Supporting Australian Mathematics Project

A guide for teachers – Years 11 and 12

Probability and statistics: Module 24

Inference for proportions

Education Services Australia

AMSI
AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Inference for proportions – A guide for teachers (Years 11–12)

Dr Sue Finch, University of Melbourne
Professor Ian Gordon, University of Melbourne

Editor: Dr Jane Pitkethly, La Trobe University

Illustrations and web design: Catherine Tan, Michael Shaw

Full bibliographic details are available from Education Services Australia.

Published by Education Services Australia
PO Box 177
Carlton South Vic 3053
Australia

Tel: (03) 9207 9600
Fax: (03) 9910 9800
Email: info@esa.edu.au
Website: www.esa.edu.au

This publication is funded by the Australian Government Department of Education, Employment and Workplace Relations.

Supporting Australian Mathematics Project

Australian Mathematical Sciences Institute
Building 161
The University of Melbourne
VIC 3010
Email: enquiries@amsi.org.au
Website: www.amsi.org.au

Inference for proportions

Assumed knowledge

The content of the modules:

- *Discrete probability distributions*
- *Binomial distribution*
- *Exponential and normal distributions*
- *Random sampling.*

Motivation

- Why can we rely on random samples to provide information about the proportion of a population with a particular characteristic?
- Should we worry that different random samples taken from the same population will give different results?
- How variable are the results obtained from different random samples?
- How can we quantify the uncertainty (imprecision) in the results from a sample?

The module *Random sampling* introduces sampling from a binomial distribution. Underlying the binomial distribution are Bernoulli trials and Bernoulli random variables. A Bernoulli random variable is a discrete random variable that takes the value 1 with probability p and the value 0 with probability $1 - p$. If we have a random sample of n observations on a Bernoulli random variable, then the sum of the observations X has a binomial distribution with parameters n and p. A random sample of n Bernoulli observations thus gives a single observation from the $\text{Bi}(n, p)$ distribution.

The binomial distribution allows us to model sampling from an essentially infinite population of units in which a proportion p of the units have a particular characteristic. If we choose a unit at random from this population, the probability that it has the characteristic is equal to p, and the probability that it does not have the characteristic is equal to $1 - p$. If we choose n units at random, the number X with the characteristic has a binomial distribution: $X \overset{\text{d}}{=} \text{Bi}(n, p)$.

In practice, we usually do not know the value of the population proportion p, and we are interested in obtaining an estimate of p. A single observation x of X can be used to provide a *point estimate* of the unknown population proportion p: the sample proportion $\frac{x}{n}$ is an estimate of the population proportion p. There will be some imprecision associated with a single point estimate, and we would like to quantify this sensibly.

In this module, we discuss the distribution of observations from a binomial distribution to illustrate how it serves as a basis for using a sample proportion to estimate an unknown population proportion p. By considering the approximate distribution of sample proportions, we can provide a quantification of the uncertainty in an estimate of the population proportion. This is a *confidence interval* for the unknown population proportion p.

This provides methods for answering questions like:

- What is our best estimate of the proportion of Australians who plan to vote for Labor in the next federal election?
- What is the uncertainty in this estimate of the proportion of Australians who plan to vote for Labor in the next federal election?
- What is our best estimate of the proportion of physically inactive Australian adults?
- What is the uncertainty in this estimate of the proportion of physically inactive Australian adults?

Content

Using probability theory to make an inference

The module *Binomial distribution* introduces the concept of a binomial random variable. Recall that, if a random variable X has a binomial distribution with parameters n and p, we write $X \overset{\text{d}}{=} \text{Bi}(n, p)$. A binomial random variable can always be thought of as the number of successes in n independent Bernoulli trials, each with probability of success p. For this reason, the binomial distribution is often assumed to be an appropriate model when we count the number of units with a characteristic of interest in a random sample of size n, taken from a population in which the proportion of units with the characteristic is p.

In the module *Binomial distribution*, the value of p is generally assumed to be known. However, in many realistic and relevant research contexts, the value of p is not known, but we are very interested in its value, because it relates to a research question of some importance.

We have often appealed to an argument based on symmetry and appropriate random mixing (such as shaking a die in a cup) to justify particular numerical choices for probabilities: for example, that the chance of rolling a four using a fair die is $\frac{1}{6}$.

But knowing probabilities, or even having a basis for assuming particular values, is not a common scenario. The opposite is the case. We are often confronted with a situation where we believe that a binomial model is appropriate and we know the size n of the random sample of units, but we do not know p. And we would like to know it.

One of the main reasons for studying probability distributions, such as the binomial, is that this theory is the foundation for making inferences about unknown population characteristics, such as p. In general terms, this is known as **statistical inference**.

Here are two quite different contexts with the same underlying binomial structure, where it is clear that p is unknown:

1. A random sample of voters is asked about their current political preference. We are interested in using the sample to draw an inference about the proportion of the population of voters who currently prefer Labor.
2. Consider a standard drawing pin with a circular, slightly rounded head. If this drawing pin is tossed (in a similar way to a normal coin toss) and allowed to land on a flat surface, there are two ways it can finish:
   - leaning on the point of the pin (as shown on the left in figure 1)
   - lying flat with the pin pointing straight up (as shown on the right).

   What is the chance that it finishes with the pin pointing straight up?

Figure 1: The two ways a drawing pin can finish after being tossed like a coin.

It is part of the power of probability models, and their application in statistics, that such diverse problems as these two can be dealt with in the same way.

In this module, we use the binomial structure to think about the following specific inferential problem: If we have an observation from a binomial random variable with known n but unknown p, how can we make an inference about p?

The sample proportion as an estimator of p

Even without using any ideas from probability or distribution theory, it seems compelling that the sample proportion should tell us something about the population proportion. If we have a random sample from the population, the sample is representative of the population in an 'expected' sense. So we should be able to use the sample proportion as an estimate of the population proportion.

Assume that $X \overset{\text{d}}{=} \text{Bi}(n, p)$. We define the **sample proportion** to be $\hat{P} = \dfrac{X}{n}$. It is a suitable name, because $\hat{P}$ reflects the proportion in the sample with the characteristic of interest. Once we obtain an actual observation x of the random variable X, we have an actual observation $\hat{p} = \dfrac{x}{n}$ of the sample proportion.

As there is a distinction to be made between the random variable $\hat{P}$ and its corresponding observed value $\hat{p}$, we refer to the random variable as the **estimator** $\hat{P}$, and the observed value as the **estimate** $\hat{p}$; note the use of upper and lower case.

More specifically, the observed value $\hat{p}$ is referred to as a **point estimate** of p.

Example: Survey of voters

Suppose we obtain a random sample of 500 voters, and we find that 227 prefer Labor. The observed sample proportion preferring Labor is $\frac{227}{500} = 0.454$, and we say that 0.454 is a point estimate of the unknown population proportion preferring Labor.

In this example:

- p is the proportion of all Australian voters who prefer Labor
- $n = 500$ is the sample size
- the random variable X is the *number* of voters who prefer Labor in a random sample of 500 voters
- the random variable $\hat{P} = \frac{X}{500}$ is the *proportion* of voters who prefer Labor in a random sample of 500 voters
- $x = 227$ is an observation of X
- $\hat{p} = \frac{227}{500} = 0.454$ is the corresponding observation of $\hat{P}$.

In the previous example, we would be very lucky if the true population proportion turned out to be 0.454. It is much more probable that this value is different from the true population proportion, because samples vary. After all, even when we *know* the population proportion p, we do not (and should not!) expect the sample proportion to be exactly equal to p. For example, if we toss a fair coin 30 times, then obtaining exactly 15 heads is not guaranteed at all, even if it is one of the more likely outcomes.

This discussion is reminding us that the sample proportion $\hat{P}$ is actually a random variable; it varies from one sample to the next. In the next section, we explore this important fact in some detail.

The sample proportion as a random variable

The module *Random sampling* includes the example of making observations on the Bi(10, 0.5) distribution. This example is motivated by supposing that, in the population of voters, 50% prefer the Labor party, and then looking at what happens for many small samples of 10 voters.

If we count the number of Labor voters in one sample of 10 voters, we have an observation from the binomial distribution with parameters $n = 10$ and $p = 0.5$. If we do this 100 times, we have 100 observations from this binomial distribution. An example of 100 actual observations is shown in figure 2; the number of people preferring Labor is shown as a horizontal bar.

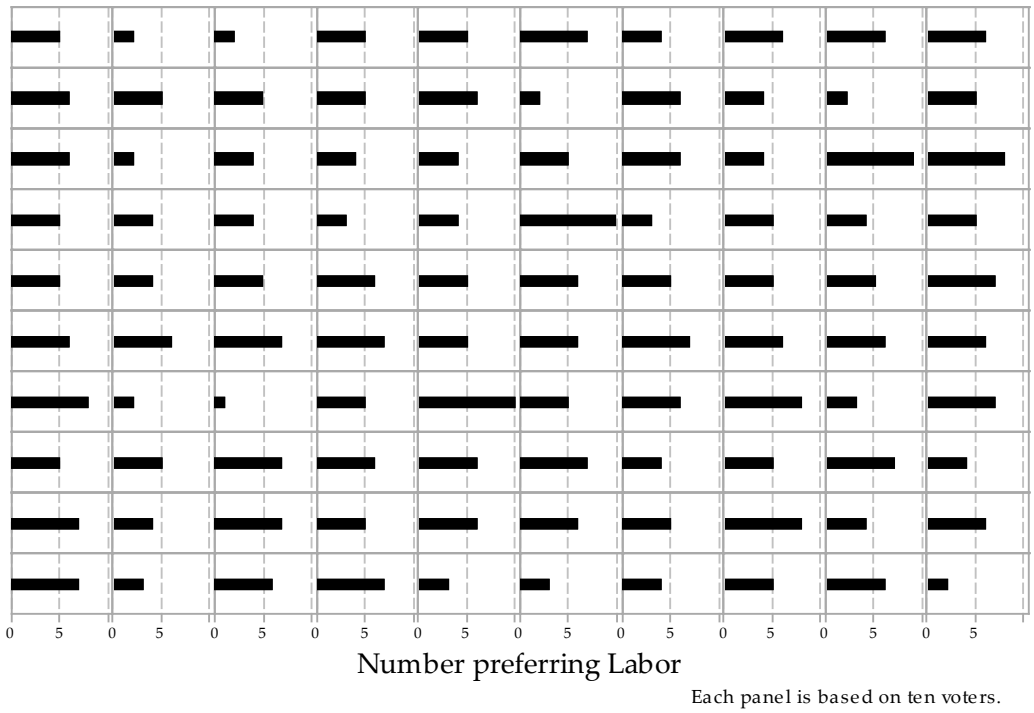

Figure 2: 100 observations from the Bi(10, 0.5) distribution.

In figure 3, the same 100 observations are represented, but the *proportion* preferring Labor is plotted, rather than the number. As each binomial observation is based on a random sample of 10 voters, the top part of figure 3 is simply a re-scaling of figure 2 with the number preferring Labor divided by 10 to provide the proportion.

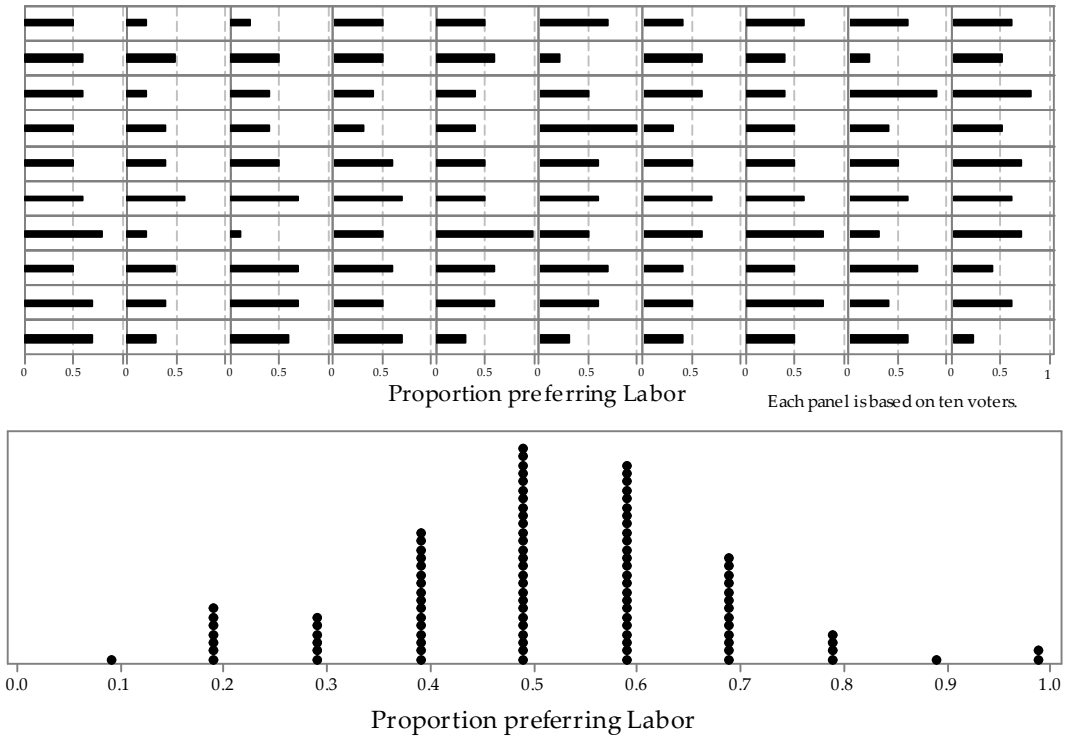

Figure 3: 100 proportions based on observations from the Bi(10, 0.5) distribution.

The bottom part of figure 3 provides an alternative representation of the same 100 observations. Each observed proportion is plotted as a dot, and the dots are stacked up when there are multiple observations of the same value.

For example, you can see that there is one dot at 0.1, and you should be able to find the corresponding single sample with a proportion of $\frac{1}{10} = 0.1$ in the top part of figure 3. Similarly, there are two dots at 1.0, corresponding to the two samples with a proportion of $\frac{10}{10} = 1.0$ in the top part of figure 3.

Since the observations shown in figure 3 are taken from a binomial distribution with $p = 0.5$, it is not surprising to find that the most frequently observed sample proportion among the 100 cases is 0.5.

Figures 2 and 3 are based on observations from the binomial distribution with $n = 10$ and $p = 0.5$. What about the distribution itself?

The top part of figure 4 shows this binomial distribution. Assuming that this is an appropriate model for the number of people preferring Labor in a sample of 10 voters, the probability of five people preferring Labor is about 0.25 (it is 0.2461 to four decimal places). We can think of this as meaning that, in the long run, among many samples of size 10, the proportion of samples in which five voters prefer Labor will be 24.61%.

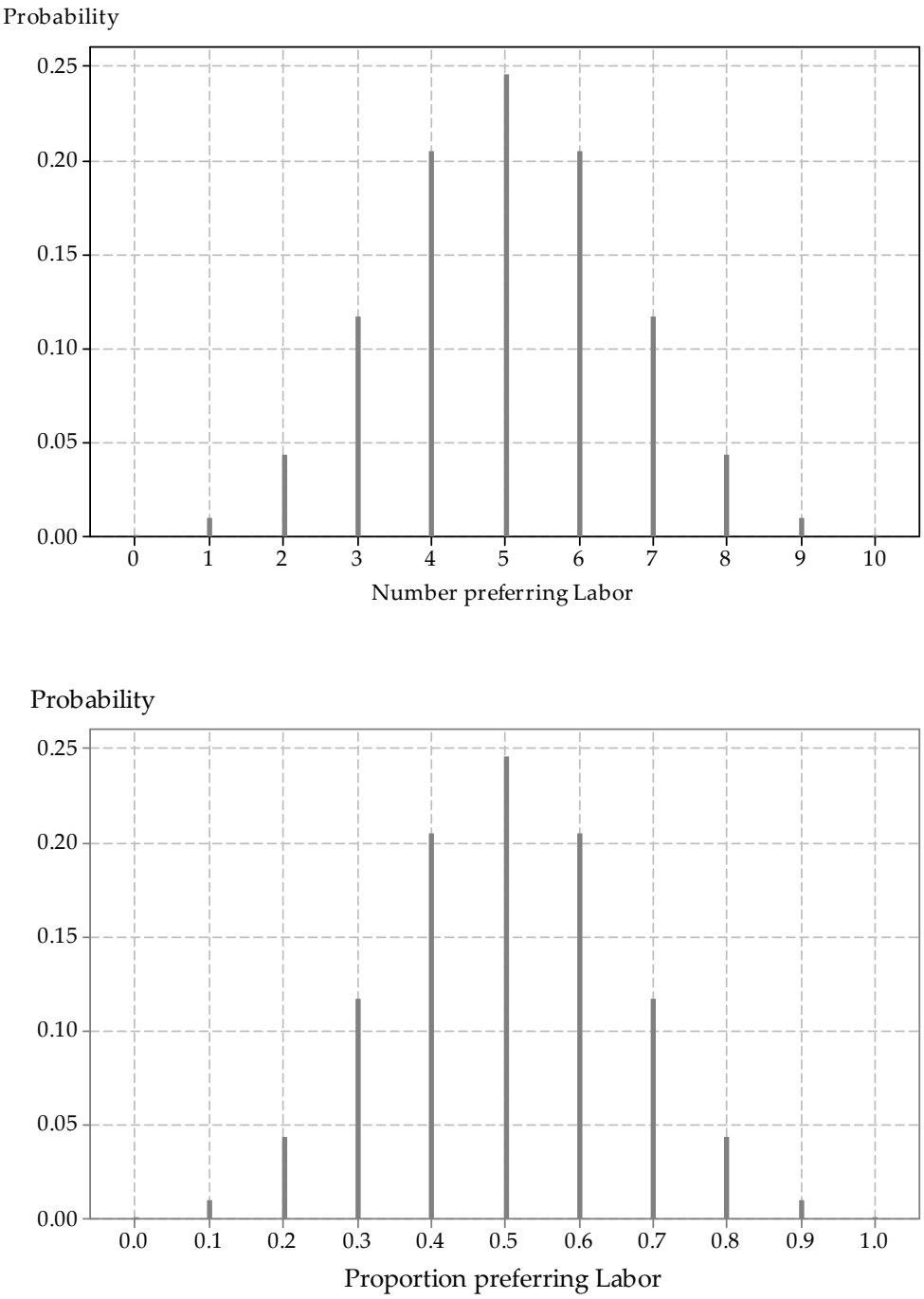

Figure 4: The Bi(10,0.5) distribution (top), and the distribution of sample proportions for observations from the Bi(10,0.5) distribution (bottom).

The top part of figure 4 corresponds closely to the bottom part; the latter shows the true distribution of sample proportions, each based on an observation from the Bi(10,0.5) distribution. The distribution in the bottom part is the same shape as the Bi(10,0.5) distribution in the top part; it is not the Bi(10,0.5) distribution because it has a different scale on the horizontal axis — the scale based on proportions.

Think about the distribution shown in the bottom part of figure 4 some more. It tells us about the true pattern of repeated sample proportions based on observations from Bi(10, 0.5). About one quarter of the sample proportions will be 'right on the money': they will be exactly equal to the true population proportion, $p = 0.5$. If the sample proportion is not exactly 0.5, it is quite likely that it is close to 0.5: the next most likely sample proportions are 0.4 and 0.6 (equally likely).

It is unlikely that the sample proportion will be a long way from the true proportion. The worst possible outcomes (furthest from the population proportion $p = 0.5$) are sample proportions of $\frac{0}{10} = 0$ or $\frac{10}{10} = 1$. They have a very small probability of occurring. Each of the outcomes 0 and 1 has probability equal to 0.0010. They are not going to occur very often, in the long run. And after all, when we looked at 100 sample proportions, none of them was equal to 0, and only two of them were equal to 1 (figure 3).

All of this shows us that the sample proportion $\hat{P} = \frac{X}{n}$ is itself a random variable. In a sense, this is obvious, since it follows directly from its definition: it is a simple function of the binomial random variable X. So $\hat{P}$ has a distribution. It has a mean and a variance: what are they?

Mean and variance of the sample proportion

We will use the following general result about the mean and variance of a linear transformation of a random variable.

If X is a discrete random variable and $Y = aX + b$, then

- $\mathrm{E}(Y) = a\,\mathrm{E}(X) + b$
- $\mathrm{var}(Y) = a^2\,\mathrm{var}(X)$
- $\mathrm{sd}(Y) = |a|\,\mathrm{sd}(X)$.

The first part is proved in the module *Discrete probability distributions*. The second part can be proved using a similar approach, and then the third part follows.

Now assume that $X \overset{\mathrm{d}}{=} \mathrm{Bi}(n, p)$. From the module *Binomial distribution*, we know that $\mathrm{E}(X) = np$ and $\mathrm{var}(X) = np(1-p)$. It follows that

$$\mathrm{E}(\hat{P}) = \mathrm{E}\left(\frac{X}{n}\right) = \frac{1}{n}\,\mathrm{E}(X) = \frac{1}{n} \times np = p.$$

So the distribution of the sample proportion $\hat{P}$ is centred around p. For statistical inference, this is highly desirable. It means that in the long run, on average, the sample proportion will neither over-estimate nor under-estimate the true value of p. Of course, a specific estimate will be out by a bit; but in the long run, the estimates average out to the true proportion p.

What about the variance? Using the general result above:

$$\begin{aligned}\text{var}(\hat{P}) &= \text{var}\left(\frac{X}{n}\right)\\ &= \left(\frac{1}{n}\right)^2 \text{var}(X)\\ &= \frac{1}{n^2}\, np(1-p)\\ &= \frac{p(1-p)}{n}.\end{aligned}$$

It follows that the standard deviation of $\hat{P}$ is given by

$$\text{sd}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}.$$

The presence of n in the denominator of the variance of the sample proportion $\hat{P}$ is very important. It means that, for larger samples, the spread of the distribution of $\hat{P}$ will be smaller. This is a good thing: Since the distribution is centred on p, the smaller the variance, the more likely it is that sample proportions will be close to p. We explore this further in the section *More on the distribution of sample proportions*.

In summary, for population proportion p and sample size n, the mean, variance and standard deviation of the sample proportion $\hat{P}$ are as follows:

$$\begin{aligned}\text{E}(\hat{P}) &= p\\ \text{var}(\hat{P}) &= \frac{p(1-p)}{n}\\ \text{sd}(\hat{P}) &= \sqrt{\frac{p(1-p)}{n}}.\end{aligned}$$

To illustrate these results, we look at what happens in our voting-preference example when we increase the sample size from 10 voters to 100 voters.

Figure 5 shows the distribution of sample proportions based on observations from the Bi$(n, 0.5)$ distribution for $n = 10$ and for $n = 100$. When $n = 100$, there are 101 possible values for the proportion of people voting Labor.

Notice what has changed between $n = 10$ and $n = 100$, and also what has not changed. What has not changed is that the distribution of the sample proportion $\hat{P}$ is still centred around the population proportion $p = 0.5$. This is true for any value of n. What has changed is that, for $n = 100$, the distribution is much more narrowly concentrated around the mean $p = 0.5$. When $n = 100$, it is more likely that a sample proportion will be close to $p = 0.5$, compared to when $n = 10$.

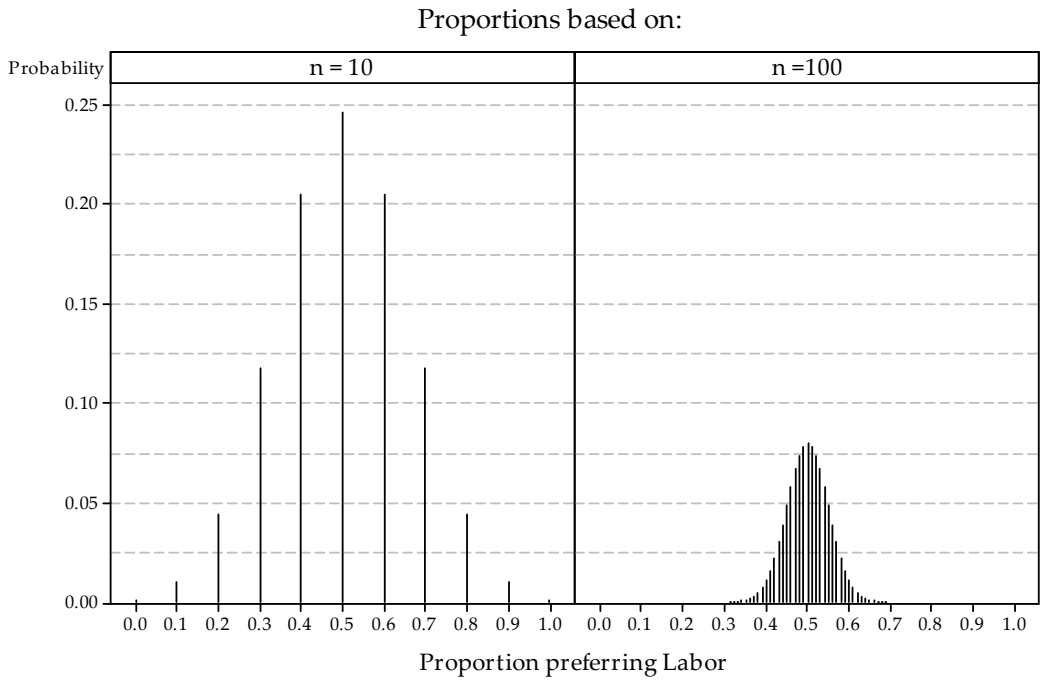

Figure 5: True distributions of sample proportions for observations from the Bi(10,0.5) distribution (left) and the Bi(100,0.5) distribution (right).

We have discussed the sample proportion $\hat{P}$ as providing a point estimate of p. We now develop this idea further, moving towards making a detailed inference about p: finding an interval which we are 'confident' contains p.

Population parameters and sample estimates

In the module *Random sampling*, the distinction between a population and a sample is described. Over the previous sections, we have considered taking samples of size n from a population in which the true proportion of people preferring Labor is 0.5. In considering this example, and others from probability, it is common to see the following reaction: 'But how do you know that the true proportion preferring Labor is 0.5?'

In many examples in the module *Probability*, there were assumptions made about specific probabilities, and the implications were explored. This is important, because we do need to understand the rules of probability and the nature of random variables and distributions. One of the main reasons for understanding this theoretical material is that it is the foundation for making inferences in real-life situations that we care about, such as 'What is the true proportion preferring Labor?' and 'How precise is our estimate?'

The actual proportion of people in the population preferring Labor is an example of a **population parameter**. It is important to make the distinction between this population parameter and a sample *estimate*. In practice (unlike in our voting-preference example), we are interested in finding out about an *unknown* population parameter; this is a proportion of the population, and has a fixed value. We collect data from a random sample in order to obtain a sample estimate of the population parameter. As we have seen, different samples from the same population do not all give the same estimate: rather, they will vary.

The unknown population parameter, the true proportion, is p. An estimate we obtain from a single sample, the observed sample proportion, is the point estimate $\hat{p}$. The aim of the methods we describe later in this module is to *infer* something about the parameter of a population from the sample. This is an **inference** because there is uncertainty about the parameter. We can, however, quantify this uncertainty.

The uncertainty involved in using sample proportions to estimate population proportions can be understood by considering the distribution of sample proportions when we sample repeatedly from the same population. Here we think of the sample proportion as a random variable. It varies from sample to sample and has a distribution. By understanding the distributional properties of the sample proportion $\hat{P}$ as an *estimator* of the population proportion, we can quantify the uncertainty in a sample estimate of a population parameter.

More on the distribution of sample proportions

As in the previous sections, we are assuming that $X \overset{\text{d}}{=} \text{Bi}(n, p)$. We have seen that the sample proportion $\hat{P} = \frac{X}{n}$ is a random variable, and so has a distribution. We found that

$$\text{E}(\hat{P}) = p$$

$$\text{var}(\hat{P}) = \frac{p(1-p)}{n}$$

$$\text{sd}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}.$$

The fact that $\text{var}(\hat{P}) = \dfrac{p(1-p)}{n}$ illustrates that, for a given value of n, the distribution of sample proportions will be more spread out when p is close to 0.5, and less spread out when p is close to 0 or 1. For example:

- if $p = 0.5$ and $n = 10$, then $\hat{P}$ has variance 0.025 and standard deviation 0.1581
- if $p = 0.1$ and $n = 10$, or if $p = 0.9$ and $n = 10$, then $\hat{P}$ has variance 0.009 and standard deviation 0.0949.

Exercise 1

The following table gives the standard deviation of $\hat{P}$ for various values of p and n. Complete the table by calculating the missing standard deviations, to two decimal places.

Standard deviation of $\hat{P}$

n	$p = 0.1$	$p = 0.3$	$p = 0.5$	$p = 0.7$	$p = 0.9$
10	0.09		0.16		0.09
50					
100					

The dependence of the spread of the distribution of sample proportions on the true proportion p is illustrated in figure 6, where we consider the distribution of $\hat{P} = \frac{X}{40}$, the proportion of successes from samples of size 40. Figure 6 also shows that the distribution of sample proportions is more symmetric for values of p closer to 0.5.

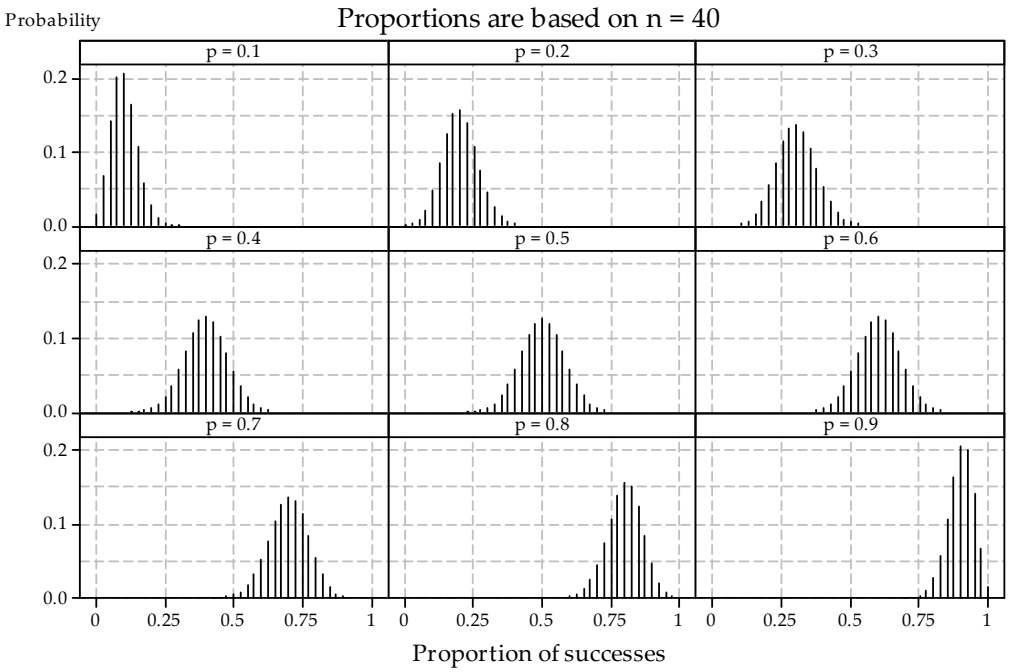

Figure 6: True distributions of sample proportions $\hat{P}$ for observations from the Bi$(40, p)$ distribution, for various values of p.

Figure 7 shows six distributions of sample proportions based on varying sample size n, but the same population parameter $p = 0.9$. As we saw in figure 5, as the sample size increases, there are more possible values for the sample proportion. Two other features of figure 7 are important. As we would expect, the spread of the distributions decreases as the sample size increases. Additionally, the symmetry of the distributions increases with sample size.

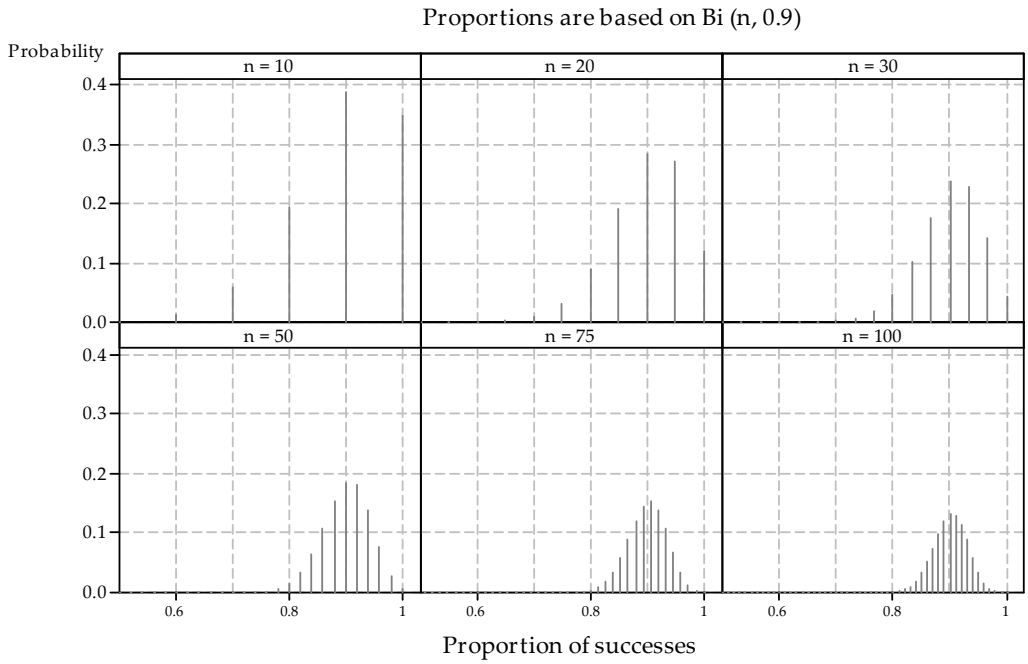

Figure 7: True distributions of sample proportions $\hat{P}$ for observations from the Bi(n,0.9) distribution, for various values of n.

The distribution of sample proportions for large sample sizes

What does the distribution of the sample proportion look like when the sample size is large? As we have seen in figure 7, even when p is as large as 0.9, if the sample size n is large, the distribution of $\hat{P}$ looks quite symmetric: Look at the three distributions in the second row, for $n = 50$, $n = 75$ and $n = 100$.

In figure 8, a sample size of $n = 100$ has been used throughout. With this large sample size, the distributions across the range of true proportions from 0.1 to 0.9 are quite symmetric, and clearly much more symmetric than the examples we have seen in previous figures when n is not large. With the large number of possibilities for the sample proportion (101 different values), the distributions are reminiscent of a continuous distribution. The shape of each distribution is symmetric, and like a Normal distribution. Of course, the distribution cannot actually be a Normal distribution, because the Normal distribution is continuous, and the distribution of sample proportions is discrete. But visually it appears that a Normal distribution would be quite a good approximation. This visual impression is correct, as we now demonstrate.

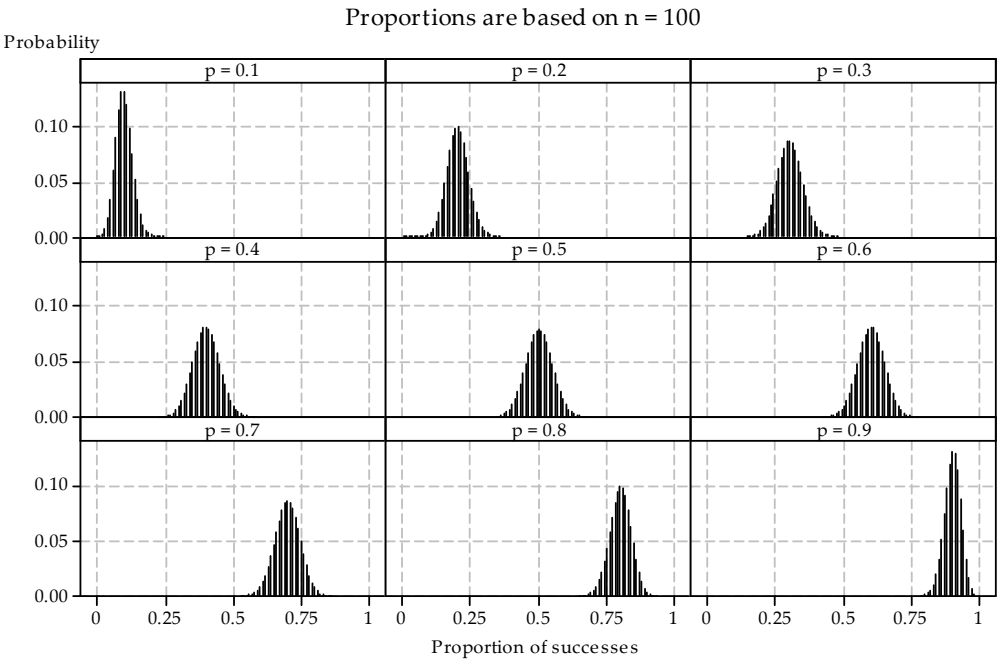

Figure 8: True distributions of sample proportions $\hat{P}$ for observations from the $\text{Bi}(100, p)$ distribution, for various values of p.

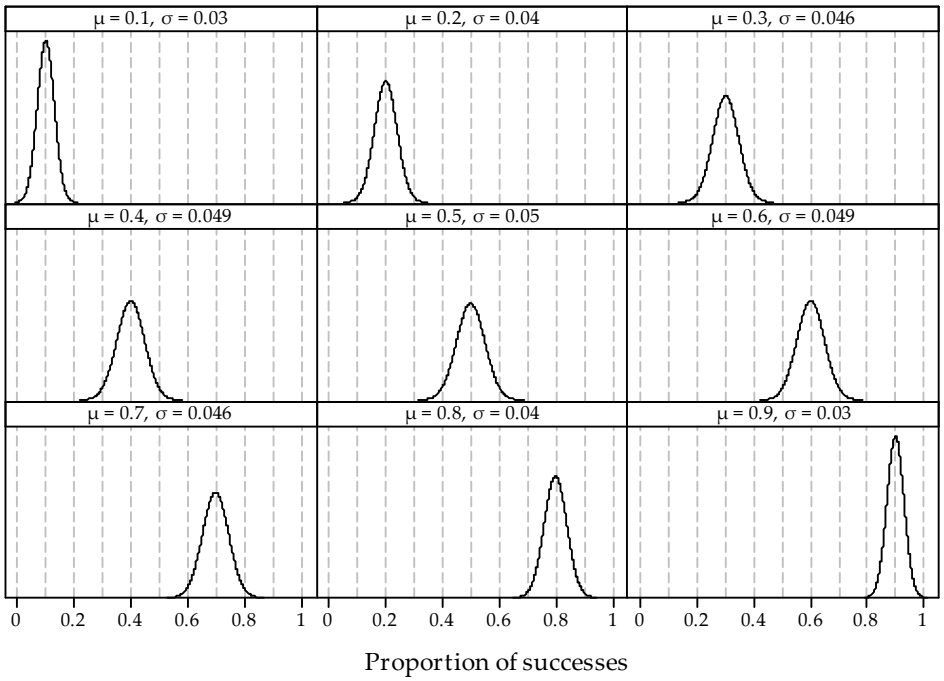

Figure 9: Normal distributions with means and standard deviations corresponding to those of the distributions of sample proportions in figure 8.

The nine Normal distributions shown in figure 9 have means and standard deviations corresponding to those of the distributions of sample proportions in figure 8. For example, the top-left panel in figure 8 shows the distribution of $\hat{P}$ for $n = 100$ and $p = 0.1$, so $\mathrm{E}(\hat{P}) = p = 0.1$ and $\mathrm{sd}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}} = 0.03$. Hence, the top-left Normal distribution in figure 9 has mean $\mu = 0.1$ and standard deviation $\sigma = 0.03$.

These two figures illustrate how the distribution of sample proportions can be approximated by a Normal distribution for large sample sizes.

Figure 10 shows the distribution of sample proportions based on $n = 1000$ and $p = 0.5$. Here we see an even closer approximation to 'continuity' and a Normal distribution. Again, the distribution cannot actually be Normal, because the proportions can only take discrete values. But consider how close together the discrete values now are, when n is so large. The gaps between the spikes (representing the probabilities) are only 0.001 apart, because the proportions can take values such as $0.500, 0.501, 0.502, \ldots$. So the appearance of a Normal distribution is stronger than any of the examples we have seen for smaller sample sizes.

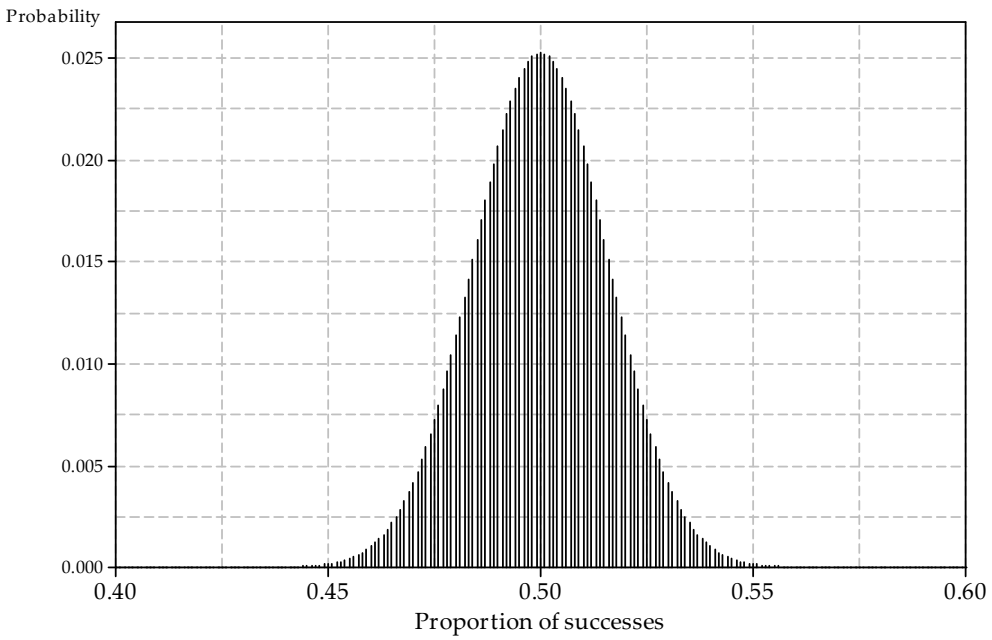

Figure 10: Distribution of the sample proportion $\hat{P}$ from the Bi(1000, 0.5) distribution.

The Normal approximation described here is used later, when we obtain an approximate *confidence interval* for the unknown p, based on an observation from the binomial distribution $\mathrm{Bi}(n, p)$, for large n. Before getting to the practicalities, however, we consider some very important general ideas about confidence intervals.

Confidence intervals

This section deals with fundamental aspects of confidence intervals. In the next section, we will deal with obtaining a confidence interval for the specific case we are considering. But it is important first to understand confidence intervals conceptually.

An observed sample proportion $\hat{p}$ is a single point or value that provides us with an estimate of the true proportion of interest in the population. For this reason, it is called a *point estimate*. In some sense, we are not interested in the particular value of the sample proportion *per se*, but rather we are interested in the information it provides us about the population. It provides an estimate of the population parameter of interest; in this case, the population proportion p.

While the proportion from the sample will provide us with the best estimate of the population proportion, it is unlikely that the sample value will be exactly equal to the parameter being estimated. Hence, the sample estimate is most useful if it is combined with some information about its precision.

Suppose, for example, we want to estimate p, the proportion of Australian adults aged 18–24 who feel that it is acceptable for a breakup to be conveyed via email, phone or text message (for brevity, we refer to this as an 'impersonal' breakup), and that we have the results of two different surveys on hand, each of them based on random samples. The first survey provides an estimate of the proportion equal to 0.17 (17%), while the second survey provides the estimate 0.28 (28%). These estimates may seem inconsistent, and it may be unclear which we might prefer to rely on. However, if the first survey result is likely to be within ± 0.07 of the true value of p, and the second survey result is likely to be within ± 0.25 of the true value of p, then the first result is more precise than the second.

By describing the first survey result as 0.17 ± 0.07, we are specifying an interval or range of values (from $0.17-0.07$ to $0.17+0.07$) within which we have confidence that the true value of p lies. The interval has a lower bound and an upper bound: 0.10 and 0.24, respectively. This interval is an indicator of the precision of an estimate of the population proportion and is called a **confidence interval**. Here the confidence interval is $(0.10, 0.24)$.

Although we are discussing the specific context of an inference on p, most of the ideas discussed in this section apply to confidence intervals for any unknown population parameter.

'Confidence' has a particular meaning in this context, which we now describe.

Confidence level

When working out a confidence interval, we must first decide on a 'degree' or 'level' of confidence. This is quantified by the confidence level. We want to be very confident, so it makes sense to have a high confidence level. In most applications, the confidence level used is 95%. This is a very strong tradition. You may wonder: Why don't we use 100%, and work out a 100% confidence interval? This question is asked in exercise 3.

A confidence interval will always be obtained from a random variable, so the interval itself can be thought of as a random interval. It varies from one sample to the next, just as a random variable does.

The confidence level specifies the long-run percentage or proportion of confidence intervals containing the true value of the parameter: in this context, p. Illustrating this idea requires a simulation or a thought experiment. In practice, we typically do not have a long run of repeated samples at all. We have a *single sample* of size n, and we calculate $\hat{p}$ and a single confidence interval to characterise the precision in the result. Any *actual* interval either contains or does not contain the true value of the parameter p, although we don't know whether it does or not, because we don't know the value of p. For example, we don't know whether the interval 0.10 to 0.24, for the proportion of Australian adults aged 18–24 who feel that impersonal breakups are acceptable, contains the true value p. The confidence level of 95% being used here does not mean that the chance of this particular interval containing p is 95%.

To illustrate the meaning of the confidence level, think about the true proportion p of Australian adults aged 18–24 who feel that impersonal breakups are acceptable. As we observed, we don't know the value of p; estimating p is what we are trying to do! Assume, *just for the purposes of this discussion*, that $p = 0.20$ (20%). The first survey described above is based on a random sample of 100 adults aged 18–24, and 17% indicated that an impersonal breakup is acceptable. We can imagine repeating the process used in the first survey many times, sampling different adults each time, while maintaining random sampling and a sample size of $n = 100$. Each time we will observe a different sample proportion.

Figure 11 shows 100 such surveys, with the first survey result closest to the horizontal axis. For each survey, the estimate of the proportion of interest is plotted as a dot in the centre of a line. The line shows the 95% confidence interval for the particular survey. For the first survey, the line showing the 95% confidence interval is from 0.10 to 0.24.

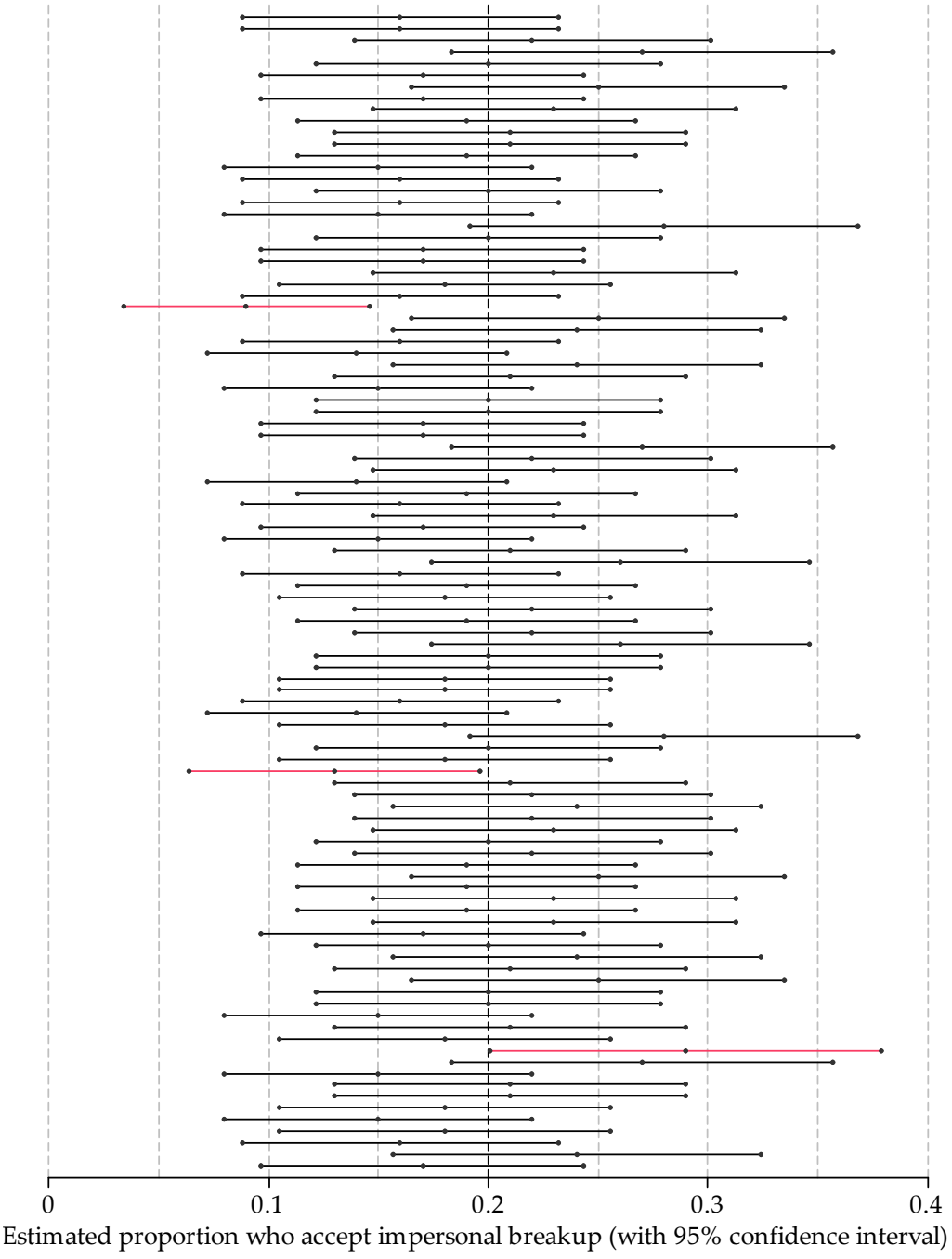

Figure 11: One hundred surveys, based on random samples of 100 adults aged 18–24, estimating the proportion who regard impersonal breakups as acceptable, showing the point estimate and the 95% confidence interval in each case.

Figure 11 shows a darker vertical gridline, corresponding to the true value of p, namely $p = 0.20$. Most of the confidence intervals are colored black, but a small number are red; these are the confidence intervals that do not include the true value of 0.2. In total, three of the one hundred intervals are red. There are two on the low side, where all values in the interval are less than 0.2, and one on the high side, where the lower bound of the interval is greater than 0.2. In this small simulation, 97% of the intervals include the true value. We expect that 95% of the 95% confidence intervals will include the true value; with much larger simulations, the percentage would be very close to 95%.

Thinking about this more formally, if we define $\text{CI}_{0.95}(\hat{P})$ to be the 95% confidence interval that is based on the random variable $\hat{P}$, then we can write

$$\Pr[p \in \text{CI}_{0.95}(\hat{P})] = 0.95.$$

We have expressed this in the specific context we are considering, but the point is general: A 95% confidence interval (considered as a random interval) for an unknown parameter has a probability of 0.95 of containing the parameter.

Exercise 2

Suppose we have m independent 95% confidence intervals for an unknown parameter.

a Define Y to be the number of intervals that include the unknown parameter value. What is the distribution of Y? (*Hint.* Think of the m intervals as a sequence of Bernoulli trials.)

b Hence, what is $\text{E}(Y)$?

c Now assume $m = 100$.

- **i** Find the chance that exactly 95 of the intervals include the parameter, that is, find $\Pr(Y = 95)$.
- **ii** Find the chance that at least 95 of the intervals include the parameter.

Varying the confidence level

Figure 12 represents the same 100 surveys of 100 Australian adults aged 18–24. For each survey, the estimate of the proportion of interest is plotted as a dot in the centre of a line which this time shows the 50% confidence interval. The central dot is the point estimate. Hence, because they are the same surveys, the central dots in figure 12 are at the same positions as those in figure 11.

However, the confidence intervals are much narrower. As before, the confidence intervals shown in black contain the true value of the parameter, namely $p = 0.2$, and those shown in red do not. The 50% confidence intervals are narrow and therefore may appear precise. But figure 12 indicates that this is at a price. About half of them do not include the true value of interest.

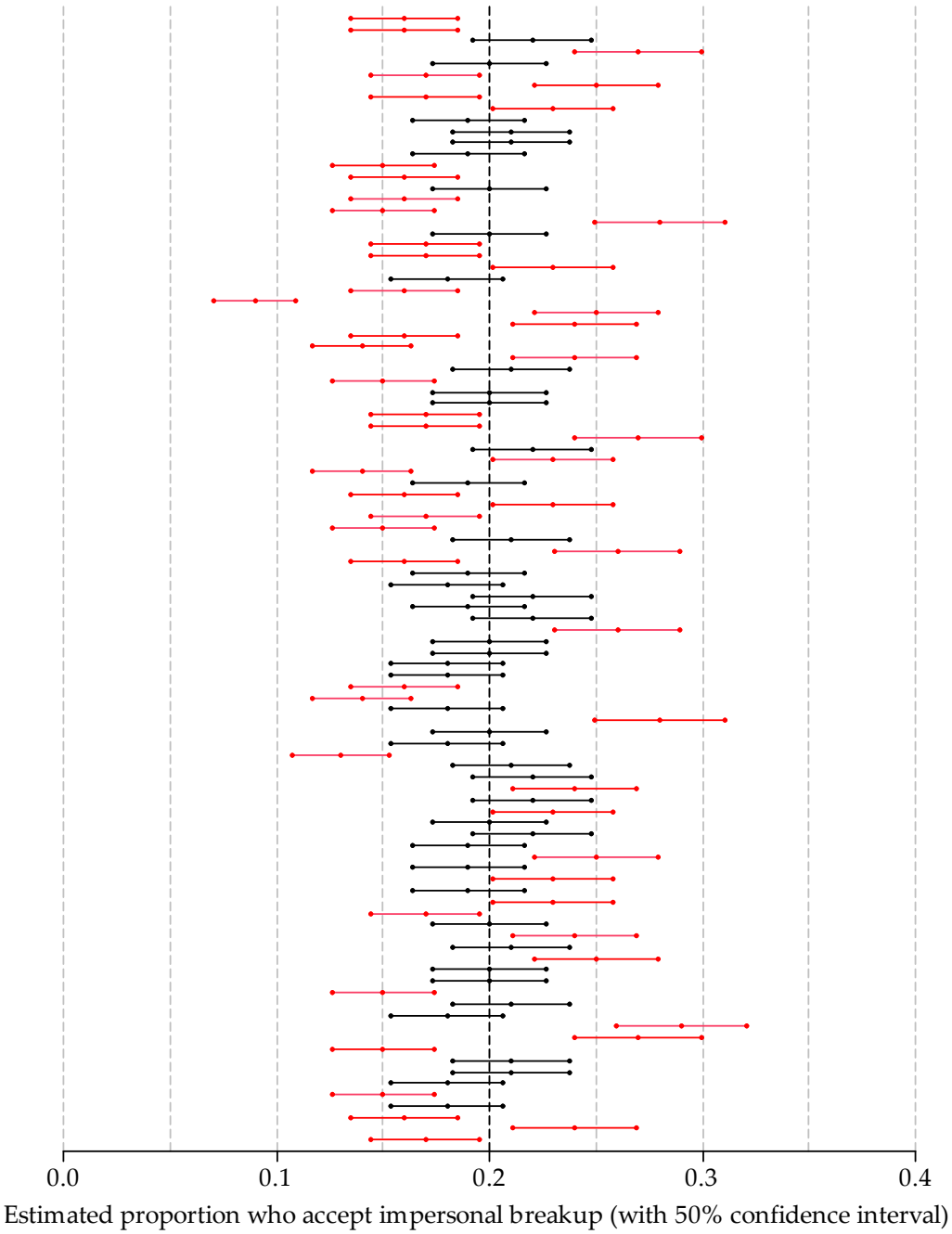

Figure 12: One hundred surveys, based on random samples of 100 adults aged 18–24, estimating the proportion who regard impersonal breakups as acceptable, showing the point estimate and the 50% confidence interval in each case.

Another way to see the effect of varying the confidence level is to examine confidence intervals with different confidence levels for the same survey. This is shown in figure 13, using the first survey of 100 Australian adults aged 18–24. The confidence intervals have different confidence levels. Since the same survey is represented in each case, the point estimate is the same, but the confidence intervals have different widths.

Figure 13: Confidence intervals from the same data, but with different confidence levels.

Exercise 3

Consider the following figure, which shows the 50% and 95% confidence intervals for the proportion of Australian adults aged 18–24 who feel that impersonal breakups are acceptable, based on the first survey. Sketch the 100% confidence interval on the figure.

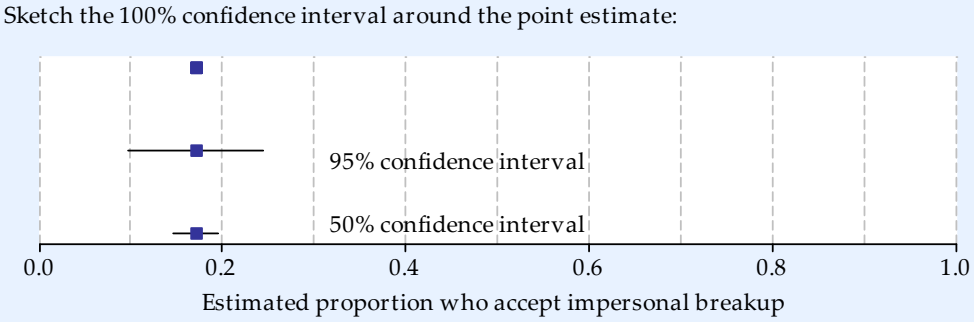

Figure 14: Estimate of the proportion who regard impersonal breakups as acceptable from one survey.

We have discussed the general properties of a confidence interval. We now turn to the practical issue of obtaining a confidence interval for the unknown population proportion p.

Calculating confidence intervals

An approximate standard Normal distribution

We have seen how the distribution of sample proportions approximates a Normal distribution, for large n. The module *Exponential and normal distributions* shows how any Normal distribution can be standardised, in the following way, to give a standard Normal distribution:

$$\text{If } Y \overset{\text{d}}{=} \text{N}(\mu, \sigma^2) \text{ and } Z = \frac{Y - \mu}{\sigma}, \text{ then } Z \overset{\text{d}}{=} \text{N}(0, 1).$$

The **standard Normal distribution** has mean 0 and variance 1. A random variable with this distribution is usually denoted by Z. That is, $Z \overset{\text{d}}{=} \text{N}(0, 1)$.

Consider then a standardisation of $\hat{P}$. We know that $\text{E}(\hat{P}) = p$ and $\text{sd}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$, and we know that $\hat{P}$ is approximately Normally distributed if n is large. What is the distribution of the random variable

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

obtained by standardising $\hat{P}$?

This is illustrated in stages by figures 15, 16 and 17. The distributions in the same position represent the same values of p and n, as shown by the labelling of the panels.

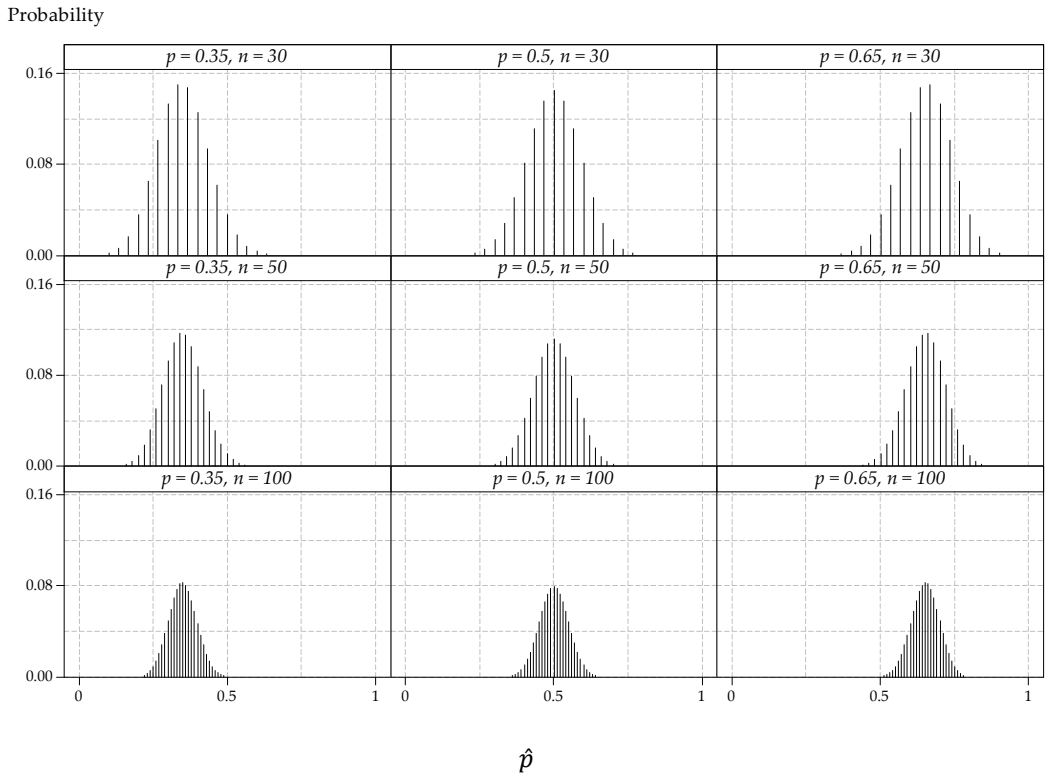

Figure 15: Distribution of $\hat{P}$, for various values of p and n.

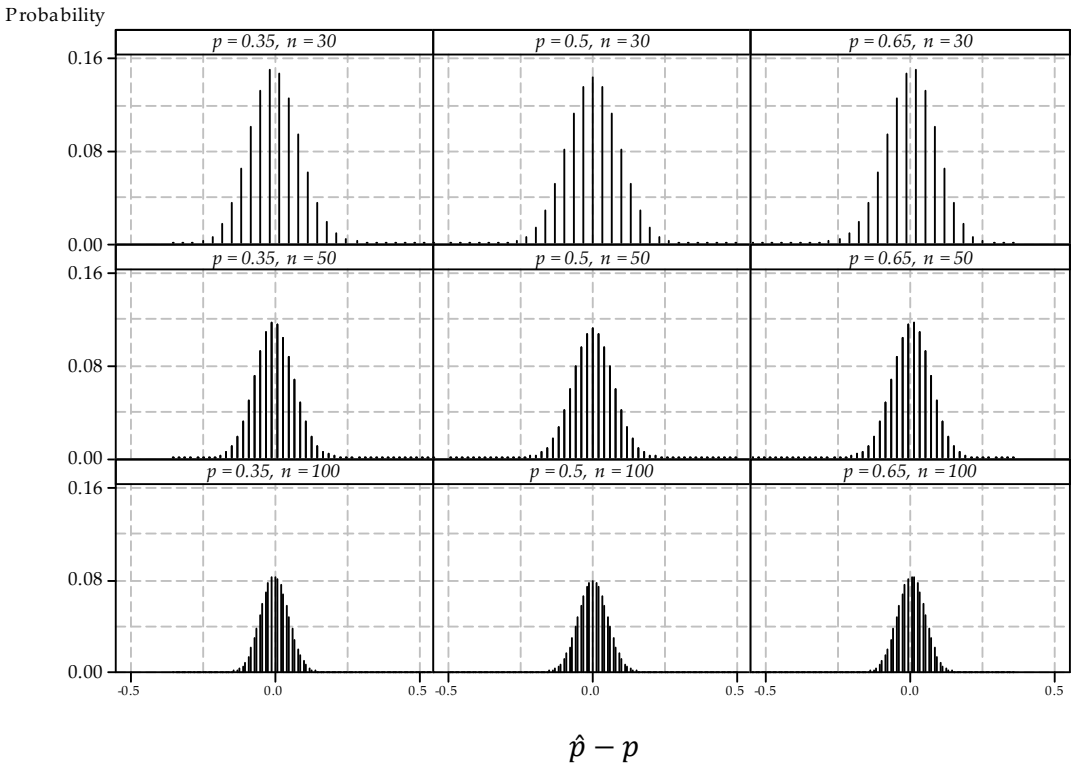

Figure 16: Distribution of $\hat{P} - p$, for various values of p and n.

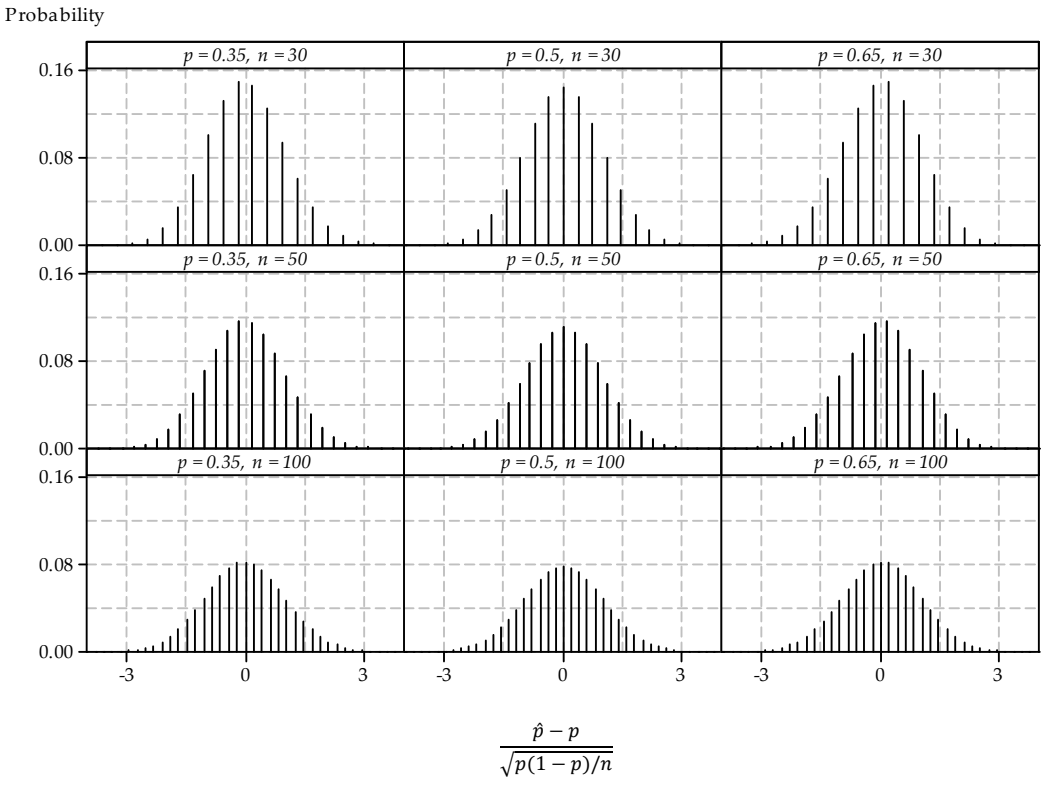

Figure 17: Distribution of the standardisation of $\hat{P}$, for various values of p and n.

Figure 15 shows the distribution of the sample proportion $\hat{P}$ in nine different cases, one for each combination of $p = 0.35, 0.5, 0.65$ and $n = 30, 50, 100$.

Figure 16 shows the distribution of $\hat{P} - p$ in each of the nine cases. Using the general fact that $\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b$, we have $\mathrm{E}(\hat{P} - p) = p - p = 0$. Further, the general result $\mathrm{var}(aX + b) = a^2\,\mathrm{var}(X)$ tells us that transforming a random variable by adding a constant leaves the variance unchanged. Hence, $\mathrm{var}(\hat{P} - p) = \mathrm{var}(\hat{P})$ and so $\mathrm{sd}(\hat{P} - p) = \mathrm{sd}(\hat{P})$.

So by subtracting the mean p from $\hat{P}$, we find that all the distributions are centred at 0, but the spread of each distribution is the same as that of the corresponding distribution in figure 15.

Figure 17 shows the distributions obtained when we complete the standardisation, by dividing the random variable $\hat{P} - p$ by its standard deviation. We are now looking at the distribution of

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

Now all the distributions have the same centre and spread. More specifically,

$$\mathrm{E}\left(\frac{\hat{P} - p}{\sqrt{\frac{1}{n}p(1-p)}}\right) = 0$$

$$\mathrm{sd}\left(\frac{\hat{P} - p}{\sqrt{\frac{1}{n}p(1-p)}}\right) = 1.$$

Exercise 4

Confirm that

a $\mathrm{E}\left(\dfrac{\hat{P} - p}{\sqrt{\frac{1}{n}p(1-p)}}\right) = 0$ **b** $\mathrm{sd}\left(\dfrac{\hat{P} - p}{\sqrt{\frac{1}{n}p(1-p)}}\right) = 1.$

We know the mean and standard deviation of the standardised random variable, and it is approximately Normally distributed. Putting all this together we can say that, for large n,

$$\frac{\hat{P} - p}{\sqrt{\frac{1}{n}p(1-p)}} \overset{\mathrm{d}}{\approx} \mathrm{N}(0, 1).$$

That is, for large n, the distribution of the standardised random variable is approximately the standard Normal distribution.

This standardisation proves crucial in obtaining a *confidence interval* for the unknown p, when we have an observation from the binomial distribution $\mathrm{Bi}(n, p)$, as we now show.

Calculating a 95% confidence interval with the Normal approximation

The crucial point to see in figure 17 is that all of the distributions are approximately the same, regardless of the values of p and n. This is especially important, since we don't know the value of p; we are trying to estimate it.

As discussed in the module *Exponential and normal distributions*, for a Normal random variable $X \overset{\mathrm{d}}{=} \mathrm{N}(\mu, \sigma^2)$, about 95% of the distribution is within two standard deviations of the mean. That is,

$$\Pr(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95.$$

Hence, for a random variable with the standard Normal distribution, $Z \overset{\mathrm{d}}{=} \mathrm{N}(0,1)$, we have $\Pr(-2 < Z < 2) \approx 0.95$. To be more precise:

$$\Pr(-1.96 < Z < 1.96) = 0.95.$$

This is illustrated in figure 18.

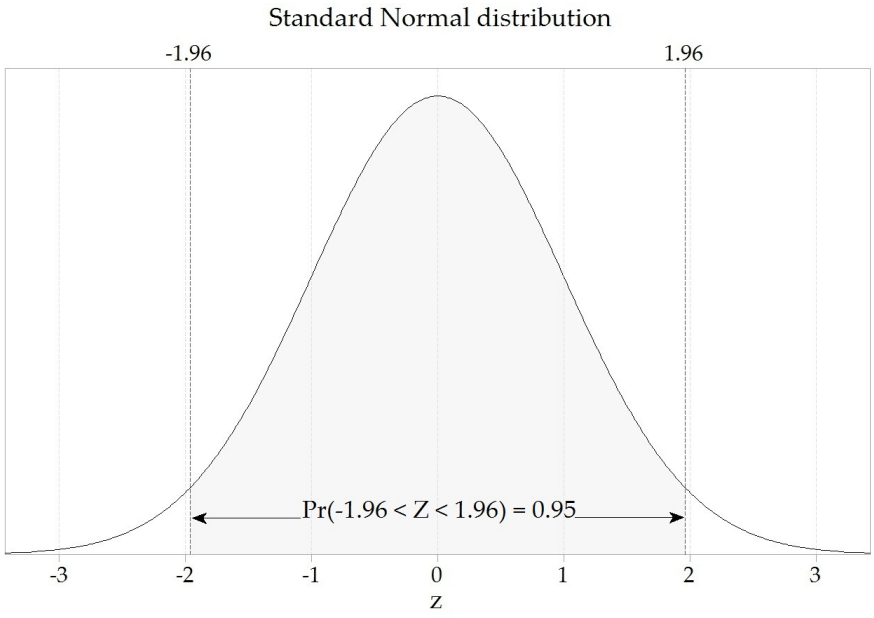

Figure 18: The standard Normal distribution, $Z \overset{\mathrm{d}}{=} \mathrm{N}(0,1)$.

Figure 19 shows the distribution of the sample proportion $\hat{P}$ for sample size $n = 100$ and population proportion $p = 0.5$ (or, equivalently, for observations from the $\mathrm{Bi}(100, 0.5)$ distribution). The standard deviation of $\hat{P}$ is shown in the figure. Only a small percentage of the sample proportions are more than two standard deviations away from $p = 0.5$. About 95% of the sample proportions are within two standard deviations of p.

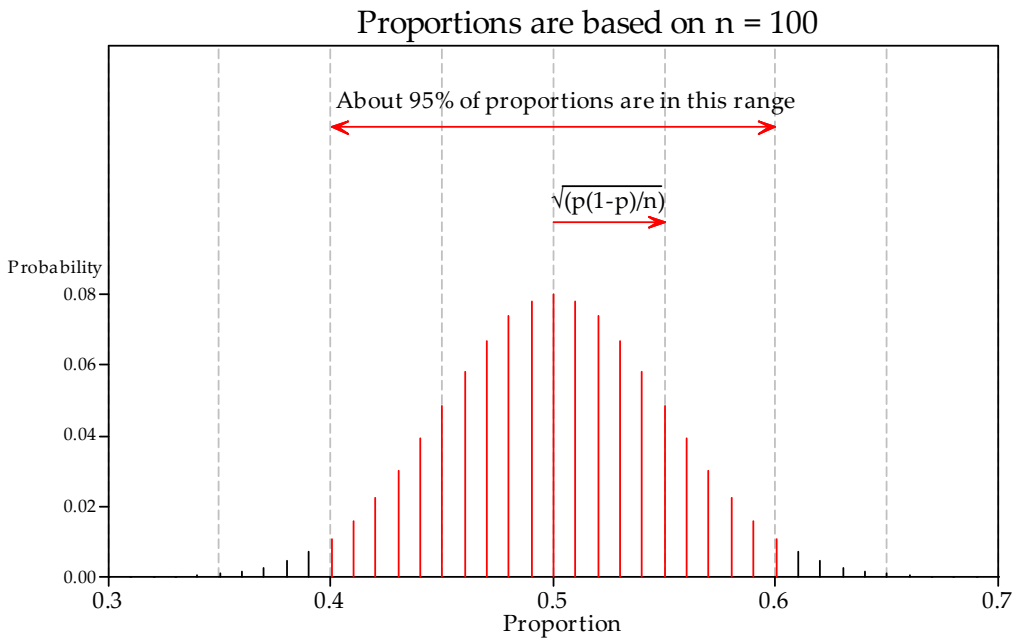

Figure 19: Distribution of the sample proportion based on $n = 100$ and $p = 0.5$.

Similarly, for each of the distributions of standardised sample proportions in figure 17, about 95% of the distribution is within two standard deviations of the mean. Since each of the standardised sample proportions has a distribution that can be approximated by the standard Normal distribution, we can state that, for large n,

$$\Pr\left(-1.96 < \frac{\hat{P} - p}{\sqrt{\frac{1}{n}p(1-p)}} < 1.96\right) \approx 0.95.$$

We multiply through by $\sqrt{\frac{p(1-p)}{n}}$ to obtain

$$\Pr\left(-1.96\sqrt{\frac{p(1-p)}{n}} < \hat{P} - p < 1.96\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95.$$

In other words, the distance between $\hat{P}$ and p will be no more than $1.96\sqrt{\frac{p(1-p)}{n}}$ for 95% of sample proportions.

One further rearrangement gives

$$\Pr\left(\hat{P} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95.$$

It is really important to reflect on this probability statement. Note that it has p in the centre of the inequalities. The population parameter p does not vary: it is fixed, but unknown. The random element in this probability statement is the random interval around p.

This forms the basis for the approximate 95% confidence interval for the true proportion p. In a given case, we have just a single observation from the $\mathrm{Bi}(n,p)$ distribution. We then find the observed value $\hat{p}$ of $\hat{P}$, and obtain the observed value of the random interval, which we call the 95% confidence interval.

This gives us a 95% confidence interval for p:

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

However, a problem remains. The unknown parameter is still present in the formula! This seems unfortunate, to say the least; the very parameter we are seeking to estimate is present in a formula that reflects the precision of our estimate.

The solution is to make a further approximation and substitute $\hat{p}$ for p in the expression for the standard deviation of $\hat{P}$.

Hence, an approximate 95% confidence interval for p is given by

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

or, equivalently,

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

It is reasonable at this point to ask: How do we know that the Normal approximation is adequate when we have substituted $\hat{p}$ for p, as described? We provide an informal answer to this sensible question in figure 20, by showing the distributions of

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

for the same values of p and n as in figure 17. Reassuringly, the two figures 17 and 20 are almost indistinguishable. It is true that, for large n,

$$\frac{\hat{P} - p}{\sqrt{\frac{1}{n}\hat{P}(1-\hat{P})}} \overset{\mathrm{d}}{\approx} \mathrm{N}(0,1).$$

This confirms that, for large n, the confidence interval $(*)$ derived above is a reasonable approximation.

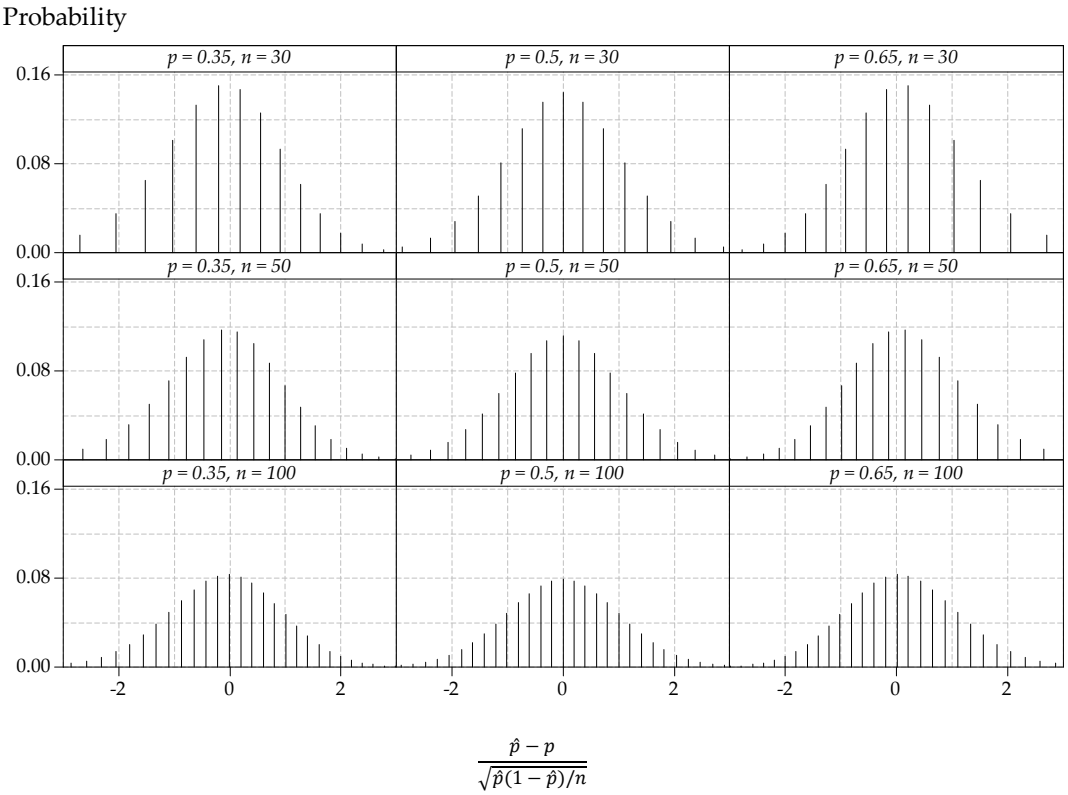

Figure 20: Distribution of $\frac{\hat{P}-p}{\sqrt{\frac{1}{n}\hat{P}(1-\hat{P})}}$, for various values of p and n; compare with figure 17.

Keep in mind that the value of 1.96 in the calculation of the confidence interval comes from the use of the standard Normal distribution, and corresponds to a central area of 95%. This is the appropriate factor because the chosen level of confidence is 95%.

We have used the phrase 'for large n' frequently, and relied on visual impressions from various distributions to get a sense of what may be adequately large for the approximation to be satisfactory in practice. In fact, the adequacy depends on both n and p. A guideline often given is that:

> If $X \overset{\text{d}}{=} \text{Bi}(n, p)$ and the observation we are using for the approximate confidence interval is x, then we require both x and $n - x$ to be greater than 10.

Example: Mobile-phone use among children

A recent large survey of a random sample of Australian children asked about mobile-phone ownership in three age groups. The following table shows the number of mobile-phone owners and the total number of children surveyed for each age group.

Mobile-phone ownership

Age group (years)	Number of mobile owners	Number surveyed
5–8	50	2150
9–11	544	2530
12–14	918	1250

Calculate an approximate 95% confidence interval for the true proportion of mobile-phone owners in each group.

Solution

These data easily meet the guideline for the Normal approximation to be adequate.

In the 5–8 age group, we have $\hat{p} = \frac{50}{2150} = 0.0233$ and so

$$1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.0233(1-0.0233)}{2150}} = 0.00637.$$

Hence, the 95% confidence interval is 0.0233 ± 0.00637, or $(0.0169, 0.0296)$. In percentage terms, the confidence interval is 1.69% to 2.96%.

It is important to learn the real meaning of this confidence interval. As discussed in the section *Confidence intervals*, we cannot really say that the chance that the unknown percentage is between 1.69% and 2.96% is equal to 0.95. Once we have the data and the actual calculated interval, there is no randomness involved: the unknown percentage is a fixed number, not a random variable. Rather, we can say that we have calculated an interval using a process that, in a long run of repeated instances of the study under the same circumstances, would produce intervals that contained the unknown percentage in 95% of cases, on average.

For the 9–11 age group, we have $\hat{p} = \frac{544}{2530} = 0.215$ and so

$$1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.215(1-0.215)}{2530}} = 0.0160.$$

Hence, the 95% confidence interval is 0.215 ± 0.0160, or $(0.199, 0.231)$. In percentage terms, the confidence interval is 19.9% to 23.1%.

For the 12–14 age group, we have $\hat{p} = \frac{918}{1250} = 0.734$ and so

$$1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.734(1-0.734)}{1250}} = 0.0245.$$

Hence, the 95% confidence interval is 0.734 ± 0.0245, or $(0.710, 0.759)$. In percentage terms, the confidence interval is 71.0% to 75.9%.

Exercise 5

Casey buys a Venus chocolate bar every day for 180 days, during a promotion promising that 'one is six wrappers is a winner'. From these 180 purchases, Casey gets 20 winning wrappers.

a What is the proportion of winning wrappers Casey expects to get, if the advertised claim is true? How many winning wrappers would this imply, for Casey?

b What is the proportion of winning wrappers in Casey's sample?

c Find an approximate 95% confidence interval for the true proportion of winning wrappers, based on Casey's sample of Venus bars.

d Casey feels he has missed out, and suspects that the true proportion of winners is not one in six. Comment on this, based on Casey's sample of Venus bars.

e What assumptions have been made about Casey's sample of Venus bar wrappers?

More on calculating confidence intervals

Calculating a C% confidence interval with the Normal approximation

We have focussed so far on 95% confidence intervals, which is the confidence level that is used most commonly. The general form of an approximate C% confidence interval for a population proportion is

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where the value of z is appropriate for the confidence level. For a 95% confidence interval, we use $z = 1.96$, while for a 90% confidence interval, for example, we use $z = 1.64$.

In general, for a C% confidence interval, we need to find the value of z that satisfies

$$\Pr(-z < Z < z) = \frac{C}{100}, \quad \text{where } Z \overset{\text{d}}{=} \text{N}(0,1).$$

Figure 21 shows the required value of z as a function of the confidence level.

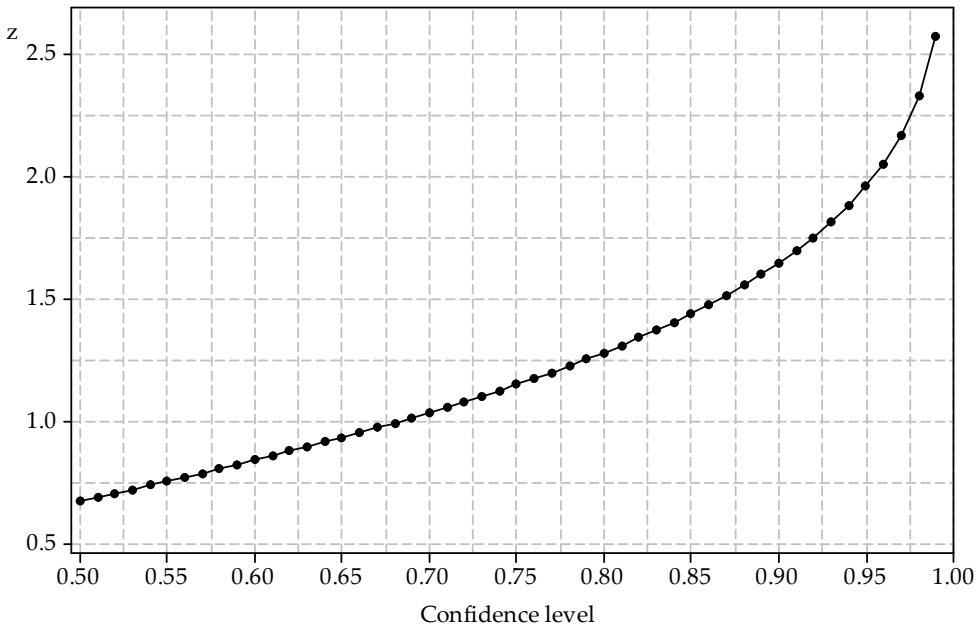

Figure 21: The relationship between the confidence level and the value of z in the formula for an approximate confidence interval.

The following figure is a repeat of figure 13. It shows confidence intervals based on the same estimated proportion, but with different confidence levels. The larger confidence levels lead to wider confidence intervals.

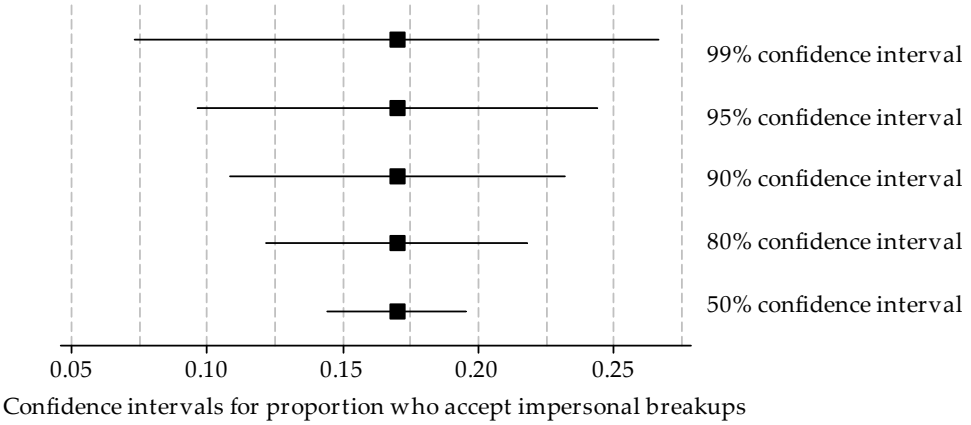

Figure 22: Confidence intervals from the same data, but with different confidence levels.

The distance from the sample estimate $\hat{p}$ to the endpoints of the confidence interval is

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

The quantity E is referred to as the **margin of error**. The margin of error is half the width of the confidence interval. Sometimes confidence intervals are reported as $\hat{p} \pm E$; this means the bounds of the interval are not directly stated, but must be calculated.

We have seen in figure 22 that the margin of error is larger when the confidence level is larger. This is because the value of z from the standard Normal distribution will be larger when the confidence level is larger.

Example: Mobile-phone use among children

Continuing with the mobile-phone example, consider the 12–14 age group. Calculate an approximate 90% confidence interval for the true proportion of mobile-phone owners in this group.

Solution

From the table in the initial mobile-phone example, we have $\hat{p} = \frac{918}{1250} = 0.734$. For a 90% confidence interval, we use $z = 1.64$, and so the margin of error is

$$1.64\sqrt{\frac{0.734(1-0.734)}{1250}} = 0.0205.$$

Hence, the 90% confidence interval is 0.734 ± 0.0205, or $(0.714, 0.755)$. In percentage terms, the confidence interval is 71.4% to 75.5%.

Exercise 6

Consider Casey's sample of Venus bars from exercise 5. He obtained a random sample of 180 wrappers, and found that 20 were winners. Rather than a 95% confidence interval for the true proportion of winning wrappers, consider a 99% confidence interval.

a Without calculating the 99% confidence interval, guess the lower and upper bounds.

b Find the value of the factor z from the standard Normal distribution for a 99% confidence interval (if necessary, by reading it off the graph in figure 21). Consider the ratio of the values of z for the 99% and 95% confidence intervals, and estimate the lower and upper bounds of the 99% confidence interval.

c Calculate the approximate 99% confidence interval for the true proportion of winning wrappers, based on Casey's sample of Venus bars.

d Consider Casey's suspicion that the true proportion of winners is not one in six. Comment on this, based on the 99% confidence interval.

Maximum margin of error

In the module *Binomial distribution*, we noted that if $X \overset{\text{d}}{=} \text{Bi}(n, p)$, then the variance is largest (for a given value of n) when $p = \frac{1}{2}$, in which case $\text{var}(X) = n \times \frac{1}{2} \times \frac{1}{2} = \frac{n}{4}$.

This has the direct consequence that, when estimating p, the variance of $\hat{P}$ is largest when $p = 0.5$, and is equal to $\frac{0.25}{n}$. This is perhaps slightly unfortunate for political polling in particular, since such surveys are quite often estimating a characteristic (such as political preference) which is present in about half of the population.

However, there is some good news for the pollsters: While they may be in the realm of least precise inferences, they know how bad it can get. For a random sample of size n, the standard deviation of $\hat{P}$ cannot be bigger than $\frac{0.5}{\sqrt{n}}$, and hence the margin of error for a 95% confidence interval is at most $1.96\,\frac{0.5}{\sqrt{n}} = \frac{0.98}{\sqrt{n}}$.

To make the reporting of such polls succinct, this fact is sometimes exploited. The report simply uses the maximum margin of error for the given sample size n, knowing that this is conservative: the precision will be as claimed if the estimated proportion $\hat{p}$ is 0.5 (the percentage is 50%), and better than claimed otherwise.

Exercise 7

A Nielsen Poll published on 17 February 2013 reported that, in a two-party vote, 56% of voters prefer the Coalition (44% prefer ALP). The report indicates that the approximate margin of error is at most 2.6%.

a Based on this margin of error, find the 95% confidence interval for the true proportion of voters preferring the Coalition. (Assume the margin of error provided is for a 95% level of confidence.)

b Based on this margin of error, approximately how many voters were surveyed?

c Why might Nielsen provide a single (maximum) margin of error in reporting on a variety of different outcomes (two-party-preferred vote, approval of the Prime Minister, approval of the Opposition Leader)?

d The approval of the Prime Minister was reported to be 40%. Find a 95% confidence interval for the true approval of the Prime Minister, based on the margin of error provided. Will this confidence interval be conservative (wider than it should be) or not (narrower than it should be)? Explain why.

When to use the Normal approximation

A guideline for when to use the Normal approximation for a confidence interval for p was given in the previous section: both x and $n - x$ should be greater than 10. These conditions are generally met for the illustrative data in figure 23, based on $n = 100$.

In contrast, the confidence intervals shown in figure 24 are based on data when $n = 10$; in no case is both x and $n - x$ greater than 10. In four of the ten cases presented, the confidence intervals are nonsense: either the lower or the upper bound is outside the range 0 to 1. Of course, proportions must be in the range 0 to 1. This shows that these intervals are wrong, and it indicates that we should not trust the approximation when n is this small.

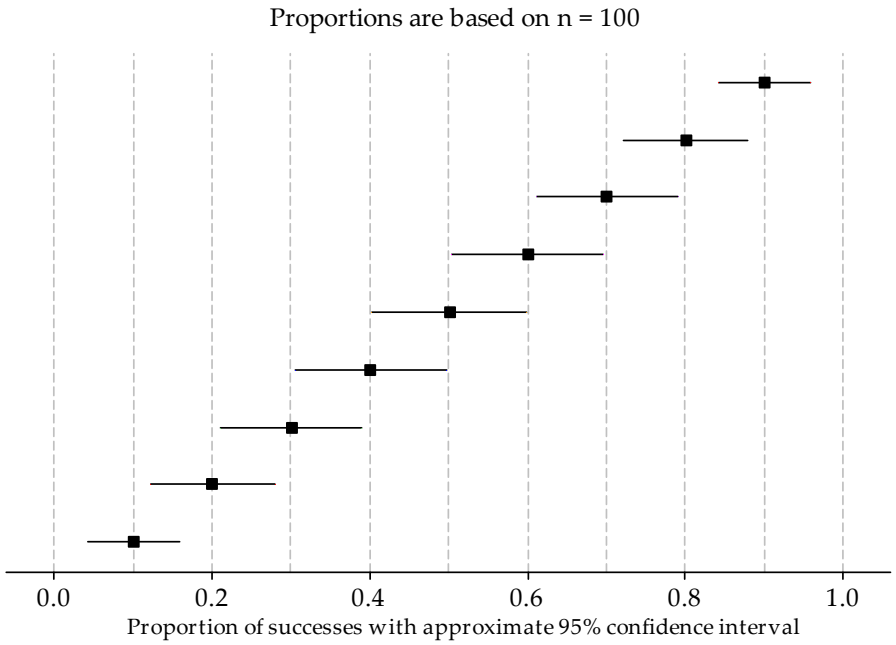

Figure 23: Approximate 95% confidence intervals for various estimates $\hat{p}$, with $n = 100$.

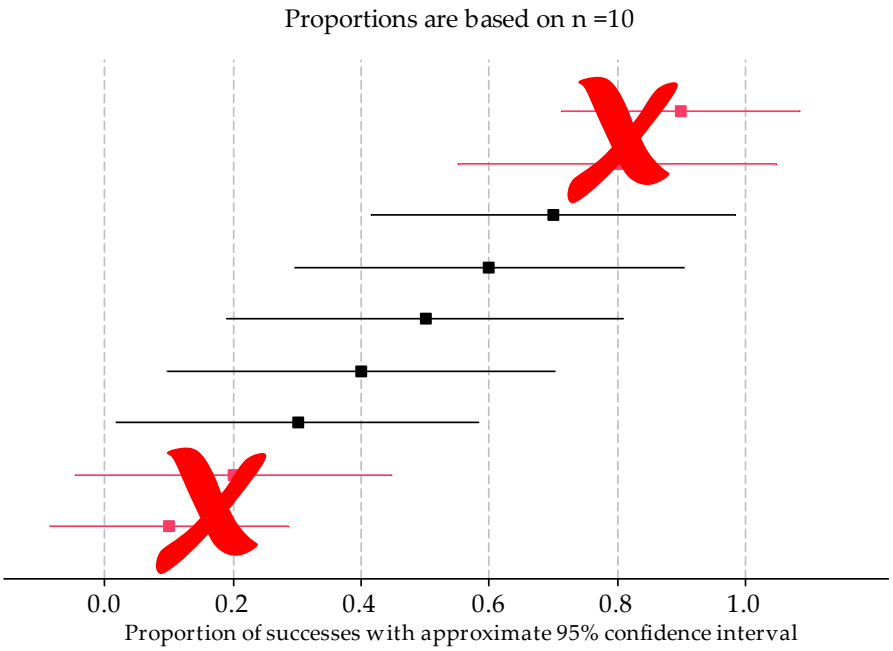

Figure 24: Approximate 95% confidence intervals for various estimates $\hat{p}$, with $n = 10$.

You may wonder whether it is possible to find a confidence interval for p when n is small, by avoiding the Normal approximation. The answer is that there is a method that uses only the binomial distribution, and does not approximate. It is beyond the scope of the curriculum.

Exercise 8

This exercise asks you to estimate π, using a statistical approach. Suppose you were aware that the area of a circle is $A = kr^2$, where r is the radius of the circle and k is a constant. Suppose you also knew that the equation for a circle centred at the origin is $x^2 + y^2 = r^2$. So, you knew a lot about circles ... but not the value of π.

a Consider the unit square in the real plane, with corners at $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$, and consider the circle of radius 1 centred at the origin. What proportion of the area of the square is covered by the circle, in terms of k? Define this proportion to be p.

b Use the following approach to estimate p, and hence k.

- **i** In an Excel spreadsheet, in columns A and B, store the variables x and y. In each column, put 10 000 observations from the $U(0,1)$ distribution. Recall that this is achieved by entering =RAND() in the first cell, and then filling down the column for 10 000 rows of data.
- **ii** In column C, calculate $x^2 + y^2$.
- **iii** In column D, evaluate whether or not $x^2 + y^2 < 1$, and store a '1' when this condition is satisfied, and a '0' otherwise.
- **iv** Use the 10 000 binary observations in column D to determine the proportion of the randomly generated points that are inside the circle. A simple way to do this is to average the values in column D. This is an estimate of p.

c Report the point estimate and approximate 95% confidence interval for p, based on this sample of size $n = 10\,000$.

d In fact, as you will realise, in estimating p you have estimated $\frac{\pi}{4} = 0.7854$. How precise is your estimate? Does your 95% confidence interval include the true value?

e Based on your 95% confidence interval for p, what is your point estimate and approximate 95% confidence interval for k?

Answers to exercises

Exercise 1

The following table gives $\text{sd}(\hat{P})$ for various values of p and n, to two decimal places. Note the symmetry in the table: $\text{sd}(\hat{P})$ is the same for $p = \theta$ and $p = 1 - \theta$.

n	$p = 0.1$	$p = 0.3$	$p = 0.5$	$p = 0.7$	$p = 0.9$
10	0.09	0.14	0.16	0.14	0.09
50	0.04	0.06	0.07	0.06	0.04
100	0.03	0.05	0.05	0.05	0.03

Exercise 2

a We can think of the m intervals as a sequence of m independent Bernoulli trials, with each trial having probability of success $p = 0.95$. Then Y is the number of successes in the m trials, where a success is counted whenever the confidence interval includes the unknown parameter value. So Y has a binomial distribution with parameters $n = m$ and $p = 0.95$, that is, $Y \overset{\text{d}}{=} \text{Bi}(m, 0.95)$.

b $\text{E}(Y) = 0.95m$.

c Now assume $m = 100$. Then $Y \overset{\text{d}}{=} \text{Bi}(100, 0.95)$.

- **i** The chance that exactly 95 intervals include the parameter is $\Pr(Y = 95) = 0.18$. This can be obtained in Excel using =BINOM.DIST(95, 100, 0.95, FALSE).
- **ii** The chance that at least 95 intervals include the parameter is $\Pr(Y \geq 95) = 0.62$. This can be obtained in Excel by summing the probabilities for values of Y from 95 to 100. But there is also a more direct method. To obtain $\Pr(Y = y)$ in Excel, we use FALSE as the fourth argument. If we use TRUE instead, the result is the cumulative probability $\Pr(Y \leq y)$. Note that $\Pr(Y \geq 95) = 1 - \Pr(Y \leq 94)$. We can find $\Pr(Y \leq 94)$ in Excel using =BINOM.DIST(94, 100, 0.95, TRUE). We obtain $\Pr(Y \geq 95) = 1 - \Pr(Y \leq 94) = 1 - 0.384 = 0.62$.

Exercise 3

A 100% confidence interval would mean that, in the long run, 100% of confidence intervals would include the true parameter value. In the case of estimating a proportion, we can be certain that the true proportion is between 0 and 1; hence, the 100% confidence interval is $(0, 1)$. That is the only way we could guarantee that every single confidence interval includes the true value. Of course, this is not a useful confidence interval in any practical sense. This reminds us, however, why we choose a confidence level less than 100%.

Exercise 4

a We use the general result that $\mathrm{E}(aX+b) = a\,\mathrm{E}(X)+b$, for constants a and b. Note that p and $\sqrt{\frac{p(1-p)}{n}}$ are constants, and so

$$\mathrm{E}\left(\frac{\hat{P}-p}{\sqrt{\frac{1}{n}p(1-p)}}\right) = \frac{\mathrm{E}(\hat{P}-p)}{\sqrt{\frac{1}{n}p(1-p)}} = \frac{\mathrm{E}(\hat{P})-p}{\sqrt{\frac{1}{n}p(1-p)}} = \frac{p-p}{\sqrt{\frac{1}{n}p(1-p)}} = 0.$$

b Using the general result $\mathrm{var}(aX+b) = a^2\,\mathrm{var}(X)$, we have

$$\mathrm{var}\left(\frac{\hat{P}-p}{\sqrt{\frac{1}{n}p(1-p)}}\right) = \frac{\mathrm{var}(\hat{P}-p)}{\frac{1}{n}p(1-p)} = \frac{\mathrm{var}(\hat{P})}{\frac{1}{n}p(1-p)} = \frac{\frac{1}{n}p(1-p)}{\frac{1}{n}p(1-p)} = 1.$$

The variance is 1, and so the standard deviation is 1.

Exercise 5

First we note that these data satisfy the guideline for the Normal approximation to be adequate: $x = 20$ and $n = 180$, so both x and $n-x$ are greater than 10.

a If the advertised claim is true, the proportion of winning wrappers Casey can expect to get (in a long-run average) is $\frac{1}{6} = 0.167$. So the expected number of winning wrappers in 180 purchases is $\frac{180}{6} = 30$.

b The proportion of winning wrappers in Casey's sample is $\frac{20}{180} = \frac{1}{9} = 0.111$.

c We have $\hat{p} = \frac{1}{9} = 0.111$, so

$$1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.111(1-0.111)}{180}} = 0.0459.$$

Thus the approximate 95% confidence interval is 0.1111 ± 0.0459, or $(0.0652, 0.1570)$. In percentage terms, the confidence interval is 6.52% to 15.70%.

d The 95% confidence interval for the true proportion is $(0.065, 0.157)$; these are values for the true proportion that are consistent with Casey's observation of 20 winning wrappers in a sample of 180 wrappers. The expected proportion of 0.167, according to the advertised claim, is outside the confidence interval; it is greater than the upper bound. Casey's sample of Venus bars provides some basis for being suspicious.

e The method used for finding the confidence interval assumes that Casey's sample of Venus bar wrappers is a random sample from the population of Venus bars produced for the promotion. Of course, if Casey buy Venus bars from shops with some old, pre-promotion stock, he would not expect to get $\frac{1}{6}$ winners. We assume that the 180 Bernoulli trials are independent; that is, Casey's success (or failure) in finding a winning wrapper on one day is not related to his success (or failure) on another day. In assessing this assumption, we need to think about the distribution of winning

wrappers and Casey's buying patterns. For example: Are bars with winning wrappers randomly mixed among all bars? Is there a limit on the number of winners per box? Does Casey always buy from the same place?

Exercise 6

a The bounds of the 99% confidence interval will be further from the point estimate than the bounds of the 95% confidence interval. Your estimate for the lower bound of the 99% confidence interval should be less than 0.065, and your estimate for the upper bound should be greater than 0.157.

b The value of the factor z from the standard Normal distribution for a 99% confidence interval is 2.576. (Reading it from figure 21 gives 2.6.) The ratio of the values of z for the 99% and 95% confidence intervals is $\frac{2.576}{1.96} = 1.3$. Hence, the margin of error for the 99% confidence interval will be 1.3 times greater than the margin of error for the 95% confidence interval. It will be about 0.06, making the 99% confidence interval about $(0.05, 0.17)$.

c We have $\hat{p} = \frac{1}{9} = 0.111$, so

$$2.576\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2.576\sqrt{\frac{0.111(1-0.111)}{180}} = 0.0603.$$

Hence, the 99% confidence interval is 0.1111 ± 0.0603, or $(0.0508, 0.1714)$. In percentage terms, the confidence interval is 5.08% to 17.14%.

d The 99% confidence interval includes the claimed true proportion of 16.7%.

Exercise 7

a Based on the margin of error provided, the approximate 95% confidence interval is 0.56 ± 0.026, or (0.534, 0.586); in percentage terms, it is (53.4%, 58.6%).

b For a 95% confidence interval, the margin of error is $1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. The maximum margin of error occurs when $\hat{p} = 0.5$. We have

$$1.96\sqrt{\frac{0.5(1-0.5)}{n}} = 0.026$$

$$\implies \sqrt{\frac{0.5 \times 0.5}{n}} = \frac{0.026}{1.96}$$

$$\implies \frac{0.25}{n} = \left(\frac{0.026}{1.96}\right)^2$$

$$\implies n = 0.25\left(\frac{1.96}{0.026}\right)^2 = 1421 \quad \text{(to the nearest whole number)}.$$

Hence, the sample size is 1421.

c Even if the sample size remains constant for the different outcomes that are reported, the margin of error will vary because it depends on $\hat{p}$. The reporting of the uncertainty in the survey results is simplified by reporting the margin of error that is the maximum for the sample size involved.

d Based on the margin of error provided, the approximate 95% confidence interval is 0.40 ± 0.026, or $(0.374, 0.426)$. This confidence interval is conservative (wider than it should be), because it is calculated using the maximum margin of error. The maximum margin of error arises when $\hat{p} = 0.5$, but in this case $\hat{p} = 0.4$. So the actual margin of error, based on the Normal approximation, is less than 0.026. Not by much, however, as $1.96\sqrt{\frac{0.4 \times 0.6}{1421}} = 0.025$.

Exercise 8

Since the circle has radius 1, it has area equal to k. As one quarter of the circle is in the unit square, the proportion of the area of the square that is covered by the circle is equal to $p = \frac{k}{4}$.

Because you are simulating data, there is no unique answer. If you can set up the Excel spreadsheet so it calculates everything (including the point estimate and approximate 95% confidence interval) based on formulas typed in cells, then hitting the F9 key will produce a new simulation and a different confidence interval; you can use the F9 key many times to see how often your confidence interval includes the true value.

The true value being estimated is $\frac{\pi}{4} = 0.7854$. The following table gives the point estimates and approximate 95% confidence intervals from five independent simulations.

Point estimate $\hat{p}$	Approximate 95% CI
0.7843	(0.7762, 0.7924)
0.7898	(0.7818, 0.7978)
0.7847	(0.7766, 0.7928)
0.7928	(0.7849, 0.8007)
0.7847	(0.7766, 0.7928)

As it happens, these 95% confidence intervals all include the true value.

An inference for $\frac{k}{4}$ can be converted into an inference for k by multiplying through by 4. For example, for the first result in the table, the point estimate for k is $4 \times 0.7843 = 3.137$, and the approximate 95% confidence interval for k is $(3.105, 3.169)$.

This is a different method for approximating π from the one used by Archimedes ...


0
1
2
3
4
5
6
7
8
9
10
11
12