_B}{\sigma_A}\right), \qquad b = 1 - \frac{m}{N}\left(1 - \rho\frac{\sigma_A}{\sigma_B}\right).$$

Hence the posterior distribution is normal with

$$\begin{aligned}\text{E}\big[\bar{y}^A - \bar{y}^B \,\big|\, y_{\text{obs}}\big] &= \bar{y}^A_{\text{obs}} - \bar{y}^B_{\text{obs}},\\ \text{Var}\big[\bar{y}^A - \bar{y}^B \,\big|\, y_{\text{obs}}\big] &= \frac{a^2\sigma_A^2 + b^2\sigma_B^2}{m} + V_1.\end{aligned}$$

(Check: for $\rho = 1$, $\sigma_A = \sigma_B = \sigma$ one gets $a = b = 1$, $V_1 = 0$, and the variance is $2\sigma^2/m = (2/n)(\sigma_A^2 + \sigma_B^2)$, as it must be since then $d_i \equiv \mu_A - \mu_B$ for every unit.)
(c) ρ enters only through a, b and V_1, all of whose ρ-dependent pieces carry a factor m/N or $1/N$; and ρ is not identified by the data, because no unit ever has both y_i^A and y_i^B observed, so its posterior distribution equals its prior distribution. In practice this means that a finite-population causal inference in which the experiment is a substantial fraction of the population is sensitive to an assumption the data cannot check, and must be reported as a sensitivity analysis over $\rho \in [-1, 1]$ (the extreme $\rho = 1$, additive unit-level effects, gives the smallest posterior variance). In the limit of large N/n we have $a, b \to 1$ and $V_1 \to 0$, and the estimand itself converges to the superpopulation contrast $\mu_A - \mu_B$, a function of the two marginal distributions only; the joint distribution, and hence ρ, drops out and (8.17) is recovered.

Problem 8.9. Cluster sampling:
(a) Discuss the analysis of one-stage and two-stage cluster sampling designs using the notation of this chapter. What is the role of hierarchical models in analysis of data gathered by one- and two-stage cluster sampling?
(b) Discuss the analysis of cluster sampling in which the clusters were sampled with probability proportional to some measure of size, where the measure of size is known for all clusters, sampled and unsampled. In what way do the measures of size enter into the Bayesian analysis?
See Kish (1965) and Lohr (2009) for thoughtful presentations of classical methods for design and analysis of such data.

Solution. (a) Both designs are ignorable given the cluster indicators and the cluster sizes, so the analysis is entirely a modeling problem: put a parameter at every level of clustering, then average the posterior predictive distribution over the unobserved units.
Use the notation of Section 8.3: units i in clusters $j = 1, \ldots, K$, with N_j units in cluster j, $N = \sum_j N_j$, and the estimand $\bar{Y} = \frac{1}{N}\sum_{j=1}^K N_j \bar{y}_{\cdot j}$.
(i) *One-stage.* A simple random sample of J clusters is drawn and *every* unit in each sampled cluster is measured. Then y_{mis} consists of whole clusters: $\bar{y}_{\cdot j}$ is known exactly for the J sampled clusters and completely unknown for the other $K - J$. Writing $\bar{y}_{\cdot j} \,|\, \theta_j$ and $\theta_j \sim \text{N}(\alpha, \tau^2)$ as in (8.10), the posterior draws of (α, τ) from the J observed cluster means are pushed through $\theta_j \,|\, \alpha, \tau$ for the unsampled j, and $\bar{Y}$ is reassembled with the known weights N_j.
(ii) *Two-stage.* A sample of n_j of the N_j units is drawn within each sampled cluster, so there are now two kinds of missing data: unsampled units inside sampled clusters and entire unsampled clusters. This is exactly the two-level structure (8.9)–(8.10): $\bar{y}_{\cdot j} \,|\, \theta_j \sim \text{N}(\theta_j, \sigma_j^2)$ within clusters and $\theta_j \sim \text{N}(\alpha, \tau^2)$ across clusters. Within a sampled cluster the unsampled units are imputed from $p(y_{ij}|\theta_j)$ with θ_j drawn from its posterior distribution — partially pooled toward α; the unsampled clusters are imputed as in (i).
The hierarchical model is what makes both steps possible. In case (i) it is the *only* source of information about the $K - J$ unsampled clusters: an exchangeable population distribution for θ_j is what licenses extrapolation to them, and τ — the model's version of the design effect — is estimated from the

observed spread of the cluster means. In case (ii) it additionally supplies the partial pooling that stabilizes the θ_j when the n_j are small. Fitting each cluster separately ($\tau = \infty$) leaves unsampled clusters undefined; pooling completely ($\tau = 0$) is the i.i.d. model and understates the posterior uncertainty about $\bar{Y}$.
(b) The measures of size enter as *covariates in the model*, not as weights in the estimate. Let M_j be the known measure of size for cluster j, with sampling probability proportional to M_j. This design is ignorable only conditional on M: the sampled clusters are a biased (size-biased) selection, so $p(\theta_j)$ and $p(\theta_j \,|\, \text{sampled})$ differ unless θ_j is modeled given M_j. Following the principle that all information used in the design must be included in the analysis, model the cluster parameter conditionally — as in (8.10),

$$\theta_j \sim \mathrm{N}(\alpha + \beta M_j,\ \tau^2),$$

or with a more flexible function of M_j if the data warrant it. The design is then ignorable for this model (Section 8.2), and the likelihood is just the product over sampled clusters.
The M_j then enter the *finite-population* step a second time: because M_j is known for the unsampled clusters as well, each unsampled θ_j is drawn from $\mathrm{N}(\alpha + \beta M_j, \tau^2)$ using that cluster's own M_j, and $\bar{Y}$ is assembled with the known N_j. When only a small fraction of clusters is sampled, this collapses to the approximation (8.11),

$$\bar{Y} \approx \frac{1}{N} \sum_{k=1}^{K} N_k(\alpha + \beta M_k),$$

which depends on the hyperparameters and on the population distribution of (M_k, N_k), but not on the individual θ_k. So where the classical probability-proportional-to-size estimator uses the M_j as inverse-probability weights $1/\pi_j$ on the observations, the Bayesian analysis uses them as regression predictors, and the weighting happens implicitly when the predictive draws for the unsampled clusters are summed with weights N_j.

Problem 8.10. Cluster sampling: Suppose data have been collected using cluster sampling, but the details of the sampling have been lost, so it is not known which units in the sample came from common clusters.
(a) Explain why an exchangeable but not independent and identically distributed model is appropriate.
(b) Suppose the clusters are of equal size, with A clusters, each of size B, and the data came from a simple random sample of a clusters, with a simple random sample of b units within each cluster. Under what limits of a, A, b, and B can we ignore the cluster sampling in the analysis?

Solution. (a) Exchangeable because the lost labels leave our information symmetric in the n sampled units; not i.i.d. because the units are still clustered, and clustering induces a positive correlation that no relabeling destroys.
Formally, let $c = (c_1, \ldots, c_n)$ be the (unknown) cluster memberships and θ the cluster parameters. The complete model is $p(y|c,\theta) = \prod_i p(y_i|\theta_{c_i})$, a product *given* c. Having lost c we must average over it,

$$p(y_1, \ldots, y_n) = \int \sum_c p(c) \prod_{i=1}^{n} p(y_i|\theta_{c_i})\ p(\theta)\, d\theta,$$

and since the design makes $p(c)$ invariant under permutations of the sample labels, this mixture is invariant under permutations of $(y_1, \ldots, y_n)$: the model is exchangeable, and it is a mixture of i.i.d. models rather than an i.i.d. model, which is de Finetti's representation (Section 5.2). That the mixture is nondegenerate shows up in the correlation: with a one-way variance-components model $y_i = \mu + \alpha_{c_i} + \epsilon_i$, $\mathrm{Var}(\alpha) = \tau^2$, $\mathrm{Var}(\epsilon) = \sigma^2$, and intraclass correlation $\rho_I = \tau^2/(\tau^2 + \sigma^2)$,

$$\mathrm{Corr}(y_i, y_{i'}) = \rho_I\ \Pr(c_i = c_{i'}) = \rho_I\, \frac{b-1}{n-1}\ >\ 0 \qquad (i \neq i'),$$

the same for every pair, which is what exchangeability demands and what independence forbids. Ignoring this and fitting an i.i.d. model understates the posterior variance of the population mean.

(b) Only when $(b-1)\rho_I \approx 0$. With $n = ab$ and the model of part (a),

$$\mathrm{Var}(\bar{y}) = \frac{\tau^2}{a} + \frac{\sigma^2}{ab} = \frac{\tau^2+\sigma^2}{n}\big[1 + (b-1)\rho_I\big],$$

so the design effect relative to a simple random sample of the same size n is $1 + (b-1)\rho_I$, and this is the entire content of the question. It is free of a and of A, and free of B except through ρ_I. Hence the limits are:

- (i) $b = 1$: one unit per sampled cluster. No two sampled units share a cluster, the design effect is 1, and if in addition $a/A \to 0$ the without-replacement selection of clusters is negligible and the a observations may be treated as i.i.d. draws. (For a/A not small, the usual finite-population correction $1 - a/A$ applies, as for a simple random sample of units.)
- (ii) $B = 1$, equivalently $\tau^2 \to 0$, equivalently $\rho_I \to 0$: the clusters carry no information, and the design is a simple random sample whatever b is.
- (iii) $a = A$ and $b = B$: a census, and no inference is required.

Shrinking the sampling fractions is *not* among them: $b/B \to 0$ or $a/A \to 0$ leaves $1 + (b-1)\rho_I$ untouched.

Problem 8.11. Capture-recapture (see Seber, 1992, and Barry et al., 2003): a statistician/fisherman is interested in N, the number of fish in a certain pond. He catches 100 fish, tags them, and throws them back. A few days later, he returns and catches fish until he has caught 20 tagged fish, at which point he has also caught 70 untagged fish. (That is, the second sample has 20 tagged fish out of 90 total.)
(a) Assuming that all fish are sampled independently and with equal probability, give the posterior distribution for N based on a noninformative prior distribution. (You can give the density in unnormalized form.)
(b) Briefly discuss your prior distribution and also make sure your posterior distribution is proper.
(c) Give the probability that the next fish caught by the fisherman is tagged. Write the result as a sum or integral — you do not need to evaluate it, but the result should not be a function of N.
(d) The statistician/fisherman checks his second catch of fish and realizes that, of the 20 'tagged' fish, 15 are definitely tagged, but the other 5 may be tagged — he is not sure. Include this aspect of missing data in your model and give the new joint posterior density for all parameters (in unnormalized form).

Solution. (a) The posterior density is

$$p(N \mid y) \;\propto\; p(N)\,\frac{\binom{N-100}{70}}{\binom{N}{90}}, \qquad N \geq 170.$$

The sampling rule was inverse (stop at the 20th tagged fish), so the likelihood is the negative hypergeometric probability that the first 89 draws yield 19 tagged and 70 untagged and the 90th draw is tagged:

$$p(y \mid N) = \frac{\binom{100}{19}\binom{N-100}{70}}{\binom{N}{89}} \cdot \frac{81}{N-89}.$$

Since $\binom{N}{89}(N-89) = 90\binom{N}{90}$, this equals $\frac{81}{90}\binom{100}{19}\binom{N-100}{70}/\binom{N}{90}$, and the factor in front is free of N. The stopping rule is thus ignorable: the likelihood is proportional in N to the ordinary hypergeometric likelihood $\binom{100}{20}\binom{N-100}{70}/\binom{N}{90}$ one would write for a fixed catch of 90.
(b) Take $p(N) \propto 1$ on $N \geq 170$ (or $p(N) \propto 1/N$, the scale-invariant choice); both are improper, so properness of the posterior distribution must be checked. As $N \to \infty$,

$$\frac{\binom{N-100}{70}}{\binom{N}{90}} \;\sim\; \frac{N^{70}/70!}{N^{90}/90!} \;=\; \frac{90!}{70!}\,N^{-20},$$

and $\sum_N N^{-20} < \infty$; the posterior distribution is proper under either prior (indeed under any prior with $p(N) = O(N^{18})$). The tail decays like N^{-20} because 20 tagged fish were recaptured; had he recaptured only one, the tail would be N^{-1} and the posterior distribution improper under a flat prior.

Summing the unnormalized density numerically from $N = 170$: under $p(N) \propto 1$ the mode is 449 (the classical Lincoln–Petersen value $100 \cdot 90/20 = 450$), the mean 489.5, the standard deviation 95.9, the central 95% interval [345, 717]; under $p(N) \propto 1/N$, mode 436, mean 472.7, interval [337, 683] — the two noninformative priors agree to well within a posterior standard deviation.
(c) Averaging over the posterior distribution of N. The 90 fish of the second catch are held, leaving $N - 90$ fish in the pond of which $100 - 20 = 80$ are tagged, so

$$\Pr(\text{next fish tagged} \,|\, y) = \sum_{N=170}^{\infty} p(N \,|\, y) \, \frac{80}{N - 90} \;=\; 0.211$$

under $p(N) \propto 1$ (0.220 under $p(N) \propto 1/N$). If instead the second catch was thrown back, replace $80/(N - 90)$ by $100/N$, giving 0.212.
(d) Let $z \in \{0, 1, \ldots, 5\}$ be the number of the five ambiguous fish that are truly tagged, so the catch of 90 contains $t = 15 + z$ tagged fish. Model the reading mechanism as independent per fish: a tagged fish is unreadable with probability ϕ, an untagged fish is (mis)recorded as ambiguous with probability ψ. Then z of the t tagged fish and $5 - z$ of the $90 - t = 75 - z$ untagged fish were ambiguous, and the joint posterior density is

$$\begin{aligned} p(N, \phi, \psi, z \,|\, y) \;\propto\; & p(N)\, p(\phi)\, p(\psi)\, \frac{\binom{100}{15+z}\binom{N-100}{75-z}}{\binom{N}{90}} \\ & \times \binom{15+z}{z} \phi^{z} (1-\phi)^{15} \\ & \times \binom{75-z}{5-z} \psi^{5-z} (1-\psi)^{70}, \end{aligned}$$

for $z = 0, \ldots, 5$ on the support $N - 100 \geq 75 - z$, that is $N \geq 175 - z$; the catch size 90 is treated as fixed here, since the field count of 20 tagged that triggered the stopping rule is itself now uncertain. Here z is the missing datum and (ϕ, ψ) the parameters of the missingness mechanism; they are only weakly identified by these data and need genuinely informative Beta prior distributions — an untagged fish is rarely mistaken for a tagged one, so $p(\psi)$ should be concentrated near 0. Summing out the six values of z gives $p(N, \phi, \psi \,|\, y)$.

Problem 8.12. Sampling with unequal probabilities: Table 8.7 summarizes the opinion poll discussed in the examples in Sections 3.4 and 8.3, with the responses classified by presidential preference and number of telephone lines in the household.
Table 8.7. Respondents to the CBS telephone survey classified by opinion and number of residential telephone lines (category ? indicates no response to the number of phone lines question).

Preference	1 line	2	3	4	?
Bush	557	38	4	3	7
Dukakis	427	27	1	0	3
No opinion/other	87	1	0	0	7

We shall analyze these data assuming that the probability of reaching a household is proportional to the number of telephone lines. Pretend that the responding households are a simple random sample of telephone numbers; that is, ignore the stratification discussed in Section 8.3 and ignore all nonresponse issues.
(a) Set up parametric models for (i) preference given number of telephone lines, and (ii) distribution of number of telephone lines in the population. (Hint: for (i), consider the parameterization (8.8).)
(b) What assumptions did you make about households with no telephone lines and households that did not respond to the 'number of phone lines' question?
(c) Write the joint posterior distribution of all parameters in your model.
(d) Draw 1000 simulations from the joint distribution. (Use approximate computational methods.)
(e) Compute the mean preferences for Bush, Dukakis, and no opinion/other in the population of

households (not phone numbers!) and display a histogram for the difference in support between Bush and Dukakis. Compare to Figure 3.2 and discuss any differences.
(f) Check the fit of your model to the data using posterior predictive checks.
(g) Explore the sensitivity of your results to your assumptions.

Solution. (a) Two ingredients: a logistic model for preference given lines, and a multinomial (Dirichlet) model for the population distribution of lines.
(i) For a household with $L \in \{1, 2, 3, 4\}$ lines let $\theta_L = (\theta_{1L}, \theta_{2L}, \theta_{3L})$ be the probabilities of Bush, Dukakis and no opinion/other. Reparameterize as in (8.8),

$$\alpha_{1L} = \frac{\theta_{1L}}{\theta_{1L} + \theta_{2L}}, \qquad \alpha_{2L} = 1 - \theta_{3L},$$

so that α_{1L} is the Bush share among those expressing a preference and α_{2L} the probability of expressing a preference. The $L = 3, 4$ columns have 5 and 3 respondents, so a separate free pair per column is hopeless; put the logits linear in L,

$$\begin{aligned} \operatorname{logit} \alpha_{1L} &= \beta_1 + \gamma_1 (L - 1), \\ \operatorname{logit} \alpha_{2L} &= \beta_2 + \gamma_2 (L - 1), \end{aligned}$$

with $p(\beta_1, \gamma_1, \beta_2, \gamma_2) \propto 1$. Recovering $\theta_{1L} = \alpha_{2L}\alpha_{1L}$, $\theta_{2L} = \alpha_{2L}(1 - \alpha_{1L})$, $\theta_{3L} = 1 - \alpha_{2L}$.
(ii) Let $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ be the proportions of *households* (in the telephone population) with L lines, with $\pi \sim \text{Dirichlet}(1, 1, 1, 1)$. Because a household is reached with probability proportional to L, the probability that a *respondent* household has L lines is

$$q_L = \frac{L\,\pi_L}{\sum_{m=1}^{4} m\,\pi_m},$$

and this reweighting is the whole content of the unequal selection probabilities here. The design is ignorable conditional on L (Section 8.3, unequal probabilities of selection) because L is fully observed in the sample and π, the population distribution of L, is a parameter of the model.
(b) Two assumptions, both unavoidable and both restricting the estimand.

- *Zero-line households.* They have selection probability $0 \cdot \pi_0 = 0$ and contribute nothing to the likelihood, so π_0 is not identified. The estimand is therefore redefined as the mean preference among *telephone* households, and π is a distribution over $L \geq 1$. Extending to all households requires an untestable assumption, e.g. that nontelephone households vote like one-line households.
- *Missing L.* The 17 households in the '?' column are assumed missing at random given preference: a household of type (j, L) fails to report L with a probability r_j depending on j but not on L. Its contribution to the likelihood is then $\sum_L q_L \theta_{jL}$ times a factor free of (π, β, γ), so the mechanism is ignorable (Section 8.2). This matters: the '?' households are 1.1% of Bush respondents and 7% of no-opinion respondents, so the missingness is clearly related to j, and assuming it unrelated to L *given* j is the substantive assumption.

(c) With n_{jL} the observed counts and $m_j = (7, 3, 7)$ the '?' counts,

$$p(\pi, \beta, \gamma \,|\, y) \;\propto\; \prod_{j=1}^{3} \prod_{L=1}^{4} \left(q_L\,\theta_{jL}\right)^{n_{jL}} \prod_{j=1}^{3} \Big(\sum_{L=1}^{4} q_L\,\theta_{jL} \Big)^{m_j},$$

where $q_L = L\pi_L / \sum_m m\pi_m$ and θ_{jL} is the function of (β, γ) above, times the $\text{Dirichlet}(1, 1, 1, 1)$ density for π. This is a 7-dimensional posterior distribution (π has 3 free coordinates).
(d) Random-walk Metropolis (Section 11.2) on $(\log \pi_L/\pi_1,\ \beta_1, \gamma_1, \beta_2, \gamma_2)$, 4×10^5 iterations with acceptance rate 0.12, first quarter discarded, thinned to 1000 draws. The fitted line effects are $\gamma_1 = 0.30 \pm 0.21$ (Bush share rises weakly with number of lines) and $\gamma_2 = 2.5 \pm 1.3$ (multi-line households almost always express a preference: only 1 of the 74 respondents with $L \geq 2$ had no opinion). The implied Bush shares θ_{1L} have posterior means 0.516, 0.627, 0.694, 0.744 for $L = 1, 2, 3, 4$. The population line distribution is

$$\mathrm{E}[\pi \,|\, y] = (0.968,\ 0.030,\ 0.0018,\ 0.0009),$$

against a raw respondent distribution of (0.935, 0.058, 0.0044, 0.0026): dividing by L and renormalizing is exactly the design correction.
(e) The household-level estimands are

$$\Theta_j = \sum_{L=1}^{4} \pi_L \, \theta_{jL}, \qquad \mathrm{E}[\Theta \,|\, y] = (0.520,\ 0.396,\ 0.084),$$

with posterior standard deviations (0.015, 0.014, 0.009). For the contrast,

$$(\Theta_{\text{Bush}} - \Theta_{\text{Dukakis}}) \,|\, y: \quad \text{mean } 0.124,\ \text{sd } 0.028,\ 95\% \text{ interval } [0.071,\ 0.183],$$

and $\Pr(\Theta_{\text{Bush}} > \Theta_{\text{Dukakis}} \,|\, y) > 0.999$; 1000 thinned draws locate the posterior mean to about ± 0.002, so every figure quoted above is good to its last digit. The histogram is a symmetric unimodal distribution centered at 0.124; compare the unweighted Dirichlet analysis of these same 1162 respondents, which gives 0.130 ± 0.028. Weighting by $1/L$ therefore shifts the Bush lead *down* by about 0.006 — one-fifth of a posterior standard deviation — because multi-line households lean Bush and are over-sampled. Figure 3.2, built from the 1447 respondents of Section 3.4 ($y_1 = 727$, $y_2 = 583$, $y_3 = 137$) rather than the 1162 tabulated here, is centered at $144/1450 = 0.099$ with a similar spread; the gap between 0.099 and 0.130 is a difference of *sample*, not of method, and is five times the unequal-probability correction. The phone-line weighting is thus a second-order adjustment, because 96.8% of households have a single line.
(f) The model fits. Take $T(y, \theta) = \sum_{\text{cells}} (y - \mathrm{E}\, y)^2 / \mathrm{E}\, y$ over the 3×4 table of the 1145 fully classified respondents (the '?' column is uninformative here, since its three expectations are fit exactly by the free r_j), and for each of the 1000 draws generate y^{rep} from the multinomial posterior predictive distribution:

$$\Pr\left(T(y^{\text{rep}}, \theta) \geq T(y, \theta) \,\middle|\, y\right) = 0.44 \quad \text{(Monte Carlo s.e. } 0.016),$$

with a median observed discrepancy of 9.1 on 12 cells. No evidence of misfit: in particular the linear-logit restriction on α_{1L} and α_{2L} is not contradicted by the sparse $L = 3, 4$ columns.
(g) Nothing inside the model matters; the one assumption that does lies outside it.
Refitting with (i) free $(\alpha_{1L}, \alpha_{2L})$ in each column in place of the linear logits, (ii) a Dirichlet$(\frac{1}{4}, \ldots, \frac{1}{4})$ prior distribution on π, (iii) all 17 '?' households assigned to $L = 1$, and (iv) all assigned to $L = 2$, the posterior mean of $\Theta_{\text{Bush}} - \Theta_{\text{Dukakis}}$ stays within 0.124–0.125 with standard deviation 0.027–0.029. Weakening the proportionality itself — selection probability proportional to $L^{1/2}$ or to L^2 rather than to L — gives 0.128 and 0.122. All of this is within half a posterior standard deviation, for the same reason throughout: the $L \geq 2$ cells hold 74 of 1145 respondents.
The binding assumption is the one identified in (b): the estimand covers telephone households only, and no reweighting of these data can reach the rest. If a fraction w of households have no telephone and their Bush–Dukakis gap differs by Δ, the all-household estimand shifts by $w\Delta$ — for $w = 0.07$ and $\Delta = 0.1$ that is 0.007, larger than every within-model sensitivity above and comparable to the entire phone-line correction.

Problem 8.13. Sampling with unequal probabilities (continued): Table 8.8 summarizes the opinion poll discussed in the examples in Sections 3.4 and 8.3, with the responses classified by vote preference, size of household, and number of telephone lines in the household.
Table 8.8. Respondents to the CBS telephone survey classified by opinion, number of residential telephone lines (category ? indicates no response to the number of phone lines question), and number of adults in the household (category ? includes all responses greater than 8 as well as nonresponses).

Adults	Preference	1	2	3	4	?
1	Bush	124	3	0	2	2
1	Dukakis	134	2	0	0	0
1	No opinion/other	32	0	0	0	1
2	Bush	332	21	3	0	5
2	Dukakis	229	15	0	0	3
2	No opinion/other	47	0	0	0	6
3	Bush	71	9	1	0	0
3	Dukakis	47	7	1	0	0
3	No opinion/other	4	1	0	0	0
4	Bush	23	4	0	1	0
4	Dukakis	11	3	0	0	0
4	No opinion/other	3	0	0	0	0
5	Bush	3	0	0	0	0
5	Dukakis	4	0	0	0	0
5	No opinion/other	1	0	0	0	0
6	Bush	1	0	0	0	0
6	Dukakis	1	0	0	0	0
6	No opinion/other	0	0	0	0	0
7	Bush	2	0	0	0	0
7	Dukakis	0	0	0	0	0
7	No opinion/other	0	0	0	0	0
8	Bush	1	0	0	0	0
8	Dukakis	0	0	0	0	0
8	No opinion/other	0	0	0	0	0
?	Bush	0	1	0	0	0
?	Dukakis	1	0	0	0	0
?	No opinion/other	0	0	0	0	0

Analyze these data assuming that the probability of reaching an individual is proportional to the number of telephone lines and inversely proportional to the number of persons in the household. Use this additional information to obtain inferences for the mean preferences for Bush, Dukakis, and no opinion/other among individuals, rather than households, answering the analogous versions of questions (a)–(g) in the previous exercise. Compare to your results for the previous exercise and explain the differences. (A complete analysis would require the data also cross-classified by the 16 strata in Table 8.2 as well as demographic data such as sex and age that affect the probability of nonresponse.)

Solution. Bush leads Dukakis by 0.156 ± 0.030 among individuals, against 0.124 ± 0.028 among households: weighting toward large households moves the lead *up* by about one posterior standard deviation, because larger households lean Bush.

(a) *Models.* Let π_{SL} be the population proportion of telephone households with S adults and L lines, $S = 1, \ldots, 8$, $L = 1, \ldots, 4$, and let θ_{jSL} be the preference probabilities within such a household, parameterized as in (8.8) and Exercise 8.12 with

$$\begin{aligned}\text{logit}\,\alpha_{1SL} &= \beta_1 + \gamma_1(L-1) + \delta_1(S-2),\\ \text{logit}\,\alpha_{2SL} &= \beta_2 + \gamma_2(L-1) + \delta_2(S-2).\end{aligned}$$

The key simplification is in the selection probabilities. A household with S adults contains S individuals; each is reached with probability proportional to L/S; so the expected number of respondents from cell (S, L) is proportional to

$$\underbrace{S\,\pi_{SL}}_{\text{individuals}} \times \underbrace{\frac{L}{S}}_{\text{per-person prob.}} = L\,\pi_{SL},$$

and the sample cell probabilities are $q_{SL} = L\pi_{SL} / \sum_{S'L'} L'\pi_{S'L'}$ — *exactly the same reweighting as in 8.12*. The $1/S$ factor cancels the S individuals per household, which is just the statement that one adult per household is interviewed. Prior: $q \sim \text{Dirichlet}(1/32, \ldots, 1/32)$ on the 32 cells (so that the many empty cells are not inflated by prior pseudo-counts), with $\pi_{SL} \propto q_{SL}/L$, and $p(\beta, \gamma, \delta) \propto 1$.
Where S does enter is the *estimand*. The individual-level mean preferences are

$$\Theta_j^{\text{ind}} = \frac{\sum_{S,L} S\, \pi_{SL}\, \theta_{jSL}}{\sum_{S,L} S\, \pi_{SL}},$$

versus $\Theta_j^{\text{hh}} = \sum_{S,L} \pi_{SL}\theta_{jSL}$ in 8.12. So relative to a naive analysis of the respondents, each respondent carries weight proportional to S/L: $1/L$ to undo the multiple-line oversampling, S to represent the other adults in the household.
(b) *Assumptions.* As in 8.12, households with $L = 0$ are excluded and the estimand is confined to adults in telephone households. Additionally: (i) the responding adult's preference is taken as a draw from θ_{jSL}, i.e. adults within a household are exchangeable given (S, L) — if spouses correlate politically, and they do, this understates the posterior variance and, more seriously, makes θ_{jSL} the distribution of the *selected* adult rather than of a random adult, which is only the same thing if which adult answers the phone is unrelated to preference; (ii) the 17 '?' entries for L and the 2 '?' entries for S are missing at random given the other two variables and are imputed within the model; (iii) the printed '?' row for adults lumps nonresponse together with every household size above 8 (Table 8.8 caption), and those 2 respondents are imputed within $S \le 8$, so the population is taken to contain no household with more than 8 adults.
(c) *Joint posterior.* With n_{jSL} the fully observed counts, m^L_{jS} the counts missing L, and m^S_{jL} the counts missing S,

$$\begin{aligned} p(q, \beta, \gamma, \delta \,|\, y) \;\propto\; p(q) \prod_{j,S,L} \big(q_{SL}\theta_{jSL}\big)^{n_{jSL}} \\ \times \prod_{j,S} \Big(\sum_{L} q_{SL}\theta_{jSL}\Big)^{m^L_{jS}} \prod_{j,L} \Big(\sum_{S} q_{SL}\theta_{jSL}\Big)^{m^S_{jL}}. \end{aligned}$$

(d) *Simulation.* Gibbs sampler with data augmentation (Section 11.1): impute the missing L and S from their multinomial conditionals; draw q from its conjugate Dirichlet conditional given the completed counts c_{SL}; update (β, γ, δ) by Metropolis. 1.2×10^5 iterations, acceptance rate 0.25, first sixth discarded, thinned to 1000 draws. The fitted coefficients are

$$\begin{aligned} \gamma_1 &= 0.25 \pm 0.22, \qquad & \delta_1 &= 0.19 \pm 0.08, \\ \gamma_2 &= 2.2 \pm 1.2, & \delta_2 &= 0.31 \pm 0.15. \end{aligned}$$

$\delta_1 > 0$ is the driver: Bush support rises with household size, visible already in the raw margins (Bush share among respondents is 0.434 in one-adult households, 0.550 with two adults, 0.574 with three, 0.622 with four). The posterior mean number of adults per telephone household is 1.97.
(e) *Estimands.*

$$\mathrm{E}\big[\Theta^{\text{ind}} \,\big|\, y\big] = (0.540,\ 0.384,\ 0.077),$$

with posterior standard deviations (0.016, 0.015, 0.008), and

$$(\Theta^{\text{ind}}_{\text{Bush}} - \Theta^{\text{ind}}_{\text{Dukakis}}) \,|\, y:\ \text{mean } 0.156,\ \text{sd } 0.030,\ 95\%\ [0.099,\ 0.212],$$

with $\Pr(\text{Bush ahead} \,|\, y) > 0.999$ and a Monte Carlo error of about ± 0.002 on the posterior means. Evaluating the *household* estimand from the same fitted model gives 0.124 ± 0.028, reproducing 8.12 and confirming that the two analyses differ only in the estimand, not in the likelihood.
Explanation of the difference. The entire $+0.032$ gap between the individual and household estimands comes from the factor S in the numerator of Θ^{ind} together with $\delta_1 = 0.19 > 0$: individual weighting shifts mass from one-adult households, which favored Dukakis (0.458 to 0.434), to two-or-more-adult households, which favored Bush (0.550 to 0.377 at $S = 2$), and there are about two adults per household

so the shift is substantial. The $1/L$ correction, worth -0.006, is already inside the household figure 0.124 and is an order of magnitude smaller — because 97% of households have one line but only 26% have one adult. A third, minor effect: the no-opinion rate falls from 0.108 among one-adult households to 0.035–0.073 among larger ones, so the individual-level residual category is also smaller (0.077 against 0.084). The posterior standard deviation grows from 0.028 to 0.030, the price of the unequal weights.
(f) *Model checking.* Check the two margins the model actually parameterizes, with $T(y,\theta) = \sum_{\text{cells}}(y - \mathrm{E}\,y)^2/\mathrm{E}\,y$:

$$\text{preference} \times \text{adults (24 cells):} \quad p_B = 0.33, \quad T(y,\theta) \text{ median } 24.7,$$
$$\text{preference} \times \text{lines (12 cells):} \quad p_B = 0.38, \quad T(y,\theta) \text{ median } 10.3,$$

where $p_B = \Pr(T(y^{\text{rep}},\theta) \geq T(y,\theta)\,|\,y)$ (Monte Carlo s.e. 0.015). Both margins fit. On the full $8\times3\times4$ table the same statistic gives $p_B = 0.24$, with median $T(y,\theta) = 59.3$ against a replicate median of 40.7, but there the statistic is dominated by cells whose expectations are a small fraction of one count, so it is unstable and has little power against the thing most likely to be wrong — linearity of the logits in S and L.
(g) *Sensitivity.* Unlike 8.12, one modeling choice here moves the answer: the prior distribution on the 32 cells of q. With Dirichlet$(1,\ldots,1)$ in place of Dirichlet$(\frac{1}{32},\ldots,\frac{1}{32})$ the individual-level lead rises from 0.156 ± 0.030 to 0.169 ± 0.031 — a shift of nearly half a posterior standard deviation, because a pseudo-count of 1 in each empty large-S cell is amplified by the factor S in Θ^{ind} and large households lean Bush. Adding a quadratic term in S to both logits gives 0.158 ± 0.031, and extending the range of S to 10 gives 0.155 ± 0.030: those are immaterial. The individual-versus-household gap of $+0.032$ is therefore robust in sign and order of magnitude but only good to about ±0.015, and the telephone-household restriction of 8.12(g) still bounds everything from outside.

Problem 8.14. Rounded data: the last two columns of Table 2.2 on page 59 give data on passenger airline deaths and deaths per passenger mile flown. We would like to divide these to obtain the number of passenger miles flown in each year, but the 'per mile' data are rounded. (For the purposes of this exercise, ignore the column in the table labeled 'Fatal accidents.')
Table 2.2. Worldwide airline fatalities, 1976–1985. Death rate is passenger deaths per 100 million passenger miles.

Year	Fatal accidents	Passenger deaths	Death rate
1976	24	734	0.19
1977	25	516	0.12
1978	31	754	0.15
1979	31	877	0.16
1980	22	814	0.14
1981	21	362	0.06
1982	26	764	0.13
1983	20	809	0.13
1984	16	223	0.03
1985	22	1066	0.15

(a) Using just the data from 1976 (734 deaths, 0.19 deaths per 100 million passenger miles), obtain inference for the number of passenger miles flown in 1976. Give a 95% posterior interval (you may do this by simulation). Clearly specify your model and your prior distribution.
(b) Apply your method to obtain intervals for the number of passenger miles flown each year until 1985, analyzing the data from each year separately.
(c) Now create a model that allows you to use data from all the years to estimate jointly the number of passenger miles flown each year. Estimate the model and give 95% intervals for each year. (Use approximate computational methods.)
(d) Describe how you would use the results of this analysis to get a better answer for Exercise 2.13.

Solution. (a) $[3.77 \times 10^{11},\ 3.96 \times 10^{11}]$ passenger miles, with median 3.86×10^{11}.
Measure x in units of 10^8 passenger miles, so that the death rate is $\rho = y/x$ with $y = 734$ known

exactly. The rounded value $r = 0.19$ is a deterministic function of x, so the likelihood is an indicator:

$$p(r \mid x) = \mathbf{1}\big\{0.185 \le 734/x < 0.195\big\} = \mathbf{1}\big\{3764.1 < x \le 3967.6\big\}.$$

Take $p(x) \propto 1/x$, the scale-invariant prior distribution on a positive quantity of unknown magnitude (equivalently $p(\rho) \propto 1/\rho$: with $x = y/\rho$ the Jacobian is $|dx/d\rho| = y/\rho^2$, and $(1/x)(y/\rho^2) = 1/\rho$). It is improper, but the indicator truncates it to a bounded interval, so the posterior distribution is proper — it is simply the prior restricted to the rounding window, $\log x \sim \mathrm{U}[\log 3764.1,\ \log 3967.6]$, giving the 95% central interval [3769, 3962] in units of 10^8 miles and median 3864. The prior matters not at all here: a flat $p(x) \propto 1$ gives [3769, 3963].

(b) The same computation year by year. Intervals in units of 10^8 passenger miles:

Year	2.5%	median	97.5%	rel. width
1976	3769	3864	3962	0.050
1977	4137	4304	4478	0.079
1978	4873	5029	5191	0.063
1979	5323	5484	5649	0.059
1980	5624	5818	6019	0.068
1981	5593	6054	6554	0.159
1982	5670	5881	6100	0.073
1983	6004	6228	6460	0.073
1984	6425	7539	8845	0.321
1985	6889	7111	7339	0.063

The relative width is essentially $0.01/r_t$: rounding to two decimals is a $\pm 2.6\%$ statement when $r = 0.19$ but a $\pm 17\%$ statement when $r = 0.03$. The 1984 interval is nearly useless on its own, and the naive point estimate $223/0.03 = 7433$ is badly placed — it sits above the 1985 figure, which is implausible for a growing series.

(c) Borrow strength through a smoothness prior on the trajectory. Airline traffic is a very smooth series, so penalize second differences of $\log x_t$:

$$\log x_t - 2\log x_{t-1} + \log x_{t-2} \sim \mathrm{N}(0, \sigma^2), \quad t = 3, \dots, 10,$$

with $p(\log x_1, \log x_2, \log \sigma) \propto 1$, multiplied by the ten rounding indicators $\mathbf{1}\{L_t \le \log x_t \le H_t\}$ from part (a). This is a locally linear (random-walk-on-the-growth-rate) model; it says the growth rate changes slowly, not that growth is constant.

The flat prior distribution on $\log \sigma$ is improper and would ordinarily put an unnormalizable spike at $\sigma = 0$, where all eight second differences vanish. Here it does not: linear programming shows no quadratic $a + bt + ct^2$ lies inside all ten windows simultaneously (the largest achievable slack is -0.016 on the log scale), so the second differences cannot all be driven to zero, σ is bounded away from 0, and the posterior distribution is proper.

Gibbs sampler: each $\log x_t$ has a truncated normal full conditional (normal from the Gaussian quadratic form, truncated to its rounding window), and σ^2 has the usual scaled inverse-χ^2 conditional given the second differences. 2×10^5 iterations, first tenth discarded, thinned to 9000 draws; $\mathrm{E}[\sigma \mid y] = 0.042$, and independent runs reproduce the interval endpoints below to within about 10 in these units.

Year	2.5%	median	97.5%	width vs (b)
1976	3770	3871	3962	0.99
1977	4214	4397	4481	0.79
1978	4870	4967	5157	0.90
1979	5323	5445	5636	0.96
1980	5620	5742	5986	0.93
1981	5625	5879	6167	0.56
1982	5726	5999	6107	0.89
1983	6018	6255	6454	0.96
1984	6397	6643	6992	0.25
1985	6888	7093	7337	1.00

The years with tight rounding windows are unchanged — the constraint already says everything — while 1984 loses three-quarters of its interval width and its median falls from 7539 to 6643, now correctly interpolating between 1983 and 1985. 1981 loses 44%. A cruder joint model, $\log x_t = \alpha + \beta(t - 1980.5) + \epsilon_t$ with i.i.d. ϵ_t, would achieve almost nothing: the least-squares residual standard deviation about a straight line through the ten window midpoints is 0.068 on the log scale, about the median width of the rounding windows themselves (0.074), so the trend line carries no usable information about individual years. The smoothness prior works because it constrains *curvature*, not *level*.

(d) Replace the plug-in exposures $\hat{x}_t = y_t/r_t$ of Exercise 2.13 by the posterior draws $x_t^{(\ell)}$ from (c), and average. Concretely, 2.13 asks for a Poisson model with rate θ and exposure proportional to passenger miles; write the joint posterior distribution of everything,

$$p(\theta, x \mid y, r) \;\propto\; p(\theta)\, p(x) \prod_{t=1}^{10} \text{Poisson}(y_t \mid \theta x_t)\, \mathbf{1}\{\text{round}(y_t/x_t) = r_t\},$$

and fit it by adding a θ step to the sampler of (c) — or, more simply, by multiple imputation: for each draw $x^{(\ell)}$, compute $p(\theta \mid y, x^{(\ell)})$, which is $\text{Gamma}(\alpha + \sum_t y_t,\ \beta + \sum_t x_t^{(\ell)})$ under a conjugate prior, draw a θ, then draw the 1986 predictive count from $\text{Poisson}(\theta \cdot 8000)$. The mixture over ℓ is the correct predictive distribution.

Two things this buys. First, honest uncertainty: the exposures are known only to within a few percent, which for the individual-year fits of 2.13(a),(c) is of the same order as the Poisson noise $\sqrt{734}/734 = 3.7\%$, so ignoring it understates the posterior variance. Second, and more important, it removes a bias: the naive $\hat{x}_{1984} = 7433$ is 12% above the joint-model estimate 6643, and any 2.13 analysis using it correspondingly understates the 1984 death rate. The joint model also removes the circularity in 2.13 of using y_t both as the Poisson outcome and, through y_t/r_t, as the exposure — here the exposure is a parameter informed by the rounding constraint and by the smoothness of the series, not a transformation of the outcome.

8.3 Exercises 8.15–8.17

Problem 8.15. Sequential treatment assignment: consider a medical study with two treatments, in which the subjects enter the study one at a time. As the subjects enter, they must be assigned treatments. Efron (1971) evaluates the following 'biased-coin' design for assigning treatments: each subject is assigned a treatment at random with probability of receiving treatment depending on the treatment assignments of the subjects who have previously arrived. If equal numbers of previous subjects have received each treatment, then the current subject is given the probability $\frac{1}{2}$ of receiving each treatment; otherwise, he or she is given the probability p of receiving the treatment that has been assigned to *fewer* of the previous subjects, where p is a fixed value between $\frac{1}{2}$ and 1.

(a) What covariate must be recorded on the subjects for this design to be ignorable?

(b) Outline how you would analyze data collected under this design.

(c) To what aspects of your model is this design sensitive?

(d) Discuss in Bayesian terms the advantages and disadvantages of the biased-coin design over the following alternatives: (i) independent randomization (that is, $p = \frac{1}{2}$ in the above design), (ii) randomized blocks where the blocks consist of successive pairs of subjects (that is, $p = 1$ in the above design). Be aware of the practical complications discussed in Section 8.5.

Solution. (a) The order of entry into the study: record for each subject i the time (or rank) t_i at which he or she arrived. The assignment distribution is not invariant to permutations of the unit indexes, so the unit labels alone do not determine the assignment probabilities: with $n = 3$ and treatments A, B,

$$\Pr(ABA) = \tfrac{1}{2} \cdot p \cdot \tfrac{1}{2} = \tfrac{p}{4},$$
$$\Pr(AAB) = \tfrac{1}{2} \cdot (1 - p) \cdot p,$$

which are unequal unless $p = \frac{1}{2}$: the same multiset of assignments has different probabilities depending on the arrival order.

Given $x = t$, write $I = (I_1, \ldots, I_n)$ for the assignment vector and $D_{i-1} = \#\{j < i : I_j = A\} - \#\{j < i : I_j = B\}$. Then

$$p(I \mid x, y, \phi) \;=\; \prod_{i=1}^{n} q(I_i \mid D_{i-1}), \qquad q(A \mid d) = \begin{cases} \frac{1}{2}, & d = 0, \\ 1 - p, & d > 0, \\ p, & d < 0, \end{cases}$$

which is free of y and involves no unknown parameter (p is fixed and known). The data are thus missing at random with ϕ trivially distinct from ω, so (8.3) equals (8.2) and the design is ignorable – indeed strongly ignorable and known, $p(I \mid x, y, \phi) = p(I \mid x)$. The propensity scores will not do as a substitute for t: $\Pr(I_i = A \mid x) = \frac{1}{2}$ for every i by the $A \leftrightarrow B$ symmetry, the same scores as complete randomization (Section 8.2).

(b) Because the design is ignorable given t, model the data directly and never model I: work with $p(\omega \mid x, y_{\text{obs}}) \propto p(\omega \mid x)\, p(y_{\text{obs}} \mid x, \omega)$, which is (8.3). Let $T_i = 1$ if subject i received treatment A; fit, for example,

$$y_i \mid \omega \sim \mathrm{N}\big(\alpha + \theta T_i + f(t_i),\ \sigma^2\big)\,,$$

with f a low-order spline, random-walk, or period-effect term in entry time, plus any genuine baseline covariates and, if warranted, a $\theta \times t$ interaction. Superpopulation inference is the posterior for θ; for the finite-population estimand, draw ω from its posterior, impute the missing potential outcomes y_i^{mis} from $p(y_{\text{mis}} \mid x, y_{\text{obs}}, \omega)$, and report the posterior distribution of $\frac{1}{n}\sum_i (y_i^A - y_i^B)$. Since the design is known, posterior predictive checks can replicate I as well as y: draw a fresh biased-coin sequence, draw y^{rep}, and check residual drift and serial correlation against the arrival order.

(c) The model for the dependence of y on entry time, $f(t)$, and in particular on any component of it at the frequency at which the design alternates. Smooth drift is handled well: an omitted trend of slope β biases θ by $\beta(\bar{t}_A - \bar{t}_B)$, and the design keeps $|D_i|$ small and hence keeps $\bar{t}_A - \bar{t}_B$ smaller than complete randomization does, so if t is dropped altogether the posterior mean of θ stays reasonable while its posterior standard deviation comes out too large. High-frequency structure is the danger: coding assignments as ± 1, the lag-one autocorrelation is $-\frac{1}{2} \cdot \frac{n}{n-1} = -0.505$ at $p = 1$, $n = 100$, and tends to 0 as $p \downarrow \frac{1}{2}$, so as $p \to 1$ the treatment indicator becomes a deterministic alternation within successive pairs and θ is aliased with any "first versus second subject of a pair" effect (time of day, position in a clinic session). The design is also sensitive to the premise that the coin depends only on past *assignments*: dependence on past outcomes costs strong ignorability (Section 8.2), while clinician adjustment based on the current subject's unrecorded characteristics costs ignorability outright (Section 8.5, complication 1), as would choosing p itself in light of beliefs about the treatment effect (complication 3, ϕ and ω not distinct).

(d) All three designs are ignorable as specified, so for a fixed model and fixed data the posterior distribution of ω and of the finite-population estimands is the same for all of them (Section 8.5); the comparison is entirely about expected precision, sensitivity to modeling assumptions, and the practical complications.

(i) Independent randomization, $p = \frac{1}{2}$, is ignorable and known with *no* covariates, $p(I \mid x, y, \phi) = p(I)$ as in (8.4), which is the simplest possible situation: units are exchangeable, entry order need not be recorded, and replicated assignment vectors for posterior predictive checking are unconstrained. Its cost is imbalance. With $n = 100$ the expected imbalance is $\mathrm{E}|n_A - n_B| = 7.96$ and $\Pr(|n_A - n_B| \le 2) = 0.24$, against 1.33 and 0.88 at $p = \frac{2}{3}$ (exact, from the birth-death chain for $|D_i|$). Unequal and time-imbalanced groups make inference about θ lean much harder on the assumed form of $f(t)$ and of any covariate adjustment.

(ii) The $p = 1$ design is randomized blocks of size two: exactly $n_A = n_B$, and within-pair contrasts that are exactly balanced against arbitrary drift, hence the highest precision under a model with pair effects. But every even-numbered assignment is determined by the odd-numbered one, so the investigator knows the next treatment before the next patient is enrolled and can select or defer patients accordingly – dependence on unrecorded covariates, and therefore a nonignorable mechanism whose selection effects must be modelled, "at best a difficult enterprise with heightened sensitivity to model assumptions" (Section 8.5). It also fixes the aliasing of (c) at its worst and admits only $2^{n/2}$

assignment vectors, which thins the set of hypothetical replications available for model checking. (The 2nd-edition solutions manual says this design has only two possible assignment vectors; since the pairs are randomized independently there are $2^{n/2}$.)
The biased coin with $\frac{1}{2} < p < 1$ buys nearly all of the balance of (ii) while keeping every assignment genuinely random: no assignment probability is ever 0 or 1, so no treatment can be predicted with certainty in advance, cheating is impeded, and the replicate distribution of I stays rich enough for posterior predictive checks. Its price is precisely (a): the units are no longer exchangeable, entry order must be recorded and modelled, and the "no covariates" simplicity of (i) is lost.

Problem 8.16. Randomized experiment with noncompliance: Table 8.5 on page 223 gives data from the study of vitamin A in Indonesia described in the example on page 223 (see Imbens and Rubin, 1997). In that experiment, I_i indicates assignment to the vitamin A treatment, $U_{\text{obs},i}$ indicates whether the supplement was actually taken, and $Y_{\text{obs},i}$ indicates survival (1 = survived). Vitamin A was available only to those assigned it, so $C_i(0) = 0$ for every unit: the two principal strata are compliers ($C_i(1) = 1$) and never-takers ($C_i(1) = 0$), and the stratum is observed only in the assigned-treatment arm. The data are

Category	Assignment $I_{\text{obs},i}$	Exposure $U_{\text{obs},i}$	Survival $Y_{\text{obs},i}$	Units in category
Complier or never-taker	0	0	0	74
Complier or never-taker	0	0	1	11514
Never-taker	1	0	0	34
Never-taker	1	0	1	2385
Complier	1	1	0	12
Complier	1	1	1	9663

(a) Is treatment assignment ignorable? Strongly ignorable? Known?
(b) Estimate the intention-to-treat effect: that is, the effect of assigning the treatment, irrespective of compliance, on the entire population.
(c) Give the simple instrumental variables estimate of the average effect of the treatment for the compliers.
(d) Write the likelihood, assuming compliance status is known for all units.

Solution. (a) Yes, yes, and yes. The assignment I was completely randomized, so

$$p(I \mid X, Y, \phi) = p(I),$$

which is exactly form (8.4): ignorable and known, with no unknown ϕ and no dependence on covariates or outcomes. It is a fortiori strongly ignorable, since $p(I \mid X, Y, \phi) = p(I) = p(I \mid X)$.
The exposure is a different matter: $U_{\text{obs},i} = I_i C_i(1)$ depends on the unobserved stratum $C_i(1)$, so it is not a covariate, and one stratifies on $(C(1), C(0))$ instead (Section 8.6).
(b) The two assignment arms have

$$n_0 = 74 + 11514 = 11588, \quad \bar{y}_0 = \tfrac{11514}{11588} = 0.99361,$$
$$n_1 = 2419 + 9675 = 12094, \quad \bar{y}_1 = \tfrac{12048}{12094} = 0.99620,$$

so the intention-to-treat effect $\overline{Y^1} - \overline{Y^0}$ of (8.18) is estimated by

$$\bar{y}_1 - \bar{y}_0 = 0.0025824,$$

an increase in survival of 2.58 per 1000 infants assigned the supplement. Randomization makes this unbiased for the population ITT, and the binomial standard error is

$$\sqrt{\frac{\bar{y}_0(1-\bar{y}_0)}{n_0} + \frac{\bar{y}_1(1-\bar{y}_1)}{n_1}} = 0.00093,$$

so the ITT is about 2.8 standard errors from zero.

(c) The proportion of compliers is estimated from the randomized treatment arm, where compliance status is observed:

$$\hat{p}_c = \frac{9675}{12094} = 0.79998.$$

The exclusion restriction (no effect of assignment on never-takers, NACE = 0) then gives the instrumental variables estimate (8.19),

$$\widehat{\text{CACE}} = \frac{\bar{y}_1 - \bar{y}_0}{\hat{p}_c} = \frac{0.0025824}{0.79998} = 0.0032280,$$

an increase in survival of 3.23 per 1000 compliers, with standard error roughly $0.00093/0.80 = 0.00116$ (treating $\hat{p}_c$ as fixed).

(d) With compliance known for every unit, the likelihood in $\omega = (p_c, \theta_{c0}, \theta_{c1}, \theta_{n0}, \theta_{n1})$ is

$$L(\omega) = p_c^{N_{c0}+N_{c1}} (1-p_c)^{N_{n0}+N_{n1}} \times \prod_{s\in\{c,n\}} \prod_{z\in\{0,1\}} \theta_{sz}^{M_{sz}} (1-\theta_{sz})^{N_{sz}-M_{sz}},$$

a binomial for p_c times four independent binomials, conjugate under independent beta priors. Here the units are exchangeable, the stratum $C_i \in \{c, n\}$ (complier, never-taker) has $\Pr(C_i = c) = p_c$, and $Y_i \mid C_i = s, I_i = z \sim \text{Bernoulli}(\theta_{sz})$; N_{sz} counts the units in stratum s assigned z and M_{sz} the survivors among them. The assignment factor drops out because the design is ignorable and known by (a). The estimands are $\text{CACE} = \theta_{c1} - \theta_{c0}$ and $\text{NACE} = \theta_{n1} - \theta_{n0}$, and the exclusion restriction is the constraint $\theta_{n1} = \theta_{n0}$.

Table 8.5 supplies only $(N_{n1}, M_{n1}) = (2419, 2385)$ and $(N_{c1}, M_{c1}) = (9675, 9663)$: in the control arm C_i is missing, so the cells (N_{c0}, M_{c0}) and (N_{n0}, M_{n0}) are unobserved subject to $N_{c0} + N_{n0} = 11588$, $M_{c0} + M_{n0} = 11514$, and the observed-data likelihood is the two-component mixture obtained by summing the display over $C_i \in \{c, n\}$ for each control unit.

Problem 8.17. Data structure and data analysis:

An experiment is performed comparing two treatments applied to the growth of cell cultures. The cultures are in dishes, with six cultures to a dish. The researcher applies each treatment to five dishes in a simple unpaired design and then considers two analyses: (i) $n = 30$ for each treatment, assuming there is no dependence among the outcomes within a dish, so that the observations for each treatment can be considered as 30 independent data points; or (ii) $n = 5$ for each treatment, allowing for the possibility of dependence by using, for each dish, the mean of the six outcomes within the dish and then modeling the 5 dishes as independent.

In either case, assume the researcher is doing the simple classical estimate. Thus, the estimated treatment effect is the same under either analysis—it is the average of the measurements under one treatment minus the average under the other. The only difference is whether the measurements are considered as clustered.

The researcher suspects that the outcomes within each dish are independent: the cell cultures are far enough apart that they are not physically interacting, and the experiment is done carefully enough that there is no reason to suspect there are 'dish effects.' That said, the data are clustered, so dish effects are a possibility. Further suppose that it is completely reasonable to consider the different dishes as independent.

The researcher reasons as follows: the advantage of method (i) is that, with 30 observations instead of just 5 per treatment, the standard errors of the estimates should be much smaller. On the other hand, method (ii) seems like the safer approach as it should work even if there are within-dish correlations or dish effects.

(a) Which of these two analyses should the researcher do? Explain your answer in two or three sentences.

(b) Write a model that you might use for a Bayesian analysis of this problem.

Solution. (a) Analysis (ii). The treatment was applied to *dishes*, not to cultures, so the dish is part of the design and the analysis must condition on it (Section 8.8); analysis (i) does not, and its standard errors are correct only if the between-dish variance is exactly zero, which the researcher's physical argument makes plausible but cannot establish.

Write $y_{tij} = \mu_t + \alpha_{ti} + \epsilon_{tij}$ for treatment t, dish $i = 1, \dots, 5$, culture $j = 1, \dots, 6$, with $\text{var}(\alpha_{ti}) = \tau^2$, $\text{var}(\epsilon_{tij}) = \sigma^2$ and intraclass correlation $\rho = \tau^2/(\tau^2 + \sigma^2)$. Both analyses estimate $\bar{y}_{2\cdot\cdot} - \bar{y}_{1\cdot\cdot}$, whose variance is $\frac{1}{5}(\tau^2 + \sigma^2/6)$, while the 30-point sample variance has $E[s^2] = \sigma^2 + \frac{24}{29}\tau^2$ (the sum of squares splits as $25\sigma^2$ within dishes plus $24\tau^2 + 4\sigma^2$ between). So (i)'s nominal variance is too small by

$$\frac{\sigma^2 + 6\tau^2}{\sigma^2 + \frac{24}{29}\tau^2} = 1.26,\ 1.53,\ 2.07 \qquad \text{at } \rho = 0.05,\ 0.10,\ 0.20,$$

i.e. standard errors understated by 12%, 24%, 44%. Against that, if $\tau = 0$ the only price of (ii) is degrees of freedom, 8 instead of 58, widening the expected interval by

$$\frac{t_{.975,8}\, c_8}{t_{.975,58}\, c_{58}} = \frac{2.235}{1.993} = 1.12, \qquad c_\nu = \sqrt{2/\nu}\,\Gamma(\tfrac{\nu+1}{2})/\Gamma(\tfrac{\nu}{2}).$$

The hoped-for gain from $n = 30$ is thus at most 12%, because with equal dish sizes the dish means $\bar{y}_{ti\cdot}$ are sufficient for μ_t given (τ, σ): the within-dish spread informs σ^2, not θ.

(b) A hierarchical normal model with the dish as the grouping level, the five dishes within a treatment being exchangeable (Section 5.2), which is what licenses a common τ. For $t = 1, 2$, $i = 1, \dots, 5$, $j = 1, \dots, 6$,

$$\begin{aligned} y_{tij} \mid \alpha_{ti}, \sigma &\sim \text{N}(\alpha_{ti},\, \sigma^2), \\ \alpha_{ti} \mid \mu_t, \tau &\sim \text{N}(\mu_t,\, \tau^2), \\ \mu_1 = \mu - \tfrac{\theta}{2}, &\quad \mu_2 = \mu + \tfrac{\theta}{2}, \end{aligned}$$

so θ is the treatment effect. Take $p(\mu, \theta, \log\sigma) \propto 1$ as in Section 5.4 (the 50 within-dish degrees of freedom pin σ down) and a half-Cauchy$(0, A)$ prior on τ with A on the scale of the outcome (Section 5.7), which with only ten dishes is safer than the uniform prior, though that too gives a proper posterior here. Because dishes were randomized to treatments, the assignment is ignorable given this model and no further design factor is needed (Section 8.4).

Marginalizing over α, the dish means are independent with

$$\begin{aligned} \bar{y}_{ti\cdot} \mid \mu_t, \tau, \sigma &\sim \text{N}\Big(\mu_t,\ \tau^2 + \tfrac{\sigma^2}{6}\Big), \\ \theta \mid y, \tau, \sigma &\sim \text{N}\Big(\bar{y}_{2\cdot\cdot} - \bar{y}_{1\cdot\cdot},\ \tfrac{2}{5}\big(\tau^2 + \tfrac{\sigma^2}{6}\big)\Big), \end{aligned}$$

a variance equal to analysis (i)'s at $\tau = 0$ and to analysis (ii)'s for $\tau \gg \sigma$, so the posterior averages over τ rather than forcing the choice.

9 Decision Analysis

Contents

9.1 Exercises 9.1–9.3

Problem 9.1. Basic decision analysis: Widgets cost \$2 each to manufacture and you can sell them for \$3. Your forecast for the market for widgets is (approximately) normally distributed with mean 10,000 and standard deviation 5,000. How many widgets should you manufacture in order to maximize your expected net profit?

Solution. Manufacture $n^* = \mu + \sigma\,\Phi^{-1}(1/3) = 10{,}000 - 5000(0.4307) \approx 7850$ widgets, for an expected net profit of $\mu - 3\sigma\varphi(z_{1/3}) \approx 4546$ dollars, i.e. \$4546.
Let $D \sim \mathrm{N}(\mu, \sigma^2)$ with $\mu = 10{,}000$, $\sigma = 5000$ be the market, and let n be the number manufactured. All n are paid for, and only $\min(n, D)$ are sold, so the net profit is

$$g(n, D) = 3\min(n, D) - 2n.$$

This is a utility function linear in dollars (BDA3 Section 9.1), so the decision is to maximize $\mathrm{E}[g(n, D)] = 3\,\mathrm{E}[\min(n, D)] - 2n$ over n.
Since $\frac{\partial}{\partial n}\min(n, D) = 1\{D > n\}$ and the integrand is bounded by 1, differentiation under the integral gives

$$\frac{d}{dn}\mathrm{E}[g(n, D)] = 3\ \Pr(D > n) - 2,$$

which is strictly decreasing in n: the expected profit is concave and the unique maximizer solves

$$\Pr(D > n^*) = \tfrac{2}{3}, \qquad \text{i.e.} \qquad \Phi\left(\frac{n^* - \mu}{\sigma}\right) = \tfrac{1}{3}.$$

(The critical fractile 1/3 is the ratio of the \$1 profit forgone on an unmet sale to the \$1 forgone plus the \$2 sunk on an unsold widget.) With $z_{1/3} = \Phi^{-1}(1/3) = -0.43073$,

$$n^* = 10{,}000 + 5000(-0.43073) = 7846.4,$$

so about 7850 widgets.
For the value attained, write $\mathrm{E}[\min(n, D)] = n - \mathrm{E}[(n - D)^+]$ and use the normal partial-expectation identity $\mathrm{E}[(n-D)^+] = (n-\mu)\Phi(z) + \sigma\varphi(z)$ with $z = (n-\mu)/\sigma$ (Check!). At $n = n^*$, where $\Phi(z) = 1/3$ and $n^* - \mu = \sigma z$,

$$\begin{aligned}
\mathrm{E}[g(n^*, D)] &= n^* - 3\big[(n^* - \mu)\Phi(z) + \sigma\varphi(z)\big] \\
&= \mu + \sigma z - \sigma z - 3\sigma\varphi(z_{1/3}) \\
&= \mu - 3\sigma\varphi(z_{1/3}) \\
&= 10{,}000 - 3(5000)(0.36360) = 4546.0 \text{ dollars.}
\end{aligned}$$

Problem 9.2. Conditional probability and elementary decision theory: Oscar has lost his dog; there is a 70% probability it is in forest A and a 30% chance it is in forest B. If the dog is in forest A and Oscar looks there for a day, he has a 50% chance of finding the dog. If the dog is in forest B and Oscar looks there for a day, he has an 80% chance of finding the dog.
(a) If Oscar can search only one forest for a day, where should he look to maximize his probability of finding the dog? What is the probability that the dog is still lost after the search?
(b) Assume Oscar made the rational decision and the dog is still lost (and is still in the same forest as yesterday). Where should he search for the dog on the second day? What is the probability that the dog is still lost at the end of the second day?
(c) Again assume Oscar makes the rational decision on the second day and the dog is still lost (and is still in the same forest). Where should he search on the third day? What is the probability that the dog is still lost at the end of the third day?
(d) (Expected value of additional information.) You will now figure out the expected value of

knowing, at the beginning, which forest the dog is in. Suppose Oscar will search for at most three days, with the following payoffs: -1 if the dog is found in one day, -2 if the dog is found on the second day, -3 if the dog is found on the third day, and -10 otherwise.
i. What is Oscar's expected payoff without the additional information?
ii. What is Oscar's expected payoff if he knows the dog is in forest A?
iii. What is Oscar's expected payoff if he knows the dog is in forest B?
iv. Before the search begins, how much should Oscar be willing to pay to be told which forest his dog is in?

Solution. (a) Forest A; the dog is still lost with probability 0.65.
Write L_d for the event that the dog is still lost at the end of day d and carry the joint probabilities $\pi_d(F) = \Pr(\text{dog in } F,\, L_d)$ rather than the posteriors, the normalization cancelling in every comparison: $\pi_0(A) = 0.7$, $\pi_0(B) = 0.3$, with per-day success $q_A = 0.5$, $q_B = 0.8$. Searching F finds the dog with probability $\pi(F)q_F$, so on day 1

$$\pi_0(A)q_A = 0.7(0.5) = 0.35 \quad \text{versus} \quad \pi_0(B)q_B = 0.3(0.8) = 0.24,$$

and Oscar searches A. Then $\Pr(L_1) = 1 - 0.35 = 0.65$, with $\pi_1(A) = 0.7(0.5) = 0.35$ and $\pi_1(B) = 0.3$ unchanged.
(b) Forest B; $\Pr(L_2) = 0.41$.
By Bayes' rule the posterior after the failed search of A is

$$\Pr(A \mid L_1) = \frac{0.35}{0.65} = \frac{7}{13} = 0.538, \qquad \Pr(B \mid L_1) = \frac{0.30}{0.65} = \frac{6}{13} = 0.462,$$

so the day-2 success probabilities are $\frac{7}{13}(0.5) = 0.269$ for A against $\frac{6}{13}(0.8) = 0.369$ for B: search B. Unconditionally, $\pi_2(A) = 0.35$ and $\pi_2(B) = 0.3(0.2) = 0.06$, so

$$\Pr(L_2) = 0.35 + 0.06 = 0.41.$$

(c) Forest A; $\Pr(L_3) = 0.235$.
Now $\Pr(A \mid L_2) = 0.35/0.41 = 0.854$ and $\Pr(B \mid L_2) = 0.06/0.41 = 0.146$, giving $0.854(0.5) = 0.427$ for A against $0.146(0.8) = 0.117$ for B: search A. Then $\pi_3(A) = 0.175$, $\pi_3(B) = 0.06$, and

$$\Pr(L_3) = 0.175 + 0.06 = 0.235.$$

(d) The policy (A, B, A) of parts (a)-(c) induces the day-of-discovery distribution obtained by differencing $1, \Pr(L_1), \Pr(L_2), \Pr(L_3)$:

outcome	probability	payoff
found day 1	0.350	-1
found day 2	0.240	-2
found day 3	0.175	-3
never found	0.235	-10

i. Hence

$$\begin{aligned}\text{E[payoff]} &= -1(0.350) - 2(0.240) - 3(0.175) - 10(0.235) \\ &= -0.350 - 0.480 - 0.525 - 2.350 = -3.705.\end{aligned}$$

Searching stops only on success, so a policy here is just a sequence of forests, and enumerating all $2^3 = 8$ of them confirms that the myopic path ABA is optimal for this payoff as well; the nearest rivals are AAB at -3.770 and BAA at -3.815.
ii. Told the dog is in A, Oscar searches A all three days and each day succeeds independently with probability 0.5, so the day of discovery is geometric truncated at 3:

$$\begin{aligned}\text{E[payoff} \mid A] &= -1(0.5) - 2(0.25) - 3(0.125) - 10(0.125) \\ &= -2.625.\end{aligned}$$

iii. Told the dog is in B, he searches B throughout with per-day success 0.8:

$$\begin{aligned}\mathrm{E}[\text{payoff} \mid B] &= -1(0.8) - 2(0.16) - 3(0.032) - 10(0.008)\\ &= -1.296.\end{aligned}$$

iv. Before the search, the informant's report is A with probability 0.7 and B with probability 0.3, so the expected payoff under perfect information is

$$0.7(-2.625) + 0.3(-1.296) = -2.2263,$$

and the expected value of the information is

$$-2.2263 - (-3.705) = 1.4787.$$

Oscar should be willing to pay up to 1.48 payoff units for the report.

Problem 9.3. Decision analysis:
(a) Formulate an example from earlier in this book as a decision problem. (For example, in the bioassay example of Section 3.7, there can be a cost of setting up a new experiment, a cost per rat in the experiment, and a benefit to estimating the dose-response curve more accurately. Similarly, in the meta-analysis example in Section 5.6, there can be a cost per study, a cost per patient in the study, and a benefit to accurately estimating the efficacy of beta-blockers.)
(b) Set up a utility function and determine the expected utility for each decision option within the framework you have set up.
(c) Explore the sensitivity of the results of your decision analysis to the assumptions you have made in setting up the decision problem.
The bioassay data of Table 3.1, used below, are twenty animals tested at four dose levels:

dose, x_i (log g/ml)	number of animals, n_i	number of deaths, y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

Solution. (a) The decision is the size of a follow-up bioassay: run m additional rats at each of the four doses of Table 3.1, $m \in \{0, 2, 4, \ldots, 40\}$, at a setup cost of \$2000 (waived if $m = 0$) plus \$50 per rat. The first three steps of BDA3 Section 9.1 then read as follows, the fourth being part (b).
Decisions. $d = m$, with $N = 4m$ the number of new animals.
Outcomes. The model of Section 3.7: $y_i \sim \mathrm{Bin}(n_i, \theta_i)$ with $\operatorname{logit}\theta_i = \alpha + \beta x_i$ and a uniform prior on (α, β), the animals being exchangeable within a dose group and the new ones exchangeable with the old. The new counts $\tilde{y}_i \sim \mathrm{Bin}(m, \theta_i)$ have as their predictive distribution the posterior predictive $p(\tilde{y} \mid y)$ from the current 20 rats.
Utility. The scientific goal of Section 3.7 is the LD50, $\lambda = -\alpha/\beta$ (defined, as there, conditional on $\beta > 0$, which carries posterior probability 0.99999 on the grid below). A final point estimate $\hat{\lambda}$ incurs a dollar loss $K(\hat{\lambda} - \lambda)^2$ with $K = 10^6$ dollars per squared log-dose unit, so with the optimal estimate (the posterior mean) the expected terminal loss is $K \operatorname{var}(\lambda \mid y, \tilde{y})$. Total utility:

$$U(m, \tilde{y}) = -K \operatorname{var}(\lambda \mid y, \tilde{y}) - \big(2000 + 50N\big)1\{m > 0\}.$$

(b) Averaging over the predictive distribution of $\tilde{y}$ gives the preposterior expected utility

$$\mathrm{E}[U(m)] = -K\, \mathrm{E}_{\tilde{y}|y}\big[\operatorname{var}(\lambda \mid y, \tilde{y})\big] - \big(2000 + 50N\big)1\{m > 0\}.$$

The inner posterior variances are computed on the Section 3.7 grid, $\alpha \in [-5, 10]$, $\beta \in [-10, 40]$, 301×401 points, and the outer expectation by 2500 draws $(\alpha, \beta) \sim p(\alpha, \beta \mid y)$ with $\tilde{y}$ simulated from each; the current posterior has $\mathrm{E}[\lambda \mid y] = -0.107$ and $\operatorname{var}(\lambda \mid y) = 0.00896$ (posterior sd 0.095), matching Figure 3.4.

m	$N = 4m$	E[var(λ)]	root	E[U] (\$)
0	0	0.00896	0.095	-8964
2	8	0.00658	0.081	-8981
4	16	0.00504	0.071	-7841
6	24	0.00414	0.064	-7344
8	32	0.00360	0.060	-7200
10	40	0.00319	0.056	-7185
12	48	0.00287	0.054	-7271
16	64	0.00226	0.048	-7463
20	80	0.00206	0.045	-8056
30	120	0.00143	0.038	-9429
40	160	0.00111	0.033	-11111

(Monte Carlo standard errors on the variances are at most 0.00017, that is \$170 on the utility scale.) The maximum is at $m^* = 10$, a follow-up of $N = 40$ rats, with E[U] $= -7185$ dollars against -8964 for doing nothing; the utility is flat across $N = 32$ to 48, whose entries sit within one simulation standard error of each other, so what is certified is a follow-up of forty-odd rats worth about \$1800 more than doing nothing.

The shape of the table is explained by the one-line approximation

$$\mathrm{E}\big[\mathrm{var}(\lambda \mid y, \tilde{y})\big] \;\approx\; \mathrm{var}(\lambda \mid y)\,\frac{n_0}{n_0 + N}, \qquad n_0 = 20,$$

accurate to within 15% across the whole table – the new rats buy precision much as if they were prior sample size. Maximizing $-Kv_0n_0/(n_0 + N) - c_0 - c_1N$ over N gives the closed form

$$N^* = \sqrt{\frac{Kv_0n_0}{c_1}} - n_0 = \sqrt{\frac{10^6(0.00896)(20)}{50}} - 20 = 39.9,$$

in agreement with the tabulated $N^* = 40$.

(c) Sensitivity. The qualitative conclusion depends almost entirely on K, the dollar value placed on precision, and the entry cost c_0; the optimal size scales as $K^{1/2}$ by the formula above. Entries are the optimal N^* and the net gain in dollars over no experiment, the base case being $c_0 = 2000$, $c_1 = 50$:

K (\$)	base	$c_1 = 25$	$c_1 = 100$	$c_0 = 0$	$c_0 = 5000$
250,000	0/0	0/0	0/0	8/196	0/0
500,000	0/0	0/0	0/0	24/1210	0/0
1,000,000	40/1779	64/3101	24/420	40/3779	0/0
2,000,000	64/8203	96/10158	40/5557	64/10203	64/5203
4,000,000	96/22315	160/25412	64/18405	96/24315	96/19315

Halving K to \$500{,}000 reverses the decision – the \$2000 setup cost then exceeds the value of any attainable variance reduction, even though a 24-rat follow-up would still cut the LD50 variance by more than half. Raising c_0 to \$5000 likewise kills the experiment at $K = 10^6$ while leaving N^* unchanged at $K = 2 \times 10^6$: a fixed cost affects whether to experiment, never how large. The marginal cost, by contrast, leaves the decision to experiment intact at $K = 10^6$ but not its scale, N^* running from 64 down to 24 as c_1 goes from \$25 to \$100.

The allocation matters less than the size. Holding $N = 40$ fixed, two doses bracketing the current LD50 estimate, 20 rats each at -0.30 and -0.05, cut E[var($\lambda \mid y, \tilde{y}$)] by 13% relative to 10 at each of the four doses of Table 3.1 – worth \$400 at $K = 10^6$, less than is gained by choosing m well. Putting all 40 at the single dose -0.107 is no better than the original design, a single dose being unable to identify the slope, and 20 each at the extreme doses -0.86 and 0.73 more than doubles the residual variance: those doses are nearly saturated and so say little about where the curve crosses one-half.

The most consequential assumption is the shape of the terminal loss, not any of its constants. Since $\lambda = -\alpha/\beta$ has Cauchy-like tails as $\beta \to 0$, var($\lambda \mid \cdot$) is driven by the rare simulated datasets still consistent with a flat dose-response curve, and quadratic loss pays heavily to remove them. Replacing it by a loss linear in the posterior sd, calibrated to agree at $m = 0$ (that is, $K' = 94{,}679$ dollars per log-dose unit of sd), leaves the best interior design essentially where it was, $N = 32$ rats with

> $\mathrm{E}[\mathrm{sd}(\lambda \mid y, \tilde{y})] \approx 0.057$, but brings its expected utility to within a few tens of dollars of the -8964 of doing nothing – a gap smaller than the simulation error. The \$1800 margin is a feature of the quadratic loss, and the recommendation to run the follow-up should be reported as resting on it.

10 Introduction to Bayesian Computation

Contents

10.1 Exercises 10.1–10.7

Problem 10.1. Number of simulation draws: Suppose the scalar variable θ is approximately normally distributed in a posterior distribution that is summarized by n independent simulation draws. How large does n have to be so that the 2.5% and 97.5% quantiles of θ are specified to an accuracy of $0.1\,\mathrm{sd}(\theta|y)$?
(a) Figure this out mathematically, without using simulation.
(b) Check your answer using simulation and show your results.

Solution. $n \approx 714$.
(a) For n independent draws from a density f with distribution function F, the sample p-quantile $\hat{q}_p$ has the standard asymptotic normal limit

$$\mathrm{sd}(\hat{q}_p) \;\approx\; \frac{1}{f(q_p)}\sqrt{\frac{p(1-p)}{n}}, \qquad q_p = F^{-1}(p).$$

Write $\sigma = \mathrm{sd}(\theta|y)$. Normality gives $q_{0.025} = \mu - 1.96\sigma$ and $f(q_{0.025}) = \phi(1.96)/\sigma$ with $\phi(1.96) = 0.05845$, and by symmetry the same at $p = 0.975$. Hence

$$\begin{aligned}\mathrm{sd}(\hat{q}_{0.025}) &\approx \frac{\sigma}{\phi(1.96)}\sqrt{\frac{(0.025)(0.975)}{n}} \\ &= \frac{0.15613}{0.05845}\,\frac{\sigma}{\sqrt{n}} \;=\; \frac{2.671\,\sigma}{\sqrt{n}}.\end{aligned}$$

Setting $2.671/\sqrt{n} = 0.1$ gives

$$n \;=\; (26.71)^2 \;=\; 713.6,$$

so about $n = 714$ draws (in practice one would round up to 1000).
(b) Taking $\mu = 0, \sigma = 1$ without loss of generality and repeating the experiment 20,000 times at each n, the Monte Carlo standard deviations of the two estimated quantiles are

n	sd of $\hat{q}_{0.025}$	sd of $\hat{q}_{0.975}$
400	0.130	0.130
714	0.098	0.099
1000	0.084	0.083

against the predicted $2.671/\sqrt{n} = 0.134, 0.100, 0.084$. The asymptotic formula is very slightly conservative at these sample sizes; at $n = 714$ it predicts 0.100 against the observed 0.098.

Problem 10.2. Number of simulation draws: suppose you are interested in inference for the parameter θ_1 in a multivariate posterior distribution, $p(\theta|y)$. You draw 100 independent values θ from the posterior distribution of θ and find that the posterior density for θ_1 is approximately normal with mean of about 8 and standard deviation of about 4.
(a) Using the average of the 100 draws of θ_1 to estimate the posterior mean, $\mathrm{E}(\theta_1|y)$, what is the approximate standard deviation due to simulation variability?
(b) About how many simulation draws would you need to reduce the simulation standard deviation of the posterior mean to 0.1 (thus justifying the presentation of results to one decimal place)?
(c) A more usual summary of the posterior distribution of θ_1 is a 95% central posterior interval. Based on the data from 100 draws, what are the approximate simulation standard deviations of the estimated 2.5% and 97.5% quantiles of the posterior distribution? (Recall that the posterior density is approximately normal.)
(d) About how many simulation draws would you need to reduce the simulation standard deviations of the 2.5% and 97.5% quantiles to 0.1?
(e) In the eight-schools example of Section 5.5, we simulated 200 posterior draws. What are the approximate simulation standard deviations of the 2.5% and 97.5% quantiles for school A in Table

5.3? (Table 5.3 reports, for school A, a 2.5% point of -2, a median of 10 and a 97.5% point of 31.)
(f) Why was it not necessary, in practice, to simulate more than 200 draws for the SAT coaching example?

Solution. (a) $\mathrm{sd}(\bar{\theta}_1) = \sigma/\sqrt{S} = 4/\sqrt{100} = 0.4$.
(b) $4/\sqrt{S} = 0.1$ gives $S = 1600$.
(c) By the quantile formula derived in Exercise 10.1, at $p = 0.025$ and $p = 0.975$ for a normal posterior,

$$\mathrm{sd}(\hat{q}) \;\approx\; \frac{2.671\,\sigma}{\sqrt{S}} \;=\; \frac{2.671 \cdot 4}{\sqrt{100}} \;=\; 1.07,$$

the same for both endpoints by symmetry: the nominal interval [0.2, 15.8] is pinned down only to about ± 1 at each end.
(d) $2.671 \cdot 4/\sqrt{S} = 0.1$ gives $S = (106.9)^2 \approx 11{,}400$, i.e. of order 10^4 draws — seven times more than the 1600 needed for the mean.
(e) Table 5.3 gives school A a 95% interval of $[-2, 31]$, so $\sigma_A \approx 33/(2 \cdot 1.96) = 8.4$. With $S = 200$,

$$\mathrm{sd}(\hat{q}_{0.025}) \approx \mathrm{sd}(\hat{q}_{0.975}) \approx \frac{2.671 \cdot 8.4}{\sqrt{200}} \;=\; 1.6.$$

(The posterior for θ_1 is right-skewed, so 1.6 understates the noise at the upper end somewhat and overstates it at the lower end.)
(f) Because 1.6 is negligible next to the posterior standard deviation 8.4 that the interval is reporting. Section 10.5 confirms it: $S = 10{,}000$ draws give the 95% interval $[-2, 31]$ and the 50% interval $[6, 15]$, against $[-2, 31]$ and $[7, 16]$ from 200 draws.

Problem 10.3. Posterior computations for the binomial model: suppose $y_1 \sim \mathrm{Bin}(n_1, p_1)$ is the number of successfully treated patients under an experimental new drug, and $y_2 \sim \mathrm{Bin}(n_2, p_2)$ is the number of successfully treated patients under the standard treatment. Assume that y_1 and y_2 are independent and assume independent beta prior densities for the two probabilities of success. Let $n_1 = 10$, $y_1 = 6$, and $n_2 = 20$, $y_2 = 10$. Repeat the following for several different beta prior specifications.
(a) Use simulation to find a 95% posterior interval for $p_1 - p_2$ and the posterior probability that $p_1 > p_2$.
(b) Numerically integrate to estimate the posterior probability that $p_1 > p_2$.

Solution. Beta-binomial conjugacy (Section 2.4) makes the posteriors independent betas: with priors $p_i \sim \mathrm{Beta}(\alpha_i, \beta_i)$,

$$p_1|y \sim \mathrm{Beta}(\alpha_1 + 6,\ \beta_1 + 4), \qquad p_2|y \sim \mathrm{Beta}(\alpha_2 + 10,\ \beta_2 + 10),$$

and $p_1 \perp p_2$ given y. (The Haldane choice $\alpha_i = \beta_i = 0$ is an improper prior; the posterior is still proper here because $0 < y_i < n_i$ for both groups.) The $\mathrm{E}(d|y)$ column below is exact, $\frac{\alpha_1+6}{\alpha_1+\beta_1+10} - \frac{\alpha_2+10}{\alpha_2+\beta_2+20}$.
(a) Drawing $S = 200{,}000$ independent pairs (p_1^s, p_2^s) from these two betas and forming $d^s = p_1^s - p_2^s$ gives, for four common prior choices (used for both p_1 and p_2):

prior	$p_1 \mid y$	$p_2 \mid y$	$\mathrm{E}(d \mid y)$	95% interval for $p_1 - p_2$	$\Pr(p_1 > p_2 \mid y)$
Beta(0, 0) (Haldane)	Beta(6, 4)	Beta(10, 10)	0.100	[−0.267, 0.445]	0.708
Beta(0.5, 0.5) (Jeffreys)	Beta(6.5, 4.5)	Beta(10.5, 10.5)	0.091	[−0.263, 0.426]	0.697
Beta(1, 1) (uniform)	Beta(7, 5)	Beta(11, 11)	0.083	[−0.258, 0.411]	0.686
Beta(2, 2)	Beta(8, 6)	Beta(12, 12)	0.071	[−0.249, 0.383]	0.671
Beta(10, 10)	Beta(16, 14)	Beta(20, 20)	0.033	[−0.200, 0.265]	0.611

The simulation standard error on each probability is $\sqrt{(0.69)(0.31)/S} = 0.001$, and on the interval endpoints (Exercise 10.1) about $2.671\,\mathrm{sd}(d|y)/\sqrt{S} \approx 0.001$.

(b) Conditioning on p_1 and using independence,

$$\Pr(p_1 > p_2|y) = \int_0^1 \text{Beta}(p \,|\, \alpha_1{+}6, \beta_1{+}4)\; I_p(\alpha_2{+}10, \beta_2{+}10)\, dp,$$

where $I_p(a, b)$ is the incomplete beta (the Beta(a, b) cdf). Adaptive quadrature on $[0, 1]$ evaluates this to an absolute error below 10^{-9}:

$$\begin{aligned} \text{Beta}(0,0) &: 0.70725, \qquad \text{Beta}(0.5,0.5) : 0.69606, \\ \text{Beta}(1,1) &: 0.68638, \qquad \text{Beta}(2,2) : 0.67044, \\ \text{Beta}(10,10) &: 0.61051. \end{aligned}$$

Each agrees with the corresponding simulation value to within one simulation standard error, and the answer is insensitive to the prior over the whole weakly informative range (0.67 to 0.71), only the 20-pseudo-observation Beta$(10, 10)$ pulling it appreciably toward $1/2$.

Problem 10.4. Rejection sampling:
(a) Prove that rejection sampling gives draws from $p(\theta|y)$.
(b) Why is the boundedness condition on $p(\theta|y)/q(\theta)$ necessary for rejection sampling?
(Recall the algorithm of Section 10.3: given a positive function $g(\theta)$ with finite integral $c = \int g$, defined wherever $p(\theta|y) > 0$, and a constant M with $p(\theta|y)/g(\theta) \le M$ for all θ: (1) draw θ from the density proportional to g; (2) accept θ with probability $p(\theta|y)/(Mg(\theta))$, otherwise return to step 1.)

Solution. (a) Run one round of the algorithm: draw $\theta \sim g/c$ and, independently, $U \sim \text{U}(0, 1)$, and accept iff $U \le p(\theta|y)/(Mg(\theta))$ (a legitimate rule since the bound M makes the right side lie in $[0, 1]$). Then for any set A,

$$\begin{aligned} \Pr(\theta \in A,\ \text{accept}) &= \int_A \frac{g(\theta)}{c} \cdot \frac{p(\theta|y)}{Mg(\theta)}\, d\theta \\ &= \frac{1}{cM} \int_A p(\theta|y)\, d\theta. \end{aligned}$$

Taking A to be the whole parameter space gives $\Pr(\text{accept}) = 1/(cM)$, which is positive, so the algorithm terminates with probability 1 (the number of rounds is geometric with mean cM). Dividing,

$$\Pr(\theta \in A \mid \text{accept}) = \int_A p(\theta|y)\, d\theta,$$

and since the rounds are i.i.d. this is exactly the law of the first accepted draw. Hence the output has density $p(\theta|y)$.
(b) Because the number in step 2 must be a probability. (The exercise writes $q(\theta)$ for the approximating density that Section 10.3 calls $g(\theta)$; q elsewhere in the chapter denotes an unnormalized target.) If $p(\theta|y)/g(\theta)$ is unbounded then no finite M exists and the algorithm is undefined; if one nevertheless uses a finite M smaller than the supremum, then the acceptance rule effectively becomes $\min\{p(\theta|y)/(Mg(\theta)), 1\}$ and the display in (a) gives accepted draws with density

$$\propto\ \min\{p(\theta|y),\, Mg(\theta)\}\ \neq\ p(\theta|y),$$

which caps the target on the region $\{p > Mg\}$ — exactly where g's tails are too light — and so silently under-represents it there.

Problem 10.5. Rejection sampling and importance sampling: Consider the model, $y_j \sim \text{Binomial}(n_j, \theta_j)$, where $\theta_j = \text{logit}^{-1}(\alpha + \beta x_j)$, for $j = 1, \ldots, J$, and with independent prior distributions, $\alpha \sim t_4(0, 2^2)$ and $\beta \sim t_4(0, 1)$. Suppose $J = 10$, the x_j values are randomly drawn from a U$(0, 1)$ distribution, and $n_j \sim \text{Poisson}^+(5)$, where Poisson$^+$ is the Poisson distribution restricted

to positive values.
(a) Sample a dataset at random from the model.
(b) Use rejection sampling to get 1000 independent posterior draws from (α, β).
(c) Approximate the posterior density for (α, β) by a normal centered at the posterior mode with covariance matrix fit to the curvature at the mode.
(d) Take 1000 draws from the two-dimensional t_4 distribution with that center and scale matrix and use importance sampling to estimate $\mathrm{E}(\alpha|y)$ and $\mathrm{E}(\beta|y)$.
(e) Compute an estimate of effective sample size for importance sampling using (10.4) on page 266,

$$S_{\text{eff}} = \frac{1}{\sum_{s=1}^{S}(\tilde{w}(\theta^s))^2}, \qquad \tilde{w}(\theta^s) = \frac{w(\theta^s)}{\sum_{s'} w(\theta^{s'})}.$$

Solution. (a) Drawing (α, β) from the prior gave $\alpha = -0.933$, $\beta = 1.325$; the x_j, n_j and resulting y_j are

j	1	2	3	4	5	6	7	8	9	10
x_j	0.179	0.640	0.467	0.371	0.355	0.791	0.905	0.177	0.653	0.298
n_j	8	5	4	8	3	2	7	10	6	3
y_j	2	2	2	3	0	2	4	3	3	1

The unnormalized log posterior is

$$\log q(\alpha, \beta|y) = \sum_{j=1}^{10} \left[y_j \eta_j - n_j \log\left(1 + e^{\eta_j}\right) \right] + \log t_4(\alpha; 0, 2^2) + \log t_4(\beta; 0, 1), \qquad \eta_j = \alpha + \beta x_j.$$

(b) Take $g =$ the prior itself. Then the importance ratio is exactly the likelihood,

$$\frac{q(\alpha, \beta|y)}{g(\alpha, \beta)} = \prod_j p(y_j|\alpha, \beta) \;\le\; M := e^{\ell_{\max}},$$

a bound available in closed form from the maximized log likelihood, here $\ell_{\max} = -35.533$ at the MLE $(\hat{\alpha}, \hat{\beta}) = (-1.441, 2.191)$. So: draw $\alpha \sim t_4(0, 4)$ and $\beta \sim t_4(0, 1)$, accept with probability $\exp(\ell(\alpha, \beta) - \ell_{\max})$. This is a legitimate rejection sampler by Exercise 10.4(a), and it needs no tuning; its acceptance rate is $p(y)/e^{\ell_{\max}}$, observed to be 2.87% (about 35 proposals per accepted draw). The 1000 accepted draws give

	$\mathrm{E}(\cdot \mid y)$	$\mathrm{sd}(\cdot \mid y)$	95% interval
α	-0.93	0.47	$[-1.97, -0.05]$
β	1.10	0.88	$[-0.49, 3.08]$

A reference run of the same sampler with 14,359 accepted draws gives $\mathrm{E}(\alpha|y) = -0.920$, $\mathrm{E}(\beta|y) = 1.090$, sd = 0.475 and 0.854, so the 1000-draw answers are correct to within their Monte Carlo errors of $0.47/\sqrt{1000} = 0.015$ and 0.027.
(c) Maximizing $\log q$ numerically and inverting minus the Hessian there (the normal approximation of Section 4.1) gives

$$\hat{\theta} = \begin{pmatrix} -0.829 \\ 0.911 \end{pmatrix}, \qquad V = \left[-\nabla^2 \log q(\hat{\theta}|y) \right]^{-1} = \begin{pmatrix} 0.200 & -0.281 \\ -0.281 & 0.628 \end{pmatrix},$$

i.e. marginal standard deviations 0.447 and 0.792 with correlation -0.79 — the usual strong negative dependence between intercept and slope when the x_j are not centered.
(d) Drawing $S = 1000$ values from $t_4(\hat{\theta}, V)$ and weighting by $w^s = q(\theta^s|y)/g(\theta^s)$ as in (10.3),

$$\hat{\mathrm{E}}(\alpha|y) = -0.908, \qquad \hat{\mathrm{E}}(\beta|y) = 1.078,$$

against the reference values -0.920 and 1.090; the weighted standard deviations are 0.487 and 0.880. Across ten independent repetitions of this 1000-draw run the two estimates have Monte Carlo standard deviations 0.017 and 0.031, so both agree with the reference.

(e) The largest normalized weight is only 0.0018, and (10.4) gives

$$S_{\text{eff}} = \frac{1}{\sum_s (\tilde{w}^s)^2} = 898$$

out of $S = 1000$ (the ten repetitions range from 882 to 903): the t_4 approximation covers the posterior so well that importance sampling costs almost nothing in efficiency, in contrast to the 2.87% acceptance rate of the prior-based rejection sampler in (b).

Problem 10.6. Importance sampling when the importance weights are well behaved: consider a univariate posterior distribution, $p(\theta|y)$, which we wish to approximate and then calculate moments of, using importance sampling from an unnormalized density, $g(\theta)$. Suppose the posterior distribution is normal, and the approximation is t_3 with mode and curvature matched to the posterior density.
(a) Draw a sample of size $S = 100$ from the approximate density and compute the importance ratios. Plot a histogram of the log importance ratios.
(b) Estimate $\mathrm{E}(\theta|y)$ and $\mathrm{var}(\theta|y)$ using importance sampling. Compare to the true values.
(c) Repeat (a) and (b) for $S = 10{,}000$.
(d) Using the sample obtained in (c), compute an estimate of effective sample size using (10.4) on page 266, $S_{\text{eff}} = 1/\sum_s (\tilde{w}(\theta^s))^2$ with $\tilde{w}$ the normalized weights.

Solution. The proposal is $g = t_3(0, 4/3)$ and the weights are well behaved: $S_{\text{eff}}/S = 87\%$. Take $p(\theta|y) = \mathrm{N}(0, 1)$ without loss of generality; matching curvature at the common mode 0, for $t_\nu(0, s^2)$,

$$\frac{d^2}{d\theta^2} \log t_\nu \Big|_0 = -\frac{\nu+1}{\nu s^2} \stackrel{!}{=} -1 \quad \Longrightarrow \quad s^2 = \frac{\nu+1}{\nu} = \frac{4}{3}$$

for $\nu = 3$. So $g = t_3(0, 4/3)$, which has $\text{sd} = \sqrt{3s^2} = 2$ — twice the target's, and with polynomial tails that dominate the normal's everywhere far out. The importance ratios are

$$w(\theta) = \frac{\mathrm{N}(\theta \,|\, 0, 1)}{t_3(\theta \,|\, 0, 4/3)}.$$

(a) With $S = 100$ the log ratios spread over $[-63.5, 0.23]$, but the histogram is sharply concentrated at the top of that range with a long left tail: the extreme values are the very *small* weights produced by the occasional draw from the t_3 tail, where the normal density is negligible. (Figure `bda3-ch10-is-logweights-normal-t3`.) The weights are bounded above by $\sup_\theta w(\theta) = 1.253$ (attained at $\theta = 0$, since the normal tail vanishes faster than the t_3 tail), and

$$\mathrm{E}_g[w^2] = \int \frac{p(\theta)^2}{g(\theta)}\, d\theta = 1.1493 < \infty,$$

so no single draw can dominate; here the largest normalized weight is 0.012.
(b) With that sample, $\hat{\mathrm{E}}(\theta|y) = -0.072$ and $\hat{\mathrm{var}}(\theta|y) = 1.042$, against the true values 0 and 1. Over 1000 independent replications of the $S = 100$ experiment, the mean estimate is -0.007 (sd 0.095) and 0.992 (sd 0.114): both estimators are essentially unbiased with the $O(S^{-1/2})$ error one expects.
(c) With $S = 10{,}000$: $\hat{\mathrm{E}}(\theta|y) = 0.005$ and $\hat{\mathrm{var}}(\theta|y) = 1.020$, errors down by the expected factor of ten. The log-ratio histogram has the same shape with a longer left tail (minimum -518), which is irrelevant — as Section 10.4 notes, small importance ratios have little influence on (10.3).
(d) The largest normalized weight in the $S = 10{,}000$ sample is 1.1×10^{-4}, and

$$S_{\text{eff}} = \frac{1}{\sum_s (\tilde{w}^s)^2} = 8704,$$

i.e. 87% efficiency, matching the limit $S/\mathrm{E}_g[w^2] = 10{,}000/1.1493 = 8701$ predicted by the finite weight variance found in (a).

Problem 10.7. Importance sampling when the importance weights are too variable: repeat the previous exercise, but with a t_3 posterior distribution and a normal approximation. Explain why the estimates of $\text{var}(\theta|y)$ are systematically too low.
That is: let $p(\theta|y) = t_3(0, 1)$ and let g be the normal density with mode and curvature matched to it. (a) Draw $S = 100$ from g, compute the importance ratios, and plot a histogram of the log ratios. (b) Estimate $\text{E}(\theta|y)$ and $\text{var}(\theta|y)$ and compare to the true values. (c) Repeat for $S = 10{,}000$. (d) Compute S_{eff} from (10.4) for the sample in (c).

Solution. The estimates are too low because half of a t_3's variance sits in the tails $|\theta| > 4$, where the matched normal proposal essentially never puts a draw; quantified at the end.
Matching curvature at 0 as in Exercise 10.6, now in the other direction: $-1/\sigma^2 = -(\nu+1)/\nu = -4/3$, so $g = \text{N}(0, 3/4)$ and

$$w(\theta) \;\propto\; \left(1 + \tfrac{\theta^2}{3}\right)^{-2} e^{2\theta^2/3},$$

which is *unbounded*: it grows like $e^{2\theta^2/3}/\theta^4$. The true values are $\text{E}(\theta|y) = 0$ and $\text{var}(\theta|y) = \nu/(\nu-2) = 3$.
(a) With $S = 100$ the log ratios run over $[-0.23, 2.65]$, now with a long *right* tail — the dangerous direction; the minimum $\log w(0) = -0.23$ is attained at the mode. (Figure `bda3-ch10-is-logweights-t3-normal`.) The largest normalized weight is 0.13: a single draw carries an eighth of the estimate.
(b) $\hat{\text{E}}(\theta|y) = 0.115$, $\hat{\text{var}}(\theta|y) = 2.19$ against the true 0 and 3.
(c) With $S = 10{,}000$: $\hat{\text{E}}(\theta|y) = -0.105$, $\hat{\text{var}}(\theta|y) = 1.58$. Over 1000 replications the variance estimate averages 1.26 at $S = 100$ and 1.66 at $S = 10{,}000$, with medians 1.11 and 1.50, and 97–98% of replications fall below the true value 3 at *both* sample sizes.
(d) Largest normalized weight 0.012 and

$$S_{\text{eff}} = \frac{1}{\sum_s (\tilde{w}^s)^2} = 2529$$

out of 10,000. The number is worth little: across replications it swings between about 700 and 6000, and S_{eff} is only meaningful when the weights have finite variance (Section 10.4), which here they do not —

$$\text{E}_g[w^2] = \int \frac{p(\theta)^2}{g(\theta)}\, d\theta = \infty,$$

since p^2/g diverges like $e^{2\theta^2/3}/\theta^8$.
Why the systematic underestimate: split $\text{var}(\theta|y) = \int \theta^2 p(\theta)\, d\theta = 3$ at $|\theta| = 4$,

$$\int_{|\theta|>4} \theta^2\, t_3(\theta)\, d\theta = 1.477 \qquad (49\% \text{ of the total}),$$

while

$$\Pr_g(|\theta| > 4) = 2\,\Phi\left(\frac{-4}{\sqrt{3/4}}\right) = 3.9 \times 10^{-6},$$

so even $S = 10{,}000$ draws from g yield an expected 0.04 draws there. In almost every run that region contributes nothing and the estimate recovers only the central half of the variance; in the rare run that does land a draw there, the enormous weight w makes the estimate leap upward. The estimator is consistent — $\text{E}_g[w\theta^2] = 3 < \infty$, so the ratio (10.3) converges to 3 — but its sampling distribution is so right-skewed that almost all realizations, and the replication mean itself, sit far below 3. Increasing S barely helps: with $\text{E}_g[w^2] = \infty$ there is no $\sqrt{S}$ central limit theorem, and a hundredfold increase in S moved the mean estimate only from 1.26 to 1.66.

10.2 Exercises 10.8–10.8

Problem 10.8. Importance resampling with and without replacement:
(a) Consider the bioassay example introduced in Section 3.7. Use importance resampling to approximate draws from the posterior distribution of the parameters (α, β), using the normal approximation of Section 4.1 as the starting distribution. Sample $S = 10{,}000$ from the approximate distribution, and resample without replacement $k = 1000$ samples. Compare your simulations of (α, β) to Figure 3.3b and discuss any discrepancies.
(b) Comment on the distribution of the simulated importance ratios.
(c) Repeat part (a) using importance sampling with replacement. Discuss how the results differ.
The bioassay data of Table 3.1 (Racine et al., 1986) are four groups of animals given different doses of a toxin:

Dose, x_i (log g/ml)	Animals, n_i	Deaths, y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

with model $y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$, $\text{logit}(\theta_i) = \alpha + \beta x_i$, and a uniform prior density on (α, β).
Figure 3.3b is a scatterplot of 1000 draws from this posterior, computed on a grid over $(\alpha, \beta) \in [-5, 10] \times [-10, 40]$: a cloud centred near $(1, 10)$, elongated along a ridge of positive correlation, bounded below by $\beta \approx 0$ and with a pronounced skew — a long, thin plume — toward the upper right, reaching $\beta \approx 35$.
The normal approximation of Section 4.1 is $p(\alpha, \beta \mid y) \approx \text{N}(\hat{\theta}, [I(\hat{\theta})]^{-1})$ of (4.2), evaluated at the posterior mode $(\hat{\alpha}, \hat{\beta}) = (0.8, 7.7)$; its contours and a scatterplot of 1000 draws appear in Figure 4.1.

Solution. Importance resampling recovers most of the skew that the normal approximation misses: the $k = 1000$ resampled draws have mean $(\alpha, \beta) = (1.22, 10.62)$ without replacement and $(1.30, 11.43)$ with, against the exact posterior mean $(1.31, 11.64)$ and the starting distribution's $(0.85, 7.75)$.
The starting density is the normal approximation (4.2), $g = \text{N}(\hat{\theta}, [I(\hat{\theta})]^{-1})$, at the mode of the Section 3.7 posterior:

$$\hat{\theta} = \begin{pmatrix} 0.847 \\ 7.749 \end{pmatrix}, \qquad [I(\hat{\theta})]^{-1} = \begin{pmatrix} 1.039 & 3.547 \\ 3.547 & 23.752 \end{pmatrix},$$

i.e. standard deviations 1.02 and 4.87 with correlation 0.71, reproducing the $(0.8, 7.7)$ with standard errors 1.0 and 4.9 of page 76. With $p(\cdot \mid y)$ the unnormalized posterior, put

$$w(\theta^s) = \frac{p(\theta^s \mid y)}{g(\theta^s)}, \qquad \tilde{w}(\theta^s) = w(\theta^s) \Big/ \sum_{s'=1}^{S} w(\theta^{s'}),$$

draw $S = 10{,}000$ values from g, and resample $k = 1000$ of them with probabilities $\tilde{w}$ by the three-step recipe of Section 10.4. The standard of comparison is the exact posterior on a fine grid over $[-5, 12] \times [-10, 60]$; every simulation summary below is a mean over 500 independent replications of the whole procedure, with Monte Carlo standard error under 0.1 on each entry of the table in (a).
(a) The resample matches Figure 3.3b in location, in the positive (α, β) ridge, and in the hard lower boundary at $\beta \approx 0$:

Summary	Exact posterior	Normal approx. g	SIR, no repl.	SIR, with repl.
$\text{E}(\alpha)$	1.31	0.85	1.22	1.30
$\text{E}(\beta)$	11.64	7.75	10.62	11.43
$\text{sd}(\alpha)$	1.10	1.02	1.02	1.08
$\text{sd}(\beta)$	5.77	4.87	4.62	5.33
97.5% point of β	25.4	17.3	20.7	23.6
$\Pr(\beta > 0)$	> 0.999	0.944	1.000	1.000
Median LD50 $(-\alpha/\beta)$	-0.112	-0.112	-0.114	-0.112

Two discrepancies remain. First, $\Pr(\beta > 0)$ is repaired exactly: the 5.6% of proposal draws with $\beta \le 0$ carry total normalized weight 4×10^{-6}, so none is ever resampled, and the LD50 becomes well behaved

(central 95% interval $[-0.28, 0.11]$ against the exact $[-0.28, 0.10]$, versus the range $[-12.4, 5.4]$ spanned by the 950 draws with $\beta > 0$ from the normal approximation, page 87). Second, the extreme upper-right plume is thinner than in Figure 3.3b: g supplies almost no draws with $\beta > 30$, and resampling can only reweight the points it is given, never create new ones. Hence $\text{sd}(\beta) = 4.62$ against the exact 5.77, and the 97.5% point of β falls short by 4.7.

(b) The log ratios are unimodal and extremely left-skewed, with a short but consequential right tail. Centring each replication at its maximum, the averaged percentiles of $\log w$ are

pct	0	1	5	25	50	75	95	99	100
$\log w$	-90.8	-33.5	-13.3	-5.05	-4.79	-4.67	-4.07	-3.26	0

The interquartile range is only 0.38 on the log scale, so the bulk of the ratios is nearly constant — g is a good approximation over the core. The huge left tail is harmless (page 266: small ratios have little influence on (10.2)); what matters is that the largest weight is typically 95 times the mean and carries about 1% of the total, the ten largest 3.7% and the hundred largest 10%. By (10.4),

$$S_{\text{eff}} = \frac{1}{\sum_{s=1}^{S} (\tilde{w}(\theta^s))^2},$$

with median 2730 over the replications but a 5%–95% range of $[280, 4490]$: the estimate is itself very noisy, exactly the caveat issued on page 266, because whether one draw lands far out in the skewed plume changes $\max_s \tilde{w}$ by an order of magnitude. A generalized Pareto fit to the largest $3\sqrt{S} = 300$ ratios gives shape $\hat{k} \approx 0.64$ on average, exceeding 0.7 in 27% of replications — a variance finite but only just, since g has lighter tails than the posterior in the upper-right direction. As $k = 1000$ sits below the typical S_{eff}, the resample is usable but not comfortably so.

(c) With replacement the resample is distributionally better and cosmetically worse. Better, because it is an honest draw from the discrete distribution $\{\theta^s, \tilde{w}(\theta^s)\}$: its means $(1.30, 11.43)$ reproduce the importance-sampling estimate (10.2) to two decimals, and $\text{sd}(\beta) = 5.33$ and the 97.5% point 23.6 are markedly closer to the exact 5.77 and 25.4. Worse, because the 1000 draws contain on average only 890 distinct values, with median largest multiplicity 10 and above 56 in the worst 5% of replications. The mechanism is the exclusion in step 2 of Section 10.4: without replacement every draw is capped at one copy, whereas its fair share is $k\tilde{w}(\theta^s)$. On average 24 draws have $k\tilde{w}(\theta^s) > 1$, carrying 7% of the total weight, and these are precisely the points in the upper-right plume; truncating them pulls the resample back toward g, whence the systematically low $\text{E}(\beta)$, $\text{sd}(\beta)$ and upper quantile in column three above. The footnote on page 266 grants without-replacement only "a more desirable intermediate approximation somewhere between the starting and target densities," which is a virtue only if one wants k distinct points; for posterior summaries, with replacement wins, as the P.S. on that page now recommends.

11 Basics of Markov Chain Simulation

Contents

11.1 Exercises 11.1–11.7

Problem 11.1. Metropolis-Hastings algorithm: Show that the stationary distribution for the Metropolis-Hastings algorithm is, in fact, the target distribution, $p(\omega|y)$.

Solution. The chain satisfies detailed balance with respect to $p(\omega|y)$, and detailed balance implies stationarity.
From $\omega^{t-1} = \omega_a$ one draws $\omega^* \sim J_t(\cdot|\omega_a)$ and accepts it with probability $\min(1, r)$, r being the ratio (11.2); so for $\omega_b \neq \omega_a$ the transition density is

$$K_t(\omega_b|\omega_a) = J_t(\omega_b|\omega_a)\,\min\big(1, r(\omega_b, \omega_a)\big), \qquad r(\omega_b, \omega_a) = \frac{p(\omega_b|y)J_t(\omega_a|\omega_b)}{p(\omega_a|y)J_t(\omega_b|\omega_a)},$$

the remaining mass sitting as an atom at ω_a. (r is always defined: the jump can occur only if $p(\omega_a|y)$ and $J_t(\omega_b|\omega_a)$ are both nonzero.)
Label any two points, as on page 280, so that $p(\omega_b|y)J_t(\omega_a|\omega_b) \geq p(\omega_a|y)J_t(\omega_b|\omega_a)$, i.e. $r(\omega_b, \omega_a) \geq 1 \geq r(\omega_a, \omega_b) = 1/r(\omega_b, \omega_a)$. Then

$$\begin{aligned} p(\omega_a|y)K_t(\omega_b|\omega_a) &= p(\omega_a|y)J_t(\omega_b|\omega_a), \\ p(\omega_b|y)K_t(\omega_a|\omega_b) &= p(\omega_b|y)J_t(\omega_a|\omega_b)\,r(\omega_a, \omega_b) \\ &= p(\omega_a|y)J_t(\omega_b|\omega_a), \end{aligned}$$

and the labelling was arbitrary, so $p(\omega_a|y)K_t(\omega_b|\omega_a) = p(\omega_b|y)K_t(\omega_a|\omega_b)$ for every pair, trivially so on the diagonal. Integrating over ω_a,

$$\begin{aligned} \int p(\omega_a|y)K_t(\omega_b|\omega_a)\,d\omega_a &= p(\omega_b|y)\int K_t(\omega_a|\omega_b)\,d\omega_a \\ &= p(\omega_b|y), \end{aligned}$$

since $K_t(\cdot|\omega_b)$ is a probability measure once the rejection atom is counted. Hence $\omega^{t-1} \sim p(\omega|y)$ implies $\omega^t \sim p(\omega|y)$. It is the unique stationary distribution, and the chain converges to it, provided the chain is irreducible, aperiodic and not transient (page 279) — irreducibility being exactly the requirement that J_t give positive probability of eventually reaching any state from any other.

Problem 11.2. Metropolis algorithm: Replicate the computations for the bioassay example of Section 3.7 using the Metropolis algorithm. Be sure to define your starting points and your jumping rule. Compute with log-densities (see page 261). Run the simulations long enough for approximate convergence.
The bioassay data of Table 3.1: twenty animals were divided into four groups of $n_i = 5$, each group given a different dose x_i of a toxin (log g/ml), and y_i deaths recorded.

Dose, x_i (log g/ml)	Animals, n_i	Deaths, y_i
−0.86	5	0
−0.30	5	1
−0.05	5	3
0.73	5	5

The model of Section 3.7 is $y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$ with $\text{logit}(\theta_i) = \alpha + \beta x_i$, and a uniform prior density on (α, β).

Solution. A random-walk Metropolis chain in (α, β) with normal jumps of scale $(1, 4)$ reproduces the grid simulations of Section 3.7: posterior medians $\alpha = 1.26$, $\beta = 10.7$, and LD50 median -0.115 with 95 percent interval $[-0.28, 0.09]$.

Log-density. With the uniform prior, the unnormalized log posterior is, from (3.19)-(3.20),

$$\begin{aligned}\log p(\alpha,\beta|y) &\doteq \sum_{i=1}^{4}\Big[y_i\log\theta_i + (n_i-y_i)\log(1-\theta_i)\Big]\\ &= \sum_{i=1}^{4}\Big[y_i\eta_i - n_i\log\big(1+e^{\eta_i}\big)\Big], \qquad \eta_i=\alpha+\beta x_i,\end{aligned}$$

the second line following from $\log\theta_i = \eta_i - \log(1+e^{\eta_i})$ and $\log(1-\theta_i) = -\log(1+e^{\eta_i})$. Everything is evaluated on the log scale and $\log(1+e^{\eta})$ is computed as $\max(\eta,0)+\log(1+e^{-|\eta|})$, so no overflow occurs for the large β values that the posterior reaches (page 261).
Jumping rule. Symmetric bivariate normal steps centred at the current draw,

$$J\big((\alpha^*,\beta^*)\,\big|\,(\alpha^{t-1},\beta^{t-1})\big) = \mathrm{N}\left(\begin{pmatrix}\alpha^{t-1}\\ \beta^{t-1}\end{pmatrix}, \begin{pmatrix}1^2 & 0\\ 0 & 4^2\end{pmatrix}\right),$$

the scales being crude posterior standard deviations read off the contour plot of Figure 3.3. Being symmetric, the ratio (11.2) reduces to (11.1),

$$\log r = \log p(\alpha^*,\beta^*|y) - \log p(\alpha^{t-1},\beta^{t-1}|y),$$

and ω^* is accepted when $\log u < \log r$ for $u\sim\mathrm{U}(0,1)$.
Starting points. Five overdispersed starts, chosen to bracket the contour plot of Figure 3.3 in all directions:

$$(\alpha,\beta)\in\{(-3,0),\,(3,0),\,(0,30),\,(0,-5),\,(5,30)\}.$$

Run. Five chains of 10,000 iterations, first half discarded as warm-up, then each remaining half split in two, so $m=10$, $n=2500$. The average acceptance rate was 0.52. Convergence diagnostics (11.4) and (11.8):

$$\widehat{R}_\alpha = 1.005, \quad \hat{n}_{\mathrm{eff},\alpha} = 885; \qquad \widehat{R}_\beta = 1.004, \quad \hat{n}_{\mathrm{eff},\beta} = 766,$$

well inside the thresholds $\widehat{R}<1.1$ and $\hat{n}_{\mathrm{eff}} > 5m = 50$ of page 287.
Results (25,000 saved draws):

Estimand	2.5%	25%	median	75%	97.5%
α	−0.51	0.59	1.26	2.03	3.80
β	3.64	7.32	10.72	15.03	25.23
LD50 $= -\alpha/\beta$	−0.281	−0.165	−0.115	−0.063	0.094

These match the 1000 grid draws of Figure 3.3b and the LD50 histogram of Figure 3.4 to within simulation error; the posterior correlation of α and β is 0.65, the ridge visible in Figure 3.3a. All 25,000 draws had $\beta>0$, as on page 77, so the LD50 summary is meaningful.

Problem 11.3. Gibbs sampling: Table 11.4 contains quality control measurements from 6 machines in a factory. Quality control measurements are expensive and time-consuming, so only 5 measurements were done for each machine. In addition to the existing machines, we are interested in the quality of another machine (the seventh machine).

Machine	Measurements
1	83, 92, 92, 46, 67
2	117, 109, 114, 104, 87
3	101, 93, 92, 86, 67
4	105, 119, 116, 102, 116
5	79, 97, 103, 79, 92
6	57, 92, 104, 77, 100

Table 11.4: Quality control measurements from 6 machines in a factory.
Implement a separate, a pooled and hierarchical Gaussian model with common variance described

in Section 11.6. Run the simulations long enough for approximate convergence. Using each of three models—separate, pooled, and hierarchical—report: (i) the posterior distribution of the mean of the quality measurements of the sixth machine, (ii) the predictive distribution for another quality measurement of the sixth machine, and (iii) the posterior distribution of the mean of the quality measurements of the seventh machine.

Solution. The three models differ chiefly on the seventh machine: separate says nothing at all, pooled says $\theta_7 = \mu$ with posterior 95 percent interval $[86, 100]$, and hierarchical gives the honest answer $[52, 134]$.

Data: $J = 6$ machines, $n_j = 5$ each, $n = 30$. Group means $\bar{y}_{\cdot j} = (76.0, 106.2, 87.8, 111.6, 90.0, 86.0)$, grand mean $\bar{y}_{\cdot\cdot} = 92.93$.

(A) Separate model. $y_{ij} \sim \mathrm{N}(\theta_j, \sigma^2)$ with $p(\theta_1, \ldots, \theta_6, \log\sigma) \propto 1$. This is Section 11.6 with $\tau = \infty$, so (11.9)-(11.11) degenerate and the Gibbs steps are

$$\theta_j \,|\, \sigma, y \sim \mathrm{N}\big(\bar{y}_{\cdot j},\ \sigma^2/n_j\big), \qquad \sigma^2 \,|\, \theta, y \sim \text{Inv}{-}\chi^2\big(n, \hat{\sigma}^2\big),$$

with $\hat{\sigma}^2$ as in (11.15). The seventh machine has no data and an improper flat prior, so its posterior is that same improper flat distribution: **no inference is possible**.

(B) Pooled model. $y_{ij} \sim \mathrm{N}(\mu, \sigma^2)$ with $p(\mu, \log\sigma) \propto 1$, i.e. $\theta_1 = \cdots = \theta_7 = \mu$. Gibbs steps

$$\mu \,|\, \sigma, y \sim \mathrm{N}\big(\bar{y}_{\cdot\cdot}, \sigma^2/n\big), \qquad \sigma^2 \,|\, \mu, y \sim \text{Inv}{-}\chi^2\big(n,\ \tilde{\sigma}^2\big), \quad \tilde{\sigma}^2 = \tfrac{1}{n}\sum_{i,j}(y_{ij} - \mu)^2.$$

(C) Hierarchical model. Exactly Section 11.6: $y_{ij} \sim \mathrm{N}(\theta_j, \sigma^2)$, $\theta_j \sim \mathrm{N}(\mu, \tau^2)$, $p(\mu, \log\sigma, \tau) \propto 1$. The Gibbs sampler cycles (11.9)-(11.11) for the θ_j, (11.12)-(11.13) for μ, (11.14)-(11.15) for σ^2, and (11.16)-(11.17) for τ^2. The seventh machine is exchangeable with the six observed ones, so $\theta_7 | \mu, \tau \sim \mathrm{N}(\mu, \tau^2)$ is drawn alongside each Gibbs iteration.

Computation. Five chains of 4000 iterations for each model, first half discarded as warm-up, starting the θ_j at randomly chosen data points from group j and μ at their average (page 289); $\widehat{R} \le 1.01$ for every estimand, including $\log\tau$. Posterior quantiles from the 10,000 saved draws:

Model	Estimand	2.5%	25%	median	75%	97.5%
Separate	θ_6	72.7	81.6	86.0	90.3	98.9
Separate	$\tilde{y}_6$	52.7	75.3	85.9	96.4	118.3
Separate	θ_7	—	—	—	—	—
Pooled	$\theta_6 = \mu$	86.3	90.7	92.9	95.2	99.8
Pooled	$\tilde{y}_6$	54.8	80.7	93.1	105.2	129.4
Pooled	$\theta_7 = \mu$	86.3	90.7	92.9	95.2	99.8
Hierarchical	θ_6	75.3	83.7	87.7	91.8	99.6
Hierarchical	$\tilde{y}_6$	56.2	77.0	87.8	98.7	120.4
Hierarchical	θ_7	52.1	83.2	93.1	103.1	134.5

Hyperparameters under the hierarchical model: μ has median 92.9 with 95 percent interval $[77.1, 108.9]$, σ median 14.7 $[11.2, 20.4]$, τ median 14.0 $[3.6, 39.6]$.

(i) The separate model uses only machine 6's five measurements and centres θ_6 at $\bar{y}_{\cdot 6} = 86.0$; the pooled model forces $\theta_6 = \mu = 92.9$ and is far too confident; the hierarchical estimate 87.7 sits between them, shrunk about one quarter of the way toward μ, in line with the shrinkage factor $(\sigma^2/n_j)/(\tau^2 + \sigma^2/n_j)$, which is 0.18 at the median $(\sigma, \tau) = (14.7, 14.0)$ and larger when averaged over the small-τ part of the posterior.

(ii) The predictive intervals for $\tilde{y}_6$ are all roughly ± 32 wide because they are dominated by $\sigma \approx 15$; the pooled one is both shifted upward and inflated, since pooling six genuinely different machines forces σ up to a median of 18.2.

(iii) Only the hierarchical model answers the seventh-machine question: its interval $[52, 134]$ is more than four times wider than the pooled interval, because it must account for both the uncertainty in μ and the genuine machine-to-machine spread τ.

Problem 11.4. Gibbs sampling: Extend the model in Exercise 11.3 by adding a hierarchical model for the variances of the machine quality measurements. Use an Inv-χ^2 prior distribution for variances with unknown scale σ_0^2 and fixed degrees of freedom. (The data do not contain enough information for determining the degrees of freedom, so inference for that hyperparameter would depend very strongly on its prior distribution in any case.) The conditional distribution of σ_0^2 is not of simple form, but you can sample from its distribution, for example, using grid sampling.

Solution. Give each machine its own σ_j^2 drawn from $\text{Inv}{-}\chi^2(\nu, \sigma_0^2)$ with $\nu = 4$ fixed; the conditionals for θ, μ, τ^2 and the σ_j^2 all stay conjugate, and σ_0 is drawn on a grid.
Model.

$$\begin{aligned} y_{ij} \,|\, \theta_j, \sigma_j^2 &\sim \text{N}(\theta_j, \sigma_j^2), \qquad i = 1, \dots, n_j,\ j = 1, \dots, J, \\ \theta_j \,|\, \mu, \tau^2 &\sim \text{N}(\mu, \tau^2), \\ \sigma_j^2 \,|\, \sigma_0^2 &\sim \text{Inv}{-}\chi^2(\nu, \sigma_0^2), \quad \nu = 4 \text{ fixed}, \end{aligned}$$

with $p(\mu, \tau, \sigma_0) \propto 1$ (uniform on τ and on σ_0, following Section 11.6 and Chapter 5; a uniform prior on $\log \tau$ would make the posterior improper).
Gibbs steps. Relative to Exercise 11.3(C), (ii) is unchanged and the rest are:
(i) θ_j, as in (11.9)-(11.11) but with σ_j in place of σ:

$$\theta_j \,|\, \cdot \sim \text{N}(\hat\theta_j, V_{\theta_j}), \quad V_{\theta_j} = \Big(\tfrac{1}{\tau^2} + \tfrac{n_j}{\sigma_j^2}\Big)^{-1}, \quad \hat\theta_j = V_{\theta_j}\Big(\tfrac{\mu}{\tau^2} + \tfrac{n_j \bar y_{\cdot j}}{\sigma_j^2}\Big).$$

(ii) $\mu \,|\, \cdot \sim \text{N}(\bar\theta, \tau^2/J)$ and $\tau^2 \,|\, \cdot \sim \text{Inv}{-}\chi^2(J-1, \hat\tau^2)$, exactly (11.12)-(11.13) and (11.16)-(11.17).
(iii) σ_j^2 is a normal variance with known mean θ_j and a conjugate $\text{Inv}{-}\chi^2(\nu, \sigma_0^2)$ prior, so by the conjugacy of Section 2.6,

$$\sigma_j^2 \,|\, \cdot \sim \text{Inv}{-}\chi^2\!\left(\nu + n_j,\ \frac{\nu\sigma_0^2 + \sum_i (y_{ij} - \theta_j)^2}{\nu + n_j}\right).$$

(iv) σ_0. Collecting the J $\text{Inv}{-}\chi^2(\nu, \sigma_0^2)$ densities, each proportional to $(\sigma_0^2)^{\nu/2} \exp(-\nu\sigma_0^2/2\sigma_j^2)$,

$$p(\sigma_0 \,|\, \theta, \sigma^2, \mu, \tau, y) \ \propto\ \sigma_0^{J\nu} \exp\!\left(-\frac{\nu\sigma_0^2}{2} \sum_{j=1}^{J} \sigma_j^{-2}\right), \qquad \sigma_0 > 0.$$

This is evaluated on a grid $\sigma_0 \in (0.5, 80]$ of 4000 points, normalized, and sampled by inverse-CDF as instructed — though with a uniform prior on σ_0 the change of variable $u = \sigma_0^2$ turns the display into $u^{(J\nu+1)/2-1} e^{-au}$ with $a = \frac{\nu}{2}\sum_j \sigma_j^{-2}$, so the conditional is in fact the simple form $\sigma_0^2 \,|\, \cdot \sim \text{Gamma}\big(\frac{J\nu+1}{2}, a\big)$, and grid and exact draws agree in every quantile.
Predictions. For a new measurement from machine 6, draw $\tilde y_6 \sim \text{N}(\theta_6, \sigma_6^2)$. For the unobserved seventh machine, draw $\theta_7 \sim \text{N}(\mu, \tau^2)$ and $\sigma_7^2 \sim \text{Inv}{-}\chi^2(\nu, \sigma_0^2)$ at each iteration.
Results. Five chains of 8000 iterations, first half discarded; $\widehat{R} \le 1.001$ throughout. Posterior quantiles from 20,000 draws:

Estimand	2.5%	25%	median	75%	97.5%
θ_6	73.3	83.7	88.7	93.7	104.6
$\tilde y_6$	43.8	75.0	88.6	102.2	134.8
θ_7	52.0	83.9	93.7	103.1	134.2
μ	76.5	89.1	93.6	98.1	109.4
τ	2.5	9.3	13.5	19.2	40.4
σ_6	11.3	15.5	18.8	23.5	38.8
σ_0	8.2	12.3	15.7	20.6	37.1
σ_7	6.9	12.5	17.8	26.3	60.2

Posterior median σ_j by machine: 19.7, 15.2, 15.4, 13.7, 14.4, 18.8, against raw within-machine standard deviations 19.6, 11.8, 12.8, 7.6, 10.8, 19.2 — substantial shrinkage toward σ_0, as $\nu = 4$ is comparable to $n_j - 1 = 4$.

Machine 6 is one of the two noisy machines, so its own variance is now estimated at $\sigma_6 \approx 18.8$ rather than the pooled $\sigma \approx 14.7$ of Exercise 11.3, and its predictive interval for $\tilde{y}_6$ widens from $[56, 120]$ to $[44, 135]$; θ_7 is essentially unchanged, since μ and τ barely move.

Problem 11.5. Monitoring convergence:
(a) Prove that $\widehat{\text{var}}^+(\psi|y)$ as defined in (11.3) is an unbiased estimate of the marginal posterior variance of ψ, if the starting distribution for the Markov chain simulation algorithm is the same as the target distribution, and if the m parallel sequences are computed independently. (Hint: show that $\widehat{\text{var}}^+(\psi|y)$ can be expressed as the average of the halved squared differences between simulations ψ from different sequences, and that each of these has expectation equal to the posterior variance.)
(b) Determine the conditions under which $\widehat{\text{var}}^+(\psi|y)$ approaches the marginal posterior variance of ψ in the limit as the lengths n of the simulated chains approach ∞.
Here, with ψ_{ij} $(i = 1, \dots, n;\ j = 1, \dots, m)$ the simulations of a scalar estimand,

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\psi}_{\cdot j} - \bar{\psi}_{\cdot\cdot})^2,$$

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2, \qquad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \bar{\psi}_{\cdot j})^2,$$

and $\widehat{\text{var}}^+(\psi|y) = \frac{n-1}{n} W + \frac{1}{n} B$.

Solution. (a) $\widehat{\text{var}}^+$ is exactly the average of $\frac{1}{2}(\psi_{ij} - \psi_{i'j'})^2$ over all $m(m-1)n^2$ ordered pairs of draws lying in **different** sequences. (The printed ϕ in both parts is a typo for ψ.)
Write $Q = \sum_{i,j} \psi_{ij}^2$ and $P = \sum_j \bar{\psi}_{\cdot j}^2$. Since $\sum_{i,j} (\psi_{ij} - \bar{\psi}_{\cdot j})^2 = Q - nP$ and $\sum_j (\bar{\psi}_{\cdot j} - \bar{\psi}_{\cdot\cdot})^2 = P - m\bar{\psi}_{\cdot\cdot}^2$,

$$\begin{aligned} \widehat{\text{var}}^+ &= \frac{1}{mn}\big(Q - nP\big) + \frac{1}{m-1}\big(P - m\bar{\psi}_{\cdot\cdot}^2\big) \\ &= \frac{Q}{mn} + \frac{P}{m(m-1)} - \frac{m}{m-1}\,\bar{\psi}_{\cdot\cdot}^2. \end{aligned}$$

Now let

$$D = \frac{1}{m(m-1)n^2} \sum_{j \neq j'} \sum_{i=1}^{n} \sum_{i'=1}^{n} \tfrac{1}{2}\big(\psi_{ij} - \psi_{i'j'}\big)^2.$$

Expanding $\frac{1}{2}(a-b)^2 = \frac{1}{2}a^2 + \frac{1}{2}b^2 - ab$, the two square terms each contribute $\frac{1}{2}(m-1)n\,Q$, and the cross term is

$$\sum_{j \neq j'} \big(n\bar{\psi}_{\cdot j}\big)\big(n\bar{\psi}_{\cdot j'}\big) = n^2\Big[\big(\textstyle\sum_j \bar{\psi}_{\cdot j}\big)^2 - P\Big] = n^2\big[m^2\bar{\psi}_{\cdot\cdot}^2 - P\big],$$

so that

$$D = \frac{(m-1)nQ - n^2m^2\bar{\psi}_{\cdot\cdot}^2 + n^2P}{m(m-1)n^2} = \frac{Q}{mn} + \frac{P}{m(m-1)} - \frac{m}{m-1}\bar{\psi}_{\cdot\cdot}^2.$$

Hence $\widehat{\text{var}}^+(\psi|y) = D$ identically, for every realization. (Check!)
If the starting distribution is the target, then every iterate is marginally distributed as $p(\psi|y)$ (this is Exercise 11.1: the target is stationary, so stationarity propagates from $t = 0$), and draws from different sequences are independent because the m sequences are run independently. So for $j \neq j'$,

$$\text{E}\big[\tfrac{1}{2}(\psi_{ij} - \psi_{i'j'})^2\big] = \tfrac{1}{2}\big[\text{var}(\psi_{ij}) + \text{var}(\psi_{i'j'})\big] = \text{var}(\psi|y),$$

the cross term vanishing since $\text{E}\,\psi_{ij} = \text{E}\,\psi_{i'j'}$ and the two are independent. Averaging, $\text{E}\,\widehat{\text{var}}^+(\psi|y) = \text{E}\,D = \text{var}(\psi|y)$.
(b) $\widehat{\text{var}}^+ \to \text{var}(\psi|y)$, from any starting distribution, provided (i) each sequence is an irreducible, aperiodic, Harris recurrent Markov chain with stationary distribution $p(\omega|y)$, and (ii) $\text{E}(\psi^2|y) < \infty$.

These are exactly the hypotheses of the ergodic theorem applied to ψ and to ψ^2 along each chain, so for every j,

$$\bar{\psi}_{\cdot j} \to \mathrm{E}(\psi|y), \qquad s_j^2 \to \mathrm{var}(\psi|y) \quad \text{a.s.;}$$

Harris recurrence is what makes these limits independent of the starting distribution. Hence, splitting $\widehat{\mathrm{var}}^+ = \frac{n-1}{n} W + \frac{B}{n}$,

$$\frac{n-1}{n} W \to \mathrm{var}(\psi|y), \qquad \frac{B}{n} = \frac{1}{m-1} \sum_{j=1}^{m} \big(\bar{\psi}_{\cdot j} - \bar{\psi}_{\cdot\cdot}\big)^2 \to 0,$$

the second because all m chain means share the one limit; no central limit theorem is needed for this, only the ergodic theorem. Adding, $\widehat{\mathrm{var}}^+ \to \mathrm{var}(\psi|y)$ and $\widehat{R} \to 1$ in (11.4).
Both conditions bind. If the chain is reducible — each sequence trapped in a different mode — the $\bar{\psi}_{\cdot j}$ converge to different limits, B grows linearly in n, and B/n tends to a positive constant, so $\widehat{\mathrm{var}}^+$ overestimates forever; that is the failure Figure 11.3a exposes. If $\mathrm{E}(\psi^2|y) = \infty$, neither W nor B/n need settle.

Problem 11.6. Effective sample size:
(a) Derive the asymptotic formula (11.5) for the variance of the average of correlated simulations,

$$\lim_{n\to\infty} mn\,\mathrm{var}(\bar{\psi}_{\cdot\cdot}) = \Big(1 + 2\sum_{t=1}^{\infty} \rho_t\Big)\mathrm{var}(\psi|y),$$

where ρ_t is the autocorrelation of the sequence ψ at lag t.
(b) Implement a Markov chain simulation for some example and plot $\hat{n}_{\mathrm{eff}}$ from (11.8) over time. Is $\hat{n}_{\mathrm{eff}}$ stable? Does it gradually increase as a function of number of iterations, as one would hope?

Solution. (a) Sum the covariance matrix of one sequence and count the lags.
Assume the m sequences are independent and stationary with the target as marginal, so $\mathrm{var}(\psi_{ij}) = \mathrm{var}(\psi|y)$ and $\mathrm{cov}(\psi_{ij}, \psi_{i+t,j}) = \rho_t\,\mathrm{var}(\psi|y)$ for all i. Independence of the sequences gives

$$\mathrm{var}(\bar{\psi}_{\cdot\cdot}) = \mathrm{var}\Big(\tfrac{1}{m}\sum_{j=1}^{m} \bar{\psi}_{\cdot j}\Big) = \tfrac{1}{m}\,\mathrm{var}(\bar{\psi}_{\cdot 1}).$$

Within one sequence, collecting the n diagonal terms and the $2(n-t)$ terms at each lag t,

$$\begin{aligned}\mathrm{var}(\bar{\psi}_{\cdot 1}) &= \frac{1}{n^2}\Big[n\,\mathrm{var}(\psi|y) + 2\sum_{t=1}^{n-1}(n-t)\,\rho_t\,\mathrm{var}(\psi|y)\Big] \\ &= \frac{\mathrm{var}(\psi|y)}{n}\Big[1 + 2\sum_{t=1}^{n-1}\Big(1 - \frac{t}{n}\Big)\rho_t\Big].\end{aligned}$$

Therefore

$$mn\,\mathrm{var}(\bar{\psi}_{\cdot\cdot}) = n\,\mathrm{var}(\bar{\psi}_{\cdot 1}) = \Big[1 + 2\sum_{t=1}^{n-1}\Big(1 - \frac{t}{n}\Big)\rho_t\Big]\mathrm{var}(\psi|y).$$

If $\sum_{t\ge 1} |\rho_t| < \infty$ then $|(1 - t/n)\rho_t| \le |\rho_t|$ and $(1 - t/n)\rho_t \to \rho_t$ for each t, so dominated convergence over the counting measure on t gives the limit (11.5). (Summability is needed: with $\rho_t \equiv \rho > 0$ the bracket grows like n.)
(b) Yes on both counts, once the chains have converged: $\hat{n}_{\mathrm{eff}}$ settles into a fixed proportion of the saved draws and thereafter grows linearly.
Take the bioassay Metropolis simulation of Exercise 11.2 — five chains, normal jumps of scale $(1, 4)$, the same five overdispersed starting points — and run it to 20,000 iterations per chain. At each checkpoint N, discard the first $N/2$ iterations as warm-up, split the remainder in two ($m = 10$ chains of $n = N/4$), and evaluate $\widehat{R}$ from (11.4) and $\hat{n}_{\mathrm{eff}}$ from (11.7)-(11.8).

N	$\hat{n}^{\alpha}_{\text{eff}}$	$\widehat{R}^{\alpha}$	$\hat{n}^{\beta}_{\text{eff}}$	$\widehat{R}^{\beta}$
100	18	1.294	15	1.371
200	31	1.143	12	1.300
400	38	1.111	46	1.079
700	125	1.020	75	1.073
1000	83	1.021	66	1.052
2000	249	1.038	180	1.060
4000	393	1.023	389	1.025
6000	741	1.012	558	1.009
10000	976	1.003	862	1.003
14000	1607	1.006	1264	1.007
20000	2495	1.002	1705	1.002

Two regimes are visible. Before convergence ($N \le 1000$, $\widehat{R}$ still above 1.05 for β) the estimate is erratic and even decreases — at $N = 700$ it reads 125 for α, at $N = 1000$ only 83 — because $\widehat{\text{var}}^{+}$ in (11.7) is then badly inflated by the starting values, the estimated ρ_t are noisy, and the truncation lag T in (11.8) is unstable. Nothing about $\hat{n}_{\text{eff}}$ should be trusted while $\widehat{R} > 1.1$.

After convergence the ratio $\hat{n}_{\text{eff}}/(mn)$ stabilises at about 0.05 for α and 0.035 for β (integrated autocorrelation times roughly 20 and 30), and $\hat{n}_{\text{eff}}$ then grows essentially linearly in N: $741 \to 976 \to 1607 \to 2495$ as N goes $6000 \to 10000 \to 14000 \to 20000$. That is what (11.5) predicts, since $\text{var}(\bar{\psi}_{\cdot\cdot})$ decays like $1/(mn)$ with $1 + 2\sum \rho_t$ fixed. The growth is not monotone, though: $\hat{n}_{\text{eff}}$ is itself a noisy estimate, because the truncation point T in (11.8) and the tail estimates $\hat{\rho}_t$ both fluctuate, and repeat runs show single-checkpoint dips of 20 to 30 percent even at $\widehat{R} = 1.002$.

Problem 11.7. Analysis of survey data: Section 8.3 presents an analysis of a stratified sample survey using a hierarchical model on the stratum probabilities.

(a) Perform the computations for the simple nonhierarchical model described in the example.

(b) Using the Metropolis algorithm, perform the computations for the hierarchical model, using the results from part (a) as a starting distribution. Check by comparing your simulations to the results in Figure 8.1b.

The data are the CBS News survey of 1447 adults of Table 8.2, divided into $J = 16$ strata cross-classified by region and residential density. Sampling is assumed proportional, so $N_j/N \doteq n_j/n$.

Stratum, j	Bush, y_{1j}/n_j	Dukakis, y_{2j}/n_j	no opinion, y_{3j}/n_j	n_j/n
Northeast, I	0.30	0.62	0.08	0.032
Northeast, II	0.50	0.48	0.02	0.032
Northeast, III	0.47	0.41	0.12	0.115
Northeast, IV	0.46	0.52	0.02	0.048
Midwest, I	0.40	0.49	0.11	0.032
Midwest, II	0.45	0.45	0.10	0.065
Midwest, III	0.51	0.39	0.10	0.080
Midwest, IV	0.55	0.34	0.11	0.100
South, I	0.57	0.29	0.14	0.015
South, II	0.47	0.41	0.12	0.066
South, III	0.52	0.40	0.08	0.068
South, IV	0.56	0.35	0.09	0.126
West, I	0.50	0.47	0.03	0.023
West, II	0.53	0.35	0.12	0.053
West, III	0.54	0.37	0.09	0.086
West, IV	0.56	0.36	0.08	0.057

The models of Section 8.3 are $y_{\text{obs}\,j} \sim \text{Multin}(n_j; \theta_{1j}, \theta_{2j}, \theta_{3j})$ within each stratum; the nonhierarchical model puts independent Dirichlet$(1,1,1)$ priors on the 16 vectors, while the hierarchical model reparametrizes by (8.8), $\alpha_{1j} = \theta_{1j}/(\theta_{1j} + \theta_{2j})$ and $\alpha_{2j} = 1 - \theta_{3j}$, sets $\beta_{kj} = \text{logit}(\alpha_{kj})$, and takes (β_{1j}, β_{2j}) i.i.d. bivariate normal with mean (μ_1, μ_2), standard deviations (τ_1, τ_2) and correlation ρ,

with a uniform prior on those five hyperparameters. The estimand is (8.7), $\sum_{j=1}^{16}(N_j/N)(\theta_{1j}-\theta_{2j})$.

Solution. Posterior median of (8.7): 0.099 under the nonhierarchical model, 0.102 under the hierarchical model, with posterior standard deviations 0.0241 and 0.0247 — the hierarchical distribution shifted slightly up and slightly wider, as Figure 8.1 shows.

Counts. Table 8.2 reports only proportions, so the cell counts are reconstructed as $n_j = \text{round}(1447\, n_j/n)$ and $y_{kj} = \text{round}(n_j\, y_{kj}/n_j)$, adjusted by at most one in the largest cell so that each row sums to n_j. This gives n_j ranging from 22 (South, I) to 182 (South, IV), totalling 1442, and raw $\alpha_{1j} = y_{1j}/(y_{1j}+y_{2j})$ ranging from 0.33 to 0.68, against the 0.33 to 0.67 quoted on page 209 — so the reconstruction is right to within its rounding.

(a) Nonhierarchical model. Dirichlet-multinomial conjugacy (Section 3.4) gives independent exact posteriors, no MCMC needed:

$$(\theta_{1j},\theta_{2j},\theta_{3j})\,|\,y \sim \text{Dirichlet}(y_{1j}+1,\ y_{2j}+1,\ y_{3j}+1), \qquad j=1,\dots,16.$$

Drawing 20,000 independent sets of 16 such vectors and forming (8.7) with $N_j/N = n_j/n$:

$$\sum_j \frac{N_j}{N}(\theta_{1j}-\theta_{2j}) \ : \quad \text{median } 0.099,\ \ \text{sd } 0.0241,\ \ 95\%\ [0.051,\ 0.147].$$

The median sits just below the raw weighted difference 0.1017 because the Dirichlet$(1,1,1)$ prior adds one voter to each of the three categories in each of the 16 strata, pulling each stratum toward equal shares (page 209). Page 208 quotes 0.097 for this median; the 0.002 gap is the residue of reconstructing integer counts from two-decimal proportions, which is the accuracy limit for everything below.

(b) Hierarchical model. Write $\theta_{1j} = \alpha_{1j}\alpha_{2j}$, $\theta_{2j} = (1-\alpha_{1j})\alpha_{2j}$, $\theta_{3j} = 1-\alpha_{2j}$ with $\alpha_{kj} = \text{logit}^{-1}(\beta_{kj})$. The prior is placed directly on β, so no Jacobian is needed in the likelihood, and

$$\begin{aligned}\log p(\beta,\mu,\tau,\rho|y) \doteq \sum_{j=1}^{16}\Big[&y_{1j}\log(\alpha_{1j}\alpha_{2j}) + y_{2j}\log\big((1-\alpha_{1j})\alpha_{2j}\big)\\ &+ y_{3j}\log(1-\alpha_{2j})\Big]\\ &- 16\log(\tau_1\tau_2) - 8\log(1-\rho^2) - \tfrac{1}{2}\sum_{j=1}^{16} Q_j,\end{aligned}$$

where, with $d_{kj} = \beta_{kj} - \mu_k$,

$$Q_j = \frac{1}{1-\rho^2}\left(\frac{d_{1j}^2}{\tau_1^2} - \frac{2\rho\, d_{1j}d_{2j}}{\tau_1\tau_2} + \frac{d_{2j}^2}{\tau_2^2}\right).$$

Algorithm (Metropolis as a building block inside a Gibbs structure, Section 11.3). One sweep:

(i) for $j = 1,\dots,16$ and $k = 1,2$, a one-dimensional symmetric normal Metropolis jump on β_{kj}, accepted by (11.1) on the log scale;

(ii) μ by Gibbs — with a uniform prior, $\mu\,|\,\beta,\tau,\rho,y \sim \text{N}\big(\bar\beta, \Sigma/16\big)$, where Σ is the 2×2 hyperparameter covariance;

(iii) five rounds of one-dimensional Metropolis jumps on $(\log\tau_1, \log\tau_2, \operatorname{arctanh}\rho)$; jumping on the unconstrained scale requires the Jacobian terms $+\log\tau_1 + \log\tau_2 + \log(1-\rho^2)$ in the target, which restores the intended uniform prior on (τ_1,τ_2,ρ) itself.

The hyperparameters are updated five times per sweep because the funnel geometry near $\tau_k = 0$ is the bottleneck: with one update per sweep the chains move between the small-τ and large-τ regions too slowly and $\widehat{R}$ for $\log\tau_1$ stays above 1.1.

Starting distribution and run. Each chain starts at a draw from part (a): sample the 16 Dirichlet posteriors, map to $(\alpha_{1j},\alpha_{2j})$ by (8.8), set $\beta_{kj} = \text{logit}(\alpha_{kj})$, and take μ, τ, ρ as the empirical mean, standard deviations and correlation of those 16 points. Six chains, 4000 adaptive warm-up sweeps

(proposal scales tuned to 0.44 acceptance, then frozen) followed by 20,000 saved sweeps. All $\widehat{R} \le 1.03$ (τ_1, τ_2 monitored on the log scale).

Results.

Estimand	2.5%	25%	median	75%	97.5%
Stratum 1, $\alpha_{1,1}$	0.37	0.47	0.51	0.55	0.58
Stratum 2, $\alpha_{1,2}$	0.45	0.52	0.54	0.56	0.61
Stratum 8, $\alpha_{1,8}$	0.52	0.56	0.58	0.60	0.65
Stratum 16, $\alpha_{1,16}$	0.50	0.55	0.56	0.59	0.64
Stratum 16, $\alpha_{2,16}$	0.88	0.90	0.91	0.92	0.94
$\text{logit}^{-1}(\mu_1)$	0.51	0.54	0.55	0.56	0.59
$\text{logit}^{-1}(\mu_2)$	0.89	0.90	0.91	0.92	0.93
τ_1	0.02	0.09	0.15	0.22	0.38
τ_2	0.01	0.07	0.14	0.23	0.48
ρ	−0.97	−0.66	−0.23	0.30	0.93

The posterior medians of the α_{1j} run from 0.51 to 0.58, heavy shrinkage from the raw range 0.33-0.68, with $\text{logit}^{-1}(\mu_1) = 0.55$, $\text{logit}^{-1}(\mu_2) = 0.91$ and a negative median for ρ with enormous spread. For the finite-population estimand (8.7),

$$\text{median } 0.102, \quad \text{sd } 0.0247, \quad 95\% \ [0.054, 0.151],$$

against 0.099 and 0.0241 in part (a): higher median and slightly more variability, the two features the text attributes to Figure 8.1b. The mechanism is the one described on page 210 — strata 1 and 5, where support for Bush is lowest, have among the smallest samples ($n_1 = n_5 = 46$) and so are pulled hardest toward the grand mean.

Discrepancy flag. Every hyperparameter above comes out smaller than in Table 8.3: τ_1 median 0.15 against 0.23, τ_2 0.14 against 0.28, ρ −0.23 against −0.44, and correspondingly more shrinkage in the α_{1j} (stratum 1 median 0.51 against 0.48, range 0.51-0.58 against 0.48-0.59); the estimand median is 0.102 against the 0.11 quoted on page 210. The mathematics favors the smaller values. Indeed the printed 0.11 is inconsistent with Table 8.3 itself: substituting that table's own α_{1j} medians and the observed α_{2j} into (8.7) gives 0.099. Replacing the multinomial likelihood by its normal approximation $\hat{\beta}_{1j} \sim \text{N}(\beta_{1j}, y_{1j}^{-1} + y_{2j}^{-1})$ turns the β_1 component into the model of Section 5.4, whose marginal posterior for τ_1 is evaluated exactly on a grid by (5.21) with no MCMC at all; that grid gives quantiles $(0.01, 0.07, 0.135, 0.21, 0.37)$, matching the Metropolis output. The moment estimate $\hat{\tau}_1^2 = \text{var}(\hat{\beta}_{1j}) - \overline{v_j} = 0.120 - 0.071$, i.e. $\hat{\tau}_1 = 0.22$, is the likelihood-based value that Table 8.3 reports as a median; the posterior median falls below it because with only 16 strata the likelihood for τ_1 stays essentially flat down to 0, and a sampler that fails to reach small τ — the funnel bottleneck noted above — reproduces Table 8.3 instead.

12 Computationally Efficient Markov Chain Simulation

Contents

12.1 Exercises 12.1–12.4

Problem 12.1. Efficient Metropolis jumping rules: Repeat the computation for Exercise 11.2 using the adaptive algorithm given in Section 12.2.
(Exercise 11.2 asks: replicate the computations for the bioassay example of Section 3.7 using the Metropolis algorithm, defining starting points and jumping rule, computing with log-densities, and running long enough for approximate convergence. The bioassay data of Table 1.3 are

Dose x_j (log g/ml)	Animals n_j	Deaths y_j
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

with model $y_j \sim \text{Binomial}(n_j, \theta_j)$, $\text{logit}(\theta_j) = \alpha + \beta x_j$, and a uniform prior on (α, β).
The adaptive algorithm of Section 12.2 is: (1) start the parallel simulations with a normal random-walk jumping rule shaped like an estimate of the target, namely the curvature matrix at the posterior mode scaled by $2.4/\sqrt{d}$; (2) after some number of simulations, reset the jumping covariance proportional to the posterior covariance estimated from the simulations so far, and increase or decrease the scale to push the acceptance rate toward the optimal 0.44 in one dimension, 0.23 in high dimensions. All tuning is confined to an adaptive phase; only the subsequent fixed phase is used for inference.)

Solution. The adaptive phase converges on the jumping kernel

$$J(\theta^* \mid \theta^{t-1}) = \text{N}\left(\theta^{t-1}, \begin{pmatrix} 2.54 & 7.18 \\ 7.18 & 57.01 \end{pmatrix}\right),$$

which accepts 38% of jumps and reproduces the Chapter 3 grid posterior to within simulation error. The log posterior, computed on the log scale as page 261 demands, is

$$\log p(\alpha, \beta \mid y) = \sum_{j=1}^{4} \Big[y_j \eta_j - n_j \log\big(1 + e^{\eta_j}\big) \Big] + \text{const}, \qquad \eta_j = \alpha + \beta x_j.$$

Adaptive phase, step 1. The posterior mode is $(\hat\alpha, \hat\beta) = (0.847, 7.749)$ and the inverse of the negative Hessian there is

$$\hat\Sigma = \begin{pmatrix} 1.039 & 3.546 \\ 3.546 & 23.74 \end{pmatrix},$$

so the initial jumping rule is $\text{N}(\theta^{t-1}, c^2\hat\Sigma)$ with $c = 2.4/\sqrt{2} = 1.70$. Five chains are started at $(\hat\alpha, \hat\beta)$ plus independent normal noise of standard deviation $(2, 5)$, which is overdispersed relative to the posterior.
Step 2. Eight adaptation rounds of 500 iterations each; after each round $\hat\Sigma$ is replaced by the sample covariance of the second half of the pooled draws and c is multiplied by $\exp\{1.5(\hat r - 0.35)\}$, where $\hat r$ is the round's acceptance rate and 0.35 is the target for $d = 2$ (interpolating between the 0.44 and 0.23 of Section 12.2):

Round	$\hat r$	c	sd(α)	sd(β)	corr
1	0.390	1.803	1.11	5.41	0.63
2	0.310	1.698	1.15	5.42	0.66
4	0.315	1.530	1.13	5.66	0.70
6	0.370	1.650	1.03	5.90	0.62
8	0.327	1.565	1.02	4.83	0.60

The scale settles at $c = 1.57$, essentially the theoretical $2.4/\sqrt{d} = 1.70$, and the shape settles at the posterior covariance, which is more elongated than the modal curvature $\hat\Sigma$ (posterior sd(β) ≈ 5.5 against $\sqrt{23.74} = 4.87$), the long right tail in β being invisible to the mode.
Fixed phase. The adapted rule is then frozen and five chains are run for 10,000 iterations, the first half of each discarded. The acceptance rate is 0.381 and, with the split-$\hat R$ and n_{eff} of Section 11.4,

	$\hat R$	n_{eff}	2.5%	25%	median	75%	97.5%
α	1.003	2433	-0.58	0.54	1.20	1.95	3.81
β	1.002	1818	3.46	7.25	10.56	14.86	26.25

from 25,000 saved draws, i.e. an efficiency of 0.10 per draw for α, consistent with the $0.3/d = 0.15$ quoted in Section 12.2 for a normal target.
Agreement with Chapter 3. Numerical integration of $p(\alpha, \beta \mid y)$ on the $(\alpha, \beta) \in [-5, 10] \times [-10, 40]$ grid of Figure 3.3 gives quantiles $(-0.59, 0.54, 1.22, 1.98, 3.72)$ for α and $(3.46, 7.37, 10.67, 14.83, 25.29)$ for β, matching the table above to within Monte Carlo error. For the LD50 $= -\alpha/\beta$ the simulations give a 95% interval $[-0.276, 0.105]$ and median -0.111, against the grid's $[-0.276, 0.103]$ and -0.112, and $\Pr(\beta > 0 \mid y) > 0.9999$ under both, so the conditioning on $\beta > 0$ in Section 3.7 is immaterial.

Problem 12.2. Simulated tempering: Consider the Cauchy model, $y_i \sim \text{Cauchy}(\theta, 1)$, $i = 1, \ldots, n$, with uniform prior on θ, and two data points, $y_1 = 1.3$, $y_2 = 15.0$.
(a) Graph the posterior density.
(b) Program the Metropolis algorithm for this problem using a symmetric Cauchy jumping distribution. Tune the scale parameter of the jumping distribution appropriately.
(c) Program simulated tempering with a ladder of 10 inverse-temperatures, $0.1, \ldots, 1$.
(d) Compare your answers in (b) and (c) to the graph in (a).

Solution. The posterior is the symmetric bimodal density

$$p(\theta \mid y) = \frac{1}{Z} \frac{1}{\big(1 + (\theta - 1.3)^2\big)\big(1 + (\theta - 15.0)^2\big)}, \qquad Z = 0.032778,$$

and both samplers reproduce it, the plain Cauchy-jump Metropolis algorithm about four times more cheaply than tempering.
(a) Since $p(\theta \mid y)$ depends on θ only through $(\theta - 1.3)^2$ and $(\theta - 15)^2$, it is exactly symmetric about $\bar{y} = 8.15$: the reflection $\theta \mapsto 16.3 - \theta$ swaps the two factors. Hence the two modes carry exactly half the mass each, the posterior mean and median are both 8.15, and

$$\Pr(\theta < 8.15 \mid y) = \tfrac{1}{2}$$

exactly – a sharp target against which to judge any sampler's mode mixing. The modes are at $\theta = 1.3734$ and $\theta = 14.9266$, where the density is 0.1625, and the antimode at 8.15, where it is 0.01328; the trough is thus only a factor 12.2 below the peaks, and the central quantiles are

$$(2.5\%, 25\%, 50\%, 75\%, 97.5\%) = (-1.45,\ 1.61,\ 8.15,\ 14.69,\ 17.75).$$

Plotted on $[-10, 25]$ the density is two equal spikes of height 0.163 at $\theta = 1.37$ and 14.93, separated by a shallow trough of height 0.0133, with tails decaying like θ^{-4}.
(b) Jumping rule $J(\theta^* \mid \theta^{t-1}) = \text{Cauchy}(\theta^{t-1}, s)$, which is symmetric, so the acceptance ratio is the plain density ratio. Four chains of 20,000 iterations, two started at each mode:

s	accept	$\hat{R}$	n_{eff}	$\widehat{\Pr}(\theta < 8.15)$	mode crossings/chain
0.5	0.765	1.004	711	0.529	367
2	0.525	1.003	2091	0.500	950
4	0.400	1.002	3913	0.487	1343
8	0.286	1.001	5157	0.503	1642
10	0.257	1.001	5438	0.506	1693
12	0.230	1.001	5409	0.486	1652
20	0.163	1.001	5048	0.488	1392

The efficiency is flat over $s \in [8, 12]$; take $s = 10$, acceptance 0.26, $n_{\text{eff}} = 5438$ out of 80,000 draws, so $\widehat{\Pr}(\theta < 8.15) = 0.506$ has Monte Carlo standard error 0.007. Even $s = 0.5$ crosses the trough 367 times per chain, because a Cauchy$(0, s)$ jump clears the mode separation 13.55 with probability

$$\Pr(|\text{jump}| > 13.55) = \tfrac{2}{\pi} \arctan(s/13.55) = 0.0235 \quad \text{at } s = 0.5,$$

which over 20,000 iterations is 470 attempted crossings.

(c) Simulated tempering augments θ with a level indicator $k \in \{1, \dots, 10\}$, $\beta_k = k/10$, and targets

$$p(\theta, k) \;\propto\; c_k \left[q(\theta)\right]^{\beta_k}, \qquad q(\theta) = \prod_{i=1}^{2} \left(1 + (\theta - y_i)^2\right)^{-1}.$$

The ladder needs a support restriction: q^β decays like $|\theta|^{-4\beta}$, so it is integrable on $\mathbb{R}$ only for $\beta > 1/4$; the ladder rungs $\beta = 0.1, 0.2$ are improper under the uniform prior. Restrict θ to $[-50, 65]$, which contains all the posterior mass to machine precision, and the ladder is well defined. The normalizing constants

$$Z_k = \int_{-50}^{65} q(\theta)^{\beta_k}\, d\theta$$

are computed by quadrature,

$$\log Z_k = (3.564,\ 2.500,\ 1.551,\ 0.701,\ -0.075,\ -0.798,\ -1.485,\ -2.147,\ -2.790,\ -3.418),$$

and we set $c_k = 1/Z_k$ so that every level is visited equally often.
Each iteration is (i) a Cauchy-jump Metropolis update of θ at the current level, with scale $s_k = \min(10/\beta_k, 40)$, then (ii) a proposal $k \to k \pm 1$ with probability $\frac{1}{2}$ each, accepted with probability

$$\min\left(1,\ \frac{c_{k^*}}{c_k}\left[q(\theta)\right]^{\beta_{k^*} - \beta_k}\right).$$

Four chains of 200,000 iterations give within-level acceptance 0.320, level-move acceptance 0.912, occupancy $(0.100, 0.099, 0.099, 0.100, 0.101, 0.101, 0.100, 0.100, 0.100, 0.099)$ – flat, as the choice $c_k = 1/Z_k$ requires – and about 950 round trips from $\beta = 0.1$ to $\beta = 1$ per chain. Retaining only the 79,001 draws with $k = 10$ gives $\hat{R} = 1.000$, $n_{\text{eff}} = 14{,}861$, $\widehat{\Pr}(\theta < 8.15) = 0.491$ (standard error 0.004), and quantiles

$$(-1.49,\ 1.63,\ 8.81,\ 14.70,\ 17.81).$$

(d) Both match (a), and tempering costs about four times as much per effective draw. Against the exact $(-1.45, 1.61, 8.15, 14.69, 17.75)$, the $s = 10$ Metropolis run gives $(-1.34, 1.54, 7.68, 14.68, 17.74)$ and the tempered run the tuple in (c). The median is the one loose entry in each, off by 0.47 and 0.66, but the median's Monte Carlo standard error is $1/(2p(8.15)\sqrt{n_{\text{eff}}})$, here 0.51 and 0.31 because the antimode density is only 0.0133: deviations of 0.9 and 2.1 standard errors, not bias. Both estimates of the exact $\Pr(\theta < 8.15) = \frac{1}{2}$ sit within two standard errors (0.506 ± 0.007, 0.491 ± 0.004). On cost, 14,861 effective draws per 800,000 iterations is 0.019 each against 5438 per 80,000, or 0.068, for the tuned Cauchy Metropolis: nine tenths of the tempered iterations sit at discarded levels, and the heavy-tailed kernel already tunnels between the modes without them.

Problem 12.3. Hamiltonian Monte Carlo: Program HMC in R for the bioassay logistic regression example from Chapter 3.
(a) Code the gradients analytically and numerically and check that the two programs give the same result.
(b) Pick reasonable starting values for the mass matrix, step size, and number of steps.
(c) Tune the algorithm to an approximate 65% acceptance rate.
(d) Run 4 chains long enough so that each has an effective sample size of at least 100. How many iterations did you need?
(e) Check that your inferences are consistent with those from the direct approach in Chapter 3.
(The bioassay data of Table 1.3 are

Dose x_j (log g/ml)	Animals n_j	Deaths y_j
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

with $y_j \sim \text{Binomial}(n_j, \theta_j)$, $\text{logit}(\theta_j) = \alpha + \beta x_j$ and a uniform prior on (α, β).)

Solution. Tuned to $\epsilon = 0.72$, $L = 3$ in the metric of the posterior standard deviations, HMC accepts 64% of trajectories and needs only 400 iterations per chain to clear $n_{\text{eff}} = 100$.
(a) With $\eta_j = \alpha + \beta x_j$ and $\theta_j = \text{logit}^{-1}(\eta_j)$,

$$\log p(\alpha, \beta \mid y) = \sum_{j=1}^{4} \Big[y_j \eta_j - n_j \log\big(1 + e^{\eta_j}\big) \Big] + \text{const},$$

$$\frac{\partial \log p}{\partial \alpha} = \sum_{j=1}^{4} (y_j - n_j \theta_j), \qquad \frac{\partial \log p}{\partial \beta} = \sum_{j=1}^{4} x_j (y_j - n_j \theta_j),$$

using $\frac{d}{d\eta} \log(1 + e^{\eta}) = \text{logit}^{-1}(\eta)$. Against central differences with $h = 10^{-5}$, e.g. at $(\alpha, \beta) = (0.8, 7.0)$ the analytic gradient is $(-0.137388,\ 0.056907)$ and the numerical one agrees to 1.6×10^{-10}; at $(-1, 2)$, $(3, 20)$ and $(0, 0)$ the discrepancies are 3.8×10^{-11}, 8.0×10^{-10} and 3.8×10^{-11}. (Check!)
(b) Mass matrix. Section 12.4 sets the momentum scales to a crude estimate of the scale of the target, so take the diagonal

$$M = \text{diag}\big(1/1.1^2,\ 1/5.5^2\big),$$

so that the momenta $\phi \sim \text{N}(0, M)$ and the leapfrog updates $\theta \leftarrow \theta + \epsilon M^{-1} \phi$ move in units of a posterior standard deviation in each coordinate. Step size and steps: Section 12.4 sets $\epsilon L = 1$ with default $\epsilon = 0.1$, $L = 10$, which is the starting point here; part (c) then pushes it to $\epsilon = 0.72$, $L = 3$, i.e. $\epsilon L = 2.2$, the small L being affordable because $d = 2$. Starting values: four draws from the normal approximation $\text{N}(\hat{\theta}, \hat{\Sigma})$ at the posterior mode $\hat{\theta} = (0.847,\ 7.749)$ with the curvature matrix $\hat{\Sigma}$ of Exercise 12.1, rejecting any draw with $\beta \le 1$, which is harmless because $\Pr(\beta > 0 \mid y) > 0.9999$. A far-out start does not destabilize the leapfrog integrator here – the gradient of part (a) is bounded, $|\partial \log p / \partial \alpha| \le \sum_j n_j = 20$ and $|\partial \log p / \partial \beta| \le \sum_j n_j |x_j| = 9.7$ – it only lengthens the warm-up.
(c) Holding $L = 3$ and scanning ϵ:

ϵ	0.66	0.70	0.72	0.74	0.78
accept	0.720	0.672	0.643	0.617	0.567

so $\epsilon = 0.72$ gives the required 65%.
(d) 400 iterations per chain, half discarded as warm-up. Over 40 independent replications of 4 chains at $\epsilon = 0.72$, $L = 3$:

iterations/chain	reps with $\hat{R} < 1.1$ and $n_{\text{eff}} > 100$	median $\min(n^{\alpha}_{\text{eff}}, n^{\beta}_{\text{eff}})$
100	1/40	61
200	28/40	122
400	39/40	264
1000	40/40	638

At 400 iterations all but one replication passed; at 200 the median n_{eff} already exceeds 100 but three runs in ten had not. The binding constraint is β, whose n_{eff} runs well below that of α.
(e) A run of 4 chains × 1000 iterations (500 saved each) has acceptance 0.641 and

	$\hat{R}$	n_{eff}	2.5%	25%	median	75%	97.5%
α	1.004	912	-0.70	0.59	1.23	2.02	3.88
β	1.009	547	3.41	7.11	10.62	15.00	26.47

against the Chapter 3 grid values $(-0.59,\ 0.54,\ 1.22,\ 1.98,\ 3.72)$ for α and $(3.46,\ 7.37,\ 10.67,\ 14.83,\ 25.29)$ for β: consistent throughout, the tail quantiles being the noisiest, as $n_{\text{eff}} \approx 550$ implies. The LD50 $= -\alpha/\beta$ has posterior median -0.113 and 95% interval $[-0.294,\ 0.111]$, matching the grid's -0.112 and $[-0.276,\ 0.103]$ and the value reported in Section 3.7.

Problem 12.4. Coverage of intervals and rejection sampling: Consider the following model: $y_j \sim \text{Binomial}(n_j, \theta_j)$, where $\theta_j = \text{logit}^{-1}(\alpha + \beta x_j)$, for $j = 1, \ldots, J$, and with independent prior distributions, $\alpha \sim t_4(0, 2^2)$ and $\beta \sim t_4(0, 1)$. Assume $J = 10$, the x_j values are randomly drawn from a $\text{U}(-1, 1)$ distribution, and $n_j \sim \text{Poisson}^+(5)$, where Poisson^+ is the Poisson distribution restricted to positive values.
(a) Sample a dataset at random from the model, estimate α and β using Stan, and make a graph simultaneously displaying the data, the fitted model, and uncertainty in the fit (shown via a set of inverse logit curves that are thin and gray).
(b) Did Stan's posterior 50% interval for α contain its true value? How about β?
(c) Use rejection sampling to get 1000 independent posterior draws from (α, β).

Solution. Yes and yes: the 50% intervals $[0.53, 0.95]$ for α and $[-0.39, 0.20]$ for β cover the true $(0.800, 0.129)$, and rejection sampling off the prior itself accepts 5.2% of proposals, so 19,140 prior draws yield the 1000 independent posterior draws. (The printed $\text{U}(1, 1)$ is a typo for $\text{U}(-1, 1)$; $t_4(0, 2^2)$ means scale 2, i.e. $\alpha = 2T$ with $T \sim t_4$.)
(a) One draw from the model gives $\alpha = 0.800$, $\beta = 0.1288$ and

j	1	2	3	4	5	6	7	8	9	10
x_j	0.290	0.901	0.455	-0.898	-0.425	-0.984	-0.707	-0.739	0.146	-0.807
n_j	3	5	8	4	9	6	5	3	4	3
y_j	1	3	6	4	5	4	2	3	4	2

The log posterior is

$$\log p(\alpha, \beta \mid y) = \sum_{j=1}^{10} \Big[y_j \eta_j - n_j \log\big(1 + e^{\eta_j}\big) \Big] - \tfrac{5}{2} \log\Big(1 + \tfrac{\alpha^2}{16}\Big) - \tfrac{5}{2} \log\Big(1 + \tfrac{\beta^2}{4}\Big) + \text{const},$$

with $\eta_j = \alpha + \beta x_j$, whose gradient adds $-\frac{5\alpha/16}{1+\alpha^2/16}$ and $-\frac{5\beta/4}{1+\beta^2/4}$ to the two likelihood scores of Exercise 12.3. In place of Stan the posterior is sampled with the HMC program of Exercise 12.3 – the same algorithm Stan runs – carrying these extra prior terms, with $M = \text{diag}(1/0.33^2, 1/0.44^2)$, $\epsilon = 0.7$, $L = 4$, four chains of 4000 iterations, half discarded. Acceptance is 0.963, $\hat{R} = 1.000$ for both parameters, and n_{eff} equals the 8000 saved draws to within the estimator's resolution – HMC produces essentially independent draws on this smooth two-dimensional posterior. Posterior quantiles:

	2.5%	25%	50%	75%	97.5%	mean
α	0.137	0.533	0.735	0.948	1.391	0.742
β	-0.969	-0.393	-0.093	0.204	0.777	-0.096

Numerical integration on a 1201×1201 grid over $[-8, 10] \times [-10, 12]$ confirms these: $(0.130, 0.520, 0.730, 0.955, 1.390)$ and $(-0.943, -0.393, -0.100, 0.193, 0.762)$. The graph plots the ten observed proportions y_j/n_j against x_j with area proportional to n_j, overlays one hundred thin gray curves $\text{logit}^{-1}(\alpha^{(s)} + \beta^{(s)} x)$ at posterior draws, and adds the posterior-mean curve in black and the true curve dashed. The gray fan is wide and its slopes take both signs: with $\sum n_j = 50$ and $|x_j| \le 1$ the data pin down β only to $\text{sd}(\beta \mid y) = 0.435$ against the prior's $\sqrt{2}$, leaving $\Pr(\beta > 0 \mid y) = 0.41$, so the sign of the slope is undetermined even though the true value is positive.
(b) Yes for both. The 50% interval for α is $[0.533, 0.948]$ and the true value is 0.800; for β it is $[-0.393, 0.204]$ and the true value is 0.1288. Repeating the whole experiment – draw (α, β) from the prior, draw x, n, y, form the posterior 50% interval – over 1000 replications gives coverage 0.505 for α and 0.496 for β, each within one Monte Carlo standard error (0.016) of the 0.50 that Bayesian intervals must attain when averaged over the prior.
(c) Take the prior itself as the proposal. Since the likelihood is a product of binomial probabilities,

$$p(\alpha, \beta \mid y) \;\propto\; p(\alpha)p(\beta)L(\alpha, \beta) \;\le\; L_{\max}\, p(\alpha)p(\beta), \qquad L_{\max} = \sup_{\alpha,\beta} L(\alpha, \beta),$$

so $g(\alpha, \beta) = p(\alpha)p(\beta)$ and $M = L_{\max}$ satisfy the envelope condition of Section 10.3 exactly, with no numerical search for a bound. Here $L_{\max}$ is attained at the maximum likelihood estimate

(0.7274, −0.1150), with $\log L_{\max} = -31.3145$. The algorithm is: draw $\alpha = 2T_1$, $\beta = T_2$ with $T_1, T_2 \sim t_4$ independent, draw $u \sim \mathrm{U}(0,1)$, and accept (α, β) if

$$\log u \;<\; \log L(\alpha, \beta) - \log L_{\max}.$$

Accepted draws are exactly i.i.d. from $p(\alpha, \beta \mid y)$, and the acceptance probability is $p(y)/L_{\max}$, which quadrature puts at 0.05215. Out of 400,000 proposals, 20,899 were accepted, a rate of 0.0522; the first 1000 of them are the required sample, so $1000/0.0522 = 19{,}140$ proposals suffice in expectation. Their quantiles,

$$(0.139,\ 0.524,\ 0.733,\ 0.951,\ 1.380) \quad \text{and} \quad (-0.936,\ -0.381,\ -0.092,\ 0.199,\ 0.763),$$

agree with the HMC and grid values above.

13 Modal and Distributional Approximations

Contents

13.1 Exercises 13.1–13.7

Problem 13.1. Multimodality: Consider a simple one-parameter model of independent data, $y_i \sim \text{Cauchy}(\theta, 1)$, $i = 1, \ldots, n$, with uniform prior density on θ. Suppose $n = 2$.
(a) Prove that the posterior distribution is proper.
(b) Under what conditions will the posterior density be unimodal?

Solution. (a) The unnormalized posterior integrates to $2\pi/(4 + (y_1 - y_2)^2) < \infty$.
With the uniform prior, $q(\theta|y) = \prod_{i=1}^{2}\big(1 + (y_i - \theta)^2\big)^{-1}$, and since each factor is at most 1,

$$\int_{-\infty}^{\infty} q(\theta|y)\, d\theta \;\le\; \int_{-\infty}^{\infty} \frac{d\theta}{1 + (y_1 - \theta)^2} \;=\; \pi \;<\; \infty.$$

The exact value follows from the convolution identity for Cauchy densities (Appendix A: a sum of independent Cauchy$(0, 1)$ variables is Cauchy$(0, 2)$). Writing f for the standard Cauchy density and substituting $u = y_1 - \theta$,

$$\begin{aligned}\int q(\theta|y)\, d\theta &= \pi^2 \int f(y_1 - \theta) f(y_2 - \theta)\, d\theta \\ &= \pi^2 (f * f)(y_1 - y_2) = \frac{2\pi}{4 + (y_1 - y_2)^2},\end{aligned}$$

using $f(-x) = f(x)$. Hence $p(\theta|y)$ is proper for every observed pair (y_1, y_2).
(b) Unimodal exactly when $|y_1 - y_2| \le 2$.
Differentiate the log posterior and put $a = y_1 - \theta$, $b = y_2 - \theta$:

$$\begin{aligned}\tfrac{1}{2}\frac{d}{d\theta}\log p(\theta|y) &= \frac{a}{1 + a^2} + \frac{b}{1 + b^2} \\ &= \frac{a(1 + b^2) + b(1 + a^2)}{(1 + a^2)(1 + b^2)} = \frac{(a + b)(1 + ab)}{(1 + a^2)(1 + b^2)}.\end{aligned}$$

So the stationary points are the roots of $a + b = 0$ and of $ab = -1$, that is

$$\theta = \frac{y_1 + y_2}{2} \quad \text{and} \quad \theta = \frac{y_1 + y_2}{2} \pm \frac{\sqrt{(y_1 - y_2)^2 - 4}}{2},$$

call the latter pair $\theta_\pm$; they are real only when $|y_1 - y_2| \ge 2$.
(i) $|y_1 - y_2| < 2$: the midpoint is the unique stationary point, and since $p(\theta|y) \to 0$ as $\theta \to \pm\infty$ it is the mode. Unimodal.
(ii) $|y_1 - y_2| = 2$: all three roots collapse to $\theta = (y_1 + y_2)/2$, still the only stationary point. Unimodal (with a flat, fourth-order maximum).
(iii) $|y_1 - y_2| > 2$: three distinct stationary points, $\theta_- < \frac{y_1 + y_2}{2} < \theta_+$. The sign of $(a + b)(1 + ab)$ runs $+, -, +, -$ as θ crosses them in increasing order (Check!), so $\theta_\pm$ are modes and the midpoint is a local minimum. Bimodal, the two modes of equal height by the symmetry of $p(\theta|y)$ about the midpoint.

Problem 13.2. Normal approximation and importance resampling:
(a) Repeat Exercise 3.12 using the normal approximation to produce posterior simulations for (α, β).
(b) Use importance resampling to improve on the normal approximation.
(c) Compute the importance ratios for your simulations. Plot a histogram of the importance ratios and comment on their distribution. Compute an estimate of effective sample size using (10.4) on page 266.
Exercise 3.12 fits the Poisson regression $y_t \sim \text{Poisson}(\alpha + \beta t)$ to the counts of fatal accidents on scheduled airline flights, $t = 1976, \ldots, 1985$ (Table 2.2). It asks for a noninformative prior distribution for (α, β); take the uniform density on the region where $\alpha + \beta t > 0$ for all t in the data. It also asks

for the posterior distribution of the expected number of fatal accidents in 1986, $\alpha + 1986\beta$, and a 95% predictive interval for the number of fatal accidents in 1986.

Year	Fatal accidents	Passenger deaths	Death rate
1976	24	734	0.19
1977	25	516	0.12
1978	31	754	0.15
1979	31	877	0.16
1980	22	814	0.14
1981	21	362	0.06
1982	26	764	0.13
1983	20	809	0.13
1984	16	223	0.03
1985	22	1066	0.15

Solution. The normal approximation is already very accurate here: the importance ratios have standard deviation 0.18 on the log scale and (10.4) gives $S_{\text{eff}}/S = 0.97$, so importance resampling barely moves the answer.

(a) Normal approximation. Write $\mu_t = \alpha + \beta t$ and reparameterize by $\alpha^* = \alpha + 1980.5\,\beta$, so that $\mu_t = \alpha^* + \beta x_t$ with $x_t = t - 1980.5$ and $\sum_t x_t = 0$; this is a linear bijection, so the normal approximation and all importance ratios are unchanged by it. The log posterior and its derivatives are

$$\log p(\alpha^*, \beta | y) = \sum_t \big(-\mu_t + y_t \log \mu_t\big) + \text{const},$$
$$\frac{\partial \log p}{\partial \gamma_k} = \sum_t \Big(\frac{y_t}{\mu_t} - 1\Big) z_{tk},$$
$$\frac{\partial^2 \log p}{\partial \gamma_k \partial \gamma_l} = -\sum_t \frac{y_t}{\mu_t^2}\, z_{tk} z_{tl},$$

with $\gamma = (\alpha^*, \beta)$ and $z_t = (1, x_t)$. Newton's method on the two score equations converges in a few steps to

$$\hat\alpha^* = 23.800, \qquad \hat\beta = -0.94933,$$

equivalently $(\hat\alpha, \hat\beta) = (1903.95, -0.94933)$ on the calendar-year scale. The value $\hat\alpha^* = \bar y$ is exact: multiplying the first score equation by $\hat\alpha^*$, the second by $\hat\beta$, and adding gives $\sum_t \mu_t(y_t/\mu_t) = \hat\alpha^* n$, i.e. $\sum_t y_t = n\hat\alpha^*$.

The curvature matrix at the mode inverts to

$$V_\gamma = \begin{pmatrix} 2.3800 & -0.09493 \\ -0.09493 & 0.29203 \end{pmatrix},$$

so $\text{sd}(\alpha^*) = 1.543$, $\text{sd}(\beta) = 0.540$, correlation -0.114. Drawing $S = 10{,}000$ values of (α^*, β) from $\text{N}(\hat\gamma, V_\gamma)$ – and mapping back by $\alpha = \alpha^* - 1980.5\beta$ – gives the approximate posterior simulations. Every draw satisfied $\mu_t > 0$ throughout 1976–1986, so the constraint is inactive.

(b) Importance resampling. The importance ratios are

$$w(\gamma^s) = \frac{p(y|\gamma^s)}{\exp\big(-\frac{1}{2}(\gamma^s - \hat\gamma)' V_\gamma^{-1} (\gamma^s - \hat\gamma)\big)},$$

and resampling 1000 of the 10,000 draws without replacement with probabilities proportional to w yields

$$\text{E}(\alpha^*|y) = 24.00, \quad \text{sd}(\alpha^*|y) = 1.57,$$
$$\text{E}(\beta|y) = -0.94, \quad \text{sd}(\beta|y) = 0.54,$$

with Monte Carlo standard errors $1.57/\sqrt{1000} = 0.05$ and 0.02; a fine grid quadrature of the exact posterior confirms $\text{E}(\alpha^*|y) = 24.000$, $\text{E}(\beta|y) = -0.940$, sd = 1.549 and 0.540. Only α^* moves appreciably from its modal value (23.80 → 24.00, about 0.13 posterior standard deviations), which is the

skewness correction the normal approximation misses; the shift in β is smaller than the Monte Carlo error. For the 1986 expectation $\alpha + 1986\beta = \alpha^* + 5.5\beta$, the resampled draws give posterior mean 18.9 with 95% interval [12.9, 25.6] (quadrature: 18.8, [12.9, 25.4]), against [12.4, 24.7] from the raw normal approximation; drawing $\tilde{y}_{1986} \sim \text{Poisson}(\alpha + 1986\beta)$ for each resampled draw gives the 95% posterior predictive interval [10, 31] for the number of fatal accidents in 1986.
(c) Importance ratios. The ratios pile up tightly: $\text{sd}(\log w) = 0.18$, with the central 98% of $\log w$ spanning only 1.2, so a histogram of w is a sharp spike at the bulk value with thin tails both ways – a short right tail of draws where the Poisson likelihood exceeds its Gaussian approximation, and a longer left tail from draws far out in the (α^*, β) tails where it falls off faster. Over six independent batches of $S = 10{,}000$, $\max_s w^s/\bar{w}$ ran between 3.5 and 6 (it grows slowly with S, as an extreme-order statistic must), so no single draw dominates. By (10.4), with $\tilde{w}^s = w^s / \sum_r w^r$,

$$S_{\text{eff}} = \frac{1}{\sum_{s=1}^{S} (\tilde{w}^s)^2} \approx 9700 \quad (S = 10{,}000),$$

i.e. about 97% efficiency, varying by only ± 20 across batches; for a single batch of $S = 1000$ draws the same formula gives $S_{\text{eff}} \approx 970$.

Problem 13.3. Mode-based approximation: Consider the model, $y_j \sim \text{Binomial}(n_j, \theta_j)$, where $\theta_j = \text{logit}^{-1}(\alpha + \beta x_j)$, for $j = 1, \ldots, J$, and with independent prior distributions, $\alpha \sim t_4(0, 2^2)$ and $\beta \sim t_4(0, 1)$. Suppose $J = 10$, the x_j values are randomly drawn from a $\text{U}(0, 1)$ distribution, and $n_j \sim \text{Poisson}^+(5)$, where Poisson^+ is the Poisson distribution restricted to positive values.
(a) Sample a dataset at random from the model.
(b) Use rejection sampling to get 1000 independent posterior draws from (α, β).
(c) Approximate the posterior density for (α, β) by a normal centered at the posterior mode with covariance matrix fit to the curvature at the mode.
(d) Take 1000 draws from the two-dimensional t_4 distribution with that center and scale matrix and use importance sampling to estimate $\text{E}(\alpha|y)$ and $\text{E}(\beta|y)$.

Solution. $\text{E}(\alpha|y) = -0.85$ and $\text{E}(\beta|y) = -0.42$; the t_4 proposal centered at the mode with the curvature scale matrix accepts 59% of its draws in rejection sampling and has importance-sampling efficiency $S_{\text{eff}}/S = 0.91$.
(a) Drawing $x_j \sim \text{U}(0, 1)$, $n_j \sim \text{Poisson}^+(5)$, then (α, β) from their t_4 priors and $y_j \sim \text{Binomial}(n_j, \text{logit}^{-1}(\alpha + \beta x_j))$ gave true values $(\alpha, \beta) = (-0.179, -1.075)$ and

j	1	2	3	4	5	6	7	8	9	10
x_j	0.778	0.238	0.824	0.966	0.973	0.453	0.609	0.776	0.642	0.722
n_j	2	5	5	3	7	6	5	3	4	5
y_j	0	1	2	0	1	2	2	2	1	0

with $\sum_j n_j = 45$ trials in all – a very small dataset, so the posterior is appreciably non-normal.
(b) Rejection sampling. Both the proposal used here and the approximation asked for in (c) are built from the posterior mode and the curvature there. With $\eta_j = \alpha + \beta x_j$ and $z_j = (1, x_j)'$, the unnormalized log posterior and its derivatives are

$$\begin{aligned}
\log q(\alpha, \beta|y) &= \sum_j \left[y_j \eta_j - n_j \log(1 + e^{\eta_j}) \right] \\
&\quad - \tfrac{5}{2} \log\Big(1 + \tfrac{\alpha^2}{16}\Big) - \tfrac{5}{2} \log\Big(1 + \tfrac{\beta^2}{4}\Big), \\
\nabla \log q &= \sum_j (y_j - n_j \theta_j) z_j + \nabla \log p(\alpha) p(\beta), \\
\nabla^2 \log q &= -\sum_j n_j \theta_j (1 - \theta_j) z_j z_j' + \text{diag}(c_\alpha, c_\beta),
\end{aligned}$$

where for a $t_\nu(0, s^2)$ prior $\frac{d^2}{dv^2} \log p(v) = -(\nu + 1)(\nu s^2 - v^2)/(\nu s^2 + v^2)^2$, giving c_α at $s^2 = 4$ and c_β at

$s^2 = 1$, both with $\nu = 4$. Newton's method gives

$$(\hat{\alpha}, \hat{\beta}) = (-0.875,\ -0.327), \qquad V_\gamma = \begin{pmatrix} 0.3654 & -0.3820 \\ -0.3820 & 0.5823 \end{pmatrix},$$

with marginal standard deviations 0.604 and 0.763 and correlation -0.828, the usual intercept-slope trade-off for uncentered x.
Use as proposal the bivariate t_4 with center $\hat{\gamma}$ and scale matrix V_γ,

$$g(\gamma) = \left(1 + \tfrac{1}{4}(\gamma - \hat{\gamma})' V_\gamma^{-1} (\gamma - \hat{\gamma})\right)^{-3},$$

whose tails dominate those of q: the data are not separable (group $j = 2$ alone has $0 < y_j < n_j$, so no line classifies all 45 Bernoulli trials correctly), hence the likelihood decays exponentially in every direction while g decays only polynomially, and $q/g \to 0$ with the supremum attained. Maximizing $\log q - \log g$ numerically over the plane – the maximum is attained at $(0.248, -2.239)$, not at the mode, because the posterior is skewed relative to the proposal – gives the bound

$$\sup_\gamma \frac{q(\gamma|y)}{g(\gamma)} = M, \qquad \log M = -24.513,$$

with q and g exactly the two unnormalized kernels displayed above (the acceptance rate below is invariant to how each is scaled).
Accepting each draw $\gamma^s \sim g$ with probability $q(\gamma^s|y)/(Mg(\gamma^s))$ (Section 10.3) then yields exact posterior draws, with observed acceptance rate 0.591. From 1000 accepted draws,

$$\begin{aligned} \mathrm{E}(\alpha|y) &= -0.848, \quad \mathrm{sd}(\alpha|y) = 0.679, \\ \mathrm{E}(\beta|y) &= -0.419, \quad \mathrm{sd}(\beta|y) = 0.891, \end{aligned}$$

with Monte Carlo standard errors 0.021 and 0.028, and 95% intervals $[-2.19, 0.49]$ for α and $[-2.30, 1.28]$ for β. Grid quadrature of q confirms the exact values -0.853, -0.420, 0.669, 0.877 and $[-2.17, 0.48]$, $[-2.25, 1.24]$.
(c) Mode-based normal approximation. From the mode and curvature computed above,

$$p_{\text{approx}}(\alpha, \beta|y) = \mathrm{N}\left(\begin{pmatrix} -0.875 \\ -0.327 \end{pmatrix},\ V_\gamma\right).$$

Against the exact draws of (b), it locates the distribution well but is too tight and too symmetric: the exact posterior mean lies above the mode in α (-0.853 vs -0.875) and below it in β (-0.420 vs -0.327), and the exact standard deviations 0.669 and 0.877 exceed the curvature-based 0.604 and 0.763 by about 11% and 15%.
(d) Importance sampling. With the same t_4 proposal and $S = 1000$ draws, the ratios $w^s = q(\gamma^s|y)/g(\gamma^s)$ give

$$\mathrm{E}(\alpha|y) \approx \frac{\sum_s w^s \alpha^s}{\sum_s w^s} = -0.806, \qquad \mathrm{E}(\beta|y) \approx -0.472,$$

with weighted standard deviations 0.674 and 0.884, and effective sample size $S_{\text{eff}} = 908$ from (10.4) – a well-behaved weight distribution, as the acceptance rate in (b) already indicated. The Monte Carlo error at $S = 1000$ is about $0.67/\sqrt{908} = 0.022$ for α and 0.029 for β, and the strong negative posterior correlation makes opposite-signed errors of that size in the two coordinates likely; the quadrature values are -0.853 and -0.420.

Problem 13.4. Analytic approximation to a subset of the parameters: suppose that the joint posterior distribution $p(\theta_1, \theta_2|y)$ is of interest and that it is known that the t provides an adequate approximation to the conditional distribution, $p(\theta_1|\theta_2, y)$. Show that both the normal and t approaches described in the last paragraph of Section 13.5 lead to the same answer.

For reference, that construction approximates the marginal density of θ_2 by the ratio (13.9),

$$p_{\text{approx}}(\theta_2|y) = \frac{p(\theta_1,\theta_2|y)}{p_{\text{approx}}(\theta_1|\theta_2,y)},$$

with θ_1 set to the center $\hat\theta_1(\theta_2)$ of the approximating conditional distribution, whose scale matrix is $V_{\theta_1}(\theta_2)$; for the d-dimensional normal approximation this yields (13.10),

$$p_{\text{approx}}(\theta_2|y) \propto p\big(\hat\theta_1(\theta_2),\theta_2|y\big)\,\big|V_{\theta_1}(\theta_2)\big|^{1/2}.$$

Solution. Both give $p_{\text{approx}}(\theta_2|y) \propto p(\hat\theta_1(\theta_2),\theta_2|y)\,|V_{\theta_1}(\theta_2)|^{1/2}$, because a normal density and a t_ν density both equal a constant times $|V|^{-1/2}$ **at their own center**, and in (13.9) that is the only place they are ever evaluated.
Explicitly, in d dimensions with center $\hat\theta_1 = \hat\theta_1(\theta_2)$ and scale matrix $V_{\theta_1} = V_{\theta_1}(\theta_2)$,

$$\begin{aligned}
\text{N}(\theta_1|\hat\theta_1, V_{\theta_1})\Big|_{\theta_1=\hat\theta_1} &= (2\pi)^{-d/2}|V_{\theta_1}|^{-1/2},\\
t_\nu(\theta_1|\hat\theta_1, V_{\theta_1})\Big|_{\theta_1=\hat\theta_1} &= \frac{\Gamma\big(\frac{\nu+d}{2}\big)}{\Gamma\big(\frac{\nu}{2}\big)(\nu\pi)^{d/2}}|V_{\theta_1}|^{-1/2},
\end{aligned}$$

the t because its quadratic form $(\theta_1-\hat\theta_1)'V_{\theta_1}^{-1}(\theta_1-\hat\theta_1)$ vanishes at $\theta_1=\hat\theta_1$, killing the factor $\big(1+\frac{1}{\nu}(\cdot)\big)^{-(\nu+d)/2}$. Substituting each into the denominator of (13.9) at $\theta_1=\hat\theta_1(\theta_2)$,

$$\begin{aligned}
p^{\text{N}}_{\text{approx}}(\theta_2|y) &= (2\pi)^{d/2}\,p\big(\hat\theta_1(\theta_2),\theta_2|y\big)\,|V_{\theta_1}(\theta_2)|^{1/2},\\
p^{t}_{\text{approx}}(\theta_2|y) &= \frac{\Gamma(\frac{\nu}{2})(\nu\pi)^{d/2}}{\Gamma(\frac{\nu+d}{2})}\,p\big(\hat\theta_1(\theta_2),\theta_2|y\big)\,|V_{\theta_1}(\theta_2)|^{1/2}.
\end{aligned}$$

The two differ only by the constant factor $(2\pi)^{d/2}\Gamma(\frac{\nu+d}{2})/\big(\Gamma(\frac{\nu}{2})(\nu\pi)^{d/2}\big)$, which does not involve θ_2 and so is absorbed when $p_{\text{approx}}(\theta_2|y)$ is normalized over θ_2. Hence the two approaches give identical approximate marginal posterior densities for θ_2, which is (13.10). The one hypothesis being used: ν and d are held fixed as θ_2 varies – only the center and scale matrix may depend on θ_2, as the book's parenthetical notation $\hat\theta_1(\theta_2)$, $V_{\theta_1}(\theta_2)$ records.

Problem 13.5. Estimating the number of unseen species (see Fisher, Corbet, and Williams, 1943, Efron and Thisted, 1976, and Seber, 1992): suppose that during an animal trapping expedition the number of times an animal from species i is caught is $x_i \sim \text{Poisson}(\lambda_i)$. For parts (a)–(d) of this problem, assume a Gamma(α,β) prior distribution for the λ_i's, with a uniform hyperprior distribution on (α,β). The only observed data are y_k, the number of species observed exactly k times during a trapping expedition, for $k = 1,2,3,\ldots$
(a) Write the distribution $p(x_i|\alpha,\beta)$.
(b) Use the distribution of x_i to derive a multinomial distribution for y given that there are a total of N species.
(c) Suppose that we are given

$$y = (118, 74, 44, 24, 29, 22, 20, 14, 20, 15, 12, 14, 6, 12, 6, 9, 9, 6, 10, 10, 11, 5, 3, 3),$$

so that 118 species were observed only once, 74 species were observed twice, and so forth, with a total of 496 species observed and 3266 animals caught. Write down the likelihood for y using the multinomial distribution with 24 cells (ignoring unseen species). Use any method to find the mode of α,β and an approximate second derivative matrix.
(d) Derive an estimate and approximate 95% posterior interval for the number of additional species that would be observed if 10,000 more animals were caught.

(e) Evaluate the fit of the model to the data using appropriate posterior predictive checks.
(f) Discuss the sensitivity of the inference in (d) to each of the model assumptions.

Solution. About 111 additional species, with 95% interval [85, 136] – but part (e) shows the gamma mixture badly misfits the upper tail, so that interval should not be taken at face value.
(a) Negative binomial: with the Gamma(α,β) density in the rate parameterization $p(\lambda) = \beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}/\Gamma(\alpha)$,

$$\begin{aligned} p(x|\alpha,\beta) &= \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!}\,\frac{\beta^\alpha\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)}\,d\lambda \\ &= \frac{\Gamma(x+\alpha)}{\Gamma(\alpha)\,x!}\Big(\frac{\beta}{\beta+1}\Big)^\alpha\Big(\frac{1}{\beta+1}\Big)^x, \quad x = 0,1,2,\dots \end{aligned}$$

Write $p_k = p(x=k|\alpha,\beta)$; in particular $p_0 = (\beta/(\beta+1))^\alpha$, and more generally $\mathrm{E}(e^{-s\lambda}) = (\beta/(\beta+s))^\alpha$, which is all that part (d) needs.
(b) The N species have independent λ_i and hence, marginally, independent and identically distributed counts x_i with the above distribution. Classifying the N species by their count, $y_k = \#\{i : x_i = k\}$, is exactly a multinomial allocation:

$$(y_0, y_1, y_2, \dots) \mid N,\alpha,\beta \;\sim\; \mathrm{Multin}\big(N;\,(p_0,p_1,p_2,\dots)\big),$$

with $y_0 = N - \sum_{k\ge1} y_k$ the (unobserved) number of species never caught.
(c) Only species with $x_i \ge 1$ are observed, so conditioning on observation replaces p_k by $p_k/(1-p_0)$ and the observed total $\sum_{k\ge1} y_k = 496$ by a fixed sample size; the cells $k > 24$ are empty and contribute a factor of 1. Hence the 24-cell multinomial likelihood, ignoring unseen species,

$$p(y|\alpha,\beta) \;=\; \frac{496!}{\prod_{k=1}^{24} y_k!}\prod_{k=1}^{24}\left[\frac{1}{1-p_0}\,\frac{\Gamma(k+\alpha)}{\Gamma(\alpha)k!}\Big(\tfrac{\beta}{\beta+1}\Big)^\alpha\Big(\tfrac{1}{\beta+1}\Big)^k\right]^{y_k},$$

and with the uniform hyperprior $p(\alpha,\beta|y) \propto p(y|\alpha,\beta)$. Maximizing numerically,

$$(\hat\alpha,\hat\beta) = (0.4695,\ 0.1070),$$

at which $p_0 = 0.334$. The second derivative matrix of the log posterior at the mode and its negative inverse are

$$\frac{d^2\log p}{d(\alpha,\beta)^2} = \begin{pmatrix} -493.0 & 2828.4 \\ 2828.4 & -21699.7 \end{pmatrix},$$

$$V_{(\alpha,\beta)} = \begin{pmatrix} 0.008043 & 0.0010484 \\ 0.0010484 & 0.00018273 \end{pmatrix},$$

i.e. standard deviations 0.090 and 0.0135 with correlation 0.865. Because both parameters are positive and strongly correlated, the approximation is much better on the log scale: with the Jacobian, $p(\log\alpha,\log\beta|y) \propto \alpha\beta\,p(y|\alpha,\beta)$, whose mode is at $(\log\hat\alpha,\log\hat\beta) = (-0.701,-2.199)$, that is $(\alpha,\beta) = (0.496,\ 0.111)$, with

$$V_{(\log\alpha,\log\beta)} = \begin{pmatrix} 0.03317 & 0.01943 \\ 0.01943 & 0.01525 \end{pmatrix}$$

(standard deviations 0.182 and 0.124, correlation 0.864). All simulations below use this log-scale normal approximation.
(d) Catching 10,000 more animals is $t = 10000/3266 = 3.062$ times the original effort, so species i is caught Poisson$(t\lambda_i)$ times in the new expedition, independently of the first. A species is **new** if it was missed before and caught now, which marginally has probability

$$\mathrm{E}\big[e^{-\lambda}(1-e^{-t\lambda})\big] = \Big(\tfrac{\beta}{\beta+1}\Big)^\alpha - \Big(\tfrac{\beta}{\beta+1+t}\Big)^\alpha =: p_0 - q.$$

Since 496 species were seen out of N, $N = 496/(1-p_0)$, and the expected number of new species is

$$\phi(\alpha,\beta) = 496\,\frac{p_0 - q}{1-p_0} = 496\,\frac{(\frac{\beta}{\beta+1})^\alpha - (\frac{\beta}{\beta+1+t})^\alpha}{1-(\frac{\beta}{\beta+1})^\alpha}.$$

At the mode $\phi = 115.2$. Propagating the posterior uncertainty through 20,000 draws from the log-scale normal approximation gives posterior mean 111, median 112, and central 95% interval [85, 136] additional species; with $\mathrm{sd}(\phi|y) = 13$ the Monte Carlo error in the mean is 0.1, so the reported digits are all real. Adding the binomial variability of which unseen species happen to be caught (given (α,β), the count is Binomial$(N-496,\, 1-((\beta+1)/(\beta+1+t))^\alpha)$) widens this only to [82, 141]: parameter uncertainty dominates. The same draws give a total species richness N with posterior mean 733 and 95% interval [637, 858].

(e) Replicating $y^{\mathrm{rep}} \sim \mathrm{Multin}(496; \{p_k/(1-p_0)\}_{k\geq 1})$ at each of 4000 posterior draws – the untruncated conditional distribution, so the cells $k > 24$ that were ignored in (c) are now put at risk – and comparing with the data (p-values are $\Pr(T(y^{\mathrm{rep}}) \geq T(y))$, Monte Carlo error 0.008):

Test quantity	Observed	Replicated mean	95% interval	p-value
y_1 (species seen once)	118	103.0	[80, 126]	0.11
Total animals $\sum_k k\,y_k$	3266	3269	[2836, 3748]	0.49
Largest count observed	24	53.9	[36, 83]	1.00
Species caught > 24 times	0	16.3	[7, 27]	1.00

The first two check out; the last two do not, and a χ^2 discrepancy on the bins $1, 2, 3, 4, 5, 6\text{–}7, 8\text{–}10, 11\text{–}15, 16\text{–}24, >24$ has posterior predictive p-value 0.0000 ($T(y,\alpha,\beta) = 63.2$ on average against $T(y^{\mathrm{rep}},\alpha,\beta) = 9.1$). The misfit is entirely in the upper tail: the model expects 33.6 species with 16–24 catches and 16.7 with more than 24, whereas the data have 66 and 0. A gamma mixture cannot simultaneously produce that many species in the teens and a hard stop at 24; the data look as if the reported table was itself truncated at 24 catches. The fit for $k \leq 3$ is also imperfect ($118, 74, 44$ observed against $103, 70, 52$ expected).

(f) Sensitivity of the estimate in (d).

(i) **The gamma mixing distribution.** Critical, because ϕ enters only through $\mathrm{E}[e^{-\lambda}]$ and $\mathrm{E}[e^{-(1+t)\lambda}]$ – expectations dominated by the small-λ region, about which the data say almost nothing beyond y_1 and y_2. Given the tail misfit in (e), a mixing family with a lighter right tail and more mass near zero would fit better and give a different ϕ; the interval $[85, 136]$ reflects uncertainty only **within** the gamma family. The classical log-series analysis of such data is the $\alpha \to 0$ limit of this family, which the present fit rejects outright ($\hat\alpha = 0.47$, standard deviation 0.09).

(ii) **Extrapolation length.** For small α, $\phi \approx 496 \log\frac{\beta+1+t}{\beta+1} / \log\frac{\beta+1}{\beta}$, logarithmic in the effort, and the dependence on the mixing family grows with t. At $t \ll 1$ the answer is nearly model-free ($\phi \approx t\,y_1$, since $p_0 - q \approx t\,\mathrm{E}[\lambda e^{-\lambda}] = t\,p_1$); at $t = 3.06$ it is not.

(iii) **Poisson catches and independence across species.** Traps saturate, animals are social, and catches cluster in time and space. Overdispersion within a species, or positive dependence between species, inflates y_1 relative to what the model attributes to genuinely rare species, and so inflates ϕ.

(iv) **Constant rates over time.** λ_i is the same in both expeditions. Seasonal or spatial change, or depletion caused by the first expedition, invalidates the calculation, and the model cannot signal it.

(v) **The uniform hyperprior and the normal approximation.** Least important: refitting on the $(\log\alpha, \log\beta)$ scale – a different uniform prior – moves the mode from $(0.470, 0.107)$ to $(0.496, 0.111)$ and ϕ by about three species, a twentieth of the width of the interval in (d).

Problem 13.6. Derivation of the monotone convergence of EM algorithm: prove that the function $\mathrm{E}_{\mathrm{old}} \log p(\gamma|\phi, y)$ in (13.5) is maximized at $\phi = \phi^{\mathrm{old}}$. (Hint: express the expectation as an integral and apply Jensen's inequality to the convex logarithm function.)

Here $\theta = (\gamma, \phi)$, $\mathrm{E}_{\mathrm{old}}$ denotes averaging over γ under $p(\gamma|\phi^{\mathrm{old}}, y)$, and (13.5) is the identity

$$\log p(\phi|y) = \mathrm{E}_{\mathrm{old}}\big(\log p(\gamma,\phi|y)\big) - \mathrm{E}_{\mathrm{old}}\big(\log p(\gamma|\phi,y)\big).$$

Solution. The difference is minus a Kullback-Leibler divergence, hence at most zero:

$$\begin{aligned}
&\mathrm{E}_{\mathrm{old}} \log p(\gamma|\phi, y) - \mathrm{E}_{\mathrm{old}} \log p(\gamma|\phi^{\mathrm{old}}, y) \\
&\quad = \int p(\gamma|\phi^{\mathrm{old}}, y)\, \log \frac{p(\gamma|\phi, y)}{p(\gamma|\phi^{\mathrm{old}}, y)}\, d\gamma \\
&\quad \le \log \int p(\gamma|\phi^{\mathrm{old}}, y)\, \frac{p(\gamma|\phi, y)}{p(\gamma|\phi^{\mathrm{old}}, y)}\, d\gamma \\
&\quad = \log \int p(\gamma|\phi, y)\, d\gamma \;=\; \log 1 \;=\; 0,
\end{aligned}$$

for every ϕ, so $\phi \mapsto \mathrm{E}_{\mathrm{old}} \log p(\gamma|\phi, y)$ attains its maximum at $\phi = \phi^{\mathrm{old}}$. The inequality is Jensen's applied to the **concave** function log – the printed hint calls it convex, an erratum – and the random variable $R = p(\gamma|\phi, y)/p(\gamma|\phi^{\mathrm{old}}, y)$ under $p(\gamma|\phi^{\mathrm{old}}, y)$, whose hypothesis is only that ER exist, which the third line verifies. That line needs $p(\gamma|\phi, y)$ absolutely continuous with respect to $p(\gamma|\phi^{\mathrm{old}}, y)$; without it the integral is at most 1 and the conclusion is unchanged.

Problem 13.7. Conditional maximization for the hierarchical normal model: show that the conditional modes of σ and τ associated with (11.14) and (11.16), respectively, are correct.
In the hierarchical normal model of Section 11.6, $y_{ij} \sim \mathrm{N}(\theta_j, \sigma^2)$ for $i = 1, \dots, n_j$ and $j = 1, \dots, J$, with $\theta_j \sim \mathrm{N}(\mu, \tau^2)$, $n = \sum_j n_j$, and $p(\mu, \log\sigma, \log\tau) \propto \tau$. The relevant conditional posterior distributions are

$$\sigma^2|\theta, \mu, \tau, y \sim \text{Inv-}\chi^2(n, \hat\sigma^2), \qquad \hat\sigma^2 = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \theta_j)^2,$$

$$\tau^2|\theta, \mu, \sigma, y \sim \text{Inv-}\chi^2(J-1, \hat\tau^2), \quad \hat\tau^2 = \frac{1}{J-1}\sum_{j=1}^{J}(\theta_j - \mu)^2,$$

which are (11.14)–(11.15) and (11.16)–(11.17). Section 13.6 performs conditional maximization on the $(\theta, \mu, \log\sigma, \log\tau)$ scale and asserts that the conditional modes are $\log\hat\sigma$ and $\log\hat\tau$, with no factor such as the $\frac{n}{n+2}$ that appears in the conditional mode of σ^2 itself.

Solution. On the log scale the Inv-$\chi^2(\nu, s^2)$ density has mode exactly at s, and both claims are the cases $(\nu, s^2) = (n, \hat\sigma^2)$ and $(J-1, \hat\tau^2)$.
Let $\sigma^2 \sim \text{Inv-}\chi^2(\nu, s^2)$, so $p(\sigma^2) \propto (\sigma^2)^{-(\nu/2+1)} e^{-\nu s^2/(2\sigma^2)}$ (Appendix A). Transforming by $\sigma^2 = e^{2\log\sigma}$, the Jacobian is $d\sigma^2/d\log\sigma = 2\sigma^2$, so

$$\begin{aligned}
p(\log\sigma) &\propto \sigma^2\,(\sigma^2)^{-(\nu/2+1)} e^{-\nu s^2/(2\sigma^2)} = \sigma^{-\nu} e^{-\nu s^2/(2\sigma^2)}, \\
\frac{d}{d\log\sigma}\log p(\log\sigma) &= -\nu + \frac{\nu s^2}{\sigma^2},
\end{aligned}$$

using $d(\sigma^{-2})/d\log\sigma = -2\sigma^{-2}$. This vanishes exactly at $\sigma^2 = s^2$, and

$$\frac{d^2}{d(\log\sigma)^2}\log p(\log\sigma) = -\frac{2\nu s^2}{\sigma^2} = -2\nu < 0 \quad \text{there},$$

so $\log\sigma = \log s$ is the unique mode. (For contrast, on the σ^2 scale $\frac{d}{d\sigma^2}\log p = -(\nu/2+1)/\sigma^2 + \nu s^2/(2\sigma^4) = 0$ gives $\sigma^2 = \frac{\nu}{\nu+2}s^2$; the Jacobian factor $2\sigma^2$ is exactly what cancels the extra -1 in the exponent and removes the $\frac{\nu}{\nu+2}$.)
Applying this to (11.14) with $\nu = n$, $s^2 = \hat\sigma^2$: the conditional mode of $\log\sigma$ given (θ, μ, τ, y) is $\log\hat\sigma$ with $\hat\sigma^2 = \frac{1}{n}\sum_j\sum_i(y_{ij} - \theta_j)^2$, as in (11.15). Applying it to (11.16) with $\nu = J-1$, $s^2 = \hat\tau^2$: the conditional mode of $\log\tau$ given (θ, μ, σ, y) is $\log\hat\tau$ with $\hat\tau^2 = \frac{1}{J-1}\sum_j(\theta_j - \mu)^2$, as in (11.17). The degrees of freedom differ ($J-1$ rather than J) because the prior $p(\mu, \log\sigma, \log\tau) \propto \tau$ contributes one extra factor of τ, and the argument above is insensitive to ν – the mode is at s for every $\nu > 0$.

13.2 Exercises 13.8–13.11

Problem 13.8. Joint posterior modes for hierarchical models:
(a) Show that the posterior density for the coagulation example from Table 11.2 on page 288 has a degenerate mode at $\tau = 0$ and $\theta_j = \mu$ for all j.
(b) The rest of this exercise demonstrates that the degenerate mode represents a small part of the posterior distribution. First estimate an upper bound on the integral of the unnormalized posterior density in the neighborhood of the degenerate mode. (Approximate the integrand so that the integral is analytically tractable.)
(c) Now approximate the integral of the unnormalized posterior density in the neighborhood of the other mode using the density at the mode and the second derivative matrix of the log posterior density at the mode.
(d) Finally, estimate an upper bound on the posterior mass in the neighborhood of the degenerate mode.
The data of Table 11.2 (coagulation time in seconds for blood drawn from 24 animals randomly allocated to four different diets) are:

Diet	Measurements
A	62, 60, 63, 59
B	63, 67, 71, 64, 65, 66
C	68, 66, 71, 67, 68, 68
D	56, 62, 60, 61, 63, 64, 63, 59

The model of Section 11.6 is $y_{ij} \mid \theta_j, \sigma \sim \mathrm{N}(\theta_j, \sigma^2)$ for $i = 1, \dots, n_j$, $j = 1, \dots, J$, with $\theta_j \mid \mu, \tau \sim \mathrm{N}(\mu, \tau^2)$ and a uniform prior density on $(\mu, \log\sigma, \tau)$; equivalently $p(\mu, \log\sigma, \log\tau) \propto \tau$. Thus $J = 4$, $(n_1, n_2, n_3, n_4) = (4, 6, 6, 8)$, $n = 24$. Table 13.1 on page 327 reports the joint posterior mode as $\theta = (61.29, 65.87, 67.73, 61.15)$, $\mu = 64.01$, $\sigma = 2.17$, $\tau = 3.31$, with $\log p(\theta, \mu, \log\sigma, \log\tau \mid y) = -61.42$.

Solution. (a) Along the ray $\theta_1 = \cdots = \theta_J = \mu$ the joint density is proportional to $\tau^{1-J} = \tau^{-3}$, which diverges as $\tau \to 0$. Explicitly, from the joint density of Section 11.6,

$$p(\theta, \mu, \log\sigma, \log\tau \mid y) \propto \tau \prod_{j=1}^{J} \mathrm{N}(\theta_j \mid \mu, \tau^2) \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathrm{N}(y_{ij} \mid \theta_j, \sigma^2),$$

$$\text{and at } \theta_1 = \cdots = \theta_J = \mu, \quad \propto (2\pi)^{-J/2}\, \tau^{1-J} \prod_{j,i} \mathrm{N}(y_{ij} \mid \mu, \sigma^2),$$

since $\prod_j \mathrm{N}(\mu \mid \mu, \tau^2) = (2\pi\tau^2)^{-J/2}$ and the leading τ is the Jacobian from $p(\mu, \log\sigma, \tau) \propto 1$. The remaining factor is free of τ and maximized at $\mu = \bar{y}_{\cdot\cdot} = 64$, $\sigma^2 = \frac{1}{n}\sum_{ij}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = 340/24$, so the density increases without bound as $\tau \downarrow 0$ along that ray; the limit point sits on the boundary of the parameter space, so the mode is degenerate rather than stationary.
(b) $\displaystyle\int_{\{\tau \le \epsilon\}} q \;=\; \frac{\epsilon}{2\sqrt{n}}\, \Gamma\Big(\frac{n-1}{2}\Big)(\pi T)^{-(n-1)/2} \;=\; 1.80 \times 10^{-29}\, \epsilon$, where q denotes the unnormalized density of (a) with all normal normalizing constants retained and $T = \sum_{ij}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = 340$.
Integrate θ out exactly first. Writing $S = \sum_{ij}(y_{ij} - \bar{y}_{\cdot j})^2 = 112$ and completing the square in θ_j,

$$\prod_i \mathrm{N}(y_{ij} \mid \theta_j, \sigma^2) = (2\pi\sigma^2)^{-n_j/2} e^{-\frac{1}{2\sigma^2}\sum_i (y_{ij} - \bar{y}_{\cdot j})^2}$$

$$\times \Big(\tfrac{2\pi\sigma^2}{n_j}\Big)^{1/2} \mathrm{N}\big(\theta_j \mid \bar{y}_{\cdot j}, \tfrac{\sigma^2}{n_j}\big),$$

so that $\int \mathrm{N}(\theta_j \mid \mu, \tau^2)\, \mathrm{N}(\theta_j \mid \bar{y}_{\cdot j}, \sigma^2/n_j)\, d\theta_j = \mathrm{N}(\bar{y}_{\cdot j} \mid \mu, \tau^2 + \sigma^2/n_j)$ and

$$\int q\, d\theta = \tau\, (2\pi)^{-\frac{n-J}{2}} \sigma^{-(n-J)} \Big(\prod_j n_j\Big)^{-1/2} e^{-S/(2\sigma^2)}$$

$$\times \prod_{j=1}^{J} \mathrm{N}\big(\bar{y}_{\cdot j} \,\big|\, \mu, \tau^2 + \tfrac{\sigma^2}{n_j}\big).$$

This is exact. The approximation is now to take $\tau \to 0$ in the smooth factors, i.e. replace $\tau^2 + \sigma^2/n_j$ by σ^2/n_j, which collapses the display to

$$\int q\, d\theta \;\approx\; \tau\,(2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{T + n(\bar{y}_{..} - \mu)^2}{2\sigma^2}\Big),$$

because $S + \sum_j n_j(\bar{y}_{.j} - \mu)^2 = \sum_{ij}(y_{ij} - \mu)^2 = T + n(\bar{y}_{..} - \mu)^2$. Now $\int d\mu$ gives $(2\pi\sigma^2/n)^{1/2}$, and with $u = T/(2\sigma^2)$,

$$\int_0^\infty \sigma^{-(n-1)} e^{-T/(2\sigma^2)}\, \frac{d\sigma}{\sigma} = \tfrac{1}{2}\Big(\tfrac{2}{T}\Big)^{(n-1)/2} \Gamma\Big(\tfrac{n-1}{2}\Big),$$

while $\int_{-\infty}^{\log\epsilon} \tau\, d\log\tau = \epsilon$. Multiplying the three factors and absorbing $(2\pi)^{(1-n)/2}(2/T)^{(n-1)/2} = (\pi T)^{-(n-1)/2}$ gives the stated closed form; with $n = 24$, $T = 340$ the coefficient is 1.80×10^{-29}.
The collapse $\tau^2 + \sigma^2/n_j \mapsto \sigma^2/n_j$ is exact only in the limit, so the result is an estimate rather than a strict bound: quadrature of the exact display gives 4.80×10^{-30} against 4.50×10^{-30} at $\epsilon = 0.25$, and 7.56×10^{-29} against 1.80×10^{-29} at $\epsilon = 1$, where ϵ is no longer small beside $\sigma/\sqrt{n_j} \approx 1.1$.

(c) $\displaystyle\int_{\text{near joint mode}} q \;\approx\; q(\hat\theta, \hat\mu, \log\hat\sigma, \log\hat\tau)\,(2\pi)^{d/2}\,|\hat\Sigma|^{1/2} = 8.19 \times 10^{-26}$, with $d = J + 3 = 7$.
Here $\hat\Sigma = [-\nabla^2 \log q]^{-1}$ at the mode. Numerical maximization of $\log q$ reproduces Table 13.1 to the digits shown, $\hat\theta = (61.292, 65.867, 67.733, 61.154)$, $\hat\mu = 64.012$, $\hat\sigma = 2.170$, $\hat\tau = 3.309$, with $\log q = -61.4193$ under the constant $-\frac{n+J}{2}\log(2\pi)$ (this is the normalization used in Table 13.1). The second derivative matrix in the coordinates $(\theta_1, \ldots, \theta_4, \mu, \log\sigma, \log\tau)$ has $\log\det(-\nabla^2 \log q) = 5.555$, so $|\hat\Sigma|^{1/2} = e^{-2.777} = 0.0623$ and

$$\log \int q \approx -61.4193 + \tfrac{7}{2}\log(2\pi) - \tfrac{1}{2}(5.555) = -57.764.$$

(d) At most a few parts in 10^4. Taking the neighborhood of the degenerate mode to be $\tau < 1$ second (against $\hat\tau = 3.3$, $\hat\sigma = 2.2$), the ratio of (b) to (c) is

$$\Pr(\tau < 1 \mid y) \;\lesssim\; \frac{1.80 \times 10^{-29}}{8.19 \times 10^{-26}} = 2.2 \times 10^{-4};$$

two-dimensional quadrature of the marginal density of (b) over the whole space gives $\int q = 1.58 \times 10^{-25}$ and hence the exact $7.56 \times 10^{-29}/1.58 \times 10^{-25} = 4.8 \times 10^{-4}$, or 3.0×10^{-5} for $\tau < 0.25$. Either way the degenerate mode carries under one tenth of one percent of the posterior distribution.

Problem 13.9. EM algorithm:
(a) For the hierarchical normal model in Section 13.6, derive the expressions (13.14) for μ^{new}, σ^{new}, and τ^{new}.
(b) Pick values for the hyperparameters in this model, then simulate fake data, then apply EM to estimate the model. Compare the EM estimate to the assumed true model.
For reference, the model of Section 13.6 has $y_{ij} \mid \theta_j, \sigma \sim \text{N}(\theta_j, \sigma^2)$, $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, $\theta_j \mid \mu, \tau \sim \text{N}(\mu, \tau^2)$, and $p(\mu, \log\sigma, \log\tau) \propto \tau$, so that by (13.13)

$$\begin{aligned}\log p(\theta, \mu, \log\sigma, \log\tau \mid y) = &-n\log\sigma - (J-1)\log\tau - \frac{1}{2\tau^2}\sum_{j=1}^{J}(\theta_j - \mu)^2 \\ &- \frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \theta_j)^2 + \text{constant},\end{aligned}$$

and the target of (13.14) is

$$\mu^{\text{new}} = \frac{1}{J}\sum_{j=1}^{J}\hat\theta_j,$$
$$\sigma^{\text{new}} = \Big[\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\big((y_{ij}-\hat\theta_j)^2 + V_{\theta j}\big)\Big]^{1/2},$$
$$\tau^{\text{new}} = \Big[\frac{1}{J-1}\sum_{j=1}^{J}\big((\hat\theta_j-\mu^{\text{new}})^2 + V_{\theta j}\big)\Big]^{1/2},$$

where $\hat\theta_j$ and $V_{\theta j}$ are given by (11.10) and (11.11) evaluated at $(\mu, \log\sigma, \log\tau)^{\text{old}}$.

Solution. (a) Set the three partial derivatives of the E-step objective to zero. The E-step of Section 13.6 gives $E_{\text{old}}((\theta_j-\mu)^2) = (\hat\theta_j-\mu)^2 + V_{\theta j}$ and $E_{\text{old}}((y_{ij}-\theta_j)^2) = (y_{ij}-\hat\theta_j)^2 + V_{\theta j}$, so, writing $\lambda = \log\sigma$ and $\gamma = \log\tau$, the function maximized in the M-step is

$$Q(\mu,\lambda,\gamma) = -n\lambda - (J-1)\gamma - \frac{e^{-2\gamma}}{2}\sum_{j=1}^{J}\big((\hat\theta_j-\mu)^2 + V_{\theta j}\big)$$
$$- \frac{e^{-2\lambda}}{2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\big((y_{ij}-\hat\theta_j)^2 + V_{\theta j}\big).$$

Abbreviate $A = \sum_j\sum_i((y_{ij}-\hat\theta_j)^2 + V_{\theta j})$ and $B(\mu) = \sum_j((\hat\theta_j-\mu)^2 + V_{\theta j})$; both are constants of the M-step apart from the indicated μ. Then

$$\frac{\partial Q}{\partial\mu} = e^{-2\gamma}\sum_{j=1}^{J}(\hat\theta_j-\mu) = 0 \qquad \Longrightarrow\ \mu = \frac{1}{J}\sum_{j=1}^{J}\hat\theta_j,$$
$$\frac{\partial Q}{\partial\lambda} = -n + e^{-2\lambda}A = 0 \qquad \Longrightarrow\ \sigma^2 = e^{2\lambda} = \frac{A}{n},$$
$$\frac{\partial Q}{\partial\gamma} = -(J-1) + e^{-2\gamma}B(\mu) = 0 \qquad \Longrightarrow\ \tau^2 = e^{2\gamma} = \frac{B(\mu)}{J-1},$$

which are exactly (13.14). The μ equation is free of λ and γ, and B is minimized at that same μ, so this stationary point is the joint maximizer and τ^{new} is evaluated at μ^{new}; there $\partial^2Q/\partial\mu^2 = -Je^{-2\gamma}$, $\partial^2Q/\partial\lambda^2 = -2n$, $\partial^2Q/\partial\gamma^2 = -2(J-1)$ and the cross-derivatives vanish (Check!). The divisor $J-1$ rather than J is where the prior enters: $p(\tau)\propto 1$ contributes $+\gamma$ to the log density, turning $-J\gamma$ into $-(J-1)\gamma$, as in (11.16).

(b) EM recovers the assumed hyperparameters to within simulation error. Take $J = 20$ groups of $n_j = 5$, so $n = 100$, and true $(\mu,\sigma,\tau) = (5,2,3)$; the realized $\theta_1,\dots,\theta_{20}$ have mean 4.907 and sample standard deviation 3.466, and the pooled within-group standard deviation of the simulated y is 1.9595. Starting deliberately far away at $(\mu,\sigma,\tau) = (0,5,0.5)$:

iteration	μ	σ	τ	$\log p(\mu,\log\sigma,\log\tau \mid y)$
0 (start)	0.0000	5.0000	0.5000	-218.274
1	0.2306	5.8980	0.5268	-211.768
2	0.4074	5.7805	0.5462	-209.287
5	1.0261	5.2924	0.6204	-200.285
10	2.5843	4.1111	0.9342	-175.182
20	4.8420	1.9587	3.3356	-126.835
500	4.8420	1.9586	3.3356	-126.835

(the log marginal density is (13.12) up to an additive constant, and it increased at every one of the 500 iterations). The marginal posterior mode $(\hat\mu,\hat\sigma,\hat\tau) = (4.842, 1.959, 3.336)$ sits within 0.07, 0.001 and 0.13 of the realized mean of the θ_j, the pooled within-group standard deviation, and the realized

standard deviation of the θ_j, all far inside the sampling errors $\tau/\sqrt{J} \approx 0.67$ and $\tau/\sqrt{2(J-1)} \approx 0.49$. That $\hat{\sigma}$ reproduces the pooled estimate $\sqrt{S/(n-J)}$ to three decimals is not a coincidence: at the fixed point $n\hat{\sigma}^2 = S + \sum_j (n_j V_{\theta j} + n_j(\bar{y}_{.j} - \hat{\theta}_j)^2)$, and the two bracketed terms average to $\sigma^2(1-B_j)$ and $\sigma^2 B_j$ for the shrinkage factor $B_j = (\sigma^2/n_j)/(\sigma^2/n_j + \tau^2)$, contributing $J\sigma^2$ in total.

Problem 13.10. Variational Bayes: Consider probit regression, which is just like logistic except that the function logit^{-1} is replaced by the normal cumulative distribution function. Set up and program variational Bayes for a probit regression with two coefficients (that is, $\Pr(y_i = 1) = \Phi(a + bx_i)$, for $i = 1, \ldots, n$), using the latent-data formulation (so that $z_i \sim \text{N}(a + bx_i, 1)$ and $y_i = 1$ if $z_i > 0$ and 0 otherwise):
(a) Write the log posterior density (up to an arbitrary constant), $p(a, b, z \mid y)$.
(b) Assuming a variational approximation g that is independent in its $n + 2$ dimensions, determine the functional form of each of the factors in g.
(c) Write the steps of the variational Bayes algorithm and program them in R.
Take a uniform prior density on (a, b).

Solution. (a) A truncated Gaussian:

$$\log p(a, b, z \mid y) = -\frac{1}{2}\sum_{i=1}^{n}(z_i - a - bx_i)^2 + \sum_{i=1}^{n}\log \mathbf{1}_{A_i}(z_i) + \text{constant},$$

where $A_i = (0, \infty)$ if $y_i = 1$ and $A_i = (-\infty, 0]$ if $y_i = 0$; equivalently the density is $\exp(-\frac{1}{2}\sum_i (z_i - a - bx_i)^2)$ on the orthant $\{\text{sign}(z_i) = \text{sign}(2y_i - 1)\ \forall i\}$ and 0 outside it. Marginalizing z returns $\prod_i \Phi(a + bx_i)^{y_i}(1 - \Phi(a + bx_i))^{1-y_i}$, since $\Pr(z_i > 0) = \Phi(a + bx_i)$.
(b) Two normals and n truncated normals. Write $g(a, b, z) = g_a(a) g_b(b) \prod_i g_i(z_i)$ and $s_i = 2y_i - 1$.
(i) For a: averaging (a) over b and z leaves $-\frac{1}{2}\sum_i E((z_i - a - bx_i)^2)$, a quadratic in a with coefficient $-\frac{n}{2}a^2$, so

$$g_a(a) = \text{N}(a \mid M_a, S_a^2), \qquad S_a^2 = \frac{1}{n}, \qquad M_a = \frac{1}{n}\sum_{i=1}^{n}\big(E(z_i) - M_b x_i\big).$$

(ii) For b: the same average is a quadratic in b with coefficient $-\frac{1}{2}(\sum_i x_i^2)b^2$, so

$$g_b(b) = \text{N}(b \mid M_b, S_b^2), \quad S_b^2 = \frac{1}{\sum_i x_i^2}, \quad M_b = \frac{\sum_i x_i(E(z_i) - M_a)}{\sum_i x_i^2}.$$

(iii) For z_i: the indicator $\mathbf{1}_{A_i}(z_i)$ involves no other coordinate, so it survives the averaging intact and

$$g_i(z_i) \propto \exp\Big(-\tfrac{1}{2}(z_i - m_i)^2\Big)\mathbf{1}_{A_i}(z_i), \qquad m_i = M_a + M_b x_i,$$

a $\text{N}(m_i, 1)$ density truncated to the half-line selected by y_i. Its mean is the inverse Mills ratio correction

$$E(z_i) = m_i + s_i \frac{\phi(m_i)}{\Phi(s_i m_i)}.$$

Note that S_a and S_b are constants, free of the iteration: the log density in (a) is exactly quadratic in (a, b) with Hessian $-X^T X$ for $X = (\mathbf{1}, x)$, and the mean-field updates see only its diagonal, n and $\sum_i x_i^2$.
(c) Cycle (iii), (i), (ii) to convergence. Explicitly, initialize $M_a = M_b = 0$ and repeat:

1. $m_i \leftarrow M_a + M_b x_i$ and $E(z_i) \leftarrow m_i + s_i\phi(m_i)/\Phi(s_i m_i)$ for $i = 1, \ldots, n$;
2. $M_a \leftarrow \overline{E(z)} - M_b \bar{x}$;
3. $M_b \leftarrow \sum_i x_i(E(z_i) - M_a)\big/\sum_i x_i^2$;

and report $g_a = \text{N}(M_a, 1/n)$, $g_b = \text{N}(M_b, 1/\sum_i x_i^2)$, $g_i = \text{N}(M_a + M_b x_i, 1)$ truncated as in (b)(iii).
Numerically, with $n = 50$, $x_i \sim \text{U}(-2, 2)$ and data simulated from $a = -0.5$, $b = 1.5$ (giving $\sum_i y_i = 21$, $\bar{x} = -0.0751$, $\sum_i x_i^2 = 75.70$), the iterates are

iteration	M_a	M_b
1	-0.1277	0.4990
2	-0.1740	0.7268
5	-0.2551	1.0250
10	-0.3298	1.1969
50	-0.4038	1.3246
200	-0.404205	1.325250

with $S_a = 1/\sqrt{50} = 0.1414$ and $S_b = 1/\sqrt{75.70} = 0.1149$ throughout. The fixed point agrees to six decimals with the probit maximum likelihood estimate, that is, with the posterior mode under the uniform prior; steps 1-3 are precisely the E- and M-steps of the classical probit EM algorithm, and the updates for M_a, M_b never reference S_a, S_b.
The spread, however, is badly understated. Two-dimensional quadrature of $\prod_i \Phi(a+bx_i)^{y_i}(1-\Phi(a+bx_i))^{1-y_i}$ gives

$$E(a \mid y) = -0.454, \quad \mathrm{sd}(a \mid y) = 0.296,$$
$$E(b \mid y) = 1.426, \quad \mathrm{sd}(b \mid y) = 0.321, \quad \mathrm{corr}(a, b \mid y) = -0.417,$$

against which $S_a = 0.141$ and $S_b = 0.115$ are too small by factors 2.1 and 2.8, and the correlation is replaced by 0; the mode-centered normal approximation of Section 4.1 gives 0.284, 0.303 and -0.390. The reason is structural: factorizing z away from (a, b) makes the variational precision for (a, b) the complete-data precision $X^T X$, the precision one would have if z were observed, so all uncertainty contributed by the unknown z is discarded.

Problem 13.11. Unknown normalizing functions: compute the normalizing factor for the following unnormalized sampling density,

$$p(y \mid \mu, A, B, C) \propto \exp\Big(-\tfrac{1}{2}\big(A(y-\mu)^6 + B(y-\mu)^4 + C(y-\mu)^2\big)\Big),$$

as a function of A, B, C. (Hint: it will help to integrate out analytically as many of the parameters as you can.)

Solution.

$$Z(A,B,C) = \frac{1}{3}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\frac{(-B/2)^j}{j!}\frac{(-C/2)^k}{k!} \times \Big(\frac{2}{A}\Big)^{\frac{4j+2k+1}{6}}\Gamma\Big(\frac{4j+2k+1}{6}\Big), \qquad A > 0,$$

so that $p(y \mid \mu, A, B, C) = Z(A,B,C)^{-1}\exp(-\frac{1}{2}(A(y-\mu)^6 + B(y-\mu)^4 + C(y-\mu)^2))$.
Two analytic reductions come first. Substituting $u = y - \mu$ shows that μ is a pure location parameter and drops out, so the normalizing function is

$$Z(A,B,C) = \int_{-\infty}^{\infty} \exp\Big(-\tfrac{1}{2}(Au^6 + Bu^4 + Cu^2)\Big)\, du,$$

finite exactly when $A > 0$, or $A = 0$ and $B > 0$, or $A = B = 0$ and $C > 0$. Second, the integrand is invariant in form under $u = sw$, so one of the three arguments can be scaled away: taking $s = C^{-1/2}$ (for $C > 0$) or $s = A^{-1/6}$,

$$Z(A,B,C) = C^{-1/2}\, G\Big(\frac{A}{C^3}, \frac{B}{C^2}\Big),$$
$$G(\alpha,\beta) = \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\alpha w^6 + \beta w^4 + w^2)}\, dw,$$

with $G(0,0) = \sqrt{2\pi}$; the choice $s = A^{-1/6}$ gives the companion form $Z = A^{-1/6}H(BA^{-2/3}, CA^{-1/3})$ for $H(\beta,\gamma) = \int e^{-\frac{1}{2}(w^6 + \beta w^4 + \gamma w^2)}dw$, which is the one to use when $C \le 0$. This is the point of the hint: for a Metropolis run over (μ, A, B, C) one need only tabulate a function of two arguments, not four.

For the exact series, expand the quartic and quadratic terms about the sextic one, which alone controls the tails:

$$Z = \sum_{j,k\ge 0} \frac{(-B/2)^j}{j!}\,\frac{(-C/2)^k}{k!} \int_{-\infty}^{\infty} u^{4j+2k} e^{-Au^6/2}\,du,$$

and with $\lambda = A/2$, $t = \lambda u^{2m}$ for the general pure power,

$$\int_{-\infty}^{\infty} u^{2p} e^{-\lambda u^{2m}}\,du = \frac{1}{m}\,\lambda^{-\frac{2p+1}{2m}}\,\Gamma\Big(\frac{2p+1}{2m}\Big),$$

which at $m = 3$, $2p = 4j + 2k$ gives the displayed answer. The series converges absolutely for every real B and C whenever $A > 0$: the ratio of consecutive terms in j is $O\big(\Gamma(\frac{2(j+1)}{3})/[(j+1)\Gamma(\frac{2j}{3})]\big) = O(j^{2/3}/j) \to 0$, and in k it is $O(k^{1/3}/k) \to 0$, the Γ growing only like the factorial of $2j/3$ against $j!$. Summing 80×80 terms reproduces adaptive quadrature to ten digits at $(A, B, C) = (1, 1, 1)$, where $Z = 1.5743533548$; at $(0.5, 2, 3)$, $Z = 1.2084957331$; and at $(2, -1, 1)$, $Z = 1.7547732727$, so negative B or C is harmless. The degenerate cases collapse to single terms: $Z = \frac{1}{3}(2/A)^{1/6}\Gamma(\frac{1}{6})$ when $B = C = 0$, $Z = \frac{1}{2}(2/B)^{1/4}\Gamma(\frac{1}{4})$ when $A = C = 0$, and $Z = \sqrt{2\pi/C}$ when $A = B = 0$.

14 Introduction to Regression Models

Contents

14.1 Exercises 14.1–14.7

Problem 14.1. Analysis of radon measurements:
(a) Fit a linear regression to the logarithms of the radon measurements in Table 7.3, with indicator variables for the three counties and for whether a measurement was recorded on the first floor. Summarize your posterior inferences in nontechnical terms.
(b) Suppose another house is sampled at random from Blue Earth County. Sketch the posterior predictive distribution for its radon measurement and give a 95% predictive interval. Express the interval on the original (unlogged) scale. (Hint: you must consider the separate possibilities of basement or first-floor measurement.)
Table 7.3 gives short-term measurements of radon concentration (in picoCuries/liter) in a sample of houses in three counties in Minnesota. All measurements were recorded on the basement level of the houses except those listed in the last column, which were recorded on the first floor. (The printed table lists each county's measurements in a single sequence, with the first-floor ones flagged by an asterisk; they are separated here for legibility.)

County	Basement measurements (pCi/L)	First floor (pCi/L)
Blue Earth	5.0, 13.0, 7.2, 6.8, 12.8, 9.5, 6.0, 3.8, 1.8, 6.9, 4.7, 9.5	5.8, 14.3
Clay	12.9, 2.6, 26.6, 1.5, 13.0, 8.8, 19.5, 9.0, 13.1, 3.6	0.9, 3.5, 2.5, 6.9
Goodhue	14.3, 7.6, 2.6, 43.5, 4.9, 3.5, 4.8, 5.6, 3.5, 3.9, 6.7	6.9, 9.8

Solution. Radon is about 7 pCi/L in a Blue Earth basement, roughly 30% lower on the first floor, essentially the same in all three counties, with a house-to-house geometric standard deviation of about 2.2.
(a) Take y_i to be the measurement in house i, $n = 41$, and fit the ordinary linear regression (14.1) to $\log y$ with no constant term and $k = 4$ columns,

$$X_i = (F_i,\ B_i,\ C_i,\ G_i),$$

where $F_i = 1$ if the measurement was on the first floor and the last three are indicators for Blue Earth, Clay and Goodhue. Under the noninformative prior (14.2), the posterior is (14.3)–(14.7). Least squares gives

$$\hat{\beta} = (-0.328,\ 1.956,\ 1.876,\ 1.918), \qquad s = 0.790,$$

on $n - k = 37$ degrees of freedom, with standard errors $s\sqrt{(V_\beta)_{jj}} = (0.316, 0.216, 0.229, 0.224)$. (The design is balanced enough that X^TX is nearly diagonal: its diagonal is $(8, 14, 14, 13)$, the number of first-floor measurements and the three county sample sizes.)
Drawing $\sigma^2 \sim \text{Inv-}\chi^2(37, s^2)$ and then $\beta \mid \sigma \sim \text{N}(\hat{\beta}, V_\beta\sigma^2)$, 2×10^5 times, and exponentiating, gives posterior medians with central 50% intervals:

Quantity	25%	50%	75%
Blue Earth, basement: e^{β_B}	6.1	7.1	8.2
Clay, basement: e^{β_C}	5.6	6.5	7.6
Goodhue, basement: e^{β_G}	5.8	6.8	7.9
Blue Earth, first floor: $e^{\beta_B+\beta_F}$	4.0	5.1	6.4
Clay, first floor: $e^{\beta_C+\beta_F}$	3.8	4.7	5.8
Goodhue, first floor: $e^{\beta_G+\beta_F}$	3.9	4.9	6.2
first-floor factor: e^{β_F}	0.58	0.72	0.89
geometric sd: e^{σ}	2.09	2.22	2.37

In nontechnical terms: a typical house in any of these three counties has a basement radon reading near 6.5–7 pCi/L, and the three counties are indistinguishable at this sample size (the largest gap between two county coefficients is 0.08, about a quarter of the 0.30 standard error of that difference). Measuring on the first floor rather than in the basement multiplies the reading by an estimated 0.72, with 95% interval $[0.38, 1.36]$ — a reduction of roughly a quarter to a third, but not established by these data alone ($\Pr(\beta_F < 0 \mid y) = 0.85$). Most of the variation is between houses, not between counties or floors: within one county and floor the geometric standard deviation is about 2.2, so a house reading five times another is unremarkable.

(b) Let $\tilde{y}$ be the new measurement, $\tilde{F}$ its first-floor indicator, which is not given. The four-step draw is

$$\begin{aligned}
&\text{(i) } (\beta, \sigma^2) \sim p(\beta, \sigma^2 \mid y) \text{ as in (a);}\\
&\text{(ii) } \tilde{F} \mid \theta \sim \text{Bernoulli}(\theta), \quad \theta \mid y \sim \text{Beta}(3, 13);\\
&\text{(iii) } \log \tilde{y} \mid \beta, \sigma, \tilde{F} \sim \text{N}(\beta_B + \tilde{F}\beta_F,\ \sigma^2);\\
&\text{(iv) } \tilde{y} = e^{\log \tilde{y}}.
\end{aligned}$$

Step (ii) needs a model for the floor: taking the Blue Earth houses as a simple random sample with an exchangeable Bernoulli(θ) floor indicator and a uniform prior on θ, the 2 first-floor out of 14 sampled houses give $\theta \mid y \sim \text{Beta}(3, 13)$.
The resulting posterior predictive distribution is a mixture of two t_{37} densities on the log scale, hence strongly right-skewed on the original scale: sketched as a histogram it rises to a mode between 3 and 4 pCi/L and trails off slowly, with about 2% of the mass beyond 40. Simulation gives quantiles

$$\begin{aligned}
2.5\% : 1.2, \qquad 25\% : 3.8, \quad 50\% : 6.7,\\
75\% : 11.8, \quad 97.5\% : 36.0,
\end{aligned}$$

(Across 20 independent runs of 2×10^5 draws the Monte Carlo standard error on every one of these quantiles is under 0.6% of its value.) The central 95% predictive interval is [1.2, 36] pCi/L; conditioning instead on the floor gives [1.4, 37] for a basement measurement and [0.9, 29] for a first-floor one, so the floor contributes little beside σ.
The published second-edition manual reports $0.8, 2.8, 5.1, 9.0, 33.3$ here; that answer uses the Clay County coefficient rather than Blue Earth and assigns the new house probability 3/16 of having a basement instead of 13/16, so it is overruled.

Problem 14.2. Causal inference using regression: discuss the difference between finite-population and superpopulation inference for the incumbency advantage example of Section 14.3.
In that example, for a given election year the units i are the contested U.S. congressional district elections, y_i is the incumbent party's share of the two-party vote, and the treatment indicator is $R_i = 1$ if the incumbent officeholder runs for reelection and 0 otherwise (an open seat). The control variables are the incumbent party's vote share in the previous election and $P_i = \pm 1$ for whether the Democrats or Republicans hold the seat, plus a constant term. The theoretical incumbency advantage in district i is defined in (14.10) as

$$y^I_{\text{complete}\,i} - y^O_{\text{complete}\,i},$$

the difference of the two potential outcomes, of which only the one corresponding to the observed R_i is seen; the aggregate incumbency advantage for a legislature is the average of these differences over all districts in that general election.

Solution. The aggregate incumbency advantage (14.10) is a finite-population estimand — an average over the actual districts of one election year, in each of which exactly one of the two potential outcomes is missing — whereas the regression coefficient of R reported in Table 14.1 and Figure 14.2 is a superpopulation estimand, a parameter of the model $p(y \mid X, \beta, \sigma)$; the first requires imputing $y^O_{\text{complete}\,i}$ for every district where the incumbent ran and $y^I_{\text{complete}\,i}$ for every open seat, the second does not.
Write the regression as $y_i = \beta_R R_i + X_{i,-R}\beta_{-R} + \epsilon_i$ with $\epsilon_i \mid \sigma \sim \text{N}(0, \sigma^2)$. The superpopulation quantity is simply

$$\beta_R,$$

whose posterior is (14.3)–(14.7); for 1988 its median is 0.114 with 95% interval $[0.084, 0.144]$. The finite-population quantity is

$$\Delta = \frac{1}{n}\sum_{i=1}^{n}\left(y^I_{\text{complete}\,i} - y^O_{\text{complete}\,i}\right),$$

and to simulate it one draws (β, σ) from the posterior and then draws the missing half of each pair from the model, exactly the two-step multiple-imputation scheme of Section 8.2: superpopulation inference $p(\beta, \sigma \mid y)$, then finite-population inference $p(y_{\text{mis}} \mid y, \beta, \sigma)$.
Three differences matter here.
(i) **Estimand versus parameter.** Under the model as fitted — additive, with a single coefficient on R and no interactions — each imputed pair satisfies

$$y^I_{\text{complete}\,i} - y^O_{\text{complete}\,i} = \beta_R + (\epsilon^I_i - \epsilon^O_i),$$

so $\Delta = \beta_R + \overline{\epsilon^I - \epsilon^O}$. In each district one of the two errors is fixed by the observed y_i and only the other is drawn, so with independent errors the imputation adds variance σ^2/n to Δ; with $\sigma \approx 0.066$ (Table 14.1) and $n \approx 300$ contested districts that is a standard deviation of 0.004, against a posterior sd for β_R of $(0.144 - 0.084)/3.92 \approx 0.015$, so the posterior for Δ is about 3% wider than that for β_R and has the same center. If instead one asserts unit-level additivity, $\epsilon^I_i = \epsilon^O_i$, then $\Delta = \beta_R$ identically. Either way Table 14.1 and Figure 14.2 serve as inference for both.
(ii) **Robustness.** Δ depends only on the conditional distribution of y at the n values of X actually in that legislature, whereas β_R is a statement about the infinite population of district elections generated by $p(y \mid X, \beta, \sigma)$. Since exactly half of the $2n$ potential outcomes are observed here, Δ is the less model-dependent of the two: departures from linearity and normality are diluted by the observed half of each pair, while β_R inherits them in full. The outlying residuals tabulated in Table 14.2 make this a real, not rhetorical, advantage.
(iii) **Which one answers the question.** "What proportion of the vote is incumbency worth?" asked of a particular Congress is a finite-population question, and the year-to-year movement in Figure 14.2 argues for reading each year's estimate as a statement about that year's legislature. But (14.10) averages over **all** districts in the general election, while the regression uses only districts contested by both parties in both elections (Table 14.1), excluding the 100 to 150 uncontested seats; there y^I and y^O are both unobserved and the control variable is undefined, so extending Δ to them is extrapolation. Hence the defensible finite-population estimand is the average over contested districts, and β_R remains the natural summary to carry across years — as in the posterior mean 0.050, interval $[0.035, 0.065]$, for the increase from the 1950s to the 1980s.

Problem 14.3. Ordinary linear regression: derive the formulas for $\hat\beta$ and V_β in (14.4)–(14.5) for the posterior distribution of the regression parameters.
That is, for the ordinary linear regression model (14.1), $y \mid \beta, \sigma, X \sim \text{N}(X\beta, \sigma^2 I)$ with X an $n \times k$ matrix of rank k, under the noninformative prior (14.2), $p(\beta, \sigma^2 \mid X) \propto \sigma^{-2}$, show that

$$\beta \mid \sigma, y \sim \text{N}(\hat\beta,\ V_\beta \sigma^2), \qquad \hat\beta = (X^T X)^{-1} X^T y, \qquad V_\beta = (X^T X)^{-1}.$$

Solution. Complete the square. Since the prior (14.2) does not depend on β,

$$p(\beta \mid \sigma, y) \propto p(y \mid \beta, \sigma) \propto \exp\Big(-\tfrac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta)\Big).$$

Put $\hat\beta = (X^T X)^{-1} X^T y$, which exists because $\text{rank}(X) = k$ makes $X^T X$ positive definite, and note $X^T X \hat\beta = X^T y$. Then

$$\begin{aligned}(\beta - \hat\beta)^T X^T X (\beta - \hat\beta) &= \beta^T X^T X \beta - 2\beta^T X^T X \hat\beta + \hat\beta^T X^T X \hat\beta \\ &= \beta^T X^T X \beta - 2\beta^T X^T y + \hat\beta^T X^T y,\end{aligned}$$

so that

$$\begin{aligned}(y - X\beta)^T (y - X\beta) &= y^T y - 2\beta^T X^T y + \beta^T X^T X \beta \\ &= (\beta - \hat\beta)^T X^T X (\beta - \hat\beta) + y^T y - \hat\beta^T X^T y.\end{aligned}$$

The last two terms are free of β, hence

$$p(\beta \mid \sigma, y) \propto \exp\Big(-\tfrac{1}{2\sigma^2}(\beta - \hat\beta)^T (X^T X)(\beta - \hat\beta)\Big),$$

the exponential of a quadratic form in β with positive definite matrix X^TX/σ^2: a normal density with mean $\hat\beta$ and variance matrix $(X^TX)^{-1}\sigma^2$. Comparing with (14.3) gives $V_\beta = (X^TX)^{-1}$.

Problem 14.4. Ordinary linear regression: derive the formula for s^2 in (14.7) for the posterior distribution of the regression parameters.
That is, in the setting of Exercise 14.3, show that

$$\sigma^2 \mid y \sim \text{Inv-}\chi^2(n-k,\ s^2), \qquad s^2 = \frac{1}{n-k}(y - X\hat\beta)^T(y - X\hat\beta).$$

Solution. Divide, as in the display preceding (14.6): $p(\sigma^2 \mid y) = p(\beta, \sigma^2 \mid y)/p(\beta \mid \sigma^2, y)$, evaluated at any convenient β — take $\beta = \hat\beta$, at which the second factor is just its normalizing constant.
By (14.1)–(14.2) the numerator is

$$p(\beta, \sigma^2 \mid y) \propto \sigma^{-2}(\sigma^2)^{-n/2} \exp\Big(-\tfrac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\Big),$$

and by Exercise 14.3 the denominator is

$$p(\beta \mid \sigma^2, y) = (2\pi)^{-k/2}|V_\beta|^{-1/2}(\sigma^2)^{-k/2} \exp\Big(-\tfrac{1}{2\sigma^2}(\beta - \hat\beta)^T V_\beta^{-1}(\beta - \hat\beta)\Big).$$

At $\beta = \hat\beta$ both exponentials simplify, $|V_\beta|$ is free of σ^2, and

$$p(\sigma^2 \mid y) \propto (\sigma^2)^{-(n-k)/2-1} \exp\Big(-\tfrac{1}{2\sigma^2}(y - X\hat\beta)^T(y - X\hat\beta)\Big).$$

Matching this against the scaled inverse-χ^2 density of Appendix A,

$$p(\sigma^2) \propto (\sigma^2)^{-\nu/2-1} \exp\big(-\nu s^2/(2\sigma^2)\big),$$

gives $\nu = n - k$ and $\nu s^2 = (y - X\hat\beta)^T(y - X\hat\beta)$, which is (14.7).
Method (2): substituting the completed square of Exercise 14.3 into the numerator at a general β leaves

$$p(\sigma^2 \mid y) \propto (\sigma^2)^{-(n-k)/2-1} \exp\big(-(n-k)s^2/(2\sigma^2)\big)$$

after the two β-dependent factors cancel exactly — which they must, since the ratio cannot depend on β. (Check!)

Problem 14.5. Analysis of the milk production data: consider how to analyze data from the cow experiment described in Section 8.4. Specifically:
(a) Discuss the role of the treatment assignment mechanism for the appropriate analysis from a Bayesian perspective.
(b) Discuss why you would focus on finite-population inferences for these 50 cows or on superpopulation inferences for the hypothetical population of cows from which these 50 are conceptualized as a random sample. Either focus is legitimate, and a reasonable answer might be that one is easier than the other, but if this is your answer, say why it is true.
The experiment of Section 8.4: 50 cows, four diets (treatments) corresponding to different levels of a feed additive (methionine hydroxy analog), and six outcomes related to the amount of milk fat produced by each cow. Three covariates were recorded before treatment assignment: lactation number (seasons of lactation), age, and initial weight of cow. Cows were initially assigned to treatments completely at random; the distributions of the three covariates were then checked for balance across treatment groups, several randomizations were tried, and the one giving the best balance with respect to the three covariates was chosen. In the notation of Chapter 8, the covariates x form a 50×3 matrix and the complete data y a 50×24 matrix, of which only one of the four possible subvectors of dimension 6 is observed for each cow.

Solution. (a) The assignment mechanism is ignorable but unknown, so the analysis may model y given x and the treatment indicator and ignore $p(I \mid x, y, \phi)$ entirely — provided x is in the model.
Ignorability holds because the re-randomization decisions were functions of the observed covariates x (lactation number, age, initial weight) alone: they did not depend on the outcomes y, nor on unrecorded variables such as the cows' physical appearance or the times at which they entered the study. It is unknown because the rule for deciding whether to re-randomize was never written down, so $p(I \mid x, \phi)$ is unavailable; but by the ignorability result of Section 8.2 the inference is unaffected, since for fixed model and data the posterior distribution of ω and of finite-population estimands is the same for all ignorable data collection mechanisms.
Two conditions carry real content. First, **distinct parameters**: the analysis assumes ϕ and ω are a priori independent. This would fail if the experimenter's choice among candidate randomizations were driven by beliefs about treatment efficacy rather than by covariate balance alone; reasonable violations should not much affect inferences here (Exercise 8.3).
Second, and this is the operative point, ignorability holds only **conditional on** x. Because the selected randomization was chosen for balance on the three covariates, the assignment is not independent of x, so a model for y given treatment alone, marginalizing over x, is not justified by the argument above. The minimal valid analysis is therefore a model for mean daily milk fat conditional on the treatment **and** the three pre-treatment variables — a linear additive normal regression after suitable transformations, since milk-fat measurements are positive and right-skewed, fitted by (14.3)–(14.7). Better still, model the six-dimensional post-treatment outcome jointly on treatment and the three covariates as a multivariate normal regression, since the six measures are correlated and the scientific question concerns all of them.
(The balancing step is no Bayesian necessity, but it improves the design by making inferences less sensitive to the form in which x enters the regression; a randomized block design on the covariates would have achieved the same with a known assignment mechanism, as Section 8.5 notes.)
(b) Superpopulation, on both grounds: it is what the experiment is about, and it is the easier of the two.
The scientific question is the effect of the additive on cows in general — there was no particular interest in the 50 cows that happened to be in the study — so the estimands of interest are the regression coefficients β on the four treatment levels, whose posterior comes directly from (14.3)–(14.7). A finite-population estimand such as

$$\Delta_{j1} = \frac{1}{50}\sum_{i=1}^{50}\bigl(y^{(j)}_{\text{complete}\,i} - y^{(1)}_{\text{complete}\,i}\bigr)$$

would require imputing y_{mis} — and here y_{mis} is three quarters of the 50×24 complete-data array, since only one of four 6-vectors is observed per cow.
That fraction is exactly why the usual robustness argument for finite-population inference (Section 8.2) does not apply. That argument requires that a large fraction of the potential outcomes be observed, so that the estimand leans on data rather than on the additivity and linearity assumptions. Here, of the 100 unit-level errors entering the contrast Δ_{j1}, only about $2 \times 12.5 = 25$ are determined by observed residuals and about 75 must be drawn from the model; the imputation therefore rests on the same additivity assumption as $\beta_j - \beta_1$ does, and buys no protection. Counting variances in the additive model $y_i^{(t)} = \mu + \beta_t + x_i\gamma + \epsilon_i^{(t)}$ with $\epsilon \sim \text{N}(0, \sigma^2)$,

$$\Delta_{j1} = \beta_j - \beta_1 + \frac{1}{50}\sum_{i=1}^{50}\bigl(\epsilon_i^{(j)} - \epsilon_i^{(1)}\bigr),$$

so the extra posterior variance is about $75\sigma^2/50^2 = 0.03\sigma^2$, against $\text{var}(\beta_j - \beta_1 \mid y) \approx \sigma^2(1/12.5 + 1/12.5) = 0.16\sigma^2$: the finite-population interval is roughly 9% wider and centered in the same place. So the two inferences agree numerically, the finite-population one costs an extra imputation step, and the superpopulation one is the one the experimenter asked for.

Problem 14.6. Ordinary linear regression: derive the conditions that the posterior distribution is proper in Section 14.2.
That is, for the model (14.1), $y \mid \beta, \sigma, X \sim \mathrm{N}(X\beta, \sigma^2 I)$ with X an $n \times k$ matrix, and the improper prior (14.2), $p(\beta, \sigma^2 \mid X) \propto \sigma^{-2}$, show that $p(\beta, \sigma^2 \mid y)$ has a finite integral if and only if (1) $n > k$ and (2) the rank of X equals k.

Solution. Integrate β out first, then σ^2; each step fails for exactly one of the two conditions.
The unnormalized posterior is

$$q(\beta, \sigma^2) = (\sigma^2)^{-n/2-1} \exp\Big(-\tfrac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\Big).$$

(i) *The β integral.* Write P for the orthogonal projection onto the column space of X, so that

$$(y - X\beta)^T(y - X\beta) = \|(I - P)y\|^2 + \|Py - X\beta\|^2.$$

If $\mathrm{rank}(X) = r < k$, pick $0 \neq v \in \ker X$; then $q(\beta + tv, \sigma^2) = q(\beta, \sigma^2)$ for all $t \in \mathbb{R}$, so $\int q\, d\beta = \infty$ for every σ^2 and the posterior is improper. If $\mathrm{rank}(X) = k$ then X^TX is positive definite, the completed square of Exercise 14.3 applies, and the Gaussian integral

$$\int_{\mathbb{R}^k} \exp\Big(-\tfrac{(\beta-\hat\beta)^T X^T X(\beta-\hat\beta)}{2\sigma^2}\Big) d\beta = (2\pi\sigma^2)^{k/2} |X^TX|^{-1/2}$$

is finite. Hence condition (2) is necessary and sufficient for $\int q\, d\beta < \infty$, and it gives

$$p(\sigma^2 \mid y) \propto (\sigma^2)^{-(n-k)/2-1} \exp\big(-(n-k)s^2/(2\sigma^2)\big),$$

with $(n-k)s^2 = \|(I-P)y\|^2 = (y - X\hat\beta)^T(y - X\hat\beta)$ as in (14.7). Note $\mathrm{rank}(X) = k$ forces $k \leq n$, so (2) already excludes $n < k$.
(ii) *The σ^2 integral.* Substituting $u = (n-k)s^2/(2\sigma^2)$,

$$\int_0^\infty (\sigma^2)^{-\nu/2-1} e^{-\nu s^2/(2\sigma^2)}\, d\sigma^2 = \Gamma\big(\tfrac{\nu}{2}\big) \Big(\tfrac{\nu s^2}{2}\Big)^{-\nu/2}, \qquad \nu = n - k,$$

which is finite precisely when $\nu > 0$ and $\nu s^2 > 0$. The two ends of the range fail separately: for $\nu \leq 0$ the integrand is not integrable at $\sigma^2 \to \infty$, and for $s^2 = 0$ it is $(\sigma^2)^{-\nu/2-1}$ throughout, not integrable at $\sigma^2 \to 0$. Given (2), $n = k$ makes X square and invertible, so $\|(I-P)y\|^2 = 0$ and both failures occur at once; the posterior is then $p(\sigma^2 \mid y) \propto (\sigma^2)^{-1}$, improper. And $n > k$ with $\mathrm{rank}(X) = k$ gives $s^2 > 0$ unless y lies exactly in the k-dimensional column space of X, an event of probability zero under (14.1). So the posterior is proper exactly when $n > k$ and $\mathrm{rank}(X) = k$.

Problem 14.7. Posterior predictive distribution for ordinary linear regression: show that $p(\tilde y \mid \sigma, y)$ is a normal density. (Hint: first show that $p(\tilde y, \beta \mid \sigma, y)$ is the exponential of a quadratic form in $(\tilde y, \beta)$ and is thus a normal density.)
Here, in the setting of Section 14.2, the new data satisfy $\tilde y \mid \beta, \sigma, \tilde X \sim \mathrm{N}(\tilde X\beta, \sigma^2 I)$ for an observed $\tilde n \times k$ matrix $\tilde X$ of explanatory variables, independently of y given (β, σ), and $\beta \mid \sigma, y \sim \mathrm{N}(\hat\beta, V_\beta\sigma^2)$ by (14.3)–(14.5).

Solution. Normal, with mean $\tilde X\hat\beta$ and variance $(I + \tilde X V_\beta \tilde X^T)\sigma^2$.
Since $\tilde y$ depends on y only through (β, σ),

$$\begin{aligned} p(\tilde y, \beta \mid \sigma, y) &= p(\tilde y \mid \beta, \sigma)\, p(\beta \mid \sigma, y) \\ &\propto \exp\Big(-\tfrac{1}{2\sigma^2}\big[(\tilde y - \tilde X\beta)^T(\tilde y - \tilde X\beta) + (\beta - \hat\beta)^T V_\beta^{-1}(\beta - \hat\beta)\big]\Big), \end{aligned}$$

a quadratic form in the stacked vector $(\tilde y, \beta)$. Expanding the bracket, the coefficient matrix of the

quadratic part is

$$Q = \frac{1}{\sigma^2}\begin{pmatrix} I & -\tilde{X} \\ -\tilde{X}^T & \tilde{X}^T\tilde{X} + V_\beta^{-1} \end{pmatrix},$$

the linear and constant parts being absorbed by a shift of centre. The form is positive definite: for $(u, v) \neq 0$,

$$\sigma^2(u,v)^T Q(u,v) = \|u - \tilde{X}v\|^2 + v^T V_\beta^{-1} v > 0,$$

since $V_\beta^{-1} = X^T X$ is positive definite (Exercise 14.6, condition (2)), so the two terms vanish together only if $v = 0$ and then $u = 0$. Hence the exponential is integrable and $p(\tilde{y}, \beta \mid \sigma, y)$ is a $(\tilde{n} + k)$-dimensional normal density with variance matrix Q^{-1}.
Every marginal of a multivariate normal is normal (Appendix A), so $p(\tilde{y} \mid \sigma, y) = \int p(\tilde{y}, \beta \mid \sigma, y)\, d\beta$ is a normal density. Its moments then follow from (2.7) and (2.8) as in the derivation of (14.8)–(14.9):

$$\begin{aligned} \mathrm{E}(\tilde{y} \mid \sigma, y) &= \mathrm{E}\big(\mathrm{E}(\tilde{y} \mid \beta, \sigma, y) \mid \sigma, y\big) = \mathrm{E}(\tilde{X}\beta \mid \sigma, y) = \tilde{X}\hat{\beta}, \\ \mathrm{var}(\tilde{y} \mid \sigma, y) &= \mathrm{E}(\sigma^2 I \mid \sigma, y) + \mathrm{var}(\tilde{X}\beta \mid \sigma, y) = (I + \tilde{X} V_\beta \tilde{X}^T)\sigma^2. \end{aligned}$$

Method (2): $\tilde{y} = \tilde{X}\beta + \tilde{\epsilon}$ with $\tilde{\epsilon} \mid \sigma \sim \mathrm{N}(0, \sigma^2 I)$ independent of $\beta \mid \sigma, y \sim \mathrm{N}(\hat{\beta}, V_\beta \sigma^2)$, and a sum of independent normal vectors is normal with the sum of the means and of the variance matrices — giving the same $\tilde{X}\hat{\beta}$ and $(I + \tilde{X} V_\beta \tilde{X}^T)\sigma^2$ at once.

14.2 Exercises 14.8–14.14

Problem 14.8. Expression of prior information as additional data: give an algebraic proof of (14.24). That is, for the linear model $y \mid \beta \sim \mathrm{N}(X\beta, \Sigma_y)$ with X an $n \times k$ matrix of known predictors and Σ_y known, and with the conjugate normal prior distribution $\beta \sim \mathrm{N}(\beta_0, \Sigma_\beta)$ (Σ_β^{-1} of rank k), show that the resulting posterior distribution for β is identical to the posterior distribution obtained from the *noninformative* weighted regression of y_* on X_* with known variance matrix Σ_*, where

$$y_* = \begin{pmatrix} y \\ \beta_0 \end{pmatrix}, \qquad X_* = \begin{pmatrix} X \\ I_k \end{pmatrix}, \qquad \Sigma_* = \begin{pmatrix} \Sigma_y & 0 \\ 0 & \Sigma_\beta \end{pmatrix}.$$

Solution. The two quadratic forms in β are literally the same one, because Σ_* is block diagonal:

$$\begin{aligned} (y_* - X_*\beta)^T \Sigma_*^{-1} (y_* - X_*\beta) &= \begin{pmatrix} y - X\beta \\ \beta_0 - \beta \end{pmatrix}^T \begin{pmatrix} \Sigma_y^{-1} & 0 \\ 0 & \Sigma_\beta^{-1} \end{pmatrix} \begin{pmatrix} y - X\beta \\ \beta_0 - \beta \end{pmatrix} \\ &= (y - X\beta)^T \Sigma_y^{-1} (y - X\beta) \\ &\quad + (\beta - \beta_0)^T \Sigma_\beta^{-1} (\beta - \beta_0). \end{aligned}$$

The right side is exactly -2 times the log of the informative-prior posterior kernel,

$$p(\beta \mid y) \propto p(y \mid \beta)\, p(\beta) \propto \exp\Big(-\tfrac{1}{2}\big[(y - X\beta)^T \Sigma_y^{-1} (y - X\beta) + (\beta - \beta_0)^T \Sigma_\beta^{-1} (\beta - \beta_0)\big] \Big),$$

while the left side is -2 times the log of the augmented-regression posterior kernel $p(\beta \mid y_*) \propto \exp\big(-\tfrac{1}{2}(y_* - X_*\beta)^T \Sigma_*^{-1} (y_* - X_*\beta)\big)$ under the flat prior $p(\beta) \propto 1$ on the augmented regression. Since the normalizing constants of two densities proportional to the same function of β agree, the two posterior distributions coincide.
Reading off the parameters: both are $\mathrm{N}(\hat{\beta}_*, V_{\beta_*})$ with, by (14.12)–(14.13) applied to the augmented data,

$$\begin{aligned} V_{\beta_*}^{-1} &= X_*^T \Sigma_*^{-1} X_* = X^T \Sigma_y^{-1} X + \Sigma_\beta^{-1}, \\ V_{\beta_*}^{-1} \hat{\beta}_* &= X_*^T \Sigma_*^{-1} y_* \\ &= X^T \Sigma_y^{-1} y + \Sigma_\beta^{-1} \beta_0, \end{aligned}$$

which is the familiar precision-weighted combination of the data estimate and the prior mean. The flat-prior step is legitimate because X_* has full column rank k (its lower block is I_k), so the noninformative

weighted regression has a proper posterior even when X does not — this is the hypothesis $\text{rank}(\Sigma_\beta^{-1}) = k$ of the stated result.

Problem 14.9. Lasso regularization:
(a) Write the (unnormalized) posterior density for regression coefficients with a lasso prior with parameter λ. Suppose the least-squares estimate is $\hat{\beta}$ with covariance matrix $V_\beta \sigma^2$. (For simplicity, treat the data variance σ as known.)
(b) Suppose β is one-dimensional. Give the lasso (posterior mode) estimate of β.
(c) Suppose β is multidimensional. Explain why, in general, the lasso estimate cannot simply be found by pulling the least-squares estimate of each coefficient toward zero.

Solution. (a) The lasso prior is the independent double-exponential (Laplace) distribution $p(\beta) = \prod_{j=1}^k \frac{\lambda}{2} \exp(-\lambda|\beta_j|)$. With σ known, the likelihood depends on β only through the least-squares quadratic form $p(y \mid \beta, \sigma) \propto \exp\big(-\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T V_\beta^{-1}(\beta - \hat{\beta})\big)$, where $V_\beta = (X^T X)^{-1}$ by (14.5). Hence

$$p(\beta \mid y, \sigma) \propto \exp\Big(-\tfrac{1}{2\sigma^2}(\beta - \hat{\beta})^T V_\beta^{-1}(\beta - \hat{\beta}) - \lambda \sum_{j=1}^k |\beta_j|\Big).$$

The posterior mode therefore minimizes $\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T V_\beta^{-1}(\beta - \hat{\beta}) + \lambda \sum_j |\beta_j|$, which is the usual lasso criterion $\|y - X\beta\|^2/(2\sigma^2) + \lambda\|\beta\|_1$ up to a constant.
(b) Soft thresholding at $\lambda\sigma^2 V_\beta$:

$$\hat{\beta}_{\text{lasso}} = \text{sign}(\hat{\beta})\,\big(|\hat{\beta}| - \lambda\sigma^2 V_\beta\big)_+, \qquad V_\beta = 1/\textstyle\sum_i x_i^2.$$

Indeed, for $\beta > 0$ the criterion $f(\beta) = (\beta - \hat{\beta})^2/(2\sigma^2 V_\beta) + \lambda\beta$ has $f'(\beta) = (\beta - \hat{\beta})/(\sigma^2 V_\beta) + \lambda$, giving the stationary point $\beta = \hat{\beta} - \lambda\sigma^2 V_\beta$, which is admissible exactly when $\hat{\beta} > \lambda\sigma^2 V_\beta$; symmetrically for $\beta < 0$; and when $|\hat{\beta}| \le \lambda\sigma^2 V_\beta$ the subdifferential of f at 0, $[-\hat{\beta}/(\sigma^2 V_\beta) - \lambda,\ -\hat{\beta}/(\sigma^2 V_\beta) + \lambda]$, contains 0, so the mode is at $\beta = 0$. (The criterion is convex, so the stationary point is the global minimum.)
(c) Because $V_\beta^{-1} = X^T X$ is not diagonal, the quadratic form couples the coordinates. The subgradient (Karush–Kuhn–Tucker) conditions for the mode are

$$\tfrac{1}{\sigma^2}\big[V_\beta^{-1}(\beta - \hat{\beta})\big]_j = -\lambda\, s_j, \qquad s_j = \text{sign}(\beta_j) \text{ if } \beta_j \neq 0, \quad s_j \in [-1, 1] \text{ if } \beta_j = 0,$$

so coordinate j of the solution is determined by the *whole* vector β, not by $\hat{\beta}_j$ alone. Only in the orthogonal-design case $X^T X \propto I$ do these conditions decouple into the k separate one-dimensional problems of part (b), for which coordinatewise soft thresholding is exact. Concretely, once the lasso drives some coefficients to zero the survivors are refit in the reduced model, which can move them *away* from zero, and no coordinatewise shrinkage of $\hat{\beta}$ can reproduce that.

Problem 14.10. Lasso regularization: Find a linear regression example, in an application area of interest to you, with many predictors.
(a) Do a lasso regression, that is, a Bayesian regression using a double-exponential prior distribution with hyperparameter λ estimated from data, summarizing by the posterior mode of the coefficients.
(b) Get uncertainty in this estimate by bootstrapping the data 100 times and repeating the above step for each bootstrap sample.
(c) Now fit a fully Bayesian lasso (that is, the same model as in (a) but assigning a hyperprior to λ and obtaining simulations over the entire posterior distribution).
(d) Compare your inferences in (a), (b), and (c).

Solution. Data: the diabetes-progression regression of Efron, Hastie, Johnstone and Tibshirani (2004), $n = 442$ patients, $k = 10$ predictors — age, sex, body-mass index, mean arterial blood pressure, and six blood-serum measurements $s_1, \ldots, s_6$ — with y a quantitative measure of disease progression one

year after baseline. Predictors are standardized to mean 0, standard deviation 1 and y centered, so no intercept is needed; two serum pairs are strongly collinear ($\mathrm{corr}(s_1, s_2) = 0.897$), which is what makes the shrinkage interesting. Least squares gives $s_{\mathrm{ols}} = 54.2$, used as the known σ in (a) and (b). Throughout, the mode minimizes the criterion of exercise 14.9, $\|y - X\beta\|^2/(2\sigma^2) + \lambda\|\beta\|_1$.

(a) With λ chosen by 10-fold cross-validation, $\hat{\lambda} = 0.178$, and the posterior mode is (three coefficients exactly zero)

predictor	$\hat{\beta}_{\mathrm{ols}}$	lasso mode
age	-0.48	0.00
sex	-11.41	-9.00
bmi	24.73	24.79
bp	15.43	13.92
s_1	-37.68	-4.46
s_2	22.68	0.00
s_3	4.81	-10.52
s_4	8.42	0.00
s_5	35.73	24.19
s_6	3.22	2.41

Note $\hat{\beta}_{s_3}$ changing *sign* between least squares and the mode: the collinear serum block is being refit, exactly as anticipated in 14.9(c).

(b) Resampling the 442 rows with replacement 100 times and rerunning (a) — with λ re-selected by cross-validation in each replicate — gives

predictor	boot. median	95% interval	$\Pr(\hat{\beta}_j = 0)$
age	0.00	$[-3.8,\ 6.4]$	0.13
sex	-10.88	$[-16.9,\ -5.3]$	0.00
bmi	25.09	$[18.9,\ 30.4]$	0.00
bp	14.99	$[8.8,\ 21.0]$	0.00
s_1	-13.14	$[-68.3,\ 0.0]$	0.16
s_2	0.00	$[-8.3,\ 41.3]$	0.40
s_3	-3.62	$[-16.7,\ 22.1]$	0.16
s_4	6.70	$[-6.4,\ 21.9]$	0.27
s_5	28.38	$[18.2,\ 46.4]$	0.00
s_6	2.93	$[-0.8,\ 8.2]$	0.15

The median selected λ across replicates was 0.033, a factor of five below the 0.178 of part (a). This is not a real disagreement: the cross-validation curve is almost perfectly flat, its estimated mean squared error varying by about 0.6% (2987 to 3004) across the whole range $\lambda \in [0.007, 0.33]$, so $\hat{\lambda}$ is barely determined; and bootstrap resamples contain duplicated rows, which leak across cross-validation folds and bias the selected penalty downward.

(c) Fully Bayesian lasso via the scale-mixture representation of Park and Casella (2008): the double-exponential prior is a normal mixed over an exponential scale,

$$\beta_j \mid \sigma^2, \tau_j^2 \sim \mathrm{N}(0, \sigma^2\tau_j^2), \qquad \tau_j^2 \sim \mathrm{Exp}(\lambda_0^2/2), \qquad p(\sigma^2) \propto 1/\sigma^2, \qquad \lambda_0^2 \sim \mathrm{Gamma}(1, 1),$$

so that marginally $p(\beta_j \mid \sigma) \propto \exp(-\lambda_0|\beta_j|/\sigma)$, i.e. $\lambda = \lambda_0/\sigma$ in the parametrization of (a). All full conditionals are standard, giving a Gibbs sampler:

$$\begin{aligned}
\beta \mid \tau, \sigma^2, y &\sim \mathrm{N}\big(A^{-1}X^Ty,\ \sigma^2A^{-1}\big), \quad A = X^TX + D_\tau^{-1},\ D_\tau = \mathrm{diag}(\tau_j^2),\\
\sigma^2 \mid \beta, \tau, y &\sim \text{Inv-gamma}\Big(\tfrac{n-1+k}{2},\ \tfrac{1}{2}\big[\|y - X\beta\|^2 + \beta^TD_\tau^{-1}\beta\big]\Big),\\
1/\tau_j^2 \mid \beta, \sigma^2, \lambda_0 &\sim \text{Inv-Gaussian}\big(\sqrt{\lambda_0^2\sigma^2/\beta_j^2},\ \lambda_0^2\big),\\
\lambda_0^2 \mid \tau &\sim \mathrm{Gamma}\big(k + 1,\ \tfrac{1}{2}\textstyle\sum_j \tau_j^2 + 1\big).
\end{aligned}$$

Two chains of 60,000 draws after 10,000 burn-in give $\widehat{R} \le 1.001$ for every parameter and Monte Carlo standard errors below 0.1 on the medians below (largest, 0.08, for the collinear β_{s_1}) — negligible against the posterior widths. Posterior medians and central 95% intervals:

predictor	median	95% interval	$\Pr(\beta_j > 0 \mid y)$
age	-0.29	$[-5.7,\ 5.1]$	0.46
sex	-10.91	$[-16.6,\ -5.2]$	0.00
bmi	24.85	$[18.7,\ 31.1]$	1.00
bp	15.08	$[9.1,\ 21.1]$	1.00
s_1	-17.19	$[-49.5,\ 6.5]$	0.08
s_2	6.40	$[-13.4,\ 33.4]$	0.72
s_3	-3.71	$[-18.1,\ 12.0]$	0.31
s_4	5.70	$[-7.0,\ 19.4]$	0.81
s_5	28.21	$[16.6,\ 42.1]$	1.00
s_6	3.15	$[-2.7,\ 9.2]$	0.85

with σ having posterior median 53.9 and $\lambda = \lambda_0/\sigma$ posterior median 0.035. Varying (r, δ) in $\lambda_0^2 \sim \text{Gamma}(r, \delta)$ over $\{(1,1), (0.1, 0.1), (10^{-3}, 10^{-3}), (1, 0.01)\}$ moves the posterior median of λ across $\{0.035, 0.064, 0.089, 0.103\}$, leaves the strongly identified coefficients alone (β_{bmi} median $24.83, 24.85, 24.86, 24.86$), and moves β_{s_1} from -17.2 to -7.4.

(d) Three comparisons.

(i) *Point estimates.* Mode, bootstrap median, and posterior median agree closely on the four coefficients the data determine sharply (sex, bmi, bp, s_5) — all within two-thirds of a bootstrap standard error. They disagree substantially inside the collinear serum block: the mode reports $\hat{\beta}_{s_1} = -4.5$ with $\hat{\beta}_{s_2} = 0$, while the posterior median reports -17.2 and $+6.4$. The posterior distribution of $(\beta_{s_1}, \beta_{s_2})$ has correlation -0.939; a ridge of nearly equal posterior density runs along $\beta_{s_1} + \beta_{s_2} \approx \text{const}$, and the mode simply picks one arbitrary point on it while the median averages over it.

(ii) *Sparsity.* The mode sets three coefficients exactly to zero; the fully Bayesian posterior sets none to zero, since the posterior density of β_j is absolutely continuous. Sparsity is a property of the mode, not of the posterior distribution. The bootstrap gives the intermediate summary $\Pr(\hat{\beta}_j = 0)$, which ranges from 0 to 0.40 and behaves like a crude inclusion probability.

(iii) *Uncertainty.* Bootstrap and fully Bayesian intervals are similar in width for the well-determined coefficients (for bmi, $[18.9, 30.4]$ versus $[18.7, 31.1]$), but the bootstrap intervals are distorted where selection bites: β_{s_1} has bootstrap interval $[-68.3, 0.0]$, truncated at zero by the selection event, whereas the posterior interval $[-49.5, 6.5]$ crosses zero smoothly. The bootstrap conflates two sources of variation — sampling variability in y *and* variability in the selected $\hat{\lambda}$ — and it conditions on a point estimate of λ within each replicate rather than averaging over λ, so it does not propagate uncertainty in λ coherently. The fully Bayesian analysis does, at the price of giving up the sparse summary.

Problem 14.11. Straight-line fitting with variation in x and y: suppose we wish to model two variables, x and y, as having an underlying linear relation with added errors. That is, with data $(x, y)_i$, $i = 1, \ldots, n$, we model

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} u_i \\ v_i \end{pmatrix}, \Sigma \right), \qquad v_i = a + b u_i.$$

The goal is to estimate the underlying regression parameters, (a, b).

(a) Assume that the values u_i follow a normal distribution with mean μ and variance τ^2. Write the likelihood of the data given the parameters; you can do this by integrating over $u_1, \ldots, u_n$ or by working with the multivariate normal distribution.

(b) Discuss reasonable noninformative prior distributions on (a, b).

See Snedecor and Cochran (1989, p. 173) for an approximate solution, and Gull (1989b) for a Bayesian treatment of the problem of fitting a line with errors in both variables.

Solution. (a) The likelihood is a product of n identical bivariate normal densities,

$$p(x, y \mid a, b, \mu, \tau^2, \Sigma) = \prod_{i=1}^{n} \text{N}\left(\begin{pmatrix} x_i \\ y_i \end{pmatrix} \;\middle|\; m,\ \Sigma + \tau^2 c c^T \right), \quad m = \begin{pmatrix} \mu \\ a + b\mu \end{pmatrix},\ c = \begin{pmatrix} 1 \\ b \end{pmatrix}.$$

This follows from the multivariate normal route with no integration. Write the latent mean as an affine function of the single latent scalar u_i,

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} 0 \\ a \end{pmatrix} + \begin{pmatrix} 1 \\ b \end{pmatrix} u_i = \alpha + c\, u_i,$$

so that $(x_i, y_i)^T = \alpha + c u_i + \epsilon_i$ with $\epsilon_i \sim \mathrm{N}(0, \Sigma)$ independent of $u_i \sim \mathrm{N}(\mu, \tau^2)$. A linear function of independent normals is normal, with

$$\mathrm{E}\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \alpha + c\mu = \begin{pmatrix} \mu \\ a + b\mu \end{pmatrix},$$
$$\mathrm{var}\begin{pmatrix} x_i \\ y_i \end{pmatrix} = c\,\tau^2 c^T + \Sigma = \tau^2 \begin{pmatrix} 1 & b \\ b & b^2 \end{pmatrix} + \Sigma,$$

and independence across i since (u_i, ϵ_i) are. Explicitly, with $\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$, the marginal covariance is

$$V = \begin{pmatrix} \tau^2 + \sigma_x^2 & b\tau^2 + \rho\sigma_x\sigma_y \\ b\tau^2 + \rho\sigma_x\sigma_y & b^2\tau^2 + \sigma_y^2 \end{pmatrix},$$

and the log-likelihood is, writing $\bar{z}$ and S for the sample mean vector and the sample covariance $\frac{1}{n}\sum_i (z_i - \bar{z})(z_i - \bar{z})^T$ of $z_i = (x_i, y_i)^T$,

$$\log p(x, y \mid \cdot) = -\tfrac{n}{2} \log |V| - \tfrac{n}{2} \operatorname{tr}(V^{-1} S) - \tfrac{n}{2} (\bar{z} - m)^T V^{-1} (\bar{z} - m) + \text{const.}$$

So $(\bar{z}, S)$ is sufficient. (By integration one gets the same answer: the integrand is exp of a quadratic form in u_i, and completing the square in u_i and integrating gives the displayed bivariate normal — the multivariate route avoids the algebra.)

Identifiability is the substantive point here. The observable model is a bivariate normal with 5 free parameters (two means, three covariances), while $(a, b, \mu, \tau^2, \Sigma)$ has $2 + 2 + 3 = 7$: the model is *not* identified if Σ is unknown. Some restriction on Σ is required — for example Σ known, Σ known up to a scalar multiple (a known variance *ratio*, as in 14.12(c)), $\Sigma = \sigma^2 I$, or Σ diagonal with one of σ_x^2, σ_y^2 known. Under $\Sigma = \sigma^2 I$ the count is 5 against 5 and the model is exactly saturated; Σ diagonal with both entries free leaves 6 parameters for 5 observable moments and is still unidentified. In a Bayesian analysis the posterior distribution is then improper under a noninformative prior, and honest practice is either to fix the ratio or to give σ_y^2/σ_x^2 a proper prior distribution and report that the inference for b depends on it.

(b) Uniform on (a, b) is the default, but the good choice is

$$p(a, b) \propto (1 + b^2)^{-3/2},$$

which is uniform in the line's angle and perpendicular offset, and is therefore invariant to rotations of the (x, y) plane — the natural symmetry here, because the problem treats x and y on the same footing and nothing distinguishes "regress y on x" from "regress x on y." Reparametrize the line $v = a + bu$ by $b = \tan\theta$, $\theta \in (-\pi/2, \pi/2)$, and the signed perpendicular distance from the origin $c = a\cos\theta$. Then $db = \sec^2\theta\, d\theta$ and, at fixed θ, $da = \sec\theta\, dc$, so

$$da\, db = \sec^3\theta\; dc\, d\theta = (1 + b^2)^{3/2}\, dc\, d\theta,$$

and $p(c, \theta) \propto 1$ pulls back to $p(a, b) \propto (1 + b^2)^{-3/2}$. A flat prior on (a, b), by contrast, puts overwhelming weight on near-vertical lines ($\int^\infty db$ diverges while the angle range is bounded) and so is not noninformative in any useful sense; it shifts b away from zero, by $3b/[(1 + b^2)\,\mathrm{prec}(b)]$ to first order. Both priors are improper, so propriety of the posterior distribution must be checked; with Σ restricted as in (a) and $n \geq 3$ distinct points, $(1 + b^2)^{-3/2}$ is in fact a *proper* prior for b after normalization, since $\int_{-\infty}^{\infty} (1 + b^2)^{-3/2}\, db = 2$, so only a remains improper and the posterior distribution is proper whenever the likelihood is integrable in a, which it is (the likelihood decays like a normal density in a for fixed b). For the remaining parameters take $p(\mu) \propto 1$, $p(\log\tau) \propto 1$, and $p(\log\sigma) \propto 1$ under $\Sigma = \sigma^2 I$.

Problem 14.12. Straight-line fitting with variation in x and y: you will use the model developed in the previous exercise to analyze the data on body mass and metabolic rate of dogs in Table 14.3, assuming an approximate linear relation on the logarithmic scale. In this case, the errors in Σ are presumably caused primarily by failures in the model and variation among dogs rather than 'measurement error.'

Table 14.3 (data from the earliest study of metabolic rate and body surface area, measured on a set of dogs; from Schmidt-Nielsen, 1984, p. 78):

Body mass (kg)	Body surface (cm^2)	Metabolic rate (kcal/day)
31.2	10750	1113
24.0	8805	982
19.8	7500	908
18.2	7662	842
9.6	5286	626
6.5	3724	430
3.2	2423	281

(a) Assume that log body mass and log metabolic rate have independent 'errors' of equal variance, σ^2. Assuming a noninformative prior distribution, compute posterior simulations of the parameters.
(b) Summarize the posterior inference for b and explain the meaning of the result on the original, untransformed scale.
(c) How does your inference for b change if you assume a variance ratio of 2?

Solution. (a) b has posterior median 0.614 with 95% central interval $[0.542, 0.686]$.

Set $x_i = \log(\text{mass}_i)$ and $y_i = \log(\text{metabolic rate}_i)$, $n = 7$. By 14.11(a) with $\Sigma = \sigma^2 I$, the likelihood is

$$p(x, y \mid a, b, \mu, \tau, \sigma) = \prod_{i=1}^{n} \mathrm{N}\left(\begin{pmatrix} x_i \\ y_i \end{pmatrix} \;\middle|\; \begin{pmatrix} \mu \\ a + b\mu \end{pmatrix},\ \sigma^2 I + \tau^2 \begin{pmatrix} 1 & b \\ b & b^2 \end{pmatrix} \right),$$

so the sufficient statistics are the sample mean vector and the sample covariance matrix of $z_i = (x_i, y_i)^T$, which are

$$\bar{z} = \begin{pmatrix} 2.5432 \\ 6.5133 \end{pmatrix}, \qquad S = \begin{pmatrix} 0.56667 & 0.34721 \\ 0.34721 & 0.21492 \end{pmatrix},$$

using the divisor n. The prior distribution is the noninformative one of 14.11(b), $p(a, b, \mu, \log\sigma, \log\tau) \propto (1 + b^2)^{-3/2}$; with Σ restricted to $\sigma^2 I$ the five parameters are exactly identified by the five observable moments.

Four random-walk Metropolis chains of 600,000 iterations each (acceptance rate 0.11, second two-thirds retained and thinned by 15, $\widehat{R} = 1.000$ for every parameter) give

parameter	median	95% interval
a	4.952	[4.761, 5.142]
b	0.614	[0.542, 0.686]
μ	2.538	[1.785, 3.279]
σ	0.0506	[0.029, 0.117]
τ	0.858	[0.517, 1.793]

The estimate sits between the two ordinary least-squares slopes that bracket it, $S_{xy}/S_{xx} = 0.6127$ (regressing y on x) and $S_{yy}/S_{xy} = 0.6190$ (regressing x on y), and agrees with the orthogonal (Deming) estimate

$$\hat{b}_{\perp} = \frac{S_{yy} - S_{xx} + \sqrt{(S_{yy} - S_{xx})^2 + 4S_{xy}^2}}{2S_{xy}} = 0.6144.$$

(b) b is the allometric exponent: on the original scale the fitted relation is

$$\text{metabolic rate} \;=\; e^{a}\,(\text{body mass})^{b} \text{ kcal/day},$$

with e^a having posterior median 142 kcal/day and 95% interval $[117, 171]$ — the metabolic rate of a hypothetical 1 kg dog — and b as above. Doubling a dog's body mass multiplies its metabolic

rate by 2^b, with posterior median 1.53 and 95% interval $[1.46, 1.61]$: metabolic rate grows appreciably more slowly than mass, so larger dogs burn fewer kcal per kilogram. The posterior distribution puts $\Pr(b < 1 \mid y) > 0.9999$, $\Pr(b < 3/4 \mid y) = 0.997$ and $\Pr(b < 2/3 \mid y) = 0.94$: sublinearity is unambiguous, these seven dogs are inconsistent with Kleiber's 3/4-power law, and a 2/3 (surface-area) exponent is only marginally tenable.

(c) Essentially not at all: with $\Sigma = \sigma^2\,\mathrm{diag}(1, 2)$ the posterior median of b is 0.613 with 95% interval $[0.542, 0.686]$, against 0.614 and $[0.542, 0.686]$ in (a); with ratio 1/2 it is 0.615, $[0.547, 0.691]$.

ratio σ_y^2/σ_x^2	median b	95% interval	median σ
1/2	0.615	[0.547, 0.691]	0.0632
1	0.614	[0.542, 0.686]	0.0506
2	0.613	[0.542, 0.686]	0.0385

The reason is that the errors are tiny compared with the spread of the underlying u_i: $\hat\sigma \approx 0.05$ against $\hat\tau \approx 0.86$, so $\sigma^2/\tau^2 \approx 0.0035$. The variance ratio enters the slope only through the attenuation correction, which is of this order — visible in the Deming estimates $\hat b_\perp = 0.6154, 0.6144, 0.6137$ at ratios $1/2, 1, 2$ against the naive least-squares 0.6127. All lie well inside a posterior standard deviation ($\mathrm{sd}(b \mid y) = 0.037$). What the ratio does change is the *allocation* of the residual variation between the two coordinates, hence the posterior for σ (which moves inversely, as it must to keep $\sigma^2(1+r)$ roughly fixed), but not the estimated direction of the line.

Problem 14.13. Straight-line fitting with variation in x_1, x_2, and y: adapt the model used in the previous exercise to the problem of estimating an underlying linear relation with two predictors. Estimate the relation of log metabolic rate to log body mass and log body surface area using the data in Table 14.3 (reproduced in exercise 14.12: body masses $31.2, 24.0, 19.8, 18.2, 9.6, 6.5, 3.2$ kg, body surfaces $10750, 8805, 7500, 7662, 5286, 3724, 2423$ cm^2, and metabolic rates $1113, 982, 908, 842, 626, 430, 281$ kcal/day for the same seven dogs). How does the near-collinearity of the two predictor variables affect your inference?

Solution. The two coefficients are individually almost uninformed — b_1 has 95% posterior interval $[-10.2, 9.4]$ and b_2 has $[-13.6, 16.8]$ — while the single combination $b_1 + 0.6455\,b_2$ is pinned down to $[0.385, 0.829]$, with median 0.613, the very value exercise 14.12 obtained from log body mass alone. Near-collinearity has converted two estimable parameters into one.

The model. Let $u_i = (u_{i1}, u_{i2})^T$ be the underlying log body mass and log body surface of dog i and $v_i = a + b_1 u_{i1} + b_2 u_{i2}$, with

$$u_i \sim \mathrm{N}_2(\mu, T), \qquad \begin{pmatrix} x_{i1} \\ x_{i2} \\ y_i \end{pmatrix} \sim \mathrm{N}_3\left(\begin{pmatrix} u_{i1} \\ u_{i2} \\ v_i \end{pmatrix},\ \Sigma \right), \quad \Sigma = \sigma^2 I_3.$$

Exactly as in 14.11(a), $(u_{i1}, u_{i2}, v_i)^T = \alpha + A u_i$ with $\alpha = (0, 0, a)^T$ and

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ b_1 & b_2 \end{pmatrix},$$

so marginalizing u_i gives independent trivariate normals

$$z_i \equiv \begin{pmatrix} x_{i1} \\ x_{i2} \\ y_i \end{pmatrix} \sim \mathrm{N}_3(m, V), \qquad m = \begin{pmatrix} \mu_1 \\ \mu_2 \\ a + b^T\mu \end{pmatrix}, \qquad V = ATA^T + \sigma^2 I_3.$$

The rotation-invariant prior of 14.11(b) generalizes to a plane in $\mathbb{R}^3$ as $p(a, b) \propto (1+|b|^2)^{-2}$: parametrizing the plane $v = a + b^T u$ by its unit normal $n = (-b, 1)^T/\sqrt{1+|b|^2}$ and perpendicular offset $c = a/\sqrt{1+|b|^2}$, the gnomonic projection gives $dn = (1+|b|^2)^{-3/2} db$ and $dc = (1+|b|^2)^{-1/2} da$, whence $dn\,dc = (1+|b|^2)^{-2}\,da\,db$. (For p predictors the exponent is $-(p+2)/2$, recovering $-3/2$ at $p = 1$.)

The transformed data. With $x_1 = \log(\text{mass})$, $x_2 = \log(\text{surface})$, $y = \log(\text{metabolic rate})$, $n = 7$: $\text{corr}(x_1, x_2) = 0.99725$, and

$$\bar{z} = \begin{pmatrix} 2.5432 \\ 8.6887 \\ 6.5133 \end{pmatrix}, \quad S = \begin{pmatrix} 0.56667 & 0.36576 & 0.34721 \\ 0.36576 & 0.23738 & 0.22497 \\ 0.34721 & 0.22497 & 0.21492 \end{pmatrix}.$$

Computation. The model is exactly saturated — 9 structural parameters $(a, b_1, b_2, \mu_1, \mu_2, T_{11}, T_{12}, T_{22}, \sigma^2)$ against the $3 + 6 = 9$ free parameters of $\text{N}_3(m, V)$ — and the map is a bijection that can be inverted in closed form. Since the eigenvalues of V are those of the rank-2 matrix ATA^T shifted by σ^2,

$$\sigma^2 = \lambda_{\min}(V), \quad W \equiv V - \sigma^2 I_3, \quad T = W_{[1:2,\,1:2]}, \quad b = T^{-1} W_{[1:2,\,3]},$$

$$\mu = m_{[1:2]}, \qquad a = m_3 - b^T \mu,$$

the required consistency $W_{33} = b^T T b$ being automatic because $\text{rank}(W) = 2$ forces the Schur complement to vanish. So rather than run Markov chain Monte Carlo in the structural parameters — where the near-singularity of T makes the geometry vicious; a random-walk Metropolis attempt gave $\widehat{R} > 2$ — we place the standard noninformative prior distribution on the *observable* parameters, $p(m, V) \propto |V|^{-2}$, draw

$$V \mid z \sim \text{Inv-Wishart}\Big(n - 1,\ \textstyle\sum_i (z_i - \bar{z})(z_i - \bar{z})^T\Big), \qquad m \mid V, z \sim \text{N}_3(\bar{z},\ V/n)$$

exactly (Section 3.6), and transform each of 40,000 independent draws. This is a Jeffreys prior for the saturated model, and it is the honest choice here: importance reweighting these draws to the structural prior $(1 + |b|^2)^{-2} |T|^{-3/2} \sigma^{-2}$ gave an effective sample size of only 9, so the structural prior would require a purpose-built sampler, and only the extreme tails of b_1, b_2 — already reported as essentially unbounded — would be affected.

Results. Posterior medians with 50% and 95% central intervals:

parameter	median	50% interval	95% interval
b_1 (log mass)	-0.287	$[-1.005,\ 0.278]$	$[-10.22,\ 9.40]$
b_2 (log surf)	1.393	$[0.522,\ 2.500]$	$[-13.59,\ 16.78]$
a	-4.86		$[-112.8,\ 100.8]$
σ	0.0285		$[0.018,\ 0.048]$
μ_1	2.544		$[1.49,\ 3.58]$
μ_2	8.689		$[8.01,\ 9.36]$

with $\text{corr}(b_1, b_2 \mid z) = -0.9998$, $\Pr(b_1 > 0 \mid z) = 0.35$ and $\Pr(b_2 > 0 \mid z) = 0.82$. For comparison, ordinary least squares of y on $(1, x_1, x_2)$ gives $(\hat{a}, \hat{b}_1, \hat{b}_2) = (0.290, 0.185, 0.662)$ with standard errors $(3.91, 0.36, 0.55)$ and $\text{corr}(\hat{b}_1, \hat{b}_2) = -0.9972$; the errors-in-variables treatment inflates the uncertainty further, because it must also apportion variation between Σ and T.

The effect of near-collinearity. Three consequences, all visible above.

(i) *The likelihood has a ridge, not a peak, in* (b_1, b_2). Because $x_2 \approx \gamma_0 + \gamma_1 x_1$ with $\gamma_1 = S_{x_1 x_2}/S_{x_1 x_1} = 0.6455$, the fitted surface depends on b essentially only through $b_1 + \gamma_1 b_2$; moving along $(\delta, -\delta/\gamma_1)$ changes the fit by almost nothing. The posterior correlation -0.9998 and the posterior distribution of T — whose implied correlation between the underlying u_{i1} and u_{i2} has 95% interval $[0.987, 1.000]$ — are the same fact seen twice: the seven dogs do not contain a single dog whose surface area is unusual for its mass.

(ii) *Individual coefficients are not interpretable, but the combination is.* The combination along the ridge has posterior distribution

$$b_1 + 0.6455\, b_2 \mid z: \quad \text{median } 0.613, \quad 50\%\ [0.590,\ 0.635], \quad 95\%\ [0.385,\ 0.829],$$

a 50% interval of width 0.046, against width 1.28 for b_1 alone — a factor of 28. Its median 0.613 reproduces the single-predictor answer $b = 0.614$ of exercise 14.12 to three decimals, as it must: adding a second predictor that is a near-affine function of the first adds no information, it only redistributes the same total slope between two labels. The signs of the medians ($b_1 < 0 < b_2$) are an artifact of this

redistribution and should not be read as "metabolic rate decreases with body mass at fixed surface area."
(iii) *Reporting.* Hence the summary must be either the one-predictor analysis of 14.12 or the *joint* posterior distribution of (b_1, b_2) — never two marginal intervals, which conceal that the pair is determined only to within one dimension.

Problem 14.14. Heaped data and regression: As part of a public health study, the ages and heights were recorded for several hundred young children in Africa. The goal was to see how many children were too short, given their age. File `reported_ages.png` shows a histogram of the reported ages (in months). The spikes at 12 months, etc., suggest that some (but not all) of the ages are rounded (it is not obvious whether they are being rounded up or down). You can assume that some ages are reported exactly, some are rounded to the nearest 6 months, and some are rounded to the nearest 12 months.
(a) Set up a model for these data. You can use the notation a_i for the reported ages and h_i for the heights (which are measured, essentially exactly).
(b) Write the likelihood for your model.
(c) Describe how you would estimate the parameters in your model.

Solution. (a) Treat the true age A_i as a latent continuous variable and the coarseness of each child's reporting as a latent discrete variable; the reported age a_i is then a deterministic, known function of the two.
True ages. $A_i \sim p(A \mid \psi)$, independently, for a smooth density on the age range of the study. Take a flexible form — a mixture of a few normals, or a piecewise-constant density on half-year bins with a random-walk smoothing prior on the log bin heights — flexible enough to fit the *envelope* of the histogram but incapable of reproducing spikes at isolated integers. Smoothness of $p(A \mid \psi)$ is the assumption that does all the work in this problem, as noted at the end of (b).
Coarseness. $G_i \in \{1, 6, 12\}$ with $\Pr(G_i = g \mid \pi) = \pi_g$, $\pi_1 + \pi_6 + \pi_{12} = 1$, independent of A_i and of h_i given A_i.
Rounding. Let $c \in [0, 1]$ be an unknown rounding offset, common to all children: given $G_i = g$, the report a_i is the unique multiple of g with

$$A_i \in R_g(a_i) \equiv \big[\, a_i - cg,\ a_i + (1-c)g \,\big).$$

Thus $c = \frac{1}{2}$ is round-to-nearest, $c = 0$ is round-down (age last attained), and $c = 1$ is round-up; making c a parameter is how the model answers the question the exercise says is not obvious from the picture. Note $R_g(a) = \emptyset$ unless a is a multiple of g, so a reported age of 13 months can only have come from $G_i = 1$, an age of 18 from $G_i \in \{1, 6\}$, and an age of 24 from any of the three.
Heights. $h_i \mid A_i \sim \mathrm{N}\big(f(A_i),\, \sigma^2(A_i)\big)$, where f is a growth curve — children this young grow fast and nonlinearly, so use a cubic spline in A (or in $\sqrt{A}$) with a smoothness penalty rather than a straight line, and let $\log \sigma(A)$ be linear or spline in A since height variability widens with age. Collect the height parameters as θ.
(b) Marginalizing the two latent variables, and writing $\nmid\{g \mid a_i\}$ for the indicator that g divides a_i,

$$p(a_i, h_i \mid \pi, c, \theta, \psi) = \sum_{g \in \{1,6,12\}} \pi_g \nmid\{g \mid a_i\} \int_{a_i - cg}^{a_i + (1-c)g} \mathrm{N}\big(h_i \mid f(A), \sigma^2(A)\big)\, p(A \mid \psi)\, dA,$$

and the full likelihood is the product over the independent children,

$$p(a, h \mid \pi, c, \theta, \psi) = \prod_{i=1}^{n} \sum_{g \in \{1,6,12\}} \pi_g \nmid\{g \mid a_i\} \int_{R_g(a_i)} \mathrm{N}\big(h_i \mid f(A), \sigma^2(A)\big)\, p(A \mid \psi)\, dA.$$

This is a proper likelihood: summing over the admissible values of a_i and integrating over h_i gives $\sum_g \pi_g \int p(A \mid \psi)\, dA = 1$, because for each g the intervals $R_g(a)$ over the multiples a of g partition the line.

Two features of the likelihood deserve notice. First, it is a finite mixture whose mixing weights are *data dependent*: a child reported at 13 months contributes the single term $\pi_1 \int_{13-c}^{14-c}(\cdot)$, whereas a child reported at 24 contributes all three terms, with the $g = 12$ term integrating over a whole year of possible true ages. Second, the information about the parameters comes from two separate places. The reported-age margin identifies π: for a a multiple of 12, the expected count is

$$n\Big[\pi_1 P_\psi\big(R_1(a)\big) + \pi_6 P_\psi\big(R_6(a)\big) + \pi_{12} P_\psi\big(R_{12}(a)\big)\Big],$$

which exceeds the smooth background $n\,\pi_1 P_\psi(R_1(a))$ by the spike the histogram displays; matching the excess at multiples of 6 that are not multiples of 12 against the excess at multiples of 12 separates π_6 from π_{12}. The heights identify c: if $c \neq \frac{1}{2}$, the children reported at a spike value a are systematically older (if c is small) or younger (if c is large) than a, and so systematically taller or shorter than $f(a)$; comparing the heights of children at the spikes with those of children at nearby non-multiples of 6 estimates the shift. Without the smoothness restriction on $p(A \mid \psi)$ nothing is identified — $\pi_1 = 1$ with a true age distribution having atoms at the multiples of 12 fits the data exactly.

(c) Markov chain Monte Carlo, marginalizing G_i analytically and A_i numerically.

Method (1): direct evaluation of the marginal likelihood. Each integral is over an interval of length at most 12 months, so a fixed Gauss–Legendre rule with 16–32 nodes per interval evaluates it to far better than Monte Carlo accuracy. This turns the expression in (b) into a smooth, differentiable function of (π, c, θ, ψ) and lets Hamiltonian Monte Carlo (Stan) sample the joint posterior distribution directly, with no discrete or latent-continuous parameters at all. Priors: $\pi \sim \mathrm{Dirichlet}(1,1,1)$; $c \sim \mathrm{U}(0,1)$; weakly informative normal priors on the spline coefficients of f and $\log\sigma$ with half-normal priors on their smoothing scales; the same for the log bin heights of $p(A \mid \psi)$. This is the method to use.

Method (2): data augmentation. Keep (A_i, G_i) as parameters and run a Gibbs sampler: $G_i \mid A_i, a_i, \pi$ is a three-category draw with weights $\pi_g \mathbb{1}\{A_i \in R_g(a_i)\}$; $A_i \mid G_i, a_i, h_i$ has density proportional to $\mathrm{N}(h_i \mid f(A), \sigma^2(A))\, p(A \mid \psi)$ truncated to the known interval $R_{G_i}(a_i)$, drawn by slice sampling or a grid approximation on that interval; $\pi \mid G$ is Dirichlet conjugate; θ is updated as in an ordinary spline regression with the current A_i as predictors; c is updated by Metropolis (note c moves the truncation limits, so it mixes slowly here — the reason to prefer Method (1)).

The quantity the study actually wants is not a parameter but the count

$$T = \sum_{i=1}^{n} \mathbb{1}\big\{h_i < f_{\mathrm{ref}}(A_i) - 2\,s_{\mathrm{ref}}(A_i)\big\}$$

for a reference growth standard $(f_{\mathrm{ref}}, s_{\mathrm{ref}})$; since T depends on the unobserved A_i, report its posterior distribution by averaging over the sampler's draws of A_i (Method (2) supplies these directly; under Method (1) draw A_i from its conditional, the truncated density above, at each retained iteration).

15 Hierarchical Linear Models

Contents

15.1 Exercises 15.1–15.6

Problem 15.1. Varying-coefficients models: express the educational testing example of Section 5.5 as a hierarchical linear model with eight observations and known observation variances. Draw simulations from the posterior distribution using the methods described in this chapter.
The data of Table 5.2 are the estimated coaching effects y_j and their standard errors σ_j from eight separate randomized experiments on SAT-V coaching:

School	Estimated effect y_j	Standard error σ_j
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Solution. The model is (15.1) with $J = 8$ and design matrix $X = I_8$: writing $\beta = (\beta_1, \dots, \beta_8)$ for the eight school effects,

$$y \mid \beta \sim \mathrm{N}(X\beta,\, \Sigma_y), \qquad \Sigma_y = \mathrm{diag}(\sigma_1^2, \dots, \sigma_8^2),$$

with Σ_y known, and the batch of coefficients exchangeable,

$$\beta \mid \alpha, \sigma_\beta \sim \mathrm{N}(\mathbf{1}\alpha,\, \sigma_\beta^2 I_8), \qquad p(\alpha, \sigma_\beta) \propto 1.$$

This is exactly the (θ, μ, τ) model of Section 5.5 relabelled $(\beta, \alpha, \sigma_\beta)$, as noted on p. 382: $n = 8$ "observations," one per school, with no predictors other than the eight indicators.
By Section 15.3 the population distribution is extra data. Stacking $y_* = (y_1, \dots, y_8, 0, \dots, 0)^T \in \mathbb{R}^{16}$ against the 16×9 design matrix and 16×16 variance matrix

$$X_* = \begin{pmatrix} I_8 & 0 \\ I_8 & -\mathbf{1} \end{pmatrix}, \qquad \Sigma_* = \begin{pmatrix} \Sigma_y & 0 \\ 0 & \sigma_\beta^2 I_8 \end{pmatrix},$$

acting on the parameter vector (β, α), the model becomes a single weighted linear regression $y_* \sim \mathrm{N}(X_*(\beta, \alpha),\, \Sigma_*)$. The all-at-once Gibbs sampler of Section 15.5 alternates

$$(\beta, \alpha) \mid \sigma_\beta, y \sim \mathrm{N}\big(\hat{V}_\beta X_*^T \Sigma_*^{-1} y_*,\ \hat{V}_\beta\big), \quad \hat{V}_\beta = (X_*^T \Sigma_*^{-1} X_*)^{-1},$$
$$\sigma_\beta^2 \mid \beta, \alpha, y \sim \text{Inv-}\chi^2\big(J - 1,\ \tfrac{1}{J-1}\textstyle\sum_j (\beta_j - \alpha)^2\big).$$

The degrees of freedom are $J - 1 = 7$, not J, because the uniform prior on σ_β contributes one factor of σ_β (Section 5.4).
Because σ_β here is estimated near zero, I ran the parameter-expanded version (15.9): $\beta_j = \alpha + \zeta\gamma_j$, $\gamma_j \sim \mathrm{N}(0, \tau^2)$, $\sigma_\beta = |\zeta|\tau$, updating γ, then (α, ζ) by weighted regression of y on $(\mathbf{1}, \gamma)$, then $\tau^2 \sim \text{Inv-}\chi^2(7, \sum_j \gamma_j^2/7)$. Four sequences of 20{,}000 draws, second halves retained: $\hat{R} < 1.01$ for every β_j, α and σ_β. Posterior quantiles (which reproduce Table 5.3):

Parameter	2.5%	25%	median	75%	97.5%
β_1	-2.1	6.0	10.2	15.5	31.6
β_2	-4.6	4.0	7.8	11.8	20.6
β_3	-11.5	2.0	6.7	10.9	20.5
β_4	-5.6	3.7	7.7	11.7	20.9
β_5	-8.8	1.4	5.6	9.4	16.4
β_6	-8.5	2.3	6.5	10.4	18.6
β_7	-1.1	6.1	10.1	14.6	26.1
β_8	-7.1	3.9	8.2	12.7	25.5
α	-2.1	4.7	7.9	11.2	18.2
σ_β	0.2	2.5	5.2	9.1	20.5

School A's effect is shrunk from 28 to a posterior median of 10.2, with $\Pr(\beta_1 > \beta_3 \mid y) = 0.68$ and $\Pr(\beta_1 \text{ largest} \mid y) = 0.25$.

Problem 15.2. Fitting a hierarchical model for a two-way array: Table 8.6 displays the yields of penicillin produced by four manufacturing processes (treatments), each applied in five different conditions (blocks); four runs were made within each block, with treatments assigned to runs at random.

Block	A	B	C	D
1	89	88	97	94
2	84	77	92	79
3	81	87	87	85
4	87	92	89	84
5	79	81	80	88

(a) Fit a standard analysis of variance model to the randomized block data discussed in Exercise 8.5, that is, a linear regression with a constant term, indicators for all but one of the blocks, and all but one of the treatments.
(b) Summarize posterior inference for the (superpopulation) average penicillin yields, averaging over the block conditions, under each of the four treatments. Under this measure, what is the probability that each of the treatments is best? Give a 95% posterior interval for the difference in yield between the best and worst treatments.
(c) Set up a hierarchical extension of the model, in which you have indicators for all five blocks and all five treatments, and the block and treatment indicators are two sets of varying coefficients. Explain why the means for the block and treatment indicator groups should be fixed at zero. Write the joint distribution of all model parameters (including the hierarchical parameters).
(d) Compute the posterior mode of the three variance components of your model in (c) using EM. Construct a normal approximation about the mode and use this to obtain posterior inferences for all parameters and answer the questions in (b). (Hint: you can use the general regression framework or extend the procedure in Section 13.6.)
(e) Check the fit of your model to the data. Discuss the relevance of the randomized block design to your check; how would the posterior predictive simulations change if you were told that the treatments had been assigned by complete randomization?
(f) Obtain draws from the actual posterior distribution using the Gibbs sampler, using your results from (d) to obtain starting points. Run multiple sequences and monitor the convergence of the simulations by computing $\hat{R}$ for all parameters in the model.
(g) Discuss how your inferences in (b), (d), and (e) differ.

Solution. (The printed part (c) says "all five treatments"; Table 8.6 has four, and I solve the intended statement with four treatment indicators.)
(a) Least squares on the 20×8 design gives $\hat{\beta} = (90, -9, -7, -4, -10, 1, 5, 2)$ with residual $s = 4.34$ on $n - k = 12$ degrees of freedom; the entries are the intercept (block 1, treatment A), the block 2–5 offsets, and the B, C, D offsets. With $p(\beta, \log\sigma) \propto 1$, Section 14.2 gives

$$\sigma^2 \mid y \sim \text{Inv-}\chi^2(12,\ 18.83), \qquad \beta \mid \sigma, y \sim \text{N}\Big(\hat{\beta},\, V_\beta \sigma^2\Big),$$

so $\beta \mid y$ is multivariate t_{12} with standard errors 2.75 for the intercept and the treatment offsets and 3.07 for the block offsets.
(b) Averaging the fitted surface over the five blocks, the estimand for treatment t is $\mu_t = \beta_0 + \frac{1}{5}\sum_{b=1}^{5}\beta_b^{\text{block}} + \beta_t^{\text{treat}}$, and the treatment contrasts $\mu_t - \mu_{t'}$ are exactly the treatment coefficients. From 200,000 draws:

Treatment	Posterior median μ_t	sd	Pr(best)
A	84.0	2.13	0.026
B	85.0	2.13	0.059
C	89.0	2.13	0.791
D	86.0	2.13	0.125

C is best and A worst; $\mu_C - \mu_A$ has posterior median 5.0 and 95% interval $[-1.0,\ 11.0]$, that is, $5.0 \pm t_{12,0.975}\sqrt{2/5}\,s$. The interval covers zero, so even the largest observed treatment gap is not established.

(c) Following Section 15.6, index the cells by block $b = 1, \ldots, 5$ and treatment $t = 1, \ldots, 4$ and write

$$y_{bt} \mid \mu, \beta, \tau, \sigma \sim \mathrm{N}(\mu + \beta_b + \tau_t,\ \sigma^2),$$

with the two batches of varying coefficients

$$\beta_b \sim \mathrm{N}(0, \sigma_\beta^2), \qquad \tau_t \sim \mathrm{N}(0, \sigma_\tau^2).$$

The batch means must be fixed at 0 because they are not identified: replacing (μ, β, τ) by $(\mu + c + d,\ \beta - c,\ \tau - d)$ leaves every cell mean unchanged, so only $\mu + \mathrm{E}(\beta) + \mathrm{E}(\tau)$ is estimable. Freeing the two batch means would add a two-dimensional flat ridge to the posterior, and with a uniform prior on μ the posterior would be improper. The constant term already carries the overall level; this is exactly the centring argued on p. 396. The joint distribution, with uniform priors on μ and on the three standard deviations, is

$$\begin{aligned} p(\mu, \beta, \tau, \sigma, \sigma_\beta, \sigma_\tau \mid y) & \\ \propto \prod_{b=1}^{5} \prod_{t=1}^{4} \mathrm{N}(y_{bt} \mid \mu + \beta_b + \tau_t,\ \sigma^2) & \\ \times \prod_{b=1}^{5} \mathrm{N}(\beta_b \mid 0, \sigma_\beta^2) \prod_{t=1}^{4} \mathrm{N}(\tau_t \mid 0, \sigma_\tau^2). & \end{aligned}$$

(d) EM with (μ, β, τ) as missing data. The complete-data mode is available in closed form, so the M-step is

$$\sigma^2 = \tfrac{1}{20}\,\mathrm{E}\Big[\sum_{b,t}(y_{bt} - \mu - \beta_b - \tau_t)^2\Big],$$

$$\sigma_\beta^2 = \tfrac{1}{5}\,\mathrm{E}\Big[\sum_{b}\beta_b^2\Big], \qquad \sigma_\tau^2 = \tfrac{1}{4}\,\mathrm{E}\Big[\sum_{t}\tau_t^2\Big],$$

the divisors being n, J_β, J_τ because with $p(\sigma, \sigma_\beta, \sigma_\tau) \propto 1$ the M-step maximizes on the standard-deviation scale: $\sigma^{-n}\exp(-\mathrm{SS}/2\sigma^2)$ peaks at $\sigma^2 = \mathrm{SS}/n$. The E-step is one multivariate normal calculation: $(\mu, \beta, \tau) \mid y, \sigma, \sigma_\beta, \sigma_\tau$ is normal with precision $W^TW/\sigma^2 + \mathrm{diag}(0, \sigma_\beta^{-2}I_5, \sigma_\tau^{-2}I_4)$, and each expectation is a squared mean plus a trace of the relevant block of the variance matrix. EM converges to

$$(\sigma,\ \sigma_\beta,\ \sigma_\tau) = (4.34,\ 3.43,\ 0.95),$$

which I confirmed by direct maximization of the marginal posterior density. The balanced design makes these coincide exactly with the classical moment estimates $\sqrt{(66 - 18.83)/4} = 3.43$ and $\sqrt{(23.33 - 18.83)/5} = 0.95$ from the block and treatment mean squares.

The posterior is far from normal in these coordinates, so I approximate on the log scale. The mode of $p(\log\sigma, \log\sigma_\beta, \log\sigma_\tau \mid y)$ is at $(4.37, 4.38, 2.50)$ with approximate log-scale standard deviations $(0.20, 0.51, 0.74)$ from the numerical second-derivative matrix; drawing variance components from this normal approximation and then the ten coefficients from their exact conditional normal gives

Treatment	Median $\mu + \tau_t$	sd	Pr(best)
A	84.8	3.07	0.056
B	85.4	3.02	0.098
C	87.7	3.14	0.681
D	86.0	3.01	0.164

with $\mu_C - \mu_A$ having median 2.7 and 95% interval $[-1.4,\ 8.3]$, and 95% intervals $[2.97, 6.43]$ for σ, $[1.63, 11.6]$ for σ_β, $[0.59, 10.5]$ for σ_τ.

(e) The model fits. Replicating $y_{bt}^{\mathrm{rep}} \sim \mathrm{N}(\mu + \beta_b + \tau_t, \sigma^2)$ at each posterior draw of (f),

$$T_1 = \max_{b,t} |y_{bt} - \mu - \beta_b - \tau_t|/\sigma, \qquad p = 0.59,$$

$$T_2 = \text{Tukey one-df nonadditivity } F, \qquad p = 0.76,$$

$$T_3 = \text{Kendall } W \text{ of treatment ranks across blocks}, \qquad p = 0.58.$$

T_2 and T_3 are the checks with teeth here, since the additive model asserts no block-by-treatment interaction and there is no within-cell replication to estimate one.

The randomized block design matters because it is what makes the design ignorable conditional on the block and treatment indicators (Exercise 8.5(b)): the replications above condition on the realized 4×5 layout, which is legitimate precisely because $p(I \mid x)$ does not depend on y or on the parameters. Under complete randomization the modeling and the posterior would be unchanged (Exercise 8.6(a)) — the same x is still being conditioned on — but a complete check would also draw I^{rep}, and the balance of the observed layout (each treatment appearing exactly once in each block) would itself become a test quantity with a small predictive probability under complete randomization. Under randomized blocks that balance is forced by the design and carries no information about the model.

(f) Starting the four sequences from the normal approximation of (d), the Gibbs sampler alternates the ten coefficients (one weighted regression, as in Section 15.5) with

$$\begin{aligned}\sigma^2 &\sim \text{Inv-}\chi^2\big(19,\ \tfrac{1}{19}\widehat{\text{SS}}\big),\\ \sigma_\beta^2 &\sim \text{Inv-}\chi^2\big(4,\ \tfrac{1}{4}\textstyle\sum_b \beta_b^2\big),\\ \sigma_\tau^2 &\sim \text{Inv-}\chi^2\big(3,\ \tfrac{1}{3}\textstyle\sum_t \tau_t^2\big),\end{aligned}$$

the degrees of freedom being $n-1$, $J_\beta - 1$, $J_\tau - 1$ under the uniform priors on the standard deviations. Four sequences of 40,000, second halves retained: $\hat{R} \le 1.00$ for all thirteen parameters.

Parameter	2.5%	median	97.5%
μ	77.5	86.0	94.3
σ	3.23	4.66	7.31
σ_β	0.57	4.32	16.4
σ_τ	0.13	2.21	15.2

The superpopulation treatment means $\mu + \tau_t$ have medians 85.0, 85.5, 87.4, 86.0 with posterior sd 3.7, $\Pr(\text{best}) = 0.084, 0.123, 0.610, 0.183$, and $\mu_C - \mu_A$ has median 2.1, 95% interval $[-1.5,\ 8.5]$. Averaging over the five blocks actually used, $\mu + \bar{\beta} + \tau_t$, gives the same medians with posterior sd 1.8 and the same contrast interval.

(g) The nonhierarchical fit of (b) gives the contrasts their raw least-squares values; the hierarchical fit shrinks them by roughly 60% toward zero ($\mu_C - \mu_A$ from 5.0 to 2.1), because with only four treatments and σ_τ estimated below σ the batch of treatment effects is pooled heavily. Correspondingly $\Pr(C \text{ best})$ falls from 0.79 to 0.61. The normal approximation of (d) reproduces the Gibbs answers to within about 0.6 on the contrasts and 0.07 on the probabilities, but it understates the posterior spread of σ_β and σ_τ — with 4 and 3 degrees of freedom these have long right tails that no normal approximation on any scale will capture — and so slightly overstates $\Pr(C \text{ best})$. The model checks of (e) do not distinguish the two models: the data are consistent with additivity either way, and the difference between (b) and (f) is a difference in what is assumed about the batch of treatment effects, not in fit.

Problem 15.3. Regression with many explanatory variables: Table 15.2 displays data from a designed experiment for a chemical process. In using these data to illustrate various approaches to selection and estimation of regression coefficients, Marquardt and Snee (1975) assume a quadratic regression form; that is, a linear relation between the expectation of the untransformed outcome, y, and the variables x_1, x_2, x_3, their two-way interactions, x_1x_2, x_1x_3, x_2x_3, and their squares, x_1^2, x_2^2, x_3^2.

Reactor temperature (°C), x_1	Ratio of H2 to n-heptane (mole ratio), x_2	Contact time (sec), x_3	Conversion of n-heptane to acetylene (%), y
1300	7.5	0.0120	49.0
1300	9.0	0.0120	50.2
1300	11.0	0.0115	50.5
1300	13.5	0.0130	48.5
1300	17.0	0.0135	47.5
1300	23.0	0.0120	44.5
1200	5.3	0.0400	28.0
1200	7.5	0.0380	31.5
1200	11.0	0.0320	34.5
1200	13.5	0.0260	35.0
1200	17.0	0.0340	38.0
1200	23.0	0.0410	38.5
1100	5.3	0.0840	15.0
1100	7.5	0.0980	17.0
1100	11.0	0.0920	20.5
1100	17.0	0.0860	19.5

(a) Fit an ordinary linear regression model (that is, nonhierarchical with a uniform prior distribution on the coefficients), including a constant term and the nine explanatory variables above.
(b) Fit a mixed-effects linear regression model with a uniform prior distribution on the constant term and a shared normal prior distribution on the coefficients of the nine variables above. If you use iterative simulation in your computations, be sure to use multiple sequences and monitor their joint convergence.
(c) Discuss the differences between the inferences in (a) and (b). Interpret the differences in terms of the hierarchical variance parameter. Do you agree with Marquardt and Snee that the inferences from (a) are unacceptable?
(d) Repeat (a), but with a t_4 prior distribution on the nine variables.
(e) Discuss other models for the regression coefficients.

Solution. Throughout, x_1, x_2, x_3 are centred and scaled to unit standard deviation before the products and squares are formed, and each of the nine derived predictors is then itself centred and scaled. Without this the shared population distribution of (b) is meaningless — a batch of coefficients can only be exchangeable if the predictors are on a common scale — and the raw design matrix is in any case numerically hopeless.

(a) Least squares, $n = 16$, $k = 10$, 6 residual degrees of freedom, $s = 1.27$, $R^2 = 0.996$. With $p(\beta, \log\sigma) \propto 1$ the posterior is the t_6 of Section 14.2 centred at

Predictor	$\hat{\beta}$	se
constant	35.48	0.32
x_1	3.07	6.33
x_2	2.00	0.43
x_3	-11.54	8.52
x_1x_2	-9.28	1.82
x_1x_3	-45.86	26.48
x_2x_3	-8.67	1.95
x_1^2	-22.63	13.72
x_2^2	-1.54	0.58
x_3^2	-21.46	11.12

The design is badly collinear (x_1 and x_3 are nearly a deterministic function of each other, and the condition number of X^TX is 4.3×10^4), so individual coefficients are estimated to within tens of units on a response ranging over 35 points with $\sigma = 1.3$.

(b) The mixed-effects model of Section 15.1 with $J_1 = 1$ noninformative coefficient and $J_2 = 9$ exchangeable ones,

$$y \mid \beta, \sigma \sim \mathrm{N}(X\beta, \sigma^2 I), \quad \beta_j \sim \mathrm{N}(0, \sigma_\beta^2)\ (j = 1, \ldots, 9),$$

with $p(\beta_0, \sigma, \sigma_\beta) \propto 1$. I used the parameter-expanded Gibbs sampler (15.9), $\beta_j = \zeta\gamma_j$ with $\gamma_j \sim \mathrm{N}(0, \tau^2)$ and $\sigma_\beta = |\zeta|\tau$; four sequences of 60,000, second halves retained, $\hat{R} \le 1.00$ throughout. Posterior medians and 95% intervals:

Predictor	2.5%	median	97.5%
constant	34.5	35.48	36.5
x_1	3.79	9.24	15.22
x_2	0.40	1.69	2.92
x_3	-9.7	-2.94	4.64
x_1x_2	-11.2	-6.44	-0.77
x_1x_3	-13.7	-1.65	7.71
x_2x_3	-10.2	-5.24	0.68
x_1^2	-6.97	-0.39	5.38
x_2^2	-2.30	-0.86	0.73
x_3^2	-8.29	-1.04	4.98
σ	1.02	1.75	3.56
σ_β	2.99	5.57	12.09

(c) Every coefficient is pulled toward zero and every interval narrows, dramatically so in the collinear directions: $\hat{\beta}_{x_1x_3}$ moves from -45.9 ± 26.5 to a median of -1.7 with interval half-width 11, and $\hat{\beta}_{x_1^2}$ from -22.6 to -0.4. The exception is x_1, whose coefficient grows from 3.1 to 9.2: shrinking the quadratic and interaction terms in the x_1/x_3 direction transfers that signal to the linear term. The hierarchical variance parameter says why. The nine least-squares estimates have root mean square 19.4, but the posterior for σ_β is concentrated near 5.6: the data say the nine underlying coefficients are of typical size 5–6, so estimates near -46 are read as noise in a poorly determined direction and are pooled back. Equivalently (Section 15.3), the population distribution acts as nine extra observations $\beta_j \approx 0$ with standard error σ_β, which is exactly the ridge-type information the collinear design lacks.

Only half agreeing with Marquardt and Snee: the inferences of (a) are not incoherent, and their intervals honestly report the collinearity. What is unacceptable is treating the least-squares point estimates as estimates — they are not, since the posterior in (a) puts the coefficients anywhere in a huge region. The defect is in the uniform prior distribution, which asserts that a coefficient of -46 is as plausible a priori as one of 0 on a standardized predictor for a response with $\sigma = 1.3$; the remedy is the population distribution of (b), not the variable selection that Marquardt and Snee reach for.

(d) The same model with the population distribution replaced by $\beta_j \sim t_4(0, \sigma_\beta^2)$, implemented as the scale mixture $\beta_j \sim \mathrm{N}(0, V_j)$, $V_j \sim \text{Inv-}\chi^2(4, \sigma_\beta^2)$ (Section 14.7), with the same parameter expansion. Four sequences of 60,000, $\hat{R} \leq 1.001$; medians and 95% intervals:

Predictor	2.5%	median	97.5%
x_1	3.31	10.11	15.38
x_2	0.28	1.66	2.88
x_3	-10.2	-1.87	4.64
x_1x_2	-10.8	-5.48	-0.02
x_1x_3	-12.0	-0.76	6.80
x_2x_3	-9.8	-4.21	1.34
x_1^2	-6.24	-0.23	4.76
x_2^2	-2.26	-0.84	0.77
x_3^2	-7.28	-0.45	4.68
σ	1.04	1.83	3.70
σ_β	1.46	4.09	10.47

The long tails buy what one would expect and no more: σ_β drops from 5.6 to 4.1, the small coefficients shrink further — by up to about half, and hardly at all for $\beta_{x_2^2}$, which the data already pin down — and the one large coefficient β_{x_1} is allowed to grow from 9.2 to 10.1. With nine coefficients there is little information to distinguish t_4 from normal, and the fits are practically the same.

(e) Four directions, roughly in order of how much they would change the answer. (i) Batch the nine coefficients by order — main effects, interactions, squares — with a separate σ_m for each, which is the analysis of variance structure of Section 15.6 and encodes the usual belief that higher-order terms are smaller; with $J_m = 3$ per batch the σ_m are barely identified and a uniform prior on them is not usable, so this needs a weakly informative prior such as half-Cauchy. (ii) Keep one batch but make the prior weakly informative rather than estimated, for example $\beta_j \sim \mathrm{N}(0, 5^2)$ on the standardized

scale, which is close to what (b) estimates and avoids the funnel entirely. (iii) Prior distributions that respect functional marginality: shrink an interaction toward zero more strongly when its main effects are small. (iv) Change the likelihood rather than the prior — y here is a percentage conversion, so a logit or log transformation, or an actual reaction-kinetics model in (x_1, x_2, x_3), would make a quadratic surface in the untransformed y unnecessary; a second-order polynomial in three highly collinear inputs is a very indirect description of this experiment.

Problem 15.4. Analysis of variance: the data of Table 5.2 are the estimated coaching effects $y_j = 28, 8, -3, 7, -1, 1, 18, 12$ with standard errors $\sigma_j = 15, 10, 16, 11, 9, 11, 10, 18$ for schools A through H.
(a) Create an analysis of variance plot for the educational testing example in Chapter 5, assuming that there were exactly 60 students in the study in each school, with 30 receiving the treatment and 30 receiving the control.
(b) Discuss the relevance of the finite-population and superpopulation standard deviation for each source of variation.

Solution. (a) At the level of individual students the model is the two-way layout of Section 15.6,

$$y_i = \mu + \beta^{\text{sch}}_{j(i)} + \beta^{\text{trt}}_{k(i)} + \beta^{\text{int}}_{j(i)k(i)} + \epsilon_i, \qquad i = 1, \ldots, 480,$$

with $j = 1, \ldots, 8$, $k = 1, 2$ for treatment and control, and the usual zero-sum constraints, giving $(df) = 7, 1, 7$ and $480 - 16 = 464$ for the four rows. The Chapter 5 data are the within-school treatment-minus-control differences, so

$$\theta_j = (\beta^{\text{trt}}_1 - \beta^{\text{trt}}_2) + (\beta^{\text{int}}_{j1} - \beta^{\text{int}}_{j2}) = 2\beta^{\text{trt}}_1 + 2\beta^{\text{int}}_{j1},$$

whence $\beta^{\text{trt}}_1 = \bar{\theta}/2$ and $\beta^{\text{int}}_{j1} = (\theta_j - \bar{\theta})/2$. Substituting into (15.10),

$$s_{\text{trt}} = \sqrt{\textstyle\sum_k (\beta^{\text{trt}}_k)^2} = \frac{|\bar{\theta}|}{\sqrt{2}}, \qquad s_{\text{int}} = \sqrt{\frac{1}{7}\textstyle\sum_{j,k} (\beta^{\text{int}}_{jk})^2} = \sqrt{\frac{1}{14}\textstyle\sum_j (\theta_j - \bar{\theta})^2}.$$

Both are functions of θ alone, so the 200,000 posterior draws of Exercise 15.1 give them directly. For the error row, σ_j is the standard error of a difference of two means of 30 students, so $\sigma_j^2 = s_{y,j}^2(1/30 + 1/30)$ and the individual-level standard deviation is $s_{y,j} = \sqrt{15}\,\sigma_j$, ranging from 34.9 (school E) to 69.7 (school H), with root mean square 49.9. Since the σ_j are taken as known, this row is known too and gets no interval.

Source	(df)	s_m 50% interval	s_m 95% interval	median
treatment	1	[3.6, 7.6]	[0.6, 11.4]	5.6
school x treatment	7	[1.5, 5.4]	[0.2, 10.1]	3.2
error	464	—	—	49.9
school	7	not identified	not identified	—

The display itself is this table drawn as bars — one row per source, 50% and 95% intervals, all on a common horizontal axis in SAT-V points starting at zero, in the style of Figure 15.4. Its message is the message of Chapter 5 in one picture: on a scale where individual students vary by about 50 points, the average coaching effect is around 5 and the school-to-school variation in that effect is smaller still and could be zero. The school main effect is a genuine row of the table but is not estimable from these data, because β^{sch}_j cancels out of every within-school difference θ_j; estimating it would require the school means, which Table 5.2 does not report.
(b) For the school-by-treatment row the two are genuinely different questions and both are answerable. The finite-population s_{int} asks how much these eight coaching programs actually differed from each other, and is estimated reasonably well (median 3.2, 95% interval $[0.2, 10.1]$); the superpopulation $\sigma_{\text{int}} = \tau/\sqrt{2}$ asks how much a ninth school's program would differ from the average, and is much more uncertain (median 3.7, 95% interval $[0.2, 14.5]$), since a variance component on 7 degrees of freedom has a long right tail. Which one to report depends on the use: the finite-population quantity for describing the experiment that was run, the superpopulation quantity for predicting a new school.

For the treatment row, with one degree of freedom, only the finite-population quantity is meaningful. This is precisely the extreme case discussed on p. 397: $s_{\text{trt}} = |\bar{\theta}|/\sqrt{2}$ is a function of the single well-estimated contrast $\bar{\theta}$, whereas σ_{trt} is being estimated by something proportional to a χ^2_1 and describes a population of treatments that does not exist — there is only one treatment here. For the error row the distinction is empty in the other direction: with 464 degrees of freedom the finite-population and superpopulation standard deviations coincide to within a fraction of a point. And for the school main effect, where nothing is identified, neither is available.

Problem 15.5. Modeling correlation matrices:
(a) Show that the determinant of a correlation matrix R is a quadratic function of any of its elements. (This fact can be used in setting up a Gibbs sampler for multivariate models.)
(b) Suppose that the off-diagonal elements of a 3×3 correlation matrix are 0.4, 0.8, and r. Determine the range of possible values of r.
(c) Suppose all the off-diagonal elements of a d-dimensional correlation matrix R are equal to the same value, r. Prove that R is positive definite if and only if $-1/(d-1) < r < 1$.

Solution. (a) Fix $a \neq b$ and let $r = R_{ab} = R_{ba}$; in the Leibniz expansion

$$\det R = \sum_{\pi} \operatorname{sgn}(\pi) \prod_{i=1}^{d} R_{i,\pi(i)},$$

a permutation π supplies the factor R_{ab} only if $\pi(a) = b$ and the factor R_{ba} only if $\pi(b) = a$, so no product contains r more than twice and $\det R = Ar^2 + Br + C$.
The degree-2 terms are those with $\pi(a) = b$ and $\pi(b) = a$, that is, $\pi = (a\,b)\pi'$ with π' a permutation of the remaining $d-2$ indices and $\operatorname{sgn}(\pi) = -\operatorname{sgn}(\pi')$; hence

$$A = -\det R_{-ab},$$

where R_{-ab} deletes rows and columns a and b. Order the indices so that a, b are last. Then the first $d-1$ leading principal minors are free of r, so if they are positive (which is necessary for R to be positive definite at any r) the admissible set is $\{r : \det R > 0\}$; and R_{-ab} is then positive definite, so $A < 0$ and that set is the open interval between the two roots. That is what makes the Gibbs step feasible: r is drawn from a univariate distribution truncated to an interval computed from three determinant evaluations.
(b) Expanding along the first row,

$$\begin{aligned}\det R &= \det \begin{pmatrix} 1 & 0.4 & 0.8 \\ 0.4 & 1 & r \\ 0.8 & r & 1 \end{pmatrix} \\ &= (1 - r^2) - 0.4(0.4 - 0.8r) + 0.8(0.4r - 0.8) \\ &= -r^2 + \tfrac{16}{25}r + \tfrac{1}{5},\end{aligned}$$

confirming (a). The other leading principal minors are 1 and $1 - 0.4^2 = 0.84$, both positive and free of r, so R is positive definite exactly when $\det R > 0$, that is, between the roots:

$$r \in \left(\frac{8 - 3\sqrt{21}}{25},\ \frac{8 + 3\sqrt{21}}{25} \right) = (-0.2299,\ 0.8699).$$

(c) Write $R = (1-r)I_d + r\,\mathbf{1}\mathbf{1}^T$. The eigenvalues are read off immediately: $\mathbf{1}$ is an eigenvector with eigenvalue $1 + (d-1)r$, and every vector orthogonal to $\mathbf{1}$ is an eigenvector with eigenvalue $1 - r$, of multiplicity $d-1$. Hence

$$\begin{aligned} R \succ 0 &\iff 1 + (d-1)r > 0 \ \text{ and } \ 1 - r > 0 \\ &\iff -\frac{1}{d-1} < r < 1. \end{aligned}$$

Problem 15.6. Analysis of a two-way stratified sample survey: Section 8.3 and Exercise 11.7 present an analysis of a stratified sample survey using a hierarchical model on the stratum probabilities. That analysis is not fully appropriate because it ignores the two-way structure of the stratification, treating the 16 strata as exchangeable.

The data are a CBS News survey of 1447 adults, divided into 16 strata cross-classified by region (Northeast, Midwest, South, West) and density of residential area (place sizes I–IV). Sampling is assumed proportional, so $N_j/N \approx n_j/n$.

Stratum, j	Bush	Dukakis	no opinion	n_j/n
Northeast, I	0.30	0.62	0.08	0.032
Northeast, II	0.50	0.48	0.02	0.032
Northeast, III	0.47	0.41	0.12	0.115
Northeast, IV	0.46	0.52	0.02	0.048
Midwest, I	0.40	0.49	0.11	0.032
Midwest, II	0.45	0.45	0.10	0.065
Midwest, III	0.51	0.39	0.10	0.080
Midwest, IV	0.55	0.34	0.11	0.100
South, I	0.57	0.29	0.14	0.015
South, II	0.47	0.41	0.12	0.066
South, III	0.52	0.40	0.08	0.068
South, IV	0.56	0.35	0.09	0.126
West, I	0.50	0.47	0.03	0.023
West, II	0.53	0.35	0.12	0.053
West, III	0.54	0.37	0.09	0.086
West, IV	0.56	0.36	0.08	0.057

Following (8.8), write $\phi_{1j} = \theta_{1j}/(\theta_{1j} + \theta_{2j})$ for the probability of preferring Bush given that a preference is expressed, and $\phi_{2j} = 1 - \theta_{3j}$ for the probability of expressing a preference.

(a) Set up a linear model for logit(ϕ) with three groups of varying coefficients, for the four regions, the four place sizes, and the 16 strata.

(b) Simplify the model by assuming that the ϕ_{1j}'s are independent of the ϕ_{2j}'s. This separates the problem into two generalized linear models, one estimating Bush vs. Dukakis preferences, the other estimating 'no opinion' preferences. Perform the computations for this model to yield posterior simulations for all parameters.

(c) Expand to a multivariate model by allowing the ϕ_{1j}'s and ϕ_{2j}'s to be correlated. Perform the computations under this model, using the results from Exercise 11.7 and part (b) above to construct starting distributions.

(d) Compare your results to those from the simpler model treating the 16 strata as exchangeable.

Solution. Table 8.2 reports rounded proportions, so I recovered counts as $n_j = \text{round}(1447\, n_j/n)$ and $y_{kj} = \text{round}(n_j p_{kj})$, adjusted in the largest cell to make the rows sum; this gives $n = 1442$ and a raw weighted Bush-minus-Dukakis margin of 0.1016, against the 0.097 of Figure 8.1a. Inferences below should be read modulo that reconstruction.

(a) Index stratum j by its region $r(j) \in \{1, \ldots, 4\}$ and place size $s(j) \in \{1, \ldots, 4\}$. Keeping the multinomial likelihood $y_j \sim \text{Multin}(n_j; \theta_{1j}, \theta_{2j}, \theta_{3j})$ of Section 8.3 and the reparameterization (8.8), put a two-way analysis of variance on each logit: for $k = 1, 2$,

$$\text{logit}(\phi_{kj}) = \mu_k + a_{k,r(j)} + b_{k,s(j)} + c_{kj},$$

with three batches of varying coefficients (Section 15.6)

$$a_{k,r} \sim \text{N}(0, \sigma_{ak}^2), \quad b_{k,s} \sim \text{N}(0, \sigma_{bk}^2), \quad c_{kj} \sim \text{N}(0, \sigma_{ck}^2),$$

over $r = 1, \ldots, 4$, $s = 1, \ldots, 4$, $j = 1, \ldots, 16$. The last batch is the region-by-size interaction, which is also the residual, since the design is a full 4×4 factorial with no replication. Each batch mean is fixed at 0 for the reason given in Exercise 15.2(c): only $\mu_k + \text{E}(a) + \text{E}(b) + \text{E}(c)$ is estimable. In general

(c_{1j}, c_{2j}) is bivariate normal with correlation ρ; parts (b) and (c) are $\rho = 0$ and ρ free. Priors uniform on μ_k, on each standard deviation, and on ρ.
(b) With $\rho = 0$ the multinomial factors as

$$p(y_j \mid \theta_j) \propto \text{Bin}(n_j - y_{3j} \mid n_j, \phi_{2j}) \, \text{Bin}(y_{1j} \mid y_{1j} + y_{2j}, \phi_{1j}),$$

so the two logits appear in disjoint factors and the posterior separates into two independent varying-coefficient logistic regressions. I ran Metropolis-within-Gibbs on each: normal random-walk updates for μ_k and for each coefficient, exact inverse-χ^2 draws for the three standard deviations, plus a joint shift move ($\mu_k \to \mu_k + d$ with one batch shifted by $-d$) to break the ridge between the constant and the batch levels. Four sequences of 150,000, second halves retained, acceptance rates near 0.55: $\hat{R} \leq 1.003$ for all 62 parameters.

Parameter	2.5%	median	97.5%
μ_1	-0.53	0.183	0.85
σ_{a1} (region)	0.01	0.211	1.27
σ_{b1} (size)	0.01	0.209	1.35
σ_{c1} (stratum)	0.01	0.094	0.33
μ_2	1.60	2.309	2.96
σ_{a2} (region)	0.01	0.178	1.44
σ_{b2} (size)	0.01	0.159	1.19
σ_{c2} (stratum)	0.00	0.165	0.58

The Bush-versus-Dukakis logit shows real region and size variation (σ_{a1}, σ_{b1} near 0.21, that is, about 5 percentage points) and almost no residual stratum variation (σ_{c1} near 0.09): the two-way additive structure absorbs essentially all of the between-stratum differences in partisan preference. The estimand (8.7), $\sum_j (N_j/N)\phi_{2j}(2\phi_{1j} - 1)$, has posterior median 0.102 with 95% interval $[0.053, 0.151]$ and posterior sd 0.0248.
(c) Letting $(c_{1j}, c_{2j}) \sim \text{N}(0, \Sigma_c)$ with Σ_c parameterized by $(\sigma_{c1}, \sigma_{c2}, \rho)$ couples the two logits. The stratum coefficients are now updated as bivariate blocks and $(\log \sigma_{c1}, \log \sigma_{c2}, \rho)$ by random walk, started from the part (b) output and from Exercise 11.7's posterior for ρ. Four sequences of 60,000, $\hat{R} \leq 1.02$:

$$\rho: \ \text{median} \ -0.19, \ 95\% \ [-0.96, \ 0.94],$$

with σ_{c1} median 0.090 and σ_{c2} median 0.166, unchanged from (b), and the estimand at median 0.102, 95% interval $[0.053, 0.150]$. The correlation is not estimable here: after region and place size are in the model there is almost nothing left in c_{1j} to correlate with anything (σ_{c1} is essentially zero), so ρ is close to its uniform prior. The multivariate extension is free but buys nothing.
(d) Fitting the exchangeable model of Section 8.3 and Exercise 11.7 to the same reconstructed counts — the same sampler with the region and size batches deleted — gives ρ with median -0.31 and 95% interval $[-1.00, 0.88]$, reproducing the book's report of a negative median with substantial variability, and the same estimand: median 0.102, 95% interval $[0.054, 0.150]$. Two discrepancies with the book's own account of this model: it reports τ_1 with median 0.23 against the 0.15 here, which is why its ϕ_{1j} medians span 0.48–0.59 rather than the narrower range below (pinning (τ_1, τ_2, ρ) at the Table 8.3 medians does reproduce 0.47–0.60); and it reports the estimand (8.7) moving up to 0.11 under the hierarchical model, which I cannot reproduce — the median stays at 0.102 even with the hyperparameters pinned at its own values.
The difference is entirely at the stratum level. Posterior medians of ϕ_{1j}:

Model	range of median ϕ_{1j}
raw proportions	0.333 – 0.684
exchangeable (8.3)	0.516 – 0.576
two-way (b), (c)	0.445 – 0.600

The exchangeable model shrinks all 16 strata to a common value because each stratum has only $n_j \approx$ 22–182 respondents and nothing else is known about it. The two-way model knows that Northeast I is in the same region as three other strata and the same place size as three others, and pools along those two margins instead of pooling everything toward one point; so Northeast I, which is low in both its row and its column, stays low (0.445) rather than being dragged to 0.52, and South IV stays high

(0.600). That is the correct correction: the strata are not exchangeable, they are exchangeable within region and within place size.

None of this moves the population aggregate. All three models put the Bush-minus-Dukakis margin at 0.102 ± 0.025, because the estimand is a weighted average over all 16 strata and the models differ only in how they redistribute a fixed total among them. The two-way model matters if the stratum-level probabilities are of interest — for instance for small-area estimation, or for reweighting to a population whose region-by-size composition differs from the sample's — and not otherwise.

16 Generalized Linear Models

Contents

16.1 Exercises 16.1–16.7

Problem 16.1. Normal approximation for generalized linear models: derive equations (16.4). Context (Section 16.2, page 410). The log-likelihood of a generalized linear model factors over the n units as $\prod_{i=1}^{n} \exp(L(y_i \mid \eta_i, \phi))$, with linear predictor $\eta = X\beta$ and dispersion parameter ϕ. Each factor is to be approximated by a normal density in η_i,

$$L(y_i \mid \eta_i, \phi) \approx -\frac{1}{2\sigma_i^2}(z_i - \eta_i)^2 + \text{constant},$$

that is, unit i is replaced by a pseudodatum z_i observed with pseudovariance σ_i^2. Writing $(\hat{\beta}, \hat{\phi})$ for the center of the approximation and $\hat{\eta}_i = (X\hat{\beta})_i$, determine z_i and σ_i^2 by matching the first- and second-order terms of the Taylor series of $L(y_i \mid \eta_i, \phi)$ about $\hat{\eta}_i$, and thereby establish

$$z_i = \hat{\eta}_i - \frac{L'(y_i \mid \hat{\eta}_i, \hat{\phi})}{L''(y_i \mid \hat{\eta}_i, \hat{\phi})}, \qquad \sigma_i^2 = -\frac{1}{L''(y_i \mid \hat{\eta}_i, \hat{\phi})},$$

where L' and L'' denote $dL/d\eta$ and $d^2L/d\eta^2$.

Solution. Match coefficients of $u = \eta_i - \hat{\eta}_i$ in the two quadratics. Writing $L' = L'(y_i \mid \hat{\eta}_i, \hat{\phi})$ and $L'' = L''(y_i \mid \hat{\eta}_i, \hat{\phi})$, the second-order Taylor expansion of the exact log-likelihood factor is

$$L(y_i \mid \eta_i, \phi) \approx L(y_i \mid \hat{\eta}_i, \hat{\phi}) + L'u + \tfrac{1}{2}L''u^2.$$

The proposed normal approximation, expanded in the same variable via $z_i - \eta_i = (z_i - \hat{\eta}_i) - u$, is

$$-\frac{(z_i - \eta_i)^2}{2\sigma_i^2} + \text{const} = -\frac{(z_i - \hat{\eta}_i)^2}{2\sigma_i^2} + \frac{z_i - \hat{\eta}_i}{\sigma_i^2}\,u - \frac{u^2}{2\sigma_i^2} + \text{const}.$$

The constant terms are irrelevant (they are absorbed into the normalizing factor), so the two expansions agree to second order if and only if the coefficients of u^2 and of u agree:

$$\begin{aligned} u^2: &\quad \tfrac{1}{2}L'' = -\frac{1}{2\sigma_i^2} &&\implies \quad \sigma_i^2 = -\frac{1}{L''}, \\ u^1: &\quad L' = \frac{z_i - \hat{\eta}_i}{\sigma_i^2} = -L''\,(z_i - \hat{\eta}_i) &&\implies \quad z_i = \hat{\eta}_i - \frac{L'}{L''}, \end{aligned}$$

which are equations (16.4). The division requires $L'' \neq 0$, and $\sigma_i^2 > 0$ requires $L'' < 0$: for the standard families with canonical link L is concave in η_i (the binomial-logistic case of page 411 has $L'' = -n_i e^{\eta_i}/(1 + e^{\eta_i})^2 < 0$), and in general the approximation is centered at a mode, where $L'' < 0$.

Problem 16.2. Computation for a simple generalized linear model:
(a) Express the bioassay example of Section 3.7 as a generalized linear model and obtain posterior simulations using the computational techniques presented in Section 16.2.
(b) Fit a probit regression model instead of the logit (you should be able to use essentially the same steps after altering the likelihood appropriately). Discuss any changes in the posterior inferences.
The data are those of Table 3.1: at each of four dose levels x_i (on the log g/ml scale), n_i animals were exposed and y_i of them died.

Dose, x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
−0.86	5	0
−0.30	5	1
−0.05	5	3
0.73	5	5

The model of Section 3.7 is $y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$ independently, with (3.14) $\text{logit}(\theta_i) = \alpha + \beta x_i$ and the noninformative prior density $p(\alpha, \beta) \propto 1$. The quantity of interest is the LD50, x such that $\theta = 1/2$, that is, $-\alpha/\beta$.

Solution. (a) This is the binomial generalized linear model with logistic link, $\eta = X\beta$, $X = (\mathbf{1}\ x)$ a 4×2 matrix, $\beta = (\alpha, \beta)$, and a uniform prior density; no dispersion parameter is present (page 411). By the binomial-logistic pseudodata of page 411, with $\hat{p}_i = e^{\hat{\eta}_i}/(1 + e^{\hat{\eta}_i})$,

$$z_i = \hat{\eta}_i + \frac{y_i - n_i\hat{p}_i}{n_i\hat{p}_i(1-\hat{p}_i)}, \qquad \sigma_i^2 = \frac{1}{n_i\hat{p}_i(1-\hat{p}_i)},$$

and one step of weighted least squares of z on X with weights $w_i = \sigma_i^{-2}$ updates $\hat{\beta}$. Started at $\hat{\beta} = 0$ the iteration converges in 8 steps to the posterior mode and curvature

$$\hat{\beta} = \begin{pmatrix} 0.847 \\ 7.749 \end{pmatrix}, \qquad V_\beta = \left(X^T \text{diag}(w_i) X\right)^{-1} = \begin{pmatrix} 1.039 & 3.546 \\ 3.546 & 23.744 \end{pmatrix}.$$

Thus $p(\alpha, \beta \mid y) \approx \text{N}(\hat{\beta}, V_\beta)$ with $\text{sd}(\alpha) = 1.02$, $\text{sd}(\beta) = 4.87$ and correlation 0.71. Drawing 10^5 values from this normal approximation gives $\Pr(\beta > 0 \mid y) \approx 0.94$ and a median LD50 of -0.114.
The normal approximation is not adequate here (only 5 animals per dose; the likelihood has the long upper tail in β visible in Figure 3.3), so, as recommended at the end of Section 16.2, correct it by importance resampling (Section 10.4): draw 2×10^5 values of (α, β) from a t_4 with location $\hat{\beta}$ and scale matrix $2V_\beta$, weight them by $p(y \mid \alpha, \beta)/p_t(\alpha, \beta)$, and resample 20,000 without replacement (effective sample size of the weights 1.3×10^5). The resulting draws have

$$\begin{aligned} \text{E}(\alpha \mid y) &= 1.31, \quad \text{sd}(\alpha \mid y) = 1.11, \\ \text{E}(\beta \mid y) &= 11.6, \quad \text{sd}(\beta \mid y) = 5.73, \end{aligned}$$

agreeing to two digits with a direct grid evaluation of $p(\alpha, \beta \mid y)$ over $[-8, 16] \times [-12, 80]$ (which gives 1.315, 1.102, 11.64, 5.77), and $\Pr(\beta > 0 \mid y) > 0.9999$. The posterior median of the LD50 $= -\alpha/\beta$ is -0.111 with central 95% interval $[-0.28, 0.11]$, reproducing the values reported in Section 3.7.
(b) Only L and its derivatives change. With Φ and ϕ the standard normal distribution and density functions, $L(y_i \mid \eta_i) = y_i \log \Phi(\eta_i) + (n_i - y_i) \log(1 - \Phi(\eta_i))$, so, writing $D_i = y_i/\Phi(\eta_i) - (n_i - y_i)/(1 - \Phi(\eta_i))$,

$$\begin{aligned} L' &= \phi(\eta_i)\, D_i, \\ L'' &= -\phi(\eta_i)^2 \left[\frac{y_i}{\Phi(\eta_i)^2} + \frac{n_i - y_i}{(1-\Phi(\eta_i))^2}\right] - \eta_i \phi(\eta_i) D_i, \end{aligned}$$

and (16.4) supplies z_i, σ_i^2 exactly as before. The same 8-step iteration gives

$$\hat{\beta} = \begin{pmatrix} 0.484 \\ 4.459 \end{pmatrix}, \qquad V_\beta = \begin{pmatrix} 0.434 & 1.575 \\ 1.575 & 9.733 \end{pmatrix},$$

and the same importance resampling step yields $\text{E}(\alpha \mid y) = 0.73$, sd $= 0.60$, $\text{E}(\beta \mid y) = 6.12$, sd $= 2.75$ (grid check: 0.725, 0.602, 6.124, 2.751).
The coefficients change substantially, the inferences do not. The ratio of the posterior means, $11.6/6.12 = 1.90$, is the familiar rescaling between the logistic and normal tolerance distributions (the logistic has standard deviation $\pi/\sqrt{3} = 1.81$), and it cancels in every statement about the data: the posterior mean fitted death probabilities at the four doses are $(0.008, 0.152, 0.648, 0.993)$ under the logit and $(0.007, 0.170, 0.644, 0.993)$ under the probit, and the LD50 has posterior median -0.117 with 95% interval $[-0.285, 0.116]$, against -0.111 and $[-0.280, 0.106]$ for the logit. With four dose levels and 20 animals the data cannot distinguish the two links, and neither should any scientific conclusion.

Problem 16.3. Overdispersed models:
(a) Express the bioassay example of Section 3.7 as a generalized linear model, but replacing (3.14) by

$$\text{logit}(\omega_i) \sim \text{N}(\alpha + \beta x_i,\ \sigma^2),$$

so that the logistic regression holds approximately but not exactly. Set up a noninformative prior distribution and obtain posterior simulations of (α, β, σ) under this model. Discuss the effect that this model expansion has on scientific inferences for this problem.
(b) Repeat (a) with the following hypothetical data: $n = (5000, 5000, 5000, 5000)$, $y = (500, 1000, 3000, 4500)$, and x unchanged from the first column of Table 3.1.
The bioassay data of Table 3.1 are

Dose, x_i (log g/ml)	n_i	y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

and equation (3.14) is $\text{logit}(\theta_i) = \alpha + \beta x_i$, with $y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$.

Solution. The expanded model is the hierarchical binomial generalized linear model

$$\begin{aligned} y_i \mid \omega_i &\sim \text{Bin}(n_i, \omega_i), \\ \text{logit}(\omega_i) = \eta_i &\sim \text{N}(\alpha + \beta x_i, \sigma^2), \qquad i = 1, \dots, 4, \end{aligned}$$

with η a vector of four latent linear predictors; σ is the overdispersion parameter of Section 16.1. Because each η_i enters exactly one likelihood factor, the latent vector can be integrated out one dimension at a time,

$$p(y \mid \alpha, \beta, \sigma) = \prod_{i=1}^{4} \int \text{Bin}\big(y_i \mid n_i, \text{logit}^{-1}(\eta)\big)\, \text{N}(\eta \mid \alpha + \beta x_i, \sigma^2)\, d\eta,$$

leaving a three-dimensional posterior in $(\alpha, \beta, \log\sigma)$. Each one-dimensional integral is evaluated on a fine trapezoid grid in η rather than by Gauss-Hermite quadrature: in part (b) the binomial factor has width $(n_i \hat{p}_i(1-\hat{p}_i))^{-1/2} \approx 0.04$ in η, far narrower than the spacing of Gauss-Hermite nodes once σ exceeds about 0.2, and Gauss-Hermite then silently loses the integral.
(a) A flat prior density on σ gives an improper posterior, so it must be replaced. The two extreme groups are on the boundary, $y_1 = 0$ and $y_4 = n_4$, and their factors do not vanish as the latent mean runs off to $\mp\infty$; substituting $(\alpha, \beta) = \sigma(a, b)$, whose Jacobian is σ^2, and using $\int \text{Bin}(y_i \mid 5, \text{logit}^{-1}(\eta))\, d\eta < \infty$ for $y_2 = 1$ and $y_3 = 3$, each interior factor contributes σ^{-1} and

$$\iint p(y \mid \alpha, \beta, \sigma)\, d\alpha\, d\beta \longrightarrow c > 0 \qquad (\sigma \to \infty),$$

numerically flat at $c \approx 1.8$ from $\sigma = 50$ onward. Hence $\int p(y \mid \alpha, \beta, \sigma)\, d\alpha\, d\beta\, d\sigma = \infty$: with $J = 4$ groups, two of them uninformative about σ, there is nothing left to estimate σ from, the difficulty of Section 5.4 for small J in its sharpest form. Keep $p(\alpha, \beta) \propto 1$ and take the weakly informative $\sigma \sim \text{half-Cauchy}(0, 2.5)$ of Section 16.3.
That prior makes the posterior proper but only just: it decays like σ^{-2}, so $\text{E}(\sigma \mid y)$, $\text{E}(\alpha \mid y)$ and $\text{E}(\beta \mid y)$ are all infinite and only quantiles may be reported. Quadrature over $(\alpha, \beta, \log\sigma)$ (the latent integrals on a fine η grid, with the two boundary groups' tails added in closed form, and $\log\sigma$ carried out to 2×10^5) gives

Parameter	2.5%	25%	Median	75%
α	-3.1	0.7	2.1	4.6
β	3.8	10.3	16.7	30
σ	0.09	0.8	2.1	4.6

with upper 2.5% points of order 45 for α and σ and 240 for β, fixed by the prior tail rather than by the data and correspondingly unstable.

The expansion costs precision without changing direction. The dose effect stays positive, $\Pr(\beta > 0 \mid y) = 0.997$, but the median of β rises from 10.6 in Exercise 16.2 to 16.7, and the LD50 $= -\alpha/\beta$ has median -0.13 with 95% interval $[-0.44, 0.24]$ against $[-0.28, 0.11]$ under (3.14). The data say nothing against the simple model – the logistic fit of Exercise 16.2 matches the four observed proportions to Pearson $X^2 = 0.03$ on 2 degrees of freedom – so the posterior for σ is the prior filtered through a likelihood that is flat in σ, and the widened intervals are the price of admitting that four binomial observations of size 5 cannot verify a functional form.

(b) With $n_i = 5000$ the empirical logits are known to within ± 0.05 and the lack of fit is unmistakable:

$$\text{logit}(y_i/n_i) = (-2.197,\ -1.386,\ 0.405,\ 2.197),$$

whose least squares fit on x leaves residuals $(0.18, -0.62, 0.45, -0.01)$, root mean square 0.56 on 2 degrees of freedom. The ordinary logistic regression of (3.14) gives $\hat{\alpha} = 0.133$ (± 0.019), $\hat{\beta} = 3.039$ (± 0.044) with Pearson $X^2 = 571$ on 2 degrees of freedom: the model is decisively rejected.

Now no group is on the boundary, each $\int \text{Bin}(y_i \mid n_i, \text{logit}^{-1}(\eta))\, d\eta = n_i/(y_i(n_i - y_i))$ is finite, and the same rescaling gives $\int\int p\, d\alpha\, d\beta \to 3.6 \times 10^{-13}\sigma^{-2}$, integrable: the flat prior on σ is proper here. It is still not usable, because a σ^{-2} tail leaves σ, α and β again without moments. Quadrature in $(\alpha, \beta, \log\sigma)$ out to $\sigma = 10^5$ gives, under $\sigma \sim$ half-Cauchy$(0, 2.5)$ and under $p(\sigma) \propto 1$:

Parameter	Median	95% interval	Median (flat)	95% interval (flat)
α	0.10	$[-1.45,\ 1.64]$	0.09	$[-5.1,\ 5.2]$
β	2.89	$[0.27,\ 5.56]$	2.89	$[-5.8,\ 11.5]$
σ	0.86	$[0.33,\ 4.02]$	1.15	$[0.36,\ 25.1]$
LD50	-0.04	$[-0.75,\ 0.75]$	-0.04	$[-1.5,\ 1.8]$

Here σ is genuinely estimated away from zero, $\Pr(\sigma > 0.5 \mid y) = 0.83$, and the medians are identical under the two priors; only the tails differ, and they differ by a factor of six in σ, since four groups never pin down an upper limit on the between-group variance.

The contrast with the classical fit is the point. Increasing n_i from 5 to 5000 does not buy a thousandfold increase in precision about the dose-response line: the binomial standard error of $\hat{\beta}$ is 0.044, but the posterior sd under the overdispersed model is 1.47, thirty times larger, because with four dose levels the accuracy of the line is limited by the scatter of the four true logit(ω_i) about it and not by the sampling error within groups. The LD50 interval widens from $[-0.056, -0.032]$ (binomial, delta method) to $[-0.75, 0.75]$: when the logistic regression is only approximately true, replication within a dose level buys almost nothing and only additional dose levels help.

Problem 16.4. Computation for a hierarchical generalized linear model:

(a) Express the rat tumor example of Section 5.1 as a generalized linear model and obtain posterior simulations using the computational techniques presented in Section 16.2.

(b) Use the posterior simulations to check the fit of the model.

The data (Table 5.1, from Tarone, 1982) are tumor incidences y_j/n_j in 70 historical control groups of female F344 rats together with a current group, 71 groups in all:

j	y_j/n_j
1–10	0/20, 0/20, 0/20, 0/20, 0/20, 0/20, 0/20, 0/19, 0/19, 0/19
11–20	0/19, 0/18, 0/18, 0/17, 1/20, 1/20, 1/20, 1/20, 1/19, 1/19
21–30	1/18, 1/18, 2/25, 2/24, 2/23, 2/20, 2/20, 2/20, 2/20, 2/20
31–40	2/20, 1/10, 5/49, 2/19, 5/46, 3/27, 2/17, 7/49, 7/47, 3/20
41–50	3/20, 2/13, 9/48, 10/50, 4/20, 4/20, 4/20, 4/20, 4/20, 4/20
51–60	4/20, 10/48, 4/19, 4/19, 4/19, 5/22, 11/46, 12/49, 5/20, 5/20
61–70	6/23, 5/19, 6/22, 6/20, 6/20, 6/20, 16/52, 15/47, 15/46, 9/24
71	4/14 (current experiment)

In Section 5.1 these are modeled as $y_j \mid \theta_j \sim \text{Bin}(n_j, \theta_j)$ with θ_j drawn from a common population distribution.

Solution. (a) As a generalized linear model this is the binomial family with logistic link, one indicator per experiment, and a hierarchical normal prior distribution on the coefficients, as described under 'Hierarchical models' in Section 16.2:

$$y_j \mid \alpha_j \sim \text{Bin}\big(n_j,\ \text{logit}^{-1}(\alpha_j)\big),$$
$$\alpha_j \mid \mu, \tau \sim \text{N}(\mu, \tau^2), \qquad j = 1, \ldots, 71,$$

with $p(\mu, \tau) \propto 1$. This replaces the conjugate Beta(α, β) population distribution of Section 5.3 by a normal one on the logit scale; it is the formulation that would survive the addition of explanatory variables, which the beta-binomial trick would not (page 409).
Section 16.2 computation. At the current guess $\hat{\alpha}$, with $\hat{p}_j = \text{logit}^{-1}(\hat{\alpha}_j)$, formula (16.4) for the binomial-logistic model gives

$$z_j = \hat{\alpha}_j + \frac{y_j - n_j\hat{p}_j}{n_j\hat{p}_j(1-\hat{p}_j)}, \qquad \sigma_j^2 = \frac{1}{n_j\hat{p}_j(1-\hat{p}_j)},$$

so that the approximate model is the normal hierarchical model of Section 5.4, $z_j \sim \text{N}(\alpha_j, \sigma_j^2)$, $\alpha_j \sim \text{N}(\mu, \tau^2)$, for which

$$p(\mu \mid \tau, z) = \text{N}(\hat{\mu}, V_\mu),$$
$$p(\tau \mid z) \propto V_\mu^{1/2} \prod_{j=1}^{71} \text{N}\big(z_j \mid \hat{\mu}, \sigma_j^2 + \tau^2\big),$$

equations (5.20) and (5.21), where $V_\mu^{-1} = \sum_j (\sigma_j^2 + \tau^2)^{-1}$ and $\hat{\mu} = V_\mu \sum_j z_j(\sigma_j^2 + \tau^2)^{-1}$, and $\alpha_j \mid \mu, \tau, z$ is normal with mean $(z_j/\sigma_j^2 + \mu/\tau^2)/(\sigma_j^{-2} + \tau^{-2})$. Alternating the pseudodata step with the hierarchical step converges in six iterations to

$$\hat{\mu} = -1.855\ (\pm 0.105), \qquad \hat{\tau} = 0.63,$$

from which simulations are drawn in the order (μ, τ), then $\alpha \mid \mu, \tau$, as in Section 5.4.
Because each α_j appears in exactly one likelihood factor, the approximation can be dispensed with entirely and the answer obtained essentially exactly: marginalizing each α_j by 160-point Gauss-Hermite quadrature leaves the two-dimensional

$$p(\mu, \tau \mid y) \ \propto \ \prod_{j=1}^{71} \int \text{Bin}\big(y_j \mid n_j, \text{logit}^{-1}(\alpha)\big) \times \text{N}(\alpha \mid \mu, \tau^2)\, d\alpha,$$

evaluated on a 251×200 grid over $[-3.5, -1] \times [0.01, 2]$ and sampled as in Section 5.3. This gives

$$\text{E}(\mu \mid y) = -1.95, \quad \text{sd}(\mu \mid y) = 0.13, \quad 95\%\text{: } [-2.22, -1.72],$$
$$\text{E}(\tau \mid y) = 0.71, \quad \text{sd}(\tau \mid y) = 0.13, \quad 95\%\text{: } [0.47,\ 1.00],$$

with joint mode $(-1.94, 0.67)$: the normal-approximation answer of the previous paragraph is close but slightly optimistic, understating τ and the uncertainty in μ by about 10 and 20 percent. Drawing α_j from its exact conditional density for each of 4000 hyperparameter draws and transforming, $\theta_j = \text{logit}^{-1}(\alpha_j)$, the current experiment is shrunk from its raw rate $4/14 = 0.286$ to posterior median 0.194 with 95% interval [0.08, 0.38], matching the beta-binomial analysis of Section 5.3; experiment 67 $(16/52 = 0.308)$ is shrunk only to 0.265, and experiment 1 (0/20) is pulled up to 0.066. A tumor rate in a new experiment, $\text{logit}^{-1}(\alpha_{72})$ with $\alpha_{72} \sim \text{N}(\mu, \tau^2)$, has posterior predictive median 0.126 and 95% interval [0.03, 0.38].
(b) Simulate $y_j^{\text{rep}} \sim \text{Bin}(n_j, \theta_j)$ for each of the 4000 posterior draws and compare test quantities (Section 6.3):

Test quantity $T(y)$	Observed	95% posterior predictive interval	$p_B = \Pr(T(y^{\text{rep}}) \geq T(y) \mid y)$
mean of y_j/n_j	0.138	[0.120, 0.167]	0.64
sd of y_j/n_j	0.104	[0.085, 0.131]	0.56
$\max_j y_j/n_j$	0.375	[0.348, 0.625]	0.89
number of j with $y_j = 0$	14	[3, 15]	0.069

and, for the realized discrepancy $T(y,\theta) = \sum_j (y_j - n_j\theta_j)^2/(n_j\theta_j(1-\theta_j))$, $p_B = 0.38$. With 4000 draws the Monte Carlo standard error of each p_B is at most 0.008.
The model fits the location, spread and upper tail of the data, and the chi-square discrepancy gives no sign of residual over- or underdispersion. The one strained feature is the lower tail: 14 of the 71 experiments recorded no tumors at all, and only 7% of replicated datasets produce that many zeros, the replications typically producing 8 or 9. The logistic-normal population distribution has to push μ down and τ up to manufacture the observed pile-up at zero, and it cannot quite do so; a population distribution with more mass at small θ (a beta with small parameters, or a mixture with a low-rate component) would accommodate these groups better. With $p_B = 0.068$ this is a mild warning, not a refutation.

Problem 16.5. Poisson model with overdispersion: Find data on counts of some event given some predictor.
(a) Fit a standard Poisson regression model relating the log of the expected count linearly to the predictor.
(b) Perform some model checking on the simple model proposed in (a), and see if there is evidence of overdispersion.
(c) Fit a hierarchical model assuming independent normally distributed errors.
(d) Is there evidence that this model provides a better fit to the data?
(e) Experiment with other forms of hierarchical model, in particular a mixture model that assumes a discrete prior distribution on two or three points for the errors, and perhaps also a t prior distribution.
Explore the fit of the various models to the data and examine the sensitivity of inferences to the assumptions.
See Hinde (1982) for background on these models.
The data used below are the U.S. House committee bill-assignment counts distributed with the statsmodels package: for each of the 20 standing committees, the number of bills assigned in the first 100 days of the 104th House, with committee size, number of subcommittees, staff, a high-prestige indicator, and the number of bills assigned in the first 100 days of the previous (103rd) House.

Committee	BILLS104	SIZE	SUBS	STAFF	PRESTIGE	BILLS103
1	6	58	13	109	1	9
2	23	42	0	39	1	101
3	44	13	2	25	1	54
4	355	39	5	23	1	542
5	125	51	5	61	0	101
6	131	43	5	69	0	158
7	271	49	4	79	0	196
8	63	44	3	68	0	40
9	149	51	7	99	0	72
10	253	35	5	56	0	168
11	81	49	5	46	0	60
12	89	55	7	48	0	75
13	142	44	5	58	0	98
14	155	61	6	74	0	69
15	27	50	4	58	0	25
16	8	43	4	29	0	9
17	28	33	3	36	0	41
18	68	12	0	24	0	233
19	1	10	0	9	0	0
20	4	16	2	24	0	2

Solution. Throughout, y_i is BILLS104 and the single predictor is $x_i = \log(1 + \text{BILLS103}_i)$, centered at its mean 3.92; $n = 20$.

(a) The Poisson generalized linear model with log link (16.2) is

$$y_i \sim \text{Poisson}\big(e^{\eta_i}\big), \qquad \eta_i = \beta_1 + \beta_2 x_i,$$

with $p(\beta) \propto 1$. Metropolis sampling (four chains, adapted normal jumping kernel, acceptance 0.34, $\hat{R} = 1.00$) gives $\beta_1 = 4.173$ (± 0.032) and $\beta_2 = 0.755$ (± 0.024): a committee receiving twice as many bills in the 103rd House is predicted to receive $2^{0.755} = 1.69$ times as many in the 104th.
(b) The model is hopeless. The Pearson statistic is

$$X^2 = \sum_{i=1}^{20} \frac{(y_i - e^{\eta_i})^2}{e^{\eta_i}} = 435 \text{ at the posterior median},$$

against 18 degrees of freedom; as a posterior predictive check (Section 6.3), replicated datasets give X^2 in the 95% interval [9, 34] and $p_B < 0.0003$ (4000 draws, none as extreme). Standardized residuals $(y_i - \hat{\mu}_i)/\hat{\mu}_i^{1/2}$ run from -9.6 to $+7.9$. The trouble is structural: the Poisson model ties the variance to the mean, while the counts here range from 1 to 355, and no amount of adjustment of β can produce residual variation of the observed size. This is overdispersion of the kind described at the end of Section 16.1, and its practical consequence is the absurdly small standard errors in (a).
(c) Add a normal error term for each data point, as in the police-stops model of Section 16.4:

$$y_i \sim \text{Poisson}\big(e^{\eta_i + \epsilon_i}\big), \qquad \epsilon_i \sim \text{N}(0, \sigma^2),$$

with $p(\beta, \sigma) \propto 1$. Since ϵ_i enters one likelihood factor only, it is integrated out by 60-node Gauss-Hermite quadrature centered and scaled at the conditional mode of ϵ_i (plain Gauss-Hermite is not accurate here: for the largest counts the Poisson factor has width $y_i^{-1/2} \approx 0.05$ in ϵ, below the node spacing),

$$p(y_i \mid \beta, \sigma) = \int \text{Poisson}\big(y_i \mid e^{\eta_i + \epsilon}\big) \times \text{N}(\epsilon \mid 0, \sigma^2)\, d\epsilon,$$

leaving a three-dimensional posterior for $(\beta_1, \beta_2, \log \sigma)$, sampled by Metropolis (acceptance 0.23, $\hat{R} \leq 1.001$):

$$\beta_1 = 3.89\ (\pm 0.16), \quad \beta_2 = 0.96\ (\pm 0.12), \quad \sigma:\ \text{median } 0.62,\ 95\%\ [0.44,\ 0.93].$$

The regression slope moves to 0.96 and its standard error grows fivefold. The overdispersion is substantial in absolute terms as well: $\sigma = 0.62$ means the true rates scatter about the regression line by a factor of $e^{\pm 0.62}$, that is between $\times 0.54$ and $\times 1.86$.
(d) Yes, decisively. WAIC (Section 7.2), computed from the marginal pointwise densities $p(y_i \mid \beta, \sigma)$:

Model	WAIC	p_{WAIC}
Poisson	637.4	39.0
normal errors (lognormal-Poisson)	197.9	2.4
t_4 errors	197.8	2.7
two-point mixture	289.6	18.6
three-point mixture	224.6	18.5

a difference of 440 in WAIC. (The value $p_{\text{WAIC}} = 39$ for the Poisson model, nearly twice the number of observations, is itself a diagnostic that the model is badly misspecified and that the WAIC approximation is not to be trusted for it; the comparison is not close enough for that to matter.) The predictive checks agree: the largest count, 355, is unremarkable under the expanded model ($p_B = 0.81$), and the chi-square discrepancy is no longer extreme.
(e) Three further error distributions, all fitted the same way with the ϵ_i marginalized.
(i) Two-point mixture: $\epsilon_i = 0$ with probability λ and $\epsilon_i = \delta$ with probability $1-\lambda$ (with β_1 absorbing the first location, which identifies the model). The posterior puts δ at -1.25, 95% interval $[-1.40, -1.10]$, with $\lambda = 0.55$: the data are described as two groups of committees whose rates differ by a factor of $e^{1.25} = 3.5$. WAIC 289.6.
(ii) Three-point mixture: adding a third support point improves matters (WAIC 224.6, offsets -1.03 and $+0.47$ relative to the first point), but it is still 27 units worse than the continuous model.

(iii) t_4 errors, $\epsilon_i \sim t_4(0, \sigma)$: WAIC 197.8, indistinguishable from the normal, with σ median 0.48, 95% interval [0.30, 0.79] (smaller than the normal σ, as it must be, since the t_4 scale is not the standard deviation).
The inferences one actually cares about are stable across the continuous specifications and unstable across the discrete ones. The slope β_2 is 0.96 ($\pm$0.12) under normal errors and 0.99 ($\pm$0.12) under t_4 errors, but 0.60 ($\pm$0.03) under the two-point mixture and 0.68 ($\pm$0.08) under the three-point mixture: a discrete error distribution with a handful of support points is doing the work of the regression, absorbing into its mixture components variation that properly belongs on the regression line, and the resulting slope is both attenuated and spuriously precise. With 20 observations the data cannot support a discrete latent structure, and the continuous error models – normal or t, it makes no difference here – are the right default.

Problem 16.6. Fake-data simulation: Consider the following discrete-data model: $y_i \sim \text{Poisson}(e^{X_i\beta})$, $i = 1, \ldots, n$, with independent Cauchy prior distributions with location 0 and scale 2.5 on the elements of β.
(a) Write a program in R to apply the Metropolis algorithm for β given data X, y. Your program should work with any number of predictors (that is, X can be any matrix with the same number of rows as the length of y).
(b) Simulate fake data from the model for a case with 50 data points and 3 predictors and run your program. Plot the posterior simulations from multiple chains and monitor convergence.
(c) Now suppose you want to allow for overdispersion. In a sentence or two, explain why it typically makes sense to fit an overdispersed model in this setting.
(d) Write a new model using the negative binomial distribution that is a sensible extension of the above Poisson model. Be careful about parameters and transformations! Write the model rigorously in statistical notation, and write an R function to compute the log posterior density. (We are not, however, asking you to program the Metropolis algorithm for this model or fit it to data.)

Solution. (a) The program needs only the unnormalized log posterior density

$$\log p(\beta \mid X, y) = \sum_{i=1}^{n}\Big(y_i X_i\beta - e^{X_i\beta} - \log(y_i!)\Big) - \sum_{j=1}^{k}\log\Big(1 + \big(\beta_j/2.5\big)^2\Big),$$

which is a single vectorized expression in $X\beta$ and so works for any X with $\text{nrow}(X) = \text{length}(y)$, and a symmetric jumping rule. Take the jump $\beta^* \sim \text{N}(\beta^{t-1}, c^2\Sigma)$; because the proposal is symmetric the Metropolis ratio is $r = \exp\{\log p(\beta^* \mid X, y) - \log p(\beta^{t-1} \mid X, y)\}$ and $\beta^t = \beta^*$ with probability $\min(r, 1)$ (Section 11.2). For Σ use the inverse Fisher information at the posterior mode, $\Sigma = (X^T \text{diag}(e^{X_i\hat\beta})X)^{-1}$, which is the V_β of the normal approximation of Section 16.2 for the Poisson model, and $c = 2.4/\sqrt{k}$ (Section 11.9). Guard the exponential by returning $-\infty$ when any $X_i\beta$ overflows.
(b) With $n = 50$, $k = 3$ (a constant and two standard normal predictors), and $\beta^{\text{true}} = (1.5, 0.8, -0.5)$, the simulated counts ranged from 0 to 45 with mean 6.1. The mode is (1.450, 0.835, $-$0.446) with normal-approximation standard deviations (0.076, 0.052, 0.058); four chains of length 20,000 started from overdispersed points $\hat\beta + 2\text{N}(0, I)$ and with the second halves retained gave acceptance rate 0.32 and

$$\hat R = (1.0003,\ 1.0002,\ 1.0003),$$

with effective sample sizes of roughly 1000 per chain (Section 11.5). The trace plots of the four chains overlap completely after a few hundred iterations. The posterior summaries recover the truth:

Parameter	True	Mean	sd	95% interval
β_1	1.5	1.446	0.077	[1.293, 1.594]
β_2	0.8	0.836	0.052	[0.734, 0.940]
β_3	-0.5	-0.447	0.058	[-0.564, -0.338]

all three intervals containing the values used to simulate the data.

(c) Because the Poisson distribution has only one parameter, it forces $\mathrm{var}(y_i) = \mathrm{E}(y_i)$, and real counts almost never satisfy this: any predictor left out of X, and any heterogeneity among units, inflates the variance beyond the mean. Fitting the Poisson model when the data are overdispersed leaves $\hat{\beta}$ roughly unbiased but makes its standard errors far too small (Exercise 16.5 (b)), so the overdispersed model is what protects the inferences one reports.
(d) Replace the Poisson by a negative binomial with the same mean function and a separate overdispersion parameter:

$$y_i \mid \beta, \phi \sim \text{Neg-bin}(\mu_i, \phi), \qquad \mu_i = e^{X_i\beta},$$
$$p(y_i \mid \beta, \phi) = \frac{\Gamma(y_i + \phi)}{\Gamma(\phi)\, y_i!} \left(\frac{\phi}{\phi + \mu_i}\right)^{\phi} \left(\frac{\mu_i}{\mu_i + \phi}\right)^{y_i},$$

so that $\mathrm{E}(y_i) = \mu_i$ and $\mathrm{var}(y_i) = \mu_i + \mu_i^2/\phi$; equivalently $y_i \sim \text{Poisson}(\mu_i\lambda_i)$ with $\lambda_i \sim \text{Gamma}(\phi, \phi)$, the gamma analogue of the normal error term of Exercise 16.5 (c). The parameterization matters. Do not put a prior distribution on ϕ, which is unbounded and in the wrong direction (the Poisson limit is $\phi \to \infty$); put it on

$$\xi = \phi^{-1/2} \in [0, \infty), \qquad \mathrm{var}(y_i) = \mu_i\big(1 + \xi^2\mu_i\big),$$

so that ξ is the coefficient of variation of the gamma errors, $\xi = 0$ is exactly the Poisson model, and the parameter lives on a scale comparable to the β_j. Matching the prior distributions of the Poisson model,

$$\beta_j \sim \text{Cauchy}(0, 2.5) \;\; (j = 1, \ldots, k), \qquad \xi \sim \text{Cauchy}^{+}(0, 2.5),$$

and sample on $\zeta = \log\xi \in \mathbb{R}$ so that the Metropolis algorithm runs on an unconstrained space, which contributes the Jacobian $|d\xi/d\zeta| = \xi$.
The log posterior function takes the vector (β, ζ) together with X and y, sets $\xi = e^{\zeta}$, $\phi = \xi^{-2}$, $\eta = X\beta$, $\mu = e^{\eta}$, and returns

$$\begin{aligned}\log p(\beta, \zeta \mid X, y) = \sum_{i=1}^{n} \Big[&\log\Gamma(y_i + \phi) - \log\Gamma(\phi) - \log(y_i!) \\ &+ \phi\big(\log\phi - \log(\phi + \mu_i)\big) + y_i\big(\eta_i - \log(\mu_i + \phi)\big)\Big] \\ &- \sum_{j=1}^{k} \log\Big(1 + (\beta_j/2.5)^2\Big) - \log\Big(1 + (\xi/2.5)^2\Big) + \zeta,\end{aligned}$$

the last term being the Jacobian; in R this is one line of `lgamma` calls applied to the vectors `y`, `mu`, so the same Metropolis driver written in (a) fits this model unchanged once Σ is enlarged to $(k+1)\times(k+1)$.

Problem 16.7. Paired comparisons: consider the subset of the chess data in Table 16.3.
(a) Perform a simple exploratory analysis of the data to estimate the relative abilities of the players.
(b) Using some relatively simple (but reasonable) model, estimate the probability that player i wins if he plays White against player j, for each pair of players, (i, j).
(c) Fit the model described in Section 16.6 and use it to estimate these probabilities.
Table 16.3 gives results of games between eight of the 29 players of the 1988–1989 World Cup of chess, aggregated over all six tournaments. Rows index the player with the White pieces, columns the player with the Black pieces, and each entry is wins-losses-draws from the point of view of the White player; for example, playing White against Kasparov, Karpov had one win, no losses and one draw.

White pieces	Kar	Kas	Kor	Lju	Sei	Sho	Spa	Tal
Karpov		1-0-1	1-0-0	0-1-1	1-0-0	0-0-2	0-0-0	0-0-0
Kasparov	0-0-0		1-0-0	0-0-0	0-0-1	1-0-0	1-0-0	0-0-2
Korchnoi	0-0-1	0-2-0		0-0-0	0-1-0	0-0-2	0-0-1	0-0-0
Ljubojevic	0-1-0	0-1-1	0-0-2		0-1-0	0-0-1	0-0-2	0-0-1
Seirawan	0-1-1	0-0-1	1-1-0	0-2-0		0-0-0	0-0-0	1-0-0
Short	0-0-1	0-2-0	0-0-0	1-0-1	2-0-1		0-0-1	1-0-0
Spassky	0-1-0	0-0-2	0-0-1	0-0-0	0-0-1	0-0-1		0-0-0
Tal	0-0-2	0-0-0	0-0-3	0-0-0	0-0-1	0-0-0	0-0-1	

The model of Section 16.6, equation (16.15), gives for a game in which i moves first (has White) against j,

$$p_{ij1} = \Pr(i \text{ defeats } j \mid \theta) = e^{\alpha_i}/\Delta_{ij},$$
$$p_{ij2} = \Pr(j \text{ defeats } i \mid \theta) = e^{\alpha_j+\gamma}/\Delta_{ij},$$
$$p_{ij3} = \Pr(i \text{ draws with } j \mid \theta) = e^{\delta+\frac{1}{2}(\alpha_i+\alpha_j+\gamma)}/\Delta_{ij},$$

with $\Delta_{ij} = e^{\alpha_i} + e^{\alpha_j+\gamma} + e^{\delta+\frac{1}{2}(\alpha_i+\alpha_j+\gamma)}$, where α_i is the ability of player i, γ determines the relative advantage or disadvantage of moving first, and δ determines the probability of a draw.

Solution. (a) Score percentage, counting a draw as half a point and pooling the two colors:

Player	Games	Wins	Losses	Draws	Points	Score rate
Kasparov	17	8	1	8	12.0	0.706
Karpov	16	6	1	9	10.5	0.656
Short	17	4	3	10	9.0	0.529
Ljubojevic	16	3	4	9	7.5	0.469
Spassky	12	0	2	10	5.0	0.417
Tal	12	0	2	10	5.0	0.417
Seirawan	17	4	7	6	7.0	0.412
Korchnoi	17	1	6	10	6.0	0.353

Kasparov and Karpov are clearly the strongest of the eight, Korchnoi and Seirawan the weakest, and the middle is not separated. Three features of the subset matter for what follows. First, 36 of the 62 games (58%) were drawn, so the score rates are all pulled toward 1/2 and carry little information. Second, Spassky and Tal won no games at all and Korchnoi won one, so an unpenalized maximum likelihood fit of any Bradley-Terry model will be pushed toward the separation of Section 16.3; a weakly informative Cauchy$(0, 2.5)$ prior distribution on each ability is used below for exactly this reason. Finally, White won 12 games and Black 14: the first-move advantage, real over the full 789-game dataset, is invisible in this subset.

(b) A simple two-part model: take the probability of a draw to be a constant d, and, conditional on the game being decisive, use the Bradley-Terry model with a white-pieces bonus w,

$$\Pr(i \text{ wins} \mid \text{decisive}) = \text{logit}^{-1}(\alpha_i - \alpha_j + w),$$
$$\Pr(i \text{ wins}) = (1-d)\,\text{logit}^{-1}(\alpha_i - \alpha_j + w).$$

Estimate d by the overall draw rate, $\hat{d} = 36/62 = 0.58$, and fit the logistic regression to the 26 decisive games with $\sum_i \alpha_i = 0$ and Cauchy$(0, 2.5)$ priors on the α_i and on w (Metropolis, four chains, $\hat{R} \le 1.002$). The posterior means of the abilities are Kasparov 3.1, Karpov 2.8, Short 1.7, Ljubojevic 0.4, Seirawan -0.6, Korchnoi -1.6, Spassky -2.0, Tal -3.8, and $w = -0.69$ with 95% interval $[-2.09,\ 0.59]$, that is, no detectable advantage to moving first. The resulting $\Pr(i \text{ wins as White against } j)$, posterior means, rows $= i$ (White):

White / Black	Kar	Kas	Kor	Lju	Sei	Sho	Spa	Tal
Karpov		0.14	0.39	0.32	0.36	0.24	0.36	0.40
Kasparov	0.19		0.40	0.34	0.38	0.26	0.37	0.41
Korchnoi	0.01	0.01		0.06	0.10	0.03	0.17	0.26
Ljubojevic	0.05	0.04	0.29		0.23	0.09	0.27	0.35
Seirawan	0.02	0.02	0.23	0.09		0.05	0.22	0.32
Short	0.10	0.07	0.36	0.25	0.32		0.32	0.39
Spassky	0.03	0.02	0.18	0.09	0.13	0.05		0.26
Tal	0.01	0.01	0.10	0.04	0.06	0.02	0.12	

These are already implausible at the extremes, giving Tal a 1% chance of beating Karpov and Korchnoi a 1% chance of beating Kasparov, because the abilities are estimated from 26 decisive games and the zero-win players are dragged down as far as the Cauchy prior allows ($\mathrm{sd}(\alpha_{\mathrm{Tal}}) = 3.2$).

(c) Fit (16.15) directly as a multinomial likelihood in the counts $y_{ij} = (y_{ij1}, y_{ij2}, y_{ij3})$, which by (16.14) is equivalent to the Poisson generalized linear model (16.16) with the nuisance parameters A_{ij} and the offset $\log n_{ij}$; the multinomial form avoids carrying the 29×28 nuisance parameters. Parameters are $\alpha_1, \ldots, \alpha_8$ with $\sum_i \alpha_i = 0$ in place of the $\alpha_1 = 0$ of Section 16.6, together with γ and δ, all given Cauchy$(0, 2.5)$ prior distributions (Metropolis, four chains, acceptance 0.28, $\hat{R} \le 1.004$):

$$\gamma = 0.10\ (\pm 0.46), \quad 95\%\ [-0.80,\ 1.01],$$
$$\delta = 1.46\ (\pm 0.31), \quad 95\%\ [0.87,\ 2.10],$$

with abilities Kasparov 2.02 ($\pm$0.84), Karpov 1.55 ($\pm$0.81), Short 0.46 ($\pm$0.71), Ljubojevic -0.21 ($\pm$0.73), Spassky -0.63 ($\pm$0.85), Seirawan -0.84 ($\pm$0.72), Tal -0.83 ($\pm$0.85), Korchnoi -1.53 ($\pm$0.79). Since γ multiplies the second mover, the first-move advantage is $-\gamma$, estimated at -0.10 in log odds: nothing, as the raw 12-14 split already indicated. The estimated probabilities p_{ij1}, posterior means:

White / Black	Kar	Kas	Kor	Lju	Sei	Sho	Spa	Tal
Karpov		0.14	0.49	0.34	0.41	0.27	0.39	0.41
Kasparov	0.21		0.55	0.39	0.47	0.32	0.44	0.47
Korchnoi	0.03	0.02		0.09	0.12	0.06	0.11	0.13
Ljubojevic	0.07	0.05	0.29		0.22	0.12	0.21	0.22
Seirawan	0.05	0.04	0.23	0.13		0.09	0.16	0.17
Short	0.10	0.08	0.36	0.23	0.29		0.27	0.29
Spassky	0.06	0.04	0.25	0.14	0.19	0.10		0.19
Tal	0.05	0.04	0.23	0.13	0.17	0.09	0.16	

The two fits agree on the ordering but not on the extremes, and (16.15) is the more credible of the two. Because δ is estimated jointly with the abilities rather than fixed at the raw draw rate, drawn games now carry information about ability – a draw between i and j is evidence that α_i and α_j are close, through the term $\frac{1}{2}(\alpha_i + \alpha_j + \gamma)$ – so the ten draws of Spassky and of Tal place them in the middle of the field (-0.63 and -0.83) instead of at the bottom, and their posterior standard deviations fall from 3.0–3.2 to 0.85. The fitted draw probabilities range from 0.40 (Kasparov-Korchnoi, the most mismatched pair) to 0.64, averaging 0.58 against the observed 0.581, and the model correctly makes draws less likely the larger the ability gap. What remains unresolved is that 62 games among eight players cannot distinguish the middle six: every ability has a posterior standard deviation of about 0.8, so only the gap between the Kasparov-Karpov pair and the rest is established by these data.

16.2 Exercises 16.8–16.11

Problem 16.8. Iterative proportional fitting:
(a) Prove that the IPF algorithm increases the posterior density for γ at each step.
(b) Prove that the Bayesian IPF algorithm is in fact a Gibbs sampler for the parameters γ_j.
For reference, the setting of Section 16.7. The cells $i = 1, \ldots, n$ of a (possibly multiway) contingency table carry counts $y_i \mid \mu_i \sim \mathrm{Poisson}(\mu_i)$, independently, and the loglinear model constrains $\log \mu =$

$X\beta$ with X a known $n \times J$ matrix all of whose entries are zeros and ones. In multiplicative form $\gamma_j = \exp(\beta_j)$, so that

$$\mu_i = \prod_{j=1}^{J} \gamma_j^{x_{ij}}, \qquad \gamma_j > 0.$$

The prior distribution is the conjugate Dirichlet-like family (16.17), $p(\mu) \propto \prod_{i=1}^{n} \mu_i^{k_i - 1}$, supported on the loglinear surface, and

$$y_{j+} \;=\; \sum_{i=1}^{n} x_{ij}\,(y_i + k_i)$$

is the margin of the table corresponding to the jth column of X (assumed positive). A single IPF step alters one parameter,

$$\gamma_j^{\text{new}} \;=\; \frac{y_{j+}}{\sum_{i=1}^{n} x_{ij}\mu_i^{\text{old}}}\;\gamma_j^{\text{old}},$$

after which the expected counts are rescaled to stay on the surface,

$$\mu_i^{\text{new}} \;=\; \mu_i^{\text{old}} \left(\frac{\gamma_j^{\text{new}}}{\gamma_j^{\text{old}}} \right)^{x_{ij}}, \tag{16.18}$$

and the two steps are repeated cycling through $j = 1, \ldots, J$. The Bayesian IPF step replaces the deterministic update by

$$\gamma_j^{\text{new}} \;=\; \frac{A}{2y_{j+}}\;\frac{y_{j+}}{\sum_{i=1}^{n} x_{ij}\mu_i^{\text{old}}}\;\gamma_j^{\text{old}}, \qquad A \sim \chi^2_{2y_{j+}},$$

followed by the same rescaling (16.18).

Solution. Both parts are read off the single-coordinate conditional $(*)$ below: it is a gamma log density whose mode is the IPF step and whose draw is the Bayesian IPF step. Parameterize the loglinear surface by $\beta_j = \log \gamma_j$. The conjugate prior (16.17) acts exactly as Section 16.7 says it does, by supplementing the observed counts y_i with the prior cell counts k_i, so the target is

$$p(\beta \mid y) \;\propto\; \prod_{i=1}^{n} \mu_i(\beta)^{y_i + k_i} e^{-\mu_i(\beta)};$$

this is the reading that makes $y_{j+} = \sum_i x_{ij}(y_i + k_i)$ the sufficient statistic, as the next line shows. Because $\prod_i \mu_i^{y_i + k_i} = \prod_j \gamma_j^{y_{j+}}$,

$$\log p(\beta \mid y) \;=\; \sum_{j=1}^{J} y_{j+}\,\beta_j \;-\; \sum_{i=1}^{n} \mu_i(\beta) \;+\; \text{const.}$$

Fix j and hold β_{-j}. Since $x_{ij} \in \{0, 1\}$ we may write $\mu_i = \gamma_j^{x_{ij}} c_i$ with $c_i = \prod_{j' \neq j} \gamma_{j'}^{x_{ij'}}$ free of γ_j, whence $\sum_i \mu_i = c\,\gamma_j + \text{const}$ with

$$c \;=\; \sum_{i=1}^{n} x_{ij} c_i \;=\; \frac{\sum_{i=1}^{n} x_{ij}\mu_i^{\text{old}}}{\gamma_j^{\text{old}}} \;> 0,$$

and therefore the conditional log density of the single coordinate is

$$\log p(\beta_j \mid \beta_{-j}, y) \;=\; y_{j+}\,\beta_j \;-\; c\,e^{\beta_j} \;+\; \text{const.} \tag{$*$}$$

(a) The right side of $(*)$ has derivative $y_{j+} - c\,e^{\beta_j}$ and second derivative $-c\,e^{\beta_j} < 0$, so it is strictly concave with unique maximizer

$$\gamma_j \;=\; e^{\beta_j} \;=\; \frac{y_{j+}}{c} \;=\; \frac{y_{j+}}{\sum_i x_{ij}\mu_i^{\text{old}}}\;\gamma_j^{\text{old}},$$

which is precisely the IPF step; the rescaling (16.18) only propagates the identity $\mu_i = \prod_j \gamma_j^{x_{ij}}$ and changes no parameter. So IPF is exactly conditional maximization (Section 13.1) of $p(\beta \mid y)$ one coordinate at a time, whence

$$p(\beta^{\text{new}} \mid y) \;\geq\; p(\beta^{\text{old}} \mid y),$$

with equality if and only if γ_j^{old} already attains the conditional maximum, that is, if and only if the jth margin already fits, $\sum_i x_{ij}\mu_i^{\text{old}} = y_{j+}$. Strictness unless the step is null gives the assertion. The density being ascended is the one in the loglinear coordinates $\beta = \log\gamma$, which is what the exercise's phrase *posterior density for* γ must mean: on the γ scale the density carries the Jacobian $\prod_j \gamma_j^{-1}$, its jth conditional mode is at $(y_{j+}-1)/c$ rather than y_{j+}/c, and an IPF step started there strictly decreases it.

(b) Exponentiate $(*)$ and change variables to γ_j, using $d\beta_j = d\gamma_j/\gamma_j$:

$$p(\gamma_j \mid \gamma_{-j}, y) \;\propto\; \gamma_j^{y_{j+}-1} e^{-c\gamma_j}, \qquad \text{i.e.} \qquad \gamma_j \mid \gamma_{-j}, y \;\sim\; \text{Gamma}(y_{j+},\, c)$$

with c the rate. Now $A \sim \chi^2_{2y_{j+}}$ is Gamma$(y_{j+}, \frac{1}{2})$, so $A/2 \sim$ Gamma$(y_{j+}, 1)$ and, by the scaling property of the gamma family,

$$\frac{A}{2c} \;\sim\; \text{Gamma}(y_{j+},\, c).$$

The Bayesian IPF step is exactly this variable:

$$\gamma_j^{\text{new}} = \frac{y_{j+}\, A\, \gamma_j^{\text{old}}}{2y_{j+} \sum_i x_{ij}\mu_i^{\text{old}}} = \frac{A\, \gamma_j^{\text{old}}}{2\sum_i x_{ij}\mu_i^{\text{old}}} = \frac{A}{2c}.$$

So each Bayesian IPF step is an exact draw from the full conditional distribution of γ_j given γ_{-j} and y, and cycling $j = 1, \ldots, J$ (with (16.18) as bookkeeping for μ) is a systematic-scan Gibbs sampler with stationary distribution $p(\gamma \mid y)$ — equivalently $p(\mu \mid y)$ on the loglinear surface, as claimed in Section 16.7.

Problem 16.9. Improper prior distributions and proper posterior distributions: consider the hierarchical model for the meta-analysis example in Section 16.6.

(a) Show that, for any value of ρ_{12}, the posterior distribution of all the remaining parameters is proper, conditional on ρ_{12}.

(b) Show that the posterior distribution of all the parameters, including ρ_{12}, is proper.

The model of Section 16.6 is as follows. For studies $j = 1, \ldots, J$ with $J = 22$ and treatments $i = 0$ (control), 1 (treated),

$$y_{ij} \mid n_{ij}, p_{ij} \;\sim\; \text{Bin}(n_{ij}, p_{ij})$$

independently, and the $2J$ probabilities are reparameterized by (16.13),

$$\beta_{1j} = \tfrac{1}{2}\big(\text{logit}(p_{0j}) + \text{logit}(p_{1j})\big), \qquad \beta_{2j} = \text{logit}(p_{1j}) - \text{logit}(p_{0j}),$$

with the J exchangeable pairs given a bivariate normal population distribution,

$$\begin{pmatrix}\beta_{1j}\\ \beta_{2j}\end{pmatrix} \,\Big|\, \alpha, \Lambda \;\sim\; \text{N}\left(\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}, \Lambda\right), \qquad j = 1, \ldots, J,$$

independently. The hyperprior distribution is uniform on α_1, α_2 and, independently, uniform on the two variances and the correlation:

$$p(\alpha_1, \alpha_2, \Lambda_{11}, \Lambda_{22}, \rho_{12}) \;\propto\; 1 \quad \text{on} \quad \mathbb{R}^2 \times (0,\infty)^2 \times (-1, 1),$$

where $\rho_{12} = \Lambda_{12}/(\Lambda_{11}\Lambda_{22})^{1/2}$. The data are the 22 two-by-two tables of Table 5.4 (beta-blockers after myocardial infarction, Yusuf et al., 1985), each entry being deaths/total:

j	control	treated	j	control	treated
1	3/39	3/38	12	47/266	45/263
2	14/116	7/114	13	16/293	9/291
3	11/93	5/69	14	45/883	57/858
4	127/1520	102/1533	15	31/147	25/154
5	27/365	28/355	16	38/213	33/207
6	6/52	4/59	17	12/122	28/251
7	152/939	98/945	18	6/154	8/151
8	48/471	60/632	19	3/134	6/174
9	37/282	25/278	20	40/218	32/209
10	188/1921	138/1916	21	43/364	27/391
11	52/583	64/873	22	39/674	22/680

Solution. Proper in both cases, and the only facts about the data that are used are that every cell of Table 5.4 satisfies $0 < y_{ij} < n_{ij}$ (the smallest death count is 3) and that $J = 22 \geq 5$.
Write $\theta_j = (\beta_{1j}, \beta_{2j})'$, let $L_j(\theta_j) = \prod_{i=0,1} p_{ij}^{y_{ij}} (1-p_{ij})^{n_{ij}-y_{ij}}$ be the jth likelihood factor, and set

$$\bar\theta = J^{-1}\textstyle\sum_j \theta_j, \qquad S = \sum_{j=1}^{J} (\theta_j - \bar\theta)(\theta_j - \bar\theta)'.$$

A dominating bound on the likelihood. For one binomial factor with $\operatorname{logit}(p) = t$,

$$p^y (1-p)^{n-y} = \frac{e^{ty}}{(1+e^t)^n} \;\leq\; \min\bigl(e^{ty}, e^{-t(n-y)}\bigr) \;\leq\; e^{-\delta|t|}, \qquad \delta = \min(y, n-y),$$

so with $t_0 = \beta_{1j} - \beta_{2j}/2$, $t_1 = \beta_{1j} + \beta_{2j}/2$ and $\delta_j = \min_i \min(y_{ij}, n_{ij} - y_{ij}) \geq 3$,

$$\begin{aligned} |t_0| + |t_1| &= \max\bigl(|t_0 + t_1|,\ |t_1 - t_0|\bigr) \;\geq\; \tfrac12\bigl(|t_0 + t_1| + |t_1 - t_0|\bigr) \\ &= |\beta_{1j}| + \tfrac12 |\beta_{2j}|, \end{aligned}$$

so that, with $f_j(u) = e^{-\delta_j|u|}$ and $g_j(v) = e^{-\delta_j|v|/2}$,

$$L_j(\theta_j) \;\leq\; f_j(\beta_{1j})\, g_j(\beta_{2j}) :$$

the two blocks $\beta_1 = (\beta_{11}, \ldots, \beta_{1J})$ and β_2 decouple, and each coordinate decays exponentially. Nothing else about the data is used.
Integrating out α. By the usual decomposition

$$\sum_j (\theta_j - \alpha)'\Lambda^{-1}(\theta_j - \alpha) = \operatorname{tr}(\Lambda^{-1}S) + J(\bar\theta - \alpha)'\Lambda^{-1}(\bar\theta - \alpha),$$

we have

$$\int_{\mathbb{R}^2} \prod_{j=1}^{J} \mathrm{N}(\theta_j \mid \alpha, \Lambda)\, d\alpha = \frac{(2\pi)^{-(J-1)}}{J}\, |\Lambda|^{-(J-1)/2} \exp\bigl(-\tfrac12 \operatorname{tr}(\Lambda^{-1}S)\bigr)\,.$$

(a) Fix $\rho = \rho_{12} \in (-1, 1)$ and factor $\Lambda = DRD$ with $D = \operatorname{diag}(\Lambda_{11}^{1/2}, \Lambda_{22}^{1/2})$ and

$$R = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Then $|\Lambda| = (1-\rho^2)\Lambda_{11}\Lambda_{22}$, and since R^{-1} has smallest eigenvalue $(1+|\rho|)^{-1} \geq \frac12$ while $D^{-1}SD^{-1} \succeq 0$,

$$\operatorname{tr}(\Lambda^{-1}S) = \operatorname{tr}\bigl(R^{-1}D^{-1}SD^{-1}\bigr) \;\geq\; \tfrac12 \left(\frac{S_{11}}{\Lambda_{11}} + \frac{S_{22}}{\Lambda_{22}}\right).$$

Using $\int_0^\infty \lambda^{-a} e^{-b/\lambda}\, d\lambda = \Gamma(a-1) b^{-(a-1)}$ $(a > 1,\, b > 0)$ with $a = (J-1)/2$ and $b = S_{kk}/4$,

$$\int_0^\infty \int_0^\infty |\Lambda|^{-\frac{J-1}{2}} e^{-\frac12 \operatorname{tr}(\Lambda^{-1}S)}\, d\Lambda_{11} d\Lambda_{22} \;\leq\; \frac{4^{J-3}\,\Gamma\bigl(\frac{J-3}{2}\bigr)^2}{(1-\rho^2)^{(J-1)/2}\,(S_{11}S_{22})^{(J-3)/2}},$$

finite because $J = 22 > 3$. It remains to check

$$T \;=\; \int_{\mathbb{R}^{2J}} \prod_{j=1}^{J} L_j(\theta_j)\,(S_{11}S_{22})^{-(J-3)/2}\,d\theta \;<\; \infty,$$

and by the dominating bound this splits as $T \le U(f)\,U(g)$ with $U(h) = \int_{\mathbb{R}^J} \prod_j h_j(u_j)\,S(u)^{-(J-3)/2}\,du$ and $S(u) = \sum_j (u_j - \bar u)^2$. Decompose $u = m\mathbf{1} + w$ with $w \perp \mathbf{1}$, so $S(u) = |w|^2$, $|u|^2 = Jm^2 + |w|^2$, and with $\delta = \min_j \delta_j \ge 3$,

$$\sum_j \delta_j |u_j| \;\ge\; \delta|u|_1 \;\ge\; \delta|u| \;\ge\; \tfrac{\delta}{2}\big(\sqrt{J}\,|m| + |w|\big).$$

Hence, in polar coordinates on the $(J-1)$-dimensional space $\mathbf{1}^\perp$,

$$U(f) \;\le\; \int_{\mathbb{R}} e^{-\frac{\delta\sqrt{J}}{2}|m|}dm \cdot \omega_{J-2} \int_0^\infty e^{-\frac{\delta}{2}r}\, r^{(J-2)-(J-3)}\,dr \;<\; \infty,$$

the radial exponent being $(J-2)-(J-3) = 1 > -1$; the same bound with $\delta/2$ in place of δ gives $U(g) < \infty$. So the conditional posterior given ρ_{12} is proper.
(b) Now ρ_{12} must be integrated as well, so the factor $(1-\rho^2)^{-(J-1)/2}$ above is no longer allowed. Change measure on the cone of positive definite 2×2 matrices: at fixed $\Lambda_{11}, \Lambda_{22}$ we have $d\Lambda_{12} = (\Lambda_{11}\Lambda_{22})^{1/2} d\rho_{12}$, so

$$d\Lambda_{11}d\Lambda_{22}d\rho_{12} = (\Lambda_{11}\Lambda_{22})^{-1/2}\,d\Lambda_{11}d\Lambda_{12}d\Lambda_{22}, \qquad (\Lambda_{11}\Lambda_{22})^{-1/2} \le |\Lambda|^{-1/2},$$

the inequality because $|\Lambda| = \Lambda_{11}\Lambda_{22}(1-\rho_{12}^2)$. Therefore

$$\int |\Lambda|^{-\frac{J-1}{2}} e^{-\frac{1}{2}\mathrm{tr}(\Lambda^{-1}S)} d\Lambda_{11}d\Lambda_{22}d\rho_{12} \;\le\; \int_{\Lambda\succ 0} |\Lambda|^{-\frac{J}{2}} e^{-\frac{1}{2}\mathrm{tr}(\Lambda^{-1}S)}\,d\Lambda,$$

which is the inverse-Wishart normalizing integral with $d = 2$ and $(\nu + d + 1)/2 = J/2$, that is $\nu = J - 3 = 19$:

$$\int_{\Lambda\succ 0} |\Lambda|^{-\frac{\nu+3}{2}} e^{-\frac{1}{2}\mathrm{tr}(S\Lambda^{-1})} d\Lambda = 2^\nu\,\Gamma_2\big(\tfrac{\nu}{2}\big)\,|S|^{-\nu/2},$$

finite since $\nu = 19 > d - 1 = 1$ and $S \succ 0$. So it suffices that

$$T' \;=\; \int_{\mathbb{R}^{2J}} \prod_{j=1}^{J} L_j(\theta_j)\,|S|^{-(J-3)/2}\,d\theta \;<\; \infty.$$

Factor the determinant as $|S| = S_{11}\,R$, where

$$R \;=\; S_{22} - \frac{S_{12}^2}{S_{11}} \;=\; \min_{t\in\mathbb{R}} \sum_{j=1}^{J} \big[(\beta_{2j} - \bar\beta_2) - t(\beta_{1j} - \bar\beta_1)\big]^2 \;=\; |P_{\beta_1}\beta_2|^2,$$

with P_{β_1} the orthogonal projection of $\mathbb{R}^J$ onto the orthogonal complement of $\mathrm{span}\{\mathbf{1}, \beta_1\}$, a subspace of dimension $J-2$ whenever $S_{11} > 0$. Using the dominating bound again and splitting $\beta_2 = u + v$ with $u = P_{\beta_1}\beta_2$ (dimension $J-2$) and v in the complementary 2-dimensional space, so that $|\beta_2|_1 \ge |\beta_2| \ge (|u| + |v|)/\sqrt{2}$,

$$\int_{\mathbb{R}^J} \prod_j g_j(\beta_{2j})\,|P_{\beta_1}\beta_2|^{-(J-3)} d\beta_2 \;\le\; \omega_{J-3} \int_0^\infty e^{-cr} r^{(J-3)-(J-3)} dr \cdot \int_{\mathbb{R}^2} e^{-c|v|} dv,$$

with $c = \delta/(2\sqrt{2})$; both factors are finite and the bound does not depend on β_1. The outer integral is $\int_{\mathbb{R}^J} \prod_j f_j(\beta_{1j}) S_{11}^{-(J-3)/2} d\beta_1 = U(f)$, shown finite in (a). Hence $T' < \infty$ and the joint posterior distribution of $(\beta, \alpha, \Lambda_{11}, \Lambda_{22}, \rho_{12})$ is proper.

Problem 16.10. Variational Bayes for probit regression: Set up and program variational Bayes for a probit regression with two coefficients (that is, $\Pr(y_i = 1) = \Phi(a + bx_i)$, for $i = 1, \dots, n$), using the latent-data formulation (so that $z_i \sim \mathrm{N}(a + bx_i, 1)$ and $y_i = 1$ if $z_i > 0$ and 0 otherwise):
(a) Write the log posterior density (up to an arbitrary constant), $p(a, b, z \mid y)$.
(b) Assuming a variational approximation g that is independent in its $n + 2$ dimensions, determine the functional form of each of the factors in g.
(c) Write the steps of the variational Bayes algorithm and program them in R.
(Take the prior distribution for (a, b) to be uniform, as in the latent-data formulation (16.3) of Section 16.1.)

Solution. (a) Because y is a deterministic function of z, the factor $p(y \mid z)$ is an indicator, so with $s_i = 2y_i - 1$ and a uniform prior on (a, b),

$$\log p(a, b, z \mid y) \;=\; -\tfrac{1}{2}\sum_{i=1}^{n}(z_i - a - bx_i)^2 \;+\; \sum_{i=1}^{n}\log \mathbf{1}\{s_i z_i > 0\} \;+\; \text{const},$$

that is, a spherical Gaussian in the $n + 2$ variables restricted to the orthant $\{s_i z_i > 0,\ i = 1, \dots, n\}$ and $-\infty$ outside it.
(b) Take $g(a, b, z) = g_a(a)\, g_b(b) \prod_{i=1}^{n} g_i(z_i)$ and apply the coordinate ascent step (13.26)–(13.27): each factor is proportional to exp of the expectation of the log joint density over the remaining factors. Since the log joint is quadratic apart from the indicators, collecting terms gives

$$\begin{aligned}
\log g_a(a) &= -\tfrac{n}{2}a^2 + a\sum_i \big(\mathrm{E}z_i - x_i\mu_b\big) + \text{const},\\
\log g_b(b) &= -\tfrac{1}{2}b^2\sum_i x_i^2 + b\sum_i x_i\big(\mathrm{E}z_i - \mu_a\big) + \text{const},\\
\log g_i(z_i) &= -\tfrac{1}{2}(z_i - m_i)^2 + \log \mathbf{1}\{s_i z_i > 0\} + \text{const},
\end{aligned}$$

where $\mu_a = \mathrm{E}_g a$, $\mu_b = \mathrm{E}_g b$ and $m_i = \mu_a + \mu_b x_i$. Hence the factors are

$$g_a = \mathrm{N}\big(\mu_a, \tfrac{1}{n}\big), \qquad g_b = \mathrm{N}\Big(\mu_b, \tfrac{1}{\sum_i x_i^2}\Big), \qquad g_i = \mathrm{N}(m_i, 1) \text{ truncated to } s_i z_i > 0,$$

the two normal factors and the n truncated-normal factors; note that the variational variances $1/n$ and $1/\sum_i x_i^2$ are fixed constants, unchanged by the iterations, and that the required first moment of a truncated normal is

$$\mathrm{E}_{g_i} z_i \;=\; m_i + \lambda_i, \qquad \lambda_i \;=\; s_i\,\frac{\phi(m_i)}{\Phi(s_i m_i)},$$

with $\mathrm{var}_{g_i}(z_i) = 1 - \lambda_i(\lambda_i + m_i)$.
(c) The algorithm is three assignments cycled to convergence. Set $\sigma_a^2 = 1/n$ and $\sigma_b^2 = 1/\sum_i x_i^2$ once, start at $\mu_a = \mu_b = 0$, and repeat:

$$\begin{aligned}
&\text{(i)} \quad m_i \leftarrow \mu_a + \mu_b x_i, \quad \lambda_i \leftarrow s_i\phi(m_i)/\Phi(s_i m_i), \quad \widehat{z}_i \leftarrow m_i + \lambda_i,\\
&\text{(ii)} \quad \mu_a \leftarrow \frac{1}{n}\sum_i (\widehat{z}_i - \mu_b x_i),\\
&\text{(iii)} \quad \mu_b \leftarrow \Big(\sum_i x_i(\widehat{z}_i - \mu_a)\Big)\Big/\sum_i x_i^2,
\end{aligned}$$

until μ_a, μ_b stop changing. (In R: one line each with `dnorm`, `pnorm` and `sum`; the n truncated factors never need to be stored, only $\widehat{z}$.)
The fixed point is identified exactly. Writing $X = (\mathbf{1}, x)$ and $\mu = (\mu_a, \mu_b)'$, steps (ii)–(iii) say $X'(\widehat{z} - X\mu) = 0$, and $\widehat{z} - X\mu = \lambda$, so at convergence

$$\sum_{i=1}^{n} s_i\,\frac{\phi(a + bx_i)}{\Phi(s_i(a + bx_i))}\begin{pmatrix}1\\ x_i\end{pmatrix} = 0 \qquad \text{at } (a, b) = (\mu_a, \mu_b),$$

which is precisely the probit score equation. So the variational means coincide with the probit maximum likelihood estimate, which under the uniform prior is the posterior mode.
Run on the arsenic data of Exercise 16.11 ($n = 3020$, x_i the arsenic concentration, $\sum_i x_i^2 = 11993.4$), starting from $(0, 0)$, the iteration converges (change below 10^{-10}) in about 110 sweeps to

$$\mu_a = -0.178415, \quad \sigma_a = n^{-1/2} = 0.018197, \qquad \mu_b = 0.227228, \quad \sigma_b = 0.009131,$$

agreeing with the probit mode $(-0.17841528, 0.22722796)$ to eight decimal places; the fitted latent factors have $\overline{\widehat{z}} = 0.198$ and average variance 0.385.

Problem 16.11. HMC for probit regression: Millions of people in rural Bangladesh are exposed to dangerous levels of arsenic in their drinking water which they get from home wells. Several years ago a survey was conducted in a small area of Bangladesh to see if people with high arsenic levels would be willing to switch to a neighbor's well. File `wells.dat` has the data; all you need here are the variables `switch` (1 if the respondent said he or she would switch, 0 otherwise) and `arsenic` (the concentration in the respondent's home well, with anything over 0.5 considered dangerous). Apply the probit model described in Exercise 15.10 to predict the probability of switching given arsenic level. The goal is inference for the coefficients a and b.
(a) Program Hamiltonian Monte Carlo for the probit model, again using the latent-variable formulation (so you are jumping in a space of $n+2$ dimensions). Tune the algorithm and run to approximate convergence. Or, if you want less of a challenge, you can instead take the (relatively) easy way out and program Metropolis for this model and data.
(b) Check your results by running Stan.
(c) Compare to the variational Bayes inferences for a and b from Exercise 16.10.
The model is $\Pr(\texttt{switch}_i = 1) = \Phi(a + b\,\texttt{arsenic}_i)$ in the latent-data form (16.3), with a uniform prior distribution on (a, b). The file has $n = 3020$ records; the relevant summaries are:

quantity	value
n	3020
mean of `switch`	0.575
mean of `arsenic`	1.657
sd of `arsenic`	1.107
range of `arsenic`	0.51–9.65
$\sum_i \texttt{arsenic}_i$	5003.9
$\sum_i \texttt{arsenic}_i^2$	11993.4

Solution. $\mathrm{E}(a \mid y) = -0.1787$ with posterior sd 0.0427, and $\mathrm{E}(b \mid y) = 0.2275$ with posterior sd 0.0225, the two being strongly negatively correlated (-0.84). (Chapter 15 has only six exercises, so the cross-reference to Exercise 15.10 is an erratum; the model intended is the two-coefficient latent-data probit of Exercise 16.10, which is what is fitted here.)
(a) *Hamiltonian Monte Carlo in $n+2 = 3022$ dimensions.* From Exercise 16.10(a) the potential energy is

$$U(a, b, z) = \tfrac{1}{2} \sum_{i=1}^{n} r_i^2, \qquad r_i = z_i - a - b x_i,$$

on the orthant $\{s_i z_i > 0\}$ ($s_i = 2y_i - 1$) and $+\infty$ outside, with gradient

$$\frac{\partial U}{\partial a} = -\sum_i r_i, \qquad \frac{\partial U}{\partial b} = -\sum_i x_i r_i, \qquad \frac{\partial U}{\partial z_i} = r_i.$$

The hard walls $z_i = 0$ are handled by elastic reflection rather than by rejection: after each leapfrog position update, any coordinate that has crossed into the forbidden half-line is mirrored,

$$z_i \leftarrow -z_i, \qquad \varrho_i \leftarrow -\varrho_i \qquad \text{whenever } s_i z_i < 0.$$

For a free particle this *is* the exact bounce (a particle starting at $z_i > 0$ with displacement v landing at $z_i + v < 0$ hits the wall and ends at $-(z_i + v)$), one flip always suffices because each constraint is a single

half-line, and the map $(z_i, \varrho_i) \mapsto (-z_i, -\varrho_i)$ is a volume-preserving involution satisfying $NTNT = \mathrm{id}$ with N the momentum negation, so the leapfrog remains reversible and volume preserving and the Metropolis correction leaves $p(a, b, z \mid y)$ exactly invariant.

Tuning: a diagonal mass matrix $M = \mathrm{diag}(\sigma_a^{-2}, \sigma_b^{-2}, 1, \ldots, 1)$ with $\sigma_a = 0.043$, $\sigma_b = 0.023$ from a short pilot run (the z coordinates already have scale 1), and $\epsilon L \approx 1.2$ held fixed so that the (a, b) block moves about one posterior sd per trajectory. Measured acceptance rates over 300 iterations:

ϵ	0.02	0.04	0.06	0.08	0.10	0.14	0.20
L	60	30	20	15	12	9	6
acceptance	0.957	0.907	0.847	0.713	0.677	0.567	0.397

Take $\epsilon = 0.08$, $L = 15$, with ϵ jittered by $\pm 15\%$; eight chains from dispersed starts, 9000 iterations each with the first 1000 discarded (64,000 saved draws), acceptance rate 0.745:

$$a: \quad \text{mean} \ -0.17865, \quad \text{sd } 0.04318, \quad \widehat{R} = 1.0003, \quad n_{\text{eff}} = 8600,$$
$$b: \quad \text{mean} \ \ 0.22756, \quad \text{sd } 0.02258, \quad \widehat{R} = 1.0003, \quad n_{\text{eff}} = 8100,$$

with $\mathrm{corr}(a, b) = -0.841$. The scale reductions are well below the 1.1 criterion of Section 11.4, so the chains have reached approximate convergence.

(b) Stan was not available in the environment used here, so in its place the answer is checked against two references that are at least as stringent, and this substitution is flagged rather than glossed. First, z can be integrated out analytically, leaving a two-dimensional marginal posterior that is available by quadrature to essentially arbitrary accuracy:

$$p(a, b \mid y) \ \propto \ \prod_{i=1}^{n} \Phi\big(s_i(a + b x_i)\big).$$

Evaluating this on grids of 801^2, 1201^2 and 1601^2 points spanning ± 8, ± 12 and ± 16 posterior sd about the mode gives the identical answer to five decimal places. Second, the Albert and Chib (1993) Gibbs sampler, which also moves in $n + 2$ dimensions (draw $z_i \mid a, b$ from truncated normals, then $(a, b) \mid z$ from the normal linear regression), was run for 12,000 iterations discarding 2000:

quantity	HMC	quadrature	Gibbs
$\mathrm{E}(a)$	-0.17865	-0.17869	-0.17952
$\mathrm{sd}(a)$	0.04318	0.04270	0.04327
$\mathrm{E}(b)$	0.22756	0.22746	0.22796
$\mathrm{sd}(b)$	0.02258	0.02251	0.02284
$\mathrm{corr}(a, b)$	-0.841	-0.839	-0.838

The HMC standard deviations exceed the exact ones by 1.5 and 0.4 Monte Carlo standard errors respectively, taking $\mathrm{se}(\widehat{\mathrm{sd}}) \approx \mathrm{sd}/\sqrt{2n_{\text{eff}}}$, which equals 0.00033 and 0.00018 here; so the three answers agree. The posterior is very nearly normal: the curvature at the mode $(-0.178415, 0.227228)$ gives normal-approximation standard deviations 0.042697 and 0.022506, matching the quadrature values to three digits.

(c) The variational answer of Exercise 16.10 gets the location right and the spread badly wrong:

quantity	variational Bayes	HMC	exact quadrature
$\mathrm{E}(a)$	-0.17842	-0.17865	-0.17869
$\mathrm{sd}(a)$	0.01820	0.04318	0.04270
$\mathrm{E}(b)$	0.22723	0.22756	0.22746
$\mathrm{sd}(b)$	0.00913	0.02258	0.02251
$\mathrm{corr}(a, b)$	0 by construction	-0.841	-0.839

The variational means are the posterior mode exactly (Exercise 16.10), which for this nearly normal posterior is within 0.02 posterior sd of the posterior mean. The standard deviations are too small by factors 2.35 and 2.47, and the deficit factors exactly as the two independence assumptions in g:

$$\underbrace{0.04270}_{\text{exact}} \ \longrightarrow \ \underbrace{0.03275}_{z \text{ treated as known}} \ \longrightarrow \ \underbrace{0.01820}_{a \perp b \text{ imposed}},$$

and likewise $0.02251 \to 0.01643 \to 0.00913$ for b. The middle column is $\mathrm{diag}\,(X'X)^{-1/2}$ with $X = (\mathbf{1}, x)$: factorizing g between (a, b) and z replaces the posterior variance by the complete-data variance, as if the

latent z_i were observed. The last step multiplies by $(1-r^2)^{1/2} = 0.5556$ where $r = \sum_i x_i/(n\sum_i x_i^2)^{1/2} = 0.8315$ is the design correlation, since mean-field precisions are the *diagonal* of $X'X$, namely n and $\sum_i x_i^2$; this is the familiar understatement of variance by variational Bayes when the target has strongly correlated components (Section 13.7).

17 Models for Robust Inference

Contents

17.1 Exercises 17.1–17.7

Problem 17.1. Prior distributions and shrinkage: in the educational testing experiments, suppose we think that most coaching programs are almost useless, but some are strongly effective; a corresponding population distribution for the school effects is a mixture, with most of the mass near zero but some mass extending far in the positive direction; for example,

$$p(\omega_1, \ldots, \omega_8) = \prod_{j=1}^{8} \left[\lambda_1 \mathrm{N}(\omega_j \,|\, \mu_1, \tau_1^2) + \lambda_2 \mathrm{N}(\omega_j \,|\, \mu_2, \tau_2^2)\right].$$

All these parameters could be estimated from the data (as long as we restrict the parameter space, for example by setting $\mu_1 > \mu_2$), but to fix ideas, suppose that $\mu_1 = 0$, $\tau_1 = 10$, $\mu_2 = 15$, $\tau_2 = 25$, $\lambda_1 = 0.9$, and $\lambda_2 = 0.1$.
(a) Compute the posterior distribution of $(\omega_1, \ldots, \omega_8)$ under this model for the data in Table 5.2.
(b) Graph the posterior distribution for ω_8 under this model for $y_8 = 0$, 25, 50, and 100, with the same standard deviation σ_8 as given in Table 5.2. Describe qualitatively the effect of the two-component mixture prior distribution.
Table 5.2 (page 120) gives the observed effects y_j of coaching on SAT-V scores in eight schools, together with their sampling standard errors σ_j; the sampling model is $y_j \,|\, \omega_j \sim \mathrm{N}(\omega_j, \sigma_j^2)$, independently across schools:

School	y_j	σ_j
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Solution. The eight posteriors are independent two-component normal mixtures, with the same component precisions as in the conjugate case but reweighted by the marginal likelihood of each component. Because the hyperparameters are fixed, the prior factors over j and so does the likelihood; hence $p(\omega \,|\, y) = \prod_j p(\omega_j \,|\, y_j)$ with

$$p(\omega_j \,|\, y_j) \propto \mathrm{N}(y_j \,|\, \omega_j, \sigma_j^2) \sum_{k=1}^{2} \lambda_k \mathrm{N}(\omega_j \,|\, \mu_k, \tau_k^2).$$

Apply the identity behind (2.11)-(2.12) to each term,

$$\mathrm{N}(y_j \,|\, \omega_j, \sigma_j^2)\, \mathrm{N}(\omega_j \,|\, \mu_k, \tau_k^2) = \mathrm{N}(y_j \,|\, \mu_k, \tau_k^2 + \sigma_j^2)\, \mathrm{N}(\omega_j \,|\, m_{jk}, v_{jk}),$$

$$v_{jk} = \left(\tfrac{1}{\tau_k^2} + \tfrac{1}{\sigma_j^2}\right)^{-1},$$

$$m_{jk} = v_{jk} \left(\tfrac{\mu_k}{\tau_k^2} + \tfrac{y_j}{\sigma_j^2}\right),$$

so the posterior is again a mixture of the two conjugate normals, with weights

$$\tilde{\lambda}_{jk} \;=\; \frac{\lambda_k \mathrm{N}(y_j \,|\, \mu_k, \tau_k^2 + \sigma_j^2)}{\sum_l \lambda_l \mathrm{N}(y_j \,|\, \mu_l, \tau_l^2 + \sigma_j^2)}.$$

(a) Evaluating these three formulas for Table 5.2 gives the exact posterior for each school; $\tilde{\lambda}_{j2}$ is the posterior probability that school j belongs to the "strongly effective" component.

School	$\tilde{\lambda}_{j2}$	m_{j1}	$\sqrt{v_{j1}}$	m_{j2}	$\sqrt{v_{j2}}$	$\mathrm{E}(\omega_j\|y)$	$\mathrm{sd}(\omega_j\|y)$
A	0.172	8.62	8.32	24.56	12.86	11.36	11.04
B	0.062	4.00	7.07	8.97	9.28	4.31	7.33
C	0.056	-0.84	8.48	2.23	13.48	-0.67	8.86
D	0.061	3.17	7.40	8.30	10.07	3.48	7.69
E	0.045	-0.55	6.69	0.84	8.47	-0.49	6.79
F	0.050	0.45	7.40	3.27	10.07	0.59	7.58
G	0.115	9.00	7.07	17.59	9.28	9.99	7.85
H	0.081	2.83	8.74	13.02	14.61	3.65	9.75

The posterior means (last column) are computed as $\sum_k \tilde{\lambda}_{jk} m_{jk}$ and the standard deviations from $\sum_k \tilde{\lambda}_{jk}(v_{jk} + m_{jk}^2) - (\sum_k \tilde{\lambda}_{jk} m_{jk})^2$. None of the eight observations is extreme enough to activate the second component: every $\tilde{\lambda}_{j2}$ is below 0.18, so the shrinkage is essentially that of the $\mathrm{N}(0, 10^2)$ population distribution. Because that distribution is centered at a fixed 0 rather than at a hyperparameter estimated near 8, the low schools land near zero instead of near the common level of Table 5.3 (page 123): C, E and F move from posterior medians of 7, 6, 7 of 7, 5, 6 there to -0.7, -0.5, 0.6 here.
(The 2e solutions manual reports much larger second-component weights (0.30 for school A rather than 0.17); its code computes $\mathrm{N}(y_j \mid \mu_k, \sigma_j^2 + \tau_k)$ instead of $\mathrm{N}(y_j \mid \mu_k, \sigma_j^2 + \tau_k^2)$, and the weights above are the ones the displayed formula gives.)
(b) With $\sigma_8 = 18$ fixed, the two component posteriors have $\sqrt{v_{81}} = 8.74$, $m_{81} = 0.236\, y_8$ and $\sqrt{v_{82}} = 14.61$, $m_{82} = 5.12 + 0.659\, y_8$, and the weight on the second component is

y_8	$\tilde{\lambda}_{82}$	m_{81}	m_{82}	$\mathrm{E}(\omega_8\|y)$	$\mathrm{sd}(\omega_8\|y)$	$\Pr(\omega_8 > 25\|y)$
0	0.062	0.00	5.12	0.32	9.30	0.007
25	0.128	5.90	21.59	7.91	11.02	0.065
50	0.426	11.79	38.05	22.98	17.42	0.385
100	0.995	23.59	70.98	70.76	14.93	0.997

The four densities are graphed in `figs/bda3-ch17-mixture-prior-school8.svg`. Qualitatively, the mixture prior distribution makes the shrinkage nonlinear in y_8. For $y_8 = 0$ and 25 the first component dominates and ω_8 is shrunk almost to the narrow component's mean, $0.24\, y_8$ — far more severely than the single $\mathrm{N}(0, 25^2)$ prior, whose shrinkage factor is 0.66. At $y_8 = 50$ the components are nearly balanced ($\tilde{\lambda}_{82} = 0.43$): the density is unimodal at 12.7 but strongly right-skewed, with a long shoulder past 40, and $\mathrm{sd}(\omega_8 \mid y) = 17.4$ exceeds either component's, reflecting uncertainty about which kind of program school H runs. By $y_8 = 100$ the first component is untenable ($\tilde{\lambda}_{82} = 0.995$) and only the wide component's mild shrinkage survives, $\mathrm{E}(\omega_8 \mid y) = 70.8$. An extreme observation is thus absorbed by the long tail instead of inflating the population variance, and the other seven schools' estimates are untouched by it.

Problem 17.2. Poisson and negative binomial distributions: as part of their analysis of the Federalist papers, Mosteller and Wallace (1964) recorded the frequency of use of various words in selected articles by Alexander Hamilton and James Madison. The articles were divided into 247 blocks of about 200 words each, and the number of instances of various words in each block were recorded. Table 17.2 displays the results for the word 'may.'
(a) Fit the Poisson model to these data, with different parameters for each author and a noninformative prior distribution. Plot the posterior density of the Poisson mean parameter for each author.
(b) Fit the negative binomial model to these data with different parameters for each author. What is a reasonable noninformative prior distribution to use? For each author, make a contour plot of the posterior density of the two parameters and a scatterplot of the posterior simulations.
Table 17.2 Observed distribution of the word 'may' in papers of Hamilton and Madison, from Mosteller and Wallace (1964). Out of the 247 blocks of Hamilton's text studied, 128 had no instances of 'may,' 67 had one instance of 'may,' and so forth, and similarly for Madison.

Number of occurrences in a block	0	1	2	3	4	5	6	> 6
Number of blocks (Hamilton)	128	67	32	14	4	1	1	0
Number of blocks (Madison)	156	63	29	8	4	1	1	0

Solution. (a) With $p(\lambda) \propto \lambda^{-1}$ the posterior is $\lambda \,|\, y \sim \text{Gamma}(\sum_i y_i,\, n)$, by (2.15) with the limiting prior Gamma$(0, 0)$.
Table 17.2 gives $n_H = 247$ blocks with $\sum_i y_i = 200$ and $n_M = 262$ blocks with $\sum_i y_i = 172$ (the printed "247 blocks" describes Hamilton only; Madison's column sums to 262), so

$$\lambda_H \,|\, y \sim \text{Gamma}(200, 247),$$
$$\lambda_M \,|\, y \sim \text{Gamma}(172, 262).$$

Hence $\text{E}(\lambda_H|y) = 0.810$ with $\text{sd} = \sqrt{200}/247 = 0.057$ and 95% interval [0.701, 0.926]; $\text{E}(\lambda_M|y) = 0.657$, $\text{sd} = 0.050$, 95% interval [0.562, 0.758]. The two densities are plotted in `figs/bda3-ch17-poisson-may-posterior.svg`; they are nearly normal (the Gamma shape parameters are large) and barely overlap, so Hamilton uses 'may' more often than Madison.
(b) Use $\delta = 1/\beta$ and $m = \alpha/\beta$, with $p(m, \delta) \propto 1/m$ — uniform on $(\log m,\, \delta)$.
Reparameterizing is the substance of the question. In the Neg-bin(α, β) family

$$p(y_i \,|\, \alpha, \beta) = \binom{\alpha + y_i - 1}{y_i} \left(\frac{\beta}{1+\beta}\right)^{\alpha} \left(\frac{1}{1+\beta}\right)^{y_i},$$

the mean is $m = \alpha/\beta$ and the variance is $m(1 + 1/\beta) = m(1 + \delta)$, so $\delta \geq 0$ is exactly the excess of the variance-to-mean ratio over the Poisson value 1, and $\delta = 0$ is the Poisson limit described in Section 17.2. A prior flat in $\log \beta$ (equivalently in $\log \delta$) is **not** admissible here: as $\delta \to 0$ the likelihood tends to the nonzero Poisson likelihood of part (a), so a density flat on $\log \delta \in (-\infty, 0]$ gives an improper posterior. Flat in δ removes that problem, and $p(m) \propto 1/m$ is the usual scale-invariant choice for a rate. The posterior is proper at the other end too: as $\delta \to \infty$ with m fixed, $\alpha = m/\delta \to 0$ and $p(y_i = k \,|\, \cdot) = O(1/\delta)$ for each $k \geq 1$, so the likelihood decays like δ^{-119} (Hamilton) and δ^{-106} (Madison), the number of blocks with at least one 'may.'
Evaluating $p(m, \delta \,|\, y) \propto m^{-1} \prod_i p(y_i \,|\, m, \delta)$ on a grid gives, for the word 'may':

Author	$\text{E}(m\|y)$	$\text{sd}(m\|y)$	95% for m	$\text{E}(\delta\|y)$	95% for δ	mode (m, δ)	mode (α, β)
Hamilton	0.818	0.071	$[0.685, 0.963]$	0.510	$[0.215, 0.895]$	$(0.804, 0.456)$	$(1.76, 2.19)$
Madison	0.665	0.064	$[0.546, 0.798]$	0.617	$[0.291, 1.051]$	$(0.651, 0.546)$	$(1.19, 1.83)$

Contour plots of $p(m, \delta \,|\, y)$ with 1000 posterior draws overlaid (drawn from the grid, as in Section 3.7) are in `figs/bda3-ch17-negbin-may-posterior.svg`. Three features: the posterior for m is centered at the same place as in (a) but about 25% wider (sd 0.071 against 0.057), since overdispersion costs precision; the contours are elliptical but tilted, m and δ having posterior correlation 0.28 (Hamilton) and 0.34 (Madison), so slightly larger mean rates go with slightly more overdispersion; and δ is decisively bounded away from 0 — $\Pr(\delta < 0.1 \,|\, y) = 0.001$ for Hamilton and 0.0001 for Madison — matching the observed variance-to-mean ratios $1.155/0.810 = 1.43$ and $1.008/0.657 = 1.54$. The data are overdispersed relative to the Poisson, and the negative binomial detects it.

Problem 17.3. Model checking with the Poisson and binomial distributions: we examine the fit of the models in the previous exercise using posterior predictive checks.
(a) Considering the nature of the data and of likely departures from the model, what would be appropriate test statistics?
(b) Compare the observed test statistics to their posterior predictive distribution (see Section 6.3) to test the fit of the Poisson model.
(c) Perform the same test for the negative binomial model.

Solution. (The printed title says "Poisson and binomial"; the models of Exercise 17.2 are the Poisson and the negative binomial, and those are what we check.)

(a) The variance-to-mean ratio $T_1(y) = s^2/\bar{y}$, the number of empty blocks $T_2(y) = \#\{i : y_i = 0\}$, the upper-tail count $T_3(y) = \#\{i : y_i \geq 3\}$, and $T_4(y) = \max_i y_i$.
The rationale: the Poisson's one free parameter forces var = mean, and the substantively likely departure is that an author's propensity to write 'may' varies from block to block with the subject matter, which inflates the variance and, more specifically, produces **both** an excess of blocks with no occurrences and an excess of blocks with many. T_1 targets the overdispersion directly; T_2, T_3, T_4 target the two tails of the frequency table, which is where a count model is most often caught out and which the reader of Table 17.2 can check by eye. A global χ^2 discrepancy on the eight cells,

$$T_5(y, \theta) = \sum_{8 \text{ cells } c} \frac{\big(n_c(y) - n\pi_c\big)^2}{n\pi_c}, \qquad \pi_c = \Pr(c \,|\, \theta),$$

over the cells $y = 0, 1, \ldots, 6$ and $y > 6$ of Table 17.2, is also natural; it depends on θ, so it is a discrepancy measure in the sense of Section 6.3 and each replication must be scored at the same draw of θ as the observed value.
(b) The Poisson model fails on every one of these checks. Drawing 40,000 values of λ from the Gamma posteriors of Exercise 17.2(a), and for each one a replicate dataset y^{rep} of 247 (respectively 262) blocks, the posterior predictive p-values $\Pr(T(y^{\text{rep}}) \geq T(y) \,|\, y)$ are

Statistic	observed (H)	rep. mean ± sd (H)	p (H)	observed (M)	rep. mean ± sd (M)	p (M)
$s^2/\bar{y}$	1.43	1.00 ± 0.09	<0.001	1.54	1.00 ± 0.09	<0.001
$\#\{y_i = 0\}$	128	110.2 ± 10.0	0.042	156	136.1 ± 10.6	0.033
$\#\{y_i \geq 3\}$	20	12.2 ± 4.0	0.042	14	7.7 ± 3.1	0.044
$\max_i y_i$	6	4.3 ± 0.7	0.049	6	3.9 ± 0.7	0.017
χ^2	-	-	0.008	-	-	0.002

Simulation standard errors are at most 0.001, so every entry is significant at face value. The variance-to-mean ratio stands about five (Hamilton) and six (Madison) predictive standard deviations above its mean, and only 4 of the 40,000 Hamilton replicates and none of Madison's reached the observed value. The other three p-values show exactly the signature anticipated in (a) — simultaneously too many empty blocks, too many crowded ones, and too high a maximum — which no one-parameter Poisson can produce, and which the χ^2 discrepancy registers at $p = 0.008$ and 0.002.
(c) The negative binomial passes. Drawing 40,000 (m, δ) from the grid posterior of Exercise 17.2(b) and replicating the same four statistics plus the χ^2 discrepancy:

Statistic	rep. mean ± sd (H)	p (H)	rep. mean ± sd (M)	p (M)
$s^2/\bar{y}$	1.50 ± 0.25	0.58	1.61 ± 0.28	0.56
$\#\{y_i = 0\}$	127.4 ± 11.0	0.50	156.1 ± 11.2	0.52
$\#\{y_i \geq 3\}$	19.8 ± 5.6	0.50	16.2 ± 5.1	0.68
$\max_i y_i$	6.1 ± 1.5	0.62	6.0 ± 1.6	0.59
χ^2	-	0.80	-	0.77

Every p-value sits in the middle of its range (simulation standard error at most 0.004), and each observed cell count lies within one predictive standard deviation of its predictive mean. The extra parameter δ buys exactly the feature the data demand; the price, noted in Exercise 17.2(b), is a 25% wider posterior interval for the mean rate.

Problem 17.4. Robust models and model checking: fit a robust model to Newcomb's speed of light data (Figure 3.1). Check the fit of the model using appropriate techniques from Chapter 6.
Figure 3.1 is a histogram of Simon Newcomb's 66 measurements of the time required for light to travel 7442 meters, recorded as deviations from 24,800 nanoseconds (Stigler, 1977). There are two unusually low measurements, -44 and -2, and then a cluster of measurements that are approximately symmetrically distributed. The 66 values are

28, 26, 33, 24, 34, −44, 27, 16, 40, −2, 29, 22, 24, 21, 25, 30, 23,
29, 31, 19, 24, 20, 36, 32, 36, 28, 25, 21, 28, 29, 37, 25, 28, 26,
30, 32, 36, 26, 30, 22, 36, 23, 27, 27, 28, 27, 31, 27, 26, 33, 26,
32, 32, 24, 39, 28, 24, 25, 32, 25, 29, 27, 28, 29, 16, 23,

with $\bar{y} = 26.2$ and $s = 10.8$. Under the normal model of Section 3.2 with $p(\mu, \sigma^2) \propto \sigma^{-2}$ the 95% posterior interval for μ is [23.6, 28.8]; the currently accepted value corresponds to $\mu = 33.0$.

Solution. Fit $y_i \sim t_\nu(\mu, \sigma^2)$ with $p(\mu, \log\sigma) \propto 1$, and let ν be the sensitivity-analysis parameter of Section 17.4.

Computation is the Gibbs sampler of Exercise 17.7(b) on the mixture representation (17.1): $V_i \,|\, \mu, \sigma \sim \text{Inv-}\chi^2\big(\nu+1,\ (\nu\sigma^2 + (y_i-\mu)^2)/(\nu+1)\big)$ by (17.6), then $\mu \,|\, V \sim \text{N}\big(\sum(y_i/V_i)/\sum(1/V_i),\ (\sum 1/V_i)^{-1}\big)$ and $\sigma^2 \,|\, V \sim \text{Gamma}\big(n\nu/2,\ (\nu/2)\sum 1/V_i\big)$. Four chains of 40,000 draws each mix immediately ($\widehat{R} < 1.01$ for μ and $\log\sigma$).

The checks of Section 6.3 need test quantities aimed at the two features a normal model cannot produce here: an extreme low tail and an inflated sample spread. Take

$$T_1(y) = \min_i y_i, \qquad T_2(y) = \max_i y_i, \qquad T_3(y) = s,$$
$$T_4(y, \mu) = |y_{(61)} - \mu| - |y_{(6)} - \mu|,$$

the last being the asymmetry measure of Section 6.3. Lower-tail posterior predictive p-values $\Pr(T(y^{\text{rep}}) \le T(y) \,|\, y)$ from 20,000 replications (simulation standard error at most 0.004):

Model	p for min	p for max	p for s	p for asymmetry
normal	<0.001	0.004	0.505	0.793
t_4	0.004	0.185	0.989	0.648
t_2	0.094	0.082	0.749	0.645
t_1 (Cauchy)	0.592	0.008	0.082	0.587

The normal model is refuted by T_1, as in Section 6.3: no replicate dataset in 20,000 had a minimum as low as −44. The t_4 model, the default robust choice, is **also** refuted, and in two directions at once: it still cannot produce −44 ($p = 0.004$) and its predictive sample standard deviation is too small (upper-tail $p = 0.011$; the replicated s has 95% interval [4.3, 9.7] against the observed 10.75). The Cauchy errs the other way: it accommodates the minimum easily ($p = 0.59$) but predicts maxima that are too large ($p = 0.008$ for T_2) and an s whose 95% interval runs from 8.1 to 618. Only at $\nu \approx 2$ does the model reproduce all four quantities: the t_2 fit gives observed-versus-replicated

T	observed	replicated 2.5%, 50%, 97.5%
$\min_i y_i$	−44	−117.5, 1.5, 17.4
$\max_i y_i$	40	37.4, 53.2, 167.3
s	10.75	4.7, 8.3, 27.1
asymmetry	1.23	−7.0, −0.1, 6.7

(The asymmetry measure depends on μ, so 1.23 is its posterior mean at the observed order statistics $y_{(6)} = 20$, $y_{(61)} = 36$.) Integrating $p(y \,|\, \nu)$ over $(\mu, \log\sigma)$ with the same flat prior for every ν, so that the arbitrary normalizing constant cancels, confirms the choice: relative to the normal, the log marginal density rises by 34.1 at $\nu = 2$, by 33.5 at $\nu = 3$, 32.3 at $\nu = 4$, 30.8 at $\nu = 1$, and 27.0 at $\nu = 8$, so the data support ν between about 1 and 4 and decisively reject the normal tails.

Inference for μ under t_2: posterior mean 27.40 (median 27.39), standard deviation 0.61, 95% interval [26.20, 28.62], with mode (from the EM algorithm of Exercise 17.7(a)) at $(\hat{\mu}, \hat{\sigma}) = (27.39, 3.74)$. The interval is less than half as wide as the normal model's [23.6, 28.8] and shifted upward by 1.2, because at the mode the EM weight w_i of the observation −44 is only 0.006 of the largest weight, and that of −2 only 0.031.

Problem 17.5. Contamination models: construct and fit a normal mixture model to the dataset used in the previous exercise — Newcomb's 66 speed of light measurements (Figure 3.1), listed in Exercise 17.4, with $\bar{y} = 26.2$ and $s = 10.8$ and two unusually low values, -44 and -2.

Solution. Take the scale-contamination model

$$y_i \;\sim\; (1-\epsilon)\,\mathrm{N}(\mu, \sigma^2) + \epsilon\,\mathrm{N}(\mu, \kappa^2\sigma^2), \qquad \kappa > 1,$$

with $p(\mu, \log\sigma) \propto 1$, $\epsilon \sim \mathrm{Beta}(1,1)$ and $\log\kappa$ uniform on $[0, \log 50]$.
This is the finite-mixture counterpart of (17.1): a common center, a rare high-variance component, and κ playing the role that ν plays in the t. Introducing indicators $z_i = 1$ if observation i comes from the wide component makes every step conjugate, so the Gibbs sampler is immediate except for κ:

$$\begin{aligned}
\Pr(z_i = 1 \,|\, \cdot) = \frac{b_i}{a_i + b_i}, \quad a_i &= (1-\epsilon)\mathrm{N}(y_i|\mu, \sigma^2), \\
b_i &= \epsilon\,\mathrm{N}(y_i|\mu, \kappa^2\sigma^2), \\
\mu \,|\, \cdot \;\sim \mathrm{N}\left(\frac{\sum_i y_i/V_i}{\sum_i 1/V_i},\, \Big(\sum_i 1/V_i\Big)^{-1}\right), \quad V_i &= \sigma^2\kappa^{2z_i}, \\
\sigma^2 \,|\, \cdot \;\sim \text{Inv-}\chi^2\left(n,\, \tfrac{1}{n}\sum_i (y_i-\mu)^2\kappa^{-2z_i}\right), & \\
\epsilon \,|\, z \sim \mathrm{Beta}\Big(1 + \textstyle\sum_i z_i,\; 1 + n - \sum_i z_i\Big), &
\end{aligned}$$

and a Metropolis step on $\log\kappa$ with jumping standard deviation 0.3 (acceptance rate 0.79). From 60,000 draws after 6000 burn-in:

Parameter	mean	sd	2.5%, 50%, 97.5%
μ	27.72	0.66	26.43, 27.72, 29.01
σ	5.05	0.49	4.16, 5.02, 6.10
κ	12.4	7.6	4.73, 10.06, 34.86
ϵ	0.062	0.037	0.012, 0.054, 0.151

The classification is unambiguous where it matters: $\Pr(z_i = 1 \,|\, y) = 1.000$ for $y_i = -44$ and 0.999 for $y_i = -2$, at most 0.14 for every other observation, and a median of 0.009 across the 66. The posterior median $\epsilon = 0.054$ is accordingly what two contaminated observations out of 66 and the $\mathrm{Beta}(1,1)$ prior give, $3/68 = 0.044$, plus a long right tail; κ is only weakly identified, as it must be when two observations inform it, but the inference for μ is unaffected by the upper bound on κ (widening it from 50 to 200 leaves the 95% interval for μ at [26.43, 29.02] and the median κ at 10.1).
The model passes the checks of Exercise 17.4: from 20,000 replications the lower-tail posterior predictive p-values are 0.27 for $\min_i y_i$, 0.16 for $\max_i y_i$, 0.47 for s, and 0.59 for the asymmetry measure (simulation standard error at most 0.004). Compared with the t_2 fit it gives almost the same location, $\mathrm{E}(\mu \,|\, y) = 27.72$ against 27.40, but a scale $\sigma = 5.05$ that describes the 64 good measurements directly, rather than through the t's mixing distribution.

Problem 17.6. Robust models:
(a) Choose a dataset from one of the examples or exercises earlier in the book and analyze it using a robust model.
(b) Check the fit of the model using the posterior predictive distribution and appropriate test variables.
(c) Discuss how inferences changed under the robust model.

Solution. (a) Take the eight-schools dataset of Table 5.2 with the modification proposed in Section 17.1, $y_8 = 100$ in place of 12, so that

$$y = (28,\, 8,\, -3,\, 7,\, -1,\, 1,\, 18,\, 100), \qquad \sigma = (15,\, 10,\, 16,\, 11,\, 9,\, 11,\, 10,\, 18),$$

and replace the normal population distribution of Section 5.5 by $\omega_j \sim t_\nu(\mu, \tau^2)$ with $p(\mu, \tau) \propto 1$. Using (17.1) for the population distribution, $\omega_j \,|\, U_j \sim \mathrm{N}(\mu, U_j)$ with $U_j \sim \text{Inv-}\chi^2(\nu, \tau^2)$, all four Gibbs steps are conjugate:

$$\begin{aligned}
\omega_j \,|\, \cdot &\sim \mathrm{N}\left(\frac{y_j/\sigma_j^2 + \mu/U_j}{1/\sigma_j^2 + 1/U_j},\ \left(1/\sigma_j^2 + 1/U_j\right)^{-1} \right), \\
U_j \,|\, \cdot &\sim \text{Inv-}\chi^2\left(\nu + 1,\ \tfrac{\nu\tau^2 + (\omega_j - \mu)^2}{\nu + 1} \right), \\
\mu \,|\, \cdot &\sim \mathrm{N}\left(\frac{\sum_j \omega_j/U_j}{\sum_j 1/U_j},\ \left(\textstyle\sum_j 1/U_j\right)^{-1} \right), \\
\tau^2 \,|\, \cdot &\sim \mathrm{Gamma}\left(\tfrac{J\nu + 1}{2},\ \tfrac{\nu}{2} \textstyle\sum_j 1/U_j \right),
\end{aligned}$$

the last using $p(\tau) \propto 1 \Rightarrow p(\tau^2) \propto (\tau^2)^{-1/2}$. (Run on the unmodified Table 5.2 data with $\nu = 4$ this sampler reproduces Table 17.1.) With four chains of 40,000 draws after 4000 burn-in:

School	normal: mean (sd)	t_4: mean (sd)	t_1: mean (sd)
A	25.7 (13.6)	22.2 (13.0)	16.4 (11.8)
B	9.1 (9.5)	8.9 (9.0)	8.2 (7.6)
C	1.8 (14.5)	2.7 (13.3)	4.5 (10.8)
D	8.4 (10.3)	8.3 (9.7)	7.9 (8.1)
E	0.9 (8.7)	1.6 (8.4)	3.4 (7.6)
F	3.2 (10.4)	3.8 (9.8)	5.1 (8.3)
G	17.9 (9.5)	16.5 (9.1)	13.1 (8.2)
H	76.9 (20.7)	82.8 (20.2)	91.9 (19.1)
μ	18.0 (14.0)	12.6 (10.9)	8.7 (7.0)
median τ	31.1	20.1	7.3

(b) Test variable: the relative isolation of the top school,

$$T(y) = \frac{y_{(8)} - y_{(7)}}{y_{(7)} - y_{(1)}},$$

whose observed value is $(100-28)/(28-(-3)) = 2.32$. This is the right variable to look at, because the question the robust model is supposed to answer is whether one school can stand far apart from seven similar ones; ordinary statistics such as $\max_j y_j$ or the sample variance are automatically matched by the normal model, which simply inflates τ to accommodate them. The replication must therefore be of a **new set of eight schools** — draw (μ, τ) from its posterior, then $\omega_j^{\text{rep}} \sim t_\nu(\mu, \tau^2)$ and $y_j^{\text{rep}} \sim \mathrm{N}(\omega_j^{\text{rep}}, \sigma_j^2)$ — which is the form of posterior predictive check sanctioned in Section 6.4; conditioning on the fitted ω_j instead would test the likelihood, not the population distribution under scrutiny. From 8000 such replications,

$$\begin{aligned}
\Pr\left(T(y^{\text{rep}}) \geq 2.32 \,\middle|\, y\right) &= 0.001 \quad \text{(normal)}, \\
&= 0.007 \quad (t_4), \\
&= 0.092 \quad (t_1).
\end{aligned}$$

The normal population model is clearly refuted — its replicated T has 97.5th percentile 1.08, against 1.54 for t_4 and 9.0 for t_1 — and t_4 is still refuted; only the Cauchy population distribution generates datasets with one school this far out in front.

(c) Three changes, all in the direction predicted in Section 17.1. First, the estimated population spread collapses: τ drops from a median of 31 under the normal model to 20 (t_4) and 7 (t_1), because the outlying school is explained by the long tail rather than by common variance. Second, the seven ordinary schools are therefore shrunk much harder toward μ: school A moves from 25.7 to 16.4 and its posterior standard deviation from 13.6 to 11.8, and schools C, E, F all move **up** toward the others. Under the normal model with $y_8 = 100$ the seven estimates stay close to their raw values, which is exactly the nonrobustness complained of in Section 17.1. Third, school H is shrunk **less**: $\mathrm{E}(\omega_8 \,|\, y)$ rises from 76.9 to 82.8 to 91.9, and $\Pr(\omega_8 > 100 \,|\, y)$ from 0.13 to 0.20 to 0.34 — in every case below 0.5, so some shrinkage remains, as Section 17.1 says it should. The single aberrant observation no longer controls the inferences for the other seven schools, while $\Pr(\omega_1 > \omega_8 \,|\, y)$ falls from 0.011 to 0.0009.

Problem 17.7. Computation for the t model: consider the model $y_1, \dots, y_n \sim$ iid $t_\nu(\mu, \sigma^2)$, with ν fixed and a uniform prior distribution on $(\mu, \log\sigma)$.
(a) Work out the steps of the EM algorithm for finding posterior modes of $(\mu, \log\sigma)$, using the specification (17.1) and averaging over $V_1, \dots, V_n$. Clearly specify the joint posterior density, its logarithm, the function $\mathrm{E}_{\text{old}} \log p(\mu, \log\sigma, V_1, \dots, V_n | y)$, and the updating equations for the M-step.
(b) Work out the Gibbs sampler for drawing posterior simulations of $(\mu, \log\sigma, V_1, \dots, V_n)$.
(c) Illustrate the analysis with the speed of light data of Figure 3.1, using a t_2 model.
Here (17.1) is the mixture representation of the t: the model $y_i \sim t_\nu(\mu, \sigma^2)$ is equivalent to

$$y_i \,|\, V_i \sim \mathrm{N}(\mu, V_i), \qquad V_i \sim \text{Inv-}\chi^2(\nu, \sigma^2).$$

Solution. (a) The M-step is a weighted mean and a weighted scale, with weights $w_i = \mathrm{E}(1/V_i \,|\, \cdot)$:

$$\mu^{\text{new}} = \frac{\sum_i w_i y_i}{\sum_i w_i}, \qquad (\sigma^{\text{new}})^2 = \frac{n}{\sum_i w_i}.$$

In detail. With $p(\mu, \log\sigma) \propto 1$ and (17.1), the joint posterior density of the parameters and the augmented variances is

$$p(\mu, \log\sigma, V \,|\, y) \propto \prod_{i=1}^n V_i^{-1/2} e^{-(y_i-\mu)^2/(2V_i)} \, (\sigma^2)^{\nu/2} V_i^{-(\nu/2+1)} e^{-\nu\sigma^2/(2V_i)},$$

whose logarithm is

$$\log p = \text{const} + n\nu \log\sigma - \sum_{i=1}^n \left[\frac{\nu+3}{2} \log V_i + \frac{(y_i-\mu)^2 + \nu\sigma^2}{2V_i} \right].$$

(No Jacobian appears beyond this, since the prior is already flat in $\log\sigma$; each Inv-$\chi^2(\nu, \sigma^2)$ factor contributes $(\sigma^2)^{\nu/2}$, hence $\sigma^{n\nu}$ in all.)
E-step. The V_i enter only through $\log V_i$, which is free of (μ, σ), and through $1/V_i$, which enters linearly; so only the n numbers $w_i = \mathrm{E}_{\text{old}}(1/V_i)$ are needed. Since $p(V_i \,|\, y_i, \mu^{\text{old}}, \sigma^{\text{old}}, \nu)$ is the Inv-χ^2 distribution (17.6) with $\nu+1$ degrees of freedom and scale $(\nu(\sigma^{\text{old}})^2 + (y_i - \mu^{\text{old}})^2)/(\nu+1)$, and $\mathrm{E}(1/V) = 1/s^2$ for $V \sim \text{Inv-}\chi^2(a, s^2)$,

$$w_i = \frac{\nu+1}{\nu(\sigma^{\text{old}})^2 + (y_i - \mu^{\text{old}})^2},$$

exactly as in the display following (17.6), and

$$\mathrm{E}_{\text{old}} \log p = \text{const}' + n\nu \log\sigma - \frac{1}{2} \sum_{i=1}^n w_i \big[(y_i-\mu)^2 + \nu\sigma^2 \big].$$

M-step. Setting $\partial/\partial\mu = \sum_i w_i(y_i - \mu) = 0$ and $\partial/\partial \log\sigma = n\nu - \nu\sigma^2 \sum_i w_i = 0$ gives the two displays above. Observations far from μ^{old} get small w_i, so each iteration is one step of iteratively reweighted least squares, as in Section 17.5.
At a fixed point, $\sum_i w_i(\nu\sigma^2 + (y_i-\mu)^2) = n(\nu+1)$ identically, so $(\sigma^{\text{new}})^2 = \sigma^2$ forces $\sum_i w_i (y_i-\mu)^2 = n$; together with $\sum_i w_i(y_i - \mu) = 0$ these are precisely the score equations of the observed-data log posterior $-n\log\sigma - \frac{\nu+1}{2} \sum_i \log\big(1 + \frac{(y_i-\mu)^2}{\nu\sigma^2}\big)$. (Check!)
(b) Three conditional draws, cycling:

$$V_i \,|\, \mu, \sigma, y \sim \text{Inv-}\chi^2\left(\nu+1, \, \frac{\nu\sigma^2 + (y_i-\mu)^2}{\nu+1} \right) \quad \text{independently},$$

$$\mu \,|\, \sigma, V, y \sim \mathrm{N}\left(\frac{\sum_i y_i/V_i}{\sum_i 1/V_i}, \, \Big(\sum_i 1/V_i \Big)^{-1} \right),$$

$$\sigma^2 \,|\, \mu, V, y \sim \text{Gamma}\left(\frac{n\nu}{2}, \, \frac{\nu}{2} \sum_i \frac{1}{V_i} \right).$$

The first is (17.6); the second is the usual weighted-mean posterior for a normal mean with known unequal variances V_i and flat prior; the third follows from collecting the σ terms of $\log p$, which give $(\sigma^2)^{n\nu/2}\exp(-\frac{\nu\sigma^2}{2}\sum_i 1/V_i)$, times the Jacobian factor $(\sigma^2)^{-1}$ from the flat prior on $\log\sigma$ — note that σ^2 appears as a **gamma**, not inverse-gamma, variate, and that its conditional distribution depends on y and μ only through V. (For small ν, add the parameter expansion of Section 12.1.)
(c) For Newcomb's 66 measurements (listed in Exercise 17.4) with $\nu = 2$: the EM algorithm of (a), started at (median(y), 5), converges to six decimal places in 65 iterations to

$$(\hat\mu, \hat\sigma) = (27.39,\ 3.74),$$

and the Gibbs sampler of (b), four chains of 40,000 draws after 4000 burn-in ($\widehat{R} < 1.01$), gives

$$\begin{aligned} &\mathrm{E}(\mu \mid y) = 27.40, \quad \mathrm{sd} = 0.61, \quad 95\%\text{: } [26.20,\ 28.62], \\ &\mathrm{E}(\sigma \mid y) = 3.82, \quad 95\%\text{: } [2.91,\ 4.96]. \end{aligned}$$

The mode 27.39 and the posterior mean 27.40 agree to 0.01, as they should for a nearly normal posterior with $n = 66$. Against the normal-model interval [23.6, 28.8] of Section 3.2, the t_2 interval is shifted up by 1.2 and is less than half as wide, because the mode's weights w_i discount $y = -44$ by a factor of 182 and $y = -2$ by a factor of 32 relative to an observation at $\hat\mu$.

17.2 Exercises 17.8–17.8

Problem 17.8. Robustness and sensitivity analysis: repeat the computations of Section 17.4 with the dataset altered as described on page 435 so that the observation y_8 is replaced by 100. Verify that, in this case, inferences are sensitive to ν. Which values of ν have highest marginal posterior density?
The altered eight-schools data (Table 5.2 with $y_8 = 100$ in place of 12, standard errors unchanged) are

School	A	B	C	D	E	F	G	H
y_j	28	8	-3	7	-1	1	18	100
σ_j	15	10	16	11	9	11	10	18

The model of Section 17.4 is the hierarchical model with a t population distribution: $y_j \mid \theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$ with σ_j known, population distribution

$$\theta_j \mid \nu, \mu, \tau \sim t_\nu(\mu, \tau^2), \quad j = 1, \ldots, 8, \tag{17.4}$$

and hyperprior $p(\mu, \tau \mid \nu) \propto 1$; for the analysis treating ν as unknown, $1/\nu$ is given a uniform prior on $[0, 1]$.

Solution. Yes: with $y_8 = 100$ every posterior summary drifts monotonically in $1/\nu$, and the marginal posterior density of $1/\nu$ rises from the normal end to a very flat maximum at $1/\nu \approx 0.87$, that is, $\nu \approx 1.15$ – the Cauchy end of the family.
Gibbs sampler. Use the mixture parameterization (17.1) applied to the population distribution (17.4): $\theta_j \mid V_j \sim \mathrm{N}(\mu, V_j)$, $V_j \sim \mathrm{Inv}\text{-}\chi^2(\nu, \tau^2)$, which leaves $(\theta_1, \ldots, \theta_8)$ exchangeable as page 436 requires. The complete-data conditionals are all standard,

$$\begin{aligned} \theta_j \mid \cdot &\sim \mathrm{N}\left(\frac{y_j/\sigma_j^2 + \mu/V_j}{1/\sigma_j^2 + 1/V_j},\ \left(\tfrac{1}{\sigma_j^2} + \tfrac{1}{V_j}\right)^{-1}\right), \\ V_j \mid \cdot &\sim \mathrm{Inv}\text{-}\chi^2\left(\nu + 1,\ \frac{\nu\tau^2 + (\theta_j - \mu)^2}{\nu + 1}\right), \\ \mu \mid \cdot &\sim \mathrm{N}\left(\frac{\sum_j \theta_j/V_j}{\sum_j 1/V_j},\ \Big(\sum_j 1/V_j\Big)^{-1}\right), \end{aligned}$$

the second being (17.6) with the residual $y_i - X_i\beta$ read as $\theta_j - \mu$ and the t scale σ as τ. For τ, the

Inv$-\chi^2(\nu,\tau^2)$ densities contribute $(\tau^2)^{J\nu/2}\exp\{-\frac{\nu\tau^2}{2}\sum_j V_j^{-1}\}$ and $p(\tau)\propto 1$ gives $p(\tau^2)\propto(\tau^2)^{-1/2}$, so

$$\tau^2 \mid \cdot \sim \text{Gamma}\left(\tfrac{J\nu}{2}+\tfrac{1}{2},\ \tfrac{\nu}{2}\sum_{j=1}^{J} V_j^{-1}\right),$$

a gamma in τ^2 itself, not an inverse-χ^2. (For $\nu=\infty$ set $V_j\equiv\tau^2$ and revert to the Section 5.4 conditionals.)

Sensitivity. The V_j integrate out one school at a time – $m_j(\mu,\tau,\nu)=\text{E}_V\,\text{N}(y_j\mid\mu,\sigma_j^2+V)$ over $V\sim$ Inv$-\chi^2(\nu,\tau^2)$, and $\text{E}(\theta_j\mid\mu,\tau,\nu,y)$ is the matching weighted average of the conditional means above – so the Gibbs output can be replaced by an exact quadrature in (μ,τ). Posterior means for $\nu=1,2,3,5,10,30,\infty$:

ν	A	B	C	D	E	F	G	H	μ	τ
1	16.3	8.2	4.4	7.8	3.4	5.1	13.1	91.8	8.7	9.5
2	19.1	8.5	3.5	8.0	2.3	4.3	14.9	88.0	10.1	15.7
3	21.0	8.7	2.9	8.1	1.8	3.9	15.9	85.1	11.5	20.1
5	23.0	8.9	2.4	8.2	1.3	3.6	16.9	81.8	13.4	25.2
10	24.5	9.0	2.1	8.3	1.0	3.4	17.5	79.1	15.5	29.7
30	25.3	9.1	1.9	8.4	0.9	3.3	17.8	77.5	17.2	32.6
∞	25.6	9.1	1.9	8.4	0.8	3.3	17.9	76.9	18.0	34.0

Posterior standard deviations move far less (school A: 13.6 at $\nu=\infty$ against 11.8 at $\nu=1$; school H: 20.6 against 19.1), so the shift in the means is a real shift, not a change of scale.

Contrast Figure 17.1, where every curve was flat to within simulation noise. Here the drift is monotone in $1/\nu$ and, being exact, cannot be blamed on noise: the posterior mean for school A falls from 25.6 to 16.3, two-thirds of a posterior standard deviation, $\text{E}(\mu\mid y)$ halves from 18.0 to 8.7, and the quartiles of τ collapse from $(23.1, 30.3, 39.8)$ to $(3.9, 7.1, 12.0)$. The mechanism is the one described on page 436: under the t the outlier claims a large V_8 of its own, so it stops inflating τ, the other seven schools are pooled far more tightly, and the outlier is itself shrunk less – $\text{E}(\theta_8\mid y)$ rises from 76.9 under the normal to 91.8 under the Cauchy, and $\Pr(\theta_8>100\mid y)$ from 0.129 to 0.337, the latter somewhat less than 0.5 as page 436 anticipates.

Marginal posterior of ν. Since $p(\mu,\tau\mid\nu)\propto 1$ with $g(\nu)\propto 1$, the density of $1/\nu$ under its uniform prior is proportional to

$$p(y\mid\nu)=\int_0^\infty\int_{-\infty}^{\infty}\prod_{j=1}^{8} m_j(\mu,\tau,\nu)\,d\mu\,d\tau,$$

$$m_j(\mu,\tau,\nu)=\int \text{N}(y_j\mid\theta,\sigma_j^2)\,t_\nu(\theta\mid\mu,\tau^2)\,d\theta.$$

The improper $p(\mu,\tau\mid\nu)\propto 1$ is harmless here: $m_j=O(\tau^{-1})$ as $\tau\to\infty$, so the integrand is $O(\tau^{-8})$ and $p(\mu,\tau\mid\nu,y)$ is proper. Each m_j was evaluated by the V-mixture average on a 900-point logarithmic grid and the outer integral by a product rule on $\mu\in[-100,200]$, $\tau\in(0,5000]$; halving or doubling either grid leaves $\log p(y\mid\nu)$ unchanged to five decimals, so the digits below are quadrature, not simulation. Normalizing over $1/\nu\in[0,1]$:

$1/\nu$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$p(1/\nu\mid y)$	0.42	0.50	0.64	0.80	0.96	1.10	1.21	1.28	1.32	1.32	1.30

The density increases over essentially the whole range, varying by under 2 parts in 100 on $[0.75,1]$, and a parabolic fit puts its maximum at $1/\nu=0.87$, i.e. $\nu\approx 1.15$; the ratio of that maximum to the value at $\nu=\infty$ is $e^{1.156}=3.2$. Hence $\Pr(1/\nu>0.5\mid y)=0.63$ against $\Pr(1/\nu<0.25\mid y)=0.14$: the highest marginal posterior density falls at the long-tailed end, ν between 1 and about 1.3, so the answer is $\nu\approx 1$, the Cauchy. One caveat, which page 443 itself raises: since $p(\mu,\tau\mid\nu)\propto g(\nu)$ is improper, $p(\nu\mid y)$ is defined only up to the convention $g(\nu)\propto 1$, and it is precisely when inferences depend strongly on ν – as they now do – that the convention needs refining before this density is read literally.

18 Models for Missing Data

Contents

18.1 Exercises 18.1–18.3

Problem 18.1. Computation for discrete missing data: reproduce the results of Section 18.6 for the 2×2 table involving independence and attendance. You can ignore the clustering in the survey and pretend it was obtained from a simple random sample. Specifically:
(a) Use EM to obtain the posterior mode of α, the proportion who will attend and will vote 'yes.'
(b) Use SEM to obtain the asymptotic posterior variance of $\text{logit}(\alpha)$, and thereby obtain an approximate 95% interval for α.
(c) Use Markov chain simulation of the parameters and missing data to obtain the approximate posterior distribution of ω. Clearly state your starting distribution. Be sure to simulate more than one sequence and to include some diagnostics on the convergence of the sequences.
The data are Table 18.1, the $3 \times 3 \times 3$ table of the 1990 Slovenian preplebiscite survey ($n = 2074$), classified by the answers to (1) independence, (2) secession, (3) attendance, with 'don't know' (DK) treated as missing at random:

Secession	Attendance	Independence: Yes	No	Don't know
Yes	Yes	1191	8	21
Yes	No	8	0	4
Yes	Don't know	107	3	9
No	Yes	158	68	29
No	No	7	14	3
No	Don't know	18	43	31
Don't know	Yes	90	2	109
Don't know	No	1	2	25
Don't know	Don't know	19	8	96

For this exercise the secession question is ignored, so the data are the 3×3 margin (independence by attendance):

Independence / Attendance	Yes	No	Don't know
Yes	1439	16	144
No	78	16	54
Don't know	159	32	136

Here ω is the vector of cell probabilities of the complete-data 2×2 table, and as in Section 18.6 its prior distribution is Dirichlet with all four parameters equal to 0.1.

Solution. $\hat{\alpha} = 0.892$, with approximate 95% interval [0.877, 0.905] from SEM and [0.877, 0.906] from data augmentation.
Index the four cells of the complete 2×2 table by jk, with j recording 'yes' on independence and k 'yes' on attendance, so that $\alpha = \omega_{11}$. The completely classified counts are

$$(m_{11}, m_{10}, m_{01}, m_{00}) = (1439,\ 16,\ 78,\ 16),$$

and there are five patterns of partially classified observations, in the notation of Section 18.5,

$$\begin{aligned}
&r_1 = 144, \quad S_1 = \{11, 10\} && \text{(independence yes, attendance DK)},\\
&r_2 = 54, \quad S_2 = \{01, 00\} && \text{(independence no, attendance DK)},\\
&r_3 = 159, \quad S_3 = \{11, 01\} && \text{(attendance yes, independence DK)},\\
&r_4 = 32, \quad S_4 = \{10, 00\} && \text{(attendance no, independence DK)},\\
&r_5 = 136, \quad S_5 = \{11, 10, 01, 00\},
\end{aligned}$$

with $1549 + 525 = 2074$. The missing-at-random assumption of Section 18.6 is exactly what licenses allocating each r_p group within S_p in proportion to ω.

(a) The E-step allocates each partially classified group proportionally,

$$n_{jk}^{\text{old}} = m_{jk} + \sum_{p=1}^{5} r_p \, \pi_{jk\,p}^{\text{old}},$$
$$\pi_{jk\,p} = \frac{\omega_{jk} I_{jk \in S_p}}{\sum_{l \in S_p} \omega_l},$$

and the M-step for the saturated multinomial is that of Section 18.6 with four cells,

$$\omega_{jk}^{\text{new}} = \frac{n_{jk}^{\text{old}} + 0.1}{2074 + 0.4}.$$

(This is the posterior mean of the completed table rather than its mode; with Dirichlet parameters below 1 the completed-data mode is not interior, and part (b) uses the complete-data log density that this M-step does maximize.) Started from $\omega^{(0)} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ the iteration meets the tolerance 10^{-16} of Section 18.6 after 62 steps at

$$\hat{\omega} = (0.891818,\ 0.015688,\ 0.065711,\ 0.026783),$$

with expected complete-data counts

$$\hat{n} = (1849.89,\ 32.44,\ 136.21,\ 55.46).$$

So $\hat{\alpha} = 0.8918$, against 0.882 for the three-way analysis of Section 18.6, which uses the secession answers to sharpen the imputations.
(b) Following Section 18.6, apply SEM on the logit scale, in the coordinates

$$\varphi = \big(\operatorname{logit}\omega_{11},\ \operatorname{logit}\omega_{10},\ \operatorname{logit}\omega_{01}\big), \qquad \omega_{00} = 1 - \omega_{11} - \omega_{10} - \omega_{01},$$

so that the $(1,1)$ entry of the variance matrix is $\operatorname{Var}(\operatorname{logit}\alpha)$ itself. At $\hat{\varphi} = (2.10944, -4.13905, -2.65451)$ the complete-data information $-d^2 \log p(\varphi, y_{\text{mis}})/d\varphi^2$, averaged over the missing data (the complete-data log density is linear in n, so this is just its value at $\hat{n}$), is

$$I_{\text{oc}} = \begin{pmatrix} 742.59 & 115.39 & 458.76 \\ 115.39 & 50.00 & 73.43 \\ 458.76 & 73.43 & 410.91 \end{pmatrix}, \qquad V_{\text{joint}} = I_{\text{oc}}^{-1},$$

with $\operatorname{diag} V_{\text{joint}} = (0.004997, 0.031218, 0.007852)$. Numerically differentiating the EM map M of step 3 of the SEM algorithm (Section 13.5) at $\hat{\varphi}$, so that $DM_{ij} = \partial M_j / \partial \varphi_i$, gives

$$DM = \begin{pmatrix} 0.0724 & 0.8300 & 0.3056 \\ -0.0120 & 0.5039 & 0.0503 \\ -0.0474 & 0.5334 & 0.4216 \end{pmatrix},$$

with eigenvalues 0.5865, 0.2862, 0.1252; the largest is the worst fraction of missing information, and the successive errors $\max_{jk} |\omega_{jk}^{(t)} - \hat{\omega}_{jk}|$ in (a) do contract by exactly 0.5865 per step. (Check!) The SEM formula of Section 13.5,

$$V = V_{\text{joint}} + V_{\text{joint}}\, DM\, (I - DM)^{-1},$$

yields

$$V = \begin{pmatrix} 0.005751 & -0.005630 & -0.005697 \\ -0.005630 & 0.054430 & 0.000852 \\ -0.005697 & 0.000852 & 0.010639 \end{pmatrix},$$

which is symmetric and agrees to every digit shown with the directly inverted Hessian of the observed-data log posterior at $\hat{\varphi}$. Hence

$$\operatorname{sd}(\operatorname{logit}\alpha) = \sqrt{0.005751} = 0.07583,$$

and $2.10944 \pm 1.96(0.07583) = [1.96081,\ 2.25808]$ transforms back to

$$\alpha \in [0.8766,\ 0.9053].$$

No design-effect inflation is applied here, so this is narrower than the [0.857, 0.903] of Section 18.6.
(c) Data augmentation alternates two draws. Given ω, each partially classified group is reallocated by an independent multinomial draw,

$$(n_{jk} - m_{jk})_{jk \in S_p} \sim \text{Multin}\big(r_p;\ (\pi_{jk\,p})_{jk \in S_p}\big), \qquad p = 1, \ldots, 5,$$

with $\pi_{jk\,p}$ as in (a); given the completed table the conjugate draw (Section 3.4) is

$$\omega \mid n \sim \text{Dirichlet}(n_{11} + 0.1,\ n_{10} + 0.1,\ n_{01} + 0.1,\ n_{00} + 0.1).$$

Starting distribution: five independent draws $\omega^{(0)} \sim \text{Dirichlet}(1, 1, 1, 1)$, uniform on the simplex, which is overdispersed relative to the posterior distribution – its draws for α spread over essentially all of $(0, 1)$, against a posterior sd of 0.007. Each of the five chains was run 2000 iterations and its first half discarded, leaving 5000 draws. The split potential scale reductions are

$$\hat{R} = 1.000,\ 1.003,\ 1.002,\ 1.003 \qquad \text{for } \omega_{11}, \omega_{10}, \omega_{01}, \omega_{00},$$

all below the 1.1 of Section 18.6. The estimand is $\alpha = \omega_{11}$, and the draws give

$$\begin{aligned} \text{median}(\alpha) &= 0.8920, \\ \text{sd}(\alpha) &= 0.0073, \\ 95\%\ \text{interval} &= [0.8770,\ 0.9058], \end{aligned}$$

with Monte Carlo standard errors, from the spread across eight independent replications of the whole run, of 0.0001 on the median and 0.0003 on the interval endpoints. On the logit scale $\text{sd}(\text{logit}\,\alpha) = 0.0763$, against the SEM value 0.0758 of (b).

Problem 18.2. Monotone missing data: create a monotone pattern of missing data for the opinion poll data of Section 18.6 by discarding some observations. Compare the results of analyzing these data with the results given in that section.
(Work with the 2×2 version of Exercise 18.1: the independence-by-attendance margin of Table 18.1, with 'don't know' treated as missing,

Independence / Attendance	Yes	No	Don't know
Yes	1439	16	144
No	78	16	54
Don't know	159	32	136

and the Dirichlet prior distribution with all four parameters equal to 0.1.)

Solution. Discarding to monotonicity moves $\text{E}(\alpha)$ from 0.892 up to 0.905 or 0.914 according to which variable is put first, and the shift is a bias, not a loss of precision: deleting a missing-data pattern is a selection on the observed data, and it does not preserve the missing-at-random property that Section 18.6 assumes.
Factor $\omega_{jk} = \nu_j \lambda_{k|j}$ with $\nu_j = \Pr(\text{independence} = j)$. By the aggregation and neutrality properties of the Dirichlet distribution (Appendix A) the $\text{Dirichlet}(0.1, 0.1, 0.1, 0.1)$ prior distribution is exactly

$$\nu_1 \sim \text{Beta}(0.2, 0.2), \quad \lambda_{1|1} \sim \text{Beta}(0.1, 0.1), \quad \lambda_{1|0} \sim \text{Beta}(0.1, 0.1),$$

independently, so for a monotone pattern the posterior distribution factors the same way and needs no iteration (Section 18.5).
(A) Independence first. Monotonicity requires attendance to be observed only where independence is, so delete the 'independence DK, attendance observed' cells, $159 + 32 = 191$ observations, leaving

$n_A = 1883$ in the patterns

$$\begin{aligned} \text{both observed:} &\quad (1439, 16, 78, 16), \\ \text{independence only:} &\quad 144 \text{ yes}, \ 54 \text{ no}, \\ \text{neither:} &\quad 136. \end{aligned}$$

The 136 doubly missing cases contribute nothing, the $1549 + 198 = 1747$ units with independence observed contribute to ν, and only the completely classified ones contribute to λ:

$$\nu_1 \mid y \sim \text{Beta}(1599.2,\ 148.2), \qquad \lambda_{1|1} \mid y \sim \text{Beta}(1439.1,\ 16.1),$$

independently, with $\alpha = \nu_1 \lambda_{1|1}$. Its first two moments are exact,

$$\text{E}(\alpha) = 0.90506, \qquad \text{sd}(\alpha) = 0.00705,$$

and 4×10^6 draws of the product give the central 95% interval [0.8908, 0.9184].
(B) Attendance first. Now delete 'attendance DK, independence observed,' 144+54 = 198 observations, leaving $n_B = 1876$. With $\mu_j = \Pr(\text{attendance} = j)$ and $\kappa_{1|1} = \Pr(\text{independence yes} \mid \text{attend})$,

$$\mu_1 \mid y \sim \text{Beta}(1676.2,\ 64.2), \qquad \kappa_{1|1} \mid y \sim \text{Beta}(1439.1,\ 78.1),$$

and $\alpha = \mu_1 \kappa_{1|1}$ has

$$\text{E}(\alpha) = 0.91353, \qquad \text{sd}(\alpha) = 0.00694,$$

with 95% interval [0.8995, 0.9267].
Against Exercise 18.1, where all 2074 observations give $\text{E}(\alpha) = 0.8919$, $\text{sd}(\alpha) = 0.0073$ and interval [0.8770, 0.9058]: (B) is off by 3.0 full-data posterior standard deviations and its interval excludes the full-data mean, while (A) is off by 1.8 and only barely covers it. Neither monotone posterior sd exceeds the full-data sd, so none of this is lost precision.
Both factors are biased upward, because each is estimated from a subsample selected on the other variable. Let $R = 1$ indicate that independence is observed. MAR permits $\Pr(R = 1 \mid \text{indep.}, \text{att.}) = g(\text{att.})$, whence

$$\Pr(\text{indep. yes} \mid R = 1) = \frac{\sum_k \Pr(\text{indep. yes, att.} = k)\, g(k)}{\sum_{j,k} \Pr(\text{indep.} = j, \text{ att.} = k)\, g(k)},$$

which equals ν_1 only for constant g, that is only under MCAR. Numerically, in (A)

$$\begin{aligned} \hat{\nu}_1 &= \tfrac{1599}{1747} = 0.9153 \quad \text{against} \quad \hat{\omega}_{11} + \hat{\omega}_{10} = 0.9075, \\ \hat{\lambda}_{1|1} &= \tfrac{1439}{1455} = 0.9890 \quad \text{against} \quad \hat{\omega}_{11}/(\hat{\omega}_{11} + \hat{\omega}_{10}) = 0.9827, \end{aligned}$$

and in (B) $\hat{\mu}_1 = 1676/1740 = 0.9632$ against 0.9575 and $\hat{\kappa}_{1|1} = 1439/1517 = 0.9485$ against 0.9314. The monotone reduction buys a closed-form posterior distribution at the price of a bias of the order of the DK rate itself.

Problem 18.3. Practical missing-data imputation: create a miniature version of the 2010 General Social Survey (publicly available on the Internet), including the following variables: sex, age, ethnicity (use four categories), urban/suburban/rural, education (use five categories), political ideology (on a 7-point scale from 'extremely liberal' to 'extremely conservative'), and general happiness.
(a) Using just the complete cases, fit a logistic regression on whether respondents feel 'not too happy.'
(b) Impute the missing values using mi() in the mi package in R. Then take one of the completed datasets and fit a logistic regression as above.
(c) Repeat, this time imputing using aregImpute() in the Hmisc package.
(d) Repeat, this time imputing using mice() in the mice package.
(e) Briefly discuss the differences between the four inferences above.

Solution. In the GSS the missingness is a few percent and concentrated in one covariate, so the four point estimates agree to well within one standard error and the only inference that really differs is the standard error – and the one procedure that is actually wrong is the one the exercise literally prescribes, analyzing a single completed dataset. Everything below was computed on the 2010 single-year file (`GSS2010.dta` from the NORC site, $n = 2044$).

Build the extract as follows: SEX (female indicator); AGE, entered as age/10 so that coefficients read per decade; ethnicity in four categories from RACE and HISPANIC (Hispanic if HISPANIC > 1, else white/black/other from RACE); SRCBELT collapsed to urban (codes 1, 2, 5), suburban (3, 4), rural (6); DEGREE in its five categories; POLVIEWS on the 1–7 liberal-to-conservative scale, entered as numeric; outcome $y = 1$ if HAPPY is 'not too happy.' The missingness pattern in the 2010 file: POLVIEWS is missing for 71 respondents, ethnicity for 6 (through HISPANIC), happiness for 5 and age for 3; sex, place and education are complete; 1964 of the 2044 rows are complete cases. The model throughout is

$$\text{logit } \Pr(y_i = 1) = X_i\beta,$$

with white, urban, male and 'less than high school' as baselines.

(a) Complete cases: drop the 80 incomplete rows and fit by maximum likelihood – the first column below.

(b) mi(): `missing_data.frame()` then `mi()` (4 chains, 30 iterations, a chained Bayesian-GLM sampler), take one completed dataset from `complete()` and refit – the second column.

(c) aregImpute(): predictive mean matching from additive models with a bootstrap resample of the fitting data at each of 5 imputations; fill one completion in with `impute.transcan()` and refit – the third column.

(d) mice(): chained equations at its defaults ($m = 5$; predictive mean matching for the numeric columns, logistic or polytomous regression for the factors), refit on `complete(imp, 1)` – the fourth column.

Estimates (standard errors), all runs seeded:

coefficient	(a) compl. cases	(b) mi	(c) aregImpute	(d) mice
intercept	-1.627 (0.328)	-1.509 (0.315)	-1.457 (0.314)	-1.609 (0.314)
female	-0.197 (0.129)	-0.219 (0.125)	-0.218 (0.125)	-0.201 (0.125)
age/10	0.074 (0.037)	0.059 (0.036)	0.057 (0.036)	0.055 (0.036)
black	0.460 (0.175)	0.509 (0.167)	0.495 (0.167)	0.533 (0.167)
Hispanic	0.377 (0.201)	0.444 (0.191)	0.412 (0.191)	0.441 (0.191)
other	0.250 (0.333)	0.174 (0.331)	0.163 (0.331)	0.180 (0.331)
suburban	-0.233 (0.153)	-0.209 (0.147)	-0.206 (0.147)	-0.202 (0.147)
rural	-0.364 (0.223)	-0.398 (0.218)	-0.400 (0.218)	-0.407 (0.218)
high school	-0.388 (0.175)	-0.390 (0.166)	-0.398 (0.165)	-0.387 (0.166)
junior college	-0.542 (0.288)	-0.519 (0.274)	-0.525 (0.274)	-0.531 (0.274)
bachelor	-1.183 (0.249)	-1.181 (0.241)	-1.194 (0.240)	-1.146 (0.238)
graduate	-0.925 (0.276)	-0.947 (0.271)	-0.960 (0.270)	-0.941 (0.271)
ideology	0.027 (0.045)	0.020 (0.043)	0.014 (0.043)	0.046 (0.043)

Education dominates (the bachelor and graduate categories sit several standard errors below the baseline), black and Hispanic respondents report 'not too happy' more often, and ideology is well within one standard error of zero in every fit.

(e) The three imputation engines agree with each other to well within one standard error on every coefficient – for the fully observed covariates, to within a few hundredths – because they impute the same 80-odd cells under models that differ only in their internal machinery; at a 4% missingness rate that machinery cannot matter. The visible spread among (b)–(d) is concentrated exactly where it should be, in the ideology coefficient (0.014 to 0.046, against a standard error of 0.043) and in the intercept that trades off against it, because ideology carries 71 of the missing values. The complete-case fit is the mild outlier, its coefficients shifting by up to about half a standard error (age, black) as the 80 incomplete rows are restored, and its standard errors are the largest in every row, by 1–5% – at least the case-count factor $\sqrt{2044/1964} \approx 1.02$.

The consequential difference is that the standard errors from (b)–(d) are too small: they come from

one completed dataset, which treats imputed values as though observed. The correct combination is the rule of Section 18.2. With K completions giving estimates $\hat{\beta}_k$ and variances W_k,

$$\bar{\beta}_K = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k, \qquad T_K = W_K + \frac{K+1}{K}\,B_K,$$

where W_K is the average of the W_k and B_K the sample variance of the $\hat{\beta}_k$. Any single completion reports W_k, and $\mathrm{E}(W_K) < \mathrm{E}(T_K)$ whenever $B_K > 0$, so a single-completion interval is too short by construction, at the nominal level only if the imputations carry no uncertainty. Because B_K is appreciable only for coefficients whose covariate is itself incomplete, the deficit concentrates on the ideology coefficient: pooling the $K = 5$ mice completions by the rule above gives the ideology coefficient a standard error of 0.048 against 0.043 from the single completion in the table, a 12% understatement, while the pooled standard errors of the fully observed sex, place and education coefficients match their single-completion values to within a few thousandths (the intercept, which inherits part of the ideology uncertainty, is the other coefficient whose pooled standard error grows visibly, 0.314 to 0.328). Intervals should use $\bar{\beta}_K \pm t_{\mathrm{df}}\sqrt{T_K}$ with the Section 18.2 degrees of freedom

$$\mathrm{df} = (K-1)\Big(1 + \frac{K}{K+1}\frac{W_K}{B_K}\Big)^2.$$

Complete-case analysis is not exposed to this particular error: its standard errors rest on fewer data but are not too small.

19 Parametric Nonlinear Models

Contents

19.1 Exercises 19.1–19.3

Problem 19.1. Nonlinear modeling: The file `dilution.dat` contains data from the dilution assay experiment described in Section 19.1.
(a) Use Stan to fit the model described in Section 19.1.
(b) Fit the same model, but with a hierarchical mixture prior distribution on the θ_j's which includes the possibility of some true concentrations θ_j to be zero. Discuss your model, its parameters, and your hyperprior distribution for these parameters.
(c) Compare the inferences for the θ_j's from the two models above.
(d) Construct a dataset (with the same dilutions and the same number of unknowns and measurements), for which these two models yield much different inferences.
The model of Section 19.1, for reference. A plate has 96 wells; well i holds a sample at concentration x_i and returns an optical reading y_i. The calibration curve is (equation 19.1)

$$E(y \mid x, \beta) = g(x, \beta) = \beta_1 + \frac{\beta_2}{1 + (x/\beta_3)^{-\beta_4}},$$

with measurement errors of unequal variance (equation 19.2),

$$y_i \sim \mathrm{N}\left(g(x_i, \beta),\ \left(\frac{g(x_i, \beta)}{A}\right)^{2\alpha} \sigma_y^2\right),$$

where $A = 30$ and $0 \le \alpha \le 1$. The standard has known concentration θ_0 and a called-for initial dilution d_0^{init}, so that its nominal concentration is $d_0^{\text{init}}\theta_0 = 0.64$; the actual concentration x_0^{init} of that initial dilution is unknown (equation 19.3):

$$\log(x_0^{\text{init}}) \sim \mathrm{N}\big(\log(d_0^{\text{init}}\theta_0),\ (\sigma^{\text{init}})^2\big), \qquad \sigma^{\text{init}} = 0.02 \text{ (fixed)}.$$

For the unknowns there is no initial dilution, so $x_j^{\text{init}} = \theta_j$, $j = 1, \ldots, 10$, and every well is a further dilution of its sample (equation 19.4),

$$x_i = d_i \cdot x_{j(i)}^{\text{init}},$$

d_i being the dilution of well i relative to the initial dilution. Priors: $\log \beta_k \sim \mathrm{U}(-\infty, \infty)$ for $k = 1, \ldots, 4$; $\sigma_y \sim \mathrm{U}(0, \infty)$; $\alpha \sim \mathrm{U}(0, 1)$; $p(\log \theta_j) \propto 1$ for each unknown j.
The measurements displayed in Figure 19.3 (the standards column of the plate, and two of the ten unknowns) are:

Standards: conc.	dilution	y
0.64	1	101.8
0.64	1	121.4
0.32	1/2	105.2
0.32	1/2	114.1
0.16	1/4	92.7
0.16	1/4	93.3
0.08	1/8	72.4
0.08	1/8	61.1
0.04	1/16	57.6
0.04	1/16	50.0
0.02	1/32	38.5
0.02	1/32	35.1
0.01	1/64	26.6
0.01	1/64	25.0
0	0	14.7
0	0	14.2

dilution	y (Unknown 8)	y (Unknown 9)
1	19.2	49.6
1	19.5	43.8
1/3	16.1	24.0
1/3	15.8	24.1
1/9	14.9	17.3
1/9	14.8	17.6
1/27	14.3	15.6
1/27	16.0	17.1

Every estimate for Unknown 8 and the four most dilute estimates for Unknown 9 are reported by the standard assay software as "below detection limit."

Solution. (a) Fitting (19.1)–(19.4) to the 32 measurements of Figure 19.3 (16 standards, plus Unknowns 8 and 9) gives the posterior medians and 50% intervals

parameter	median	50% interval
β_1	14.68	[14.40, 14.95]
β_2	106.8	[102.8, 111.4]
β_3	0.06797	[0.0620, 0.0751]
β_4	1.135	[1.081, 1.192]
α	0.945	[0.892, 0.977]
σ_y	2.074	[1.879, 2.308]
x_0^{init}	0.6401	[0.6315, 0.6488]
θ_8	0.00421	[0.00355, 0.00490]
θ_9	0.02965	[0.02792, 0.03144]

with 95% intervals $[0.0021, 0.0064]$ for θ_8 and $[0.0247, 0.0353]$ for θ_9. (Sampling: four chains of random-walk Metropolis on $(\log\beta,\ \text{logit}\,\alpha,\ \log\sigma_y,\ \log x_0^{\text{init}},\ \log\theta_8,\ \log\theta_9)$, proposal covariance from the normal approximation at the mode, 400{,}000 iterations each, first quarter discarded, thinned by 20; $\hat{R} \le 1.02$, and the θ intervals carry about 2% Monte Carlo error in their endpoints. Section 19.1 fits all 96 wells and reports $\hat{\beta} = (14.7, 99.7, 0.054, 1.34)$, $\hat{\sigma}_y = 2.2$, $\hat{\alpha} = 0.97$; the shift in β_3, β_4 here is the eight omitted unknowns, which in the full plate pin down the midpoint of the curve.)

One flag before reading the table: $p(\log\theta_j) \propto 1$ makes this posterior **improper**, and not just formally. As $\theta_j \downarrow 0$ the mean function tends to β_1 at every dilution, so $p(y_j \mid \theta_j) \to p(y_j \mid \theta_j = 0) > 0$, a positive constant, and $\int_{-\infty}^{\log c} p(y_j \mid \theta_j = e^t)\,dt = \infty$. For Unknowns 8 and 9 that constant sits 26 and 1300 log-units below the likelihood maximum, so the flat tail would need some 10^{11} log-units of range before it carried comparable mass: the divergent tail is numerically unreachable here and the summaries above are stable. Part (c) shows what happens when it is not.

Two features of the fit matter. First, α is near 1, so the error model is essentially multiplicative and the low readings are correctly given small standard deviations (about $(15/30)^{0.95} \cdot 2.07 \approx 1.1$) rather than the common σ_y of an equal-variance model. Second, the eight readings of Unknown 8 – all "below detection limit" – determine θ_8 to within a factor of about 1.7, because the model reads the **decline** of y from dilution 1 to $\frac{1}{3}$ to $\frac{1}{9}$ through the calibration curve rather than thresholding each well separately.

(b) Replace $p(\log\theta_j) \propto 1$ by the point-mass mixture

$$\theta_j \sim \lambda\,\delta_0 + (1-\lambda)\,\text{LN}(\mu_\theta, \sigma_\theta^2), \qquad j = 1, \dots, 10,$$

i.e. $\theta_j = 0$ with probability λ and otherwise $\log\theta_j \sim \text{N}(\mu_\theta, \sigma_\theta^2)$. The three hyperparameters have the interpretations: λ, the proportion of homes on this plate with no detectable allergen; μ_θ, the geometric mean concentration among the positive samples; σ_θ, their spread on the log scale. Hyperprior: $\lambda \sim \text{U}(0,1)$, $p(\mu_\theta) \propto 1$, $\sigma_\theta \sim \text{U}(0, 10)$ – weak, but proper in λ, which is what matters, since with only 10 unknowns λ is the parameter the data barely pin down. The indicators are summed out analytically,

$$\begin{aligned} p(y_j \mid \lambda, \mu_\theta, \sigma_\theta, \beta, \alpha, \sigma_y) &= \lambda\, p(y_j \mid \theta_j = 0) \\ &\quad + (1-\lambda) \int \text{N}(t \mid \mu_\theta, \sigma_\theta^2)\, p(y_j \mid \theta_j = e^t)\,dt, \end{aligned}$$

which is what Stan requires anyway (no discrete parameters); the integral is done on a 300-point grid in $t = \log\theta$ over $[\log 10^{-6}, \log 5]$, and the posterior probability that a sample is truly zero is recovered afterwards as

$$\Pr(\theta_j = 0 \mid y) = E\left[\frac{\lambda\, p_{0j}}{\lambda\, p_{0j} + (1-\lambda)\, q_j(\mu_\theta, \sigma_\theta)}\right],$$

writing $p_{0j} = p(y_j \mid \theta_j = 0)$ and $q_j = \int \mathrm{N}(t \mid \mu_\theta, \sigma_\theta^2)\, p(y_j \mid e^t)\, dt$, the expectation being over the posterior draws of $(\lambda, \mu_\theta, \sigma_\theta, \beta, \alpha, \sigma_y)$.

The book's plate carries ten unknowns but only Unknowns 8 and 9 are printed in the text, and a mixture hyperprior on two units is vacuous. So (b)–(d) are carried out on full ten-unknown plates simulated from the part (a) calibration curve refitted to the 16 standards alone (posterior medians $\beta = (14.37, 112.1, 0.0779, 1.041)$, $\alpha = 0.80$), with noise $\sigma_y = 3.05$ – essentially that fit's posterior median $\hat\sigma_y = 3.0$, which with only 16 points sits well above the posterior mode 2.3, so the detection limit is not flattered. Each unknown gets the Figure 19.3 dilution design $d = (1, 1, \frac{1}{3}, \frac{1}{3}, \frac{1}{9}, \frac{1}{9}, \frac{1}{27}, \frac{1}{27})$, and the calibration parameters are integrated out under the standards-only posterior (400 thinned Metropolis draws; the unknowns carry negligible information about them). Everything below is reproducible: NumPy seed 19 drives the standards Metropolis (four chains of 60{,}000, second halves thinned by 50, acceptance 0.33) and the 400-draw subsample, seed 191 the plate noise; the no-pooling posteriors use a 300-point grid in $\log\theta$ over $[10^{-6}, 5]$ and the mixture a $40 \times 60 \times 40$ grid over $(\lambda, \mu_\theta, \sigma_\theta) \in (0,1) \times [\log 10^{-5}, \log 2] \times (0, 10]$.

(c) Plate 1, true concentrations $(0, 0, 0.0005, 0.002, 0.005, 0.01, 0.03, 0.08, 0.2, 0.5)$:

j	true θ_j	no-pooling median	no-pooling 95%	$\Pr(\theta_j = 0 \mid y)$	mixture median if $\neq 0$
1	0	0.000035	[1.1e-6, 0.0014]	0.50	0.00029
2	0	0.000082	[1.2e-6, 0.0019]	0.41	0.00052
3	0.0005	0.00049	[1.4e-6, 0.0034]	0.25	0.0013
4	0.002	0.00046	[1.4e-6, 0.0033]	0.26	0.0012
5	0.005	0.0048	[0.0020, 0.0080]	0.003	0.0049
6	0.01	0.0105	[0.0059, 0.016]	0.000	0.0105
7	0.03	0.0252	[0.018, 0.034]	0.000	0.0252
8	0.08	0.0695	[0.058, 0.084]	0.000	0.0694
9	0.2	0.167	[0.13, 0.21]	0.000	0.166
10	0.5	0.451	[0.37, 0.56]	0.000	0.449

The hyperparameters are estimated at $\hat\lambda = 0.16$ (95% interval $[0.013, 0.54]$), $e^{\hat\mu_\theta} = 0.012$, $\hat\sigma_\theta = 2.7$. Note Unknown 4: its true concentration 0.002 is positive, but this draw of the noise left its readings indistinguishable from Unknown 3's, and both models say so – the mixture by $\Pr(\theta_4 = 0 \mid y) = 0.26$, an honest statement that a quarter of the posterior mass is on "truly empty."

Above the detection limit – unknowns 5–10, say $\theta_j \gtrsim 0.005$ – the two models agree to within 2%: the likelihood is sharp there and the prior, mixture or flat, is irrelevant. Below it they differ in kind, not merely in their point estimates. The impropriety flagged in part (a) now bites: for unknown 1, whose eight readings are all indistinguishable from β_1, the constant $p(y_1 \mid \theta_1 = 0)$ sits at the likelihood maximum rather than 26 log-units below it, so

$$\int_{-\infty}^{\log c} p(y_1 \mid \theta_1 = e^t)\, dt = \infty$$

with the divergence reached immediately. Its quoted median 3.5×10^{-5} and lower endpoint 1.1×10^{-6} are artifacts of the grid truncation and have no inferential content: rerunning the same fit with the lower bound moved from 10^{-6} to 10^{-9} to 10^{-12} drags the median from 3.5×10^{-5} to 1.1×10^{-6} to 3.5×10^{-8} (each factor of 10^{-3} in the bound moves the median by about $10^{-3/2}$, exactly the log-uniform-tail scaling). The mixture returns the well-posed answer instead, $\Pr(\theta_1 = 0 \mid y) = 0.50$, with the remaining mass distributed as the positive component conditioned on being small – and it is the **other nine** unknowns that supply the $\lambda, \mu_\theta, \sigma_\theta$ that make the statement possible.

(d) Take the same design with all ten unknowns at or below the detection limit: true $\theta = (0, 0, 0, 0, 0, 0, 0.0002, 0.0005, 0.001, 0.002)$. The simulated readings (seed 191, continuing the same stream) are then all in the 10.4–17.2 range, i.e. all "below detection limit." The two fits:

j	true θ_j	no-pooling median	no-pooling mass at $\theta < 10^{-4}$	$\Pr(\theta_j = 0 \mid y)$
1	0	0.000027	0.70	0.80
2	0	0.000041	0.63	0.79
3	0	0.000022	0.73	0.81
4	0	0.000020	0.75	0.81
5	0	0.000036	0.65	0.79
6	0	0.000022	0.74	0.81
7	0.0002	0.000055	0.58	0.78
8	0.0005	0.000029	0.68	0.80
9	0.001	0.00051	0.34	0.73
10	0.002	0.000076	0.54	0.77

with $\hat{\lambda} = 0.83$ (95% interval $[0.07, 0.98]$) and $e^{\hat{\mu}_\theta} = 0.00025$. Here every no-pooling marginal has its flat tail within a few log-units of the likelihood maximum, so the whole left-hand half of the table is grid-dependent noise (every median would shift by orders of magnitude under the truncation experiment of part (c)), while the mixture reports $\Pr(\theta_j = 0 \mid y) \approx 0.73$–$0.81$ for every well and borrows the ten samples' shared evidence that this plate is mostly empty.

Problem 19.2. Nonlinear modeling: Table 19.1 presents data on the success rate of putts by professional golfers (see Berry, 1996, and Gelman and Nolan, 2002c).
(a) Fit a nonlinear model for the probability of success (using the binomial likelihood) as a function of distance. Does your fitted model make sense in the range of the data and over potential extrapolations?
(b) Use posterior predictive checks to assess the fit of the model.
Table 19.1. Number of attempts and successes of golf putts, by distance from the hole, for a sample of professional golfers. From Berry (1996).

distance (feet)	tries	successes
2	1443	1346
3	694	577
4	455	337
5	353	208
6	272	149
7	256	136
8	240	111
9	217	69
10	200	67
11	237	75
12	202	52
13	192	46
14	174	54
15	167	28
16	201	27
17	195	31
18	191	33
19	147	20
20	152	24

Solution. (a) Take the one-parameter geometric model

$$p(x) = \Pr\left(|\vartheta| < \sin^{-1}\frac{R-r}{x}\right) = 2\Phi\left(\frac{1}{\sigma}\sin^{-1}\frac{R-r}{x}\right) - 1,$$

where x is the distance in inches, $R = 2.125$ in is the radius of the hole, $r = 0.84$ in the radius of the ball, and the aim error is $\vartheta \sim \mathrm{N}(0, \sigma^2)$: a putt struck hard enough drops exactly when the ball's centre passes within $R - r$ of the hole's centre. The single parameter σ is the standard deviation of a

professional's angular error. With $y_j \sim \text{Bin}(n_j, p(x_j))$ independently and a flat prior on σ, the posterior is

$$p(\sigma \mid y) \propto \prod_{j=1}^{19} p(x_j)^{y_j} \, (1 - p(x_j))^{n_j - y_j},$$

a smooth one-dimensional density evaluated directly on a grid: posterior median $\sigma = 0.0266$ radians, 95% interval $[0.0259, 0.0274]$, i.e. $1.53°$ with interval $[1.48°, 1.57°]$. The observed and fitted rates:

x	n_j	y_j	observed	fitted	std. resid.
2	1443	1346	0.933	0.956	-4.2
3	694	577	0.831	0.820	0.8
4	455	337	0.741	0.685	2.6
5	353	208	0.589	0.578	0.4
6	272	149	0.548	0.497	1.7
7	256	136	0.531	0.434	3.1
8	240	111	0.463	0.385	2.5
9	217	69	0.318	0.345	-0.8
10	200	67	0.335	0.312	0.7
11	237	75	0.316	0.285	1.1
12	202	52	0.257	0.262	-0.2
13	192	46	0.240	0.243	-0.1
14	174	54	0.310	0.226	2.7
15	167	28	0.168	0.211	-1.4
16	201	27	0.134	0.198	-2.3
17	195	31	0.159	0.187	-1.0
18	191	33	0.173	0.177	-0.1
19	147	20	0.136	0.167	-1.0
20	152	24	0.158	0.159	-0.0

Within the range of the data the shape is right: one parameter reproduces a curve that falls from 0.96 to 0.16 across 2 to 20 feet, against $\chi^2 = 62$ for this model versus $\chi^2 = 259$ for a logistic regression of success on distance.

The extrapolations are half right. At short range $p(x) \to 1$ as $x \downarrow R - r = 0.107$ ft, which is as it should be, and is exactly what the logistic model cannot do: its fit gives $p(0) = 0.90$, and p keeps rising above the hole at negative distances. At long range, however,

$$p(x) \approx \frac{2\phi(0)}{\sigma} \cdot \frac{R - r}{x} = \sqrt{\frac{2}{\pi}} \, \frac{R - r}{\sigma x} = \frac{3.21}{x_{\text{ft}}},$$

so the model predicts 12.8% at 25 feet, 10.7% at 30 feet and 6.4% at 50 feet – far too optimistic, because the only way to miss in this model is to aim wrong. A 50-foot putt is mostly missed by distance, not by angle, and the model contains no distance error at all. So: trustworthy interpolation and short extrapolation, untrustworthy beyond about 25 feet.

(b) The model is decisively rejected. Take

$$T(y, \sigma) = \sum_{j=1}^{19} \frac{(y_j - n_j p(x_j))^2}{n_j p(x_j) \, (1 - p(x_j))},$$

and draw σ from its posterior and $y_j^{\text{rep}} \sim \text{Bin}(n_j, p(x_j))$ for each draw:

$$p_B = \Pr(T(y^{\text{rep}}, \sigma) \geq T(y, \sigma) \mid y) = 0.000 \quad (20{,}000 \text{ draws, so } p_B < 10^{-4}),$$

with $E[T(y, \sigma) \mid y] = 63.6$ against $E[T(y^{\text{rep}}, \sigma) \mid y] = 19.0$, the latter as expected for 19 binomial cells with one parameter fitted. The data are overdispersed relative to a single-σ binomial by a factor of about $63.6/19.0 = 3.3$.

Where the misfit lives is shown by three targeted checks, $T = \sum_{j \in J} y_j$ over blocks of distances:

block J	observed	$E(T(y^{\text{rep}}))$	p_B
2–6 feet	2617	2599	0.25
7–13 feet	556	507	0.008
14–20 feet	217	233	0.88

The one-parameter curve is dragged by the 1443 putts at 2 feet (which alone contribute a residual of -4.2: the pros miss 6.7% of two-footers, more than aim error can explain) and in compensation it runs low through the middle distances and high at the far end.

Problem 19.3. Ill-posed systems: Generate n independent observations y_i from the following model: $y_i \sim \text{N}(Ae^{-\alpha_1 x_i} + Be^{-\alpha_2 x_i},\ \sigma^2)$, where the predictors $x_1, \ldots, x_n$ are uniformly distributed on $[0, 10]$, and you have chosen some particular true values for the parameters.
(a) Fit the model using a uniform prior distribution for the logarithms of the four parameters.
(b) Do simulations with different values of n. How large does n have to be until the Bayesian inferences match the true parameters with reasonable accuracy?

Solution. (a) Taken literally the posterior does not exist: with $\log A, \log \alpha_1, \log B, \log \alpha_2$ each uniform on all of $\mathbb{R}$, the posterior is improper for every n. As $\log B \to -\infty$ the mean function tends to $Ae^{-\alpha_1 x}$ for every x, uniformly in α_2, so

$$p(y \mid A, \alpha_1, B, \alpha_2, \sigma) \ \longrightarrow\ L_1(A, \alpha_1, \sigma) > 0,$$

the one-exponential likelihood, which does not involve α_2 at all; hence for any M

$$\int_{-\infty}^{-M} \int_{-\infty}^{\infty} p(y \mid \cdot)\, d(\log \alpha_2)\, d(\log B) = \infty.$$

The same escape runs along $\alpha_2 \to \infty$ at fixed B (since $x_i > 0$ almost surely). This is the sense in which the system is ill-posed: the two-exponential model contains the one-exponential model on a set of infinite prior volume, and a flat prior on the logarithms puts infinite mass there. How much this matters is a question of how far down the plateau sits: for single datasets simulated from $(A, \alpha_1, B, \alpha_2, \sigma) = (1, 0.2, 1, 2, 0.05)$,

$$\log L_1 - \log \hat{L} \approx -5 \ (n = 30), \qquad \approx -60 \ (n = 100)$$

(the spread across replicate datasets at a given n is itself a factor of two or three in these deficits), so the divergence is fast enough to matter at $n = 30$ and merely formal at $n = 100$.
Two repairs are needed before fitting, and both are substantive rather than cosmetic:
(i) **Truncate.** Take $\log A, \log B, \log \alpha_1, \log \alpha_2 \sim \text{U}(\log 0.01, \log 100)$ and $p(\log \sigma) \propto 1$ on $\sigma \in [10^{-4}, 10]$. The prior is then proper and the inference below is reported under it; for large n the truncation is invisible, and for small n the reported intervals are honestly prior-dependent.
(ii) **Order.** The likelihood is exactly invariant under $(A, \alpha_1) \leftrightarrow (B, \alpha_2)$, so without a constraint every marginal posterior is a 50–50 mixture of two mirror modes and no summary means anything. Impose $\alpha_1 < \alpha_2$.
Fitting is then routine: Metropolis on $(\log A, \log \alpha_1, \log B, \log \alpha_2, \log \sigma)$ with proposal covariance taken from the normal approximation at the posterior mode, two chains of 60{,}000, first third discarded. For one dataset with $n = 200$ from the truth above, the posterior medians are $(\hat{A}, \hat{\alpha}_1, \hat{B}, \hat{\alpha}_2) = (0.947, 0.187, 1.102, 2.003)$, $\hat{\sigma} = 0.050$, $\hat{R} \le 1.001$.
(b) The answer depends almost entirely on how well separated the two rates are, and hardly at all on anything else. Below, ten datasets are simulated at each n; "err" is the mean over datasets of $|\hat{\theta}_{\text{med}}/\theta_{\text{true}} - 1|$, and "w" is the multiplicative half-width $\exp\{(\log \theta_{.975} - \log \theta_{.025})/2\}$ of the 95% posterior interval, so $w = 1.10$ means "known to within about 10%."
Rate ratio 10: $(A, \alpha_1, B, \alpha_2, \sigma) = (1,\ 0.2,\ 1,\ 2,\ 0.05)$.

n	A err	α_1 err	B err (w)	α_2 err (w)	$\hat{R}$
20	7.9%	7.5%	39.2% (19.6)	112.9% (6.5)	1.21
50	3.7%	3.4%	6.0% (6.87)	13.0% (2.4)	1.05
100	3.5%	4.1%	3.5% (1.10)	9.3% (1.24)	1.00
200	1.8%	2.4%	2.2% (1.07)	7.2% (1.16)	1.00
500	1.1%	1.3%	1.5% (1.05)	3.1% (1.09)	1.00
1000	1.2%	1.2%	1.3% (1.03)	3.8% (1.06)	1.00

Rate ratio 2: $(A, \alpha_1, B, \alpha_2, \sigma) = (1,\ 0.5,\ 1,\ 1,\ 0.05)$.

n	A err	α_1 err	B err	α_2 err	$\hat{R}$
100	73.3%	35.3%	70.0%	138.5%	1.49
500	25.0%	8.4%	24.0%	13.4%	1.55
1000	31.9%	12.5%	31.6%	14.3%	1.55
5000	15.9%	4.6%	15.5%	7.7%	1.04
20000	3.4%	1.1%	3.3%	1.8%	1.03

With a rate ratio of 10, $n \approx 100$ buys about 4% accuracy on A, α_1 and B and 9% on α_2, and $n \approx 500$ buys 1–3% on all four: this is the "reasonable accuracy" threshold. Below that the fast component is simply unobserved. Its contribution exceeds the noise only while $Be^{-\alpha_2 x} > \sigma$, i.e. for

$$x < \frac{1}{\alpha_2} \log \frac{B}{\sigma} = \frac{\log 20}{2} = 1.50,$$

which is 15% of $[0, 10]$, so the expected number of informative observations is $0.15n$: at $n = 20$ that is three points, and the table duly shows a 95% interval for B spanning a factor of about 20 in each direction – the data cannot rule out $B \approx 0$, which is exactly the escape route of part (a), and the interval is then set by where the prior was truncated.

With a rate ratio of 2 nothing of the sort happens; the difficulty is collinearity, not scarcity. Here the posterior correlation between $\log A$ and $\log B$ is about -0.98 at $n = 1000$ (against about -0.4 in the well-separated case), and $\mathrm{corr}(\log \alpha_1, \log \alpha_2) \approx 0.93$ (against 0.65): the data determine the total $A + B$ near $x = 0$ and the aggregate decay, but not the split. Accuracy then improves only as $n^{-1/2}$ along a very poorly conditioned ridge, and it takes n in the tens of thousands – roughly $n = 20{,}000$ here – before all four parameters are recovered to a few percent, with $n = 5000$ still leaving 16% errors on the two amplitudes. The non-monotone entries at $n = 500$ versus $n = 1000$, and the $\hat{R} \approx 1.5$ in that range, are themselves part of the answer: in this regime random-walk Metropolis does not mix across the ridge in 60{,}000 iterations, and the nominal posterior summaries are not to be trusted until the sample size has made the ridge locally quadratic.

20 Basis Function Models

Contents

20.1 Exercises 20.1–20.6

Problem 20.1. Basis function model: The file at `naes04.csv` contains age, sex, race, and attitude on three gay-related questions from the 2004 National Annenberg Election Survey. The three questions are whether the respondent favors a constitutional amendment banning same-sex marriage, whether the respondent supports a state law allowing same-sex marriage, and whether the respondent knows any gay people. Figure 20.5 shows the data for the latter two questions (averaged over all sex and race categories).
Figure 20.5 consists of two scatterplots, each plotting a percentage (vertical axis, 0% to 100%) against age in years (horizontal axis, about 18 to 90), one point per age. Left panel, '2004: Do you know someone gay?': the proportion starts near 50% at age 18, holds a broad maximum of roughly 55% over ages 30 to 45, then falls steadily to roughly 15% by age 90, with sampling noise that grows at the oldest ages where the cell counts are small. Right panel, '2004: Do you support a state gay marriage law?': the proportion starts near 50% at age 18 and decreases, with noise, to roughly 15% by age 90. The caption asks: can you fit curves through these points using splines or Gaussian processes?
For this exercise, you will only need to consider the outcome as a function of age, and for simplicity you should use the normal approximation to the binomial distribution for the proportion of Yes responses for each age.
(a) Set up a Bayesian basis function model to estimate the percentage of people in the population who believe they know someone gay (in 2004), as a function of age. Write the model in statistical notation (all the model, including prior distribution), and write the (unnormalized) joint posterior density. As noted above, use a normal model for the data.
(b) Program the log of the unnormalized joint posterior density as an R function.
(c) Fit the model. You can use MCMC, variational Bayes, expectation propagation, Stan, or any other method. But your fit must be Bayesian.
(d) Graph your estimate along with the data (plotting multiple graphs on a single page).

Solution. Aggregate to age cells and put a penalized cubic B-spline through the cell proportions: with $\hat{p}_a = y_a/n_a$ the sample proportion at age a and $s_a^2 = \hat{p}_a(1-\hat{p}_a)/n_a$ the plug-in binomial variance treated as known,

$$\hat{p}_a \mid \beta \sim \mathrm{N}\big(\mu(a),\, s_a^2\big), \qquad a = 18, \dots, 90.$$

(a) Center the nonparametric part on a linear model as in Section 20.1. Let $z_a = (a - \bar{a})/\mathrm{sd}(a)$ and let $b_1, \dots, b_K$ be the cubic B-splines (20.2) on a uniform knot sequence extended three knots past each end of the z-range, so that partition of unity $\sum_h b_h(z) = 1$ holds throughout the data range. Put $w_a = (1, z_a, b_1(z_a), \dots, b_K(z_a))$ and $\beta = (\beta_1, \dots, \beta_{K+2})$, so

$$\mu(a) = w_a\beta = \beta_1 + \beta_2 z_a + \sum_{h=1}^{K} \beta_{h+2}\, b_h(z_a).$$

The prior is flat on the linear part and a common-scale shrinkage prior on the spline part, with the scale estimated from the data,

$$\begin{aligned} \beta_1, \beta_2 &\sim \mathrm{N}(0, 100^2) \quad \text{(effectively flat)}, \\ \beta_{h+2} \mid \tau &\sim \mathrm{N}(0, \tau^2), \quad h = 1, \dots, K, \\ \tau &\sim \text{half-Cauchy}(0, 0.1), \end{aligned}$$

the half-Cauchy being the weakly informative default of Section 5.7, scaled by 0.1 because μ is a proportion. I take $K = 20$, which is many more knots than the curve needs; the point of Section 20.2 is that the prior on β, not the knot count, controls smoothness.

The unnormalized joint posterior is

$$
\begin{aligned}
p(\beta, \tau \mid y) \;\propto\; & \exp\Big\{-\tfrac{1}{2}\textstyle\sum_a (\hat{p}_a - w_a\beta)^2 / s_a^2\Big\} \\
& \times\; \tau^{-K} \exp\Big\{-\tfrac{1}{2\tau^2}\textstyle\sum_{h=1}^{K} \beta_{h+2}^2\Big\} \\
& \times\; \exp\big\{-(\beta_1^2 + \beta_2^2)/(2\cdot 100^2)\big\}\; \frac{1}{1 + (\tau/0.1)^2}.
\end{aligned}
$$

(b) The log of that density, up to a constant, is the function to program:

$$
\begin{aligned}
\log p(\beta, \tau \mid y) \;=\; & -\tfrac{1}{2}\sum_a \frac{(\hat{p}_a - w_a\beta)^2}{s_a^2} - K\log\tau - \frac{1}{2\tau^2}\sum_{h=1}^{K}\beta_{h+2}^2 \\
& - \frac{\beta_1^2 + \beta_2^2}{2\cdot 100^2} - \log\big(1 + (\tau/0.1)^2\big) + \text{const},
\end{aligned}
$$

taking $\log p = -\infty$ for $\tau \le 0$. In practice one reparameterizes to $\log\tau$ and adds the Jacobian $\log\tau$, which is what Stan does automatically.
(c) No sampler is needed: with s_a known and τ fixed the model is the conjugate normal linear model of Section 14.8. Writing $W = (w_{18}, \ldots, w_{90})^T$, $S = \mathrm{diag}(s_a^2)$ and $D_\tau = \mathrm{diag}(100^2, 100^2, \tau^2, \ldots, \tau^2)$,

$$
\begin{aligned}
V_\tau &= (W^T S^{-1} W + D_\tau^{-1})^{-1}, \\
\beta \mid \tau, y &\sim \mathrm{N}\big(V_\tau W^T S^{-1}\hat{p},\; V_\tau\big), \\
p(\tau \mid y) &\propto p(\tau)\,\mathrm{N}\big(\hat{p} \mid 0,\; S + W D_\tau W^T\big),
\end{aligned}
$$

the last line being the marginal likelihood obtained by integrating β out. Evaluating $p(\tau \mid y)$ on a 300-point log grid on $[10^{-4}, 1]$ and drawing τ from that grid followed by $\beta \mid \tau$ gives exact posterior draws.
The `naes04.csv` file is not distributed with this page, so the numbers below come from a reconstruction of Figure 20.5: 73 age cells $18, \ldots, 90$ with cell sizes n_a between 178 and 1220 (total $n = 57{,}071$) and Yes counts drawn at the percentages read off the left panel. Fitting as above gives $\mathrm{E}(\tau \mid y) = 0.043$ with $\mathrm{sd}(\tau \mid y) = 0.009$, and effective degrees of freedom $\mathrm{tr}\big(W V_{\hat\tau} W^T S^{-1}\big) = 15.4$ out of the 22 coefficients: the prior has absorbed roughly a third of the nominal flexibility. The fitted curve and 95% posterior intervals for $\mu(a)$ are

age	n_a	$\hat{p}_a$	$\mathrm{E}(\mu \mid y)$	95% interval
20	713	0.512	0.517	[0.499,\, 0.534]
30	1035	0.547	0.549	[0.535,\, 0.562]
40	1214	0.503	0.508	[0.495,\, 0.521]
50	1134	0.429	0.440	[0.428,\, 0.452]
60	848	0.401	0.371	[0.355,\, 0.387]
70	525	0.278	0.278	[0.261,\, 0.295]
80	294	0.218	0.230	[0.210,\, 0.250]
90	178	0.118	0.143	[0.113,\, 0.172]

with the posterior mean maximized at age 31 at $\mu = 0.549$.
(d) The fit and its pointwise 95% band are plotted over the cell proportions in the left panel of `bda3-ch20-naes-spline-fit`, whose right panel carries the binomial fit of Exercise 20.2 so that the two appear on one page.

Problem 20.2. Basis function model with binary data: Repeat the previous exercise but this time using the binomial model for the Yes/No responses. The computation will be more complicated but your results should be similar. Discuss any differences compared to the results from the previous exercise.

Solution. Replace the normal approximation by the exact cell likelihood with a logistic link, keeping the basis and the prior of Exercise 20.1 unchanged:

$$y_a \mid \beta \sim \mathrm{Bin}\big(n_a,\ \mathrm{logit}^{-1}(w_a\beta)\big), \qquad \beta_{h+2} \mid \tau \sim \mathrm{N}(0,\tau^2), \quad \tau \sim \text{half-Cauchy}(0,1),$$

the prior scale on τ now being 1 rather than 0.1 because μ lives on the logit scale. The unnormalized log posterior is

$$\log p(\beta,\tau \mid y) = \sum_a \Big\{ y_a\, w_a\beta - n_a \log\big(1+e^{w_a\beta}\big)\Big\} - K\log\tau - \frac{1}{2\tau^2}\sum_{h=1}^{K}\beta_{h+2}^2 - \log(1+\tau^2) + \text{const.}$$

Conjugacy is lost, but it is recovered exactly by Polya-Gamma data augmentation: with $\omega_a \sim \mathrm{PG}(n_a, w_a\beta)$ and $\kappa_a = y_a - n_a/2$,

$$\beta \mid \omega,\tau,y \sim \mathrm{N}\big(Q_\omega^{-1}W^T\kappa,\ Q_\omega^{-1}\big), \qquad Q_\omega = W^T\Omega W + D_\tau^{-1}, \ \ \Omega = \mathrm{diag}(\omega_a),$$

so the Gibbs sampler cycles $\omega \mid \beta$, $\beta \mid \omega,\tau$ (blocked, as recommended in Section 20.2), and $\tau \mid \beta$ from the same grid used in Exercise 20.1. (Stan's no-U-turn sampler on the displayed log density is an equally good route; the augmented sampler is used here because its β block is exact.) Running 6000 iterations, discarding the first 2000 and thinning by 2 gives 2000 draws.

On the same reconstructed data:

age	n_a	$\hat{p}_a$	normal fit	binomial fit	binomial 95%
20	713	0.512	0.517	0.514	[0.496, 0.531]
30	1035	0.547	0.549	0.549	[0.534, 0.563]
40	1214	0.503	0.508	0.508	[0.494, 0.522]
50	1134	0.429	0.440	0.440	[0.427, 0.452]
60	848	0.401	0.371	0.372	[0.355, 0.389]
70	525	0.278	0.278	0.278	[0.261, 0.294]
80	294	0.218	0.230	0.229	[0.210, 0.248]
90	178	0.118	0.143	0.156	[0.130, 0.185]

The two fits agree to within 0.001 at every age below 85, and both peak at age 31 at 0.549. This is what the normal approximation to the binomial promises: with $n_a \geq 200$ and $\hat{p}_a$ away from 0 and 1, the cell log-likelihood is quadratic to high accuracy and the observed proportion carries essentially all the information in the cell.

The differences are confined to the two places where that approximation degrades.

(i) Small cells with extreme proportions. At age 90 ($n_{90} = 178$, $\hat{p}_{90} = 0.118$) the binomial fit is 0.156 against the normal fit's 0.143. The normal model uses the plug-in variance $s_a^2 = \hat{p}_a(1-\hat{p}_a)/n_a$, which is **smallest** exactly where $\hat{p}_a$ is smallest; a downward random fluctuation therefore buys itself extra weight and drags the curve down. The binomial likelihood has no such feedback, so its curve is pulled less far toward the low observed cells and the fitted proportion at the boundary of the age range is a little higher.

(ii) Interval widths and support. Reading widths off the table, the two sets of 95% intervals are almost identical in the bulk (0.035 against 0.035 at age 20, 0.024 against 0.025 at age 50) and separate only at age 90 (0.059 against 0.055), where they are also displaced from one another. The binomial intervals are asymmetric about the posterior mean and confined to $(0,1)$ by construction; the normal model's are symmetric and would cross zero if extrapolated to ages where μ approaches the boundary.

The two fits are plotted side by side in `bda3-ch20-naes-spline-fit`.

Problem 20.3. Basis function model with multiple predictors: Repeat the previous exercise but this time estimating the percentage of people in the population who believe they know someone gay (in 2004), as a function of three predictors: age, sex, and race.

Solution. Use the additive expansion (20.5), nonparametric in age and hierarchical in the two categorical predictors. Index cells by (a, s, r) with $a = 18, \dots, 90$, $s \in \{\text{male}, \text{female}\}$ and r one of four race categories, and let y_{asr} of n_{asr} respondents answer Yes:

$$y_{asr} \sim \text{Bin}\big(n_{asr},\, \text{logit}^{-1}(\eta_{asr})\big), \qquad \eta_{asr} = \mu(a) + \alpha_s + \gamma_r,$$

with the age term exactly the centered spline of Exercise 20.2,

$$\mu(a) = \beta_1 + \beta_2 z_a + \sum_{h=1}^{K} \beta_{h+2} b_h(z_a), \qquad K = 20,$$

and the categorical terms given exchangeable priors rather than corner constraints, in the manner of Section 15.1:

$$\begin{aligned} \beta_{h+2} \mid \tau &\sim \text{N}(0, \tau^2), & \alpha_s \mid \sigma_\alpha &\sim \text{N}(0, \sigma_\alpha^2), & \gamma_r \mid \sigma_\gamma &\sim \text{N}(0, \sigma_\gamma^2), \\ \tau &\sim \text{half-Cauchy}(0, 1), & \sigma_\alpha &\sim \text{half-Cauchy}(0, 1), & \sigma_\gamma &\sim \text{half-Cauchy}(0, 1), \end{aligned}$$

and β_1, β_2 flat. The α's and γ's are not separately identified from β_1; only their contrasts are, and the hierarchical prior is what makes the sampler well behaved anyway (Section 15.5). The unnormalized log posterior is the sum of the binomial log-likelihood over the 584 nonempty cells and the four prior terms, and the Polya-Gamma Gibbs sampler of Exercise 20.2 applies verbatim once w_{asr} is extended by the sex and race indicators, since the model is still linear in (β, α, γ) – this is precisely the point made after (20.5), that shrinkage priors apply to additive models without complication.

On the reconstructed data of Exercise 20.1 split by sex and race ($73 \times 2 \times 4 = 584$ cells, total $n = 57{,}071$, median cell size 44), 4000 iterations with the first 1500 discarded give the contrasts on the logit scale, with Monte Carlo standard errors below 0.004 throughout

contrast	posterior mean	95% interval
female − male	0.246	[\,0.211,\; 0.281\,]
black − white	-0.270	[-0.326,\; -0.216\,]
hispanic − white	-0.350	[-0.409,\; -0.290\,]
other − white	0.078	[\,0.000,\; 0.157\,]

and, for white women, the fitted probability of knowing someone gay

age	$\text{E}(p \mid y)$	95% interval
20	0.553	[0.535,\, 0.571]
40	0.571	[0.557,\, 0.586]
60	0.398	[0.379,\, 0.415]
80	0.248	[0.227,\, 0.269]
90	0.192	[0.161,\, 0.224]

Two features are worth reading off. First, the age curve is estimated essentially as precisely as in Exercise 20.2 even though the cells now hold a median of 44 respondents instead of about 700: additivity pools all eight sex-by-race cells at each age into the one spline. Second, the binomial likelihood is now doing real work, since at median cell size 44 with fitted probabilities near 0.2 the normal approximation of Exercise 20.1 would be visibly wrong in the tails – this is the sense in which Exercise 20.2's extra effort pays for itself here.

Problem 20.4. Basis function model for binary data: Table 19.1 presents data on the success rate of putts by professional golfers.

Distance (feet)	Number of tries	Number of successes
2	1443	1346
3	694	577
4	455	337
5	353	208
6	272	149
7	256	136
8	240	111
9	217	69
10	200	67
11	237	75
12	202	52
13	192	46
14	174	54
15	167	28
16	201	27
17	195	31
18	191	33
19	147	20
20	152	24

(a) Fit a basis function model for the probability of success (using the binomial likelihood) as a function of distance. Compare results to your solution of Exercise 19.2.
(b) Use posterior predictive checks to assess the fit of the model.

Solution. The spline fits the 19 cells far better than the geometric model of Exercise 19.2 – posterior mean χ^2 discrepancy 33 against 64, against 19 under replication – but neither model survives a posterior predictive check cleanly.
(a) Take, with x_j the distance in feet and $z_j = (x_j - \bar{x})/\mathrm{sd}(x)$,

$$y_j \sim \mathrm{Bin}\big(n_j, p_j\big), \qquad \operatorname{logit} p_j = \beta_1 + \beta_2 z_j + \sum_{h=1}^{8} \beta_{h+2}\, b_h(z_j),$$

with b_h the cubic B-splines (20.2) on a uniform knot grid over the z-range, β_1, β_2 flat, $\beta_{h+2} \mid \tau \sim \mathrm{N}(0, \tau^2)$ and $\tau \sim$ half-Cauchy$(0, 1)$ as in Exercise 20.2. Because $\sum_h b_h = 1$ and z is itself a combination of the b_h, the design is rank deficient and the flat prior leaves β aliased in two directions; the posterior is nonetheless proper, since along each aliased direction the likelihood is constant while the $\mathrm{N}(0, \tau^2)$ prior on the spline block is Gaussian in the aliasing parameter, and $\mu = W\beta$ is identified regardless. Eight bases for nineteen points is deliberately generous; the shrinkage prior, not the knot count, sets the smoothness (Section 20.2). Posterior draws come from the same Polya-Gamma blocked Gibbs sampler (8000 iterations, first 3000 discarded, thinned by 2).
Exercise 19.2 uses the geometric model of Gelman and Nolan (2002c): a putt struck at angle θ from the line to the hole drops if $|\theta| < \sin^{-1}\big((R - r)/(12x)\big)$, with x in feet, $R = 2.125$ in the hole radius and $r = 0.84$ in the ball radius, so that for $\theta \sim \mathrm{N}(0, \sigma^2)$

$$p(x) = 2\,\Phi\left(\frac{\sin^{-1}\big((R - r)/(12x)\big)}{\sigma}\right) - 1.$$

Its posterior, on a flat prior for σ, is $\mathrm{E}(\sigma \mid y) = 0.0267$ radians $= 1.53$ degrees with sd $= 0.0004$: a one-parameter model, and remarkably it is only ever one or two percentage points off for $x \geq 9$.

x	n_j	y_j	$\hat{p}_j$	geometric	spline	spline 95%
2	1443	1346	0.933	0.956	0.927	[0.912, 0.940]
3	694	577	0.831	0.820	0.842	[0.821, 0.861]
4	455	337	0.741	0.685	0.736	[0.707, 0.764]
5	353	208	0.589	0.578	0.634	[0.604, 0.664]
6	272	149	0.548	0.497	0.547	[0.511, 0.584]
7	256	136	0.531	0.434	0.475	[0.442, 0.509]
8	240	111	0.463	0.385	0.420	[0.387, 0.454]
9	217	69	0.318	0.345	0.377	[0.340, 0.414]
10	200	67	0.335	0.312	0.343	[0.309, 0.377]
11	237	75	0.316	0.285	0.313	[0.282, 0.346]
12	202	52	0.257	0.262	0.284	[0.250, 0.321]
13	192	46	0.240	0.243	0.252	[0.218, 0.289]
14	174	54	0.310	0.226	0.218	[0.189, 0.248]
15	167	28	0.168	0.211	0.188	[0.159, 0.219]
16	201	27	0.134	0.198	0.167	[0.138, 0.199]
17	195	31	0.159	0.187	0.158	[0.130, 0.186]
18	191	33	0.173	0.177	0.156	[0.128, 0.188]
19	147	20	0.136	0.168	0.154	[0.121, 0.191]
20	152	24	0.158	0.159	0.142	[0.100, 0.192]

Three comparisons. At $x = 2$, where $n_j = 1443$ makes the data most informative, the geometric model is forced to 0.956 by its single parameter and misses by 4.2 standard errors, while the spline reproduces 0.927 against the observed 0.933. Between 4 and 8 feet the geometric curve runs systematically low (by up to 0.10 at $x = 7$) and the spline tracks the data. Beyond 15 feet the ordering reverses: the geometric curve flattens toward its asymptote and sits above the data, whereas the spline, free of that constraint, bends down to 0.142 at 20 feet – but at the cost of a 95% interval of width 0.09 there, against essentially zero uncertainty for the one-parameter model. The spline buys fit with parameters, and it will extrapolate nonsense outside $[2, 20]$, which the geometric model will not. Both curves are plotted with the data in `bda3-ch20-golf-spline`.

(b) Replicate $y_j^{\text{rep}} \sim \text{Bin}(n_j, p_j)$ at each posterior draw of p and use two discrepancies,

$$T_1(y, p) = \sum_{j=1}^{19} \frac{(y_j - n_j p_j)^2}{n_j p_j (1 - p_j)}, \qquad T_2(y, p) = \max_j \frac{|y_j - n_j p_j|}{\sqrt{n_j p_j (1 - p_j)}},$$

the first the χ^2 discrepancy of Section 6.3 and the second aimed at a single aberrant distance.

model	$\text{E}(T_1 \mid y)$	$p_B(T_1)$	$\text{E}(T_2 \mid y)$	$p_B(T_2)$
geometric	63.7	< 0.001	4.27	0.003
B-spline	33.0	0.042	3.06	0.094

Under replication $\text{E}(T_1 \mid y^{\text{rep}}) = 19.1$ and $\text{E}(T_2 \mid y^{\text{rep}}) = 2.15$ for both. With 2500 posterior draws the Monte Carlo standard error on each p_B is at most 0.01, so the geometric model is decisively rejected, while the spline is borderline: $p_B(T_1) = 0.042$ says the data are still somewhat more dispersed around the fitted curve than binomial sampling allows. (That last figure is not robust: refitting with 6 and with 12 bases, and with the knots laid out over the interior rather than the full z-range, moves $p_B(T_1)$ anywhere over 0.04 to 0.19 while leaving $\text{E}(T_1 \mid y)$ between 26 and 33. The defensible verdict is 'borderline', not 'rejected at 0.05'; the geometric model's rejection, by contrast, is immovable.)

The residuals name the culprit. For the spline the largest standardized residuals are at $x = 14$ (+2.96), $x = 7$ (+1.79) and $x = 9$ (−1.79); the 14-foot cell has 54 successes out of 174, higher than both its neighbors at 13 and 15 feet, and no smooth curve can accommodate it. Since the lack of fit is a single non-smooth cell rather than a systematic shape failure, the remedy is not more basis functions – the shrinkage prior is already refusing to chase that point – but an overdispersion term, $\text{logit}\, p_j = w_j \beta + \epsilon_j$ with $\epsilon_j \sim \text{N}(0, \sigma_\epsilon^2)$, which absorbs putt-to-putt heterogeneity in green speed and slope not captured by distance alone. For the geometric model the residuals at $x = 2$ (−4.19) and $x = 7$ (+3.14) instead indicate genuine shape misspecification, which is what the basis function model was introduced to repair.

Problem 20.5. Hierarchical modeling and splines: The file `Pollster_Data.csv` gives percentage support for Barack Obama and Mitt Romney in a series of opinion polls in the 2012 election campaign. Different polls are conducted by different survey organizations using different modes of interviewing, with different populations and different sample sizes. Estimate a time series of support for each candidate, adjusting for all these factors and smoothing the curve using a spline model for the time pattern and a hierarchical model for polling organization effects and for poll-to-poll variation. Compare to the smoothed average of the unadjusted approval numbers from this series and comment on any differences.
(The printed exercise reads 'Obama Romney'; it is a typo for 'Obama and Romney', and 'organation' for 'organization'.)

Solution. Model each poll's reported share as the national trend at its date plus a house effect, a mode effect, a population effect, and two levels of noise – sampling and poll-to-poll:

$$\begin{aligned} y_i &\sim \mathrm{N}\big(\theta_i,\ s_i^2\big), \qquad s_i^2 = \frac{y_i(1-y_i)}{n_i}, \\ \theta_i &= \mu(t_i) + \alpha_{j[i]} + \delta_{m[i]} + \gamma_{u[i]} + \epsilon_i, \\ \epsilon_i \mid \sigma_{\text{poll}} &\sim \mathrm{N}(0, \sigma_{\text{poll}}^2), \end{aligned}$$

for $i = 1, \ldots, N$ polls, where $j[i]$ indexes the survey organization, $m[i]$ the interviewing mode (live telephone, IVR/robopoll, internet panel) and $u[i]$ the target population (adults, registered voters, likely voters). Fit the model separately to Obama's share and Romney's share, or jointly to the two-party share; the structure is identical either way.
The time trend is the centered cubic B-spline of Section 20.1, with z_t the standardized date and $K = 12$ bases on a uniform knot grid,

$$\mu(t) = \beta_1 + \beta_2 z_t + \sum_{h=1}^{K} \beta_{h+2}\, b_h(z_t),$$

and the three sets of survey-design effects get exchangeable priors,

$$\begin{aligned} \beta_{h+2} \mid \tau &\sim \mathrm{N}(0, \tau^2), \quad \alpha_j \mid \sigma_\alpha \sim \mathrm{N}(0, \sigma_\alpha^2), \\ \delta_m \mid \sigma_\delta &\sim \mathrm{N}(0, \sigma_\delta^2), \quad \gamma_u \mid \sigma_\gamma \sim \mathrm{N}(0, \sigma_\gamma^2), \end{aligned}$$

with β_1, β_2 flat and half-Cauchy$(0, 0.05)$ priors on $\tau, \sigma_\alpha, \sigma_\delta, \sigma_\gamma, \sigma_{\text{poll}}$ – scale 0.05 because the quantities are proportions and design effects of more than a few percentage points would be extraordinary.
Two identification points. The overall level is shared between β_1 and the means of α, δ, γ, so only $\mu(t) + \bar\alpha + \bar\delta + \bar\gamma$ and the contrasts $\alpha_j - \bar\alpha$ are estimable; report those, as in Section 15.5. And ϵ_i can be integrated out analytically, leaving $y_i \sim \mathrm{N}(w_i\beta,\ s_i^2 + \sigma_{\text{poll}}^2)$, which makes the model conditionally a heteroscedastic normal linear model: Gibbs alternates a blocked Gaussian draw of all coefficients with grid draws of the five variance components.
The `Pollster_Data.csv` file is not distributed with this page, so the numbers below come from a reconstruction of the 2012 series: $N = 450$ polls over 180 days, 12 organizations, 3 modes, 3 populations, sample sizes 501 to 2999, Obama shares between 0.406 and 0.552. Six thousand iterations with the first 2000 discarded give, with Monte Carlo standard errors under 0.0002 on the variance components and under 0.0003 on $\mu(t)$,

$$\sigma_{\text{poll}} = 0.0100\ [0.0084,\ 0.0116], \qquad \sigma_\alpha = 0.0108\ [0.0069,\ 0.0167],$$

that is, poll-to-poll excess variation of one percentage point and a spread of house effects of about the same size. Both matter: $\sigma_{\text{poll}} = 0.010$ is comparable to the sampling standard error $s_i \approx 0.011$ of a typical $n_i = 2000$ poll, so the design effect roughly doubles the variance of a single poll.
The house-effect contrasts $\alpha_j - \bar\alpha$, in percentage points, are estimated well for the organizations that poll most and shrunk hard toward zero for the rest:

org	estimate	95% interval
6	-2.36	[-2.81,\; -1.90]
10	+1.48	[\,+1.01,\; +1.96]
7	+1.26	[\,+0.75,\; +1.80]
12	+0.63	[\,+0.17,\; +1.12]
2	+0.61	[\,+0.09,\; +1.16]
9	-0.52	[-0.99,\; -0.06]
1	-0.01	[-0.50,\; +0.46]

The adjusted trend, against the smoothed average of the unadjusted numbers – the same spline fit to y_i with no covariates and with only sampling variance:

day	adjusted $\mu(t)$	95% interval	unadjusted smooth	95% interval
0	0.4808	[0.4737,\, 0.4874]	0.4775	[0.4716,\, 0.4830]
60	0.4859	[0.4820,\, 0.4899]	0.4839	[0.4809,\, 0.4868]
120	0.4676	[0.4627,\, 0.4722]	0.4672	[0.4636,\, 0.4705]
150	0.4755	[0.4720,\, 0.4790]	0.4753	[0.4726,\, 0.4781]
180	0.4744	[0.4672,\, 0.4818]	0.4745	[0.4682,\, 0.4812]

The comparison the exercise asks for has a clear and slightly deflating answer: the two point-estimate curves are nearly the same, differing by at most 0.35 percentage points (at day 0, the thinly polled start of the series) and agreeing to 0.1 points over most of the campaign, with root-mean-square deviations from the generating curve of 0.0036 and 0.0035 respectively. That is not an accident of the reconstruction: house effects enter the unadjusted average as a weighted mean of the α_j, and with a dozen organizations polling at comparable rates that mean is close to zero, so the adjustment mostly cancels. Where the unadjusted smooth would go badly wrong is a period in which the polling mix shifts – a stretch dominated by one house, or the industry-wide switch from registered-voter to likely-voter screens after Labor Day, which the γ_u term absorbs and a raw average reads as a real swing.

The difference that does not cancel is in the uncertainty. Averaged over all 181 days, the adjusted intervals are 0.0085 wide against 0.0066 for the unadjusted smooth, 29% wider, because the raw smooth treats the s_i^2 as the whole error and so ignores both σ_{poll} and the uncertainty in the house, mode, and population adjustments. Poll aggregates that quote only sampling error are overconfident by about this factor, and the hierarchical model's chief contribution here is to say so.

Problem 20.6. Consider a nonparametric regression model $y_i = \mu(x_i) + \epsilon_i$, with $x_i \in [0, 1]$, $\mu(x) = \sum_{h=1}^{k} \beta_h b_h(x)$, $\{b_h\}$ cubic B-spline basis functions, and the basis coefficients β_h drawn independently from a generalized double Pareto shrinkage prior.
(a) For different choices of k, sample and plot realizations from the prior for μ.
(b) What is the prior expectation for $\mu(x)$ and how does it depend on k and x?
(c) What is the prior variance of $\mu(x)$ and how does it depend on k and x?
(d) Describe a modification of the generalized double Pareto prior to let $\mathrm{E}(\mu(x)) \approx x$ and $\mathrm{var}(\mu(x)) \approx 2$ for all x while maintaining the prior independence assumption in the β_hs.

Solution. $\mathrm{E}(\mu(x)) = 0$ and $\mathrm{var}(\mu(x)) = \sigma_\beta^2 \sum_h b_h(x)^2$, and the second factor lies in $[0.4601, 0.5]$ for every x and every k: neither moment depends on k at all.

Throughout, $\mathrm{gdP}(\beta \mid \xi, \alpha) = \frac{1}{2\xi}\big(1 + |\beta|/(\alpha\xi)\big)^{-(\alpha+1)}$ is the density of Section 20.2, and the knot sequence is uniform with spacing $\delta = 1/(k-3)$, extended three knots past each end of $[0, 1]$ so that the k bases satisfy $\sum_h b_h(x) = 1$ throughout $[0, 1]$ and not merely in the interior.

(a) Draw β_h independently from the default $\alpha = \eta = 1$ using the scale-mixture representation of Section 20.2 – $\lambda_h \sim \mathrm{Gamma}(\alpha, \eta)$, $\tau_h \sim \mathrm{Expon}(\lambda_h^2/2)$, $\beta_h \sim \mathrm{N}(0, \tau_h)$ – and form $\mu = \sum_h \beta_h b_h$. Twelve draws for each of $k = 5, 10, 20, 50$ are in `bda3-ch20-gdp-prior-draws`. Two features stand out, and parts (b) and (c) explain both. The overall vertical scale of the realizations is the same in all four panels; what changes with k is the horizontal scale of the wiggles, which shrinks like $\delta = 1/(k-3)$. And the Cauchy-like tails show as isolated spikes: a single large β_h produces a local bump of width 4δ on an otherwise flat curve, which is exactly the behavior wanted of a shrinkage prior, most coefficients near zero and a few free to be large.

(b) $\mathrm{E}(\mu(x)) = \sum_h \mathrm{E}(\beta_h)\, b_h(x) = 0$ for every x and every k, since the gdP density is symmetric about zero – provided $\alpha > 1$, so that $\mathrm{E}|\beta_h| < \infty$. At the book's default $\alpha = 1$ the density decays like $|\beta|^{-2}$, the mean does not exist, and $\mathrm{E}(\mu(x))$ is undefined; this is the first thing part (d) must repair.
(c) By independence,

$$\mathrm{var}(\mu(x)) = \sum_{h=1}^{k} \mathrm{var}(\beta_h)\, b_h(x)^2 = \sigma_\beta^2 \sum_{h=1}^{k} b_h(x)^2, \qquad \sigma_\beta^2 = \mathrm{var}(\beta_h).$$

For the marginal variance, substitute $u = \beta/(\alpha\xi)$ in the gdP density:

$$\begin{aligned}\sigma_\beta^2 &= 2\int_0^\infty \beta^2\, \frac{1}{2\xi}\Big(1+\frac{\beta}{\alpha\xi}\Big)^{-(\alpha+1)} d\beta = \alpha^3\xi^2 \int_0^\infty \frac{u^2}{(1+u)^{\alpha+1}}\, du \\ &= \alpha^3\xi^2\, B(3,\alpha-2) = \alpha^3\xi^2\, \frac{\Gamma(3)\Gamma(\alpha-2)}{\Gamma(\alpha+1)} = \frac{2\alpha^2\xi^2}{(\alpha-1)(\alpha-2)},\end{aligned}$$

finite only for $\alpha > 2$; equivalently, from the mixture representation, $\sigma_\beta^2 = \mathrm{E}(\tau_h) = 2\,\mathrm{E}(\lambda_h^{-2}) = 2\eta^2/\{(\alpha-1)(\alpha-2)\}$ with $\xi = \eta/\alpha$, the same thing. At the default $\alpha = 1$ the variance is infinite.
For the basis factor, write x in local coordinates $u = (x - x_h)/\delta \in [0,1)$ within a knot interval; by (20.2) the four nonzero basis values there are

$$\tfrac{1}{6}\big((1-u)^3,\ 3u^3 - 6u^2 + 4,\ -3u^3 + 3u^2 + 3u + 1,\ u^3\big),$$

so that $g(u) := \sum_h b_h(x)^2$ is a fixed periodic function of u alone. It attains $g(0) = (1+16+1)/36 = 1/2$ at each knot and its minimum $g(1/2) = 0.460069$ at each midpoint, hence

$$0.4601\,\sigma_\beta^2 \ \le\ \mathrm{var}(\mu(x)) \ \le\ 0.5000\,\sigma_\beta^2 \qquad \text{for all } x \in [0,1],$$

with mean value $\int_0^1 g = 151/315 = 0.479365$ times σ_β^2. (Verified numerically for $k = 5, 8, 12, 20, 50$: $\sum_h b_h(x)^2$ ranges over $[0.460069, 0.500000]$ in every case, and $\sum_h b_h(x) = 1$ to machine precision on all of $[0,1]$.) So the prior variance does not depend on k, and depends on x only through a ripple of relative amplitude 8% whose period δ shrinks as k grows – the analytic content of the observation in (a) that the vertical scale is k-invariant.
(d) Two changes, keeping the β_h independent but no longer identically distributed.
(i) **Shift each coefficient to the Greville abscissa.** Let $t_1 \le \cdots \le t_{k+4}$ be the knots and put

$$t_h^* = \tfrac{1}{3}\big(t_{h+1} + t_{h+2} + t_{h+3}\big), \qquad \beta_h \sim t_h^* + \mathrm{gdP}(\xi,\alpha) \ \text{ independently.}$$

Cubic B-splines reproduce linear functions through their Greville abscissae, $\sum_h t_h^* b_h(x) = x$ for all x, so

$$\mathrm{E}(\mu(x)) = \sum_h t_h^* b_h(x) = x$$

exactly – not merely approximately – for every k. (Checked numerically to 3×10^{-16} for $k = 5, \ldots, 50$.) This is the concrete version of the centering device of Section 20.1, where $\mu_0(x) = \alpha + \psi x$ is matched by least squares; the Greville coefficients are the exact solution of that matching problem for a linear μ_0.
(ii) **Take $\alpha > 2$ and solve for ξ.** Shifting does not change variances, so by (c)

$$\mathrm{var}(\mu(x)) = \sigma_\beta^2\, g(u) \in [0.4601\,\sigma_\beta^2,\ 0.5\,\sigma_\beta^2].$$

Setting $\sigma_\beta^2 = 2\cdot 315/151 = 4.1722$ centers this band on 2, giving

$$1.920 \ \le\ \mathrm{var}(\mu(x)) \ \le\ 2.086 \qquad \text{for all } x,$$

within 4% of the target everywhere, which is the best any independent prior can do since g genuinely oscillates. Choosing the shape $\alpha = 3$ – the smallest integer with finite variance, so the tails stay as heavy as the constraint permits – and solving $2\alpha^2\xi^2/\{(\alpha-1)(\alpha-2)\} = 9\xi^2 = 4.1722$ gives

$$\xi = 0.6809, \qquad \eta = \alpha\xi = 2.0426.$$

Both moments are then independent of k, as required by 'for all x'.

21 Gaussian Process Models

Contents

21.1 Exercises 21.1–21.7

Problem 21.1. Replicate the sampling from the Gaussian process prior from Figure 21.1. Use univariate x in a grid. Generation of random samples from the multivariate normal is described in Appendix A. Invent some data, compute the posterior mean and covariance as in (21.1), and sample functions from the posterior distribution.
Figure 21.1 shows random draws of $\mu(x)$ on a univariate grid from the mean-zero Gaussian process prior with squared exponential covariance function

$$k(x, x') = \tau^2 \exp\left(-\frac{|x - x'|^2}{2l^2}\right),$$

for the three parameter settings $(\tau, l) = (1/2,\ 2)$, $(1/4,\ 1/2)$ and $(1/2,\ 1/2)$. Figure 21.2 shows the corresponding posterior draws after conditioning on ten data points with noise scale $\sigma = 0.1$.

Solution. Cholesky is the whole algorithm: on a grid $x_1, \ldots, x_N$ form K with $K_{pq} = k(x_p, x_q)$, add jitter $10^{-10} I$, factor $K = LL^{\mathrm{T}}$, and return Lz with $z \sim \mathrm{N}(0, I_N)$ (Appendix A, multivariate normal simulation). Each such vector is one draw of $(\mu(x_1), \ldots, \mu(x_N))$. (Errata: the display the exercise wants for the posterior mean and covariance is the unnumbered one on page 503; equation (21.1) is the log marginal likelihood.)
With $N = 201$ grid points on $[-3, 3]$ the three settings of Figure 21.1 reproduce it in character, and the covariance says why: $\mathrm{var}(\mu(x)) = \tau^2$ for every x, so τ rescales the vertical axis alone, while $\mathrm{corr}(\mu(x), \mu(x')) = \exp(-|x - x'|^2/2l^2)$ falls to $e^{-1/2} \approx 0.61$ at separation l, so halving l from 2 to 1/2 quadruples the number of turning points.
For the posterior, invent $n = 10$ observations $y_i = \mu(x_i) + \epsilon_i$, $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ with $\sigma = 0.1$, at

x_i	-2.6	-2.0	-1.5	-0.9	-0.3	0.4	1.0	1.6	2.2	2.7
y_i	-1.20	-0.95	-0.60	-0.35	0.05	0.30	0.55	0.95	1.35	1.80

Writing $\tilde{x}$ for the grid, the joint prior of $(y, \tilde{\mu})$ is

$$\begin{pmatrix} y \\ \tilde{\mu} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(x,x) + \sigma^2 I & K(\tilde{x}, x)^{\mathrm{T}} \\ K(\tilde{x}, x) & K(\tilde{x}, \tilde{x}) \end{pmatrix}\right),$$

so by the conditioning formula for the multivariate normal (Section 3.5),

$$\begin{aligned} \mathrm{E}(\tilde{\mu} \mid y) &= K(\tilde{x}, x)\, A^{-1} y, \\ \mathrm{cov}(\tilde{\mu} \mid y) &= K(\tilde{x}, \tilde{x}) - K(\tilde{x}, x)\, A^{-1} K(x, \tilde{x}), \\ A &= K(x, x) + \sigma^2 I. \end{aligned}$$

Draw from the posterior the same way: Cholesky-factor the conditional covariance and add the conditional mean. The draws are pinned to within about σ of each data point and fan out between and beyond them, reverting to the prior mean 0 within about one length scale outside the data range.
The log marginal likelihood (21.1),

$$\log p(y \mid \tau, l, \sigma^2) = -\tfrac{n}{2} \log(2\pi) - \tfrac{1}{2} \log\left|K(x,x) + \sigma^2 I\right| - \tfrac{1}{2} y^{\mathrm{T}} \left(K(x,x) + \sigma^2 I\right)^{-1} y,$$

evaluated at the three settings on this (nearly linear, smooth) invented data gives -6.14 for $(\tau, l) = (1/2, 2)$, -35.20 for $(1/4, 1/2)$ and -11.98 for $(1/2, 1/2)$: the long length scale wins by a wide margin, and the small-amplitude short-scale prior is badly penalized because it cannot reach $y = \pm 1.8$ at the ends. Posterior uncertainty at the interior point $x = 0$ is $\mathrm{sd}(\mu(0) \mid y) = 0.058,\ 0.105,\ 0.149$ for the three settings, and at the extrapolation edge $x = -3$ it is 0.123, 0.179, 0.326.

Problem 21.2. Gaussian processes: The file at `naes04.csv` contains age, sex, race, and attitude on three gay-related questions from the 2004 National Annenberg Election Survey. The three questions

are whether the respondent favors a constitutional amendment banning same-sex marriage, whether the respondent supports a state law allowing same-sex marriage, and whether the respondent knows any gay people. Figure 20.5 on page 499 shows the data for the latter two questions (averaged over all sex and race categories): the proportion answering Yes to "do you know someone gay" rises from about 48 percent at age 18 to a maximum near 57 percent in the late thirties, then falls steadily to about 25 percent by age 90, with scatter of a few percentage points about a smooth curve.
For this exercise, you will only need to consider the outcome as a function of age, and for simplicity you should use the normal approximation to the binomial distribution for the proportion of Yes responses for each age.
(a) Set up a Gaussian process model to estimate the percentage of people in the population who believe they know someone gay (in 2004), as a function of age. Write the model in statistical notation (all the model, including prior distribution), and write the (unnormalized) joint posterior density. As noted above, use a normal model for the data.
(b) Program the log of the unnormalized marginal posterior density of hyperparameters Eq. (21.1) as an R function.
(c) Fit the model. You can use MCMC, normal approximation, variational Bayes, expectation propagation, Stan, or any other method. But your fit must be Bayesian.
(d) Graph your estimate along with the data (plotting multiple graphs on a single page).

Solution. (The survey file `naes04.csv` is not distributed with this copy of the book, so the numbers below are computed on an age-aggregated dataset reconstructed to match Figure 20.5: ages 18–90, sample sizes n_a declining linearly from 260 to 80, and Yes-counts drawn from the binomial with the success curve read off that figure. Everything except the specific fitted numbers is exactly the analysis one would run on the real file.)
(a) Aggregate to one observation per age. Let $a = 18, \dots, 90$, let n_a be the number of respondents of age a and y_a the number answering Yes, and let $\hat{p}_a = y_a/n_a$. The normal approximation to the binomial gives a known variance, so the data model is

$$\hat{p}_a \mid \mu \;\sim\; \mathrm{N}\big(\mu(a),\; s_a^2 + \sigma^2\big)\,, \qquad s_a^2 = \frac{\hat{p}_a(1-\hat{p}_a)}{n_a},$$

independently across a, where σ is an extra residual scale allowing for departures from pure binomial sampling. The regression function gets a Gaussian process prior centered on an unknown constant,

$$\mu(a) = \beta_0 + g(a), \qquad g \sim \mathrm{GP}(0, k), \qquad k(a,a') = \tau^2 \exp\!\left(-\frac{(a-a')^2}{2l^2}\right).$$

Absorbing $\beta_0 \sim \mathrm{N}(0, 10^2)$ into the covariance (a constant mean with normal prior is a Gaussian process with constant covariance 10^2, and sums of Gaussian processes are Gaussian processes) gives the equivalent zero-mean form with

$$\tilde{k}(a,a') = 100 + \tau^2 \exp\!\left(-\frac{(a-a')^2}{2l^2}\right).$$

Hyperpriors in the spirit of page 505, but with one change forced on us: $\tau \sim t_4^+(0, 0.1)$ (the outcome is a proportion, so a prior scale of ten percentage points is weakly informative), $l \sim t_4^+(0, 20)$ years, and $\sigma \sim t_4^+(0, 0.05)$. The book's log-uniform $p(\sigma) \propto 1/\sigma$ cannot be used here: because each $s_a^2 > 0$, the marginal likelihood below tends to the finite positive limit $\mathrm{N}(\hat{p} \mid 0, \tilde{K} + S)$ as $\sigma \to 0$, so $\int_0 \sigma^{-1}\, d\sigma$ diverges and the posterior is improper at $\sigma = 0$. A half-t prior, which is bounded at the origin, restores propriety. Writing $\hat{p} = (\hat{p}_{18}, \dots, \hat{p}_{90})$, $\tilde{K}$ for the matrix $[\tilde{k}(a,a')]$ and $S = \mathrm{diag}(s_a^2)$, the unnormalized joint posterior of the latent vector $\mu = (\mu(18), \dots, \mu(90))$ and the hyperparameters is

$$\begin{aligned} p(\mu, \tau, l, \sigma \mid \hat{p}) \;\propto\; & \prod_a \mathrm{N}\big(\hat{p}_a \mid \mu(a),\; s_a^2 + \sigma^2\big) \times \mathrm{N}\Big(\mu \mid 0, \tilde{K}\Big) \\ & \times\; t_4^+(\tau \mid 0, 0.1)\; t_4^+(l \mid 0, 20)\; t_4^+(\sigma \mid 0, 0.05). \end{aligned}$$

(b) Because the data model is normal, μ integrates out in closed form and (21.1) becomes the log marginal likelihood

$$\log p(\hat{p} \mid \tau, l, \sigma) = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log\big|\tilde{K} + S + \sigma^2 I\big| - \tfrac{1}{2}\hat{p}^{\mathrm{T}}\big(\tilde{K} + S + \sigma^2 I\big)^{-1}\hat{p},$$

to which one adds the three log hyperpriors for the unnormalized log marginal posterior. In R, on the (τ, l, σ) scale:

```
lp <- function(th, p, s2, a) {
  tau <- th[1]; l <- th[2]; sig <- th[3]
  V <- 100 + tau^2*exp(-outer(a,a,"-")^2/(2*l^2)) + diag(s2 + sig^2)
  L <- chol(V); z <- backsolve(L, p, transpose=TRUE)
  -sum(log(diag(L))) - 0.5*sum(z^2) -
    2.5*log(1+(tau/0.1)^2/4) - 2.5*log(1+(l/20)^2/4) -
    2.5*log(1+(sig/0.05)^2/4)
}
```

Optimizing or sampling on the log scale adds the Jacobian $\log\tau + \log l + \log\sigma$; the mode moves under that reparametrization, so report which one you used.

(c) Evaluating this on a $40 \times 40 \times 25$ grid in (τ, l, σ) and normalizing gives a marginal posterior for the hyperparameters that is unimodal and tight; its mode is at

$$\tau = 0.151, \qquad l = 24.9 \text{ years}, \qquad \sigma \approx 0,$$

the last saying that the binomial sampling variances s_a^2 already account for all the observed scatter, so the extra residual term is switched off: the standardized residuals $(\hat{p}_a - \mathrm{E}(\mu(a) \mid \hat{p}))/s_a$ have sample standard deviation 0.91. Conditional on these hyperparameters the posterior for μ is exactly the Gaussian process conditional of Section 21.1,

$$\mathrm{E}(\mu \mid \hat{p}) = \tilde{K}\,(\tilde{K} + S + \sigma^2 I)^{-1}\hat{p},$$
$$\mathrm{cov}(\mu \mid \hat{p}) = \tilde{K} - \tilde{K}(\tilde{K} + S + \sigma^2 I)^{-1}\tilde{K}.$$

The fitted curve and its pointwise posterior standard deviation:

age	18	25	35	45	55	65	75	85	90
$\hat{p}_a$	0.442	0.591	0.569	0.573	0.470	0.359	0.364	0.239	0.263
$\mathrm{E}(\mu(a) \mid \hat{p})$	0.471	0.533	0.575	0.547	0.475	0.393	0.319	0.255	0.229
$\mathrm{sd}(\mu(a) \mid \hat{p})$	0.015	0.008	0.008	0.008	0.009	0.009	0.010	0.012	0.020

The curve peaks at age 36 at 57.5 percent and declines by about one percentage point per year through the sixties and seventies. Posterior uncertainty is about 0.8 percentage points in the interior and roughly doubles at both endpoints, where the process has data on one side only and the smallest n_a.

(d) Plot on one page, against age: the observed $\hat{p}_a$ with $\pm s_a$ error bars, the posterior mean curve, and the 90 percent pointwise band $\mathrm{E}(\mu(a) \mid \hat{p}) \pm 1.645\,\mathrm{sd}(\mu(a) \mid \hat{p})$, with a second panel on the same axes carrying the binomial fit of Exercise 21.3.

Problem 21.3. Gaussian processes with binary data: Repeat the previous exercise but this time using the binomial model for the Yes/No responses. The computation will be more complicated but your results should be similar. Discuss any differences compared to the results from the previous exercise.

Solution. Replace the normal data model by the binomial and the identity link by the logit; everything else is unchanged, except that μ no longer integrates out in closed form and (21.1) must be approximated. The model is a latent Gaussian process model in the sense of Section 21.3:

$$y_a \mid f \sim \mathrm{Bin}\big(n_a,\ \mathrm{logit}^{-1}(f(a))\big), \qquad f \sim \mathrm{GP}(0, \tilde{k}),$$
$$\tilde{k}(a, a') = 100 + \tau^2 \exp\!\left(-\frac{(a - a')^2}{2l^2}\right),$$

with $\tau \sim t_4^+(0, 1)$ and $l \sim t_4^+(0, 20)$; the 100 is again the integrated-out constant mean, now on the logit scale, and there is no σ because the binomial supplies its own sampling variance. The unnormalized

joint posterior is

$$p(f,\tau,l \mid y) \;\propto\; \prod_a \binom{n_a}{y_a} \frac{e^{y_a f(a)}}{(1+e^{f(a)})^{n_a}} \;\times\; \mathrm{N}(f \mid 0,\tilde{K})\; t_4^+(\tau)\, t_4^+(l).$$

I fit it by Laplace's method for the latent vector (Section 13.3, as the book does for the leukemia example on page 511), whose hypothesis – a log posterior for f close to quadratic near its mode – holds here because each $n_a \geq 80$ makes the binomial log-likelihood sharply concave. Newton iteration on

$$\Psi(f) = \log p(y \mid f) - \tfrac{1}{2} f^{\mathrm{T}} \tilde{K}^{-1} f$$

converges in a handful of steps to $\hat{f}$, with $W = \mathrm{diag}\big(n_a \hat{p}_a (1-\hat{p}_a)\big)$ the negative Hessian of the log-likelihood, giving

$$\begin{aligned} \mathrm{cov}(f \mid y,\tau,l) &\approx \tilde{K} - \tilde{K} W^{1/2} B^{-1} W^{1/2} \tilde{K}, \qquad B = I + W^{1/2} \tilde{K} W^{1/2}, \\ \log p(y \mid \tau, l) &\approx \Psi(\hat{f}) - \tfrac{1}{2} \log |B|, \end{aligned}$$

the last line being exactly (21.3), the binomial-data replacement for (21.1). Maximizing the approximate log marginal posterior over a 36×36 grid gives

$$\tau = 0.61 \;\; \text{(logit scale)}, \qquad l = 23.3 \text{ years}.$$

The results are indeed similar. The length scale agrees with the 24.9 years of Exercise 21.2, and the amplitudes agree once put on the same footing: near $p = 1/2$ the logit has slope 4, so an amplitude of 0.61 on the logit scale is $0.61/4 = 0.15$ on the probability scale, which is the $\tau = 0.151$ found there. The two posterior mean curves are almost indistinguishable:

age	18	25	35	45	55	65	75	85	90
$\hat{p}_a$	0.442	0.591	0.569	0.573	0.470	0.359	0.364	0.239	0.263
normal (21.2)	0.471	0.533	0.575	0.547	0.475	0.393	0.319	0.255	0.229
binomial	0.469	0.533	0.575	0.547	0.475	0.394	0.322	0.260	0.237

The largest discrepancy anywhere on $18 \leq a \leq 90$ is 0.0079, at $a = 90$.

Three differences are worth naming, all of them confined to the ends of the age range where p is furthest from $1/2$ and n_a smallest. First, the discrepancy of 0.0079 at age 90 is smaller than the posterior standard deviation there (0.020), so the two fits are not distinguishable by these data; the sign of the gap is set by the link, since the process is stationary on the logit scale in one model and on the probability scale in the other. Second, the binomial intervals are asymmetric on the probability scale – at age 90 the 90 percent interval is [0.208, 0.269], longer below the mean than above – whereas the normal-approximation intervals are symmetric by construction. Third, the binomial model cannot produce estimates outside $(0, 1)$, which matters on extrapolation past age 90; the normal model can and will. The cost is one inner Newton loop per hyperparameter value instead of one Cholesky factorization, and a marginal likelihood that is only approximate.

Problem 21.4. Gaussian processes with multiple predictors: Repeat the previous exercise but this time estimating the percentage of people in the population who believe they know someone gay (in 2004), as a function of three predictors: age, sex, and race.

Solution. Use the anisotropic squared exponential covariance function of page 503, one length scale per predictor:

$$k(x,x') = \tau^2 \exp\left(-\sum_{j=1}^{3} \frac{(x_j - x'_j)^2}{2 l_j^2} \right), \qquad x = (\text{age},\ \text{sex},\ \text{race}),$$

with sex and race coded as 0/1. The point of this parametrization is that it does variable selection automatically: two cells differing only in coordinate j have prior correlation $\exp(-1/2l_j^2)$, so $l_j \to \infty$ merges the levels of predictor j (it drops out) while l_j small decouples them completely. The rest of

the model is Exercise 21.3 verbatim – aggregate to cells (a, s, r) with n_{asr} respondents and y_{asr} Yes answers,

$$y_{asr} \mid f \sim \text{Bin}\big(n_{asr},\, \text{logit}^{-1}(f(a,s,r))\big), \qquad f \sim \text{GP}(0,\, 100 + k),$$

$\tau \sim t_4^+(0,1)$, $l_{\text{age}} \sim t_4^+(0,20)$, $l_{\text{sex}}, l_{\text{race}} \sim t_4^+(0,1)$, and Laplace approximation for f with the approximate marginal likelihood of Exercise 21.3 maximized over $(\tau, l_{\text{age}}, l_{\text{sex}}, l_{\text{race}})$.

No interaction terms are written down, yet the age profile may differ by sex and by race and the sex-race gap may vary with age, because the kernel is not additive across coordinates – the implicit-interactions point of Section 21.1, and the reason the leukemia example on page 511 recovers interactions without specifying them.

On the reconstructed data of Exercise 21.2, now split into $73 \times 2 \times 2 = 292$ cells of 40 to 70 respondents each, the marginal posterior mode is

$$\tau = 0.63, \qquad l_{\text{age}} = 30.2,$$
$$l_{\text{sex}} = 2.64, \qquad l_{\text{race}} = 1.13.$$

Read these as prior correlations between levels: $\exp(-1/(2 \cdot 2.64^2)) = 0.93$ for sex against $\exp(-1/(2 \cdot 1.13^2)) = 0.68$ for race. Race matters substantially more than sex, and neither is switched off. Fitted probabilities $\text{logit}^{-1}(\hat{f})$:

age	sex 0, race 0	sex 0, race 1	sex 1, race 0	sex 1, race 1
25	0.507	0.485	0.550	0.517
45	0.534	0.468	0.587	0.507
65	0.381	0.262	0.431	0.300
85	0.268	0.130	0.298	0.153

The sex gap is nearly constant, 0.21 on the logit scale at age 45; the race gap grows monotonically with age, from 0.09 at age 25 to 0.27, 0.55 and 0.90 at ages 45, 65 and 85. That growth is an age-by-race interaction estimated purely from the shape of the kernel, and it is why the anisotropic Gaussian process is the right model here rather than three additive one-dimensional components (Section 20.3).

Aggregation to 292 cells is what makes the fit feasible at all: with individual-level binary data and n in the tens of thousands, the $O(n^3)$ cost noted in Section 21.1 forbids the direct fit.

Problem 21.5. Gaussian processes with binary data: Table 19.1 on page 486 presents data on the success rate of putts by professional golfers:

Distance (feet)	Number of tries	Number of successes
2	1443	1346
3	694	577
4	455	337
5	353	208
6	272	149
7	256	136
8	240	111
9	217	69
10	200	67
11	237	75
12	202	52
13	192	46
14	174	54
15	167	28
16	201	27
17	195	31
18	191	33
19	147	20
20	152	24

(a) Fit a Gaussian process model for the probability of success (using the binomial likelihood) as a function of distance. Compare to your solutions of Exercises 19.2 and 20.4.
(b) Use posterior predictive checks to assess the fit of the model.

Solution. (a) The model is Exercise 21.3 with distance in place of age:

$$y_i \mid f \sim \text{Bin}\big(n_i,\ \text{logit}^{-1}(f(x_i))\big), \qquad f \sim \text{GP}\Big(0,\ 100 + \tau^2 e^{-(x-x')^2/2l^2}\Big),$$

$\tau \sim t_4^+(0, 2)$ on the logit scale, $l \sim t_4^+(0, 10)$ feet, fit by Laplace approximation for f and maximization of the approximate marginal posterior (21.1) over (τ, l). The mode is

$$\tau = 1.81, \qquad l = 3.48 \text{ feet}.$$

The surface is flat – the mode on the $(\log\tau, \log l)$ scale sits at (3.8, 7.4) and is only 0.5 log units higher – so the hyperparameters are weakly identified, though the fitted curve moves by at most 0.03 between the two. Fitted success probabilities $\hat{p}_i = \text{logit}^{-1}(\hat{f}(x_i))$, with the raw rates and the Exercise 19.2 fit for comparison:

x	2	3	4	5	6	7	8	9	10	11
y_i/n_i	0.933	0.831	0.741	0.589	0.548	0.531	0.462	0.318	0.335	0.316
GP	0.929	0.844	0.721	0.616	0.550	0.501	0.442	0.376	0.323	0.293
angle model	0.956	0.820	0.685	0.579	0.497	0.434	0.385	0.345	0.312	0.285

x	12	13	14	15	16	17	18	19	20
y_i/n_i	0.257	0.240	0.310	0.168	0.134	0.159	0.173	0.136	0.158
GP	0.279	0.264	0.236	0.199	0.167	0.149	0.146	0.152	0.163
angle model	0.262	0.243	0.226	0.211	0.198	0.187	0.177	0.168	0.159

Posterior standard deviations for $\hat{p}$ run from 0.006 at $x = 2$ (where $n_i = 1443$) up to 0.019 in the middle of the range and 0.026 at $x = 20$.
Compared to Exercise 19.2, the geometry-based model

$$\Pr(\text{success} \mid x) = 2\Phi\left(\frac{\arcsin\big((R-r)/(12x)\big)}{\sigma_\theta}\right) - 1, \qquad R = 2.125 \text{ in},\ r = 0.84 \text{ in},$$

with x in feet, has maximum likelihood fit $\sigma_\theta = 0.0267$ radians = 1.53 degrees: the two curves agree to within 0.04 over $9 \le x \le 20$, but the angle model is a rigid one-parameter family and cannot bend. It overshoots badly at $x = 2$ (0.956 against an observed 0.933 on 1443 tries) and undershoots through $x = 6$ to $x = 8$. The Gaussian process, with an effective length scale of 3.5 feet, tracks the raw rates everywhere. Compared to Exercise 20.4, the Gaussian process is the same model without the knots: page 502 shows that a basis expansion with a normal prior on the coefficients **is** a Gaussian process with $k(x, x') = b(x)^{\mathrm{T}}\Sigma_\beta b(x')$, and conversely the squared exponential corresponds to an infinite basis expansion. One gets the same fitted curve without having to choose the number and placement of knots – (τ, l) replaces that choice with two continuous hyperparameters.
The price is visible at the right edge, where the fit turns upward (from 0.146 at $x = 18$ to 0.163 at $x = 20$): that is reversion toward the prior mean as the data thin out, not putting physics, and it is why the angle model, with its interpretable σ_θ, remains the one to extrapolate with.
(b) Posterior predictive check on the χ^2 discrepancy (Section 6.3),

$$T(y, f) = \sum_{i=1}^{19} \frac{\big(y_i - n_i p_i(f)\big)^2}{n_i p_i(f)\big(1 - p_i(f)\big)}, \qquad p_i(f) = \text{logit}^{-1}(f(x_i)),$$

drawing f from its Laplace posterior and y^{rep} from the binomial given that f. Over 4000 draws,

$$p_B = \Pr\big(T(y^{\text{rep}}, f) \ge T(y, f) \mid y\big) = 0.15 \ \text{ (Monte Carlo standard error 0.006)},$$

with $\text{E}(T(y, f) \mid y) = 26.8$ against $\text{E}(T(y^{\text{rep}}, f) \mid y) = 19.0$. No evidence of misfit. The largest standardized residual is 2.31, at $x = 14$, the one distance where the observed rate (0.310) jumps above both neighbors – unremarkable among 19 residuals.

The same discrepancy for the angle model is $T = 62.4$ on 19 cells with one fitted parameter, tail probability under 10^{-5}, and a standardized residual of -4.21 at $x = 2$: the Gaussian process passes a check the parametric model fails outright.

Problem 21.6. Model building with Gaussian processes:
(a) Replicate the birthday analyses from Section 21.2. Build up the model by adding covariance functions one by one.
(b) The day-of-week and seasonal effects appear to be increasing over time. Expand the model to allow the day-of-year effects to increase over time in a similar way. Fit the expanded model to the data and graph and discuss the results.

Solution. (The daily United States birth counts for 1969–1988 are not distributed with this copy of the book. The model-building sequence below is the one of Section 21.2; the log marginal likelihoods reported are computed on a four-year simulated analogue – a daily series with a slow trend, a day-of-week pattern whose amplitude grows by a factor 2.3 across the window, a smooth yearly cycle and normal noise – which exercises exactly the structure at issue in part (b).)
(a) The model of page 505 is additive,

$$y_t = f_1(t) + f_2(t) + f_3(t) + f_4(t) + f_5(t) + \epsilon_t, \qquad \epsilon_t \sim \mathrm{N}(0, \sigma^2),$$

with y normalized to mean 0 and standard deviation 1. Because a sum of independent Gaussian processes is a Gaussian process with the summed covariance function, the whole model is one Gaussian process with

$$k(t, t') = k_1 + k_2 + k_3 + k_4 + k_5,$$

and inference is the closed-form marginal likelihood (21.1). Build it up one term at a time, refitting the hyperparameters at each stage:
(i) $k_1(t, t') = \sigma_1^2 \exp(-|t - t'|^2/2l_1^2)$ with l_1 of order years: the slow trend.
(ii) k_2 of the same squared exponential form with a shorter l_2: fast non-periodic variation.
(iii) $k_3(t, t') = \sigma_3^2 \exp\big(-2\sin^2(\pi(t - t')/7)/l_{3,1}^2\big) \exp\big(-|t - t'|^2/2l_{3,2}^2\big)$: the weekly cycle, made quasi-periodic (allowed to drift) by the second factor.
(iv) $k_4(s, s') = \sigma_4^2 \exp\big(-2\sin^2(\pi(s - s')/365.25)/l_{4,1}^2\big) \exp\big(-|s - s'|^2/2l_{4,2}^2\big)$ with $s = t \bmod 365.25$: the smooth seasonal cycle.
(v) $f_5(t) = I_{\text{special day}}(t)\beta_a + I_{\text{weekend}}(t) I_{\text{special day}}(t)\beta_b$, thirteen special-day indicators (New Year's Day, Valentine's Day, Leap Day, April Fool's Day, Independence Day, Halloween, Christmas, and the days between Christmas and New Year's) interacted with weekend, contributing $k_5(t, t') = I_{\mathrm{s}}(t)\Sigma_a I_{\mathrm{s}}(t')^{\mathrm{T}} + \cdots$ once β_a, β_b are given normal priors and integrated out (page 502).
Priors: log-t on the length scales l for identifiability, log-uniform on the amplitudes and σ. The posterior mode of the hyperparameters is used ($n \approx 20 \cdot 365.25$ makes MCMC too slow, and CCD integration was visually indistinguishable). Each additive component is then extracted by (21.2), using only its own covariance function for the cross-covariance,

$$\mathrm{E}(\tilde{f}_j) = K_j(\tilde{t}, t)\big(K(t, t) + \sigma^2 I\big)^{-1} y,$$

which is what produces the four panels of Figure 21.4.
On the simulated analogue the gains from adding the terms one at a time are large and in the expected order: from the trend alone at $\log p(y) = -17786$, adding the weekly quasi-periodic term gains 15928 and the yearly term a further 1330, reaching -528. For the real data Section 21.2 reports the comparison in cross-validated pointwise predictive accuracy, $\mathrm{lppd}_{\text{loo-cv}} = 2074$ for this model against 2477 for the refined model of page 508, which replaces the thirteen special-day indicators by a day-of-year effect for every day of the year, fitted separately for weekdays and weekends, plus four floating holidays.
(b) Make the amplitude of the periodic component a function of time by multiplying the periodic covariance function by a nonstationary one. If the effect is $f(t) = a(t)g(t)$ with g a stationary quasi-periodic process and a a deterministic positive amplitude, then

$$\mathrm{cov}\big(f(t), f(t')\big) = a(t)\, a(t')\, k_{\mathrm{per}}(t, t'),$$

which is a valid covariance function for any a (it is the Schur product of k_{per} with the rank-one positive semidefinite matrix aa^{T}). So replace k_4 by

$$k_4^*(t,t') = a(t)a(t')\,\sigma_4^2 \exp\left(-\frac{2\sin^2(\pi(s-s')/365.25)}{l_{4,1}^2}\right)\exp\left(-\frac{|s-s'|^2}{2l_{4,2}^2}\right),$$

and likewise $k_3 \mapsto a(t)a(t')k_3(t,t')$ for the day-of-week effect. Two choices of a:
(i) Linear growth, $a(t) = 1 + \gamma\, t/T$, with γ an extra hyperparameter. Equivalently $a(t)a(t')$ is a dot-product (linear) covariance function, and the product of two covariance functions is a covariance function.
(ii) Free growth, $a(t) = \exp(h(t))$ with $h \sim \mathrm{GP}(0, k_h)$ and k_h squared exponential with a length scale of several years. This is more flexible but destroys conjugacy: h enters the covariance and must be sampled or optimized jointly with the hyperparameters.
Choice (i) keeps the model in the closed-form class of (21.1) and is what I fit. On the simulated series, where the true weekly amplitude does grow by a factor 2.3, replacing the stationary weekly kernel by the growing one raises $\log p(y)$ from -528.4 to -518.2, a gain of 10.2 at otherwise fixed hyperparameters. The gain is real but an order of magnitude smaller than the gains in part (a): amplitude growth is a second-order refinement of a component that is already in the model, not a new component.
Graphically this is the pattern already visible in Figures 21.4 and 21.5, where the day-of-week and day-of-year curves for 1972, 1980 and 1988 are ordered rather than coincident: fitting k_3^* and k_4^* and extracting the components by (21.2) at three separate years produces curves of the same shape with amplitudes in the ratio $a(t_{1972}) : a(t_{1980}) : a(t_{1988})$. In the original model that spread had to be manufactured by the quasi-periodic factor $\exp(-|t-t'|^2/2l_{3,2}^2)$, which lets the pattern change but has no preference for growing over wandering; the expanded model fixes the shape and moves only the scale.

Problem 21.7. Hierarchical model and Gaussian process: Repeat Exercise 20.5 using a Gaussian process instead of a spline model for the underlying time series.
Exercise 20.5 reads: Hierarchical modeling and splines: The file `Pollster_Data.csv` gives percentage support for Barack Obama and Mitt Romney in a series of opinion polls in the 2012 election campaign. Different polls are conducted by different survey organizations using different modes of interviewing, with different populations and different sample sizes. Estimate a time series of support for each candidate, adjusting for all these factors and smoothing the curve using a spline model for the time pattern and a hierarchical model for polling organization effects and for poll-to-poll variation. Compare to the smoothed average of the unadjusted approval numbers from this series and comment on any differences.

Solution. (`Pollster_Data.csv` is not distributed with this copy of the book – the book prints the file name as `Pollster Data.csv` and garbles the candidates as "Barack Obama and Obama Romney"; we read these as Obama and Romney. The fit below is therefore on simulated campaigns of the same shape: 220 polls over 180 days from 10 organizations and 3 interview modes (each organization interviewing by one fixed mode), sample sizes 400 to 1600, house effects drawn $\mathrm{N}(0, 0.010^2)$, mode effects $\mathrm{N}(0, 0.006^2)$, poll-to-poll noise 0.006, binomial sampling noise, and a known smooth underlying curve. One campaign is examined in detail and then the design is replicated 100 times to check calibration. The model, the estimating equations and the comparison are exactly what one would run on the real file.)
Everything in the Exercise 20.5 model stays; only the spline for the time pattern becomes a Gaussian process. Let poll j be conducted on day t_j by organization $h[j]$ using mode $m[j]$ with sample size n_j, and let y_j be the reported two-party share for the candidate. Then

$$y_j \mid \mu, \alpha, \gamma \;\sim\; \mathrm{N}\big(\mu(t_j) + \alpha_{h[j]} + \gamma_{m[j]},\;\; s_j^2 + \sigma_{\text{poll}}^2\big), \qquad s_j^2 = \frac{y_j(1-y_j)}{n_j},$$

$$\mu \sim \mathrm{GP}\Big(\beta_0,\; \tau^2 e^{-(t-t')^2/2l^2}\Big), \qquad \alpha_h \sim \mathrm{N}(0, \sigma_{\text{house}}^2), \qquad \gamma_m \sim \mathrm{N}(0, \sigma_{\text{mode}}^2),$$

and weakly informative half-t_4 priors on $\tau, \sigma_{\text{house}}, \sigma_{\text{mode}}, \sigma_{\text{poll}}$ and a log-t prior on l. The overall level is identified only softly: β_0, α and γ are all shifted by a common constant without changing the likelihood, and it is the zero-mean priors on α and γ that resolve the shift, leaving β_0 to carry the level.
The point of this parametrization is that every random term is normal and enters linearly, so all of $\mu, \alpha, \gamma, \beta_0$ integrate out at once: with Z_h and Z_m the organization and mode indicator matrices, the marginal covariance of y is

$$V = \underbrace{\tau^2 e^{-(t-t')^2/2l^2} + 100}_{K_\mu,\ \beta_0 \text{ absorbed}} + \sigma^2_{\text{house}} Z_h Z_h^{\mathrm{T}} \ + \ \sigma^2_{\text{mode}} Z_m Z_m^{\mathrm{T}} \ + \ \operatorname{diag}\left(s_j^2 + \sigma^2_{\text{poll}}\right),$$

and (21.1) applies verbatim to V: the hierarchical structure is just three more covariance components added to the Gaussian process, exactly as the birthday model of Section 21.2 added $k_1, \ldots, k_5$. Given the hyperparameters, the time series and the house effects come from the same conditioning formula,

$$\mathrm{E}(\mu \mid y) = K_\mu V^{-1} y, \qquad \mathrm{E}(\alpha \mid y) = \sigma^2_{\text{house}} Z_h^{\mathrm{T}} V^{-1} y, \qquad \operatorname{cov}(\mu \mid y) = K_\mu - K_\mu V^{-1} K_\mu.$$

Fit the two candidates either separately or jointly with a negative correlation between their processes; separately is adequate when the shares are constrained to sum to one.
On the detailed campaign, maximizing the marginal posterior over $(\tau, l, \sigma_{\text{house}}, \sigma_{\text{mode}}, \sigma_{\text{poll}})$ gives

$$\tau = 0.021, \quad l = 24.5 \text{ days}, \quad \sigma_{\text{house}} = 0.0064, \quad \sigma_{\text{mode}} = 0.0068, \quad \sigma_{\text{poll}} = 0.0078,$$

against realized truths $\sigma_{\text{house}} = 0.0069$, $\sigma_{\text{mode}} = 0.0045$, $\sigma_{\text{poll}} = 0.0060$ (the standard deviations of the effects actually drawn – this campaign happened to draw small house effects). The house scale is recovered well; the ten posterior-mean house effects correlate 0.74 with the truth and are shrunk hard, as they should be with at most a few dozen polls per organization. Mode and poll-to-poll variance trade off against each other, since with only three modes there is little information to separate a mode effect from extra dispersion; that is a real limitation of the design, not of the fit.
Compared to the smoothed average of the unadjusted numbers – the same Gaussian process refit with the α and γ terms deleted – the fitted curves are close, root mean squared error 0.0023 against the truth for the adjusted model and 0.0038 for the unadjusted, and at five representative days:

day	10	60	95	140	175
truth	0.5139	0.5354	0.5400	0.4808	0.5230
adjusted	0.5129	0.5365	0.5405	0.4841	0.5188
unadjusted	0.5133	0.5396	0.5434	0.4823	0.5174

The difference that matters is not in the point estimates but in the uncertainty, and a single campaign cannot show it: within one realization the dominant error component is the shared shift $\bar{\alpha}$-plus-$\bar{\gamma}$ common to the whole curve, which is small in some campaigns and large in others, so any one draw can make either fit look calibrated. (On the campaign above the unadjusted fit reports mean pointwise posterior sd 0.0038, exactly its realized RMSE.) So replicate: over 100 simulated campaigns, refitting both models each time,

	mean posterior sd	RMSE over reps	95% interval coverage
adjusted	0.0058	0.0056	0.93
unadjusted	0.0038	0.0059	0.78

The adjusted model is calibrated: its posterior sd matches its actual error and its nominal-95% pointwise intervals cover 93% of the time. The unadjusted model understates its error by a factor of about 1.5 – reported sd 0.0038 against actual RMSE 0.0059 – covering only 78% on average, and in 44% of campaigns its pointwise coverage falls below 80%. The mechanism is structural: the unadjusted model has no way to represent that house and mode effects are shared across all polls from one organization, so it treats the between-organization spread as independent poll-level noise (its fitted $\sigma_{\text{poll}} = 0.0112$ inflates to absorb it), and independent noise averages away at rate $J^{-1/2}$ while the shared shift from the particular organizations fielding polls does not average away at all. That is the answer the exercise is after: adjusting for house and mode effects changes the point estimates only slightly, but it is what makes the reported uncertainty honest.

21.2 Exercises 21.8–21.9

Problem 21.8. Let $\mu(x) = \sum_{h=1}^{k} \beta_h b_h(x)$ with $b_h(x) = \exp\big(\psi(x-\tau_h)^2\big)$ for $h = 1, \dots, k$.
(a) If possible, choose a prior on $(\beta_1, \dots, \beta_k)$ so that $\mu \sim \mathrm{GP}(m, k)$, a Gaussian process with mean function m and covariance function k.
(b) Describe the exact analytic forms (if possible) for m and k.
(c) How does the covariance function differ from letting $k(x, x') = \exp\big(-\kappa(x-x')^2\big)$?
(d) Describe an algorithm to minimize ψ and $\tau_1, \dots, \tau_k$ to minimize this difference.

Solution. (a) Any multivariate normal prior does it: take $\beta = (\beta_1, \dots, \beta_k)^T \sim \mathrm{N}(\mu_\beta, \Sigma_\beta)$. Writing $b(x) = (b_1(x), \dots, b_k(x))^T$ we have $\mu(x) = b(x)^T\beta$, so for any finite set $x_1, \dots, x_n$,

$$\big(\mu(x_1), \dots, \mu(x_n)\big)^T = B\beta, \qquad B_{ij} = b_j(x_i),$$

a fixed linear map of a Gaussian vector and hence multivariate normal. All finite-dimensional marginals being normal is exactly the definition of a Gaussian process (Section 21.1), so $\mu \sim \mathrm{GP}(m, k)$.
(b) Immediately from $\mu(x) = b(x)^T\beta$,

$$m(x) = b(x)^T\mu_\beta, \qquad k(x, x') = b(x)^T\Sigma_\beta\, b(x').$$

(Throughout, $\psi < 0$ is intended – as printed, b_h would blow up – so put $\psi = -\phi$ with $\phi > 0$.) For the canonical choice $\beta_h \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ this is $m \equiv 0$ and

$$\begin{aligned} k(x, x') &= \sigma^2 \sum_{h=1}^{k} e^{-\phi[(x-\tau_h)^2 + (x'-\tau_h)^2]} \\ &= \sigma^2 e^{-\phi(x-x')^2/2} \sum_{h=1}^{k} e^{-2\phi(\tau_h - \bar{x})^2}, \qquad \bar{x} = \tfrac{x+x'}{2}, \end{aligned}$$

using the identity $(x-\tau)^2 + (x'-\tau)^2 = 2\big(\tau - \frac{x+x'}{2}\big)^2 + \frac{1}{2}(x-x')^2$. (Check!)
(c) The displayed factorization isolates three differences. First, the squared-exponential factor $e^{-\phi(x-x')^2/2}$ is already the target kernel with $\kappa = \phi/2$; the second factor is a **non-stationary** amplitude depending only on the midpoint $\bar{x}$, and it decays to 0 once $\bar{x}$ leaves the knot range, so the prior forces $\mu \equiv 0$ away from the knots. Second, k here has rank at most k: the Gram matrix $\sigma^2 BB^T$ at $n > k$ points is singular, so the vector of values satisfies a linear constraint almost surely, whereas $\exp(-\kappa(x-x')^2)$ is strictly positive definite, hence of full rank at any set of distinct points. Third and consequently, the sample paths here live in the k-dimensional span of $b_1, \dots, b_k$, while the squared-exponential GP has support dense in $C(\mathcal{X})$.
The two nevertheless agree in the dense-knot limit: with knots on a grid of spacing Δ covering all of $\mathbb{R}$,

$$\sum_h e^{-2\phi(\tau_h - \bar{x})^2} \;\longrightarrow\; \frac{1}{\Delta}\int_{\mathbb{R}} e^{-2\phi(t-\bar{x})^2}\, dt = \frac{1}{\Delta}\sqrt{\frac{\pi}{2\phi}},$$

free of $\bar{x}$, so $k(x, x') \to \frac{\sigma^2}{\Delta}\sqrt{\pi/(2\phi)}\; e^{-\phi(x-x')^2/2}$. This is the squared-exponential kernel exactly, at $\kappa = \phi/2$ and unit scale, provided

$$\phi = 2\kappa, \qquad \sigma^2 = \Delta\sqrt{2\phi/\pi} = 2\Delta\sqrt{\kappa/\pi}.$$

(d) Fix a region of interest $\mathcal{X}$, a grid $x_1, \dots, x_N$ on it, and minimize the Frobenius discrepancy between Gram matrices,

$$D(\phi, \tau, \sigma^2) = \big\| \sigma^2 BB^T - K_{\mathrm{SE}} \big\|_F^2, \qquad (K_{\mathrm{SE}})_{ij} = e^{-\kappa(x_i - x_j)^2}.$$

The limit in (c) supplies a closed-form initialization:
(i) Set $\phi = 2\kappa$, which matches the distance-dependent factor exactly and leaves only the amplitude $g(\bar{x}) = \sigma^2 \sum_h e^{-2\phi(\tau_h - \bar{x})^2}$ to be flattened to 1.

(ii) Place $\tau_1, \dots, \tau_k$ equally spaced with spacing Δ on $\mathcal{X}$ widened at each end by a few multiples of $(2\phi)^{-1/2}$, so the edge taper of g falls outside $\mathcal{X}$. Poisson summation leaves only the ripple

$$\frac{\max g}{\min g} - 1 \;\approx\; 4\exp\Big(-\frac{\pi^2}{2\phi\Delta^2}\Big),$$

so already at $\Delta = (2\phi)^{-1/2}$ the amplitude is flat to $4e^{-\pi^2} = 2 \times 10^{-4}$.
(iii) Profile out σ^2, which enters D quadratically with minimizer $\sigma^2 = \langle BB^T, K_{\mathrm{SE}}\rangle_F / \|BB^T\|_F^2$; the closed form $\Delta\sqrt{2\phi/\pi}$ is the starting value.
(iv) Polish (ϕ, τ) by L-BFGS on the profiled D, whose gradients are closed-form. The floor is $\sum_{j>k} \lambda_j^2$ for $\lambda_1 \geq \lambda_2 \geq \cdots$ the eigenvalues of K_{SE}, since no rank-k matrix does better (Eckart–Young); those eigenvalues decay geometrically, so the floor sits far below what the equally spaced grid of (ii) reaches at small k and the polish is not cosmetic.

Problem 21.9. Continuing the previous problem, suppose we instead let $\mu(x) = \beta_1 + \beta_2 x$ with Gaussian priors placed on β_1 and β_2.
(a) Does this induce a Gaussian process prior on the function $\mu(x)$?
(b) Since we have a Gaussian process prior, does that mean that we can capture non-linear functions with this prior?

Solution. (a) Yes. This is Exercise 21.8(a) with $k = 2$, $b_1(x) \equiv 1$, $b_2(x) = x$: for $\beta = (\beta_1, \beta_2)^T \sim \mathrm{N}(\mu_\beta, \Sigma_\beta)$ every vector $(\mu(x_1), \dots, \mu(x_n))^T = B\beta$ with $B_{i\cdot} = (1, x_i)$ is normal, so $\mu \sim \mathrm{GP}(m, k)$ with

$$\begin{aligned} m(x) &= \mu_1 + \mu_2 x, \\ k(x, x') &= \Sigma_{11} + \Sigma_{12}(x + x') + \Sigma_{22}\, xx'. \end{aligned}$$

For independent $\beta_j \sim \mathrm{N}(0, \sigma_j^2)$ this is the linear kernel $k(x, x') = \sigma_1^2 + \sigma_2^2 xx'$, with $m \equiv 0$.
(b) No – every draw is a straight line, with probability one. Being a Gaussian process constrains the finite-dimensional distributions to be normal; it says nothing about how rich the support is, and here the prior is carried by the two-dimensional space $\mathrm{span}\{1, x\}$. The degeneracy is visible in the kernel: $k(x, x') = b(x)^T \Sigma_\beta b(x')$ has rank at most 2, so for any $n \geq 3$ points the Gram matrix is singular and, writing $x_3 = (1 - t)x_1 + t x_2$,

$$\mu(x_3) = (1 - t)\,\mu(x_1) + t\,\mu(x_2) \qquad \text{with probability 1.}$$

A prior assigning mass 0 to the non-linear functions yields a posterior assigning them mass 0 as well, for any likelihood. Curvature therefore has to be put in through the basis, and support dense in $C(\mathcal{X})$ requires a kernel of infinite rank – for instance the squared-exponential $\exp(-\kappa(x - x')^2)$, which by Exercise 21.8(c) is the $k \to \infty$, dense-knot limit of the basis construction.

22 Finite Mixture Models

Contents

22.1 Exercises 22.1–22.7

Problem 22.1. Posterior summaries: Suppose you have a mixture model with three components. For each data point you want to identify which of the three components it comes from. It would be best to use the full posterior distribution but you need a point estimate. Which of the following would you prefer: the posterior mean, the posterior median, or the posterior mode? Assume each of these is done pointwise (that is, you are getting the marginal mean, median, or mode of the latent component for each data point, not the joint mean, median, or mode for all the data points at once).

Solution. The posterior mode. The estimand is the indicator $z_i \in \{1, 2, 3\}$, an unordered label, and the mode is the only one of the three summaries that is invariant to relabelling the components – which is exactly the invariance the likelihood itself has (Section 22.3, label ambiguity).
Concretely, let $p_i = (p_{i1}, p_{i2}, p_{i3})$ be the marginal posterior of z_i. If $p_i = (0.5, 0, 0.5)$ then

$$\begin{aligned} \text{mean}(z_i) &= 1(0.5) + 2(0) + 3(0.5) = 2, \\ \text{median}(z_i) &= 2, \end{aligned}$$

and both name the one component the data have ruled out, while the mode returns 1 or 3, each a possible answer. In general the mean is not even integer-valued, so it is not a point of the parameter space at all.
Formally, the mode is the Bayes estimate under the natural loss for a classification problem, the 0–1 loss $L(a, z_i) = 1_{a \neq z_i}$, whose posterior expectation $1 - p_{ia}$ is minimized at $a = \arg\max_h p_{ih}$ (Section 22.5, final paragraph). Posterior mean and median minimize squared-error and absolute-error loss, and neither loss means anything applied to component names.

Problem 22.2. Mixtures with unspecified numbers of components: Simulate 500 data points from an equally weighted mixture of three normal distributions centered at -2, 0, and 2, each with scale parameter 1.
(a) Fit a Bayesian model of a mixture of two normal distributions to these data.
(b) Fit a Bayesian model of a mixture of three normal distributions to these data.
(c) Fit a Bayesian model of a mixture of four normal distributions to these data.
(d) Fit a Bayesian model with unspecified number of mixture components, with the total number of components being allowed to be anywhere between 1 and 6.

Solution. All four fits give the same density estimate, and no fit recovers the number three: with centers two units apart and unit scales the three components overlap so heavily that the mixture is essentially unimodal, and the likelihood carries almost no information about H.
Simulation: $n = 500$ draws with sample mean -0.06 and sample standard deviation 1.92. The model in every part is the location-scale mixture (22.11)–(22.12), fit to the standardized data,

$$\begin{aligned} y_i \mid z_i &\sim \text{N}(\mu_{z_i}, \tau_{z_i}^2), \quad \Pr(z_i = h) = \pi_h, \\ (\pi_1, \ldots, \pi_H) &\sim \text{Dirichlet}(a, \ldots, a), \\ \mu_h \mid \tau_h^2 &\sim \text{N}(0, \kappa\tau_h^2), \quad \tau_h^2 \sim \text{Inv-gamma}(3, 1), \end{aligned}$$

with $\kappa = 1$, the weakly informative default recommended in Section 22.3 for standardized data. The three-step Gibbs sampler of Section 22.3 (update z from its multinomial conditional, (μ_h, τ_h^2) from the normal/inverse-gamma conditional, π from the Dirichlet conditional) was run for 8000 iterations, the first half discarded; the density estimate is the posterior mean of $g(y) = \sum_h \pi_h \text{N}(y \mid \mu_h, \tau_h^2)$ on a grid. No monitoring of (μ_h, τ_h) was attempted, for the label-switching reason given in Section 22.3.
(a)–(c) With $a = 1$ and $H = 2, 3, 4$ the fits are statistically indistinguishable. The L_1 error below is $\int |\hat{g} - g_0|$, the integrated absolute error of the posterior mean density against the true mixture:

H	WAIC	p_{WAIC}	L_1 error
2	1399.1	4.1	0.055
3	1399.4	4.2	0.052
4	1399.1	4.1	0.053

The WAIC differences are $+0.3$ and -0.1 with paired standard errors 0.4 and 0.3, against a standard error of 22.8 for WAIC itself: nothing separates the three models. Note $p_{\text{WAIC}} \approx 4$ in all three cases – the effective number of parameters does not grow with H, because the extra components are not being used.

(d) Take $H = 6$ as the upper bound and $a = n_0/H$ with $n_0 = 1$, as prescribed in Section 22.4, and read the posterior of $H_n = \sum_{h=1}^{H} 1_{n_h>0}$ off the Gibbs output (20,000 iterations, half discarded):

k	$\Pr(H_n = k \mid y)$	$\Pr(\#\{h : \pi_h > 0.05\} = k \mid y)$
1	0.000	0.000
2	0.026	0.291
3	0.170	0.472
4	0.367	0.204
5	0.332	0.033
6	0.105	0.001

The raw count of occupied components has posterior mode 4 and puts 44% on $\{5, 6\}$, because a handful of points is always parked in a negligible cluster – the overestimation Section 22.4 warns about. Discarding components carrying less than 5% of the mass, the posterior mode is 3 with probability 0.47 and $\{2, 3\}$ carries 0.76. The density estimate averaged over this posterior has L_1 error 0.052, the same as the fixed-H fits.

The choice $a = 1/H$ is what makes this work: repeating (d) with $a = 1$ gives $\Pr(H_n = 6 \mid y) = 0.93$ and $\Pr(H_n \leq 4 \mid y) = 0.002$, since with fixed prior mass a per component the prior sample size Ha grows with the upper bound and every component gets filled, exactly as predicted in Section 22.4. The density estimate is unaffected (L_1 error 0.050).

Problem 22.3. Fitting long-tailed data with a normal mixture: Repeat the above problem, but simulating the data from a mixture of three t_4 distributions. Again fit the data using mixtures of normals. (That is: simulate 500 points from an equally weighted mixture of t_4 densities centered at -2, 0, and 2 with unit scale, then fit Bayesian normal mixtures with two, three, and four components, and a model with the number of components unspecified between 1 and 6.)

Solution. Now the extra components pay, and they are spent on scale rather than on location: the fit pairs a narrow kernel with a broad one at nearly the same place to imitate a t_4 tail, so the number of components has even less to do with the number three than in 22.2.

The data: $n = 500$ draws, sample mean -0.17, sample standard deviation 2.21, range $[-9.95, 7.68]$ (a t_4 has standard deviation $\sqrt{4/2} = 1.41$ against the 1 of 22.2, and the observed range is nearly twice as wide). The model, priors ($\mu_0 = 0$, $\kappa = 1$, $\tau_h^2 \sim$ Inv-gamma$(3, 1)$ on standardized data) and Gibbs sampler are exactly those of 22.2.

(a)–(c) With $a = 1$:

H	WAIC	p_{WAIC}	L_1 error
2	1423.0	5.2	0.145
3	1420.6	7.8	0.113
4	1419.3	7.6	0.110

Unlike 22.2 the fit genuinely improves: the integrated absolute error against the true t_4 mixture falls by a quarter from $H = 2$ to $H = 3$, and p_{WAIC} rises from 5.2 to about 7.7 – the third component is being used. WAIC agrees in direction but not decisively, $\text{WAIC}(3) - \text{WAIC}(2) = -2.0 \pm 2.6$ and $\text{WAIC}(4) - \text{WAIC}(2) = -3.1 \pm 3.2$ (paired standard errors). There is no further gain from $H = 4$.

What the components do is visible in their posterior summaries. At $H = 4$, ordering within each draw by scale, the widest component has posterior mean standard deviation 2.88 (original units) and weight 0.24, while the narrowest has 1.17; at $H = 3$ the corresponding numbers are 2.74 with weight 0.36, and 1.20. A normal mixture represents a long tail by superposing kernels of very different widths at

overlapping locations, which is precisely the locally adaptive bandwidth behavior described in Section 22.3; it is not clustering.
(d) With $H = 6$ and $a = 1/6$: WAIC 1421.2, $p_{\text{WAIC}} = 7.6$, L_1 error 0.119, and

k	$\Pr(H_n = k \mid y)$	$\Pr(\#\{h : \pi_h > 0.05\} = k \mid y)$
1	0.004	0.106
2	0.034	0.260
3	0.149	0.372
4	0.327	0.203
5	0.344	0.049
6	0.142	0.010

The posterior for the non-negligible count is more diffuse than in 22.2 – mode 3, but with 0.26 on $\{4, 5, 6\}$ and 0.11 on a single component – because with heavy-tailed kernels there is no correct finite answer: the components are approximating a shape, so the mass at 2 and 3 is carrying three clusters plus a wide tail component, not one component per cluster. This is what the book flags for the acidity data in Section 22.4: when the kernel does not match the shape of the clusters, the number of components cannot be read as the number of clusters.

Problem 22.4. Specify a finite mixture of Gaussians model for the density of the galaxy data. Assume a symmetric Dirichlet prior with parameter α for the weights on the different components, and normal inverse-gamma priors for the location and variance of each Gaussian kernel. Plot the Bayesian density estimate under squared error loss along with a simple histogram of the data, and comment on how the density estimate changes as (a) the parameter α decreases to zero, (b) the number of mixture components k increases, or (c) the variance of the normal inverse-gamma prior increases.
(The galaxy data are the $n = 82$ measured recession velocities, in km/s, of galaxies in the Corona Borealis region, the univariate dataset displayed in the top row of Figure 22.4 and summarized in Table 22.2. The values range from 9172 to 34,279 with mean 20,831 and standard deviation 4568, and are visibly grouped: seven velocities below 10,500, a large central group, and a handful above 26,000.)

Solution. The model, for data y_i standardized to mean 0 and variance 1,

$$\begin{aligned} y_i \mid z_i &\sim \mathrm{N}(\mu_{z_i}, \tau_{z_i}^2), \quad \Pr(z_i = h) = \pi_h,\ h = 1, \dots, k, \\ (\pi_1, \dots, \pi_k) &\sim \text{Dirichlet}(\alpha, \dots, \alpha), \\ \mu_h \mid \tau_h^2 &\sim \mathrm{N}(\mu_0, \kappa\tau_h^2), \quad \tau_h^2 \sim \text{Inv-gamma}(a_\tau, b_\tau), \end{aligned}$$

this being (22.10)–(22.12) with $\mu_0 = 0$, $a_\tau = 3$, $b_\tau = 1$; the Bayes estimate of the density under squared error loss is the posterior mean

$$\hat{g}(y) = \mathrm{E}\Big[\sum_{h=1}^{k} \pi_h \mathrm{N}(y \mid \mu_h, \tau_h^2) \;\Big|\; y \Big],$$

computed as the average of $g^{(s)}$ over the three-step Gibbs sampler of Section 22.3, 20,000 iterations with the first half discarded. Every entry in the tables below comes from one such run, so read them to about ± 0.1 in $\mathrm{E}(H_n \mid y)$ and ± 0.3 in WAIC; the integer mode can flip between adjacent values.
Baseline: $k = 5$, $\alpha = 1/k = 0.2$, $\kappa = 10$. Plotted with pointwise 95% intervals over a histogram of the 82 velocities, $\hat{g}$ has three well separated modes, at about 9900, 21,300 and 31,400 km/s, the central one carrying almost all the mass; this matches the top row of Figure 22.4 and the component locations -2.35, 0.10, 1.89 (standardized) of Table 22.2.
(a) Decreasing α to zero empties components and eventually oversmooths. With $k = 10$:

α	$\mathrm{E}(H_n \mid y)$	mode	$\mathrm{E}(\#\{\pi_h > 0.05\} \mid y)$	WAIC	modes of $\hat{g}$
1.000	8.44	9	5.31	192.5	2
0.500	7.00	7	4.03	191.3	2
0.100	4.33	4	2.69	189.9	3
0.010	3.16	3	2.35	189.2	3
0.001	3.02	3	2.30	189.2	3

At $\alpha = 1$ nearly every one of the ten components is occupied and $\hat{g}$ is a rough, multi-bumped curve fitted to 82 points – the behavior Section 22.4 warns of, $H_n \to n$ as k grows with α fixed. As $\alpha \downarrow 0$ the number of occupied components drops to three and stabilizes; the estimate is essentially unchanged below $\alpha \approx 0.01$ and WAIC is flat there. Taken to the limit $\alpha \to 0$ the Dirichlet degenerates on the vertices of the simplex and the model collapses to a single Gaussian, so the density estimate would eventually become the unimodal normal fit; at $n = 82$ with three clearly separated groups that limit is not reached at any α the sampler can distinguish from 0.001.

(b) Increasing k has almost no effect once k exceeds the number of groups, provided $\alpha = 1/k$:

k	$\mathrm{E}(H_n \mid y)$	mode	$\mathrm{E}(\#\{\pi_h > 0.05\} \mid y)$	WAIC	modes of $\hat{g}$
2	2.00	2	2.00	201.3	1
3	3.00	3	2.39	189.2	3
5	3.69	4	2.59	189.4	3
10	4.33	4	2.69	189.9	3
20	4.64	4	2.73	190.4	3
40	4.91	5	2.78	190.4	3

Going from $k = 2$ to $k = 3$ costs 12 units of WAIC; after that the curve is stable and the added components stay empty or negligible. The raw occupied count creeps up ($3.0 \to 4.9$) because with more available components a few points always land in a stray cluster, so H_n is mildly sensitive to the upper bound, while the count of components with weight above 0.05 is not ($2.4 \to 2.8$) and $\hat{g}$ is not at all. This is why Section 22.4 recommends $\alpha = n_0/k$ with $n_0 = 1$ and reporting the density rather than H_n.

(c) Increasing the prior variance κ of the component means first improves the fit sharply, then flattens, and slowly pushes toward fewer clusters. With $k = 5$, $\alpha = 0.2$:

κ	$\mathrm{E}(H_n \mid y)$	mode	$\mathrm{E}(\#\{\pi_h > 0.05\} \mid y)$	WAIC	modes of $\hat{g}$
1	3.72	4	3.02	199.6	1
10	3.69	4	2.59	189.4	3
100	3.34	3	2.34	187.4	3
1000	3.13	3	2.30	187.3	3

The conditionally conjugate prior ties the location to the scale, $\mathrm{var}(\mu_h \mid \tau_h^2) = \kappa\tau_h^2$, equivalently $1/\kappa$ prior observations at $\mu_0 = 0$; so at $\kappa = 1$ a tight component far from the center is heavily penalized – the seven low velocities sit 2.4 standard deviations out, and placing a narrow kernel there inflates its own $\hat{b}_{\tau h}$ by $\frac{1}{2}\frac{n_h}{1+\kappa n_h}(\bar{y}_h - \mu_0)^2$. The result (across five seeds) is a unimodal, badly oversmoothed $\hat{g}$ and WAIC worse by 10. Raising κ to 10 frees the outlying components and recovers the three groups. Beyond that the effect reverses direction: at $\kappa = 100$ and 1000 the prior for each component mean is so diffuse that the marginal likelihood penalty for occupying a new component grows, $\mathrm{E}(H_n \mid y)$ falls from 3.7 to 3.1 and the non-negligible count to 2.3, i.e. the posterior increasingly favors too few clusters – the sensitivity to "the size of the enormous variance chosen" against which Section 22.3 argues, and the reason it recommends standardizing and then keeping P_0 on the support of the data.

Problem 22.5. Consider the football point spread data of Section 1.6. Instead of assuming that the differences between score differential and point spread follow a normal distribution, fit a finite mixture of Gaussians to these data using a symmetric Dirichlet prior with hyperparameter $1/k$, where k is the number of mixture components. Run a Gibbs sampler to analyze the data, and compare the fitted distribution with that for the normal model. Comment on whether the results suggest the Gaussian density provides a good approximation.

(The data of Section 1.6 are the $n = 672$ professional American football games of the 1981, 1983 and

1984 seasons; for each game, d_i is the score differential, favorite minus underdog, minus the pre-game point spread. Section 1.6 models these as normal with mean 0 and standard deviation 14.)

Solution. Yes, excellent: the mixture buys nothing. Fit to the $n = 672$ differences (sample mean 0.070, standard deviation 13.860, skewness 0.002, excess kurtosis 0.010), the mixture's WAIC exceeds the normal model's by 0.8 ± 3.2 and the fitted density is unimodal and within 7.5% of the normal fit everywhere.

Model, on the standardized differences, with $\mu_0 = 0$, $\kappa = 10$, $\tau_h^2 \sim \text{Inv-gamma}(3, 1)$ as in 22.4:

$$\begin{aligned} d_i \mid z_i &\sim \text{N}(\mu_{z_i}, \tau_{z_i}^2), \quad \Pr(z_i = h) = \pi_h, \\ (\pi_1, \ldots, \pi_k) &\sim \text{Dirichlet}(1/k, \ldots, 1/k), \\ \mu_h \mid \tau_h^2 &\sim \text{N}(\mu_0, \kappa\tau_h^2), \end{aligned}$$

with the three-step Gibbs sampler of Section 22.3, 20,000 iterations, first half discarded. The comparison model is the normal of Section 1.6 with the noninformative prior $p(\mu, \sigma^2) \propto \sigma^{-2}$ of Section 3.2, for which WAIC was computed from 4000 draws of its exact posterior $\sigma^2 \mid d \sim \text{Inv-}\chi^2(n-1, s^2)$, $\mu \mid \sigma^2, d \sim \text{N}(\bar{d}, \sigma^2/n)$.

k	WAIC	p_{WAIC}	$\text{E}(H_n \mid y)$	$\text{E}(\#\{\pi_h > 0.05\} \mid y)$	modes of $\hat{g}$
normal	5443.4	2.00	1	1	1
3	5444.2	4.58	2.67	1.96	1
5	5444.4	5.45	3.80	2.41	1
10	5444.2	5.55	5.03	2.74	1
20	5444.7	5.78	5.60	2.75	1

Every mixture fit costs about three extra effective parameters and returns nothing: the paired WAIC difference at $k = 10$ is WAIC(mixture) − WAIC(normal) = +0.77 with standard error 3.20. The occupied count H_n drifts upward with the bound k while the number of components carrying more than 5% of the mass sits near 2.7 – extra components are splitting the single normal shape, not finding structure, and the fitted density $\hat{g}$ has a single mode for every k.

The two fitted distributions agree numerically. Against $\text{N}(0.070, 13.860^2)$, the posterior mean mixture density has integrated absolute difference $\int |\hat{g} - \hat{\phi}| = 0.053$ and maximum absolute discrepancy 7.5% of the peak height; the normal density lies inside the pointwise 95% posterior band of $\hat{g}$ at every grid point. Quantiles:

q	mixture	$\text{N}(0.07, 13.86^2)$
0.025	-27.34	-27.10
0.100	-17.45	-17.69
0.500	-0.46	0.07
0.900	18.15	17.83
0.975	27.48	27.24

The largest disagreement anywhere in the tails is a quarter of a point on a scale whose standard deviation is 13.9. The Gaussian approximation of Section 1.6 is therefore very good, as the sample skewness 0.002 and excess kurtosis 0.010 already suggested. The one respect in which the data are genuinely non-normal is invisible to a continuous mixture: the differences are discrete, taking 130 distinct values over 672 games with 18 tied at the modal value −6, and any continuous density, mixture or normal, smooths that away.

Problem 22.6. Consider the kidney cancer example of Section 2.7. Assume $y_j \sim \text{Poisson}(10 n_j \theta_j)$ with

$$\theta_j \sim \sum_{h=1}^{k} \pi_h \delta_{\theta_h^*}, \quad \theta_h^* \sim \text{Gamma}(\alpha, \beta), \quad \alpha = 20,\ \beta = 430{,}000,\ k = 25,$$

and $\pi = (\pi_1, \ldots, \pi_k) \sim \text{Dirichlet}(1/k, \ldots, 1/k)$. Comment on how this model differs from $\theta_j \sim \text{Gamma}(\alpha, \beta)$. Fit both these models and compare the results.

(In Section 2.7, y_j is the number of kidney cancer deaths in county j of the United States over 1980–1989, n_j is the county population, and θ_j is the underlying death rate in deaths per person per year; the Gamma(20, 430,000) prior has mean 4.65×10^{-5} and standard deviation 1.04×10^{-5}, and the conjugate posterior is $\theta_j \mid y_j \sim \text{Gamma}(20 + y_j,\ 430{,}000 + 10n_j)$.)

Solution. The two models give a county the same marginal prior and differ entirely in the dependence they impose across counties: the mixture ties the θ_j together, so that two counties share an identical rate with prior probability $(k+1)/(2k) = 0.52$, whereas under $\theta_j \sim \text{Gamma}(\alpha, \beta)$ the rates are independent and no strength is borrowed between counties at all.
Marginally the priors coincide. Averaging over (π, θ^*),

$$\Pr(\theta_j \in A) = \mathrm{E}\Big[\sum_{h=1}^{k} \pi_h 1_{\theta_h^* \in A}\Big] = \sum_{h=1}^{k} \frac{1}{k} \,\text{Gamma}(A \mid \alpha, \beta) = \text{Gamma}(A \mid \alpha, \beta),$$

since $\mathrm{E}(\pi_h) = 1/k$ and $\pi \perp \theta^*$. So any inference for a single isolated county is unchanged. The difference is in the joint distribution: for $i \neq j$,

$$\Pr(\theta_i = \theta_j) = \mathrm{E}\Big[\sum_{h=1}^{k} \pi_h^2\Big] = \frac{a+1}{ka+1}\Big|_{a=1/k} = \frac{k+1}{2k} = 0.52,$$

using $\mathrm{E}(\pi_h^2) = \mathrm{var}(\pi_h) + (1/k)^2$ for the symmetric Dirichlet, and the fact that a continuous Gamma(α, β) makes the atoms $\theta_1^*, \ldots, \theta_k^*$ almost surely distinct. The mixture model is thus a clustering prior: the 3071 county rates are forced onto at most $k = 25$ distinct values, and a county's estimate is informed by every other county allocated to the same atom, so a small county is no longer shrunk merely toward the fixed prior mean 4.65×10^{-5} but toward the rate of the group it is judged to belong to. Equivalently, the mixture treats the cross-county distribution of rates as an unknown to be estimated from the data, while the gamma model fixes it at Gamma(20, 430,000).
Computation is conjugate throughout. The gamma model needs none. For the mixture, the Gibbs sampler of Section 22.3 becomes

$$\begin{aligned}
\Pr(z_j = h \mid -) &\propto \pi_h \,\text{Poisson}(y_j \mid 10 n_j \theta_h^*), \\
\theta_h^* \mid - &\sim \text{Gamma}\Big(\alpha + \sum_{j:z_j=h} y_j,\ \beta + \sum_{j:z_j=h} 10 n_j\Big), \\
\pi \mid - &\sim \text{Dirichlet}(1/k + n_1, \ldots, 1/k + n_k),
\end{aligned}$$

run here for 6000 iterations with the first half discarded, θ_j reported as the posterior mean of $\theta_{z_j}^*$.
The county file of Section 2.7 is not printed in the book, so the fits below use 3071 synthetic counties: populations n_j lognormal with median 25,000 truncated to $[500,\ 8 \times 10^6]$, and either (I) true rates drawn from Gamma(20, 430,000) or (II) true rates equal to 4.0×10^{-5} for 85% of counties and 1.0×10^{-4} for the remaining 15%. Root mean squared error of the estimated rates against the truth, in units of 10^{-5}:

truth	raw $y_j/(10n_j)$	gamma	mixture	occupied atoms
(I)	2.621	0.797	0.798	12.0
(II)	2.709	1.509	1.136	5.1

Under (I), where the gamma prior is exactly right, the mixture matches it (0.798 against 0.797): it pays nothing for its extra flexibility, spending about 12 of its 25 atoms to approximate the continuous gamma. Under (II), where the rate distribution has structure the gamma cannot represent, the mixture finds it – overall error drops by a quarter – and the gain is concentrated where the data can resolve the grouping:

county size	J	gamma	mixture
$n_j < 2000$	204	2.180	2.080
$2 \times 10^4 < n_j < 3 \times 10^4$	322	1.448	0.844
$n_j > 10^6$	36	0.184	0.026

The large counties show the mechanism: each has enough deaths to say which group it belongs to, the atoms lock onto the two true rates, and the residual error nearly vanishes – a sevenfold improvement no independent-shrinkage model can produce. The smallest counties gain almost nothing, since they carry too little information to be allocated reliably and are shrunk to roughly the overall mean either way.

The costs are the usual ones: the estimates are discrete, so many counties receive numerically identical rates, which is an artifact of the model rather than a finding about the counties; and by Section 22.3 the individual θ_h^* are not interpretable without relabelling, only functionals such as $\theta_j = \theta_{z_j}^*$ and the implied distribution of rates.

Problem 22.7. What problems (if any) result from putting a noninformative prior on component-specific parameters in a finite mixture model?

Solution. The posterior is improper, and nothing in the computation tells you so.

Take the location-scale mixture (22.11) with $p(\mu_h, \tau_h^2) \propto \tau_h^{-2}$ for each h. Condition on an allocation z that leaves component h empty, $n_h = 0$ – an event of positive probability under any π with $\pi_h < 1$. That component's parameters then appear nowhere in the likelihood, so

$$\int p(y \mid z, \omega)\, p(\omega)\, d\omega \;\propto\; \int_0^\infty \!\! \int_{-\infty}^\infty \frac{d\mu_h\, d\tau_h^2}{\tau_h^2} \;=\; \infty,$$

and since $p(\omega \mid y) = \sum_z p(\omega, z \mid y)$ includes that term, the posterior has no normalizing constant. This is exactly the degeneracy identified in Section 22.2 under "Possible difficulties at a degenerate point": with $\sum_{ij} z_{ij} = 0$ there are no delayed reactions, τ has an improper prior and no data, "strictly speaking, this means that our posterior distribution is improper."

The trap is that the Gibbs sampler of Section 22.3 never complains. Every full conditional it uses is conditional on z with $n_h \geq 1$ for the components it happens to be updating, hence proper; the chain mixes, the density estimate looks sensible, and the impropriety shows up only if the sampler wanders into the degenerate region – which in the reaction-time example it did not, but only because "this degenerate point has extremely low posterior probability." That is a property of the data, not a guarantee.

Two further problems survive even if the prior is made proper but diffuse.

(i) The likelihood of a normal mixture is unbounded – the "uninteresting" modes of Section 22.1, each a component consisting of one observation with no variance. Fix $\mu_2 = y_1$ and let $\tau_2^2 \to 0$; then $\mathrm{N}(y_1 \mid \mu_2, \tau_2^2) \to \infty$ while every other factor stays bounded below, so $p(y \mid \omega) \to \infty$. A prior proportional to τ^{-2} does not damp this, so the posterior carries an infinite spike at each of the n data points, the EM/ECM maximization of Section 22.2 runs off to a degenerate mode, and MCMC can be absorbed into one. A proper inverse-gamma prior on τ_h^2 – Inv-gamma(a_τ, b_τ) with $b_\tau > 0$ – kills the spike, since $\tau^{-2a_\tau - 2} e^{-b_\tau/\tau^2} \to 0$ as $\tau^2 \to 0$.

(ii) Inference about the number of components becomes a function of the arbitrary prior variance rather than of the data. A vague-but-proper P_0 with variance V multiplies the marginal likelihood of each occupied component by a factor of order $V^{-1/2}$, so as V grows the posterior progressively empties components – the Bartlett–Lindley effect. Section 22.3 states the practical version ("it is important to avoid choosing a P_0 that is improper, or even diffuse but proper, as the results may be sensitive to the size of the enormous variance chosen"), and 22.4(c) above exhibits it: on the galaxy data with $k = 5$, raising κ from 10 to 1000 moved $\mathrm{E}(H_n \mid y)$ from 3.7 to 3.1 and the number of non-negligible components from 2.6 to 2.3, with no change in the data.

The remedy Section 22.3 prescribes is the one used throughout this chapter: standardize the data and take a proper, weakly informative P_0 placing the component parameters on the support of the data – $\mu_0 = 0$, κ of order 1, $a_\tau = 2$, $b_\tau = 4$ as a default.

22.2 Exercises 22.8–22.8

Problem 22.8. Draw 1000 samples from $\pi = (\pi_1, \ldots, \pi_k) \sim \text{Dirichlet}(1/k, \ldots, 1/k)$ for $k = 5, 10, 25, 50, 100, 1000$. For each sample, reorder the elements of π to be decreasing from largest to smallest and estimate posterior summaries of these order statistics. Plot the results and describe the behavior as k increases. Repeat this exercise for $\pi \sim \text{Dirichlet}(1, \ldots, 1)$.

Solution. Under $\alpha_j = 1/k$ the ranked weights settle down: they converge to a nondegenerate limit law, so a draw is effectively supported on three or four components however large k is; under $\alpha_j = 1$ every ranked weight shrinks like $1/k$ and the mass spreads evenly over all k components. No data enter, so these are Monte Carlo summaries of the prior itself.

Each π is generated by the standard normalization, $g_j \sim \text{Gamma}(\alpha_j, 1)$ independently and $\pi_j = g_j / \sum_h g_h$ (BDA3 Appendix A), then sorted to $\pi_{(1)} \geq \cdots \geq \pi_{(k)}$. With $S = 1000$ draws per k:

Symmetric sparse prior, $\alpha_j = 1/k$ (means, and for $\pi_{(1)}$ the median and central 90% interval):

k	$E\pi_{(1)}$	med $\pi_{(1)}$	90% int.	$E\pi_{(2)}$	$E\pi_{(3)}$	$E\sum_{j\leq 5}\pi_{(j)}$	$EN_{0.95}$	$E\,\text{ESS}$
5	0.707	0.708	[0.418, 0.984]	0.208	0.066	1.000	2.53	1.86
10	0.665	0.656	[0.372, 0.972]	0.211	0.079	0.996	2.99	2.09
25	0.635	0.619	[0.345, 0.955]	0.214	0.087	0.989	3.35	2.26
50	0.640	0.626	[0.348, 0.970]	0.203	0.085	0.984	3.49	2.29
100	0.621	0.604	[0.345, 0.947]	0.215	0.090	0.984	3.56	2.35
1000	0.624	0.604	[0.337, 0.956]	0.208	0.087	0.981	3.65	2.38

Here $N_{0.95} = \min\{m : \sum_{j\leq m}\pi_{(j)} \geq 0.95\}$ and $\text{ESS} = 1/\sum_j \pi_j^2$. The Monte Carlo standard error of $E\pi_{(1)}$ at $S = 1000$ is 0.006, and past $k = 25$ every column moves by only about one such unit, so the sequence has essentially converged. The reason is that the total mass $\sum_j \alpha_j = k \cdot (1/k) = 1$ is held fixed, and the ranked weights of $\text{Dirichlet}(1/k, \ldots, 1/k)$ converge as $k \to \infty$ to the ranked stick-breaking weights of a Dirichlet process with concentration $\alpha = 1$ (BDA3 Sections 23.1–23.2), i.e. to the Poisson–Dirichlet law PD(1). Simulating 200,000 draws at $k = 2000$ pins the limit means down to

$$E\pi_{(1)} = 0.6249, \quad E\pi_{(2)} = 0.2093,$$

$$E\pi_{(3)} = 0.0882, \quad E\pi_{(4)} = 0.0402,$$

(MC standard errors $\leq 5 \times 10^{-4}$), the first being the Golomb–Dickman constant 0.62433. Increasing k past a few dozen therefore buys no extra occupied components: it only subdivides the negligible tail.

Uniform prior, $\alpha_j = 1$:

k	$E\pi_{(1)}$	med $\pi_{(1)}$	90% int.	$E\pi_{(2)}$	$E\pi_{(3)}$	$E\sum_{j\leq 10}\pi_{(j)}$	$EN_{0.95}$	$E\,\text{ESS}$
5	0.455	0.438	[0.299, 0.673]	0.261	0.156	—	4.17	3.18
10	0.292	0.278	[0.191, 0.443]	0.193	0.143	1.000	7.75	5.76
25	0.153	0.144	[0.103, 0.234]	0.113	0.093	0.756	18.17	13.32
50	0.0908	0.0865	[0.0623, 0.1340]	0.0707	0.0610	0.517	35.71	25.71
100	0.0516	0.0496	[0.0370, 0.0716]	0.0417	0.0368	0.325	70.82	51.07
1000	0.0075	0.0072	[0.0059, 0.0100]	0.0065	0.0060	0.056	701.65	500.87

Here the means are exact in closed form, by uniform spacings:

$$E[\pi_{(j)}] \;=\; \frac{1}{k}\sum_{i=j}^{k}\frac{1}{i}, \qquad \text{so} \qquad E[\pi_{(1)}] = \frac{H_k}{k} \sim \frac{\log k}{k},$$

whose exact values $0.4567, 0.2929, 0.1526, 0.0900, 0.0519, 0.0075$ reproduce column two of the table. Everything decays at rate $1/k$, the largest weight exceeding the flat value $1/k$ only by the factor H_k (2.9 at $k = 10$, 7.5 at $k = 1000$); $E\,\text{ESS} \approx (k+1)/2$, since $E\sum_j \pi_j^2 = 2/(k+1)$; and $N_{0.95} \approx 0.70k$. So $\text{Dirichlet}(1, \ldots, 1)$ keeps all k components live, whereas $\text{Dirichlet}(1/k, \ldots, 1/k)$ is the sparse prior that makes an overfitted finite mixture behave like a Dirichlet process, the prior itself emptying the surplus components.

Plotting $E[\pi_{(j)}]$ against rank j on a log scale for $j \leq 20$ shows this at a glance: for $\alpha_j = 1/k$ the six curves lie on top of one another, while for $\alpha_j = 1$ they are the same curve displaced downward by $1/k$.

23 Dirichlet Process Models

Contents

23.1 Exercises 23.1–23.2

Problem 23.1. The following exercise is useful to gain familiarity with posterior computation and inferences for the Dirichlet process mixture of Gaussian models:
(a) Simulate data from the following mixture of normals:

$$p(y_i) \sim 0.1\,\mathrm{N}(y|-1,0.2) + 0.5\,\mathrm{N}(y|0,1) + 0.4\,\mathrm{N}(y|1,0.4), \quad i = 1,\ldots,100.$$

(b) Use the `density()` function in R to obtain a non-Bayesian estimate of the density and plot this estimate versus the true density.
(c) Apply the finite mixture model Gibbs sampler described in Chapter 22 for $k = 20$, $a = \alpha/k$, $\alpha = 1$, $\mu_0 = 0$, and $\kappa = a_\tau = b_\tau = \alpha = 1$.
(d) Run the blocked Gibbs sampler for $N = 20$ and the same hyperparameter specification.
(e) Compare the resulting density estimates.
For sufficiently many MCMC iterations and sufficiently large truncation levels k, N, the density estimates obtained via the finite Dirichlet approximation and truncated stick-breaking approximations to the DPM of Gaussians should be similar.

Solution. The two approximations agree to graphical accuracy: the posterior mean densities of (c) and (d) differ by $\sup_y |\hat{f}_k - \hat{f}_N| = 0.002$, half a percent of the modal height, against a common L_1 error of about 0.20 from the truth.
Throughout, the model is the DPM (23.5) with a location-scale normal kernel and the same conjugate normal-gamma base measure as in (23.9),

$$y_i \sim \mathrm{N}(\mu_i, \tau_i^{-1}), \quad (\mu_i,\tau_i) \sim P, \quad P \sim \mathrm{DP}(\alpha P_0),$$

$$P_0(\mu,\tau) = \mathrm{N}(\mu \,|\, \mu_0,\ \kappa\tau^{-1})\,\mathrm{Gamma}(\tau \,|\, a_\tau, b_\tau),$$

with $\mu_0 = 0$, $\kappa = a_\tau = b_\tau = 1$ and $\alpha = 1$. Every figure below comes from one simulated sample and 20,000 Gibbs iterations with the first 5,000 discarded; L_1 and squared errors against the truth f_0 are integrated on a grid of step 0.01 over $[-6, 6]$, and the batch-means (50 batches) Monte Carlo standard error of each reported L_1 error is 0.0007 or less, so the third decimal is not to be trusted.
(a) Draw $z_i \in \{1,2,3\}$ with probabilities $(0.1, 0.5, 0.4)$ and then $y_i \mid z_i \sim \mathrm{N}(\mu_{z_i}, \sigma^2_{z_i})$ with $\mu = (-1, 0, 1)$ and $\sigma^2 = (0.2, 1, 0.4)$ (the second argument of $\mathrm{N}(\cdot|\cdot,\cdot)$ is a variance, per Table A.1). The realized sample of $n = 100$ used below has $\bar{y} = 0.356$, $s = 0.966$, and ranges over $[-1.91,\ 2.64]$. The target f_0 is unimodal, with its only stationary maximum at $y = 0.807$ (height 0.385); the component at -1 carries too little mass to raise a second mode, and only thickens the left shoulder.
(b) R's `density()` uses a Gaussian kernel with Silverman's rule `bw.nrd0`,

$$h = 0.9\,\min\Big(s,\ \tfrac{\mathrm{IQR}}{1.349}\Big)\,n^{-1/5} = 0.346, \qquad \hat{f}_{\mathrm{ker}}(y) = \frac{1}{nh}\sum_{i=1}^{n} \phi\big((y - y_i)/h\big),$$

since $s = 0.966$ and $\mathrm{IQR}/1.349 = 1.431/1.349 = 1.061$. This gives

$$\int \big|\hat{f}_{\mathrm{ker}} - f_0\big|\,dy = 0.127, \qquad \int \big(\hat{f}_{\mathrm{ker}} - f_0\big)^2 dy = 0.0040,$$

a unimodal estimate that tracks f_0 closely, the bandwidth being far too large to resolve the -1 component (its own sd is $\sqrt{0.2} = 0.45$).
(c) The finite Dirichlet approximation of Chapter 22 replaces P by $\sum_{c=1}^{k} \pi_c \delta_{\omega_c^*}$ with $\pi \sim \mathrm{Dirichlet}(\alpha/k,\ldots,\alpha/k)$ and $\omega_c^* = (\mu_c^*, \tau_c^*) \overset{iid}{\sim} P_0$, which approaches $\mathrm{DP}(\alpha P_0)$ as $k \to \infty$. With $k = 20$ the Gibbs sampler cycles three conditionals. Writing $n_c = \#\{i : S_i = c\}$ and $\bar{y}_c$ for the cluster mean,

$$\begin{aligned} \Pr(S_i = c \mid -) &\propto \pi_c\,\mathrm{N}(y_i \mid \mu_c^*, \tau_c^{*-1}), \quad c = 1,\ldots,k, \\ \pi \mid - &\sim \mathrm{Dirichlet}(\alpha/k + n_1,\ldots,\alpha/k + n_k), \\ (\mu_c^*, \tau_c^*) \mid - &\sim \mathrm{N}\big(\mu_c^* \mid \hat{\mu}_c, \hat{\kappa}_c\tau_c^{*-1}\big)\,\mathrm{Gamma}\big(\tau_c^* \mid \hat{a}_{\tau c}, \hat{b}_{\tau c}\big), \end{aligned}$$

with the normal-gamma updates printed on page 556,

$$\hat\kappa_c = (\kappa^{-1} + n_c)^{-1}, \quad \hat\mu_c = \hat\kappa_c(\kappa^{-1}\mu_0 + n_c\bar y_c),$$
$$\hat a_{\tau c} = a_\tau + \tfrac{n_c}{2}, \qquad \hat b_{\tau c} = b_\tau + \tfrac{1}{2}\Big(\textstyle\sum_{i:S_i=c}(y_i - \bar y_c)^2 + \frac{n_c(\bar y_c - \mu_0)^2}{1+\kappa n_c}\Big);$$

(the book prints $(y_i - \bar y_c)$ unsquared here, plainly a typo). For $n_c = 0$ these reduce to P_0 itself, so empty clusters are drawn from the prior, as the sampler of Section 23.3 requires. Monitoring

$$f^{(t)}(y) = \sum_{c=1}^{k} \pi_c^{(t)}\, \mathrm{N}\big(y \mid \mu_c^{*(t)}, \tau_c^{*(t)-1}\big)$$

at each iteration — never the component parameters themselves, whose labels switch, as Section 23.3 warns — gives the posterior mean density $\hat f_k$ with

$$\int |\hat f_k - f_0| = 0.200, \qquad \int (\hat f_k - f_0)^2 = 0.0112.$$

The posterior number of occupied clusters has mean 3.6 (sd 1.5, 95% interval $[1, 7]$), far inside $k = 20$.
(d) The blocked Gibbs sampler truncates the stick-breaking representation (23.3) at N by setting $V_N = 1$, so $\pi_h = 0$ for $h > N$. Steps 1 and 3 are exactly as in (c) (the component update is unchanged, empty components again drawn from P_0); only the weight update differs:

$$V_c \mid - \sim \mathrm{Beta}\Big(1 + n_c,\ \alpha + \sum_{c'=c+1}^{N} n_{c'}\Big), \quad c = 1, \dots, N-1,$$

$$\pi_c = V_c \prod_{c'<c} (1 - V_{c'}), \qquad V_N \equiv 1.$$

The truncation is harmless here. In the untruncated process the sticks beyond N carry expected mass $\mathrm{E}\big[\prod_{h\le N}(1 - V_h)\big] = \{\alpha/(1+\alpha)\}^N = 2^{-20} = 9.5 \times 10^{-7}$, and the diagnostic Section 23.3 actually asks for, $S_{\max} = \max_i S_i$, sat at or below 12 at its 99th percentile. Hence

$$\int |\hat f_N - f_0| = 0.198, \qquad \int (\hat f_N - f_0)^2 = 0.0109,$$

with occupied-cluster count mean 4.0 (sd 1.6, 95% interval $[1, 7]$).
(e) Comparing the two:

estimate	L1 error	ISE	posterior mean no. clusters
kernel, $h = 0.346$	0.127	0.0040	—
finite Dirichlet, $k = 20$	0.200	0.0112	3.6
blocked Gibbs, $N = 20$	0.198	0.0109	4.0

The two Bayesian estimates are indistinguishable: $\sup_y |\hat f_k - \hat f_N| = 0.002$ against a modal height of 0.41, and the L_1 gap of 0.002 is at the edge of the Monte Carlo error. This is what the theory predicts, both k and N exceeding the occupied-cluster count by a factor of five.

Problem 23.2. To get an intuition for the impact of α and P_0, repeat the previous exercise but with:
(a) A higher value of α, such as $\alpha = 10$.
(b) A gamma hyperprior distribution on α, with $a_\alpha = b_\alpha = 0.1$.
(c) Much higher variance in the normal-gamma P_0.
(Here α is the Dirichlet process precision and $P_0(\mu, \tau) = \mathrm{N}(\mu \mid \mu_0, \kappa\tau^{-1})\,\mathrm{Gamma}(\tau \mid a_\tau, b_\tau)$ the base measure, as in Exercise 23.1.)

Solution. α and P_0 move the posterior over clusterings a great deal and the posterior over densities almost not at all: across the four specifications below the L_1 error of $\hat{f}$ stays within [0.198, 0.222] while the posterior mean number of occupied clusters runs from 1.3 to 19.6. Same data, grid, blocked sampler and 20,000/5,000 schedule as Exercise 23.1, so the Monte Carlo standard error on each L_1 figure is again about 0.0007.

(a) $\alpha = 10$. Only the weight prior changes: $\pi \sim \text{Dirichlet}(10/k, \ldots, 10/k)$ in the finite approximation, $V_c \sim \text{Beta}(1, 10)$ in the stick-breaking one. Since α multiplies P_0 against the $n-1$ existing atoms in the Polya urn (23.6), raising it by a factor of 10 raises the prior mean number of clusters from

$$\sum_{i=0}^{n-1} \frac{1}{1+i} = 5.19 \qquad \text{to} \qquad \sum_{i=0}^{n-1} \frac{10}{10+i} = 24.4.$$

The posterior follows: occupied clusters have mean 12.9 (sd 1.9) under the finite Dirichlet approximation and 14.6 (sd 1.9) under the blocked sampler, against 3.6 and 4.0 before. The density estimate moves by 0.01, to L_1 error 0.208 and 0.212, the extra components being redundant copies splitting mass that three or four already described.

At $\alpha = 10$, however, $N = 20$ is no longer a safe truncation. The untruncated tail mass is

$$\text{E}\Big[\sum_{h>N} \pi_h\Big] = \Big(\frac{\alpha}{1+\alpha}\Big)^N = \Big(\frac{10}{11}\Big)^{20} = 0.149,$$

against 9.5×10^{-7} at $\alpha = 1$, and $S_{\max}$ hits the bound 20, exactly the failure Section 23.3 tells us to watch for. Rerunning at the recommended default $N = 50$ (tail mass 0.0085) gives occupied clusters 19.6 (sd 3.5) and L_1 error 0.222: the $N = 20$ cluster count was a truncation artifact, a lower bound, while the density estimate moved by 0.01.

(b) $\alpha \sim \text{Gamma}(0.1, 0.1)$, a prior with mean $a_\alpha/b_\alpha = 1$ and standard deviation $\sqrt{a_\alpha}/b_\alpha = 3.16$. In the blocked sampler this is conditionally conjugate (Section 23.3), so one extra step suffices:

$$\alpha \mid - \;\sim\; \text{Gamma}\Big(a_\alpha + N - 1, \;\; b_\alpha - \sum_{h=1}^{N-1} \log(1 - V_h)\Big),$$

the $N-1$ free sticks being the only quantities that depend on α ($V_N \equiv 1$ carries no information about it, which is why the shape is $a_\alpha + N - 1$ and not $a_\alpha + N$). The posterior is

$$\text{E}[\alpha \mid y] = 0.39, \qquad 95\% \text{ interval } [0.06,\ 1.55],$$

so the data pull α below its prior mean of 1 and decisively below 10: this sample does not support many clusters. Occupied clusters drop to mean 2.0 (sd 1.5), and the density estimate is again essentially unchanged, L_1 error 0.204.

(c) A much more diffuse P_0. Take $\kappa = 100$ (prior variance of μ_c^* is κ/τ_c^*, so 100 times larger) with $a_\tau = b_\tau = 1$, and more aggressively $\kappa = 100$, $a_\tau = b_\tau = 0.1$ (a Gamma(0.1, 0.1) precision spreading τ over several orders of magnitude). The effect is the opposite of (a):

specification (blocked sampler, $N = 20$)	mean no. clusters	L1 error
$\alpha = 1$, $\kappa = 1$, $a_\tau = b_\tau = 1$ (baseline)	4.0	0.198
$\alpha = 10$, $\kappa = 1$, $a_\tau = b_\tau = 1$	14.6	0.212
$\alpha \sim \text{Gamma}(0.1, 0.1)$	2.0	0.204
$\alpha = 1$, $\kappa = 100$, $a_\tau = b_\tau = 1$	1.6	0.202
$\alpha = 1$, $\kappa = 100$, $a_\tau = b_\tau = 0.1$	1.3	0.204

The mechanism is the marginal likelihood $\int K(y_i|\omega)\, dP_0(\omega)$ that multiplies α in the urn probability (23.6): marginalizing μ out of P_0 at fixed τ gives $y_i \sim \text{N}(\mu_0, (1+\kappa)/\tau)$, so raising κ from 1 to 100 inflates that predictive scale by $\sqrt{101/2} = 7.1$ and divides the density of a central y_i by about the same factor. Opening a new cluster therefore costs an extra factor of roughly 7 in probability and the sampler collapses to one or two broad components, with the posterior mean density still almost unchanged — a single Gaussian at the empirical mean and variance already achieves L_1 error 0.207 on this unimodal target.