CHAPTER

19

Statistics and Probability

Statistics

People deal with large amounts of information every day. When we read newspapers, watch television or open our mail, we may be looking at information that has been organised so that we can understand it easily. For example, we can tell how much water we use at home by looking at a water bill.

When we collect, organise, represent and analyse information, we are using **statistics**. The information is collected and we call this information **data**. A person who does this kind of work is called a statistician.

Analysing statistical data can help us understand more about a group of people, the habits of animals or changes in the weather over time. This information helps us make decisions and predict outcomes.

One very important aspect of statistics is collecting representative data. This can be done by taking a census or a survey or recording data through observations. These ideas are discussed and developed using many of the techniques that have been introduced in earlier years.

19A Comparing means and medians

Mean

You have already heard of the mean and know that it is commonly called the average. We recall that to calculate the mean, we find the sum of the values and divide this by the number of values.

Mean

$$\text{mean} = \frac{\text{sum of values}}{\text{number of values}}$$

Example 1

The heights of eight people measured in centimetres are shown below:

178 187 175 183 174 180 177.5 182.5

Find the average.

Solution

$$\text{Average} = \frac{178+187+175+183+174+180+177.5+182.5}{8}$$
$$= 179.625 \text{ cm}$$

Median

The **median** is the 'middle value' when all values are arranged in order of size. Here are some numbers in order of size:

2, 2, 3, 3, 3, 4, 5, 11, 13, 18, 18, 19, 21

This data set has an odd number of values. The middle value is 5, since it has the same number of values on either side of it. Hence the median of this dataset is 5.

Here are some more numbers:

1, 3, 4, 4, 5, 6, 8, 11, 13, 13, 19, 21

This data set has an even number of values. The middle values are 6 and 8. We take the average of 6 and 8 to calculate the median.

$$\text{Median} = \frac{6+8}{2} = 7$$

Hence, the median of this data set is 7 even though it does not occur in the dataset.

ICE-EM MATHEMATICS YEAR 8
ICE-EM Mathematics 8 3ed ISBN 978-1-108-40429-7

Median

- When the data set has an odd number of values, the median is the middle value.
- When the number of values is even, the median is the average of the two middle values.

Example 2

Students measured their heights to the nearest centimetre, and recorded the results shown below.

164 168 167 158 164 154 170 175 164 168

Calculate the median.

Solution

Arrange the data in order:

154, 158, 164, 164, 164, 167, 168, 168, 170, 175

The median lies between 164 and 167, so we need to take the average of these two values.

$$\text{Median} = \frac{164 + 167}{2}$$
$$= 165.5$$

Example 3

Students measured their heights to the nearest centimetre, and recorded the results shown below.

164 168 167 158 164 154 170 175 164 168

Calculate the mean.

Solution

The mean is found by dividing the sum of the values by the number of values in the dataset.

$$\text{Mean} = \frac{(164 + 168 + 167 + 158 + 164 + 154 + 170 + 175 + 164 + 168)}{10}$$
$$= 165.2 \text{ cm}$$

The difference between the mean and median can make quite an impact. Consider the following examples of the mean and median of house prices. You might think that you could not afford to buy a house in the suburb in Example 4, based on the mean. This is because the mean is affected by the two extremely high prices. The median gives a clearer picture of what the 'average' house in this suburb might cost.

ICE-EM Mathematics 8 3ed ISBN 978-1-108-40429-7

Example 4

Listed below are some house prices achieved at auction last weekend.

$320\,000, $299\,000, $308\,000, $335\,000, $1\,005\,000, $325\,000, $985\,000

a Calculate the average house price.

b Calculate the mean, excluding the two extremely high prices.

c What is the median of all of the house prices?

Solution

a $$\text{Mean} = \frac{\text{sum of values}}{\text{number of values}}$$
$$= \frac{(320\,000 + 299\,000 + 308\,000 + 335\,000 + 1\,005\,000 + 325\,000 + 985\,000)}{7}$$
$$= \$511\,000$$

b $$\text{Mean} = \frac{320\,000 + 299\,000 + 308\,000 + 335\,000 + 325\,000}{5}$$
$$= \$317\,400$$

c Arrange the values in order. The median is the middle value.
$299\,000, $308\,000, $320\,000, $325\,000, $335\,000, $985\,000, $1\,005\,000
The median is $325\,000.

Stem-and-leaf plots

Stem-and-leaf plots were introduced in *ICE-EM Mathematics Year 7*.

Example 5

The heights of 20 students, in centimetres, are given below.

164	158	152	167	146	149	167	171	181	154
167	158	164	172	176	180	178	165	159	153

a Represent this information on a stem-and-leaf plot.

b Find the median height.

c Find the average height.

ICE-EM Mathematics 8 3ed ISBN 978-1-108-40429-7

Solution

a Since each data value contains three digits, the first two will be the stem and the last digit will be the leaf. Also, since the smallest height is 146 cm and the largest height is 181 cm, the stems will be 14, 15, 16, 17 and 18.

This produces the following stem-and-leaf plot.

14	6 9
15	2 3 4 8 8 9
16	4 [4 5] 7 7 7
17	1 2 6 8
18	1 0

17 | 2 means 172

b There are 20 items, so the median is the average of 164 and 165, which are the 10th and 11th items of the ranked data set.

$$\text{Median} = \frac{164 + 165}{2} = 164.5$$

c Mean

$$= \frac{164 + 158 + 152 + 167 + 146 + 149 + 167 + 171 + 181 + 154 + 167 + 158 + 164 + 172 + 176 + 180 + 178 + 165 + 159 + 153}{20}$$

$$= 164.05$$

In this case the mean and median are quite close.

Example 6

Form a stem-and-leaf plot for the following data and give the median and mean of the data below.

59, 57, 56, 47, 45, 43, 36, 33, 32, 31, 31, 30, 30, 25, 24, 23, 22, 11

Solution

The 18 data items have been ordered in descending order.

1	1
2	2 3 4 5
3	0 0 1 1 2 3 6
4	3 5 7
5	6 7 9

4 | 3 means 43

$$\text{Median} = \frac{31 + 32}{2} = 31.5$$

$$\text{Mean} = \frac{59 + 57 + 56 + 47 + 45 + 43 + 36 + 33 + 32 + 31 + 31 + 30 + 32 + 25 + 24 + 23 + 22 + 11}{18}$$

$$= 35\tfrac{5}{18}$$

ICE-EM Mathematics 8 3ed ISBN 978-1-108-40429-7

If a data set contains items that are clearly 'very different' to most of the items, then the median is the better choice to give you an idea of centre.

For example:

12, 17, 19, 23, 24, 34, 38

has a mean of 23.857 correct to three decimal places and the median is 23.

The data set:

12, 17, 19, 23, 24, 34, 118

has a mean of 35.286 correct to three decimal places and the median is 23.

The mean is changed by the value 78. The median does not change at all.

Exercise 19A

Example 3

1 The mean of each of the following data sets is 50. Determine the median of each of these.

a 23, 56, 37, 29, 58, 97

b 48, 49, 50, 50, 51, 52

c 0,102,58,67,71,2

Example 2, 3

2 The weights of a group of students, in kilograms, are given below.

47 46 45 46 42 41 45 41 49 46 42 43

a What is the median?

b Calculate the mean, correct to two decimal places.

3 The mean price of bags of potatoes at different supermarkets was \$7.50. The sum of the data was \$90.00. How many bags of potatoes were included in the survey?

Example 4

4 Sue spent the following amounts on her lunch each day over the course of two working weeks.

\$14 \$6 \$12 \$44.50 \$8.50 \$12.50 \$8 \$10 \$8.50 \$9.50

a Calculate the median for these data.

b Calculate the mean for these data.

c Compare the median and mean and comment on which is the better indicator of how much Sue usually spent on her daily lunch.

5 A list of data has 10 entries. Each entry is 1, 2 or 3. What could the list be if the average is:

a 1? **b** 2? **c** 3?

Example 6

6 For the stem-and-leaf plot shown, calculate the mean and find the median.

1	5 5 6 8 9
2	4 5
3	0 0 1 1 2 3 5
4	0 1 2 3 3 5

ICE-EM Mathematics 8 3ed ISBN 978-1-108-40429-7

7 The chest measurement in centimetres of 23 people is taken. The results are recorded in the stem-and-leaf plot shown.

8	8 9
9	3 3 3 4 5 6 7 9 9
10	0 1 2 3 4 6 6 7 7
11	1 1 9

a Find the median.

b Find the mean.

c Find the mean and median if the readings less than 90 and greater than 110 are not included.

8 The following list gives the area in hectares of each of the suburbs of a city.

7.5 2.2 5.2 19.2 2.4 41.3 11.3 27.6 9.0 2.3 28.4 3.2 3.6

a Find the mean and the median areas.

b Which do you think is a better measure of centre for the data set? Explain your answer.

9 The birth weights, in kilograms, of the first 20 babies born at a hospital in a selected month are as follows.

3.0 2.8 3.6 2.8 3.6 3.7 3.2 3.9 3.6 4.2
3.7 2.7 3.1 3.0 2.5 2.6 3.6 2.4 2.9 3.2

a Represent these data with a stem-and-leaf plot.

b Find the median value.

c Find the mean value.

10 **a** Find three different sets of four positive whole numbers that have a mean of 3 and a median of 2.

b Find seven different sets of five positive whole numbers that have a mean of 3 and a median of 2.

19B Sampling data

Population

An investigator usually wants to generalise about a whole class of individuals or objects. This class is called a **population**. For example, consider the following scenario. You have come up with a training strategy that you believe enables participants to increase their swimming speeds by at least 10%. You decide to set up a study to test the effectiveness of your training strategy. If your participants seem to benefit from the program, you would like to be able to say more than just that the program worked for these particular people. You would like to be able to generalise and say that it would be effective for a larger group. This larger group, the population, could be the members of a swimming club.

ICE-EM Mathematics 8 3ed ISBN 978-1-108-40429-7

Likewise, if a cure for a disease that works for one group of patients has been found, you would like to be able to conclude that it will work for all similar patients. Here the population is the class of similar patients.

An early step in your study should be to determine exactly what you want your population to be. It could be people in certain age group living in a particular state or attending your school.

You start with the population and with a question you want to answer.

Census

Sometimes it is possible to conduct your investigation for every member of the population; this is called taking a **census**.

A census of the whole population is taken in Australia every 5 years. The Australian Census aims to measure accurately the number of people in Australia on Census night wherever they are, from Australia's research hubs in Antarctica to remote Indigenous communities in northern Australia.

A census is the most comprehensive way of providing a snapshot of the people of Australia, our key characteristics and where we live. Census data support planning, decision making and funding at all levels of government, and are behind the services and facilities you use in your area every day.

It is often impossible to carry out a census. It is then necessary to look at a subset of the population.

Sampling data

The subset of the population that you choose to work with is called a **sample**. You can select a sample from a population in many different ways. Unfortunately, not all samples are equally useful if you want to be able to generalise your results.

The results of a study conducted with members of a swimming club, no matter how chosen, would not be appropriate if you wanted to make statements about the effectiveness of your swimming training program for everybody. The method of choosing a sample is very important. The best methods of sampling involve the planned introduction of chance.

Simple random sample

The simplest way of obtaining a representative sample from a population is to select a **simple random sample**. In a simple random sample, each member of the population has an equal chance of being selected; that is, no member of the population is systematically excluded from the sample nor are any particular members of the population more likely to be included. Each member of the sample is also selected **independently**; that is, the selection of a member in no way influences the selection of another member.

Uses of sampling

An example of the need for sampling is to investigate whether particular species of animals are prospering. In particular if we wanted to discover whether green tree frogs in an area of Queensland are surviving and are healthy, we might collect information about the numbers of frogs living in different areas, the weight of each frog and length of the hind legs. The type of information collected would depend on the questions we wanted to answer.

It would not be practical to collect information about all the green tree frogs in Queensland. It is not possible. Instead, we usually collect information about a smaller group within a population. The smaller group is a sample.

ICE-EM Mathematics 8 3ed ISBN 978-1-108-40429-7

Once scientists have some information about the sample, they try to make valid predictions about the entire population of frogs.

Usually, there are some numerical facts about the population that the investigators want to know, for example:

- the mean weight of a frog
- the percentage of frogs that have spots on their necks.

These cannot be determined exactly, but can only be estimated from a sample. Then an important issue is accuracy. That is, how close are the estimates going to be?

Selecting random samples

If you wanted to randomly select a sample of students in a school, you could write each student's school number on a small piece of paper, place the pieces of paper in a bucket, making sure that the slips of paper were well mixed up, and then withdraw a piece of paper. Mix up the pieces again and select the second, and so on. This could also be done with balls with numbers on them placed in a bin. Calculators and Excel have random number generators which could be used for this process.

See www.cambridge.edu.au/go for information on how to do this with Excel or a calculator.

19C Variation of means and proportions

In this chapter, consideration of variation across datasets leads us to explore the variation of sample means and of sample proportions across datasets collected or obtained under the same or similar circumstances. Sample proportions are considered for categorical data, and sample means for numerical data.

Categorical data

Categorical data are where each observation falls into one of a number of distinct categories.

Such data are everywhere in everyday life, for example:

- gender
- hair colour
- place of birth.

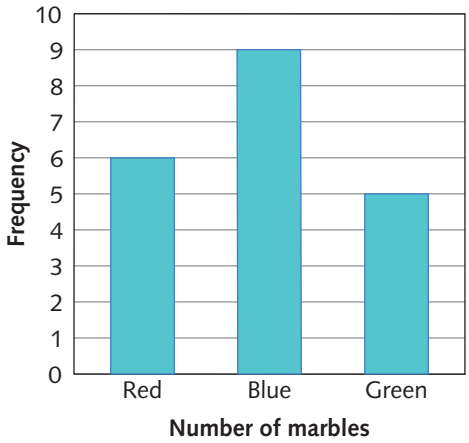
Relative frequencies for samples

A large bin contains 200 red, blue and green balls. They are thoroughly mixed up. Twenty balls are withdrawn without looking, and the colour is noted. Here the population is the bin of balls and the random sample of each ball consists of balls that have been withdrawn.

ICE-EM Mathematics 8 3ed ISBN 978-1-108-40429-7

These data obtained are categorical data. In this sample, there are 6 red balls, 9 blue balls and 5 green balls. The categories are the colours.

A column graph has been constructed for this data.



One way of describing this data is through **relative frequencies** or **proportions**.

$$\text{Relative frequency (proportion)} = \frac{\text{frequency}}{\text{size of data set}}$$

The relative frequency is sometimes given as a percentage.

In this case the sample dataset has 20 items. The frequencies and relative frequencies are shown.

Category	Red	Blue	Green
Frequency	6	9	5
Relative frequency as a percentage	30%	45%	25%

Variability

Thirty samples of 20 balls are taken. The number of red balls in each sample is recorded:

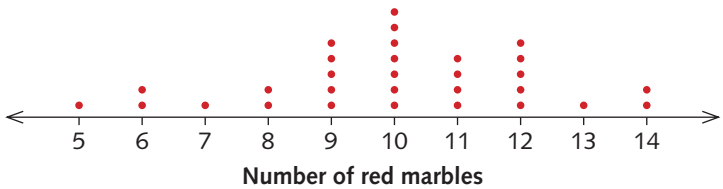
8 9 9 11 10 11 6 12 10 14 11 14 9 9 12 12
12 7 10 13 5 9 10 12 10 11 10 10 8 6

This is summarised in the following table.

Number of red balls	5	6	7	8	9	10	11	12	13	14
Frequency	1	2	1	2	5	7	4	5	1	2

This table tells us that 9 red balls were obtained from 5 of the samples, 10 red balls were obtained from 7 of the samples, and so on.

A dot plot of the results for the number of red balls obtained from the 30 samples is shown below.



ICE-EM MATHEMATICS YEAR 8
ICE-EM Mathematics 8 3ed ISBN 978-1-108-40429-7

The frequency of each proportion (relative frequency) is given in the following table.

Proportion of red balls	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%
Frequency	1	2	1	2	5	7	4	5	1	1

You can see that the sample proportion of red balls varies from 25% to 70%.

For categorical data, we are interested in relative frequencies or proportions of the different categories. If we have sample data that are representative of some general situation, we are interested in using the sample data to **estimate** proportions for the more general situation. In the case above, it would be estimating how many red balls there are in the bin.

Great care must be taken with the method of selecting the sample. At this stage we have no method for determining the accuracy of our estimates.

Numerical data

Some examples of numerical data are:

- time in minutes to get to work
- length in centimetres of the left hands of 13-year-old girls
- weight of boys who are in Year 9.

Sample means

Rods are known to have lengths between 50 cm and 100 cm. They are measured to the nearest centimetre. A random sample of size 20 is taken. The population is all of the rods.

The results are as shown below.

61 79 59 91 72 99 80 73 89 89 91 54 51 78 56 75 69 90 60 66

Mean = 74

The mean of the sample is 74 cm, correct to the nearest whole number. This provides us with an *estimate* of the mean lengths of the rods for the population.

Variability of sample means

Ten random samples of 20 rods are taken from the same population and the means recorded.

Sample number	1	2	3	4	5	6	7	8	9	10
Sample mean	74	77	70	73	74	66	76	70	74	81

The sample means vary from 66 through to 81.

The sampling procedure was also undertaken with samples of size 100. (There are a lot more than 100 rods.) The means are given correct to the nearest whole number.

Sample number	1	2	3	4	5	6	7	8	9	10
Sample mean	76	73	77	74	74	74	75	70	75	77

You can see that the means vary less with this larger sample size. The sample means vary from 70 to 77.

ICE-EM Mathematics 8 3ed ISBN 978-1-108-40429-7

Exercise 19C

1 A bin contains black, blue and green marbles. A sample of 20 marbles is taken from the bin. The results are as shown:

green, black, green, blue, green, black, green, blue, black, black, green, black, black, green, green, green, black, green, green, black

a Draw a column graph showing the frequency of each category.

b Find the relative frequency of each category.

c Find the relative frequency of each category as a percentage.

2 A second sample of 20 marbles is taken from the bin. The results are as shown in the table.

Category	Green	Blue	Black
Frequency	11	5	4

Calculate the relative frequency of each category.

3 The mass of 22 people is recorded in kilograms as shown.

77	85.5	63	80.5	79.5	94	66	69	65	58	69.5
73	74	68	80	66	54.5	64	84	73	89	94

a Two random samples of 10 people are chosen as shown. Find the mean of each sample.

i 66, 80, 63, 73, 84, 94, 69.5, 64, 79.5, 65

ii 65, 58, 94, 73, 77, 64, 89, 84, 66, 63

b List the 10 smallest masses and calculate the average.

c List the 10 largest masses and calculate the average.

Note: The average mass of the 22 people is 73.93 kg.

4 A section of the Murray River is known to contain three types of fish. A random sample of fish is taken from the river and the results are recorded. The fish are released back into the river.

Type of fish	Murray cod	Redfin	Catfish	Murray perch
Number in sample	15	10	12	3

a Draw a column graph with this information.

b Find the relative frequency of each type of fish.

ICE-EM MATHEMATICS YEAR 8
ICE-EM Mathematics 8 3ed ISBN 978-1-108-40429-7

5 A sample of red-eyed tree frogs is taken from an area surrounding a pond in central Queensland.

The lengths of the frogs in millimetres are as follows.

33 35 45 50 55 58 32 56 60 32

Find the sample mean of these data.

6 Use the data shown.

1 3 5 7 9 11 13 15 17

a Find the smallest and the largest sample mean obtainable from samples of 5.

b The average of the whole dataset is 8.5. Is there a sample of 5 which has a mean of 8.5?

7 The 1420 students at a school were asked 'Which method of transport do you use to get to school?' The responses were as follows.

Method	Train	Tram	By foot	Cycle	Car
Number	450	320	150	240	260

a Construct a table showing the relative frequency expressed as a percentage (correct to one decimal place) of each type of transport.

b Draw a column graph representing this data.

Activity 1 – Number of red counters (sample proportion)

A jar contains 100 counters. There are both red counters and black counters in the jar.

Take samples of 10 with replacement and good mixing and estimate the proportion of red counters in the jar. Compare results.

Activity 2 – Colourful Yummies (sample proportion)

Packets of Colourful Yummies come in several colours and are sold in packets that contain all of the colours.

Population: This could be the packing of Colourful Yummies by the manufacturer or a large bin of the sweet collected in the classroom. The benefit of the second is that the population proportion of a particular colour could be known.

The questions could then be:

a What is the proportion of red Colourful Yummies in a packet prepared by the manufacturer?

b What is the proportion of red Colourful Yummies in the large bin?

Sample: Samples could be prepared from the large bin or use the packages prepared by the manufacturer.

Compare the results of all students and display the results with a dot plot. Each of these results is an estimate for the proportion of red Colourful Yummies in the population.

ICE-EM Mathematics 8 3ed ISBN 978-1-108-40429-7