Section 5.2: Confidence Intervals for a Proportion

Diez, D. M., Çetinkaya-Rundel, M., Barr, C. D. (2019). OpenIntro Statistics (4th ed.). OpenIntro. https://www.openintro.org/book/os/ CC BY-SA 3.0

STAT 1201
Introduction to Probability and Statistics

5.2 Confidence intervals for a proportion

The sample proportion $\hat{p}$ provides a single plausible value for the population proportion p. However, the sample proportion isn't perfect and will have some *standard error* associated with it. When stating an estimate for the population proportion, it is better practice to provide a plausible *range of values* instead of supplying just the point estimate.

5.2.1 Capturing the population parameter

Using only a point estimate is like fishing in a murky lake with a spear. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish. A **confidence interval** is like fishing with a net, and it represents a range of plausible values where we are likely to find the population parameter.

If we report a point estimate $\hat{p}$, we probably will not hit the exact population proportion. On the other hand, if we report a range of plausible values, representing a confidence interval, we have a good shot at capturing the parameter.

GUIDED PRACTICE 5.6

G If we want to be very certain we capture the population proportion in an interval, should we use a wider interval or a smaller interval?[6]

5.2.2 Constructing a 95% confidence interval

Our sample proportion $\hat{p}$ is the most plausible value of the population proportion, so it makes sense to build a confidence interval around this point estimate. The standard error provides a guide for how large we should make the confidence interval.

The standard error represents the standard deviation of the point estimate, and when the Central Limit Theorem conditions are satisfied, the point estimate closely follows a normal distribution. In a normal distribution, 95% of the data is within 1.96 standard deviations of the mean. Using this principle, we can construct a confidence interval that extends 1.96 standard errors from the sample proportion to be **95% confident** that the interval captures the population proportion:

$$\text{point estimate} \pm 1.96 \times SE$$
$$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

But what does "95% confident" mean? Suppose we took many samples and built a 95% confidence interval from each. Then about 95% of those intervals would contain the parameter, p. Figure 5.6 shows the process of creating 25 intervals from 25 samples from the simulation in Section 5.1.2, where 24 of the resulting confidence intervals contain the simulation's population proportion of $p = 0.88$, and one interval does not.

[6] If we want to be more certain we will capture the fish, we might use a wider net. Likewise, we use a wider confidence interval if we want to be more certain that we capture the parameter.

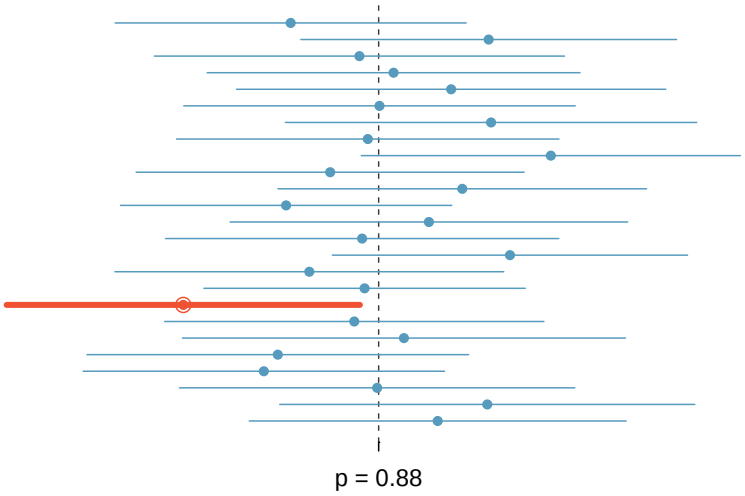

Figure 5.6: Twenty-five point estimates and confidence intervals from the simulations in Section 5.1.2. These intervals are shown relative to the population proportion $p = 0.88$. Only 1 of these 25 intervals did not capture the population proportion, and this interval has been bolded.

EXAMPLE 5.7

In Figure 5.6, one interval does not contain $p = 0.88$. Does this imply that the population proportion used in the simulation could not have been $p = 0.88$?

(E)

Just as some observations naturally occur more than 1.96 standard deviations from the mean, some point estimates will be more than 1.96 standard errors from the parameter of interest. A confidence interval only provides a plausible range of values. While we might say other values are implausible based on the data, this does not mean they are impossible.

95% CONFIDENCE INTERVAL FOR A PARAMETER

When the distribution of a point estimate qualifies for the Central Limit Theorem and therefore closely follows a normal distribution, we can construct a 95% confidence interval as

$$\text{point estimate} \pm 1.96 \times SE$$

EXAMPLE 5.8

In Section 5.1 we learned about a Pew Research poll where 88.7% of a random sample of 1000 American adults supported expanding the role of solar power. Compute and interpret a 95% confidence interval for the population proportion.

(E)

We earlier confirmed that $\hat{p}$ follows a normal distribution and has a standard error of $SE_{\hat{p}} = 0.010$. To compute the 95% confidence interval, plug the point estimate $\hat{p} = 0.887$ and standard error into the 95% confidence interval formula:

$$\hat{p} \pm 1.96 \times SE_{\hat{p}} \quad \rightarrow \quad 0.887 \pm 1.96 \times 0.010 \quad \rightarrow \quad (0.8674, 0.9066)$$

We are 95% confident that the actual proportion of American adults who support expanding solar power is between 86.7% and 90.7%. (It's common to round to the nearest percentage point or nearest tenth of a percentage point when reporting a confidence interval.)

5.2.3 Changing the confidence level

Suppose we want to consider confidence intervals where the confidence level is higher than 95%, such as a confidence level of 99%. Think back to the analogy about trying to catch a fish: if we want to be more sure that we will catch the fish, we should use a wider net. To create a 99% confidence level, we must also widen our 95% interval. On the other hand, if we want an interval with lower confidence, such as 90%, we could use a slightly narrower interval than our original 95% interval.

The 95% confidence interval structure provides guidance in how to make intervals with different confidence levels. The general 95% confidence interval for a point estimate that follows a normal distribution is

$$\text{point estimate} \ \pm \ 1.96 \times SE$$

There are three components to this interval: the point estimate, "1.96", and the standard error. The choice of $1.96 \times SE$ was based on capturing 95% of the data since the estimate is within 1.96 standard errors of the parameter about 95% of the time. The choice of 1.96 corresponds to a 95% confidence level.

GUIDED PRACTICE 5.9

(G) If X is a normally distributed random variable, what is the probability of the value X being within 2.58 standard deviations of the mean?[7]

Guided Practice 5.9 highlights that 99% of the time a normal random variable will be within 2.58 standard deviations of the mean. To create a 99% confidence interval, change 1.96 in the 95% confidence interval formula to be 2.58. That is, the formula for a 99% confidence interval is

$$\text{point estimate} \ \pm \ 2.58 \times SE$$

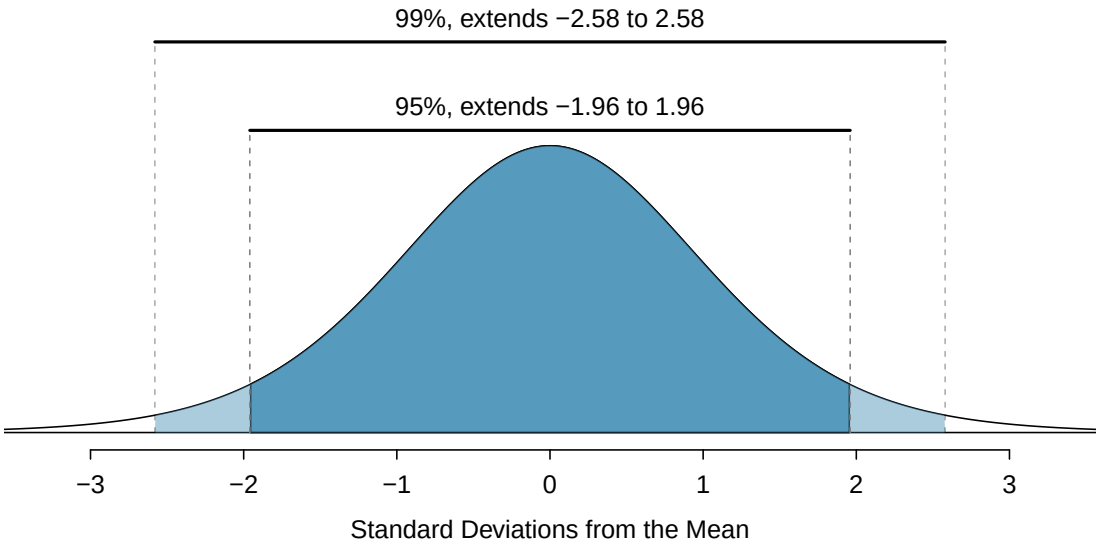

Figure 5.7: The area between -$z^\star$ and $z^\star$ increases as $z^\star$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of a normal normal distribution is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

[7] This is equivalent to asking how often the Z-score will be larger than -2.58 but less than 2.58. For a picture, see Figure 5.7. To determine this probability, we can use statistical software, a calculator, or a table to look up -2.58 and 2.58 for a normal distribution: 0.0049 and 0.9951. Thus, there is a $0.9951 - 0.0049 \approx 0.99$ probability that an unobserved normal random variable X will be within 2.58 standard deviations of μ.

This approach – using the Z-scores in the normal model to compute confidence levels – is appropriate when a point estimate such as $\hat{p}$ is associated with a normal distribution. For some other point estimates, a normal model is not a good fit; in these cases, we'll use alternative distributions that better represent the sampling distribution.

CONFIDENCE INTERVAL USING ANY CONFIDENCE LEVEL

If a point estimate closely follows a normal model with standard error SE, then a confidence interval for the population parameter is

$$\text{point estimate} \pm z^\star \times SE$$

where $z^\star$ corresponds to the confidence level selected.

Figure 5.7 provides a picture of how to identify $z^\star$ based on a confidence level. We select $z^\star$ so that the area between -$z^\star$ and $z^\star$ in the standard normal distribution, $N(0, 1)$, corresponds to the confidence level.

MARGIN OF ERROR

In a confidence interval, $z^\star \times SE$ is called the **margin of error**.

EXAMPLE 5.10

Use the data in Example 5.8 to create a 90% confidence interval for the proportion of American adults that support expanding the use of solar power. We have already verified conditions for normality.

(E) We first find $z^\star$ such that 90% of the distribution falls between -$z^\star$ and $z^\star$ in the standard normal distribution, $N(\mu = 0, \sigma = 1)$. We can do this using a graphing calculator, statistical software, or a probability table by looking for an upper tail of 5% (the other 5% is in the lower tail): $z^\star = 1.65$. The 90% confidence interval can then be computed as

$$\hat{p} \pm 1.65 \times SE_{\hat{p}} \quad \rightarrow \quad 0.887 \pm 1.65 \times 0.0100 \quad \rightarrow \quad (0.8705, 0.9035)$$

That is, we are 90% confident that 87.1% to 90.4% of American adults supported the expansion of solar power in 2018.

CONFIDENCE INTERVAL FOR A SINGLE PROPORTION

Once you've determined a one-proportion confidence interval would be helpful for an application, there are four steps to constructing the interval:

Prepare. Identify $\hat{p}$ and n, and determine what confidence level you wish to use.

Check. Verify the conditions to ensure $\hat{p}$ is nearly normal. For one-proportion confidence intervals, use $\hat{p}$ in place of p to check the success-failure condition.

Calculate. If the conditions hold, compute SE using $\hat{p}$, find $z^\star$, and construct the interval.

Conclude. Interpret the confidence interval in the context of the problem.

5.2.4 More case studies

In New York City on October 23rd, 2014, a doctor who had recently been treating Ebola patients in Guinea went to the hospital with a slight fever and was subsequently diagnosed with Ebola. Soon thereafter, an NBC 4 New York/The Wall Street Journal/Marist Poll found that 82% of New Yorkers favored a "mandatory 21-day quarantine for anyone who has come in contact with an Ebola patient". This poll included responses of 1,042 New York adults between Oct 26th and 28th, 2014.

EXAMPLE 5.11

What is the point estimate in this case, and is it reasonable to use a normal distribution to model that point estimate?

E The point estimate, based on a sample of size $n = 1042$, is $\hat{p} = 0.82$. To check whether $\hat{p}$ can be reasonably modeled using a normal distribution, we check independence (the poll is based on a simple random sample) and the success-failure condition ($1042 \times \hat{p} \approx 854$ and $1042 \times (1 - \hat{p}) \approx 188$, both easily greater than 10). With the conditions met, we are assured that the sampling distribution of $\hat{p}$ can be reasonably modeled using a normal distribution.

EXAMPLE 5.12

Estimate the standard error of $\hat{p} = 0.82$ from the Ebola survey.

E We'll use the substitution approximation of $p \approx \hat{p} = 0.82$ to compute the standard error:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.82(1-0.82)}{1042}} = 0.012$$

EXAMPLE 5.13

Construct a 95% confidence interval for p, the proportion of New York adults who supported a quarantine for anyone who has come into contact with an Ebola patient.

E Using the standard error $SE = 0.012$ from Example 5.12, the point estimate 0.82, and $z^\star = 1.96$ for a 95% confidence level, the confidence interval is

$$\text{point estimate} \pm z^\star \times SE \quad \rightarrow \quad 0.82 \pm 1.96 \times 0.012 \quad \rightarrow \quad (0.796, 0.844)$$

We are 95% confident that the proportion of New York adults in October 2014 who supported a quarantine for anyone who had come into contact with an Ebola patient was between 0.796 and 0.844.

GUIDED PRACTICE 5.14

G Answer the following two questions about the confidence interval from Example 5.13:[8]

(a) What does 95% confident mean in this context?

(b) Do you think the confidence interval is still valid for the opinions of New Yorkers today?

[8] (a) If we took many such samples and computed a 95% confidence interval for each, then about 95% of those intervals would contain the actual proportion of New York adults who supported a quarantine for anyone who has come into contact with an Ebola patient.
(b) Not necessarily. The poll was taken at a time where there was a huge public safety concern. Now that people have had some time to step back, they may have changed their opinions. We would need to run a new poll if we wanted to get an estimate of the current proportion of New York adults who would support such a quarantine period.

GUIDED PRACTICE 5.15

In the Pew Research poll about solar energy, they also inquired about other forms of energy, and 84.8% of the 1000 respondents supported expanding the use of wind turbines.[9]

(G)

(a) Is it reasonable to model the proportion of US adults who support expanding wind turbines using a normal distribution?

(b) Create a 99% confidence interval for the level of American support for expanding the use of wind turbines for power generation.

We can also construct confidence intervals for other parameters, such as a population mean. In these cases, a confidence interval would be computed in a similar way to that of a single proportion: a point estimate plus/minus some margin of error. We'll dive into these details in later chapters.

5.2.5 Interpreting confidence intervals

In each of the examples, we described the confidence intervals by putting them into the context of the data and also using somewhat formal language:

Solar. We are 90% confident that 87.1% to 90.4% of American adults support the expansion of solar power in 2018.

Ebola. We are 95% confident that the proportion of New York adults in October 2014 who supported a quarantine for anyone who had come into contact with an Ebola patient was between 0.796 and 0.844.

Wind Turbine. We are 99% confident the proportion of Americans adults that support expanding the use of wind turbines is between 81.9% and 87.7% in 2018.

First, notice that the statements are always about the population parameter, which considers *all* American adults for the energy polls or *all* New York adults for the quarantine poll.

We also avoided another common mistake: *incorrect* language might try to describe the confidence interval as capturing the population parameter with a certain probability. Making a probability interpretation is a common error: while it might be useful to think of it as a probability, the confidence level only quantifies how plausible it is that the parameter is in the given interval.

Another important consideration of confidence intervals is that they are *only about the population parameter*. A confidence interval says nothing about individual observations or point estimates. Confidence intervals only provide a plausible range for population parameters.

Lastly, keep in mind the methods we discussed only apply to sampling error, not to bias. If a data set is collected in a way that will tend to systematically under-estimate (or over-estimate) the population parameter, the techniques we have discussed will not address that problem. Instead, we rely on careful data collection procedures to help protect against bias in the examples we have considered, which is a common practice employed by data scientists to combat bias.

GUIDED PRACTICE 5.16

(G)

Consider the 90% confidence interval for the solar energy survey: 87.1% to 90.4%. If we ran the survey again, can we say that we're 90% confident that the new survey's proportion will be between 87.1% and 90.4%?[10]

[9] (a) The survey was a random sample and counts are both ≥ 10 ($1000 \times 0.848 = 848$ and $1000 \times 0.152 = 152$), so independence and the success-failure condition are satisfied, and $\hat{p} = 0.848$ can be modeled using a normal distribution. (b) Guided Practice 5.15 confirmed that $\hat{p}$ closely follows a normal distribution, so we can use the C.I. formula:

$$\text{point estimate} \pm z^\star \times SE$$

In this case, the point estimate is $\hat{p} = 0.848$. For a 99% confidence interval, $z^\star = 2.58$. Computing the standard error: $SE_{\hat{p}} = \sqrt{\frac{0.848(1-0.848)}{1000}} = 0.0114$. Finally, we compute the interval as $0.848 \pm 2.58 \times 0.0114 \rightarrow (0.8186, 0.8774)$. It is also important to *always* provide an interpretation for the interval: we are 99% confident the proportion of American adults that support expanding the use of wind turbines in 2018 is between 81.9% and 87.7%.

[10] No, a confidence interval only provides a range of plausible values for a parameter, not future point estimates.