17

Sampling and estimation

Objectives

- To understand **random samples** and how they may be obtained.
- To define the **population proportion** and the **sample proportion**.
- To introduce the concept of the sample proportion as a random variable.
- To investigate the **sampling distribution** of the sample proportion both exactly (for small samples) and through simulation.
- To use a **normal distribution** to approximate the sampling distribution of the sample proportion.
- To use the sample proportion as a **point estimate** of the population proportion.
- To find **confidence intervals** for the population proportion.
- To introduce the concept of **margin of error**, and illustrate how this varies both with level of confidence and with sample size.

There is more to a complete statistical investigation than data analysis. First, we should concern ourselves with the methods used to collect the data. In practice, the purpose of selecting a sample and analysing the information collected from the sample is to make some sort of conclusion, or inference, about the population from which the sample was drawn. Therefore we want the sample we select to be representative of this population.

For example, consider the following questions:

- What proportion of Year 12 students intend to take a gap year?
- What proportion of people aged 18–25 regularly attend church?
- What proportion of secondary students take public transport to school?

While we can answer each of these questions for a sample of people from the group, we really want to know something about the whole group. How can we generalise information gained from a sample to the population, and how confident can we be in that generalisation?

17A Populations and samples

The set of all eligible members of a group which we intend to study is called a **population**. For example, if we are interested in the IQ scores of the Year 12 students at ABC Secondary College, then this group of students could be considered a population; we could collect and analyse all the IQ scores for these students. However, if we are interested in the IQ scores of all Year 12 students across Australia, then this becomes the population.

Often, dealing with an entire population is not practical:

- The population may be too large – for example, all Year 12 students in Australia.
- The population may be hard to access – for example, all blue whales in the Pacific Ocean.
- The data collection process may be destructive – for example, testing every battery to see how long it lasts would mean that there were no batteries left to sell.

Nevertheless, we often wish to make statements about a property of a population when data about the entire population is unavailable.

The solution is to select a subset of the population – called a **sample** – in the hope that what we find out about the sample is also true about the population it comes from. Dealing with a sample is generally quicker and cheaper than dealing with the whole population, and a well-chosen sample will give much useful information about this population. How to select the sample then becomes a very important issue.

Random samples

Suppose we are interested in investigating the effect of sustained computer use on the eyesight of a group of university students. To do this we go into a lecture theatre containing the students and select all the students sitting in the front two rows as our sample. This sample may be quite inappropriate, as students who already have problems with their eyesight are more likely to be sitting at the front, and so the sample may not be typical of the population. To make valid conclusions about the population from the sample, we would like the sample to have a similar nature to the population.

While there are many sophisticated methods of selecting samples, the general principle of sample selection is that the method of choosing the sample should not favour or disfavour any subgroup of the population. Since it is not always obvious if the method of selection will favour a subgroup or not, we try to choose the sample so that every member of the population has an equal chance of being in the sample. In this way, all subgroups have a chance of being represented. The way we do this is to choose the sample at random.

A sample of size n is called a **simple random sample** if it is selected from the population in such a way that every subset of size n has an equal chance of being chosen as the sample. In particular, every member of the population must have an equal chance of being included in the sample.

To choose a sample from the group of university students, we could put the name of every student in a hat and then draw out, one at a time, the names of the students who will be in the sample.

ISBN 978-1-009-11049-5 © Michael Evans et al 2023 Cambridge University Press

Choosing the sample in an appropriate manner is critical in order to obtain usable results.

Example 1

A researcher wishes to evaluate how well the local library is catering to the needs of a town's residents. To do this she hands out a questionnaire to each person entering the library over the course of a week. Will this method result in a random sample?

Solution

Since the members of the sample are already using the library, they are possibly satisfied with the service available. Additional valuable information might well be obtained by finding out the opinion of those who do not use the library.

A better sample would be obtained by selecting at random from the town's entire population, so the sample contains both people who use the library and people who do not.

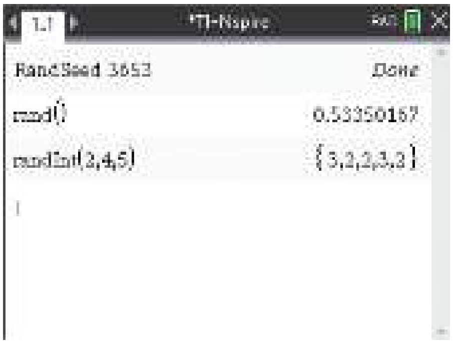
Thus, we have a very important consideration when sampling if we wish to generalise from the results of the sample.

In order to make valid conclusions about a population from a sample, we would like the sample chosen to be representative of the population as a whole. This means that all the different subgroups present in the population appear in the sample in similar proportions as they do in the population.

One very useful method for drawing random samples is to generate random numbers using a calculator or a computer.

Using the TI-Nspire

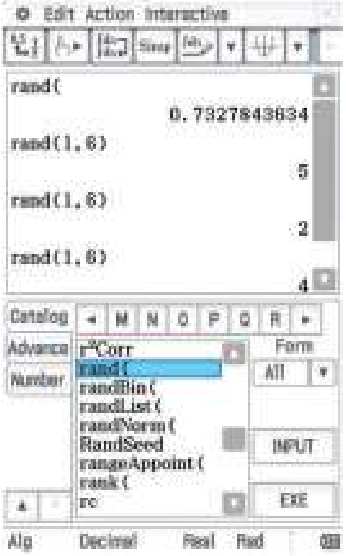
- In a **Calculator** page, go to Menu > **Probability** > **Random** > **Seed** and enter the last 4 digits of your phone number. This ensures that your random-number starting point differs from the calculator default.
- For a random number between 0 and 1, use Menu > **Probability** > **Random** > **Number**.
- For a random integer, use Menu > **Probability** > **Random** > **Integer**. To obtain five random integers between 2 and 4 inclusive, use the command randInt(2, 4, 5) as shown.



```
RandSeed 3653                Done
rand()                 0.53350167
randInt(2,4,5)        {3,2,2,3,3}
```

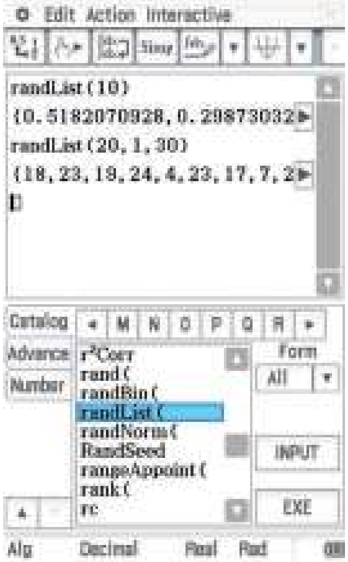
Using the Casio ClassPad

- In Main $\sqrt{\alpha}$, press the Keyboard button.
- Find and then select Catalog by first tapping ▼ at the bottom of the left sidebar.
- Scroll across the alphabet to the letter R.



```
rand(
                    0.7327843634
rand(1,6)
                               5
rand(1,6)
                               2
rand(1,6)
                               4
```

- To generate a random number between 0 and 1:
  - In Catalog, select **rand(**.
  - Tap EXE.
- To generate three random integers between 1 and 6 inclusive:
  - In Catalog, select **rand(**.
  - Type: 1, 6)
  - Tap EXE three times.
- To generate a list of 10 random numbers between 0 and 1:
  - In Catalog, select **randList(**.
  - Type: 10)
  - Tap EXE.
  - Tap ► to view all the numbers.
- To generate a list of 20 random integers between 1 and 30 inclusive:
  - In Catalog, select **randList(**.
  - Type: 20, 1, 30)
  - Tap EXE.
  - Tap ► to view all the integers.



```
randList(10)
{0.5182070928,0.29873032▶
randList(20,1,30)
{18,23,19,24,4,23,17,7,2▶
```

Example 2

Use a random number generator to select a group of six students from the following class:

- Denice
- Matt
- Teresa
- Sue
- Shanyn
- Mark
- Arnold
- Nick
- Miller
- William
- Lulu
- Darren
- Tom
- David
- Lacey
- Janelle
- Mike
- Jane
- Georgia
- Jaimie

Solution

First assign a number to each member of the class:

- Denice (1)
- Matt (2)
- Teresa (3)
- Sue (4)
- Shanyn (5)
- Mark (6)
- Arnold (7)
- Nick (8)
- Miller (9)
- William (10)
- Lulu (11)
- Darren (12)
- Tom (13)
- David (14)
- Lacey (15)
- Janelle (16)
- Mike (17)
- Jane (18)
- Georgia (19)
- Jaimie (20)

Generating six random integers from 1 to 20 gives on this occasion: 4, 19, 9, 2, 13, 14. The sample chosen is thus:

Sue, Georgia, Miller, Matt, Tom, David

Note: In this example, we want a list of six random integers without repeats. We do not add a randomly generated integer to our list if it is already in the list.

The sample proportion as a random variable

Suppose that our population of interest is the class of students from Example 2, and suppose further that we are particularly interested in the proportion of female students in the class. This is called the **population proportion** and is generally denoted by p. The population proportion p is constant for a particular population.

$$\text{Population proportion } p = \frac{\text{number in population with attribute}}{\text{population size}}$$

In this class there are 10 females, so the proportion of female students in the class is

$$p = \frac{10}{20} = \frac{1}{2}$$

Now consider the proportion of female students in the sample chosen:

Sue, Georgia, Miller, Matt, Tom, David

The proportion of females in the sample may be calculated by dividing the number of females in the sample by the sample size. In this case, there are two females in the sample, so the proportion of female students in the sample is $\frac{2}{6} = \frac{1}{3}$. This value is called the **sample proportion** and is denoted by $\hat{p}$. (We say 'p hat'.)

$$\text{Sample proportion } \hat{p} = \frac{\text{number in sample with attribute}}{\text{sample size}}$$

Note that different symbols are used for the sample proportion and the population proportion, so that we don't confuse them.

In this particular case, $\hat{p} = \frac{1}{3}$, which is not the same as the population proportion $p = \frac{1}{2}$. This does not mean there is a problem. In fact, each time a sample is selected the number of females in the sample will vary. Sometimes the sample proportion $\hat{p}$ will be $\frac{1}{2}$, and sometimes it will not.

- The population proportion p is a **population parameter**; its value is constant.
- The sample proportion $\hat{p}$ is a **sample statistic**; its value is not constant, but varies from sample to sample.

Example 3

Use a random number generator to select another group of six students from the same class, and determine the proportion of females in the sample.

- Denice (1)
- Matt (2)
- Teresa (3)
- Sue (4)
- Shanyn (5)
- Mark (6)
- Arnold (7)
- Nick (8)
- Miller (9)
- William (10)
- Lulu (11)
- Darren (12)
- Tom (13)
- David (14)
- Lacey (15)
- Janelle (16)
- Mike (17)
- Jane (18)
- Georgia (19)
- Jaimie (20)

Solution

Generating another six random integers from 1 to 20 gives 19, 3, 11, 9, 15, 1.

The sample chosen is thus:

Georgia, Teresa, Lulu, Miller, Lacey, Denice

For this sample, we have

$$\hat{p} = \frac{5}{6}$$

Since $\hat{p}$ varies according to the contents of the random samples, we can consider the sample proportions $\hat{p}$ as being the values of a random variable, which we will denote by $\hat{P}$. We investigate this idea further in the next section.

Summary 17A

- A **population** is the set of all eligible members of a group which we intend to study.
- A **sample** is a subset of the population which we select in order to make inferences about the population. Generalising from the sample to the population will not be useful unless the sample is representative of the population.
- A sample of size n is called a **simple random sample** if it is selected from the population in such a way that every subset of size n has an equal chance of being chosen as the sample. In particular, every member of the population must have an equal chance of being included in the sample.
- The **population proportion** p is the proportion of individuals in the entire population possessing a particular attribute, and is constant.
- The **sample proportion** $\hat{p}$ is the proportion of individuals in a particular sample possessing the attribute, and varies from sample to sample.
- The sample proportions $\hat{p}$ are the values of a random variable $\hat{P}$.

Exercise 17A

Example 1

1 In order to determine the sort of film in which to invest his money, a producer waits outside a theatre and asks people as they leave whether they prefer comedy, drama, horror or science fiction. Do you think this is an appropriate way of selecting a random sample of movie goers? Explain your answer.

2 A market researcher wishes to find out how people spend their leisure time. She positions herself in a shopping mall and asks shoppers as they pass to fill out a short questionnaire.

a Do you think this sample will be representative of the general population? Explain.

b How would you suggest that the sample could be chosen?

3 To investigate people’s attitudes to control of gun ownership, a television station conducts a phone-in poll, where people are asked to telephone one number if they are in favour of tighter gun control, and another if they are against. Is this an appropriate method of choosing a random sample? Give reasons for your answer.

4 A researcher wishes to select five guinea pigs at random from a large cage containing 20 guinea pigs. In order to select her sample, she reaches into the cage and (gently) pulls out five guinea pigs.

a Do you think this sample will be representative of the general population? Explain.

b How would you suggest the sample could be chosen?

5 In order to estimate how much money young people spend on takeaway food, a questionnaire is sent to several schools randomly chosen from a list of all schools in the state, to be given to a random selection of students in the school. Is this an appropriate method of choosing a random sample? Give reasons for your answer.

Example 2

6 Use a random number generator to select a random sample of size 3 from the following list of people:

- Karen
- Kate
- Trinh
- Paul
- Lina
- Alexander
- Juliet
- Craig
- Conrad
- Fairlie
- Kylie
- Edward
- Kelly
- Rani
- Maree
- Janet
- Fleur
- Connie
- Aden
- Wolfgang
- Zoe
- Cara
- Noel
- Judy
- Andrew

7 In a survey to obtain adults’ views on unemployment, people were stopped by interviewers as they came out of:

a a travel agency **b** a supermarket **c** an employment-services centre.

What is wrong with each of the methods of sampling listed here? Describe a better method of choosing the sample.

8 A marine biologist wishes to estimate the total number of crabs on a rock platform which is 10 metres square. It would be impossible to count them all individually, so she places a 1-metre-square frame at five random locations on the rock platform, and counts the number of crabs in the frame. To estimate the total number, she will multiply the average number in the frame by the total area of the rock platform.

a Explain how a random number generator could be used to select the five locations for the frame.

b Will this give a good estimate of the crab population?

9 In order to survey the attitude of parents to the current uniform requirements, the principal of a school selected 100 students at random from the school roll, and then interviewed their parents. Do you think this group of parents would form a simple random sample?

10 A television station carried out a poll to find out if the public felt that mining should be allowed in a particular area. People were asked to ring one number to register a ‘yes’ vote and another to register a ‘no’ vote. The results showed that 77% of people were in favour of mining proceeding. Comment on the results.

11 A market-research company decided to collect information concerning the way people use their leisure time by phoning a randomly chosen group of 1000 people at home between 7 p.m. and 10 p.m. on weeknights. The final report was based on the responses of only the 550 people of those sampled who could be found at home. Comment on the validity of this report.

12 In a certain school, 35% of the students travel on the school bus. A group of 100 students were selected in a random sample, and 42 of them travel on the school bus. In this example:

a What is the population?

b What is the value of the population proportion p?

c What is the value of the sample proportion $\hat{p}$?

13 Of a random sample of 100 homes, 22 were found to have central heating.

a What proportion of these homes have central heating?

b Is this the value of the population proportion p or the sample proportion $\hat{p}$?

Example 3

14 Use a random number generator to select another group of six students from the class listed below, and determine the proportion of females in the sample:

- Denice (1)
- Matt (2)
- Teresa (3)
- Sue (4)
- Shanyn (5)
- Mark (6)
- Arnold (7)
- Nick (8)
- Miller (9)
- William (10)
- Lulu (11)
- Darren (12)
- Tom (13)
- David (14)
- Lacey (15)
- Janelle (16)
- Mike (17)
- Jane (18)
- Georgia (19)
- Jaimie (20)

17B The exact distribution of the sample proportion

We have seen that the sample proportion varies from sample to sample. We can use our knowledge of probability to further develop our understanding of the sample proportion.

Sampling from a small population

Suppose we have a bag containing six blue balls and four red balls, and from the bag we take a sample of size 4. We are interested in the proportion of blue balls in the sample. We know that the population proportion is equal to $\frac{6}{10} = \frac{3}{5}$. That is,

$$p = 0.6$$

The probabilities associated with the possible values of the sample proportion $\hat{p}$ can be calculated either by direct consideration of the sample outcomes or by using our knowledge of selections. Recall that

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

is the number of different ways to select x objects from n objects.

Example 4

A bag contains six blue balls and four red balls. If we take a random sample of size 4, what is the probability that there is one blue ball in the sample ($\hat{p} = \frac{1}{4}$)?

Solution

Method 1

Consider selecting the sample by taking one ball from the bag at a time (without replacement). The favourable outcomes are RRRB, RRBR, RBRR and BRRR, with

$$\begin{aligned}&\Pr(\{\text{RRRB}, \text{RRBR}, \text{RBRR}, \text{BRRR}\})\\ &= \left(\frac{4}{10} \times \frac{3}{9} \times \frac{2}{8} \times \frac{6}{7}\right) + \left(\frac{4}{10} \times \frac{3}{9} \times \frac{6}{8} \times \frac{2}{7}\right) + \left(\frac{4}{10} \times \frac{6}{9} \times \frac{3}{8} \times \frac{2}{7}\right) + \left(\frac{6}{10} \times \frac{4}{9} \times \frac{3}{8} \times \frac{2}{7}\right)\\ &= \frac{4}{35}\end{aligned}$$

Method 2

In total, there are $\binom{10}{4} = 210$ ways to select 4 balls from 10 balls.

There are $\binom{4}{3} = 4$ ways of choosing 3 red balls from 4 red balls, and there are $\binom{6}{1} = 6$ ways of choosing one blue ball from 6 blue balls.

Thus the probability of obtaining 3 red balls and one blue ball is equal to

$$\frac{\binom{4}{3} \times \binom{6}{1}}{\binom{10}{4}} = \frac{24}{210} = \frac{4}{35}$$

The following table gives the probability of obtaining each possible sample proportion $\hat{p}$ when selecting a random sample of size 4 from the bag.

Number of blue balls in the sample (x)	0	1	2	3	4
Proportion of blue balls in the sample, $\hat{p}$	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
Probability	$\frac{1}{210}$	$\frac{24}{210}$	$\frac{90}{210}$	$\frac{80}{210}$	$\frac{15}{210}$

We can see from the table that we can consider the sample proportion as a random variable, $\hat{P}$, and we can write:

- $\Pr(\hat{P} = 0) = \frac{1}{210}$
- $\Pr\left(\hat{P} = \frac{1}{4}\right) = \frac{24}{210}$
- $\Pr\left(\hat{P} = \frac{1}{2}\right) = \frac{90}{210}$
- $\Pr\left(\hat{P} = \frac{3}{4}\right) = \frac{80}{210}$
- $\Pr(\hat{P} = 1) = \frac{15}{210}$

The possible values of $\hat{p}$ and their associated probabilities together form a probability distribution for the random variable $\hat{P}$, which can summarised as follows:

$\hat{p}$	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
$\Pr(\hat{P} = \hat{p})$	$\frac{1}{210}$	$\frac{24}{210}$	$\frac{90}{210}$	$\frac{80}{210}$	$\frac{15}{210}$

The distribution of a statistic which is calculated from a sample (such as the sample proportion) has a special name – it is called a **sampling distribution**.

Example 5

A bag contains six blue balls and four red balls. Use the sampling distribution in the previous table to determine the probability that the proportion of blue balls in a sample of size 4 is more than $\frac{1}{4}$.

Solution

$$\begin{aligned}\Pr\left(\hat{P} > \frac{1}{4}\right) &= \Pr\left(\hat{P} = \frac{1}{2}\right) + \Pr\left(\hat{P} = \frac{3}{4}\right) + \Pr(\hat{P} = 1)\\ &= \frac{90}{210} + \frac{80}{210} + \frac{15}{210}\\ &= \frac{185}{210}\\ &= \frac{37}{42}\end{aligned}$$

Sampling from a large population

Generally, when we select a sample it is from a population which is too large or too difficult to enumerate or even count – populations such as all the people in Australia, or all the cows in Texas, or all the people who will ever have asthma. When the population is so large, we assume that the probability of observing the attribute we are interested in remains constant with each selection, irrespective of prior selections for the sample.

Suppose we know that 70% of all 17-year-olds in Australia attend school. That is,

$$p = 0.7$$

We will assume that this probability remains constant for all selections for the sample.

Now consider selecting a random sample of size 4 from the population of all 17-year-olds in Australia. This time we can use our knowledge of binomial distributions to calculate the associated probability for each possible value of the sample proportion $\hat{p}$, using the probability function

$$\Pr(X = x) = \binom{4}{x} 0.7^x 0.3^{4-x} \qquad x = 0, 1, 2, 3, 4$$

The following table gives the probability of obtaining each possible sample proportion $\hat{p}$ when selecting a random sample of four 17-year-olds.

Number at school in the sample (x)	0	1	2	3	4
Proportion at school in the sample, $\hat{p}$	0	0.25	0.5	0.75	1
Probability	0.0081	0.0756	0.2646	0.4116	0.2401

Once again, we can summarise the sampling distribution of the sample proportion as follows:

$\hat{p}$	0	0.25	0.5	0.75	1
$\Pr(\hat{P} = \hat{p})$	0.0081	0.0756	0.2646	0.4116	0.2401

The population that the sample of size $n = 4$ is being taken from is such that each item selected has a probability $p = 0.7$ of success. Thus we can define the random variable

$$\hat{P} = \frac{X}{4}$$

where X is a binomial random variable with parameters $n = 4$ and $p = 0.7$. To emphasise this we can write:

x	0	1	2	3	4
$\hat{p} = \frac{x}{4}$	0	0.25	0.5	0.75	1
$\Pr(\hat{P} = \hat{p}) = \Pr(X = x)$	0.0081	0.0756	0.2646	0.4116	0.2401

Note: The probabilities for the sample proportions, $\hat{p}$, correspond to the probabilities for the numbers of successes, x.

Example 6

Use the sampling distribution in the previous table to determine the probability that, in a random sample of four Australian 17-year-olds, the proportion attending school is less than 50%.

Solution

$$\begin{aligned}\Pr(\hat{P} < 0.5) &= \Pr(\hat{P} = 0) + \Pr(\hat{P} = 0.25)\\ &= 0.0081 + 0.0756\\ &= 0.0837\end{aligned}$$

The mean and standard deviation of the sample proportion

Since the sample proportion $\hat{P}$ is a random variable with a probability distribution, we can determine values for the mean and standard deviation, as illustrated in the following example.

Example 7

Use the probability distribution to determine the mean and standard deviation of the sample proportion $\hat{P}$ from Example 6.

$\hat{p}$	0	0.25	0.5	0.75	1
$\Pr(\hat{P} = \hat{p})$	0.0081	0.0756	0.2646	0.4116	0.2401

Solution

By definition, the mean of $\hat{P}$ is given by

$$\begin{aligned}\mathrm{E}(\hat{P}) &= \sum \hat{p} \cdot \Pr(\hat{P} = \hat{p})\\ &= 0 \times 0.0081 + 0.25 \times 0.0756 + 0.5 \times 0.2646 + 0.75 \times 0.4116 + 1 \times 0.2401\\ &= 0.7\end{aligned}$$

Similarly, by definition,

$$\mathrm{sd}(\hat{P}) = \sqrt{\mathrm{E}(\hat{P}^2) - [\mathrm{E}(\hat{P})]^2}$$

We have

$$\begin{aligned}\mathrm{E}(\hat{P}^2) &= 0^2 \times 0.0081 + 0.25^2 \times 0.0756 + 0.5^2 \times 0.2646 + 0.75^2 \times 0.4116 + 1^2 \times 0.2401\\ &= 0.5425\end{aligned}$$

Thus

$$\mathrm{sd}(\hat{P}) = \sqrt{0.5425 - 0.7^2} = 0.2291$$

We can see from Example 7 that the mean of the sampling distribution in this case is actually the same as the value of the population proportion (0.7). Is this always true? Can we determine the mean and standard deviation of the sample proportion without needing to find the probability distribution?

If we are selecting a random sample of size n from a large population, then we can assume that the sample proportion is of the form

$$\hat{P} = \frac{X}{n}$$

where X is a binomial random variable with parameters n and p. From Chapter 14, the mean and variance of X are given by

$$\mathrm{E}(X) = np \qquad \text{and} \qquad \mathrm{Var}(X) = np(1-p)$$

Thus we can determine

$$\begin{aligned}\mathrm{E}(\hat{P}) &= \mathrm{E}\left(\frac{X}{n}\right)\\ &= \frac{1}{n}\mathrm{E}(X) && \text{since } \mathrm{E}(aX+b) = a\mathrm{E}(X)+b\\ &= \frac{1}{n}\times np\\ &= p\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(\hat{P}) &= \mathrm{Var}\left(\frac{X}{n}\right)\\ &= \left(\frac{1}{n}\right)^2\mathrm{Var}(X) && \text{since } \mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)\\ &= \frac{1}{n^2}\times np(1-p)\\ &= \frac{p(1-p)}{n}\end{aligned}$$

If we are selecting a random sample of size n from a large population, then the mean and standard deviation of the sample proportion $\hat{P}$ are given by

$$\mathrm{E}(\hat{P}) = p \quad \text{and} \quad \mathrm{sd}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$$

(The standard deviation of a sample statistic is called the **standard error**.)

Example 8

Use these rules to determine the mean and standard deviation of the sample proportion $\hat{P}$ from Example 6. Are they the same as those found in Example 7?

Solution

$$\mathrm{E}(\hat{P}) = p = 0.7$$

$$\mathrm{sd}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.7(1-0.7)}{4}} = 0.2291$$

These are the same as those obtained in Example 7.

ISBN 978-1-009-11049-5 Cambridge University Press

Example 9

Suppose that 70% of 17-year-olds in Australia attend school. If a random sample of size 20 is chosen from this population, find:

a the probability that the sample proportion is equal to the population proportion (0.7)

b the probability that the sample proportion lies within one standard deviation of the population proportion

c the probability that the sample proportion lies within two standard deviations of the population proportion.

Solution

a If the sample proportion is $\hat{p} = 0.7$ and the sample size is 20, then the number of school students in the sample is $0.7 \times 20 = 14$. Thus

$$\begin{aligned}\Pr(\hat{P} = 0.7) &= \Pr(X = 14)\\ &= \binom{20}{14} 0.7^{14} 0.3^{6} = 0.1916\end{aligned}$$

b We have

$$\begin{aligned}\text{sd}(\hat{P}) &= \sqrt{\frac{p(1-p)}{n}}\\ &= \sqrt{\frac{0.7(1-0.7)}{20}} = 0.1025\end{aligned}$$

Since $0.7 - 0.1025 = 0.5975$ and $0.7 + 0.1025 = 0.8025$, we find

$$\begin{aligned}\Pr(0.5975 \le \hat{P} \le 0.8025) &= \Pr(11.95 \le X \le 16.05)\\ &= \Pr(12 \le X \le 16) \qquad \text{since } X \text{ takes integer values}\\ &= 0.7795\end{aligned}$$

c Since $0.7 - 2 \times 0.1025 = 0.495$ and $0.7 + 2 \times 0.1025 = 0.905$, we find

$$\begin{aligned}\Pr(0.495 \le \hat{P} \le 0.905) &= \Pr(9.9 \le X \le 18.1)\\ &= \Pr(10 \le X \le 18)\\ &= 0.9752\end{aligned}$$

Summary 17B

- The distribution of a statistic which is calculated from a sample is called a **sampling distribution**.
- The **sample proportion** $\hat{P} = \dfrac{X}{n}$ is a random variable, where X is the number of favourable outcomes in a sample of size n.
- The distribution of $\hat{P}$ is known as the **sampling distribution** of the sample proportion.
- When the population is *small*, the sampling distribution of the sample proportion $\hat{P}$ can be determined using our knowledge of selections.

ISBN 978-1-009-11049-5

Cambridge University Press

- When the population is *large*, the sampling distribution of the sample proportion $\hat{P}$ can be determined by assuming that X is a binomial random variable with parameters n and p. In this case, the mean and standard deviation of $\hat{P}$ are given by

$$\mathrm{E}(\hat{P}) = p \quad \text{and} \quad \mathrm{sd}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$$

Skill-sheet

Exercise 17B

1 Consider a bag containing five blue and five red balls.

Example 4

Example 5

a What is p, the proportion of blue balls in the bag?

b If samples of size 3 are taken from the bag, without replacement, then a sample could contain 0, 1, 2 or 3 blue balls. What are the possible values of the sample proportion $\hat{p}$ of blue balls associated with each of these samples?

c Construct a probability distribution table which summarises the sampling distribution of the sample proportion of blue balls when samples of size 3 are taken from the bag, without replacement.

d Use the sampling distribution from **c** to determine the probability that the proportion of blue balls in the sample is more than 0.5. That is, find $\Pr(\hat{P} > 0.5)$.

2 A company employs a sales team of 20 people, consisting of 12 men and 8 women.

a What is p, the proportion of men in the sales team?

b Five salespeople are to be selected at random to attend an important conference. What are the possible values of the sample proportion $\hat{p}$ of men in the sample?

c Construct a probability distribution table which summarises the sampling distribution of the sample proportion of men when samples of size 5 are selected from the sales team.

d Use the sampling distribution from **c** to determine the probability that the proportion of men in the sample is more than 0.7.

e Find $\Pr(0 < \hat{P} < 0.7)$ and hence find $\Pr(\hat{P} < 0.7 \mid \hat{P} > 0)$.

3 A pond contains eight gold and eight black fish.

a What is p, the proportion of gold fish in the pond?

b Three fish are to be selected at random. What are the possible values of the sample proportion $\hat{p}$ of gold fish in the sample?

c Construct a probability distribution table which summarises the sampling distribution of the sample proportion of gold fish when samples of size 3 are selected from the pond.

d Use the sampling distribution from **c** to determine the probability that the proportion of gold fish in the sample is more than 0.25.

4 A random sample of three items is selected from a batch of 10 items which contains four defectives.

a What is p, the proportion of defectives in the batch?

b What are the possible values of the sample proportion $\hat{p}$ of defectives in the sample?

c Construct a probability distribution table which summarises the sampling distribution of the sample proportion of defectives in the sample.

d Use the sampling distribution from **c** to determine the probability that the proportion of defectives in the sample is more than 0.5.

e Find $\Pr(0 < \hat{P} < 0.5)$ and hence find $\Pr(\hat{P} < 0.5 \mid \hat{P} > 0)$.

Example 6

5 Suppose that a fair coin is tossed four times, and the number of heads observed.

a What is p, the probability that a head is observed when a fair coin is tossed?

b What are the possible values of the sample proportion $\hat{p}$ of heads in the sample?

c Construct a probability distribution table which summarises the sampling distribution of the sample proportion of heads in the sample.

d Use the sampling distribution from **c** to determine the probability that the proportion of heads in the sample is more than 0.7.

6 Suppose that the probability of a male child is 0.5, and that a family has five children.

a What are the possible values of the sample proportion $\hat{p}$ of male children in the family?

b Construct a probability distribution table which summarises the sampling distribution of the sample proportion of male children in the family.

c Use the sampling distribution from **b** to determine the probability that the proportion of male children in the family is less than 0.4.

d Find $\Pr(\hat{P} > 0 \mid \hat{P} < 0.8)$.

7 Suppose that, in a certain country, the probability that a person is left-handed is $\frac{1}{5}$. If four people are selected at random from that country:

a What are the possible values of the sample proportion $\hat{p}$ of left-handed people in the sample?

b Construct a probability distribution table which summarises the sampling distribution of the sample proportion of left-handed people in the sample.

c Find $\Pr(\hat{P} \geq \frac{1}{2})$.

Example 7

8 Use the sampling distribution from Question 5 to determine the mean and standard deviation of the sample proportion $\hat{P}$ of heads observed when a fair coin is tossed four times.

9 Use the sampling distribution from Question 6 to determine the mean and standard deviation of the sample proportion $\hat{P}$ of male children in a family of five children.

10 Use the sampling distribution from Question 7 to determine the mean and standard deviation of the sample proportion $\hat{P}$ of left-handed people when a sample of four people are selected.

ISBN 978-1-009-11049-5 Cambridge University Press

Example 8 **11** Suppose that the probability of rain on any day is 0.3. A random sample of 30 days is selected across the year (365 days).

a Find the probability that the proportion of rainy days in this sample is greater than 0.4.

b Find the mean and standard deviation of the sample proportion of rainy days.

12 In a certain country, it is known that 40% of people speak more than one language. The random variable $\hat{P}$ represents the proportion of people in sample of size n who speak more than one language:

a If $n = 100$, find $\Pr(\hat{P} > 0.45)$. Give your answer correct to four decimal places.

b If $n = 200$, find $\Pr(\hat{P} > 0.45)$. Give your answer correct to four decimal places.

13 An examination consists of 16 multiple-choice questions, each with four possible answers. $\hat{P}$ is the random variable representing the proportion of questions a student answers correctly if they guess the answer to every question.

a Find correct to four decimal places $\Pr\left(\hat{P} \geq \dfrac{5}{16}\right)$.

b Find correct to four decimal places $\Pr\left(\hat{P} \geq \dfrac{5}{16} \mid \hat{P} \geq \dfrac{3}{16}\right)$.

c Find the mean and standard deviation of the sample proportion of correct answers that will be achieved if a student guesses every answer.

Example 9 **14** Suppose that 65% of people in Australia support an AFL team. If a random sample of size 20 is chosen from this population, find:

a the probability that the sample proportion is equal to the population proportion

b the probability that the sample proportion lies within one standard deviation of the population proportion

c the probability that the sample proportion lies within two standard deviations of the population proportion.

17C Approximating the distribution of the sample proportion

In the previous section, we used our knowledge of probability to determine the exact distribution of the sample proportion. Working out the exact probabilities associated with a sample proportion is really only practical when the sample size is quite small (say less than 10). In practice, we are rarely working with such small samples. But we can overcome this problem by approximating the distribution of the sample proportion.

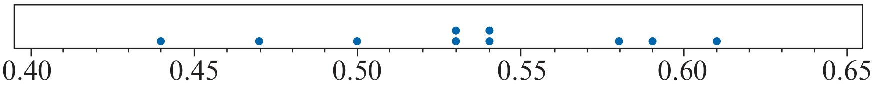
Suppose, for example, we know that 55% of people in Australia have blue eyes ($p = 0.55$) and that we are interested in the values of the sample proportion $\hat{p}$ which might be observed when samples of size 100 are drawn at random from the population.

If we select one sample of 100 people and find that 50 people have blue eyes, then the value of the sample proportion is $\hat{p} = \dfrac{50}{100} = 0.5$.

ISBN 978-1-009-11049-5 © Michael Evans et al 2023 Cambridge University Press

If a second sample of 100 people is selected and this time 58 people have blue eyes, then the value of the sample proportion for this second sample is $\hat{p} = \dfrac{58}{100} = 0.58$.

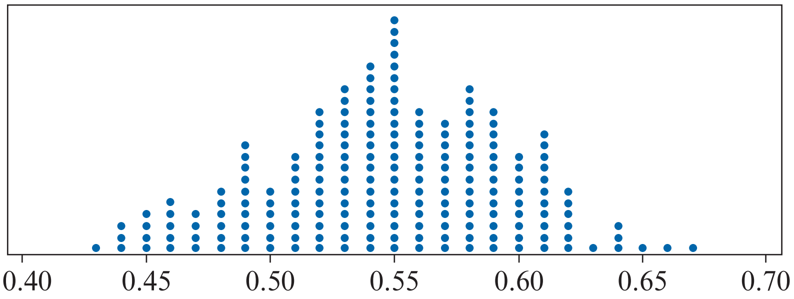
Continuing in this way, after selecting 10 samples, the values of $\hat{p}$ that are observed might look like those in the following dotplot:



It is clear that the proportion of people with blue eyes in the sample, $\hat{p}$, is varying from sample to sample: from as low as 0.44 to as high as 0.61 for these particular 10 samples.

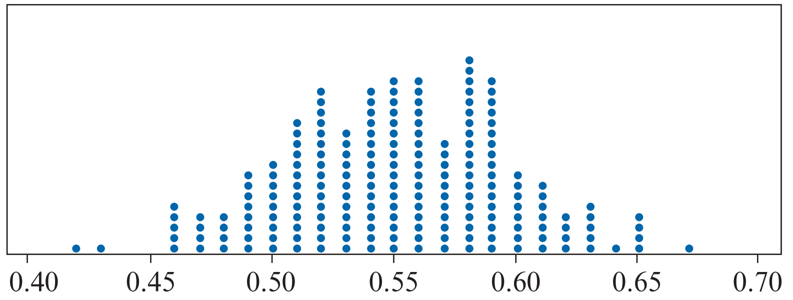
What does the distribution of the sample proportions look like if we continue with this sampling process?

The following dotplot summarises the values of $\hat{p}$ observed when 200 samples (each of size 100) were selected from a population in which the probability of having blue eyes is 0.55. We can see from the dotplot that the distribution is reasonably symmetric, centred at 0.55, and has values ranging from 0.43 to 0.67.



What does the distribution look like when another 200 samples (each of size 100) are selected at random from the same population?

The following dotplot shows the distribution obtained when this experiment was repeated. Again, the distribution is reasonably symmetric, centred at 0.55, and has values ranging from 0.42 to 0.67.



It seems reasonable to infer from these examples that, while there will be variation in the details of the distribution each time we take a collection of samples, the distribution of the values of $\hat{p}$ observed tends to conform to a predictable shape, centre and spread.

Actually, we already know from Chapter 16 that, when the sample size is large enough, the distribution of a binomial random variable is well approximated by the normal distribution. We have also seen that the rule of thumb for the normal approximation to the binomial distribution to apply is that both np and $n(1 - p)$ should be greater than 5.

The dotplots confirm the reasonableness of the normality assumption with regard to the sample proportion $\hat{P}$, which can be considered to be a linear function of a binomial random variable.

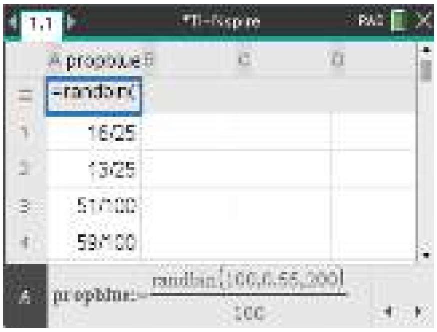
Repeated sampling can be investigated using a calculator.

Example 10

Assume that 55% of people in Australia have blue eyes. Use your calculator to illustrate a possible distribution of sample proportions $\hat{p}$ that may be obtained when 200 different samples (each of size 100) are selected from the population.

Using the TI-Nspire

- To generate the sample proportions:
  - Start from a **Lists & Spreadsheet** page.
  - Name the list ‘propblue’ in Column A.
  - In the formula cell of Column A, enter the formula using Menu > **Data** > **Random** > **Binomial** and complete as: $= \text{randbin}(100, 0.55, 200)/100$

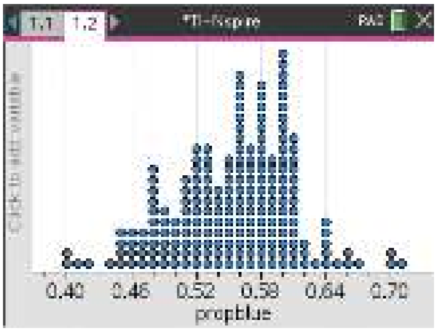


	A propblue	B	C	D
=	=randbin(			
1	16/25			
2	13/25			
3	51/100			
4	59/100			

A propblue:= randbin(100,0.55,200)/100

Note: The syntax is: randbin(*sample size*, *population proportion*, *number of samples*) To calculate as a proportion, divide by the sample size.

- To display the distribution of sample proportions:
  - Insert a **Data & Statistics** page (ctrl I or ctrl doc ▾).
  - Click on ‘Click to add variable’ on the horizontal axis and select ‘propblue’. A dotplot is displayed.



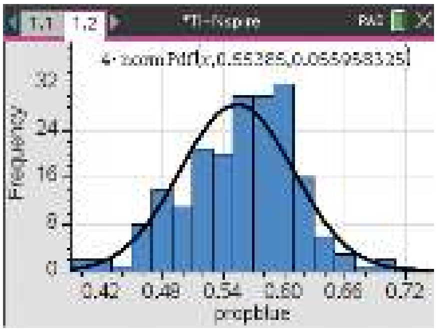
Note: You can recalculate the random sample proportions by using ctrl R while in the **Lists & Spreadsheet** page.

ISBN 978-1-009-11049-5

Cambridge University Press

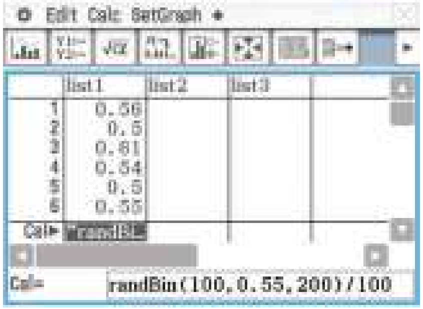
- To fit a normal curve to the distribution:
  - Menu > **Plot Type** > **Histogram**
  - Menu > **Analyze** > **Show Normal PDF**

Note: The calculated Normal PDF, based on the data set, is superimposed on the plot, showing the mean and standard deviation of the sample proportion.



Using the Casio ClassPad

- To generate the sample proportions:
  - Open the **Statistics** application.
  - Tap the 'Calculation' cell at the bottom of list1.
  - Type: randBin(100, 0.55, 200)/100
  - Tap EXE.

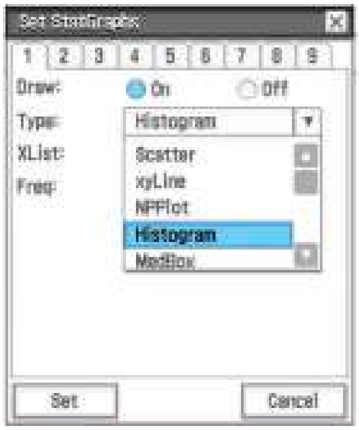


	list1	list2	list3
1	0.56		
2	0.5		
3	0.61		
4	0.54		
5	0.5		
6	0.55		
Cal▸	=randBi		

Cal= randBin(100,0.55,200)/100

Note: The syntax is: randBin(*sample size*, *population proportion*, *number of samples*)
To calculate as a proportion, divide by the sample size.

- To display the distribution of sample proportions:
  - Tap on the **Set StatGraphs** icon, select the type 'Histogram' and tap Set.
  - Tap on the graph icon in the toolbar.
  - In the **Set Interval** window, enter the values shown below and tap OK.

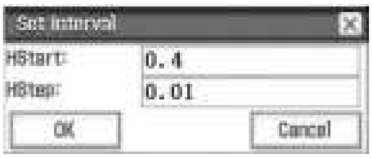


Draw: ● On ○ Off

Type: Histogram

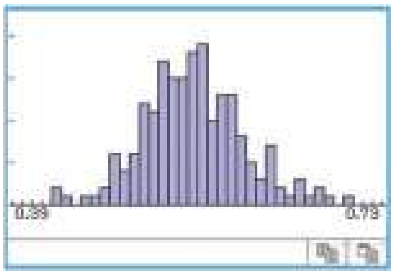
XList:

Freq:

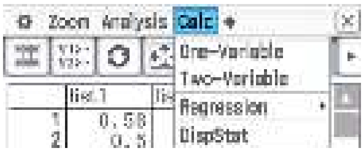


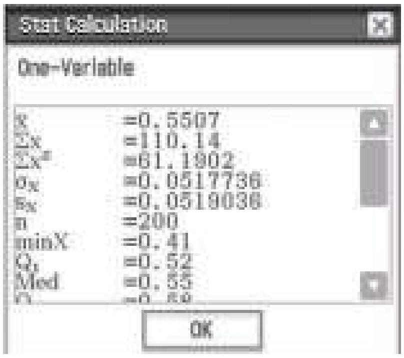
HStart: 0.4

HStep: 0.01



- To obtain statistics from the distribution, select **Calc** > **One-Variable**. Tap OK.





```
Stat Calculation
One-Variable
x̄      =0.5507
Σx     =110.14
Σx²    =61.1902
σx     =0.0517736
sx     =0.0519036
n      =200
minX   =0.41
Q₁     =0.52
Med    =0.55
              OK
```

Note: The mean of the sample proportions, $\bar{x}$, estimates the population proportion.

When the sample size n is large, the sample proportion $\hat{P}$ has an approximately normal distribution, with mean $\mu = p$ and standard deviation $\sigma = \sqrt{\dfrac{p(1-p)}{n}}$.

Thus, when samples of size $n = 100$ are selected from a population in which the proportion of people with blue eyes is $p = 0.55$, the distribution of the sample proportion $\hat{P}$ is approximately normal, with mean and standard deviation given by

$$\mu = \mathrm{E}(\hat{P}) = 0.55 \quad \text{and} \quad \sigma = \mathrm{sd}(\hat{P}) = \sqrt{\frac{0.55 \times 0.45}{100}} = 0.0497$$

Example 11

Assume that 60% of people have a driver's licence. Using the normal approximation, find the approximate probability that, in a randomly selected sample of size 200, more than 65% of people have a driver's licence.

Solution

Here $n = 200$ and $p = 0.6$. Since n is large, the distribution of $\hat{P}$ is approximately normal, with mean $\mu = p = 0.6$ and standard deviation

$$\sigma = \sqrt{\frac{0.6(1-0.6)}{200}} = 0.0346$$

Thus the probability that more than 65% of people in the sample have a driver's licence is

$$\Pr(\hat{P} > 0.65) = 0.0745 \quad \text{(correct to four decimal places)}$$

The use of a normal approximation allows to find approximate solutions to problems which could not be solved exactly using the binomial distribution, as shown in the following example.

Example 12

Suppose again that 60% of people have a driver's licence, and that random sample of size n is selected from the population. If the probability that the proportion of people in the sample with a drivers licence is less than 58% is equal to 0.3446, what size sample was chosen?

Solution

Here n is unknown, $p = 0.6$, and $\Pr(\hat{P} < 0.58) = 0.3446$. We will assume that n is large enough so that the distribution of the sample proportion $\hat{P}$ is approximately normal, with mean and standard deviation given by

$$\mu = \mathrm{E}(\hat{P}) = 0.60 \quad \text{and} \quad \sigma = \mathrm{sd}(\hat{P}) = \sqrt{\frac{0.6 \times 0.4}{n}} = \sqrt{\frac{0.24}{n}}$$

Thus $\Pr(\hat{P} < 0.58) = \Pr\left(Z < \dfrac{-0.02}{\sqrt{\frac{0.24}{n}}}\right) = 0.3446$

Using the inverse-normal facility of the calculator, enter 0.3446 and we find

$$\frac{-0.02}{\sqrt{\frac{0.24}{n}}} \approx -0.400$$

Solving this equation gives us $n = 96$

Summary 17C

When the sample size n is large, the sample proportion $\hat{P}$ has an approximately normal distribution, with mean $\mu = p$ and standard deviation $\sigma = \sqrt{\dfrac{p(1-p)}{n}}$.

Skill-sheet

Exercise 17C

In each of the following questions, use the normal approximation to the binomial distribution.

Example 11

1 Find the approximate probability that, in the next 50 tosses of a fair coin, the proportion of heads observed will be less than or equal to 0.46.

2 In a large city, 12% of the workforce are unemployed. If 300 people from the workforce are selected at random, find the approximate probability that more than 10% of the people surveyed are unemployed.

3 It is known that on average 50% of the children born at a particular hospital are female. Find the approximate probability that more than 60% of the next 25 children born at that hospital will be female.

4 A car manufacturer expects 10% of cars produced to require minor adjustments before they are certified as ready for sale. What is the approximate probability that more than 15% of the next 200 cars inspected will require minor adjustments?

5 Past records show that on average 30% of the workers at a particular company have had one or more accidents in the workplace. What is the approximate probability that less than 20% of a random sample of 50 workers have had one or more accidents?

6 Sacha is shooting at a target which she has a probability of 0.6 of hitting. What is the approximate probability that:

a the proportion of times she hits the target in her next 100 attempts is less than 0.8

b the proportion of times she hits the target in her next 100 attempts is between 0.6 and 0.8

c the proportion of times she hits the target in her next 100 attempts is between 0.7 and 0.8, given that it is more than 0.6?

7 **a** Find the approximate probability that, in the next 100 tosses of a fair coin, the proportion of heads will be between 0.4 and 0.6.

b If the probability that the proportion heads in the next n tosses is more than 0.55 is equal to 0.1, how many times was the coin tossed (give your answer to the nearest whole number?

8 A machine has a probability of 0.1 of producing a defective item.

a What is the approximate probability that, in the next batch of 1000 items produced, the proportion of defective items will be between 0.08 and 0.12?

b What is the approximate probability that, in the next batch of 1000 items produced, the proportion of defective items will be between 0.08 and 0.12, given that we know that it is greater than 0.10?

9 The proportion of voters in the population who favour Candidate A is 52%. Of a random sample of 400 voters, 230 indicated that they would vote for Candidate A at the next election.

a What is the value of the sample proportion, $\hat{p}$?

b Find the approximate probability that, in a random sample of 400 voters, the proportion who favour Candidate A is greater than or equal to this value of $\hat{p}$.

10 A manufacturer claims that 90% of their batteries will last more than 100 hours. Of a random sample of 250 batteries, 212 lasted more than 100 hours.

a What is the value of the sample proportion, $\hat{p}$?

b Find the approximate probability that, in a random sample of 250 batteries, the proportion lasting more than 100 hours is less than or equal to this value of $\hat{p}$.

c Does your answer to **b** cause you to doubt the manufacturer's claim?

Example 12

11 Suppose that in a certain community 35% of people have wavy hair. If in a sample of size n the probability that the proportion of people with wavy hair is less than 32% is equal to 0.2445, what size sample was chosen? Give n to the nearest whole number.

17D Confidence intervals for the population proportion

In practice, the reason we analyse samples is to further our understanding of the population from which they are drawn. That is, we know what is in the sample, and from that knowledge we would like to infer something about the population.

Point estimates

Suppose, for example, we wish to know the proportion of primary school children in Australia who regularly use social media. The value of the population proportion p is unknown. As already mentioned, collecting information about the whole population is generally not feasible, and so a random sample must suffice. What information can be obtained from a single sample? Certainly, the sample proportion $\hat{p}$ gives some indication of the value of the population proportion p, and can be used when we have no other information.

The value of the sample proportion $\hat{p}$ can be used to estimate the population proportion p. Since this is a single-valued estimate, it is called a **point estimate** of p.

Thus, if we select a random sample of 200 Australian primary school children and find that the proportion who use social media is 0.7, then the value $\hat{p} = 0.7$ serves as an estimate of the unknown population proportion p.

Interval estimates

The value of the sample proportion $\hat{p}$ obtained from a single sample is going to change from sample to sample, and while sometimes the value will be close to the population proportion p, at other times it will not. To use a single value to estimate p can be rather risky. What is required is an interval that we are reasonably sure contains the parameter value p.

An **interval estimate** for the population proportion p is called a **confidence interval** for p.

We have already seen that, when the sample size n is large, the sample proportion $\hat{P}$ has an approximately normal distribution with $\mu = p$ and $\sigma = \sqrt{\dfrac{p(1-p)}{n}}$.

By standardising, we can say that the distribution of the random variable

$$\frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximated by that of a standard normal random variable Z.

Consider now the values of c_1 and c_2 such that

$$\Pr(c_1 < Z < c_2) = 0.95$$

Since we would like the confidence interval to be symmetric around the value of $\hat{p}$, we can determine the value of c_2 by finding the inverse normal of 0.975 on the calculator. This gives 1.9600, correct to four decimal places. By symmetry, $c_1 = -1.9600$.

Note that we should use enough decimal places determining values using the inverse normal command so as to be able to determine the endpoints of the confidence interval to the required number of decimal places.

Thus $\Pr(-1.9600 < Z < 1.9600) = 0.95$ and therefore

$$\Pr\left(-1.9600 < \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} < 1.9600\right) \approx 0.95$$

Multiplying through gives

$$\Pr\left(-1.9600\sqrt{\frac{p(1-p)}{n}} < \hat{P} - p < 1.9600\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95$$

Further simplifying, we obtain

$$\Pr\left(\hat{P} - 1.9600\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + 1.9600\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95$$

Remember that what we want to do is to use the value of the sample proportion $\hat{p}$ obtained from a single sample to calculate an interval that we are fairly certain (say 95% certain) contains the true population proportion p (which we do not know).

In order to do this, we need to make one further approximation, and substitute $\hat{p}$ for p in our estimate of the standard deviation σ of $\hat{P}$, so that an approximate **95% confidence interval** for p is given by

$$\left(\hat{p} - 1.9600\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1.9600\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Note: In order to use this rule to calculate a confidence interval, the criteria for the normal approximation to the binomial distribution must apply. Therefore, from Chapter 16, we require both np and $n(1-p)$ to be greater than 5.

A confidence interval with level of confidence **other than** 95% can be found using the same principles. For example, since we know that

$$\Pr(-1.6449 < Z < 1.6449) = 0.90$$

then an approximate **90% confidence interval** for p is given by

$$\left(\hat{p} - 1.6449\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1.6449\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

We can express this is the following general rule.

In general, a $C\%$ confidence interval is given by

$$\left(\hat{p} - k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

where k is such that $\Pr(-k < Z < k) = \dfrac{C}{100}$, and:

- p is the population proportion (unknown)
- $\hat{p}$ is a value of the sample proportion
- n is the size of the sample from which $\hat{p}$ was calculated.

Example 13

Calculate and compare 90%, 95% and 99% confidence intervals for the proportion p of primary school children in Australia who regularly use social media, if we select a random sample of 200 children and find the sample proportion $\hat{p}$ to be 0.7.

Solution

The 90% confidence interval is

$$\left(0.7 - 1.6449\sqrt{\frac{0.7 \times 0.3}{200}},\ 0.7 + 1.6449\sqrt{\frac{0.7 \times 0.3}{200}}\right) = (0.647, 0.753)$$

The 95% confidence interval is

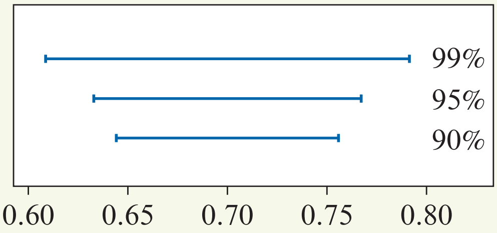
$$\left(0.7 - 1.9600\sqrt{\frac{0.7 \times 0.3}{200}},\ 0.7 + 1.9600\sqrt{\frac{0.7 \times 0.3}{200}}\right) = (0.636, 0.764)$$

The 99% confidence interval is

$$\left(0.7 - 2.5758\sqrt{\frac{0.7 \times 0.3}{200}},\ 0.7 + 2.5758\sqrt{\frac{0.7 \times 0.3}{200}}\right) = (0.617, 0.783)$$

It is helpful to use a diagram to compare these confidence intervals:

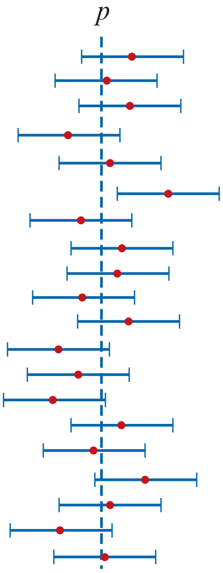
From the diagram, it can be clearly seen that the effect of being more confident means that a wider interval is required.



Interpretation of a confidence interval

The confidence interval found in Example 13 should not be interpreted as meaning that $\Pr(0.636 < p < 0.764) = 0.95$. In fact, such a statement is meaningless, as p is a constant and either does or does not lie in the stated interval.

The particular confidence interval found is just one of any number of confidence intervals which could be found for the population proportion p, each one depending on the particular value of the sample proportion $\hat{p}$. The correct interpretation of the 95% confidence interval, for example, is that we expect approximately 95% of such intervals to contain the population proportion p. Whether or not the particular confidence interval obtained contains the population proportion p is generally not known.



If we were to repeat the process of taking a sample and calculating a 95% confidence interval many times, the result would be something like that indicated in the diagram.

The diagram shows the confidence intervals obtained when 20 different samples were drawn from the same population. The round dot indicates the value of the sample estimate in each case. The 95% confidence intervals vary, because the samples themselves vary. The value of the population proportion p is indicated by the vertical line, and it is of course constant.

It is quite easy to see from the diagram that none of the values of the sample estimate is exactly the same as the population proportion, but that all the intervals except one (19 out of 20, or 95%) have captured the value of the population proportion, as would be expected in the case of a 95% confidence interval.

Simialrly, we would expect that in the long run that 90% of 90% confidence intervals, and 99% of 99% confidence intervals, will capture the true value of the population proportion.

Using a calculator to determine confidence intervals

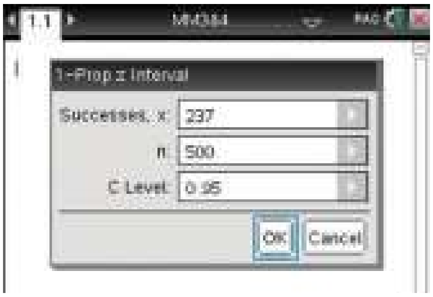
Example 14

A survey found that 237 out of 500 undergraduate university students questioned intended to take a postgraduate course in the future. Find a 95% confidence interval for the proportion of undergraduates intending to take a postgraduate course.

Using the TI-Nspire

In a **Calculator** page:

- Use (Menu) > **Statistics** > **Confidence Intervals** > **1–Prop z Interval**.
- Enter the values $x = 237$ and $n = 500$ as shown.


1-Prop z Interval

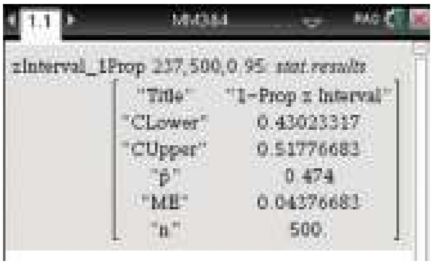
Successes, x: 237

n: 500

C Level: 0.95

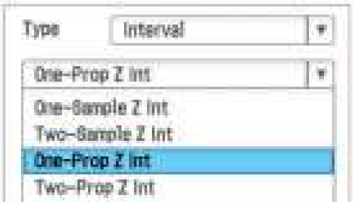
- The ‘CLower’ and ‘CUpper’ values give the 95% confidence interval $(0.43, 0.52)$.

Note: ‘ME’ stands for margin of error, which is covered in the next subsection.


```
zInterval_1Prop 237,500,0.95: stat.results
  "Title"    "1-Prop z Interval"
  "CLower"    0.43023317
  "CUpper"    0.51776683
  "p̂"         0.474
  "ME"        0.04376683
  "n"         500
```

Using the Casio ClassPad

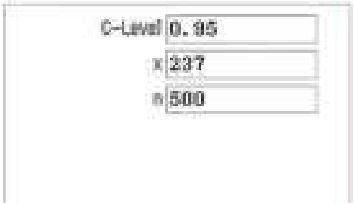
- In , go to **Calc** > **Interval**.
- Select **One–Prop Z Int** and tap Next.
- Enter the values C-Level = 0.95, $x = 237$ and $n = 500$ as shown below. Tap Next.



Type Interval

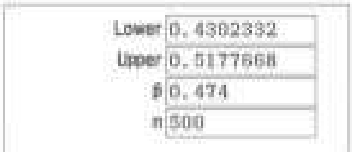
One-Prop Z Int

One-Sample Z Int
Two-Sample Z Int
One-Prop Z Int
Two-Prop Z Int



C-Level 0.95
x 237
n 500

- The ‘Lower’ and ‘Upper’ values give the 95% confidence interval (0.43, 0.52).



Lower 0.4302332
Upper 0.5177668
$\hat{p}$ 0.474
n 500

Precision and margin of error

Often we discuss the confidence interval in terms of its width or, more formally, in terms of the distance between the sample estimate and the endpoints of the confidence interval.

That is, we find it useful to make statements such as ‘we predict the proportion of people who will vote Labor in the next election as 52% ± 2%’. Here the sample estimate is 52%, and the distance between the sample estimate and the endpoints is 2%.

The distance between the sample estimate and the endpoints of the confidence interval is called the **margin of error** (M).

- For a 90% confidence interval,

$$M = 1.6449\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- For a 95% confidence interval,

$$M = 1.9600\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- For a 99% confidence interval,

$$M = 2.5758\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

We can see from this rule that the margin of error is a function of both the level of confidence, and the sample size n. Thus, one way to make the interval narrower (that is, to increase the precision of the estimate) without changing the level of confidence is to increase the sample size.

Example 15

Determine the sample size required to achieve a margin of error of 2% in an approximate 95% confidence interval for the proportion p of primary school children in Australia who use social media, if the sample proportion $\hat{p}$ is found to be 0.7.

Solution

Substituting $M = 0.02$ and $\hat{p} = 0.7$ in the expression for the margin of error gives

$$0.02 = 1.9600\sqrt{\frac{0.7 \times 0.3}{n}}$$

Solving for n:

$$\left(\frac{0.02}{1.9600}\right)^2 = \frac{0.7 \times 0.3}{n}$$

$$\therefore \quad n = 0.7 \times 0.3 \times \left(\frac{1.9600}{0.02}\right)^2 \approx 2016.84$$

Thus, to achieve a margin of error of 2%, we need a sample of size 2017.

Of course, it is highly unlikely that we will know the value of the sample proportion $\hat{p}$ before we have selected the sample. Thus it is usual to substitute an estimated value into the equation in order to determine the sample size before we select the sample. This estimate can be based on our prior knowledge of the population or on a pilot study. If we denote this estimated value for the sample proportion by p^*, we can write the margin of error for a 95% confidence interval as

$$M = 1.9600\sqrt{\frac{p^*(1 - p^*)}{n}}$$

Rearranging to make n the subject of the equation, we find

$$M^2 = 1.9600^2\left(\frac{p^*(1 - p^*)}{n}\right) \text{ and therefore } n = \left(\frac{1.9600}{M}\right)^2 p^*(1 - p^*)$$

We can find similar expressions for n for any level of confidence.

If p^* is an estimated value for the population proportion p, then

- A 90% confidence interval for a population proportion p will have margin of error approximately equal to a specified value of M when the sample size is
$$n = \left(\frac{1.6449}{M}\right)^2 p^*(1 - p^*)$$
- A 95% confidence interval for a population proportion p will have margin of error approximately equal to a specified value of M when the sample size is
$$n = \left(\frac{1.9600}{M}\right)^2 p^*(1 - p^*)$$
where p^* is an estimated value for the population proportion p.

- A 99% confidence interval for a population proportion p will have margin of error approximately equal to a specified value of M when the sample size is

$$n = \left(\frac{2.5758}{M}\right)^2 p^*(1 - p^*)$$

Summary 17D

- The value of the sample proportion $\hat{p}$ can be used to estimate the population proportion p. Since this is a single-valued estimate, it is called a **point estimate** of p.
- An **interval estimate** for the population proportion p is called a **confidence interval** for p.
- In general, an approximate $C\%$ confidence interval is given by

$$\left(\hat{p} - k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

where k is such that $\Pr(-k < Z < k) = \dfrac{C}{100}$, and:

- p is the population proportion (unknown)
- $\hat{p}$ is a value of the sample proportion
- n is the size of the sample from which $\hat{p}$ was calculated.

- The distance between the sample estimate and the endpoints of the confidence interval is called the **margin of error** (M) and, for a C% confidence interval,

$$M = k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- A 95% confidence interval for a population proportion p will have margin of error approximately equal to a specified value of M when the sample size is

$$n = \left(\frac{k}{M}\right)^2 p^*(1 - p^*)$$

where p^* is an estimated value for the population proportion p.

Skill-sheet

Exercise 17D

Example 13

1 A quality-control engineer in a factory needs to estimate the proportion of bags of potato chips packed by a certain machine that are underweight. The engineer takes a random sample of 100 bags and finds that eight of them are underweight.

a Find a point estimate for p, the proportion of bags packed by the machine that are underweight.

b Calculate and compare 90%, 95% and 99% confidence intervals for p.

2 A newspaper wants to estimate the proportion of its subscribers who believe that the government should be allowed to tap telephones without a court order. It selects a random sample of 250 subscribers, and finds that 48 of them believe that the government should have this power.

a Find a point estimate for p, the proportion of subscribers who believe that the government should be allowed to tap telephones without a court order.

b Calculate and compare 90%, 95% and 99% confidence intervals for p.

3 The lengths of stay in hospital among patients is of interest to health planners. A random sample of 100 patients was investigated, and 20 were found to have stayed longer than 7 days. If p is the proportion of patients who stay in hospital longer than 7 days.

a Find a point estimate for p.

b Calculate a 98% confidence interval for p.

Example 14

4 Given that 132 out of 400 randomly selected adult males are cigarette smokers, find a 92% confidence interval for the proportion of adult males in the population who smoke.

5 Of a random sample of 400 voters in a particular electorate, 210 indicated that they would vote for the Labor party at the next election.

a Use this information to find a 95% confidence interval for the proportion of Labor voters in the electorate.

b A random sample of 4000 voters from the same electorate was taken, and this time 2100 indicated that they would vote for Labor at the next election. Find a 95% confidence interval for the proportion of Labor voters in the electorate.

c Compare your answers to parts **a** and **b**.

6 A manufacturer claims that 90% of their batteries will last more than 50 hours.

a Of a random sample of 250 batteries, 212 lasted more than 50 hours. Use this information to find a 99% confidence interval for the proportion of batteries lasting more than 50 hours.

b An inspector requested further information. A random sample of 2500 batteries was selected and this time 2120 lasted more than 50 hours. Use this information to find a 99% confidence interval for the proportion of batteries lasting more than 50 hours.

c Compare your answers to parts **a** and **b**.

7 When a coin thought to be biased was tossed 100 times, it came up heads 60 times. Calculate and compare 90%, 95% and 99% confidence intervals for the probability of observing a head when that coin is tossed.

8 In a survey of attitudes to climate change, a total of 537 people from a random sample of 1000 people answered no to the question ‘Do you think the government is doing enough to address global warming?’ Calculate and compare 90%, 95% and 99% confidence intervals for the proportion of people in Australia who would answer no to that question.

Example 15

9 Determine the size of sample required to achieve a margin of error of 2% in an approximate 95% confidence interval when the sample proportion $\hat{p}$ is 0.8.

10 Determine the size of sample required to achieve a margin of error of 5% in an approximate 90% confidence interval when the sample proportion $\hat{p}$ is 0.2.

11 Samar is conducting a survey to estimate the proportion of people in Victoria who would support reducing the driving age to 16. He knows from previous studies that this proportion is about 30%.

a Determine the size of sample required for the survey to achieve a margin of error of 3% in an approximate 99% confidence interval for this proportion.

b Determine the size of sample required for the survey to achieve a margin of error of 2% in an approximate 99% confidence interval for this proportion.

c Compare your answers to parts **a** and **b**.

12 Bob is thinking of expanding his pizza delivery business to include a range of desserts. He would like to know the proportion of his clients who would order dessert from him, and so he intends to ask a number of his clients what they think.

a Bob thinks that the proportion of his clients who would order dessert is around 0.3. Determine the size of sample required for Bob to achieve a margin of error of 2% in an approximate 95% confidence interval for this proportion.

b Bob's business partner Phil thinks that the proportion of clients who would order dessert is around 0.5. Determine the size of sample required to achieve a margin of error of 2% in an approximate 95% confidence interval for this proportion.

c What is the effect on the margin of error if:

i Bob is correct, but they use the sample size from Phil's estimate

ii Phil is correct, but they use the sample size from Bob's estimate?

d What sample size would you recommend that Bob and Phil use?

Chapter summary

Assign-ment

Nrich

- A **population** is the set of all eligible members of a group which we intend to study.
- A **sample** is a subset of the population which we select in order to make inferences about the population. Generalising from the sample to the population will not be useful unless the sample is representative of the population.
- A sample of size n is called a **simple random sample** if it is selected from the population in such a way that every subset of size n has an equal chance of being chosen as the sample. In particular, every member of the population must have an equal chance of being included in the sample.
- The **population proportion** p is the proportion of individuals in the entire population possessing a particular attribute, and is constant.
- The **sample proportion** $\hat{p}$ is the proportion of individuals in a particular sample possessing the attribute, and varies from sample to sample.
- The sample proportion $\hat{P} = \dfrac{X}{n}$ is a random variable, where X is the number of favourable outcomes in a sample of size n. The distribution of the random variable $\hat{P}$ is known as the **sampling distribution** of the sample proportion.
- When the population is *small*, the sampling distribution of the sample proportion $\hat{P}$ can be determined using our knowledge of selections.
- When the population is *large*, the sampling distribution of the sample proportion $\hat{P}$ can be determined by assuming that X is a binomial random variable with parameters n and p. In this case, the mean and standard deviation of $\hat{P}$ are given by

$$\mathrm{E}(\hat{P}) = p \quad \text{and} \quad \mathrm{sd}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$$

- When the sample size n is large, the sample proportion $\hat{P}$ has an approximately normal distribution, with mean $\mu = p$ and standard deviation $\sigma = \sqrt{\dfrac{p(1-p)}{n}}$.
- If the value of the sample proportion $\hat{p}$ is used as an estimate of the population proportion p, then it is called a **point estimate** of p.
- In general, an approximate ***C*% confidence interval** is given by

$$\left(\hat{p} - k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

where k is such that $\Pr(-k < Z < k) = \dfrac{C}{100}$, and:
  - p is the population proportion (unknown)
  - $\hat{p}$ is a value of the sample proportion
  - n is the size of the sample from which $\hat{p}$ was calculated.
- The distance between the sample estimate and the endpoints of the confidence interval is called the **margin of error** (M) and, for a C% confidence interval,

$$M = k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

ISBN 978-1-009-11049-5 Cambridge University Press

- A 95% confidence interval for a population proportion p will have margin of error approximately equal to a specified value of M when the sample size is

$$n = \left(\frac{k}{M}\right)^2 p^*(1 - p^*)$$

where p^* is an estimated value for the population proportion p.

Technology-free questions

1 A company has 2000 employees, 700 of whom are female. A random sample of 100 employees was selected, and 40 of them were female. In this example:

a What is the population?

b What is the value of the population proportion p?

c What is the value of the sample proportion $\hat{p}$?

2 To study the effectiveness of yoga for reducing stress levels, a researcher measured the stress levels of 50 people who had just enrolled in a 10-week introductory yoga course, and then measured their stress levels at the end the course.

a Do you think that this sample will be representative of the general population? Explain your answer.

b How would you suggest that the sample could be chosen?

3 Consider a bag containing three blue and two red balls.

a What is p, the proportion of blue balls in the bag?

b Samples of size 3 are taken from the bag without replacement. If $\hat{P}$ is a random variable describing the possible values of the sample proportion $\hat{p}$ of blue balls in the sample, list the possible values that $\hat{P}$ can take.

c Find $\Pr\left(\hat{P} = \frac{1}{3}\right)$.

4 In a large population the proportion of blonde haired people is $\frac{1}{8}$. Let $\hat{P}$ be the random variable that represents that the sample proportion of blonde haired people in a sample of size n. Find the smallest integer value of n such that the standard deviation of $\hat{P}$ is less than or equal to $\frac{1}{80}$.

5 A coin is tossed 100 times, and k heads observed.

a Give a point estimate for p, the probability of observing a head when the coin is tossed.

b Write down an expression for a 95% confidence interval for p.

6 A sample of n people were asked whether they thought that income tax in Australia was too high, and 90% said yes.

a What is the value of the sample proportion $\hat{p}$?

b Write down an expression for M, the margin of error for this estimate at the 95% confidence level, in terms of n.

c If the number of people in the sample were doubled, what would be the effect on the margin of error M?

7 Let $\hat{P}$ be the random variable that represents the sample proportion of customers who pay for their shopping using cash. From a sample of size n, an approximate 95% confidence interval for the p, the population proportion of customers who pay for their shopping with cash was determined to be $\left(\dfrac{576}{1250}, \dfrac{674}{1250}\right)$.

a Find the value of $\hat{p}$ which was used to obtain this confidence interval.

b Using the fact that $1.96 = \dfrac{49}{25}$, find the size of the sample from which this 95% confidence interval was obtained.

8 Suppose that 40 independent random samples were taken from a large population, and that a 95% confidence interval for the population proportion p was computed from each of these samples.

a How many of the 95% confidence intervals would you expect to contain the population proportion p?

b Write down an expression for the probability that all 40 confidence intervals contain the population proportion p.

9 Suppose that 50 independent random samples were taken from a large population, and that a 90% confidence interval for the population proportion p was computed from each of these samples.

a How many of the 90% confidence intervals would you expect to contain the population proportion p?

b Write down an expression for the probability that at least 49 of the 50 confidence intervals contain the population proportion p.

10 A newspaper determined that an approximate 95% confidence interval for the proportion of people in Australia who regularly read the news online was $(0.50, 0.70)$.

a What was the value of $\hat{p}$ which was used to determine this confidence interval?

b What is the margin of error?

c How could the newspaper increase the precision of their study?

Multiple-choice questions

1 In order to estimate the ratio of males to females at a school, a teacher determines the number of males and the number of females in one class chosen at random. The ratio that he then calculates is called a

A sample **B** sample statistic **C** population parameter

D population **E** sample parameter

2 In a complete census of the population of a particular community, it is found that 59% of families have two or more children. Here ‘59%’ represents the value of a

A sample **B** sample statistic **C** population parameter
D population **E** sample parameter

3 From a random sample, a 95% confidence interval for the population proportion p is found to be $(0.7, 0.8)$. Which of the following statements is correct?

A the population proportion $p = 0.6$

B the probability that the population proportion p lies in the interval $(0.7, 0.8)$ is 0.95

C the probability that the population proportion p lies in the interval $(0.7, 0.8)$ is 0.05

D There is a probability of $\frac{1}{20}$ that a 95% of confidence intervals will capture the population proportion p

E more than one of these statements is correct

4 A survey showed that 15 out of a random sample of 50 football supporters attend at least one match per season. If this information is used to find a 95% confidence interval for the proportion of all football supporters who attend at least one match per season, then the margin of error will be

A 0.3 **B** 0.004 **C** 0.065 **D** 0.254 **E** 0.127

5 In a certain country it is known that 15% of golfers play left-handed. A random sample of 20 golfers it to be selected. If $\hat{P}$ is the proportion of golfers in the sample who play left handed, then (do not use a normal approximation)

$$\Pr\left(\hat{P} \geq \frac{3}{10}\right)$$

is closest to

A 0.9780 **B** 0.9326 **C** 0.0673
D 0.0219 **E** 0.3920

6 A 95% confidence interval for the proportion of people in the population who prefer to watch the news on a certain channel is given by (0.084, 0.236). The sample proportion from which this interval was constructed is

A 0.152 **B** 0.320 **C** 0.244
D 0.076 **E** 0.160

7 If the sample proportion remains unchanged, then an increase in the level of confidence will lead to a confidence interval which is

A narrower **B** wider **C** unchanged **D** asymmetric
E cannot be determined from the information given

8 Which of the following statements is true?

I The centre of a confidence interval is a population parameter.

II The bigger the margin of error, the smaller the confidence interval.

III The confidence interval is a type of point estimate.

IV A population proportion is an example of a point estimate.

A I only B II only C III only D IV only E none of these

9 If a researcher increases her sample size by a factor of 4, then the width of a 95% confidence interval would

A increase by a factor of 2 B increase by a factor of 4 C decrease by a factor of 2
D decrease by a factor of 4 E none of these

10 The Education Department in a certain state wishes to determine the percentage of teachers who are considering leaving the profession in the next two years. They believe it to be about 25%. How large a sample should be taken to find the answer to within ±3% at the 95% confidence level?

A 6 B 33 C 534 D 752 E 897

11 Which of the following statements is true?

A We use sample statistics to estimate population parameters.

B We use sample parameters to estimate population statistics.

C We use population parameters to estimate sample statistics.

D We use population statistics to estimate sample parameters.

E none of the above

12 A sampling distribution can best be described as a distribution which

A gives the possible range of values of the sample statistic

B describes how a statistic's value will change from sample to sample

C describes how samples do not give reliable estimates

D gives the distribution of the values observed in particular sample

E none of the above

13 There are 10,000 bricks in a large container, some of which are red and the rest are grey. There are more red bricks in the container than grey. Random samples of 100 bricks are selected from the container, with replacement. If $\hat{P}$ is the random variable describing the proportion or red bricks in the sample, and the standard deviation $\hat{P}$ is 0.04, then the number of red bricks in the container is

A 6000 B 7000 C 8000
D 8500 E 9000

14 How could you decrease the width of the confidence interval?

A increase the sample size

B use a smaller confidence level

C use a higher confidence level

D both A and B are correct

E both A and C are correct

Extended-response questions

1 A survey is being planned to estimate the proportion of people in Australia who think that university fees should be abolished. The organisers of the survey want the error in the approximate 95% confidence interval for this proportion to be no more than ±2%. They have no prior information about the value of the proportion.

a Plot that sample size, $n = \left(\dfrac{1.96}{M}\right)^2 p^*(1 - p^*)$, against p^* for $0 \leq p^* \leq 1$.

b For what value of p^* is the sample size the maximum?

c What value of n would you recommend be used for the survey?

d Show that the maximum sample size required for the error in an approximate 95% confidence interval to be no more than M is approximately $n = \dfrac{1}{M^2}$.

2 It is known that 70% of the students in a particular state study mathematics at Year 12. Let $\hat{P}$ be the random variable for the sampling distribution of the sample proportion when a sample of n Year 12 students is selected from that state.

a Suppose $n = 100$. Without using the normal approximation find:

i $\Pr(\hat{P} > 0.75)$

ii $\Pr(0.68 \leq \hat{P} < 0.75 | \hat{P} > 0.68)$

b If the the probability that the proportion of Year 12 students in the sample who study mathematics is less than 66% is equal to 0.0228, what size sample was chosen? Use the normal approximation, and give you answer to the nearest whole number.

3 **a** Summer is investigating the probability that a drawing pin will land point-up when tossed. She tosses the drawing pin 100 times, and finds that it lands point-up 57 times. Determine an approximate 95% confidence interval for the probability that the drawing pin lands point-up when tossed.

b Four of Summer's friends decide to repeat her investigation, each tossing the drawing pin 100 times. They each calculate an approximate 95% confidence interval based on their own data, making five confidence intervals in all.

i What is the probability that all five confidence intervals contain the true value of p, the probability that the drawing pin will land point-up when tossed?

ii What is the probability that none of the confidence intervals contain p?

iii What is the probability that at least one of the confidence intervals does not contain p?

iv How many of these five confidence intervals would you expect to contain p?

c Summer's four friends obtained the following results, each based on tossing the drawing pin 100 times and counting the number of times that it lands point-up:

- Emma 67
- Chloe 72
- Maddie 55
- Regan 60

Summer suggests that the best estimate of p would be obtained by pooling their results. Based on all the data collected, determine an approximate 95% confidence interval for p.

4 A landscape gardener wishes to estimate how many carp live in his very large ornamental lake. He is advised that the best way to do this is through capture–recapture sampling.

a Suppose that there are N carp in the lake and he captures 500 of them, tags them and then releases them back into the lake. Write down an expression for the proportion of tagged carp in the lake.

b The next day, a sample of 400 carp is captured from the lake, and he finds that there are 60 tagged carp in this sample. What is the proportion of tagged carp in the second sample?

c If the second sample is representative of the population, we expect the proportion of tagged carp in the second sample to be the same as the proportion of tagged carp in the lake. That is,

$$\frac{60}{400} \approx \frac{500}{N}$$

Use this equation to find an estimate for the number of carp in the lake.

d Show that an expression for a 95% confidence interval for the proportion of tagged carp in the lake can be written as

$$0.15 - 1.96\sqrt{\frac{0.1275}{400}} < \frac{500}{N} < 0.15 + 1.96\sqrt{\frac{0.1275}{400}}$$

e Use this inequality to find an approximate 95% confidence interval for the number of carp in the lake.