13 Sampling and confidence intervals

13.1 Overview

Many of the statistics about life in Australia are drawn from the Census of Population and Housing conducted by the Australian Bureau of Statistics every five years. A census is simply a survey of the population of a country. It collects general information such as gender, age and occupation as well as more detailed data such as ethnicity and housing situation. The most recent Australian census revealed many interesting facts about Queensland. Here's a few examples:

- Of the 4.9 million people living in the state, almost half live within the greater Brisbane area.
- The top five languages (other than English) spoken in Queensland are Mandarin, Filipino/Tagalog, Vietnamese, Cantonese and Spanish.
- Aboriginal and Torres Strait Islander People account for 4.0 percent of Queensland's population.
- The top five countries of birth (other than Australia) for people in Queensland are England, New Zealand, India, China and South Africa.

It is quite impractical and sometimes impossible to analyse the entire population for a particular situation, so a sample is usually taken. The sample needs to be representative of the population, and multiple samples should be taken where possible. The study of statistics allows us to gauge the confidence we can have in the validity of our samples for an investigation before we make any predictions or assumptions.

LEARNING SEQUENCE

13.1 Overview
13.2 Sample statistics
13.3 The distribution of $\hat{p}$
13.4 Confidence intervals
13.5 Review: exam practice

Fully worked solutions for this chapter are available in the Resources section of your eBookPLUS at www.jacplus.com.au.

13.2 Sample statistics

Suppose you were interested in the percentage of Year 12 graduates who plan to study Mathematics once they complete school. It is probably not practical to question every student. There must be a way that we can ask a smaller group and then use this information to make generalisations about the whole group.

13.2.1 Samples and populations

A **population** is a group that you want to know something about, and a **sample** is the group within the population that you collect the information from. Normally, a sample is smaller than the population; the exception is a census, where the whole population is the sample.

The number of members in a sample is called the **sample size** (symbol n), and the number of members of a population is called the **population size** (symbol N). Sometimes the population size is unknown.

WORKED EXAMPLE 1

Cameron has uploaded a popular YouTube video. He thinks that the 133 people in his year group at school have seen it, and he wants to know what they think. He decides to question 10 people. Identify the population and sample size.

THINK	WRITE
1. Cameron wants to know what the people in his year at school think. This is the population.	$N = 133$
2. He asks 10 people. This is the sample.	$n = 10$

WORKED EXAMPLE 2

A total of 137 people volunteer to take part in a medical trial. Of these, 57 are identified as suitable candidates and are given the medication. Identify the population and sample size.

THINK	WRITE
1. 57 people are given the medication. This is the sample size.	$n = 57$
2. We are interested in the group of people who might receive the drug in the future. This is the population.	The population is unknown, as we don't know how many people may be given this drug in the future.

on Resources

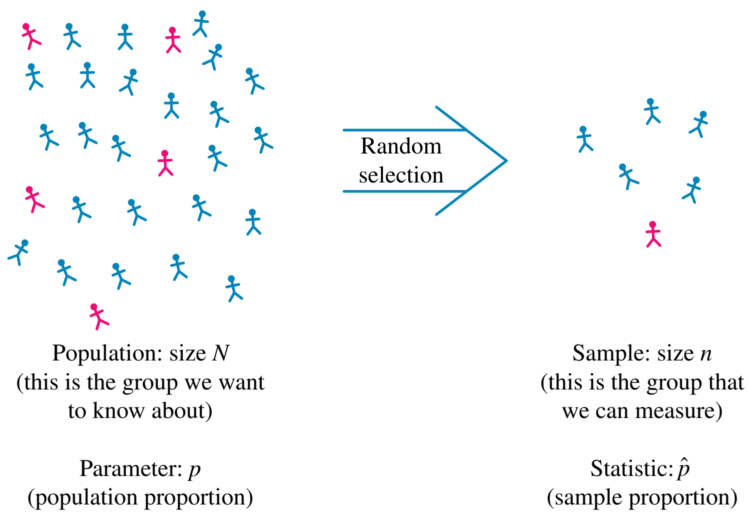
Interactivity: Population parameters and sample statistics (int-6442)

13.2.2 Statistics and parameters

A **parameter** is a characteristic of a population, whereas a **statistic** is a characteristic of a sample. This means that a statistic is always known exactly (because it is measured from the sample that has been selected). A parameter is usually estimated from a sample statistic. (The exception is if the sample is a census, in which case the parameter is known exactly.)

In this unit, we will study binomial data (that means that each data point is either yes/no or success/failure) with special regard to the proportion of successes.

The relationship between populations and samples



WORKED EXAMPLE 3

Identify the following as either sample statistics or population parameters.

a. **Forty-three per cent of voters polled say that they are in favour of banning fast food.**

b. **According to Australian Bureau of Statistics census data, the average family has 1.7 children.**

c. **Between 18% and 23% of Australians skip breakfast regularly.**

d. **Nine out of 10 children prefer cereal for breakfast.**

THINK	WRITE
a. 43% is an exact value that summarises the sample asked.	**a.** Sample statistic
b. The information comes from census data. The census questions the entire population.	**b.** Population parameter

c. 18%–23% is an estimate about the population.

d. Nine out of 10 is an exact value. It is unlikely that all children could have been asked; therefore, it is from a sample.

c. Population parameter

d. Sample statistic

13.2.3 Random samples

A good sample should be representative of the population. If we consider our initial interest in the proportion of Year 12 graduates who intend to study Mathematics once they finished school, we could use a Mathematical Methods class as a sample. This would not be a good sample because it does not represent the population — it is a very specific group.

In a **random sample**, every member of the population has the same probability of being selected. The Mathematical Methods class is not a random sample because students who don't study Mathematical Methods have no chance of being selected; furthermore, students who don't attend that particular school have no chance of being selected.

A **systematic sample** is almost as good as a random sample. In a systematic sample, every kth member of the population is sampled. For example, if $k = 20$, a customs official might choose to sample every 20th person who passes through the arrivals gate. The reason that this is almost as good as random sample is that there is an assumption that the group passing the checkpoint during the time the sample is taken is representative of the population. This assumption may not always be true; for example, people flying for business may be more likely to arrive on an early morning flight. Depending on the information you are collecting, this may influence the quality of the data.

In a **stratified random sample**, care is taken so that subgroups within a population are represented in a similar proportion in the sample. For example, if you were collecting information about students in Years 9–12 in your school, the proportions of students in each year group should be the same in the sample and the population. Within each subgroup, each member has the same chance of being selected.

A **self-selected sample**, that is one where the participants choose to participate in the survey, is almost never representative of the population. For example, television phone polls, where people phone in to answer yes or no to a question, do not accurately reflect the opinion of the population.

WORKED EXAMPLE 4

A survey is to be conducted in a middle school that has the distribution detailed in the table. It is believed that students in different year levels may respond differently, so the sample chosen should reflect the subgroups in the population (that is, it should be a stratified random sample). If a sample of 100 students is required, determine how many from each year group should be selected.

Year level	Number of students
7	174
8	123
9	147

THINK	WRITE
1. Calculate the total population size.	Total population $= 174 + 123 + 147 = 444$
2. Calculate the number of Year 7s to be surveyed.	Number of Year 7s $= \frac{174}{444} \times 100$ $= 39.1$ Survey 39 Year 7s.
3. Calculate the number of Year 8s to be surveyed.	Number of Year 8s $= \frac{123}{444} \times 100$ $= 27.7$ Survey 28 Year 8s.
4. Calculate the number of Year 9s to be surveyed.	Number of Year 9s $= \frac{147}{444} \times 100$ $= 33.1$ Survey 33 Year 9s.
5. There has been some rounding, so check that the overall sample size is still 100.	Sample size $= 39 + 28 + 33 = 100$ The sample should consist of 39 Year 7s, 28 Year 8s and 33 Year 9s.

on Resources

Interactivity: Random samples (int-6443)

13.2.4 Using technology to select a sample

If you know the population size, it should also be possible to produce a list of population members. Assign each population member a number (from 1 to N). Use the random number generator on your calculator to generate a random number between 1 and N. The population member who was allocated that number becomes the first member of the sample. Continue generating random numbers until the required number of members has been picked for the sample. If the same random number is generated more than once, ignore it and continue selecting members until the required number has been chosen.

studyon

Units 3 & 4 > Area 8 > Sequence 1 > Concepts 1 & 2

Samples and proportions Summary screen and practice questions
A random sample Summary screen and practice questions

Exercise 13.2 Sample statistics

Technology free

1. **WE1** On average, Mr Parker teaches 120 students per day. He asks one class of 30 about the amount of homework they have that night. Identify the population and sample size.
2. Bruce is able to sew the hems 100 shirts per day. Each day he checks 5 to make sure that they are suitable. Identify the population and sample size.
3. **WE2** Ms Lane plans to begin her Statistics class each year by telling her students a joke. She tests her joke on this year's class (15 students). She plans to retire in 23 years' time. Identify the population and sample size.

4. Lee-Yin is trying to perfect a recipe for cake pops. She tries 5 different versions before she settles on her favourite. She takes some samples to school and asks 9 friends what they think. Identify the population and sample size.

5. **WE3** Identify the following as either sample statistics or population parameters.
   a. Studies have shown that between 85% and 95% of lung cancers are related to smoking.
   b. About 50% of children aged between 9 and 15 years eat the recommended daily amount of fruit.
6. Identify the following as either sample statistics or population parameters.
   a. According to the 2013 census, the ratio of male births per 100 female births is 106.3.
   b. About 55% of boys and 40% of girls reported drinking at least 2 quantities of 500 mL of soft drink every day.
7. **WE4** A school has 523 boys and 621 girls. You are interested in finding out about their attitudes to sport and believe that boys and girls may respond differently. If a sample of 75 students is required, determine how many boys and how many girls should be selected.
8. In a school, 23% of the students are boarders. For this survey, it is believed that boarders and day students may respond differently. To select a sample of 90 students, how many boarders and day students should be selected?
9. You are trying out a new chocolate pudding recipe. You found 40 volunteers to taste test your new recipe compared to your normal pudding. Half of the volunteers were given a serving the new pudding first, then a serving of the old pudding. The other half were given the old pudding first and then the new pudding. The taste testers did not know the order of the puddings they were trying. The results show that 31 people prefer the new pudding recipe.

   a. What is the population size?
   b. What is the sample size?
10. You want to test a new flu vaccine on people with a history of chronic asthma. You begin with 500 volunteers and end up with 247 suitable people to test the vaccine.
    a. What is the population size?
    b. What is the sample size?
11. In a recent survey, 1 in 5 students indicated that they ate potato crisps or other salty snacks at least four times per week. Is this a sample statistic or a population parameter?
12. Around 25 to 30% of children aged 0–15 years eat confectionary at least four times a week. Is this a sample statistic or a population parameter?

13. According to the Australian Bureau of Statistics, almost a quarter (24%) of internet users did not make an online purchase or order in 2012–13. The three most commonly reported main reasons for not making an online purchase or order were: 'Has no need' (33%); 'Prefers to shop in person/ see the product' (24%); and 'Security concerns/concerned about providing credit card details online' (12%). Are these sample statistics or population parameters?
14. According to the 2011 census, there is an average of 2.6 people per household. Is this a sample statistic or a population parameter?

15. A doctor is undertaking a study about sleeping habits. She decides to ask every 10th patient about their sleeping habits.
   a. What type of sample is this?
   b. Is this a valid sampling method?

16. A morning television show conducts a viewer phone-in poll and announces that 95% of listeners believe that Australia should become a republic. Comment on the validity of this type of sample.
17. Tony took a survey by walking around the playground at lunch and asking fellow students questions. Why is this not the best sampling method?

Technology active

18. A company has 1500 staff members, of whom 60% are male; 95% of the male staff work full time, and 78% of the female staff work full time. If a sample of 80 staff is to be selected, identify the numbers of full-time male staff, part-time male staff, full-time female staff and part-time female staff that should be included in the sample.
19. Use your calculator to produce a list of 10 random numbers between 1 and 100.
20. Use your calculator to select a random sample from students in your Mathematical Methods class.

13.3 The distribution of $\hat{p}$

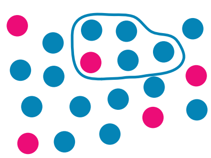
13.3.1 The sample proportion

Let us say that we are interested in the following collection of balls. As you can see in Figure 1, there are 20 balls, and $\frac{1}{4}$ of them are red. This means that the population parameter, p, is $\frac{1}{4}$ and the population size, N, is 20.

FIGURE 1

Normally we wouldn't know the population parameter, so we would choose a sample from the population and determine the sample statistic. In this case, we are going to use a sample size of 5, that is, $n = 5$.

FIGURE 2



If our sample is the group shown in Figure 2, then as there is 1 red ball, the **sample proportion** would be $\hat{p} = \frac{1}{5}$.

A different sample could have a different sample proportion. In the case shown in Figure 3, $\hat{p} = \frac{2}{5}$.
In the case shown in Figure 4, $\hat{p} = 0$.

FIGURE 3

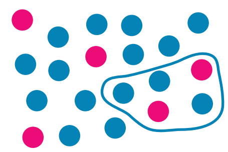
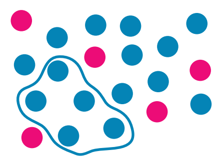


FIGURE 4



It would also be possible to have samples for which $\hat{p} = \frac{3}{5}$, $\hat{p} = \frac{4}{5}$ or $\hat{p} = 1$, although these samples are less likely to occur.

In summary,

Sample proportion

$$\hat{p} = \frac{\text{number of successful outcomes in the sample}}{\text{sample size}}$$

It might seem that using a sample does not give a good estimate about the population. However, the larger the sample size, the more likely that the sample proportions will be close to the population proportion.

on Resources

Interactivity: Distribution of $\hat{p}$ (int-6444)

WORKED EXAMPLE 5

You are trying out a new chocolate tart recipe. You found 40 volunteers to taste test your new recipe compared to your normal one. Half the volunteers were given a serving of the new tart first, then a serving of the original tart. The other half were given the original tart first and then the new one. The taste testers did not know the order of the tarts they were trying. The results show that 31 people prefer the new tart recipe. What is the sample proportion, $\hat{p}$?

THINK	WRITE
1. There are 40 volunteers. This is the sample size.	$n = 40$
2. 31 people prefer the new recipe.	Number of successes = 31
3. Calculate the sample proportion.	$\hat{p} = \frac{31}{40}$

13.3.2 Revision of binomial distributions

In a set of binomial data, each member of the population can have one of two possible values. We define one value as a success and the other value as a failure. (A success isn't necessarily a good thing, it is simply the name for the condition we are counting. For example, a success may be having a particular disease and a failure may be not having the disease).

Binomial distributions

The proportion of successes in a population is called p and is a constant value.

$$p = \frac{\text{number in the population with the favourable attribute}}{\text{population size}}$$

The proportion of failures in a population is called q, where $q = 1 - p$.
The sample size is called n.
The number of successes in the sample is called X.
The proportion of successes in the sample, $\hat{p}$, will vary from one sample to another.

$$\begin{aligned}\hat{p} &= \frac{\text{number in the sample with the favourable attribute}}{\text{sample size}} \\ &= \frac{X}{n}\end{aligned}$$

13.3.3 Sampling distribution of $\hat{p}$

Normally, you would take one sample from a population and make some inferences about the population from that sample. In this section, we are going to explore what would happen if you took lots of samples of the same size. (Assume you return each sample back to the population before selecting again.)

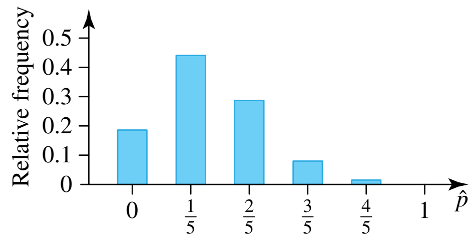
Consider our population of 20 balls (5 red and 15 blue). There are ${}^{20}C_5 = 15\,504$ possible samples that could be chosen. That is, there are 15 504 possible ways of choosing 5 balls from a population of 20 balls. A breakdown of the different samples is shown in the table, where X is the number of red balls in the sample.

X	$\hat{p}$	Number of samples	Relative frequency
0	0	${}^5C_0\,{}^{15}C_5 = 3003$	0.194
1	$\frac{1}{5}$	${}^5C_1\,{}^{15}C_4 = 6825$	0.440
2	$\frac{2}{5}$	${}^5C_2\,{}^{15}C_3 = 4550$	0.293
3	$\frac{3}{5}$	${}^5C_3\,{}^{15}C_2 = 1050$	0.068
4	$\frac{4}{5}$	${}^5C_4\,{}^{15}C_1 = 75$	0.005
5	1	${}^5C_5\,{}^{15}C_0 = 1$	6.450×10^{-5}
	Total number of samples	**15 504**	

Graphing the distribution of $\hat{p}$ against the relative frequency of $\hat{p}$ results in the following.

As the value of $\hat{p}$, the sample proportion, varies depending on the sample, these values can be considered as the values of the random variable, $\hat{P}$.

The graph of the distribution of $\hat{p}$ can also be represented in a probability distribution table. This distribution is called a **sampling distribution**.



$\hat{p}$	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1
$\mathbf{P}(\hat{p} = \hat{p})$	0.194	0.440	0.293	0.068	0.005	6.450×10^{-5}

We can calculate the average value of $\hat{p}$ as shown.

$\hat{p}$	**Frequency, f**	$f \cdot \hat{p}$
0	3 003	0
$\frac{1}{5}$	6 825	1365
$\frac{2}{5}$	4 550	1820
$\frac{3}{5}$	1 050	630
$\frac{4}{5}$	75	60
1	1	1
Totals	**15 504**	**3876**

$$\begin{aligned}\text{The average value of } \hat{p} &= \frac{3876}{15\,504}\\ &= 0.25\end{aligned}$$

For this distribution, the average value for $\hat{p}$ is equal to the population proportion, p.

Expected value of large samples

It was mentioned earlier that larger samples give better estimates of the population. The proportion of $\hat{p}$ in a large sample conforms to $\hat{P} = \frac{X}{n}$. As the sample is from a large population, X can be assumed to be a binomial variable.

$$\begin{aligned}\therefore \mathrm{E}(\hat{P}) &= \mathrm{E}\left(\frac{X}{n}\right)\\ &= \frac{1}{n}\mathrm{E}(X) \quad \left(\text{because } \frac{1}{n} \text{ is a constant}\right)\end{aligned}$$

$$= \frac{1}{n} \times np$$
$$= p$$

Variance and standard deviation

The variance and standard deviation can be found as follows.

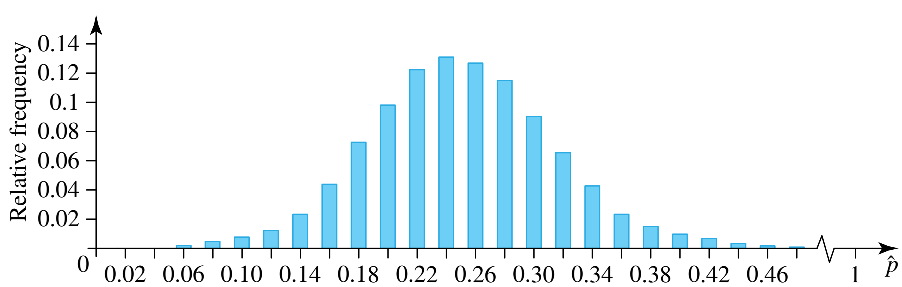
$$\begin{aligned}
\text{Var}(\hat{P}) &= \text{Var}\left(\frac{X}{n}\right) \\
&= \left(\frac{1}{n}\right)^2 \text{Var}(X) \\
&= \frac{1}{n^2} \times npq \\
&= \frac{pq}{n} \\
&= \frac{p(1-p)}{n} \\
\therefore \text{SD}(\hat{P}) &= \sqrt{\frac{p(1-p)}{n}}
\end{aligned}$$

Mean and standard deviation for large samples

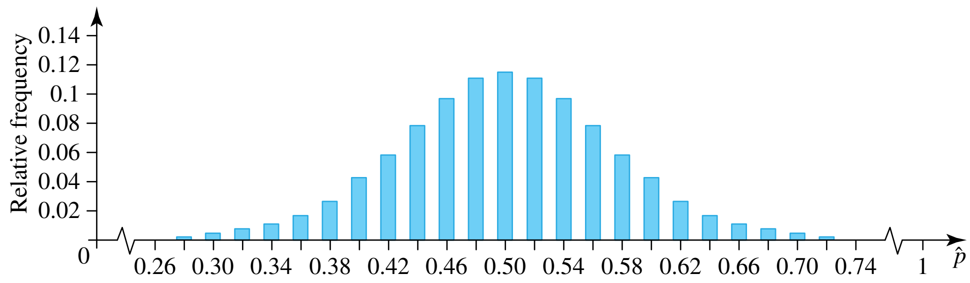
For large samples, the distribution of $\hat{p}$ is approximately normal with a mean or expected value of $\mu_{\hat{p}} = p$ and a standard deviation of $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

There are a number of different ways to decide if a sample is large. One generally accepted method that we will adopt for this section is that if $np \geq 10$, $nq \geq 10$ and $10n \leq N$, then the sample can be called large.

Consider the distribution of $\hat{p}$ when $N = 1000$, $n = 50$ and $p = 0.25$.



Also consider this distribution of $\hat{p}$ when $N = 1000$, $n = 50$ and $p = 0.5$.



As these graphs show, the value of p doesn't matter. The distribution of $\hat{p}$ is symmetrical about p.

WORKED EXAMPLE 6

Consider a population size of 1000 and a sample size of 50. If $p = 0.1$, would this still be a large sample? If not, how big would the sample need to be?

THINK	WRITE
1. Is $10n \le N$?	$n = 50$ and $N = 1000$ $10n = 500$ Therefore, $10n \le N$.
2. Is $np \ge 10$?	$p = 0.1$ $np = 0.1 \times 50$ $= 5$ $5 \not\ge 10$ The sample is not large.
3. Calculate a value for n to make a large sample by solving $np = 10$.	$np = 10$ $0.1n = 10$ $n = 100$
4. Check the other conditions.	$10n = 10 \times 100$ $= 1000$ $= N$ $nq = 100 \times 0.9$ $= 90$ $nq \ge 10$ A sample size of 100 would be needed for a large sample.

WORKED EXAMPLE 7

If $N = 600$, $n = 60$ and $p = 0.3$:

a. calculate the mean of the distribution

b. calculate the standard deviation of the distribution, correct to 2 decimal places.

THINK	WRITE
a. The mean is p.	a. $\mu_{\hat{p}} = p$ $= 0.3$

b. 1. Write the rule for the standard deviation.

2. Substitute the appropriate values and simplify.

b. $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$

$= \sqrt{\dfrac{0.3 \times (1-0.3)}{60}}$

$= 0.06$

on Resources

Interactivity: Sampling distribution of $\hat{p}$ (int-6445)

studyon

Units 3 & 4 › Area 8 › Sequence 1 › Concepts 3 & 4

The sample proportion, $\hat{p}$ Summary screen and practice questions
Sampling distribution of $\hat{p}$ Summary screen and practice questions

Exercise 13.3 The distribution of $\hat{p}$

Technology free

1. **WE5** In a 99-g bag of lollies, there were 6 green lollies out of the 15 that were counted. What is the sample proportion, $\hat{p}$?

2. Hang is interested in seedlings that can grow to more than 5 cm tall in the month of her study period. She begins with 20 seedlings and finds that 6 of them are more than 5 cm tall after the month. What is the sample proportion, $\hat{p}$?

3. **WE6** Consider a population size of 1000 and a sample size of 50. If $p = 0.85$, is this a large sample? If not, how big does the sample need to be?
4. If the population size was 10 000 and $p = 0.05$, what would be a large sample size?
5. **WE7** If $N = 500$, $n = 50$ and $p = 0.5$:
   a. calculate the mean of the distribution
   b. calculate the standard deviation of the distribution, correct to 2 decimal places.
6. If $N = 1000$, $n = 100$ and $p = 0.8$:
   a. calculate the mean of the distribution
   b. calculate the standard deviation of the distribution, correct to 2 decimal places.

7. A car manufacturer has developed a new type of bumper that is supposed to absorb impact and result in less damage than previous bumpers. The cars are tested at 25 km/h. If 30 cars are tested and only 3 are damaged, what is the proportion of undamaged cars in the sample?

8. A standard warranty lasts for 1 year. It is possible to buy an extended warranty for an additional 2 years. The insurer decides to use the sales figures from Tuesday to estimate the proportion of extended warranties sold. If 537 units were sold and 147 of them included extended warranties, estimate the proportion of sales that will include extended warranties.

Technology active

9. A Year 12 Mathematical Methods class consists of 12 girls and 9 boys. A group of 4 students is to be selected at random to represent the school at an inter-school Mathematics competition.

   a. What is the value of p, the proportion of girls in the class?
   b. What could be the possible values of the sample proportion, $\hat{p}$, of girls?
   c. Use this information to construct a probability distribution table to represent the sampling distribution of the sample proportion of girls in the small group.
   d. Determine $\mathrm{P}\left(\hat{P} > 0.6\right)$. That is, determine the probability that the proportion of girls in the small group is greater than 0.6.
   e. Determine $\mathrm{P}\left(\hat{P} > 0.5 | \hat{P} > 0.3\right)$.
10. In a particular country town, the proportion of employment in the farming industry is 0.62. Five people aged 15 years and older are selected at random from the town.
   a. What are the possible values of the sample proportion, $\hat{p}$, of workers in the farming industry?
   b. Use this information to construct a probability distribution table to represent the sampling distribution of the sample proportion of workers in the farming industry.
   c. Determine the probability that the proportion of workers in the farming industry in the sample is greater than 0.5.
11. In a population of 1.2 million, it is believed that $p = 0.01$. What would be the smallest sample size that could be considered large?
12. If $N = 1500$, $n = 150$ and $p = 0.15$, calculate the mean and standard deviation for the distribution of $\hat{p}$. Give your answers correct to 3 decimal places where appropriate.
13. If $N = 1200$, $n = 100$ and $p = 0.75$, calculate the mean and standard deviation for the distribution of $\hat{p}$. Give your answers correct to 3 decimal places where appropriate.
14. A distribution for $\hat{p}$ has a mean of 0.12 and a standard deviation of 0.0285. Calculate the population proportion and the sample size.
15. A distribution for $\hat{p}$ has a mean of 0.81 and a standard deviation of 0.0253. Calculate the population proportion and the sample size.
16. If $N = 1500$, $n = 150$ and $p = 0.15$, use technology to graph the distribution for $\hat{p}$.
17. A distribution for $\hat{p}$ has a standard deviation of 0.015. If the sample size was 510 and $\hat{p} > 0.5$, what was the population proportion, correct to 2 decimal places?
18. A distribution for $\hat{p}$ has a standard deviation of 0.0255. If the sample size was 350, what was the population proportion, correct to 2 decimal places?

13.4 Confidence intervals

You have just learned that different samples can have different values for $\hat{p}$. So what can one sample tell us about a population?

Let us say that you are interested in the proportion of the school that buys their lunch. You decide that your class is a reasonable sample and find out that 25% of the class will buy their lunch today. What can you say about the proportion of the whole school that will buy their lunch today? Assuming that your class is in fact a representative sample, you may say that around 25% of the school will buy their lunch. Is it possible to be more specific? By using **confidence intervals**, it is possible to say how confident you are that a population parameter will lie in a particular interval.

13.4.1 The normal approximation to the distribution of $\hat{p}$

We have learned that when we consider the distributions of $\hat{p}$, they are normally distributed with a $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$. As we don't know the exact value for p, the best estimate is $\hat{p}$. This means that the best estimate of the standard deviation is $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$.

We know that for normal distributions, $z = \dfrac{x-\mu}{\sigma}$. This means that, to determine the upper and lower values of z, we can use $z = \dfrac{\hat{p} \pm p}{\sigma_{\hat{p}}}$. Rearranging gives us $p = \hat{p} \pm z\sigma_{\hat{p}}$.

Approximate confidence interval

An approximate confidence interval for a population proportion is given by

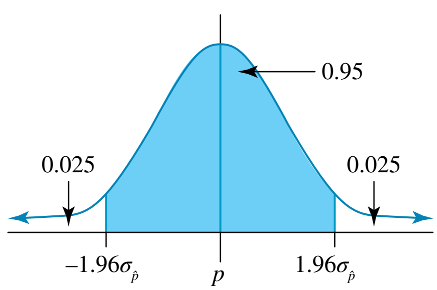
$$(\hat{p} - z\sigma_{\hat{p}}, \hat{p} + z\sigma_{\hat{p}})$$

where $\sigma_{\hat{p}} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$.

A 95% confidence interval means that 95% of the distribution is in the middle area of the distribution. This means that the tails combined contain 5% of the distribution (2.5% on each end). The z-value for this distribution is 1.96.

The confidence interval for this distribution can be expressed as

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$



WORKED EXAMPLE 8

There are 20 people in your class and 25% are planning on buying their lunch. Estimate the proportion of the school population that will purchase their lunch today. Determine a 95% confidence interval for your estimate, given $z = 1.96$.

THINK | **WRITE**

1. There are 20 people in the class. This is the sample size.
   25% are buying their lunch. This is the sample proportion.

$n = 20$
$\hat{p} = 0.25$

2. For a 95% confidence interval, $z = 1.96$.

$z = 1.96$

3. The confidence interval is $\left(\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$.
   Determine $z\sigma_{\hat{p}}$.

$$\begin{aligned} z\sigma_{\hat{p}} &= z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= 1.96\sqrt{\frac{0.25 \times 0.75}{20}} \\ &= 0.1898 \end{aligned}$$

4. Identify the 95% confidence interval by determining the upper and lower values.

$$\begin{aligned} \hat{p} - z\sigma_{\hat{p}} &= 0.25 - 0.1898 \\ &= 0.0602 \end{aligned}$$

$$\begin{aligned} \hat{p} + z\sigma_{\hat{p}} &= 0.25 + 0.1898 \\ &= 0.4398 \end{aligned}$$

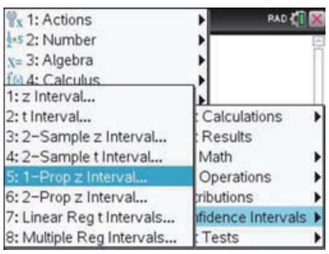
We can be 95% confident that between 6% and 44% of the population will buy their lunch today.

5. Write the answer.

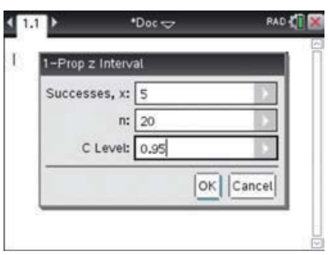
The confidence interval, CI $= (0.06, 0.44)$.

TI | THINK | **WRITE**

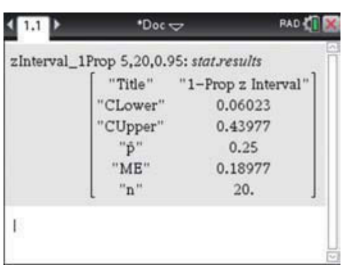
1. On a Calculator page, press MENU then select:
   6: Statistics
   6: Confidence Intervals
   5: 1-Prop z Interval ...



2. Complete the entry lines as:
   x: 5
   n: 20
   C Level: 0.95
   then press the OK button.

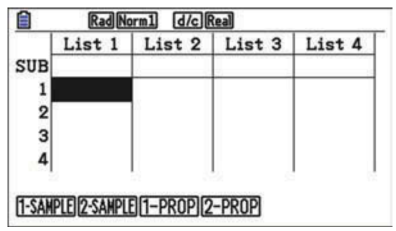


3. The answer appears on the screen.

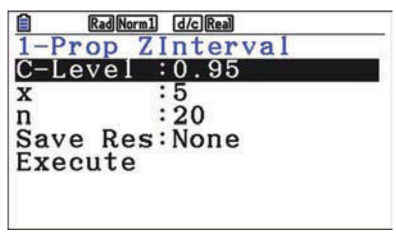


CASIO | THINK | **WRITE**

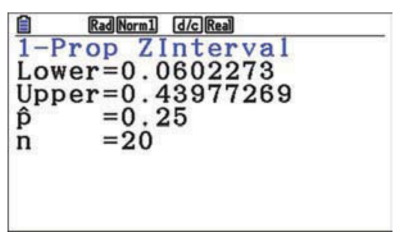
1. On a Statistics screen, press: INTR Z 1-PROP



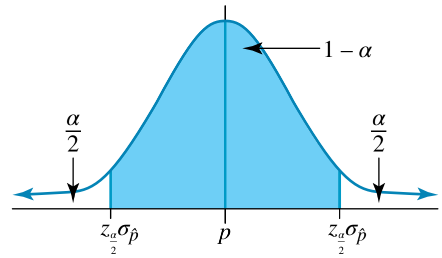
2. Complete the entry lines as:
   C Level: 0.95
   x: 5
   n: 20
   then press the EXE button.



3. The answer appears on the screen.



To determine other confidence intervals, we can talk in general about a $1 - \alpha$ confidence interval. In this case, the tails combined will have an area of α (as each tail has an area of $\frac{\alpha}{2}$). In this case, the z-value that has a tail area of $\frac{\alpha}{2}$ is used.



WORKED EXAMPLE 9

Paul samples 102 people and determines that 18 of them like drinking coconut milk. Estimate the proportion of the population that likes drinking coconut milk. Determine a 99% confidence interval for your estimate, correct to 1 decimal place.

THINK	WRITE
1. There are 102 people in the sample. This is the sample size. 18 like drinking coconut milk.	$n = 102$ $\hat{p} = \frac{18}{102}$ $= 0.18$
2. For a 99% confidence interval, determine the z-value using the inverse standard normal distribution.	For the 99% confidence interval, 1% will be in the tails, so 0.5% in each tail. Therefore, the area under the normal distribution curve to the left of z is 0.995. $z = 2.58$
3. The confidence interval is $(\hat{p} - z\sigma_{\hat{p}}, \hat{p} + z\sigma_{\hat{p}})$. Determine $z\sigma_{\hat{p}}$.	$z\sigma_{\hat{p}} = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $= 2.58\sqrt{\frac{0.18 \times 0.82}{102}}$ $= 0.098$
4. Identify the 99% confidence interval by determining the upper and lower values, correct to 1 decimal place.	$\hat{p} - z\sigma_{\hat{p}} = 0.18 - 0.098$ $= 0.082$ $\hat{p} + z\sigma_{\hat{p}} = 0.18 + 0.098$ $= 0.278$ We can be 99% confident that between 8.2% and 27.8% of the population like drinking coconut milk.

WORKED EXAMPLE 10

Grow Well are 95% sure that 30% to 40% of shoppers prefer their mulch. What sample size was needed for this level of confidence?

THINK	WRITE
1. The confidence interval is symmetric about $\hat{p} : (\hat{p} - z\sigma_{\hat{p}}, \hat{p} + z\sigma_{\hat{p}})$, so the value of $\hat{p}$ must be halfway between the upper and lower values of the confidence interval.	$\hat{p} = \frac{30 + 40}{2}$ $= 35\%$ $= 0.35$

2. State the z-value related to the 95% confidence interval.

$z = 1.96$

3. The lower value of the confidence interval, 30%, is equivalent to $\hat{p} - z\sigma_{\hat{p}}$. Substitute the appropriate values.
Note: The equation $0.4 = \hat{p} + z\sigma_{\hat{p}}$ could also have been used.

$$0.3 = \hat{p} - z\sigma_{\hat{p}}$$
$$= \hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$= 0.35 - 1.96\sqrt{\frac{0.35(1-0.35)}{n}}$$

4. Solve for n.

$n = 349.586$

5. Write the answer.

The sample size needed was 350 people.

13.4.2 Margin of error

The distance between the endpoints of the confidence interval and the sample estimate is called the **margin of error**, E.

Worked example 10 considered a 95% confidence interval, $(\hat{p} - z\sigma_{\hat{p}}, \hat{p} + z\sigma_{\hat{p}})$. In this case the margin of error would be $E = z\sigma_{\hat{p}}$.

Margin of error for a 95% confidence interval

For a 95% level of confidence,

$$E = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Note that the larger the sample size, the smaller the value of E will be. This means that one way to reduce the size of a confidence interval without changing the level of confidence is to increase the sample size.

studyon

Units 3 & 4 > Area 8 > Sequence 1 > Concept 5 > **Confidence intervals** Summary screen and practice questions

Exercise 13.4 Confidence intervals

Technology active

1. WE8 Of 30 people surveyed, 78% said that they like breakfast in bed. Estimate the proportion of the populations that like breakfast in bed. Determine a 95% confidence interval for the estimate.
2. Of the 53 people at swimming training today, 82% said that their favourite stroke is freestyle. Estimate the proportion of the population whose favourite stroke is freestyle. Determine a 95% confidence interval for the estimate.

3. WE9 Jenny samples 116 people and finds that 86% of them plan to go swimming over the summer holidays. Estimate the proportion of the population that plan to go swimming over the summer holidays. Determine a 99% confidence interval for your estimate.

4. Yuki samples 95 people and finds that 30% of them eat chocolate daily. Estimate the proportion of the population that eats chocolate daily. Determine a 90% confidence interval for your estimate.
5. **WE10** In a country town, the owners of Edie's Eatery are 95% sure that 35% to 45% of their customers love their homemade apple pie. What sample size was needed for this level of confidence?

6. If Parkers want to be 90% confident that between 75% and 85% of their customers will shop in their store for more than 2 hours, what sample size will be needed?

The following information relates to questions 7, 8 and 9.

Teleco is being criticised for its slow response time when handling complaints. The company claims that it will respond within 1 day. Of the 3760 complaints in a given week, a random sample of 250 was selected. Of these, it was found that 20 of them had not been responded to within 1 day.

7. Determine the 95% confidence interval for the proportion of claims that take more than 1 day to resolve.
8. What is the 99% confidence interval for the proportion of claims that take less than 1 day to resolve?
9. Teleco want to be 95% sure that less than 5% of their complaints take more than 1 day to resolve. What sample proportion do they need and how large does the sample need to be to support this claim?
10. A sample of 250 blood donors have their blood types recorded. Of this sample, 92 have Type A blood. What is the 90% confidence interval for the proportion of Australians who have Type A blood?

11. It is believed that 65% of people have brown hair. A random selection of 250 people were asked the colour of their hair. Applying the normal approximation, determine the probability that less than 60% of the people in the sample have brown hair.
12. Nidya is a top goal shooter. The probability of her getting a goal is 0.8. To keep her skills up, each night she has 200 shots on goal. Applying the normal approximation, determine the probability that on Monday the proportion of times she scores a goal is between 0.8 and 0.9, given that it is more than 0.65.
13. Smooth Writing are 95% sure that 25% to 35% of shoppers prefer their pen. What sample size was needed for this level of confidence?
14. An online tutoring company is 99% sure that 20% to 30% of students prefer to use their company. What sample size was needed for this level of confidence?
15. Barton's Dentistry want to be able to claim that 90% to 98% of people floss daily. They would like 99% confidence about their claim. How many people do they need to survey?
16. Tatiana is conducting a survey to estimate the proportion of Year 12 students who will take a gap year after they complete their studies. Previous surveys have shown the proportion to be approximately 15%. Determine the required size of the sample so that the margin of error for the survey is 3% in a confidence interval of approximately 95% for this proportion.
17. Bentons claim that between 85% and 95% of their customers stay for more than 2 hours when they shop. If they surveyed 100 people, how confident are they about that claim?

18. The Brisbane Lions Football Club claim that between 75% and 80% of their members remain members for at least 10 years. If they surveyed 250 people, how confident are they about that claim? Give your answer to the nearest whole number.

13.5 Review: exam practice

A summary of this chapter is available in the Resources section of your eBookPLUS at www.jacplus.com.au.

Simple familiar

1. Identify which of the following are population parameters.
  a. According to the Australian Bureau of Statistics, the unemployment rate is 6.4%.
  b. According to the 2011 census, on average there are 1.7 motor vehicles per dwelling.
  c. According to a poll in the newspaper *The Age*, 54% of Australians will vote Liberal at the next election.

2. Identify the following as either population parameters or sample statistics.
  a. According to data collected at enrolment, 97% of students speak English as their first language.
  b. Between 3% and 15% of children who leave home return again within 12 months.
  c. Three out of every 5 students brings an apple with their lunch.

3. Glen is interested in the proportion of internet users in his local town. He surveys 150 households and determines that 132 of them have internet access. What proportion of households in Glen's town have internet access?

4. A local vet decides to survey every 7th client about their pet care.
  a. What type of sample is this?
  b. Is this a valid method for this situation?

5. If a population proportion is believed to be 0.25 and samples of size 30 are chosen, what is the standard deviation of $\hat{p}$?

6. In a recent voter survey, an approximate 90% confidence interval for the proportion of people who will vote for a particular party is $(0.58, 0.66)$.
  a. What is the value of $\hat{p}$ for this confidence interval?
  b. What is the value of the margin of error?

7. Ronit wants to survey the students at his school. He wants to survey 50 students, but he believes that it is necessary to use a stratified random sample.

Gender	Middle school	Senior school
Male	253	342
Female	287	323

a. What is the population size?
b. What is the sample size?
c. How many from each subgroup should be sampled?

8. The formal committee of Brickwall State High School is trying to choose between strawberry cheesecake and passionfruit cheesecake for dessert at the formal dinner. There will be 423 students at the dinner. Of the 52 senior students sampled, 23 prefer the strawberry cheesecake. What proportion of students chose passionfruit cheesecake?

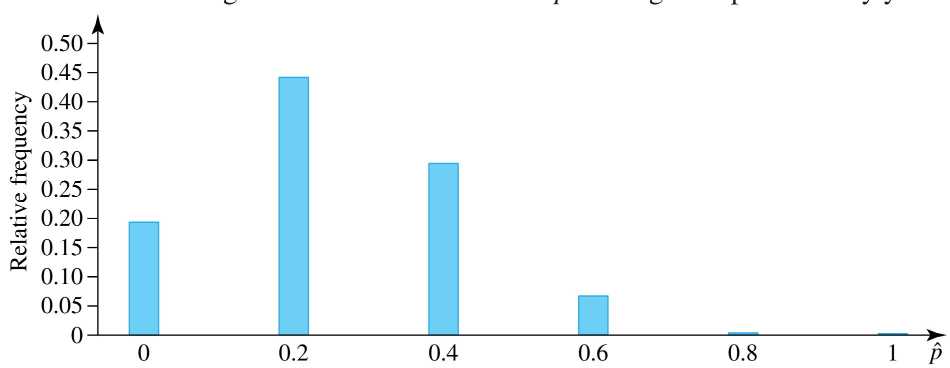
9. If $N = 2\,000$, $n = 120$ and $p = 0.37$, calculate:
a. the mean of the population
b. the standard deviation of the population, correct to 3 decimal places.

10. A textbook publishing company is 95% sure that 65%– 75% of students prefer to use their resources. What sample size was needed for this level of confidence?

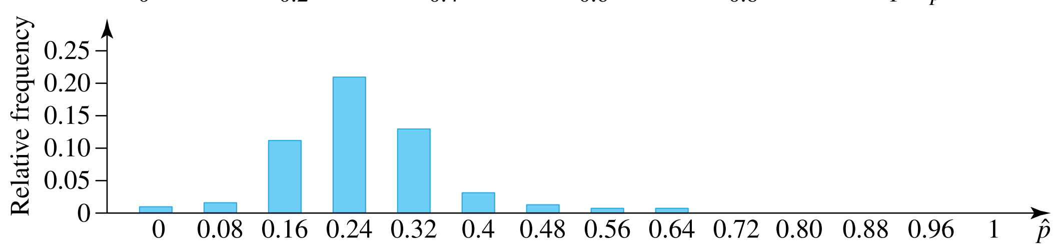
11. On a particular Friday night, 52 000 people went to Suncorp Stadium to watch the football. Every 25th person entering the stadium was asked who they thought would win. Out of the people asked, 1600 people believed that the Queensland Reds would win.
a. What is the population size?
b. What is the sample size?
c. Determine $\hat{p}$.

12. Which of the following could be a distribution for $\hat{p}$ for large samples? Justify your answer.

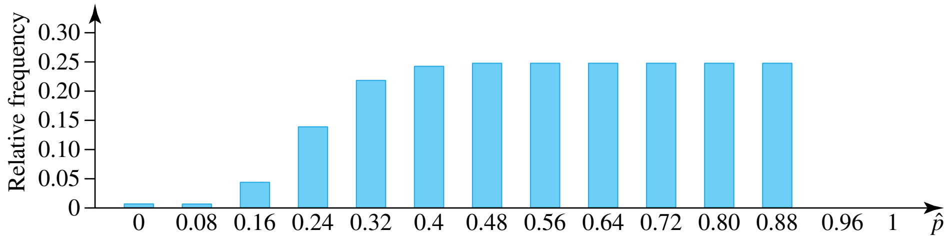
A.



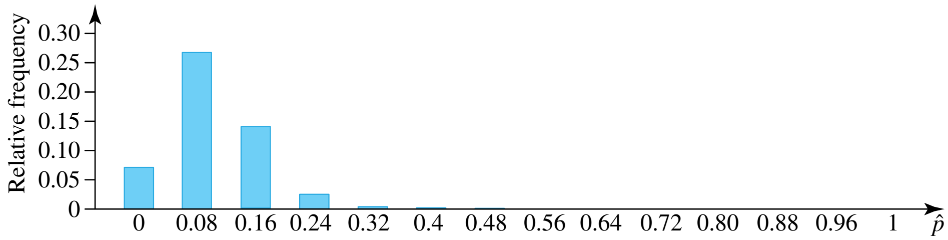
B.



C.



D.



Complex familiar

13. Each year the Year 7 class from Gympie State High visits a theme park. One hundred of the students decided to go on the monster rollercoaster, and 10 of them complained of feeling dizzy afterwards.
   a. What is the value of the sample proportion?
   b. Write an expression for the 95% confidence interval for the likelihood of feeling dizzy.
   c. Determine the margin of error, E, for the 85% confidence interval.
   d. If only 50 people had decided to go on the rollercoaster, what would be the effect on the margin of error?
14. A distribution for $\hat{p}$ has a standard deviation of 0.05. If the sample size is 50, what is the population proportion, correct to 2 decimal places?
15. Maxwell Industries surveyed 100 customers and believe that between 85% and 90% of their customers are satisfied with their level of service. How confident are they about that claim?
16. Krypton Industries are 99% certain that between 67% and 83% of people prefer their product. How many people were sampled for this level of confidence?

Complex unfamiliar

17. Every year, thousands of tourists drive the Great Ocean Road. In a recent survey of 50 people, 87% listed seeing the Twelve Apostles as the highlight of their drive. What proportion of drivers would rate the Twelve Apostles as the highlight of their drive? Give your answer with a 90% confidence level.
18. It is believed that 40% of Australians wear glasses or contact lenses. Four hundred people were randomly selected and asked about their eyesight. Applying the normal distribution, determine the probability that more than 45% of the people in the sample need to wear glasses or contact lenses.
19. The lower limit of a 95% confidence interval is 13%. If 100 people were surveyed, what is the sample proportion, correct to 2 decimal places? Solve this problem graphically.

20. Breanna, Kayley and Teagan spent the day collecting survey results from the same population. They each surveyed 100 people. Breanna found that 23% of people said Yes, Kayley found that 20% of people said Yes, and Teagan found that 19% of people said Yes. They want to obtain an estimate for the population parameter at a 95% confidence interval.

Breanna says they should each work out a confidence interval and then average them out to give the population parameter. Kayley says that they should combine their data into one sample and determine the population parameter using that parameter. Teagan says that it doesn't matter, because they will get the same results either way.

a. Is Teagan correct? (Show all your working.)
b. Who has the most reliable method? Explain your answer.
c. What is the best estimate of the population parameter at a confidence level of 95%?

studyon

Units 3 & 4 Sit exam

Answers

13 Sampling and confidence intervals

Exercise 13.2 Sample statistics

1. $N = 120$, $n = 30$
2. $N = 100$, $n = 5$
3. $n = 15$, population size is unknown
4. $n = 9$, population size is unknown
5. a. Population parameter b. Sample statistic
6. a. Population parameter b. Sample statistic
7. 34 boys and 41 girls
8. 21 boarders, 69 day students
9. a. The population size is unknown.
   b. 40
10. a. The population is people who will receive the vaccine in the future. The size is unknown.
    b. 247
11. Sample statistic
12. Population parameter
13. Sample statistics
14. Population parameter
15. a. A systematic sample with $k = 10$
    b. Yes, assuming that the order of patients is random
16. The sample is not random; therefore, the results are not likely to be random.
17. It is probably not random. Tony is likely to ask people who he knows or people who approach him.
18. Full-time male staff: 46
    Part-time male staff: 2
    Full-time female staff: 25
    Part-time female staff: 7
19. Use the random number generator on your calculator to produce numbers from 1 to 100. Keep generating numbers until you have 10 different numbers.
20. First, assign every person in your class a number, e. g. 1 to 25 if there are 25 students in your class. Decide how many students will be in your sample, e.g. 5. Then use the random number generator on your calculator to produce numbers from 1 to 25. Keep generating numbers until you have 5 different numbers. The students that were assigned these numbers are the 5 students in your random sample.

Exercise 13.3 The distribution of $\hat{p}$

1. $\frac{2}{5} = 0.4$
2. $\frac{3}{10} = 0.3$
3. This is not a large sample; $n = 67$ would be a large sample.
4. $n = 200$
5. a. 0.5
   b. 0.07
6. a. 0.8
   b. 0.04
7. $\frac{9}{10} = 0.9$
8. $\frac{147}{537} \approx 0.274$
9. a. $\frac{4}{7} \approx 0.571$
   b. $0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$
   c.

$\hat{p}$	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
$P(\hat{P} = \hat{p})$	0.021	0.168	0.397	0.331	0.083

   d. 0.414
   e. 0.510
10. a. $0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1$
    b.

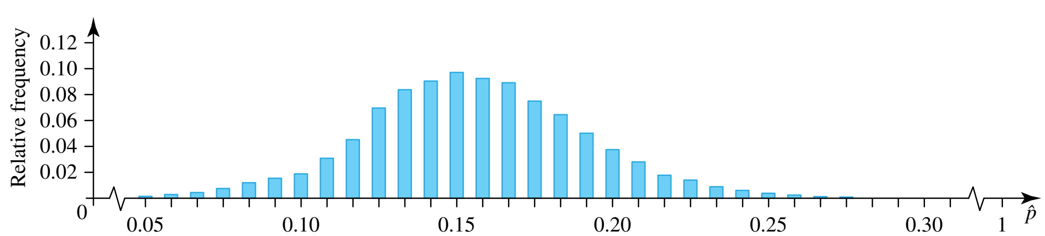
$\hat{p}$	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1
$P(\hat{P} = \hat{p})$	0.008	0.064	0.211	0.344	0.281	0.092

    c. 0.717
11. 1000
12. $\mu_{\hat{p}} = 0.15$, $\sigma_{\hat{p}} = 0.029$
13. $\mu_{\hat{p}} = 0.75$, $\sigma_{\hat{p}} = 0.043$
14. $p = 0.12$, $n = 130$
15. $p = 0.81$, $n = 240$
16. *See the figure at the foot of the page.
17. $p = 0.87$
18. $p = 0.35$ or $p = 0.65$

Exercise 13.4 Confidence intervals

1. $(0.63, 0.93)$
2. $(0.72, 0.92)$
3. $(0.78, 0.94)$
4. $(0.22, 0.38)$
5. 369
6. 173
7. $(0.46, 0.114)$
8. $(0.876, 0.964)$

*16. a.



9. $\hat{p} = 2.5\%$; $n = 150$
10. $(0.318, 0.418)$
11. 0.0487
12. 0.4998
13. 323
14. 498
15. 235
16. 544
17. 90%
18. 66%

13.5 Review: exam practice

1. **b**
2. a. Population parameter
   b. Population parameter
   c. Sample statistic
3. 0.88
4. a. Systematic sample
   b. Yes, as the clients are likely to be in a random order.
5. 0.079
6. a. 0.62 b. 0.04
7. a. 1205
   b. 50
   c. Middle school: 11 male students, 12 female students; Senior school: 14 male students, 13 female students
8. 0.56
9. a. 0.37 b. 0.044
10. 323
11. a. 52 000 b. 2080 c. 0.77
12. B
13. a. 0.1
    b. 0.1 ± 0.0432
    c. 0.0432
    d. E increases by a factor of $\sqrt{2}$.
14. $p = 0.15$ or 0.85
15. 0.55
16. 195
17. 79%–95%
18. 0.0206
19. 0.21
20. a. Teagan is not correct.
    b. Kayley's method
    c. 16%–26%