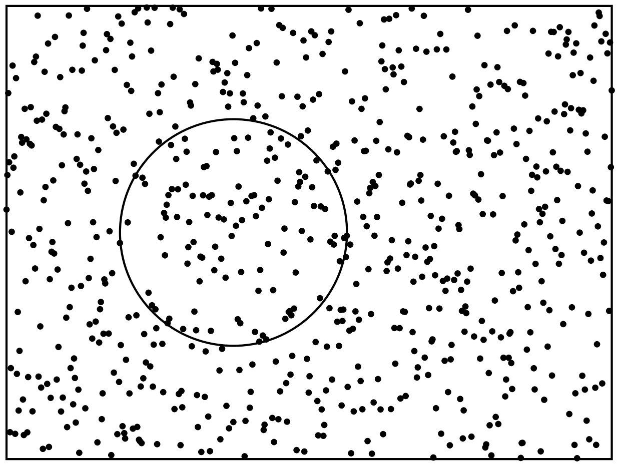
Chapter 25

Statistical Inference

Recall from Chapter 18 that

- a **population** is the whole group about which information is wanted, whereas
- a **sample** is a part of the population, usually a small part, that has been chosen to gather information about the population.



In the diagram above, the rectangle is the **population** and the circle is the **sample**.

We are especially interested in **random samples**. (Recall that a **random sample** of size n is obtained by randomly choosing n items from a population.)

In this chapter, we will learn about **statistical inference**, which

uses data collected from random samples to make conclusions about entire populations.

25.1 Population proportion and sample proportion

Suppose we want to know about the occurrence of a particular feature in a population. (In previous chapters, the items with the particular feature being considered were often referred to as "successes".)

The proportion of items in the population that have this feature is called

the **population proportion**, and is represented with the symbol $\boldsymbol{p}$.

Now suppose that a sample is taken from the population. A proportion of the items *in the sample* will have the particular feature we are interested in. This proportion is called

the **sample proportion**, and is represented with the symbol $\boldsymbol{\hat{p}}$.

Example 1. Suppose that we are interested in the proportion of people in Australia who are female. Then the **population** consists of all the people in Australia, and

- the **population proportion** of females is given by

$$p = \frac{\text{the number of females in Australia}}{\text{the total number of people in Australia}} .$$

If we choose a sample of people from the Australian population, then

- the **sample proportion** of females is given by

$$\hat{p} = \frac{\text{the number of females } \textit{in the sample}}{\text{the total number of people } \textit{in the sample}} .$$

□

Usually, it is not practical to check every item in the population (because the population is so large). Therefore,

usually the value of p *is unknown.*

However, if we consider a sample, then we can use that sample to find a value of $\hat{p}$. We might then

estimate p by using that value of $\hat{p}$.

(Notice that this fits in with our overall goal of making conclusions about entire populations by *using data collected from random samples.)*

Example 2. Consider the population of all teachers at Trinity College. Suppose that there are 300 teachers in total, of whom

200 are female teachers and 100 are male teachers.

A student does not know the proportion of female teachers at Trinity and wishes to estimate it. The student chooses a random sample of 10 teachers from the population. Suppose that this sample contains

7 female teachers and 3 male teachers.

Calculate

(a) p, the population proportion of female teachers

(b) $\hat{p}$, the sample proportion of female teachers.

Solution: (a) The population proportion of female teachers is $p = \dfrac{200}{300} = \dfrac{2}{3}$.

(b) The sample proportion of female teachers is $\hat{p} = \dfrac{7}{10}$. □

- Note that, for a population at some fixed time, the population proportion p is a *constant*, in the sense that each person who calculates the population proportion at that particular time will get the *same* answer.
- In contrast however, the sample proportion $\hat{p}$ is *not* constant. Instead, it depends on the particular random sample that is chosen. (Different people are likely to obtain different samples, which will likely have *different* sample proportions.)

The sample proportion $\hat{p}$ provides an estimate of the population proportion p. We call it a **point estimate** for p.

Continuing the previous example:

The point estimate of $\hat{p} = \frac{7}{10} = 0.7$ happens to be a *good estimate* of p, since it is *quite close* to p. (Recall that we had $p = \frac{2}{3} \approx 0.67$.) However, if the student took another random sample of 10 teachers, that second sample might contain

1 female teacher and 9 male teachers,

in which case the sample proportion of female teachers for this second sample would be $\hat{p} = \frac{1}{10}$. This is also a point estimate for p, although not an accurate one (since it is very different from p).

Note: The sample proportion from any particular sample is <u>very unlikely</u> to *equal* the population proportion.

Consider the previous example again:

The possible values of $\hat{p}$ from different samples of 10 teachers are $0,\ 0.1,\ 0.2,\ \ldots\ 0.9,\ 1$. Notice that none of these is *equal* to $\frac{2}{3}$, the value of p.

In real life, the population proportion p usually is unknown and we have to estimate it by taking a sample. In this chapter, we will use probability theory to infer a value for an unknown population proportion p from a sample proportion $\hat{p}$.

25.2 Random variable for sample proportion

If we take a random sample of size n, we can define a random variable X as

$X =$ the number of items in the sample with the particular feature we are interested in.

Then $X \in \{0, 1, 2, \ldots, n-1, n\}$.

(Note that, in unusual examples, the particular feature that we are interested in might be very *rare* in the population, in which case X might not be able to equal the largest values from this set.)

We then can define another random variable, $\hat{P}$, whose possible values are all the values of the sample proportion $\hat{p}$, as

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

Example 3. Consider a population of 100 people in which exactly 40 people like chocolate. Suppose we randomly choose a sample of three people from this population and ask them if they like chocolate. Define appropriate random variables and list their possible values.

Solution: Let X be the number of people in the sample who like chocolate, where $n = 3$.

The possible values (x) of X are $0, 1, 2, 3$.

Let $\hat{P}$ be the random variable for the sample proportion of people who like chocolate. Then $\hat{P} = \frac{X}{3}$ and the possible values $(\hat{p})$ of $\hat{P}$ are $0, \frac{1}{3}, \frac{2}{3}, 1$.

□

When we choose a sample from a population, we usually sample *without replacement*. In this case X has the hypergeometric distribution which we studied in Chapter 20.4, and so does $\hat{P}$. When the size of the population is 'large', however, the hypergeometric calculations can be troublesome. At this point, we make a key assumption:

If the population is 'large' (in comparison to the sample size),
then we may approximate X well by assuming it has a binomial distribution.
That is, the answers obtained by using the binomial formulae
(by pretending that the probability of success remains <u>constant</u>)
are similar to the answers obtained by using the hypergeometric formulae.

The next example illustrates the closeness of this approximation.

Example 4. Consider the population of 100 people in which exactly 40 people like chocolate, and again randomly choose three people (without replacement) from this population and ask them if they like chocolate. For the sample proportion variable $\hat{P}$ considered in Example 3 of this chapter calculate

(a) its probability distribution, with values given to three decimal places.

(b) its approximate distribution obtained by approximating X with a binomial variable.

Solution: (a) Again let X be the number of people in the sample who like chocolate. Then X is hypergeometric with $N = 100$, $D = 40$ and $n = 3$. The probability distribution is:

x	$\hat{p}$	$\Pr(X = x) = \Pr(\hat{P} = \hat{p})$
0	0	$\dfrac{{}^{40}\mathrm{C}_0 \times {}^{60}\mathrm{C}_3}{{}^{100}\mathrm{C}_3} = 0.212$ (3 d.p.)
1	$\frac{1}{3}$	$\dfrac{{}^{40}\mathrm{C}_1 \times {}^{60}\mathrm{C}_2}{{}^{100}\mathrm{C}_3} = 0.438$ (3 d.p.)
2	$\frac{2}{3}$	$\dfrac{{}^{40}\mathrm{C}_2 \times {}^{60}\mathrm{C}_1}{{}^{100}\mathrm{C}_3} = 0.289$ (3 d.p.)
3	1	$\dfrac{{}^{40}\mathrm{C}_3 \times {}^{60}\mathrm{C}_0}{{}^{100}\mathrm{C}_3} = 0.061$ (3 d.p.)

(b) Now we will approximate X with a *binomial* variable X^*, obtained by pretending that the probability that a person in the sample likes chocolate *remains constant* at $\frac{40}{100}$.

Then this binomial variable X^* has $p = 0.4$ and $n = 3$.

We define $\hat{P}^* = \dfrac{X^*}{3}$, and calculate the following probability distribution:

x	$\hat{p}$	$\Pr(X^* = x) = \Pr(\hat{P}^* = \hat{p})$
0	0	${}^{3}\mathrm{C}_0\,(0.4)^0\,(0.6)^3 = 0.216$
1	$\frac{1}{3}$	${}^{3}\mathrm{C}_1\,(0.4)^1\,(0.6)^2 = 0.432$
2	$\frac{2}{3}$	${}^{3}\mathrm{C}_2\,(0.4)^2\,(0.6)^1 = 0.288$
3	1	${}^{3}\mathrm{C}_3\,(0.4)^3\,(0.6)^0 = 0.064$

$\square$

Notice how *similar* the probabilities are in the two probability distributions found in this example. Thus the distribution for this random variable X is *approximated well* by the binomial distribution.

We have seen that when the population size is 'large' (compared to the size of the sample), we may approximate the hypergeometric distribution well by using the binomial distribution. That is, for a 'small' sample from a 'large' population, we may assume that $\Pr(\mathit{success})$ remains constant as we take our sample. Some books (including this one) suggest that

such an approximation is appropriate if
the sample size is less than one-tenth of the size of the population;

$$\text{that is, if } \boldsymbol{n} < \frac{\textbf{size of the population}}{\textbf{10}} .$$

Example 5. Consider a population of 500 students of whom exactly 150 like Korean drama serials. Suppose we are going to randomly sample 10 students from this population and ask them if they like Korean drama serials. Use a binomial variable to approximate, to three decimal places, the probability that the sample proportion is 0.6.

Solution: Let X be the number of students in the sample who like Korean drama serials. We are told that the sample size n is 10, and the population size is 500. Since

$$n < \frac{\text{size of the population}}{10}$$

we should be able to calculate a reasonably accurate probability by approximating X with a binomial variable, with $p = \frac{150}{500} = 0.3$ and $n = 10$.

Then, defining the sample proportion variable $\hat{P}$ by $\hat{P} = \dfrac{X}{10}$, we have

$$\Pr(\hat{P} = 0.6) = \Pr(X = 6) \approx {}^{10}\mathrm{C}_6\,(0.3)^6\,(0.7)^4 = 0.037 \quad (3\text{ d.p.}).$$

□

It turns out that the answer for this example (which we obtained by approximating X with a binomial variable) is very close to the actual probability of

$$\Pr(\hat{P} = 0.6) = \Pr(X = 6) = \frac{{}^{150}\mathrm{C}_6 \times {}^{350}\mathrm{C}_4}{{}^{500}\mathrm{C}_{10}} = 0.036\ (3\text{ d.p.}).$$

Note: In this example, the p used to represent $\Pr(\mathit{success})$ is the same as the p we have been using to represent the population proportion. These two definitions of the symbol p are referring to the same number, so we don't need to worry about confusion!

Exercises for Sections 25.1 and 25.2

1. A politician wishes to estimate the proportion of people who will vote for Labor in an upcoming Australian national election. He randomly chooses 100 students from the University of Melbourne, and asks them if they plan to vote for Labor. Do you think that the politician has correctly chosen his random sample?

2. Consider the population of all students at Trinity College. A group of 100 of these students were selected in a random sample, and it was found that 26 of them travel on a train each day to get to school. What is the sample proportion of students that take the train to school?

3. A retirement village has a population of 600 elderly people and 150 of these people are over 80 years of age. A sample of 10 people is randomly selected and three of these are over 80 years of age.
   Find:

   (a) the population proportion of people over 80 years of age.

   (b) the sample proportion of people over 80 years of age.

4. A population of 150 people has exactly 30 people who read newspapers everyday. A sample of 5 people is randomly selected from this population. Define random variables X and $\hat{P}$ for this example, and list the possible values of those variables.

5. A store has 50 apples of various types for sale. Of these, there are five apples that each weigh 60 grams. A random sample of 4 apples is selected, and the variable X is defined to be

   the number of apples in the sample which weigh 60 grams.

   (a) For the sample proportion, $\hat{P}$, calculate its hypergeometric distribution, with values given to five decimal places.

   (b) Is it sensible to approximate this distribution with a binomial distribution? If so, calculate its binomial distribution, with values given to 4 decimal places.

25.3 Measures for the random variable $\hat{P}$

In this section, we will use one situation for two examples.

> Suppose that the population proportion of females in ABigCity is $p = 0.52$. And suppose that some students don't know that, and wish to estimate the unknown p from a sample proportion $\hat{p}$.

Example 6. Jiawei chooses a random sample of 10 people in ABigCity, and this sample contains seven females and three males. Yuhan also chooses a random sample of 10, and that sample contains four females and six males.

Determine the point estimate for p from each sample.

Solution: Let X be the number of females in a sample of 10 people. Then $\hat{P} = \dfrac{X}{10}$.

For Jiawei's sample, $\hat{P}$ has the value of $\hat{p} = \frac{7}{10} = 0.7$.

For Yuhan's sample, $\hat{P}$ has the value of $\hat{p} = \frac{4}{10} = 0.4$.

Both of these are point estimates for the population proportion p.

□

Instead of comparing different samples, however, we will use probability.

We start by considering the *expected value* of $\hat{P}$ (that is $E(\hat{P})$), which is the long-run average of the values of $\hat{p}$:

Recall from the previous section that, if X represents the number of 'successes' in a sample, then X can be approximated well by assuming it has a binomial distribution, as long as

the sample size n is less than one-tenth of the size of the population.

For the rest of this chapter, we will assume that the sample size is always less than one-tenth of the population size *(even when we are considering large sample sizes), so that we can approximate the distribution with the binomial distribution. This is possible by considering* ***very large*** *population sizes.*

In that case, we have $E(X) = np$ (as stated on page 3 of Chapter 21, and on the Formula Sheet), where p is the (unknown) population proportion.

Then, since $\hat{P} = \dfrac{X}{n}$, we have $E(\hat{P}) = E\left(\dfrac{X}{n}\right) = \dfrac{1}{n}E(X) = \dfrac{1}{n}np = p.$

That is,

the long-run average of the values of $\hat{P}$ is
the population proportion p that we are wanting to estimate.

Example 7. Jiawei's classmates choose eight random samples of size 10, and obtain the following eight values for $\hat{p}$:

$$0.6, \quad 0.5, \quad 0.7, \quad 0.3, \quad 0.3, \quad 0.4, \quad 0.5, \quad 0.7 \ .$$

Yuhan's classmates choose 100 random samples of size 10, and obtain the following 100 values for $\hat{p}$:

$\hat{p}$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
frequency	0	1	4	9	18	28	24	9	6	0	1

Their friend, Bowen, also arranges with classmates to take samples. Bowen's classmates choose eight random samples of size **100**, and obtain the following eight values for $\hat{p}$:

$$0.56, \quad 0.49, \quad 0.58, \quad 0.45, \quad 0.51, \quad 0.47, \quad 0.54, \quad 0.59 \ .$$

For each experiment, calculate

(a) the average of the values for $\hat{p}$, as an estimate of the value of p , to three decimal places.

(b) the standard deviation of the values for $\hat{p}$, to four decimal places.

Solution:

(a) Jiawei's class: The average of the eight $\hat{p}$ -values is given by

$$\frac{0.6 + 0.5 + 0.7 + 0.3 + 0.3 + 0.4 + 0.5 + 0.7}{8} = \frac{4.0}{8} = 0.500 \ .$$

Yuhan's class: The average of the 100 $\hat{p}$ -values is given by

$$\frac{0 \times 0 + 0.1 \times 1 + 0.2 \times 4 + \ldots + 1 \times 1}{100} = \frac{51.3}{100} = 0.513 \ .$$

Bowen's class: The average of the eight $\hat{p}$ -values is given by

$$\frac{0.56 + 0.49 + 0.58 + 0.45 + 0.51 + 0.47 + 0.54 + 0.59}{8}$$

. $$= \frac{4.19}{8} = 0.524 \ \ (3 \text{ d.p.}) \ .$$

(b) Using the statistical functions on a calculator (as taught when learning Chapter 16), we obtain the following results:

Jiawei's class (for which $n = 10$) : For these 8 values of $\hat{p}$, $s = 0.1604$ (4 d.p.).

Yuhan's class (for which $n = 10$) : For these 100 values of $\hat{p}$, $s = 0.1568$ (4 d.p.).

Bowen's class (for which $n = 100$) : For these 8 values of $\hat{p}$, $s = 0.0518$ (4 d.p.).

□

Unlike these students, we know (because it was stated at the start of this section) that the population proportion for Example 7 is $p = 0.52$. Notice (from Example 7(a)) that

- the average (0.513) obtained for the 100 values from Yuhan's class (which had $n = 10$)

is *closer* to this population proportion 0.52 (and so is a *better estimate* of the population proportion) than is

- the average (0.5) obtained for the 8 values from Jiawei's class (which also had $n = 10$).

This illustrates that, for the same sample size n,

the average value of $\hat{p}$ from a (much) larger number of samples
will often produce a closer estimate for p.
This is because the long-run average is given by $E(\hat{P}) = p$.

Also, notice (from Example 7(a)) that the average (0.524) obtained for the 8 values from Bowen's class (which had $n = \mathbf{100}$) is even *closer* to the population proportion 0.52 . Furthermore, in Example 7(b) we see that

the *larger* sample size $(n = 100)$ used by Bowen's class
corresponds to a much *smaller* standard deviation of the values of $\hat{p}$.

To see why this situation occurs, let's return to probability and consider the variance and standard deviation of $\hat{P}$.

As seen previously, if the sample size is less than one tenth of the size of the population, then the random variable X for the number of 'successes' in a sample can be approximated well by assuming it has a binomial distribution. Then, for sample size n and (unknown) population proportion p, we have

$$\text{Var}(X) = np(1-p)$$

(as stated on page 3 of Chapter 21, and on the Formula Sheet). Thus, since $\hat{P} = \dfrac{X}{n}$,

we have
$$\begin{aligned} \text{Var}(\hat{P}) &= \text{Var}\left(\frac{X}{n}\right) \\ &= \frac{1}{n^2}\,\text{Var}(X) \quad \text{(using a formula from page 8 in Chapter 20)} \\ &= \frac{1}{n^2}\,np(1-p) \\ &= \frac{p(1-p)}{n}\,, \end{aligned}$$

and so the standard deviation for $\hat{P}$ is given by

$$\begin{aligned}\sigma &= \sqrt{\mathrm{Var}(\hat{P})} \\ &= \sqrt{\frac{p(1-p)}{n}}\,.\end{aligned}$$

From this we see that, essentially,

we can make the standard deviation of $\hat{P}$ as small as we wish,

by choosing the sample size n to be large enough

(since if n is large then this σ is small). That is,

a consequence of having a large sample size n

is that most of the values of $\hat{p}$ are close to the long-run average $E(\hat{P})$.

Note though that, even when we talk about large sample sizes, we will continue to assume that the population size is at least ten times the size of the sample.

We have just seen that if we want to obtain a good estimate for an unknown population proportion p from a sample proportion $\hat{p}$, then choosing a *large sample size* will increase our confidence that a value of $\hat{p}$ is a good point estimate for p. Unfortunately however

often we have neither the money or time to consider a very large sample.

Instead, of just considering the actual values of $\hat{p}$ obtained from our samples, it turns out to be helpful to obtain

interval estimates for p from those $\hat{p}$-values.

This will be our goal for the rest of this chapter.

Exercises for Section 25.3

1. Suppose that the proportion of adults in Australia that have an iPhone is $p = 0.4$. A student takes seven random samples of size 100 from the population of adults in Australia, and from those seven samples the student obtains the following seven sample proportions $\hat{p}$ of adults that have an iPhone:

$$0.44,\ 0.35,\ 0.43,\ 0.45,\ 0.44,\ 0.41,\ 0.34$$

   (a) Calculate the average and standard deviation of these proportions, using the ideas from Chapter 16 (and/or using the statistical functions on your calculator).

   Write your answers to two decimal places.

   (b) Compare the average you found in (a) with $E(\hat{P}) = p$.

   (c) Calculate, to four decimal places, $\sigma = \sqrt{\mathrm{Var}(\hat{P})} = \sqrt{\frac{p(1-p)}{n}}$.

   Compare σ with the standard deviation you found in (a).

   (d) Suppose that we do not know the population proportion p. We can estimate the population standard deviation $\sigma = \sqrt{\frac{p(1-p)}{n}}$ by substituting $\hat{p}$ for p. Calculate, to two decimal places,

   (i) $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ when $\hat{p} = 0.45$

   (ii) $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ when $\hat{p} = 0.34$.

   Note : $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ does not vary greatly for different values of $\hat{p}$.

25.4 Distribution of the random variable $\hat{P}$

Recall that the random variable $\hat{P}$, whose possible values are the values of the sample proportion, is given by

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

in which X is the number of 'successes' in a random sample of size n.

By approximating the distribution of X with a binomial distribution (which is appropriate as long as the sample size n is less than one tenth of the population size), we have established two measures for $\hat{P}$:

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

In these statements, p is the population parameter, which is usually unknown.

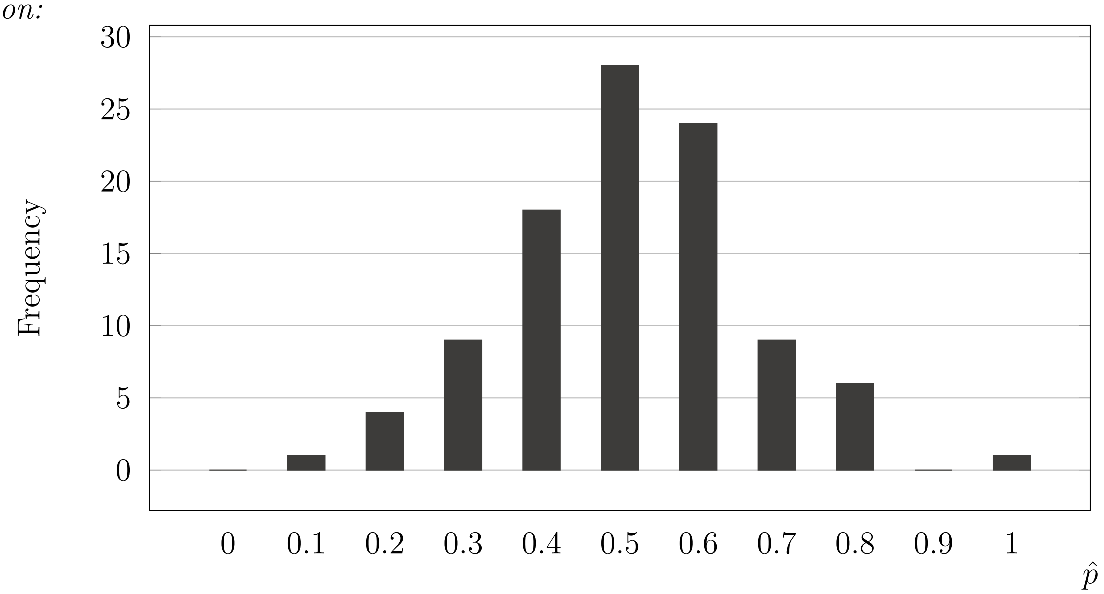
We could calculate the probabilities for $\hat{P}$ by using the binomial distribution. There is, however, a more convenient distribution.

Example 8. In Example 7, we considered the following 100 values of $\hat{p}$ from 100 random samples of size 10:

$\hat{p}$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
frequency	0	1	4	9	18	28	24	9	6	0	1

Make a frequency diagram for these 100 values of $\hat{p}$.

Solution:



□

Notice that the diagram has approximately the same shape as a normal curve.
That is, $\hat{P}$ seems to have – approximately – the normal distribution.
Indeed, we have the following result:

$$\boxed{\text{If } n \text{ is large, then the distribution of } \hat{P} \text{ is approximately normal.}}$$

A justification (which is not examinable) of this result is given on the next page. It relies on an important theorem in probability called the **Central Limit Theorem** which says that:

the *mean* of a large number of independent[1] identically-distributed variables
approximately follows a normal distribution,
even if the original variables have some other distribution.

[1]We have already seen (on page 10 of Chapter 19) what it means for *events* to be **independent**. Similarly we can define that two *random variables* X and Y are **independent** if

$$\Pr(X = x \text{ and } Y = y) \;=\; \Pr(X = x) \times \Pr(Y = y) \quad \text{for all values } x \text{ and } y\,.$$

Independent random variables X and Y satisfy

- $\Pr(Y = y \mid X = x) \;=\; \Pr(Y = y)$ for all values x and y, and
- $E(XY) \;=\; E(X)\,E(Y)$, and
- $\mathrm{Var}(XY) \;=\; \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Result: If n is large then the distribution of $\hat{P}$ is approximately normal.

^^^

Justification (non-examinable):

Provided that the population is large compared with the sample size, a random sample of n items from a population is equivalent to n random samples each containing one item.

Define n variables $X_1, X_2, \ldots, X_n$ to be the number of successes in the first, second, ..., n^{th} sample. Since each sample contains just one item, each variable can only equal 0 or 1. Each of these variables is independent, and is identically-distributed, because it is a random sample of one item from a population.

Let X be, as before, a random variable for the number of successes in a random sample of n items. Then

$$X = X_1 + X_2 + \ldots + X_n$$

$$\text{and so} \quad \hat{P} = \frac{X}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

which shows that $\hat{P}$ is the *mean* of the n random variables $X_1, X_2, \ldots, X_n$. Therefore, we can apply the Central Limit Theorem to $\hat{P}$, which enables us to conclude that $\hat{P}$ approximately follows a normal distribution, as claimed.

^^^

Combining results from this section and the previous section gives the following:

For a large sample size n (but still less than one tenth of the population size), the distribution of the sample proportion variable $\hat{P}$ satisfies:

$$\boxed{\hat{P} \sim N\left(p, \frac{p(1-p)}{n}\right) \text{ approximately.}}$$

In this, the interpretation of 'large' varies. For this chapter, 'n is large' will be valid if

$$\boxed{n\hat{p} > 15} \quad \text{and} \quad \boxed{n(1-\hat{p}) > 15.}$$

(but with n still being less than a tenth of the population size).

25.5 Confidence intervals

We will study another result from probability and then return to studying sample proportions.

Example 9. Suppose Y is normally-distributed with mean 20 and standard deviation 3.

(a) Calculate the probability that a randomly-chosen value of Y will be

within two standard deviations of the mean.

Give your answer to four decimal places.

(b) If the probability that a randomly-chosen value of Y will be

within z_k standard deviations of the mean

is 0.8558, then find z_k . Give your answer to two decimal places.

Solution:

We have $Y \sim N(20\,,\,9)$. Then $\sigma = \sqrt{9} = 3$ and so $Z = \dfrac{Y-20}{3} \sim N(0,1)$.

(a)
$$\begin{aligned}
\Pr(\mu - 2\times\sigma < Y < \mu + 2\times\sigma) &= \Pr(20 - 2\times 3 < Y < 20 + 2\times 3) \\
&= \Pr(-2 < \frac{Y-20}{3} < 2) \\
&= \Pr(-2 < Z < 2) \\
&= \Pr(Z < 2) - \Pr(Z < -2) \\
&= 0.9772 - 0.0228 \quad \text{(from standard normal tables)} \\
&= 0.9544\ .
\end{aligned}$$

(b) **Method 1:**

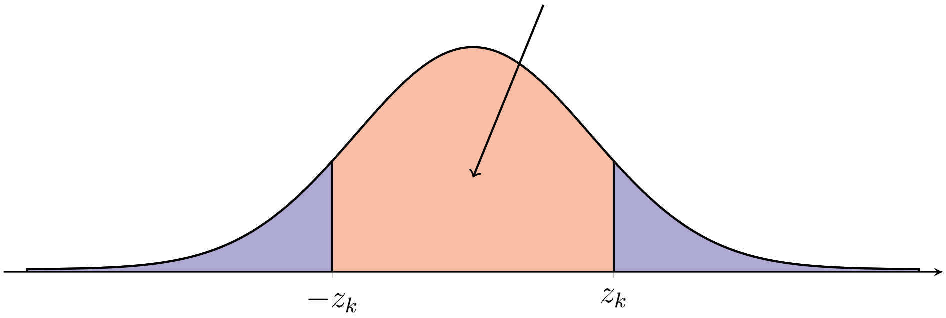
$$\begin{aligned}
\text{We seek } z_k \text{ for which } \Pr(\mu - z_k\times\sigma < Y < \mu + z_k\times\sigma) &= 0.8558. \\
\text{That is, } \Pr(20 - z_k\times 3 < Y < 20 + z_k\times 3) &= 0.8558. \\
\text{That is, } \Pr(-z_k < \frac{Y-20}{3} < z_k) &= 0.8558 \\
\text{That is, } \Pr(-z_k < Z < z_k) &= 0.8558 \\
\Pr(Z < z_k) - \Pr(Z < -z_k) &= 0.8558 \\
\Pr(Z < z_k) - \Pr(Z > z_k) &= 0.8558 \\
\Pr(Z < z_k) - [\,1 - \Pr(Z < z_k)\,] &= 0.8558 \\
2\Pr(Z < z_k) &= 1.8558 \\
\Pr(Z < z_k) &= 0.9279 \\
\text{from which } z_k &= 1.46\ .
\end{aligned}$$

Alternatively, we can find this answer quite easily by considering the normal curve (and its symmetry), as shown next.

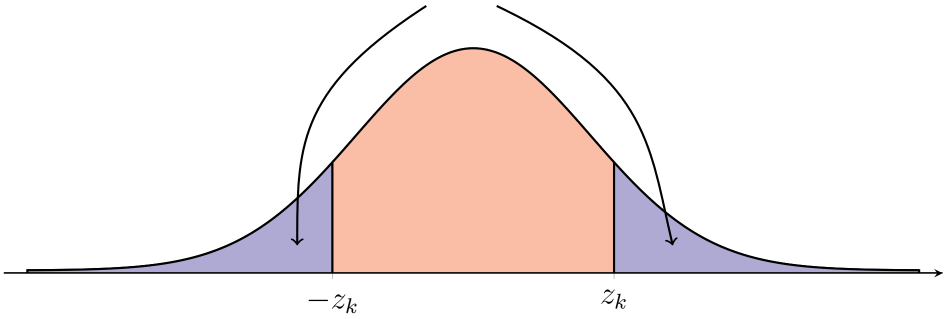
Method 2: As we saw in Method 1, we seek z_k such that

$$\Pr(-z_k < Z < z_k) = 0.8558 \text{ , where } Z = \frac{Y - 20}{3} \sim N(0,1) \text{ .}$$

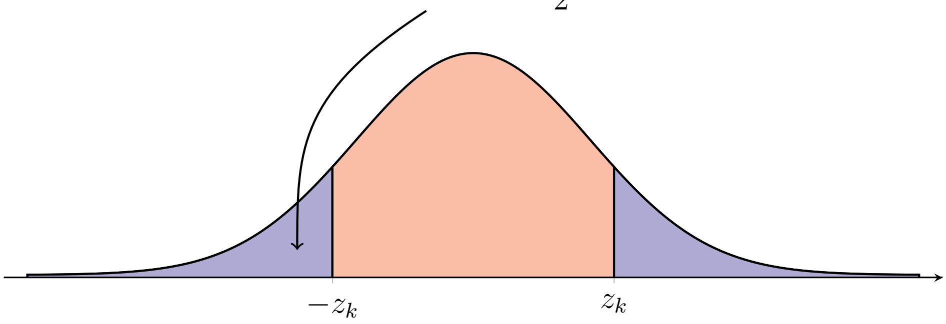
That is, we want the area of the middle region to be 0.8558 .



Thus the combined area of the two tails $= 1 - 0.8558 = 0.1442$,



and so the area of the left tail $= \dfrac{1}{2} \times 0.1442 = 0.0721$.



This third diagram shows us that $\Pr(Z < -z_k) = 0.0721$.
Then, by using standard normal tables, we find that

$$-z_k = -1.46,$$

and so we conclude that $\qquad z_k = 1.46$

(which is the same answer as obtained previously). $\square$

Let's look at the ideas from Example 9 more generally.

Suppose $Y \sim N(\mu\,, \sigma^2)$.

Let k and z_k be two numbers for which the probability that a value of Y is *within* z_k *standard deviations of the mean* is $\dfrac{k}{100}$. That is, suppose that k and z_k are related by

$$\Pr(\mu - z_k\,\sigma < \ Y \ < \mu + z_k\,\sigma) \ = \ \frac{k}{100} \ .$$

Notice that (in contrast to some later working) this statement concerns a *variable* (namely Y) in an interval with *fixed* endpoints.

In a twist, let's express this statement in an unusual format:

$$\begin{aligned}
\Pr(\mu - z_k\,\sigma < Y < \mu + z_k\,\sigma) \ &= \ \Pr(-z_k\,\sigma \ < \ Y - \mu \ < \ z_k\,\sigma) \\
&= \ \Pr(-Y - z_k\,\sigma \ < \ -\mu \ < \ -Y + z_k\,\sigma) \\
&= \ \Pr(Y + z_k\,\sigma \ > \ \mu \ > \ Y - z_k\,\sigma) \\
&= \ \Pr(Y - z_k\,\sigma \ < \ \mu \ < \ Y + z_k\,\sigma) \ = \ \frac{k}{100} \ .
\end{aligned}$$

Caution: Notice that the last line of this working now has μ (which is a *constant*) in the centre of an interval with *varying* endpoints (since those endpoints depend on the variable Y) .

Suppose that we calculate endpoints of this interval by *using a particular value of the variable* Y. Then, the mean μ is either

- *within* the calculated interval, in which case we have

$$\Pr(\mu \text{ is in the calculated interval}) \ = \ 1\,,$$

or else it is

- *not* within the calculated interval, in which case we have

$$\Pr(\mu \text{ is in the calculated interval}) \ = \ 0\,.$$

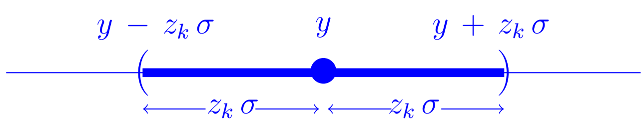
In particular, we do **not** have

$$\Pr(\mu \text{ is in the calculated interval}) \ = \ \frac{k}{100} \ .$$

We will return to this important idea repeatedly.

The unusual format above leads to the definition of a **confidence interval for the mean**. For a variable Y assumed to have a normal distribution, and for a value y, a **k% confidence interval for the mean** is defined to be

$$\boldsymbol{(y - z_k\,\sigma\,,\, y + z_k\,\sigma)}.$$



Popular percentages for confidence intervals are 90, 95, and 99. That is, 90, 95 and 99 are three popular values of k. We can use standard normal tables to calculate the values of z_k for these values of k. (See the exercises at the end of Section 25.6.) These pairings of k and z_k are summarised in the table below, which is included on the Formula Sheet.

k	90	95	99
z_k	1.65	1.96	2.58

Notice that as k increases then z_k also increases; this results in a wider confidence interval.

Example 10. Consider a random variable Y such that $Y \sim N(\mu, 0.09)$. Suppose that a random sample yields the value $y = 0.7$. Use this y–value to find an 85.58% confidence interval for the mean.

Solution: Since $Y \sim N(\mu, 0.09)$, then $\sigma = \sqrt{0.09} = 0.3$.

Also, we are given the value $y = 0.7$.

Finally, we have seen in Example 9(b) that the given k-value of 85.58 corresponds to $z_k = 1.46$.

Thus, an 85.58% confidence interval for the mean is

$$\begin{aligned}(y - z_k\,\sigma\,,\, y + z_k\,\sigma) &= (0.7 - 1.46 \times 0.3\,,\, 0.7 + 1.46 \times 0.3)\\ &= (0.262\,,\, 1.138).\end{aligned}$$

□

Example 11. Suppose Y is normally-distributed with standard deviation 9. Given the value $y = 62$, determine a 90% confidence interval for the mean.

Solution: We have $\sigma = 9$ and $y = 62$. Also, since $k = 90$, we immediately know (from the table at the top of this page) that $z_k = 1.65$.

Thus a 90% confidence interval for the mean is

$$\begin{aligned}(y - z_k\,\sigma\,,\, y + z_k\,\sigma) &= (62 - 1.65 \times 9\,,\, 62 + 1.65 \times 9)\\ &= (47.15\,,\, 76.85).\end{aligned}$$

□

Caution: We definitely should *not* write that $\Pr(47.15 < \mu < 76.85) = 0.9$.
The mean is *not* a random variable; it is a constant. Suppose that, for this example, the mean is actually 40. Then this mean is *not* in the confidence interval that we found, and thus we have $\Pr(47.15 < \mu < 76.85) = \mathbf{0}$.
Alternatively, if the mean is actually 50, then it *is* in the confidence interval we found, in which case we have $\Pr(47.15 < \mu < 76.85) = \mathbf{1}$.

We should interpret a 90% confidence interval as follows:

If we repeat the sampling experiment a large number of times,
and calculate a 90% confidence interval each time,
then 90% of these confidence intervals will contain the mean.

25.6 Confidence intervals for population proportion

We are finally ready to consider the task of inferring a value for an unknown population proportion p from a sample proportion $\hat{p}$. In particular, we will

construct an interval estimate for p *from a value of* $\hat{p}$.

Let X be the random variable for the number of 'successes' in a random sample of size n. Then (as usual) $\hat{P} = \dfrac{X}{n}$ is the random variable for the values of the sample proportion. We've already seen that if n is large (but still less than a tenth of the population size), then

$$\hat{P} \sim N\left(p\,,\ \frac{p\,(1-p)}{n}\right) \quad \text{approximately.}$$

(Recall that, for this chapter, n is 'large' if $n\,\hat{p} > 15$ and $n\,(1-\hat{p}) > 15$.)

Also, recall that, for a normally distributed variable Y, and for a value y, a **k% confidence interval for the mean** was defined to be

$$\boldsymbol{(y - z_k\,\sigma\,,\ y + z_k\,\sigma)}.$$

Therefore, since $\hat{P}$ has the normal distribution (approximately), and since p is the mean of $\hat{P}$, we can use a value of the sample proportion $\hat{p}$ to write a **k% confidence interval for the population proportion p** as

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

in which z_k is found via $\Pr(-z_k < Z < z_k) = \dfrac{k}{100}$ for $Z \sim N(0,1)$.

We need to make one last approximation:

This confidence interval formula written above relies on the value of p, but we want to use it when we *do not know* the value of p. Therefore, when we have taken a sample from the population, and used it to calculate a sample proportion $\hat{p}$, we use this value of $\hat{p}$ as an *approximation* for p in the confidence interval.

That is, for a sample proportion $\hat{p}$, an ***approximate*** **k% confidence interval for the unknown population proportion p** is

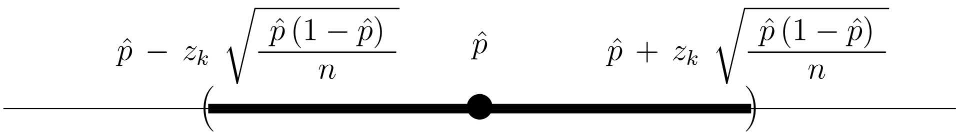
$$\boxed{\boldsymbol{\left(\hat{p} - z_k\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\ ,\ \hat{p} + z_k\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\ \right)}}$$

in which z_k is found via $\Pr(-z_k < Z < z_k) = \dfrac{k}{100}$ for $Z \sim N(0,1)$.

Notice that to find an approximate $k\%$ confidence interval for p, using

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

we need to know values of $\hat{p}$, z_k and n. Furthermore, notice that this confidence interval is centred on $\hat{p}$.



Example 12. Suppose that we select a random sample of 200 students from the University of Melbourne and find that 120 of the students have an iPad. Find an approximate 95% confidence interval for the proportion p of students from the University of Melbourne that have an iPad, writing the answer to four decimal places.

Solution: The sample size is $\boldsymbol{n = 200}$, and the sample proportion for students with an iPad is $\boldsymbol{\hat{p} = \frac{120}{200} = 0.6}$. We check that n is large enough to approximate the distribution with the normal distribution. We have

$$n\hat{p} = 200 \times 0.6 = 120 \quad \text{and} \quad n(1-\hat{p}) = 200 \times 0.4 = 80.$$

Since $n\hat{p}$ and $n(1-\hat{p})$ are both greater than 15, we conclude that it is appropriate to use a z_k -value obtained from the normal tables (or from the table of z_k -values provided on the Formula Sheet).

In this example, we want a 95% confidence interval, so we have $k = 95$. Looking at the table of z_k -values provided on the Formula Sheet, we see that $\boldsymbol{z_k = 1.96}$.

Thus an approximate 95% confidence interval for p from $\hat{p} = 0.6$ is

$$\begin{aligned}
&\left(\hat{p} - z_k\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\ ,\ \hat{p} + z_k\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right) \\
=\ &\left(0.6 - 1.96\sqrt{\frac{0.6 \times 0.4}{200}}\ ,\ 0.6 + 1.96\sqrt{\frac{0.6 \times 0.4}{200}}\right) \\
=\ &(0.6 - 0.0679\,,\ 0.6 + 0.0679) \quad \text{(to 4 d.p.)} \\
=\ &(0.5321\,,\ 0.6679) \quad \text{(to 4 d.p.)}.
\end{aligned}$$

□

Example 13. A biased die is tossed 300 times. Suppose that it shows a six 30 times. Find an approximate 95% confidence interval for the 'population proportion' p of times that the die will show a six. That is, find an approximate 95% confidence interval for the probability p that the die will show a six. Write the answer to four decimal places.

Solution: The sample size is $\boldsymbol{n} = \mathbf{300}$, and the sample proportion (for the throws that show a six) is $\hat{\boldsymbol{p}} = \frac{\mathbf{30}}{\mathbf{300}} = \mathbf{0.1}$. We check that n is large enough to approximate the distribution with the normal distribution. We have

$$n\hat{p} = 300 \times 0.1 = 30 \quad \text{and} \quad n(1-\hat{p}) = 300 \times 0.9 = 270.$$

Since $n\hat{p}$ and $n(1-\hat{p})$ are both greater than 15, we conclude that it is appropriate to use a z_k-value obtained from the normal tables (or from the table of z_k-values provided on the Formula Sheet).

In this example, we want a 95% confidence interval, so (by looking at the table of z_k-values provided on the Formula Sheet), we see that $\boldsymbol{z_k} = \mathbf{1.96}$.

Thus an approximate 95% confidence interval for p from $\hat{p} = 0.1$ is

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

$$= \left(0.1 - 1.96\sqrt{\frac{0.1 \times 0.9}{300}}\ ,\ 0.1 + 1.96\sqrt{\frac{0.1 \times 0.9}{300}}\right)$$

$$= (0.1 - 0.0339\,,\ 0.1 + 0.0339) \quad \text{(to 4 d.p.)}$$

$$= (0.0661\,,\ 0.1339) \quad \text{(to 4 d.p.)}.$$

□

Caution: We *do not know* whether the probability p is in this interval. Either

- it is, in which case $\Pr(p \text{ is in this interval}) = \mathbf{1}$,

or else

- it is not, in which case $\Pr(p \text{ is in this interval}) = \mathbf{0}$.

In particular, we do **not** have $\Pr(p \text{ is in this interval}) = 0.95$.

The way to interpret this approximate 95% confidence interval is:

Suppose we repeated the experiment a large number of times,
obtaining a sample proportion each time,
and calculated a 95% *confidence interval for* p *from each sample proportion* $\hat{p}$*,*
then approximately 95% *of those confidence intervals would contain* p*.*

Example 14. Suppose that, in the year 2000, the population proportion of females in Australia was 0.52 . Last year, the public service selected a random sample of 1000 people from Australia, and found that 492 of the people in the sample were female.

Calculate, to three decimal places,

(a) an approximate 90% confidence interval (b) an approximate 99% confidence interval

for last year's proportion of females in Australia.

Solution: The sample size is $\boldsymbol{n = 1000}$, and the sample proportion of females is $\boldsymbol{\hat{p} = \frac{492}{1000} = 0.492}$. We check that n is large enough to approximate the distribution with the normal distribution. We have

$$n\hat{p} = 1000 \times 0.492 = 492 \quad \text{and} \quad n(1-\hat{p}) = 1000 \times 0.508 = 508.$$

Since $n\hat{p}$ and $n(1-\hat{p})$ are both greater than 15, we conclude that it is appropriate to use a z_k -value obtained from the normal tables (or from the table of z_k -values provided on the Formula Sheet).

(a) We want a 90% confidence interval, so (by looking at the table of z_k -values provided on the Formula Sheet), we see that $\boldsymbol{z_k = 1.65}$.

Thus an approximate 90% confidence interval for p from $\hat{p} = 0.492$ is

$$\begin{aligned}
&\left(\hat{p} - z_k\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\;,\; \hat{p} + z_k\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right) \\
&= \left(0.492 - 1.65 \times \sqrt{\frac{0.492 \times 0.508}{1000}}\;,\; 0.492 + 1.65 \times \sqrt{\frac{0.492 \times 0.508}{1000}}\right) \\
&= (0.466\,,\,0.518) \quad \text{(to 3 d.p.)}.
\end{aligned}$$

(b) This time, we want a 99% confidence interval, so (by looking at the table of z_k -values provided on the Formula Sheet), we see that $\boldsymbol{z_k = 2.58}$.

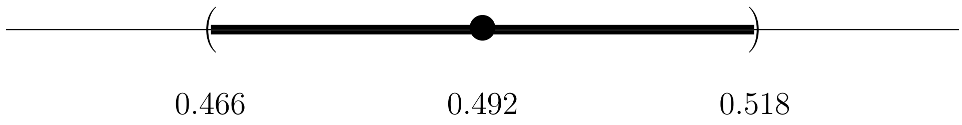
Thus an approximate 99% confidence interval for p from $\hat{p} = 0.492$ is

$$\begin{aligned}
&\left(\hat{p} - z_k\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\;,\; \hat{p} + z_k\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right) \\
&= \left(0.492 - 2.58 \times \sqrt{\frac{0.492 \times 0.508}{1000}}\;,\; 0.492 + 2.58 \times \sqrt{\frac{0.492 \times 0.508}{1000}}\right) \\
&= (0.451\,,\,0.533) \quad \text{(to 3 d.p.)}.
\end{aligned}$$

$\square$

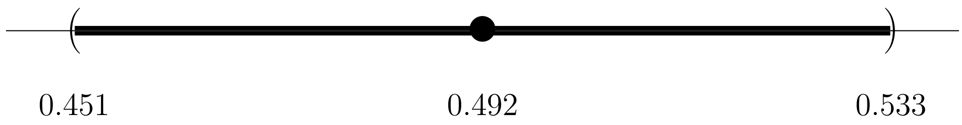
In the previous example, we used the $\hat{p}$-value of 0.492 to find

- an approximate **90%** confidence interval of $(0.466, 0.518)$,



and

- an approximate **99%** confidence interval of $(0.451, 0.533)$.



Notice that increasing the k-value from 90 to 99 resulted in a *wider* confidence interval. This observation should seem entirely sensible, since we would expect that more of the wider intervals will contain p (compared to the narrower intervals).

Exercises for Sections 25.5 and 25.6

1. Suppose Y is a normal variable. Let k and z_k be two numbers for which the probability that a value of Y is within z_k standard deviations of the mean is $\frac{k}{100}$.

   Use one of the methods from Example 9(b) to find the value z_k for which the probability that a value of Y is within z_k standard deviations of the mean is 95%. Repeat this for 90% and for 99%, making sure that your values of z_k allow the mentioned probability.

2. A survey company takes a random sample of $1\,600$ people from the adults of a large city. Among the people surveyed, it was found that 360 were unemployed. Find an approximate 95% confidence interval for the proportion p of unemployed adults in the city. Write your answer to four decimal places.

3. A biased coin is tossed 1000 times. Suppose that it lands "heads" 526 times. Find an approximate 99% confidence interval for the 'population proportion' p of times that it will show heads. That is, find an approximate 99% confidence interval for the probability p that the coin will show heads. Write your answer to four decimal places.

4. Suppose that the proportion of people in Australia that have a driving licence is 0.3 . A student takes four random samples of size 100 from the population of all people in

Australia, and finds the following four sample proportions $\hat{p}$ of people that have a driving licence:

$$0.32,\ 0.24,\ 0.36,\ 0.39\,.$$

Calculate (to four decimal places) an approximate 90% confidence interval for the proportion p of people with a driving licence, based on

(a) $\hat{p} = 0.32$ (b) $\hat{p} = 0.24$

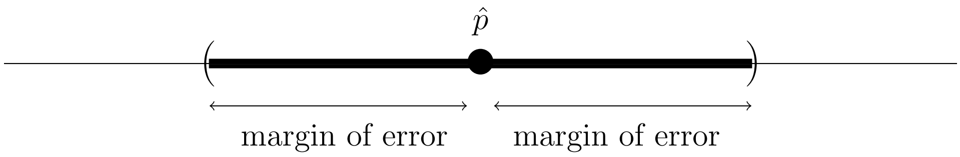
(c) $\hat{p} = 0.36$ (d) $\hat{p} = 0.39$.

Notice that sometimes a confidence interval does not contain the true value of $p = 0.3$. In fact, we would expect approximately 90% of the approximate 90% confidence intervals to contain p .

25.7 Margin of error

When we take a random sample of size n from a population, calculate the sample proportion $\hat{p}$ and use it to find a confidence interval for the unknown population proportion p, the **margin of error** is

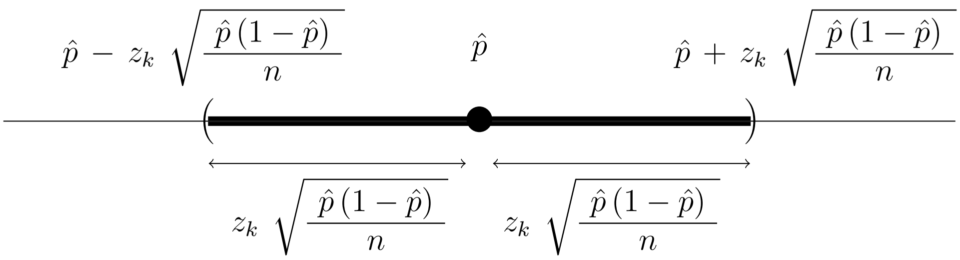
the distance between the point estimate $\hat{p}$ and an endpoint of the confidence interval.



Recall that an approximate k% confidence interval for p is given by

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

We have



Thus we see that the **margin of error for the approximate $k\%$ confidence interval** is

$$\boldsymbol{z_k \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}} \,.$$

Example 15. Suppose that we select a random sample of 200 people from Melbourne and find that 90 of the people use Facebook. Find the margin of error for an approximate 99% confidence interval for the proportion of people in Melbourne that use Facebook. Use two decimal places in the answer.

Solution: We have $\boldsymbol{n = 200}$ and $\boldsymbol{\hat{p} = \frac{90}{200} = 0.45}$.

We check that n is large enough to approximate the distribution with the normal distribution. We have

$$n\hat{p} = 200 \times 0.45 = 90 \quad \text{and} \quad n(1-\hat{p}) = 200 \times 0.55 = 110.$$

Since $n\hat{p}$ and $n(1-\hat{p})$ are both greater than 15, we conclude that it is appropriate to use a z_k-value obtained from the normal tables (or from the table of z_k-values provided on the Formula Sheet).

We want a 99% confidence interval, so (by looking at the table of z_k-values provided on the Formula Sheet), we see that $\boldsymbol{z_k = 2.58}$. Thus the margin of error is

$$z_k\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \;=\; 2.58\sqrt{\frac{0.45 \times 0.55}{200}} \;=\; 0.09 \quad \text{(to 2 d.p.)}. \qquad \square$$

Recall that the margin of error for an approximate k% confidence interval is $z_k\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$. Notice that this depends on three values, namely z_k, $\hat{p}$ and n. Let's briefly consider how varying these values will affect the margin of error:

- As seen in Exercise 1(d) on page 12, changing $\hat{p}$ has a *negligible* effect on the margin of error.
- Clearly, *increasing* z_k (which is equivalent to *increasing* k) leads to an *increased* margin of error (which corresponds to a *wider* confidence interval). This matches the observation made on page 24. Conversely, *reducing* z_k (and k) leads to a *smaller* margin of error (and a *narrower* confidence interval).
  Generally, a smaller margin of error is considered to be desirable, as a narrower interval gives a more focussed estimate of p. Unfortunately however, it is not very helpful to achieve this by considering a smaller value of k, since a lower k-value means that fewer of the confidence intervals are likely to actually contain p. Fortunately there is an alternative way of reducing the margin of error, as considered next.
- Consider a fixed value of k (thus the value of z_k is also fixed). Notice that

  $$\text{as } n \text{ } \textit{increases} \text{ then the margin of error } \; z_k\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \;\; \textit{decreases},$$

  giving a *narrower* confidence interval (and a more focussed estimate of p).

Example 16. A news website is interested in finding a 99% confidence interval for the proportion p of voters who intend to vote for Liberal in an upcoming election.

(a) Initially, the website selects a small random sample of 100 voters from Australia and finds that 40 of the voters intend to vote for Liberal in the upcoming election. Find, to three decimal places, an approximate 99% confidence interval for the proportion p of Liberal voters in Australia.

(b) The website wants to have a margin of error of at most 0.01 in the approximate 99% confidence interval. Using the sample proportion found in (a), find the smallest sample size n required.

Solution: (a) We have $\boldsymbol{n = 100}$ and $\boldsymbol{\hat{p} = \frac{40}{100} = 0.4}$. Notice that

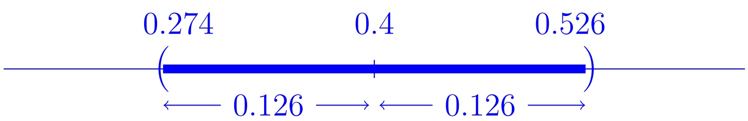
$$n\hat{p} = 100 \times 0.4 = 40 \quad \text{and} \quad n(1-\hat{p}) = 100 \times 0.6 = 60.$$

Since $n\hat{p}$ and $n(1-\hat{p})$ are both greater than 15, then it is appropriate to use $\boldsymbol{z_k = 2.58}$ (obtained from the table of z_k-values provided on the Formula Sheet). Thus an approximate 99% confidence interval for the proportion p of Liberal voters is given by

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

$$= \left(0.4 - 2.58\sqrt{\frac{0.4 \times 0.6}{100}}\,,\ 0.4 + 2.58\sqrt{\frac{0.4 \times 0.6}{100}}\right)$$

$$= (0.274\,,\ 0.526) \quad \text{(to 3 d.p.)}.$$



(Notice that the margin of error for this confidence interval is 0.126.)

(b) Suppose the website randomly chooses a sample of n voters from Australia. Then the margin of error is

$$z_k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2.58\sqrt{\frac{0.4 \times 0.6}{n}} = 2.58\sqrt{\frac{0.24}{n}}\,.$$

Since the website wants the margin of error to be at most 0.01, we have

$$\begin{aligned}
2.58\sqrt{\frac{0.24}{n}} &\leq 0.01\,. \\
\text{Then}\ \ (2.58)^2 \times \frac{0.24}{n} &\leq (0.01)^2, \\
\text{from which}\ \ n &\geq \frac{(2.58)^2 \times 0.24}{(0.01)^2}\,. \\
\text{That is,}\ \ n &\geq 15\,975.36\,.
\end{aligned}$$

Thus the smallest suitable sample size is $15\,976$.

□

Exercises for Section 25.7

1. A random sample of $1\,000$ Australian voters is taken to estimate the proportion of Australian voters that support the Greens. It turns out that 240 of the people in the sample support the Greens.

   (a) Find the sample proportion of voters that support the Greens.

   (b) Find the margin of error for an approximate 95% confidence interval for the proportion of voters that support the Greens (obtained by considering this particular sample). Answer to 3 decimal places.

   (c) Use the sample proportion found in (a) to find an approximate 95% confidence interval for the proportion of voters that support the Greens. Answer to 3 decimal places.

2. A mobile phone manufacturing company has received many complaints from its customers, and has had to replace or repair many returned mobile phones. The company wants to find an approximate 90% confidence interval for the proportion of defective phones that are produced by its factory.

   (a) A random sample of 100 phones from the factory are tested and it is found that 20 of the phones in the sample are defective. Find an approximate 90% confidence interval for the proportion of defective phones that are produced by the factory.

   (b) The manufacturer wants to have a margin of error of at most 0.02 in the approximate 90% confidence interval. Using the sample proportion found in (a), find the sample size n required.

25.8 Answers for the Chapter 25 Exercises

25.2 1. No. He needs to choose people randomly from all over Australia.

2. $\hat{p} = \frac{26}{100} = 0.26$

3. (a) $p = \frac{1}{4} = 0.25$

   (b) $\hat{p} = \frac{3}{10} = 0.3$

4. Define X to be the number of people in the sample who read newspapers daily, and define $\hat{P} = \dfrac{X}{5}$. The possible values of X are 0, 1, 2, 3, 4, 5 and the possible values of $\hat{P}$ are 0, $\frac{1}{5}$, $\frac{2}{5}$, $\frac{3}{5}$, $\frac{4}{5}$, 1.

5. (a)

x	$\hat{p}$	$\Pr(X = x) = \Pr(\hat{P} = \hat{p})$
0	0	$\dfrac{{}^5\mathrm{C}_0 \times {}^{45}\mathrm{C}_4}{{}^{50}\mathrm{C}_4} = 0.64696$
1	$\frac{1}{4}$	$\dfrac{{}^5\mathrm{C}_1 \times {}^{45}\mathrm{C}_3}{{}^{50}\mathrm{C}_4} = 0.30808$
2	$\frac{1}{2}$	$\dfrac{{}^5\mathrm{C}_2 \times {}^{45}\mathrm{C}_2}{{}^{50}\mathrm{C}_4} = 0.04299$
3	$\frac{3}{4}$	$\dfrac{{}^5\mathrm{C}_3 \times {}^{45}\mathrm{C}_1}{{}^{50}\mathrm{C}_4} = 0.00195$
4	1	$\dfrac{{}^5\mathrm{C}_4 \times {}^{45}\mathrm{C}_0}{{}^{50}\mathrm{C}_4} = 0.00002$

(b) Yes, since $n < \dfrac{\text{the size of the population}}{10}$.

x	$\hat{p}$	$\Pr(X = x) = \Pr(\hat{P} = \hat{p})$
0	0	${}^4\mathrm{C}_0\,(0.1)^0\,(0.9)^4 = 0.6561$
1	$\frac{1}{4}$	${}^4\mathrm{C}_1\,(0.1)^1\,(0.9)^3 = 0.2916$
2	$\frac{1}{2}$	${}^4\mathrm{C}_2\,(0.1)^2\,(0.9)^2 = 0.0486$
3	$\frac{3}{4}$	${}^4\mathrm{C}_3\,(0.1)^3\,(0.9)^1 = 0.0036$
4	1	${}^4\mathrm{C}_4\,(0.1)^4\,(0.9)^0 = 0.0001$

25.3 1. (a) The average is 0.41 (2 d.p.) and the standard deviation is 0.05 (2 d.p.).

(b) $E(\hat{P}) = p = 0.4$

The average found in (a) is close to this.

(c) $\sigma = 0.0490$ (4 d.p.)

The standard deviation found in (a) is close to this.

(d) (i) $\sqrt{\dfrac{0.45 \times 0.55}{100}} = 0.05$ (2 d.p.) (ii) $\sqrt{\dfrac{0.34 \times 0.66}{100}} = 0.05$ (2 d.p.)

25.6 1. The value of z_k for 95% is $z_k = 1.96$.

For 90%, use $z_k = 1.65$ (rather than 1.64).

For 99%, use $z_k = 2.58$ (rather than 2.57).

2. (0.2045 , 0.2455)

3. (0.4853 , 0.5667)

4. (a) (0.2430 , 0.3970) (b) (0.1695 , 0.3105)

(c) (0.2808 , 0.4392) (d) (0.3095 , 0.4705)

Notice that the confidence interval in (d) does not contain $p = 0.3$.

25.7 1. (a) 0.24 (b) 0.026 (c) (0.214 , 0.266)

2. (a) (0.134 , 0.266) (b) We need $n \geq 1\,089$.