7. QUANTIFYING UNCERTAINTY

R symbols, operators, and functions introduced in this chapter: `nrow()`, `predict()`, `abs()`, and `summary()$coef`.

In the previous chapters, we analyzed data to estimate different quantities of interest. For example, in chapter 2, we analyzed data from Project STAR to estimate the average causal effect of attending a small class on students' reading test scores. In chapter 3, we analyzed data from the BES survey to estimate the proportion of UK voters in favor of Brexit. In chapter 4, we analyzed data from 170 countries to predict GDP growth using night-time light emissions. Finally, in chapter 5, we analyzed data from a survey of Ukrainians to estimate the average causal effect of receiving Russian TV on respondents' voting behavior. The results we arrived at in each of these analyses are, at best, applicable only to the sample of observations we analyzed. In most instances, however, we want to generalize our conclusions to the population from which the sample of observations was drawn. To do so, we need to account for sampling variability, which introduces uncertainty and makes the sample-level estimates different from the population-level quantities of interest. In this chapter, we learn how to quantify the degree of uncertainty in our estimates. As illustrations, we revisit each of the aforementioned analyses.

7.1 ESTIMATORS AND THEIR SAMPLING DISTRIBUTIONS

As we saw in chapter 6, when analyzing data, we are usually interested in a quantity at the population level, yet we typically have access to only a sample of observations. For example, in chapter 3, we were interested in the level of support for Brexit among all UK voters, but we knew only the proportion of supporters among BES survey respondents.

A **parameter** is an unknown quantity of interest. An **estimate** is a sample-level statistic that estimates a parameter. An **estimator** is a function of observed data that is used to produce an estimate of a parameter.

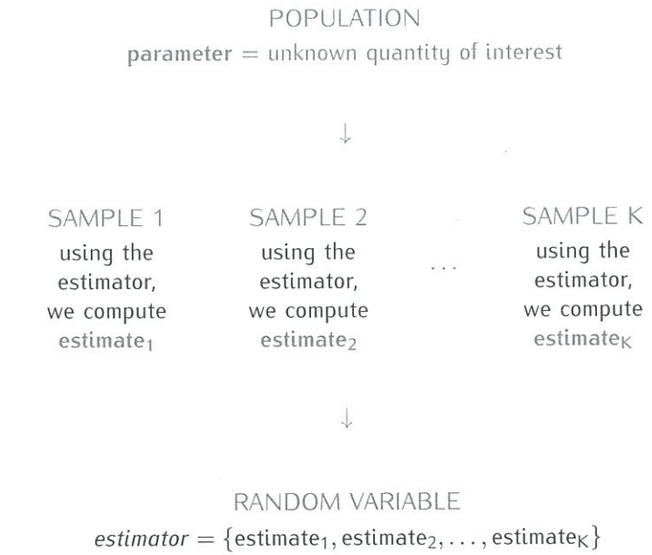
We call the unknown quantity of interest the **parameter**. (Parameters can be sample-level quantities, but we focus on population-level parameters.) We call the statistic that we compute using the sample data the **estimate**, and the formula that produces it, an **estimator**. Formally, an estimator is a function of observed data used to produce an estimate of a parameter.

In this book, we have seen how to use four estimators:

- In chapter 2, we used the difference-in-means estimator to estimate average treatment effects with a randomized experiment.
- In chapter 3, we used the sample mean to estimate population-level averages and proportions.
- In chapter 4, we used a fitted linear model to predict outcomes.
- In chapter 5, we used the coefficients of a fitted linear model to estimate average treatment effects with observational data.

In each of these analyses, we used an estimator to produce an estimate of the corresponding parameter. These estimates, however, are not necessarily identical to the parameters we are interested in. As we saw in the previous chapter, sample statistics differ from population-level parameters because each sample is only a subset of the target population, and sample statistics vary from one sample to another. In the case of the BES survey, the respondents account for only a tiny fraction of all UK voters. As a result, the sample proportion of survey respondents in favor of Brexit is not necessarily the same as the population proportion of *all* UK voters in favor of Brexit. In technical terms, our estimates have some uncertainty due to sampling variability.

Our goal, then, is to quantify the uncertainty in our estimates so that we can draw conclusions about the parameter. Since the value of an estimator varies from one sample to another, we can think of an estimator as a random variable.



The **sampling distribution** of an estimator characterizes the degree to which the estimator varies from one sample to another due to sampling variability. The **standard error** of an estimator is the estimated standard deviation of the sampling distribution of the estimator.

The **sampling distribution** of this random variable characterizes the variability of the estimator from one sample to another, and in relation to the population-level parameter. To quantify the amount of uncertainty in our estimates, then, we need to characterize this sampling distribution.

At the end of the previous chapter, we used the central limit theorem to derive the sampling distribution of the sample mean. We can do the same for the other estimators.

One implication of the central limit theorem is that all the estimators covered in this book have a sampling distribution that is approximately normal and centered at the population-level parameter. (Note that we always assume that the samples are large enough that we can reliably use the central limit theorem.)

To quantify the variation around the population-level parameter, we need to measure the spread of the sampling distribution of the estimator. As we saw in chapter 3, we can use the standard deviation of a random variable to measure the spread of its distribution. Unfortunately, in most cases we cannot compute the standard deviation of the sampling distribution of an estimator directly. Doing so would require drawing multiple samples from the target population, but we rarely have access to more than one sample. Instead, we estimate the standard deviation based on the one sample we draw. We refer to the estimated standard deviation of the sampling distribution of an estimator as the **standard error** of the estimator. (The formula for the standard error is different for each estimator; some of these formulas are quite complicated and beyond the scope of this book.)

Figure 7.1 depicts the sampling distribution of an estimator. As we can see, the standard error quantifies the degree of uncertainty of the estimator due to sampling variability. It measures the amount of variation of the estimator around the true value of the population-level parameter.

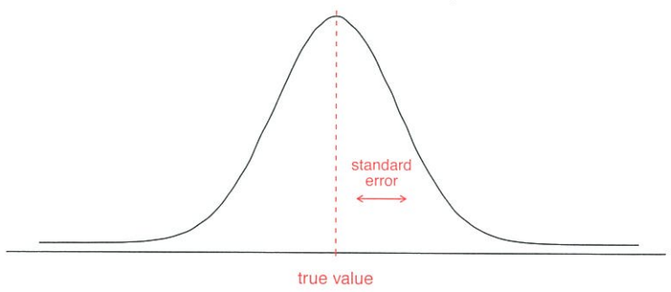
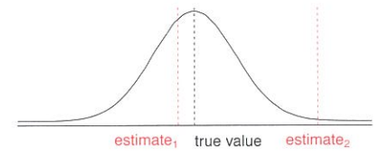


FIGURE 7.1. Sampling distribution of an estimator. All the estimators covered in this book have a sampling distribution that is approximately normal and centered at the true value of the population-level parameter. The standard error of an estimator quantifies the spread of its sampling distribution, which is a measure of the degree of uncertainty of the estimator.

Note that since we usually draw only one sample from the population, we can compute only one value of the estimator. This one estimate might be close to the true value of the parameter (as is, for example, the value of estimate$_1$ in the figure in the margin), or it might be quite far away (as is the value of estimate$_2$). When working with only one sample of data, we never know how far our estimate is from the true value since the true value is unknown.



The **estimation error** is the difference between the estimate and the true value of the parameter. The **average estimation error**, also known as bias, is the average difference between the estimate and the true value of the parameter over multiple hypothetical samples. An estimator is said to be **unbiased** if the average estimation error over multiple hypothetical samples is zero. The **standard error** is an estimate of the average size of the estimation error over multiple hypothetical samples.

The difference between the estimate and the true value is called the **estimation error**:

$$\textit{estimation error}_i = \text{estimate}_i - \text{true value}$$

where:

- $\textit{estimation error}_i$ is the estimation error for sample i
- estimate_i is the estimate for sample i
- true value is the true value of the population-level parameter.

In the hypothetical cases above, the estimation errors would be (estimate$_1$−true value) and (estimate$_2$−true value).

While we can never compute the estimation error of a particular estimate, we can calculate two helpful statistics about the estimation error using the central limit theorem.

First, we can derive the **average estimation error**, also known as bias, over multiple hypothetical samples.

FORMULA IN DETAIL

In mathematical notation, the average estimation error is:

$$\textit{average estimation error} = \mathbb{E}(\text{estimate}_i - \text{true value})$$

where:

- $\mathbb{E}$ is the population mean
- estimate_i is the estimate for sample i
- true value is the true value of the population-level parameter, which equals the population mean of the sampling distribution of the estimator.

An estimator is said to be **unbiased** if the average estimation error over multiple hypothetical samples is zero. While the details are beyond the scope of this book, it is worth noting that all the estimators covered here are unbiased estimators of their corresponding parameters. They provide, on average, accurate estimates. This is consistent with the fact that our estimators all have a sampling distribution that is centered at the true value of the population-level parameter.

Second, we can derive the average *size* of the estimation error over multiple hypothetical samples. This is actually what the standard error of the estimator measures. (As we saw in chapter 3, the standard deviation of a random variable measures the average distance of the observations to the mean; here, this average distance is equivalent to the average size of the estimation error.)

FORMULA IN DETAIL

In mathematical notation, the standard error of the estimator is:

$$\begin{aligned} \textit{standard error} &= \sqrt{\mathbb{V}(\textit{estimator})} && \text{because standard deviation} = \sqrt{\mathbb{V}(X)} \\ &= \sqrt{\mathbb{E}[(\text{estimate}_i - \mathbb{E}(\textit{estimator}))^2]} && \text{because } \mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] \\ &= \sqrt{\mathbb{E}[(\text{estimate}_i - \text{true value})^2]} && \text{because } \mathbb{E}(\textit{estimator}) = \text{true value} \end{aligned}$$

where:

- $\mathbb{V}$ is the population variance, and $\mathbb{E}$ is the population mean
- *estimator* is a random variable across multiple hypothetical samples, and estimate_i is the estimate for sample i
- true value is the true value of the population-level parameter, which equals the population mean of the sampling distribution of the estimator.

Note the difference between the average estimation error and the standard error. In the formula for the average estimation error, positive errors cancel out negative errors. By contrast, in the formula for the standard error, positive errors do not cancel out negative errors. (The errors are squared so that they are all positive. Then, after computing the average squared error, we take the square root to return to the initial unit of measurement.) As a result, these two statistics generally differ from each other. While the estimators we use have average estimation errors that equal zero, their standard errors usually do not equal zero.

Putting it all together, if we were to draw multiple samples from the target population and calculate the estimate for each sample, the random variable *estimator* would be approximately distributed as follows:

$$\textit{estimator} \overset{\text{approx.}}{\sim} N\left(\text{true value},\ (\text{standard error})^2\right)$$

where:

- *estimator* is a random variable containing the estimates from multiple hypothetical samples
- $\overset{\text{approx.}}{\sim}$ stands for "approximately distributed according to"
- N stands for "normal distribution," the first number inside the parentheses denotes the mean of the normal distribution, and the second number denotes the variance

- true value is the true value of the population-level parameter
- standard error is the estimated standard deviation of the estimator across multiple samples, so $(\text{standard error})^2$ is the estimated variance of the estimator across multiple samples.

Since the sampling distributions of all the estimators in this book can be approximated with the normal distribution, we can standardize the estimators using formula 6.1.

For each of the estimators, then, if we drew multiple samples from the same target population and computed the standardized estimate for each sample, the resulting statistic would approximately follow the standard normal distribution. (See formula 7.1.)

THE STANDARDIZED ESTIMATOR

$$\frac{\textit{estimator} - \text{true value}}{\text{standard error}} \overset{\text{approx.}}{\sim} N(0, 1)$$

where:

- *estimator* is a random variable across multiple hypothetical samples
- true value is the true value of the population-level parameter
- standard error is the estimated standard deviation of the estimator across multiple samples.

FORMULA 7.1. Formula of the standardized estimator and its distribution across multiple hypothetical samples.

As we will see in detail below, we can use this distribution to draw conclusions about a population-level parameter. In particular, we can use it for two purposes:

First, we can use the sampling distribution to compute confidence intervals. The confidence interval of an estimator provides the range of values that is likely to include the true value of the parameter. In section 7.2, we learn how to construct confidence intervals for the sample mean, the difference-in-means estimator, and the predicted value of an outcome.

Second, we can use the sampling distribution for hypothesis testing. Through hypothesis testing, we determine whether the true value of a parameter is likely to equal a particular value. For example, we may want to determine whether an average treatment effect is different from zero at the population level. In section 7.3, we learn to use hypothesis testing with the difference-in-means estimator as well as with estimated regression coefficients.

7.2 CONFIDENCE INTERVALS

A **confidence interval** provides the range of values that is likely to include the true value of the parameter.

Three levels of confidence are conventionally used in the social sciences to construct confidence intervals: 90%, 95%, and 99%. The level of confidence indicates the probability, over multiple samples, that the true value lies within the interval. With higher levels of confidence, the degree of uncertainty decreases, but the width of the confidence interval increases. The most commonly used level of confidence is 95%, so that is what we use here.

To construct a 95% confidence interval, we start with one of the properties of the standard normal distribution. As we saw in chapter 6, about 95% of the observations in the standard normal random variable fall between -1.96 and 1.96. In mathematical notation, if Z is the standard normal random variable, then:

$$P\,(\text{-}1.96 \leq Z \leq 1.96) \;\approx\; 0.95$$

Since the standardized estimator approximately follows the standard normal distribution (see formula 7.1), over multiple samples, 95% of the standardized estimators fall between -1.96 and 1.96.

$$P\left(\text{-}1.96 \leq \frac{\textit{estimator} - \text{true value}}{\text{standard error}} \leq 1.96\right) \;\approx\; 0.95$$

After moving terms around to isolate the true value, we arrive at:

$$P\,(\,\textit{estimator} - 1.96 \times \text{standard error} \leq \text{true value} \leq \textit{estimator} + 1.96 \times \text{standard error}\,) \;\approx\; 0.95$$

Given the probability above, we can define the 95% confidence interval of an estimator as shown in formula 7.2.

95% CONFIDENCE INTERVAL

$$95\%\ \text{CI} = [\ \textit{estimator} - 1.96 \times \text{standard error},\ \textit{estimator} + 1.96 \times \text{standard error}\ ]$$

where:

- *estimator* is a random variable across multiple hypothetical samples
- standard error is the estimated standard deviation of the estimator across multiple hypothetical samples.

A **confidence interval** provides the range of values that is likely to include the true value of the parameter.

TIP: We focus on this property of the standard normal distribution because we want to construct the 95% confidence interval. If we wanted to construct the 99% confidence interval instead, for example, we would start with the fact that in the standard normal distribution, about 99% of the observations are between -2.58 and 2.58. The resulting confidence interval would be much wider.

FORMULA 7.2. Formula to construct the 95% confidence interval, the 95% CI for short. In 95% of the samples, the 95% confidence interval constructed using this formula will contain the true value of the parameter.

This confidence interval provides bounds on where the true value of the parameter is likely to be. Since the confidence level of the interval is 95%, if we were to draw multiple samples from the same population, 95% of the intervals constructed using this formula should contain the true value of the parameter. In other words, the confidence level of the interval refers to the probability that the interval contains the true value *over multiple samples*.

In reality, as we have already discussed, we usually draw only one sample. As a result, we can construct only one confidence interval. This one confidence interval may or may not contain the true value. Discerning whether it does or not is impossible since we do not know the true value.

Thanks to the central limit theorem, we know that in 5% of the samples, the 95% confidence interval will *not* contain the true value of the parameter. Unfortunately, we have no way of knowing whether we happen to be analyzing one of those fringe samples. This is why it is so important to replicate social scientific studies, that is, to arrive at similar conclusions when analyzing a different sample of data from the same target population. While getting one unlucky sample occurs 5% of the time, getting two unlucky independent samples in a row occurs only 0.25% of the time.

It is important to replicate social scientific studies to confirm that we arrive at similar conclusions when analyzing a different sample from the same target population.

Now that we know the general formula for constructing confidence intervals, let's see how we can use it to construct the confidence interval for the following three estimators: (i) the sample mean, (ii) the difference-in-means estimator, and (iii) predicted outcomes from a fitted linear model.

7.2.1 CONFIDENCE INTERVAL FOR THE SAMPLE MEAN

Let's return to the analysis of chapter 3. There, we analyzed data from the BES survey conducted before the 2016 Brexit referendum to measure public opinion among the entire UK population.

TIP: The code for this chapter's analysis can be found in the "Uncertainty.R" file.

By running the following code, we (i) read and store the dataset in an object named *bes*, (ii) eliminate observations with missing data (including observations from respondents who were either undecided or did not intend to vote) and store the new dataset in an object named *bes1*, (iii) show the number of observations and the number of variables in the *bes1* dataset, and (iv) show the first six observations. (Remember to first set the working directory so that R knows where to find the CSV file.)

RECALL: If the DSS folder is saved directly on your Desktop, to set the working directory, you must run `setwd("~/Desktop/DSS")` if you have a Mac and `setwd("C:/user/Desktop/DSS")` if you have a Windows computer (where *user* is your own username). If the DSS folder is saved elsewhere, please see subsection 1.7.1 for instructions on how to set the working directory.

```
bes <- read.csv("BES.csv") # reads and stores data

bes1 <- na.omit(bes) # eliminates observations with NAs

dim(bes1) # provides dimensions of dataframe: rows, columns
## [1] 25097     4
```

```
head(bes1) # shows first observations
##   vote leave education age
## 1 leave     1         3  60
## 3 stay      0         5  73
## 4 leave     1         4  64
## 6 stay      0         4  85
## 7 leave     1         3  78
## 8 leave     1         2  51
```

As you may recall, *leave* is a binary variable that identifies Brexit supporters, that is, respondents who intended to vote "leave".

If we want to know the proportion of BES respondents who were in favor of Brexit, we can calculate the mean of *leave*, since the mean of a binary variable is equivalent to the proportion of the observations that have the characteristic identified by the variable. The mean of *leave* can be calculated by running:

```
mean(bes1$leave) # calculates the mean
## [1] 0.4718891
```

Based on the output, we can state that 47.19% of BES respondents were in favor of Brexit (0.4719×100=47.19%).

Can we infer from this that about 47% of *all* UK voters were in favor of Brexit? We cannot. This is a sample-level result. To draw conclusions at the population level, we need to take into consideration the noise introduced by sampling variability.

We can construct a measure of uncertainty for the sample mean. In particular, we can derive the 95% confidence interval by substituting in formula 7.2 the sample mean and its standard error. This results in the following 95% confidence interval:

95% CONFIDENCE INTERVAL FOR THE SAMPLE MEAN

$$\left[\overline{Y} - 1.96 \times \sqrt{\frac{var(Y)}{n}}, \quad \overline{Y} + 1.96 \times \sqrt{\frac{var(Y)}{n}} \right]$$

where:

- $\overline{Y}$ is the sample mean of Y
- $\sqrt{var(Y)/n}$ is the standard error of the sample mean
- $var(Y)$ is the sample variance of Y
- n is the number of observations in the sample.

This interval provides the range of values that is likely to contain the true value of the population mean of Y, or $\mathbb{E}(Y)$.

In the running example, to compute the confidence interval for the sample mean of *leave*, we start by computing and storing the sample size, *n*, into an object so that we can more easily operate with its value.

To compute the sample size of a dataframe, we can use the function nrow(), which stands for "number of rows." The only required argument is the name of the object where the dataset is stored. Here, to compute and store the sample size in an object named *n*, we run:

nrow() computes the number of rows of a dataframe. The only required argument is the name of the object where the dataframe is stored. Example: nrow(*data*).

```
n <- nrow(bes1) # computes and stores n
```

Now we can compute the lower limit of the interval by running:

```
## calculate lower limit of the 95% CI for sample mean
mean(bes1$leave) - 1.96 * sqrt(var(bes1$leave) / n)
## [1] 0.4657127
```

And, we can compute the upper limit by running:

```
## calculate upper limit of the 95% CI for sample mean
mean(bes1$leave) + 1.96 * sqrt(var(bes1$leave) / n)
## [1] 0.4780655
```

Based on the outputs above, we conclude that the true proportion of support for Brexit among *all* UK voters was likely to be between 46.57% and 47.81%.

There is an alternative way of expressing confidence intervals, which is popular in the world of polling. It involves using what is known as the **margin of error**, defined as half the width of the confidence interval. Using this term, we can express the confidence interval as:

estimator ± margin of error

The **margin of error** of an estimator is defined as half the width of the estimator's confidence interval. As a result, we can express the confidence interval as:

estimator ± margin of error

In this case, the margin of error equals 0.62 percentage points. (The width of the confidence interval is 47.81%−46.57%=1.24 p.p.; half of that is 0.62 p.p.) Thus, we can state that the likely proportion of support for Brexit among all UK voters was 47.19% with a margin of error of 0.62 percentage points.

RECALL: The difference between two percentages is measured in percentage points (%−%=p.p.).

TIP: Here, the 95% confidence interval can be expressed as either [46.57%, 47.81%] or 47.19% ± 0.62 p.p.

The margin of error here is small because, as we computed earlier, the BES survey has a large sample size of 25,097 observations. Most polls have a much smaller sample size, of about 1,000 observations, and as a result their margins of error are much larger. (As the sample size, *n*, decreases, the width of the confidence interval increases.) In general, the degree of uncertainty of our estimates will be larger with smaller sample sizes.

7.2.2 CONFIDENCE INTERVAL FOR THE DIFFERENCE-IN-MEANS ESTIMATOR

We can use a similar procedure to construct the confidence interval for the difference-in-means estimator. Let's return to the analysis of chapter 2. There, we analyzed data from Project STAR, an experiment in which students were randomly assigned to attend either a small class or a regular-size class.

RECALL: If the DSS folder is saved directly on your Desktop, to set the working directory, you must run `setwd("~/Desktop/DSS")` if you have a Mac and `setwd("C:/user/Desktop/DSS")` if you have a Windows computer (where *user* is your own username). If the DSS folder is saved elsewhere, please see subsection 1.7.1 for instructions on how to set the working directory.

By running the following code, we (i) read and store the dataset in an object named *star*, (ii) show the number of observations and the number of variables in the dataset, (iii) show the first six observations, and (iv) create a new binary variable named *small* identifying the students who were assigned to attend a small class. (Remember to first set the working directory.)

```
star <- read.csv("STAR.csv") # reads and stores data

dim(star) # provides dimensions of dataframe: rows, columns
## [1] 1274    4

head(star) # shows first  observations
##    classtype reading math graduated
## 1      small     578  610         1
## 2    regular     612  612         1
## 3    regular     583  606         1
## 4      small     661  648         1
## 5      small     614  636         1
## 6    regular     610  603         0

star$small <- ifelse(star$classtype=="small",
                     1, 0) # creates the treatment variable
```

As you may recall, the purpose of the analysis was to estimate the average causal effect of attending a small class on three measures of student performance: third-grade reading test scores, third-grade math test scores, and the probability of graduating from high school. We focus here on the causal effect on the reading scores.

Because the treatment was randomly assigned, we can assume that students who attended a small class were comparable before schooling to students who attended a regular-size class. As a result, we can use the difference-in-means estimator to estimate the average treatment effect.

To calculate the difference-in-means estimator for reading, we run the following piece of code:

```
## compute the difference-in-means estimator for reading
mean(star$reading[star$small==1]) -
  mean(star$reading[star$small==0])
## [1] 7.210547
```


Based on the output, we can state that among the students who participated in Project STAR, attending a small class increased performance on the third-grade reading test by an estimated 7.21 points, on average. This value is the estimated average treatment effect for the sample of 1,274 students who participated in the experiment. How about at the population level? What would have been the average causal effect of attending a small class on the entire population of students from which the sample was drawn?

We can construct a measure of uncertainty for the difference-in-means estimator. We can derive the 95% confidence interval by substituting in formula 7.2 the difference-in-means estimator and its standard error:

95% CONFIDENCE INTERVAL FOR THE DIFFERENCE-IN-MEANS ESTIMATOR

LOWER LIMIT:

$$\overline{Y}_{\substack{\text{treatment}\\\text{group}}} - \overline{Y}_{\substack{\text{control}\\\text{group}}} - 1.96 \times \sqrt{\frac{var(Y_{\text{treatment}})}{n_{\text{treatment group}}} + \frac{var(Y_{\text{control}})}{n_{\text{control group}}}}$$

UPPER LIMIT:

$$\overline{Y}_{\substack{\text{treatment}\\\text{group}}} - \overline{Y}_{\substack{\text{control}\\\text{group}}} + 1.96 \times \sqrt{\frac{var(Y_{\text{treatment}})}{n_{\text{treatment group}}} + \frac{var(Y_{\text{control}})}{n_{\text{control group}}}}$$

where:

- $\overline{Y}_{\substack{\text{treatment}\\\text{group}}} - \overline{Y}_{\substack{\text{control}\\\text{group}}}$ is the difference-in-means estimator
- $\sqrt{var(Y_{\text{treatment}})/n_{\text{treatment group}} + var(Y_{\text{control}})/n_{\text{control group}}}$ is the standard error of the difference-in-means estimator
- $var(Y_{\text{treatment}})$ and $var(Y_{\text{control}})$ are the sample variances of Y under the treatment and control conditions
- $n_{\text{treatment group}}$ and $n_{\text{control group}}$ are the number of observations in the treatment and the control groups in the sample.

To compute the confidence interval for the difference-in-means estimator, we start by creating two separate dataframes, one for the treatment group and one for the control group. This will help

[] is the operator used to extract a selection of observations from a dataframe. To its left, we specify the dataframe we want to subset. Inside the square brackets, we specify the criterion of selection. Since a dataframe is composed of two dimensions, rows and columns, we can specify a criterion of selection on one or both dimensions. First, we specify the criterion of selection of the rows, and then the criterion of selection of the columns (separated by a comma). If the first criterion is left blank, all rows are extracted, and if the second criterion is left blank, all columns are extracted. Example: *data[data$var1==1,]* extracts the observations that have a value of 1 in *var1* as well as their corresponding values in all the other variables in the dataframe *data*.

simplify our computations. To subset the original dataframe, we can use the [] operator. To its left, we specify the dataframe we want to subset, star in this case. Inside the square brackets we specify (1) the criterion of selection of the rows, and (2) the criterion of selection of the columns (in this order and separated by a comma). To extract the observations that refer to the treatment group, we use star$small==1 as the criterion of selection of rows. To extract the observations that refer to the control group, we use star$small==0 as the criterion of selection of rows. (As you may recall, we can use the relational operator == to specify a logical test.) In both cases, we leave the criterion of selection of columns blank, indicating that we want to extract all variables. To subset and store as new objects the two dataframes, then, we run:

```
## create separate dataframes for each group
treatment <- star[star$small==1, ] # for the treatment group
control <- star[star$small==0, ] # for the control group
```

Next, we can compute and store as a new object the sample size of each of the two dataframes by running:

```
## compute and store sample sizes for each group
n_t <- nrow(treatment) # for the treatment group
n_c <- nrow(control) # for the control group
```

Now, to compute the lower limit of the 95% confidence interval for the difference-in-means estimator, we run:

```
## calculate lower limit of 95% CI for diffs-in-means
mean(treatment$reading) - mean(control$reading) -
  1.96 * sqrt(var(treatment$reading) / n_t
              + var(control$reading) / n_c)
## [1] 3.167621
```

And, to compute the upper limit, we run:

```
## calculate upper limit of 95% CI for diffs-in-means
mean(treatment$reading) - mean(control$reading) +
  1.96 * sqrt(var(treatment$reading) / n_t
              + var(control$reading) / n_c)
## [1] 11.25347
```

Based on the outputs above, we conclude that the average causal effect of attending a small class on third-grade reading test scores among *all* students in the target population was likely an increase of between 3.17 and 11.25 points or, expressed differently, an increase of 7.21 ± 4.04 points. (The width of the confidence interval here is 11.25−3.17=8.08 points, and so the margin of error is 4.04 points.)

7.2.3 CONFIDENCE INTERVAL FOR PREDICTED OUTCOMES

Finally, we can use a similar procedure to construct confidence intervals for predicted outcomes. Let's return to the analysis of chapter 4, where we fitted a linear model to predict GDP growth using changes in night-time light emissions.

By running the following code, we (i) read and store the dataset in an object named *co*, (ii) show the number of observations and variables in the dataset, (iii) show the first six observations, and (iv) create our two variables of interest. (Remember to first set the working directory.)

RECALL: If the DSS folder is saved directly on your Desktop, to set the working directory, you must run setwd("~/Desktop/DSS") if you have a Mac and setwd("C:/*user*/Desktop/DSS") if you have a Windows computer (where *user* is your own username). If the DSS folder is saved elsewhere, please see subsection 1.7.1 for instructions on how to set the working directory.

```
co <- read.csv("countries.csv") # reads and stores data

dim(co) # provides dimensions of dataframe: rows, columns
## [1] 170   5

head(co) # shows first observations
##     country      gdp  prior_gdp   light  prior_light
## 1       USA   11.107      7.373   4.227        4.482
## 2     Japan  543.017    464.168  11.926       11.808
## 3   Germany    2.152      1.793  10.573        9.699
## 4     China   16.558      4.901   1.451        0.735
## 5        UK    1.098      0.754  11.856       13.392
## 6    France    1.582      1.208   8.513        6.909

## create GDP percentage change variable
co$gdp_change <-
  ((co$gdp - co$prior_gdp) / co$prior_gdp) * 100

## create light percentage change variable
co$light_change <-
  ((co$light - co$prior_light) / co$prior_light) * 100
```

As you may recall, to predict GDP growth using the percentage change in night-time light emissions, we employed the following linear model:

$$\widehat{gdp\_change}_i = \widehat{\alpha} + \widehat{\beta}\ light\_change_i \quad (i\text{=countries})$$

where:

- $\widehat{gdp\_change}_i$ is the average predicted percentage change in GDP from 1992–1993 to 2005–2006 among countries in which the value of *light_change* equals $light\_change_i$
- $light\_change_i$ is the percentage change in night-time light emissions experienced by country *i* from 1992–1993 to 2005–2006.

To fit the linear model and store it as an object, we run:

```
fit <- lm(gdp_change ~ light_change,
          data=co) # fits and stores linear model

fit # shows contents of object
##
## Call:
## lm(formula = gdp_change ~ light_change, data = co)
##
## Coefficients:
## (Intercept)  light_change
##     49.8202        0.2546
```

The fitted model is then:

$$\widehat{gdp\_change} = 49.82 + 0.25\ light\_change$$

Now, we can use this model to make predictions. For example, in chapter 4, we found that a country in which night-time light emissions increased by 20% during a 13-year period is predicted to have experienced GDP growth of about 55% in the same time period, on average (49.82+0.25×20=54.82).

Because of potential noise in the data, there is some uncertainty around this prediction. As we did in the last two subsections, we can construct a 95% confidence interval to measure this uncertainty. In this case, the math is much more complicated, so we ask R to compute it for us.

predict() makes predictions based on a fitted linear model. The only required argument is the name of the object that contains the output of the lm() function. By default, this function produces a prediction for every observation in the dataset used to fit the linear model. To produce only one prediction based on a particular value of the predictor(s), we set the optional argument newdata to equal data.frame(), where inside the parentheses we specify the value of the predictor(s). To also produce the 95% confidence interval of that one prediction, we set the optional argument interval to equal "confidence". To change the level of confidence of the interval, we would specify the optional argument level. Example: fit <- lm(*y_var* ~ *x_var*, data=*data*) then predict(fit, newdata=data.frame(*x_var*=5), interval="confidence", level=0.99).

To calculate the 95% confidence interval for a predicted outcome, we can use the function predict(), which makes predictions based on a fitted linear model. This function requires as its main argument the name of the object that contains the output of the lm() function. To specify the value of the predictor we want to use for the prediction, we use the optional argument newdata. This argument needs a dataframe, which we can create using the function data.frame(). Inside these parentheses, we specify the value of the predictor: light_change=20 in this case. Finally, if in addition to the prediction, we want R to provide the 95% confidence interval, we set the optional argument interval to equal "confidence". By default, this argument provides the interval using a level of confidence of 95%. (If we wanted a different level of confidence, we would specify the optional argument level.)

```
## compute 95% confidence interval for prediction
predict( fit , # object with lm() output
      newdata=data.frame(light_change=20), # set value of X
      interval ="confidence") # provide 95% confidence interval
##        fit       lwr       upr
## 1 54.91233 48.77123 61.05343
```


The first number R provides is the predicted outcome based on the specified (i) fitted linear model and (ii) value of the predictor. The next two numbers are the lower and upper limits of the 95% confidence interval. Based on the output above, then, we can state that the 95% confidence interval of our predicted outcome is [48.77, 61.05].

We can, therefore, conclude that a country in which night-time light emissions increased by 20% during a 13-year period would have likely experienced in the same time period an average GDP growth of between 48.77% and 61.05%, or 54.91% ± 6.14 p.p. (The width of the interval here is 61.05%−48.77%=12.28 p.p., and so the margin of error is 6.14 p.p.)

7.3 HYPOTHESIS TESTING

Hypothesis testing is a methodology that we use to determine whether a parameter is likely to equal a particular value. (There are other uses of hypothesis testing, but we focus on this specific application.) For example, we can use hypothesis testing to determine whether or not an average treatment effect is different from zero in the target population.

Hypothesis testing is a methodology we use to determine whether a parameter is likely to equal a particular value. The **null hypothesis**, H_0, is the hypothesis we would like to eventually refute; in this book, H_0: true value=0. The **alternative hypothesis**, H_1, is the hypothesis we test the null hypothesis against; in this book, H_1: true value≠0.

Hypothesis testing is based on the idea of proof by contradiction. We start by assuming the contrary of what we would like to prove and show how this assumption leads to a logical contradiction.

Specifically, we begin by defining what is known as the **null hypothesis**, denoted as H_0. This is the hypothesis we would like to eventually refute, that is, find sufficient evidence against. For example, if we are interested in whether a treatment affects an outcome, on average, at the population level, we would set the null hypothesis to state that the true value of the parameter—the average treatment effect at the population level, in this case—equals zero. This would mean that the outcome neither increases nor decreases, on average, as a result of the treatment.

In general, the null hypothesis can state that the true value equals any particular value, which we denote by θ (the Greek letter theta). In this book, however, we always set the null hypothesis to state that the true value of the parameter equals zero. In mathematical notation, the null hypothesis is:

H_0: true value $= \theta$ (in general)

H_0: true value $= 0$ (in this book)

Next, we set the **alternative hypothesis**, denoted as H_1. This is the hypothesis we test the null hypothesis against. In this book, we employ what is known as a two-sided alternative hypothesis, which states that the true value of the parameter is not θ, without restricting the parameter to being above or below θ. In particular, since we set θ to equal zero in our null hypothesis, our alternative hypothesis states that the true value of the parameter is not zero, without restricting the sign of the parameter to being positive or negative. In mathematical notation, the alternative hypothesis is:

$$H_1: \text{true value} \neq \theta \qquad \text{(in general)}$$
$$H_1: \text{true value} \neq 0 \qquad \text{(in this book)}$$

Now, let's return to the distribution of our standardized estimator over multiple hypothetical samples (formula 7.1):

$$\frac{\textit{estimator} - \text{true value}}{\text{standard error}} \overset{\text{approx.}}{\sim} N(0, 1)$$

If the null hypothesis is correct and the true value of the parameter equals θ, then we end up with:

$$\frac{\textit{estimator} - \theta}{\text{standard error}} \overset{\text{approx.}}{\sim} N(0, 1) \quad \text{(if true value}=\theta\text{)}$$

A **test statistic** is a function of observed data that can be used to test the null hypothesis. Here we use a test statistic called the **z-statistic**, whose distribution under the null hypothesis is the standard normal distribution.

This random variable is known as the **z-statistic**. The z-statistic is an example of a **test statistic**, which is a function of observed data that can be used to test the null hypothesis.

In the case of our null hypothesis, in which the true value of the parameter is set to equal zero, the test statistic and its distribution across multiple hypothetical samples are:

FORMULA 7.3. Formula of the test statistic and its distribution under the null, when the null hypothesis states that the true value of the parameter equals zero.

TEST STATISTIC

$$\text{z-statistic} = \frac{\textit{estimator}}{\text{standard error}} \overset{\text{approx.}}{\sim} N(0, 1)$$

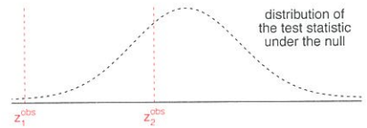
where:

- *estimator* is a random variable across multiple hypothetical samples
- standard error is the estimated standard deviation of the estimator across multiple hypothetical samples.

Suppose we were to draw multiple samples from the same target population and compute the z-statistic for each sample. Then, thanks to the central limit theorem, we know that if the null hypothesis were true, the z-statistics would approximately follow the standard normal distribution. In reality, however, we usually draw only one sample. As a result, we can observe only one realization of the z-statistic. We denote the observed value of the z-statistic as z^{obs}.

TIP: The distribution of the test statistic *under the null hypothesis* is the distribution the test statistic would approximately follow if the null hypothesis were true. Here, the distribution of the test statistic under the null hypothesis is the standard normal distribution, $N(0, 1)$.

Now we can gauge the degree of consistency between what we observe and the null hypothesis. Here is the general idea: If the observed value of the test statistic is extreme relative to the distribution of the test statistic under the null hypothesis (as is, for example, the value of z_1^{obs} in the figure in the margin), then what we observe would be highly unlikely if the null hypothesis were true. We would, thus, conclude that the null hypothesis is likely to be false. In statistical terms, we would *reject the null hypothesis*. Alternatively, if the observed value of the test statistic is typical under the null hypothesis (as is the value of z_2^{obs}), then what we observe would be likely if the null hypothesis were true. We would, in this case, not have enough evidence to claim that the null hypothesis is likely to be false. In statistical terms, we would *fail to reject the null hypothesis*. Let's add more details.



Because we know the distribution of the test statistic under the null hypothesis, we can compute the probability that we observe a value at least as extreme as the one we observed if indeed the null hypothesis is true. This probability is called the **p-value**. Here, because our alternative hypothesis is two-sided, we calculate what is known as the two-sided p-value.

The **p-value** is the probability that we observe a value of the test statistic at least as extreme as the one we actually observed if the null hypothesis is true.

The two-sided p-value computes the probability that we observe a test statistic as extreme as the one we observed in either direction of the real line. Here, it is equivalent to (i) the area under the curve of the standard normal distribution between negative infinity and $-|z^{obs}|$, plus (ii) the area under the curve of the standard normal distribution between $|z^{obs}|$ and infinity (where, again, z^{obs} is the observed value of the z-statistic). (See the shaded areas in figure 7.2.)

RECALL: The probability that Z takes a value between z_1 and z_2 is equivalent to the area under the curve of the standard normal distribution between z_1 and z_2.

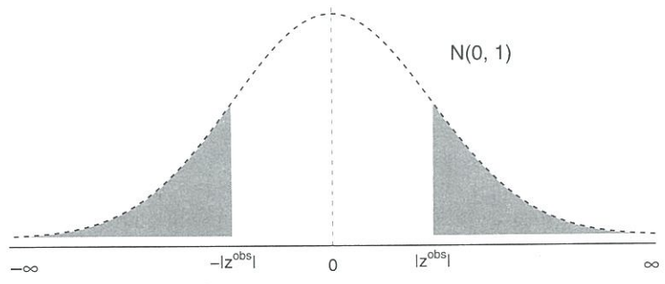


FIGURE 7.2. The two-sided p-value is the probability of observing a test statistic below $-|z^{obs}|$ plus the probability of observing a test statistic above $|z^{obs}|$ in the standard normal distribution, which is the distribution of the test statistic if the null hypothesis is true.

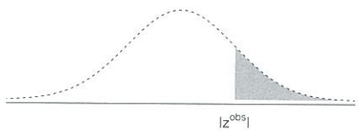
In mathematical notation, the two-sided p-value is defined as:

$$\text{two-sided p-value} = P(Z \leq -|z^{obs}|) + P(Z \geq |z^{obs}|)$$

Since the standard normal distribution is symmetric and centered at zero, the probability of a value below $-|z^{obs}|$ is the same as the probability of a value above $|z^{obs}|$. This property enables us to simplify the formula of the two-sided p-value:

$$\text{two-sided p-value} = 2 \times P(Z \leq -|z^{obs}|)$$

When calculating the two-sided p-value, we add the two above-mentioned probabilities because we consider extreme values in either direction of the real line. We use this type of p-value whenever the alternative hypothesis is two-sided, that is, when it does not constrain the sign of the parameter. Had we stated as our alternative hypothesis that the parameter is positive (because we knew for sure that it could not be negative), then we could compute a one-sided p-value, which would equal the probability that we observe a test statistic as extreme as the one we observed only in the positive direction. (See figure in the margin.)



In general, a smaller p-value provides stronger evidence against the null hypothesis. A very small p-value indicates that the observed value of the test statistic would be highly unlikely if the null hypothesis were true. Thus, when the p-value is very small, there are two possible scenarios: either (a) the null hypothesis is true and we observed something highly unlikely, or (b) the null hypothesis is not true. As the p-value decreases into extremely small magnitudes, we become increasingly confident that the null hypothesis is not true, and thus, we reject it.

The **significance level** determines the rejection threshold of the test and characterizes the probability of false rejection of the null hypothesis.

How small does the p-value need to be for us to reject the null hypothesis? We reject the null hypothesis when the p-value is equal to or smaller than what is known as the **significance level** (or just "level") of the test. Social scientists conventionally use one of three significance levels: 10%, 5%, and 1%. In this book, we use 5% as our significance level. Thus, we reject the null hypothesis when the p-value is equal to or smaller than 0.05 (or 5%), and we fail to reject the null hypothesis when the p-value is greater than 0.05 (or 5%).

RECALL: To interpret a proportion or a probability as a percentage, we multiply the decimal value by 100.

Note that through this procedure, we never *accept the null hypothesis*. Failing to reject the null hypothesis is not the same as accepting it. Just because we have not found evidence against the null hypothesis doesn't mean that we have proven it to be true. On the flip side, however, rejecting the null hypothesis *is* the same as accepting the alternative hypothesis, although we typically do not express it that way.

A result is said to be **statistically significant** at the 5% level when we can reject the null hypothesis using the 5% rejection threshold and conclude that the corresponding parameter is distinguishable from zero. Alternatively, a result is said to be **not statistically significant** at the 5% level when we fail to reject the null hypothesis using the 5% rejection threshold and conclude that the corresponding parameter is not distinguishable from zero.

A result is **statistically significant** at the 5% level when the corresponding parameter is distinguishable from zero using 5% as the rejection threshold. Conversely, a result is **not statistically significant** at the 5% level when the corresponding parameter is not distinguishable from zero using 5% as the rejection threshold.

When a result is statistically significant at the 5% level, do we know for sure that the true value of the corresponding parameter is not zero? No, we do not. A p-value of 5% does not rule out the possibility that the parameter is zero. In fact, thanks to the central limit theorem, we know that if the null hypothesis is true, in 5% of the samples drawn from the target population, we will wrongly reject the null when using a significance level of 5%. Indeed, the significance level of a test characterizes the probability of false rejection of the null hypothesis (known as *type I error*). The smaller the level used in the test, the less likely we are to falsely reject the null. The possibility of wrongly rejecting the null illustrates the importance of replicating social scientific studies to confirm their conclusions. While the probability of falsely rejecting the null hypothesis in any one sample is 5%, the probability of falsely rejecting the null twice in a row, when analyzing two independent samples of data drawn from the same target population, is only 0.25%.

The cut-off points of the test statistic used to determine whether to reject the null hypothesis are called **critical values**. If the distribution of the test statistic is well-approximated by the standard normal distribution and our alternative hypothesis is two-sided, the critical value for the 5% significance level is 1.96. This means that when we observe a z-statistic that in absolute value is greater than or equal to 1.96, we will reject the null at the 5% level, and when we observe a z-statistic that in absolute value is less than 1.96, we will fail to reject the null at the 5% level. (For an explanation, see the formula in detail below.)

A **critical value** is the cut-off point of the test statistic used to determine whether to reject the null hypothesis. If the distribution of the test statistic is well-approximated by the standard normal distribution and our alternative hypothesis is two-sided, the critical value for the 5% significance level is 1.96.

FORMULA IN DETAIL

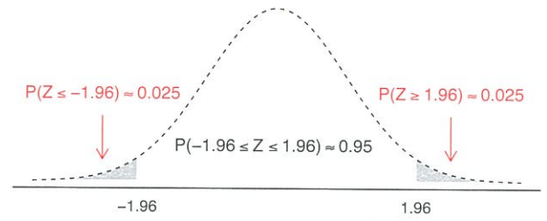
First, recall that the two-sided p-value is:

$$\text{two-sided p-value} = P(Z \leq -|z^{obs}|) + P(Z \geq |z^{obs}|)$$

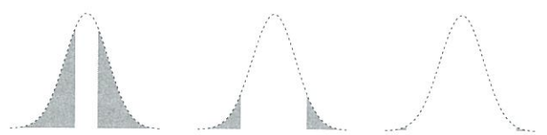
If $|z^{obs}|$ equals 1.96, the two-sided p-value will approximately equal 0.05 (or 5%):

$$\text{two-sided p-value} = P(Z \leq -1.96) + P(Z \geq 1.96) \approx 0.05$$

Here is the reasoning: As we saw in chapter 6, the probability that Z takes a value between -1.96 and 1.96 is approximately 95%. Therefore, the probability that Z takes a value less than or equal to -1.96 plus the probability that Z takes a value greater than or equal to 1.96 is approximately 5% (1−0.95=0.05).



Second, note that as $|z^{obs}|$ increases, that is, moves farther into the tails of the distribution, the associated two-sided p-value decreases because the area under the curve that measures this probability becomes smaller. (As we move from left to right in the figure below, the values of $|z^{obs}|$ increase and the associated two-sided p-values decrease.)



Taken together, if the absolute value of the z-statistic is greater than or equal to 1.96, the two-sided p-value will be less than or equal to 0.05 (or 5%), and so we will reject the null at the 5% significance level. Conversely, if the absolute value of the z-statistic is less than 1.96, the two-sided p-value will be greater than 0.05 (or 5%), and so we will fail to reject the null at the 5% significance level.

To summarize, below is the formal procedure for conducting hypothesis testing to determine whether a parameter is likely different than zero using the 5% significance level. Note, again, that you can compare either the absolute value of the observed z-statistic to 1.96 or the associated two-sided p-value to 0.05. These two procedures are mathematically equivalent and lead to the same conclusion.

HYPOTHESIS TESTING WITH 5% SIGNIFICANCE LEVEL

1. Specify null and alternative hypotheses:

$$H_0: \text{true value} = 0$$
$$H_1: \text{true value} \neq 0$$

2a. Compute observed value of the test statistic:

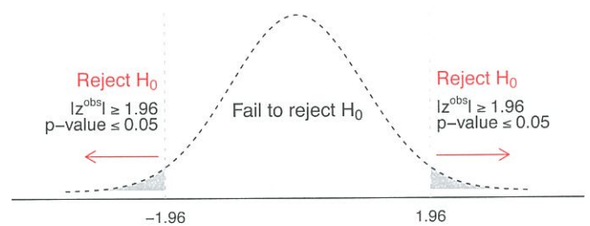
$$z^{obs} = \frac{\text{estimator}}{\text{standard error}}$$

2b. Compute associated two-sided p-value:

$$\text{two-sided p-value} = 2 \times P(Z \leq -|z^{obs}|)$$

3. Conclude:

- If $|z^{obs}| \geq 1.96$ or p-value ≤ 0.05, reject the null hypothesis and conclude that the result is statistically significant at the 5% level.
- If $|z^{obs}| < 1.96$ or p-value > 0.05, fail to reject the null hypothesis and conclude that the result is not statistically significant at the 5% level.



Now that we know the general procedure for conducting hypothesis testing, let's see how we can use it to determine whether a treatment affects an outcome, on average, at the population level. We start by learning how hypothesis testing works with the difference-in-means estimator. Then, we learn how it works with estimated regression coefficients.

7.3.1 HYPOTHESIS TESTING WITH THE DIFFERENCE-IN-MEANS ESTIMATOR

Let's return to our analysis of Project STAR. Since this was a randomized experiment, we can use the difference-in-means estimator to estimate average treatment effects.

To conduct hypothesis testing with the difference-in-means estimator, first we set the null hypothesis to state that the true value of the average treatment effect at the population level equals zero. In mathematical notation:

$$H_0: \mathbb{E}[Y_i(X_i=1) - Y_i(X_i=0)] = 0$$

where:

- $\mathbb{E}[Y_i(X_i=1) - Y_i(X_i=0)]$ is the average treatment effect at the population level, where $\mathbb{E}$ denotes the population mean
- $Y_i(X_i=1)$ and $Y_i(X_i=0)$ are the potential outcomes under the treatment and control conditions, respectively, for individual i.

Next, we set the alternative hypothesis to state that the treatment either increases or decreases the outcome, on average, at the population level. In mathematical notation:

$$H_1: \mathbb{E}[Y_i(X_i=1) - Y_i(X_i=0)] \neq 0$$

Then, using formula 7.3, we construct the following test statistic for the difference-in-means estimator:

TEST STATISTIC FOR THE DIFFERENCE-IN-MEANS ESTIMATOR

$$\text{z-statistic} = \frac{\overline{Y}_{\text{treatment group}} - \overline{Y}_{\text{control group}}}{\sqrt{\frac{var(Y_{\text{treatment}})}{n_{\text{treatment group}}} + \frac{var(Y_{\text{control}})}{n_{\text{control group}}}}}$$

where:

- $\overline{Y}_{\text{treatment group}} - \overline{Y}_{\text{control group}}$ is the difference-in-means estimator
- $\sqrt{\frac{var(Y_{\text{treatment}})}{n_{\text{treatment group}}} + \frac{var(Y_{\text{control}})}{n_{\text{control group}}}}$ is the standard error of the difference-in-means estimator
- $var(Y_{\text{treatment}})$ and $var(Y_{\text{control}})$ are the sample variances of Y under the treatment and control conditions
- $n_{\text{treatment group}}$ and $n_{\text{control group}}$ are the number of observations in the treatment and the control groups in the sample.

Now that we know what we need to compute to test the null hypothesis in this case, let's continue the analysis we started in subsection 7.2.2. To compute (and store) the observed value of the test statistic in the running example, we run:

```
## calculate and store observed value of test statistic
z_obs <- (mean(treatment$reading) -
                    mean(control$reading)) /
                     sqrt(var(treatment$reading) / n_t +
                          var(control$reading) / n_c )

z_obs # shows contents of object
## [1] 3.495654
```

In the sample of data we are analyzing, the value of the test statistic is 3.5. Since its absolute value is greater than 1.96, we can already reject the null hypothesis and conclude that the effect is statistically significant at the 5% level.

Even so, let's continue to compute the associated p-value. Since the test statistic is 3.5, the two-sided p-value is the probability that in the standard normal distribution, we observe a value less than -3.5 or greater than 3.5. This is equivalent to two times the probability that we observe a value below -3.5.

To compute p-values in R, we can use the function pnorm() in conjunction with the function abs(), which stands for "absolute value." For example, to compute the p-value here, we run:

```
## calculate the associated two-sided p-value
2 * pnorm( - abs(z_obs))
## [1] 0.0004729011
```

Based on the output above, if the null hypothesis is true, the probability of observing a test statistic equal to or larger than 3.5 (in absolute value) is 0.05% (0.0005×100=0.05%). This is an extremely small probability.

Since the p-value is smaller than 5%, we reject the null hypothesis and conclude that the effect is statistically significant at the 5% level. In other words, we conclude that attending a small class is likely to have a non-zero average causal effect on reading scores for *all* students in the target population, and not only for those who participated in Project STAR.

Note that we could have arrived at the same conclusion using the 95% confidence interval for the difference-in-means estimator we computed in subsection 7.2.2. If the 95% confidence interval of an estimator does not include zero, we will reject the null hypothesis that the corresponding parameter equals zero at the 5% level. By the same logic, if it does include zero, we will fail to reject the null hypothesis.

TIP: If you are starting a new R session here, you need to re-run the lines of code that we wrote in subsection 7.2.2 that:
- set the working directory
- read and store the dataset
- create the treatment variable
- create two separate dataframes, one for the treatment group and one for the control group
- compute and store the sample sizes of each of the two dataframes.

RECALL: pnorm() calculates the probability that the standard normal random variable, Z, takes a value *less than or equal to* the number specified inside the parentheses. Example: pnorm(0).

abs() calculates the absolute value of the argument specified inside the parentheses. Example: abs(-2).

TIP: Computing confidence intervals and conducting hypothesis testing are equivalent procedures and will lead us to the same conclusions as long as the level of confidence of the interval equals 100 minus the significance level of the test.

RELATIONSHIP BETWEEN CONFIDENCE INTERVALS AND HYPOTHESIS TESTING: If the 95% confidence interval of an estimator does not include zero, we will reject the null hypothesis that the corresponding parameter equals zero at the 5% level. By the same logic, if it does include zero, we will fail to reject the null hypothesis.

In the example at hand, the 95% confidence interval for the difference-in-means estimator was [3.17, 11.25]. Since the interval did not include zero, we could have already concluded that the effect is statistically significant at the 5% level.

7.3.2 HYPOTHESIS TESTING WITH ESTIMATED REGRESSION COEFFICIENTS

We have just learned how to use hypothesis testing to determine whether an average treatment effect is statistically significant based on the difference-in-means estimator. This procedure is useful for analyses of randomized experiments in which we do not have to worry about confounding variables.

As we saw in chapter 5, when analyzing observational data, we do worry about the presence of confounding variables obscuring the causal relationship between the treatment and the outcome. In this case, the difference-in-means estimator no longer provides a valid estimate of the average treatment effect. Instead, we can fit a multiple linear regression model in which X_1 is the treatment variable and all other X variables are the confounding variables. If the model includes all potential confounding variables as control variables, $\widehat{\beta}_1$ (the estimated coefficient affecting the treatment variable X_1) can be interpreted as a valid estimate of the average treatment effect. We can then use hypothesis testing to determine whether the effect, represented by $\widehat{\beta}_1$, is likely to be zero.

Let's return to the analysis in chapter 5 of the survey conducted after the 2014 election on a random sample of Ukrainians living in precincts within 50 kilometers of the Ukraine-Russia border.

RECALL: If the DSS folder is saved directly on your Desktop, to set the working directory, you must run `setwd("~/Desktop/DSS")` if you have a Mac and `setwd("C:/user/Desktop/DSS")` if you have a Windows computer (where *user* is your own username). If the DSS folder is saved elsewhere, please see subsection 1.7.1 for instructions on how to set the working directory.

By running the following code, we (i) read and store the dataset in an object named *uas*, (ii) show the number of observations and variables in the dataset, and (iii) show the first six observations. (Remember to first set the working directory.)

```
uas <- read.csv("UA_survey.csv") # reads and stores data

dim(uas) # provides dimensions of dataframe: rows, columns
## [1] 358   3
```



```
head(uas) # shows first observations
##   russian_tv pro_russian_vote within_25km
## 1          1                0          1
## 2          1                1          1
## 3          0                0          0
## 4          0                0          1
## 5          0                0          1
## 6          1                0          0
```

As you may recall, we were interested in estimating the effect that receiving Russian TV had on a respondent's probability of voting for a pro-Russian party in the 2014 parliamentary election. We were concerned that living in close proximity to the border was a confounding variable, given the existence of military fortifications along the border at the time.

The treatment variable was *russian_tv*, the outcome variable was *pro_russian_vote*, and the confounding variable was *within_25km*. To estimate the average treatment effect, we employed the following multiple linear regression model:

$$pro\_russian\_vote_i = \alpha + \beta_1\ russian\_tv_i + \beta_2\ within\_25km_i + \epsilon_i \quad (i\text{=respondents})$$

where:

- $pro\_russian\_vote_i$ is the binary variable that identifies whether respondent i voted for a pro-Russian party in the 2014 Ukrainian parliamentary election
- $russian\_tv_i$ is the treatment variable, which indicates whether the precinct where respondent i lives received Russian TV
- $within\_25km_i$ is the confounding variable, which indicates whether the precinct where respondent i lives is within 25 kilometers of the border
- ϵ_i is the error term for respondent i.

To fit the linear model and store it as an object, we run:

```
fit <- lm(pro_russian_vote ~ russian_tv + within_25km,
          data=uas) # fits and stores linear model

fit # shows contents of object
##
## Call:
## lm(formula = pro_russian_vote ~ russian_tv
##         + within_25km, data=uas)
##
## Coefficients:
## (Intercept)     russian_tv    within_25km
##      0.1959         0.2876        -0.2081
```

Based on the value of $\widehat{\beta}_1$ above, we estimate that, when we hold living very close to the border constant, receiving Russian TV (as compared to not receiving it) increased a respondent's probability of voting for a pro-Russian party by 29 percentage points, on average.

Does this mean that the average treatment effect is different from zero at the population level (that is, across *all* Ukrainians who live near the border with Russia)? To decide whether we have statistically significant evidence to conclude one way or the other, we use hypothesis testing.

First, we set the null hypothesis to state that receiving Russian TV had an average causal effect on Ukrainians' voting behavior at the population level of zero. In other words, we set the true value of β_1 to equal zero. (Note that when we speak of the true regression coefficient, we do not use the "hat," because it is not an estimate. We are referring to the value one would obtain if the model were fitted to the population.) In mathematical notation:

$$H_0: \beta_1 = 0$$

Next, we set the alternative hypothesis to state that the true coefficient does not equal 0; that is, receiving Russian TV either increased or decreased Ukrainian's probability of voting for a pro-Russian party, on average, at the population level. In mathematical notation:

$$H_1: \beta_1 \neq 0$$

Then, using formula 7.3, we construct the following test statistic for the estimated regression coefficient $\widehat{\beta}_1$:

TEST STATISTIC FOR $\widehat{\beta}_1$

$$\text{z-statistic} = \frac{\widehat{\beta}_1}{\text{standard error of } \widehat{\beta}_1}$$

where:

- $\widehat{\beta}_1$ is the estimated regression coefficient
- standard error of $\widehat{\beta}_1$ is the estimated standard deviation of the estimated regression coefficients over multiple samples.

In this case, we do not go into the specifics of how to compute the standard error of $\widehat{\beta}_1$ because it is rather complicated. We focus instead on how to ask R to compute it.

For this purpose, we can use the function summary(), which computes several statistics related to a fitted linear model, including the standard errors of the estimated regression coefficients. To focus on the statistics we are interested in, we can ask R to show us only the element named coef of the output from the function summary() by running summary()$coef, where inside the parentheses we specify the name of the object that contains the output of the lm() function. (Recall that we use the $ character to access a variable inside a dataframe; in general, we can use it to access an element within an object.) For example, go ahead and run:

summary()$coef provides a table with the following statistics related to a fitted linear model: estimated regression coefficients, standard errors, test statistics, and two-sided p-values. The one required argument is the output of the lm() function. Example: fit <- lm(*y_var* ~ *x_var*, data=*data*) and then summary(fit)$coef.

```
## show table with statistics related to fitted model
summary(fit)$coef
##                Estimate Std. Error   t value     Pr(>|t|)
##(Intercept)     0.19590  0.0345782  5.665602 3.0321e-08
##russian_tv      0.28759  0.0765243  3.758194 2.0002e-04
##within_25km    -0.20806  0.0768105 -2.708802 7.0798e-03
```

TIP: What does R mean by 2.0002e−04? (See p-value associated with $\widehat{\beta}_1$.) It means 0.00020002, or 2.0002×10^{-4}. When a number is either too large or too small to be displayed compactly, R uses what is known as scientific notation, where *e* stands for "times ten raised to the power of." To get a better sense of how scientific notation works, see the examples below:

$$2e{+}04 = 2 \times 10^4 = 20{,}000$$
$$0e{+}00 = 0 \times 10^0 = 0$$
$$2e{-}04 = 2 \times 10^{-4} = 0.0002$$

As shown above, R provides a table of statistics related to the fitted linear model. The first column shows the estimated coefficients: $\widehat{\alpha}$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$. The second column shows the standard errors of each of the coefficients. The third column shows the values of the test statistics for each of the coefficients. Finally, the fourth column shows the associated two-sided p-values.

Note that, by default, R does not assume that the sample size is large enough to use the central limit theorem. As a result, the distribution of the test statistic under the null hypothesis is no longer the standard normal distribution but rather a distribution called the t-distribution. Consequently, the name of the test statistic changes from z-statistic to t-statistic, although the formula remains the same. (Note that R refers to the observed value of the t-statistic as t-value.) Compared to the standard normal distribution, the t-distribution is also symmetric and bell-shaped but has fatter tails. The p-values computed by R here are slightly larger and, as a result, lead to somewhat more conservative inferences. As long as the sample is not very small, however, the difference is typically negligible. (In fact, as the sample size increases, the t-distribution converges to the standard normal distribution.) When drawing conclusions, then, we can ignore the differences and rely on the p-values provided by R in the table above.

The statistics we care about are those related to $\widehat{\beta}_1$, the estimated coefficient that affects *russian_tv*, since that is the coefficient that can be interpreted as the average treatment effect in this case.

Based on the table of results above, the value of the test statistic associated with $\widehat{\beta}_1$ is 3.76. This is indeed the result we arrive at if we divide $\widehat{\beta}_1$ by its standard error (0.2876/0.0765=3.76).

TIP: Most studies that fit linear regression models to analyze data report the estimated coefficients and their standard errors in a table similar to one of the two below:

	Estimated Coefficients	Standard Errors
Russian TV	0.2876	(0.0765)
Within 25 km	-0.2080	(0.0768)
Intercept	0.1959	(0.0346)

Or, if multiple models are fitted:

	Model 1	Model 2
Russian TV	0.1191	0.2876
	(0.045)	(0.0765)
Within 25 km		-0.2080
		(0.0768)
Intercept	0.1709	0.1959
	(0.0336)	(0.0346)

Although the values of the test statistics are not provided, they can be easily computed by dividing the estimated coefficients by their standard errors, which are usually displayed in parentheses.

Because the absolute value of the test statistic is greater than 1.96 (the critical value for the 5% level), we already have enough evidence to reject the null hypothesis at the 5% significance level and determine that the effect is statistically significant.

Even so, let's take a look at the associated p-value. Based on the table above, the two-sided p-value associated with $\widehat{\beta}_1$ is 0.0002. Thus, if the null hypothesis is true, the probability of observing a test statistic equal to or larger than 3.76 (in absolute value) is 0.02% ($0.0002 \times 100 = 0.02\%$). Since the p-value is smaller than 5%, here too we reject the null hypothesis and determine that the effect is statistically significant at the 5% level.

We conclude, then, that receiving Russian TV likely had a non-zero average causal effect on the probability of voting for a pro-Russian party in the 2014 parliamentary election for *all* Ukrainians living close to the border with Russia, not just for those who participated in the survey.

7.4 STATISTICAL VS. SCIENTIFIC SIGNIFICANCE

A result is **scientifically significant** when it is large enough to be consequential.

A common misconception is that statistical significance is equivalent to scientific significance. As we have just seen, an effect is statistically significant when it is not likely to be zero. In contrast, an effect is **scientifically significant** when its size is large enough to be consequential. Therefore, results that are statistically significant are not necessarily scientifically significant, and vice versa.

Suppose that we found that reducing class sizes had a tiny, albeit statistically distinguishable from zero, effect on test performance. This effect would be statistically significant but not scientifically significant. Based on this study, we would not recommend redirecting educational resources toward extra teachers and classroom space to implement a policy of class-size reduction.

By comparison, imagine that we found that attending a remediation program doubled the probability of graduating from high school, although the effect was found to be not distinguishable from zero due to the small size of the program. This effect would be scientifically significant but not statistically significant. Based on this study, we would at the very least recommend expanding the study by involving a larger number of students.

Typically, we aim to find results that are both statistically and scientifically significant.

7.5 SUMMARY

In this chapter, we learned to make inferences about unknown population-level quantities of interest using sample data. First, we learned to compute confidence intervals, which identify the range of values that is likely to include the true value of our quantity of interest. Then, we learned to conduct hypothesis testing to figure out whether an average causal effect is likely to be different than zero at the population level. Finally, we discussed the difference between statistical and scientific significance. Along the way, we completed some of the analyses from chapters 2 through 5. In particular, we quantified the degree of uncertainty in our estimates so that we could draw conclusions regarding all the observations in the target population and not just those in the sample of data analyzed.

With this chapter, we complete our friendly introduction to data analysis for the social sciences. We hope we have piqued your interest in data science and how it can be used to answer important questions about the real world.

7.6 CHEATSHEETS

7.6.1 CONCEPTS AND NOTATION

concept/notation	description	example(s)
parameter	unknown quantity of interest; it can be a sample-level quantity, but we focus on population-level parameters	the level of support for Brexit among all UK voters is a population-level parameter
estimate	sample-level statistic that estimates a parameter	the proportion of supporters among BES survey respondents is an estimate of the level of support for Brexit among all UK voters
estimator	function of observed data that is used to produce an estimate of a parameter	the sample mean is the estimator used to produce the estimate above
sampling distribution of an estimator	characterizes the degree to which the estimator varies from one sample to another due to sampling variability; it enables us to quantify the amount of uncertainty in our estimates; for all the estimators we use in this book: $\textit{estimator} \overset{\text{approx.}}{\sim} N\left(\text{true value}, (\text{s.e.})^2\right)$ where true value is the true value of the population-level parameter and s.e. is the standard error of the estimator	thanks to the central limit theorem, we know that if we drew multiple large samples of a random variable X, with mean $\mathbb{E}(X)$ and variance $\mathbb{V}(X)$, the sample means would approximately follow a normal distribution with mean $\mathbb{E}(X)$ and variance $\mathbb{V}(X)/n$; the sample distribution of the sample mean is thus: $\overline{X} \overset{\text{approx.}}{\sim} N\left(\mathbb{E}(X), \frac{\mathbb{V}(X)}{n}\right)$
standard error of an estimator	estimated standard deviation of the sampling distribution of the estimator; estimate of the average size of the estimation error over multiple hypothetical samples	the formula for the standard error is different for each estimator; given the sampling distribution of the sample mean above, the standard error of the sample mean is: $\sqrt{\frac{\mathbb{V}(X)}{n}}$ (recall that the standard deviation equals the square root of the variance)
estimation error	difference between the estimate and the true value of the parameter	since the true value of the parameter is unknown, we can never compute the estimation error for any one sample, but thanks to the central limit theorem, we can derive (i) the average size of the estimation error over multiple hypothetical samples (see standard error above), and (ii) the average estimation error over multiple hypothetical samples
average estimation error	also known as bias; average difference between the estimate and the true value of a parameter over multiple hypothetical samples	all the estimators covered in this book have an average estimation error equal to zero
unbiased estimator	estimator for which the average estimation error over multiple samples is zero; estimator that provides, on average, accurate results	all the estimators covered in this book are unbiased estimators of their corresponding parameters; their sampling distributions are centered at the true value of the parameter

continues on next page...

7.6.1 CONCEPTS AND NOTATION (CONTINUED)

concept/notation	description	example(s)
confidence interval	provides the range of values that is likely to include the true value of the parameter three levels of confidence are conventionally used in the social sciences to construct confidence intervals: 90%, 95%, and 99%; the level of confidence refers to the probability that the interval contains the true value of the parameter over multiple samples; the formula to construct the 95% confidence interval is: 95% CI = [*estimator*−1.96×standard error, *estimator*+1.96×standard error]	the 95% confidence interval for the sample mean of *leave* in the BES survey is: [0.4657, 0.4781] we conclude, then, that the true proportion of support for Brexit among *all* UK voters was likely to be between 46.57% and 47.81%
margin of error	defined as half the width of the estimator's confidence interval; as a result, we can express the confidence interval as: *estimator* ± margin of error	in the example above, the margin of error equals 0.62 percentage points, which is half the width of the confidence interval; we can state that the likely proportion of support for Brexit among *all* UK voters was 47.19% with a margin of error of 0.62 percentage points
hypothesis testing	methodology we use to determine whether a parameter is likely to equal a particular value	we can use hypothesis testing to determine whether or not an average treatment effect is different from zero in the target population
null hypothesis or H_0	hypothesis we would like to eventually refute; it states that the true value of the parameter equals a particular value, θ (the Greek letter theta) H_0: true value $= \theta$ in this book, our null hypothesis states that the true value of the parameter is zero H_0: true value $= 0$	in our analysis of the survey of Ukrainians, $\widehat{\beta}_1$ is the estimator we use to estimate the average causal effect of receiving Russian TV on Ukrainians' probability of voting for a pro-Russian party in the 2014 parliamentary election; our null hypothesis states that Russian TV reception had zero average causal effect on Ukrainians voting behavior; in other words, it states that β_1 (the true value of the coefficient affecting the treatment variable) equals zero $H_0: \beta_1 = 0$
alternative hypothesis or H_1	hypothesis we test the null hypothesis against; the two-sided alternative hypothesis states that the true value of the parameter is not θ, without restricting the parameter to being above or below θ H_1: true value $\neq \theta$ in this book, our alternative hypothesis states that the true value of the parameter is not zero H_1: true value $\neq 0$	in our analysis of the survey of Ukrainians, our alternative hypothesis is that receiving Russian TV had either a positive or a negative average causal effect on Ukrainians' voting behavior; in other words, it states that β_1 does not equal zero $H_1: \beta_1 \neq 0$
test statistic	function of observed data that can be used to test the null hypothesis	the z-statistic is an example of a test statistic

continues on next page...

7.6.1 CONCEPTS AND NOTATION (CONTINUED)

concept/notation	description	example(s)
z-statistic	test statistic whose distribution under the null hypothesis is the standard normal distribution; in general: $z\text{-}statistic = \frac{estimator - \theta}{\text{standard error}} \sim N(0, 1)$ in this book, since H_0: true value = 0: $z\text{-}statistic = \frac{estimator}{\text{standard error}}$ we denote the observed value of the z-statistic as z^{obs}	when the estimator is $\widehat{\beta}_1$, the formula of the test statistic is: $\text{z-statistic} = \frac{\widehat{\beta}_1}{\text{standard error of } \widehat{\beta}_1}$ in our analysis of the survey of Ukrainians, the observed value of the test statistic is: $z^{obs} = \frac{0.2876}{0.0765} = 3.76$
p-value	probability that we observe a value of the test statistic at least as extreme as the one we actually observed if the null hypothesis is true; when the null hypothesis is two-sided, we compute the two-sided p-value, which conveys the probability that we observe a test statistic as extreme as the one we observed in either direction of the real line: two-sided p-value $= 2 \times P(Z \leq -\lvert z^{obs} \rvert)$	in our analysis of the survey of Ukrainians, the two-sided p-value is: p-value $= 2 \times P(Z \leq -\lvert 3.76 \rvert) \approx 0.0002$ if the null hypothesis is true, the probability of observing a test statistic equal to or larger than 3.76 (in absolute value) is 0.02% (0.0002×100=0.02%)
significance level	determines the rejection threshold of the test and characterizes the probability of false rejection of the null hypothesis; social scientists conventionally use one of three significance levels: 10%, 5%, and 1%	if we use 5% as our significance level: we will reject the null hypothesis when the p-value is equal to or smaller than 0.05 (or 5%), and we will fail to reject the null hypothesis when the p-value is greater than 0.05 (or 5%) we will wrongly reject the null hypothesis in 5% of the samples drawn from the target population
statistical significance	a result is statistically significant at the 5% level when it is distinguishable from zero using 5% as the rejection threshold specifically, if $\lvert z^{obs} \rvert \geq 1.96$ or the two-sided p-value ≤ 0.05, we will reject the null hypothesis and conclude that the result is statistically significant at the 5% level conversely, a result is not statistically significant at the 5% level when it is not distinguishable from zero using 5% as the rejection threshold	in our analysis of the survey of Ukrainians: $\lvert z^{obs} \rvert = 3.76$ and p-value ≈ 0.0002 thus, we reject the null hypothesis and conclude that receiving Russian TV was likely to have a non-zero average causal effect on the probability of voting for a pro-Russian party in the 2014 parliamentary election for all Ukrainians, not just for those in the sample we observed
critical value	cut-off point of the test statistic used to determine whether to reject the null hypothesis	if the distribution of the test statistic is well-approximated by the standard normal distribution and our alternative hypothesis is two-sided, the critical value for the 5% significance level is 1.96; if $\lvert z^{obs} \rvert \geq 1.96$, we will reject the null hypothesis; conversely, if $\lvert z^{obs} \rvert < 1.96$, we will fail to reject the null hypothesis
scientific significance	a result is scientifically significant when it is large enough to be consequential	a result might be statistically significant but be so small as to not be scientifically significant

7.6.2 R SYMBOLS AND OPERATORS

code	description	example(s)
[]	operator used to extract a selection of observations from a dataframe; to its left, we specify the dataframe we want to subset; inside the square brackets, we specify the criterion of selection; since a dataframe is composed of two dimensions, rows and columns, we can specify a criterion of selection on one or both dimensions; first, we specify the criterion of selection of the rows, and then the criterion of selection of the columns (separated by a comma); if the first criterion is left blank, all rows are extracted, and if the second criterion is left blank, all columns are extracted; (for other uses, see pages 50, 61, and 187)	*data*[*data*$*var1*==1,] # extracts the observations that have a value of 1 in *var1* as well as their corresponding values in all the other variables in the dataframe *data*

7.6.3 R FUNCTIONS

function	description	required argument(s)	example(s)
nrow()	computes the number of rows of a dataframe	name of the object where the dataframe is stored	nrow(*data*)
predict()	makes predictions based on a fitted linear model	the name of the object that contains the output of the lm() function	fit <- lm(*y_var* ~ *x_var*, data=*data*) # stores fitted model into an object named fit
		by default, this function produces a prediction for every observation in the dataset used to fit the linear model; to produce only one prediction based on a particular value of the predictor(s), we set the optional argument newdata to equal data.frame(), where inside the parentheses we specify the value of the predictor(s) to also produce the 95% confidence interval of one prediction, we set the optional argument interval to equal "confidence"; to change the confidence level of the interval to a probability different than 95%, we would specify the optional argument level	predict(fit, newdata=data.frame(*x_var*=5), interval="confidence", level=0.99) # produces the predicted average value of *y_var* when *x_var*=5 as well as the 99% confidence interval for the given predicted outcome
abs()	calculates the absolute value	what we want to compute the absolute value of	abs(-2)
summary()$coef	provides a table with the following statistics related to a fitted linear model: estimated regression coefficients, standard errors, test statistics, and two-sided p-values	the name of the object containing the output of the lm() function	fit <- lm(*y_var* ~ *x_var*, data=*data*) # stores fitted model into an object named fit summary(fit)$coef # provides table with results