Chapter 5

Foundations for inference

Statistical inference is primarily concerned with understanding and quantifying the uncertainty of parameter estimates. While the equations and details change depending on the setting, the foundations for inference are the same throughout all of statistics.

We start with a familiar topic: the idea of using a sample proportion to estimate a population proportion. Next, we create what's called a *confidence interval*, which is a range of plausible values where we may find the true population value. Finally, we introduce the *hypothesis testing framework*, which allows us to formally evaluate claims about the population, such as whether a survey provides strong evidence that a candidate has the support of a majority of the voting population.

For videos, slides, and other resources, please visit
www.openintro.org/os

5.1 Point estimates and sampling variability

Companies such as Pew Research frequently conduct polls as a way to understand the state of public opinion or knowledge on many topics, including politics, scientific understanding, brand recognition, and more. The ultimate goal in taking a poll is generally to use the responses to estimate the opinion or knowledge of the broader population.

5.1.1 Point estimates and error

Suppose a poll suggested the US President's approval rating is 45%. We would consider 45% to be a **point estimate** of the approval rating we might see if we collected responses from the entire population. This entire-population response proportion is generally referred to as the **parameter** of interest. When the parameter is a proportion, it is often denoted by p, and we often refer to the sample proportion as $\hat{p}$ (pronounced *p-hat*[1]). Unless we collect responses from every individual in the population, p remains unknown, and we use $\hat{p}$ as our estimate of p. The difference we observe from the poll versus the parameter is called the **error** in the estimate. Generally, the error consists of two aspects: sampling error and bias.

Sampling error, sometimes called *sampling uncertainty*, describes how much an estimate will tend to vary from one sample to the next. For instance, the estimate from one sample might be 1% too low while in another it may be 3% too high. Much of statistics, including much of this book, is focused on understanding and quantifying sampling error, and we will find it useful to consider a sample's size to help us quantify this error; the **sample size** is often represented by the letter n.

Bias describes a systematic tendency to over- or under-estimate the true population value. For example, if we were taking a student poll asking about support for a new college stadium, we'd probably get a biased estimate of the stadium's level of student support by wording the question as, *Do you support your school by supporting funding for the new stadium?* We try to minimize bias through thoughtful data collection procedures, which were discussed in Chapter 1 and are the topic of many other books.

5.1.2 Understanding the variability of a point estimate

Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest.[2] If we were to take a poll of 1000 American adults on this topic, the estimate would not be perfect, but how close might we expect the sample proportion in the poll would be to 88%? We want to understand, *how does the sample proportion $\hat{p}$ behave when the true population proportion is 0.88.*[3] Let's find out! We can simulate responses we would get from a simple random sample of 1000 American adults, which is only possible because we know the actual support for expanding solar energy is 0.88. Here's how we might go about constructing such a simulation:

1. There were about 250 million American adults in 2018. On 250 million pieces of paper, write "support" on 88% of them and "not" on the other 12%.

2. Mix up the pieces of paper and pull out 1000 pieces to represent our sample of 1000 American adults.

3. Compute the fraction of the sample that say "support".

Any volunteers to conduct this simulation? Probably not. Running this simulation with 250 million pieces of paper would be time-consuming and very costly, but we can simulate it using computer

[1] Not to be confused with *phat*, the slang term used for something cool, like this book.

[2] We haven't actually conducted a census to measure this value perfectly. However, a very large sample has suggested the actual level of support is about 88%.

[3] 88% written as a proportion would be 0.88. It is common to switch between proportion and percent. However, formulas presented in this book always refer to the proportion, not the percent.

code; we've written a short program in Figure 5.1 in case you are curious what the computer code looks like. In this simulation, the sample gave a point estimate of $\hat{p}_1 = 0.894$. We know the population proportion for the simulation was $p = 0.88$, so we know the estimate had an error of $0.894 - 0.88 = +0.014$.

```
# 1. Create a set of 250 million entries, where 88% of them are "support"
#    and 12% are "not".
pop_size <- 250000000
possible_entries <- c(rep("support", 0.88 * pop_size), rep("not", 0.12 * pop_size))

# 2. Sample 1000 entries without replacement.
sampled_entries <- sample(possible_entries, size = 1000)

# 3. Compute p-hat: count the number that are "support", then divide by
#    the sample size.
sum(sampled_entries == "support") / 1000
```

Figure 5.1: For those curious, this is code for a single $\hat{p}$ simulation using the statistical software called **R**. Each line that starts with **#** is a **code comment**, which is used to describe in regular language what the code is doing. We've provided software labs in **R** at openintro.org/stat/labs for anyone interested in learning more.

One simulation isn't enough to get a great sense of the distribution of estimates we might expect in the simulation, so we should run more simulations. In a second simulation, we get $\hat{p}_2 = 0.885$, which has an error of +0.005. In another, $\hat{p}_3 = 0.878$ for an error of -0.002. And in another, an estimate of $\hat{p}_4 = 0.859$ with an error of -0.021. With the help of a computer, we've run the simulation 10,000 times and created a histogram of the results from all 10,000 simulations in Figure 5.2. This distribution of sample proportions is called a **sampling distribution**. We can characterize this sampling distribution as follows:

Center. The center of the distribution is $\bar{x}_{\hat{p}} = 0.880$, which is the same as the parameter. Notice that the simulation mimicked a simple random sample of the population, which is a straightforward sampling strategy that helps avoid sampling bias.

Spread. The standard deviation of the distribution is $s_{\hat{p}} = 0.010$. When we're talking about a sampling distribution or the variability of a point estimate, we typically use the term **standard error** rather than *standard deviation*, and the notation $SE_{\hat{p}}$ is used for the standard error associated with the sample proportion.

Shape. The distribution is symmetric and bell-shaped, and it *resembles a normal distribution.*

These findings are encouraging! When the population proportion is $p = 0.88$ and the sample size is $n = 1000$, the sample proportion $\hat{p}$ tends to give a pretty good estimate of the population proportion. We also have the interesting observation that the histogram resembles a normal distribution.

SAMPLING DISTRIBUTIONS ARE NEVER OBSERVED, BUT WE KEEP THEM IN MIND

In real-world applications, we never actually observe the sampling distribution, yet it is useful to always think of a point estimate as coming from such a hypothetical distribution. Understanding the sampling distribution will help us characterize and make sense of the point estimates that we do observe.

EXAMPLE 5.1

If we used a much smaller sample size of $n = 50$, would you guess that the standard error for $\hat{p}$ would be larger or smaller than when we used $n = 1000$?

E

Intuitively, it seems like more data is better than less data, and generally that is correct! The typical error when $p = 0.88$ and $n = 50$ would be larger than the error we would expect when $n = 1000$.

Example 5.1 highlights an important property we will see again and again: a bigger sample tends to provide a more precise point estimate than a smaller sample.

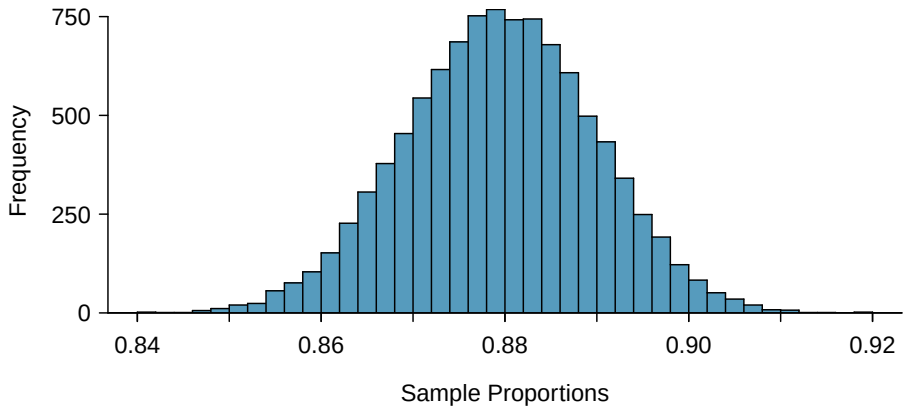


Figure 5.2: A histogram of 10,000 sample proportions, where each sample is taken from a population where the population proportion is 0.88 and the sample size is $n = 1000$.

5.1.3 Central Limit Theorem

The distribution in Figure 5.2 looks an awful lot like a normal distribution. That is no anomaly; it is the result of a general principle called the **Central Limit Theorem**.

CENTRAL LIMIT THEOREM AND THE SUCCESS-FAILURE CONDITION

When observations are independent and the sample size is sufficiently large, the sample proportion $\hat{p}$ will tend to follow a normal distribution with the following mean and standard error:

$$\mu_{\hat{p}} = p \qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

In order for the Central Limit Theorem to hold, the sample size is typically considered sufficiently large when $np \geq 10$ and $n(1-p) \geq 10$, which is called the **success-failure condition**.

The Central Limit Theorem is incredibly important, and it provides a foundation for much of statistics. As we begin applying the Central Limit Theorem, be mindful of the two technical conditions: the observations must be independent, and the sample size must be sufficiently large such that $np \geq 10$ and $n(1-p) \geq 10$.

EXAMPLE 5.2

Earlier we estimated the mean and standard error of $\hat{p}$ using simulated data when $p = 0.88$ and $n = 1000$. Confirm that the Central Limit Theorem applies and the sampling distribution is approximately normal.

(E) **Independence.** There are $n = 1000$ observations for each sample proportion $\hat{p}$, and each of those observations are independent draws. *The most common way for observations to be considered independent is if they are from a simple random sample.*

Success-failure condition. We can confirm the sample size is sufficiently large by checking the success-failure condition and confirming the two calculated values are greater than 10:

$$np = 1000 \times 0.88 = 880 \geq 10 \qquad n(1-p) = 1000 \times (1 - 0.88) = 120 \geq 10$$

The independence and success-failure conditions are both satisfied, so the Central Limit Theorem applies, and it's reasonable to model $\hat{p}$ using a normal distribution.

HOW TO VERIFY SAMPLE OBSERVATIONS ARE INDEPENDENT

Subjects in an experiment are considered independent if they undergo random assignment to the treatment groups.

If the observations are from a simple random sample, then they are independent.

If a sample is from a seemingly random process, e.g. an occasional error on an assembly line, checking independence is more difficult. In this case, use your best judgement.

An additional condition that is sometimes added for samples from a population is that they are no larger than 10% of the population. When the sample exceeds 10% of the population size, the methods we discuss tend to overestimate the sampling error slightly versus what we would get using more advanced methods.[4] This is very rarely an issue, and when it is an issue, our methods tend to be conservative, so we consider this additional check as optional.

EXAMPLE 5.3

Compute the theoretical mean and standard error of $\hat{p}$ when $p = 0.88$ and $n = 1000$, according to the Central Limit Theorem.

E

The mean of the $\hat{p}$'s is simply the population proportion: $\mu_{\hat{p}} = 0.88$.

The calculation of the standard error of $\hat{p}$ uses the following formula:

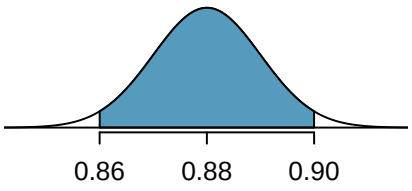
$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.88(1-0.88)}{1000}} = 0.010$$

EXAMPLE 5.4

Estimate how frequently the sample proportion $\hat{p}$ should be within 0.02 (2%) of the population value, $p = 0.88$. Based on Examples 5.2 and 5.3, we know that the distribution is approximately $N(\mu_{\hat{p}} = 0.88, SE_{\hat{p}} = 0.010)$.

E

After so much practice in Section 4.1, this normal distribution example will hopefully feel familiar! We would like to understand the fraction of $\hat{p}$'s between 0.86 and 0.90:



With $\mu_{\hat{p}} = 0.88$ and $SE_{\hat{p}} = 0.010$, we can compute the Z-score for both the left and right cutoffs:

$$Z_{0.86} = \frac{0.86 - 0.88}{0.010} = -2 \qquad Z_{0.90} = \frac{0.90 - 0.88}{0.010} = 2$$

We can use either statistical software, a graphing calculator, or a table to find the areas to the tails, and in any case we will find that they are each 0.0228. The total tail areas are $2 \times 0.0228 = 0.0456$, which leaves the shaded area of 0.9544. That is, about 95.44% of the sampling distribution in Figure 5.2 is within ± 0.02 of the population proportion, $p = 0.88$.

[4] For example, we could use what's called the **finite population correction factor**: if the sample is of size n and the population size is N, then we can multiply the typical standard error formula by $\sqrt{\frac{N-n}{N-1}}$ to obtain a smaller, more precise estimate of the actual standard error. When $n < 0.1 \times N$, this correction factor is relatively small.

GUIDED PRACTICE 5.5

In Example 5.1 we discussed how a smaller sample would tend to produce a less reliable estimate. Explain how this intuition is reflected in the formula for $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.[5]

5.1.4 Applying the Central Limit Theorem to a real-world setting

We do not actually know the population proportion unless we conduct an expensive poll of all individuals in the population. Our earlier value of $p = 0.88$ was based on a Pew Research conducted a poll of 1000 American adults that found $\hat{p} = 0.887$ of them favored expanding solar energy. The researchers might have wondered: does the sample proportion from the poll approximately follow a normal distribution? We can check the conditions from the Central Limit Theorem:

Independence. The poll is a simple random sample of American adults, which means that the observations are independent.

Success-failure condition. To check this condition, we need the population proportion, p, to check if both np and $n(1-p)$ are greater than 10. However, we do not actually know p, which is exactly why the pollsters would take a sample! In cases like these, we often use $\hat{p}$ as our next best way to check the success-failure condition:

$$n\hat{p} = 1000 \times 0.887 = 887 \qquad n(1-\hat{p}) = 1000 \times (1 - 0.887) = 113$$

The sample proportion $\hat{p}$ acts as a reasonable substitute for p during this check, and each value in this case is well above the minimum of 10.

This **substitution approximation** of using $\hat{p}$ in place of p is also useful when computing the standard error of the sample proportion:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.887(1-0.887)}{1000}} = 0.010$$

This substitution technique is sometimes referred to as the "plug-in principle". In this case, $SE_{\hat{p}}$ didn't change enough to be detected using only 3 decimal places versus when we completed the calculation with 0.88 earlier. The computed standard error tends to be reasonably stable even when observing slightly different proportions in one sample or another.

[5] Since the sample size n is in the denominator (on the bottom) of the fraction, a bigger sample size means the entire expression when calculated will tend to be smaller. That is, a larger sample size would correspond to a smaller standard error.

5.1.5 More details regarding the Central Limit Theorem

We've applied the Central Limit Theorem in numerous examples so far this chapter:

> *When observations are independent and the sample size is sufficiently large, the distribution of $\hat{p}$ resembles a normal distribution with*
>
> $$\mu_{\hat{p}} = p \qquad\qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$
>
> *The sample size is considered sufficiently large when $np \geq 10$ and $n(1-p) \geq 10$.*

In this section, we'll explore the success-failure condition and seek to better understand the Central Limit Theorem.

An interesting question to answer is, *what happens when $np < 10$ or $n(1-p) < 10$?* As we did in Section 5.1.2, we can simulate drawing samples of different sizes where, say, the true proportion is $p = 0.25$. Here's a sample of size 10:

no, no, yes, yes, no, no, no, no, no, no

In this sample, we observe a sample proportion of yeses of $\hat{p} = \frac{2}{10} = 0.2$. We can simulate many such proportions to understand the sampling distribution of $\hat{p}$ when $n = 10$ and $p = 0.25$, which we've plotted in Figure 5.3 alongside a normal distribution with the same mean and variability. These distributions have a number of important differences.

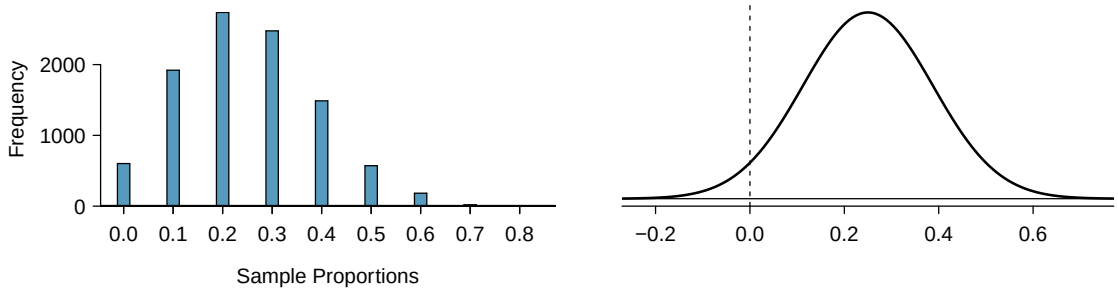


Figure 5.3: Left: simulations of $\hat{p}$ when the sample size is $n = 10$ and the population proportion is $p = 0.25$. Right: a normal distribution with the same mean (0.25) and standard deviation (0.137).

	Unimodal?	Smooth?	Symmetric?
Normal: $N(0.25, 0.14)$	**Yes**	**Yes**	**Yes**
$n = 10$, $p = 0.25$	**Yes**	*No*	*No*

Notice that the success-failure condition was not satisfied when $n = 10$ and $p = 0.25$:

$$np = 10 \times 0.25 = 2.5 \qquad\qquad n(1-p) = 10 \times 0.75 = 7.5$$

This single sampling distribution does not show that the success-failure condition is the perfect guideline, but we have found that the guideline did correctly identify that a normal distribution might not be appropriate.

We can complete several additional simulations, shown in Figures 5.4 and 5.5, and we can see some trends:

1. When either np or $n(1-p)$ is small, the distribution is more **discrete**, i.e. *not continuous*.
2. When np or $n(1-p)$ is smaller than 10, the skew in the distribution is more noteworthy.
3. The larger both np *and* $n(1-p)$, the more normal the distribution. This may be a little harder to see for the larger sample size in these plots as the variability also becomes much smaller.
4. When np and $n(1-p)$ are both very large, the distribution's discreteness is hardly evident, and the distribution looks much more like a normal distribution.

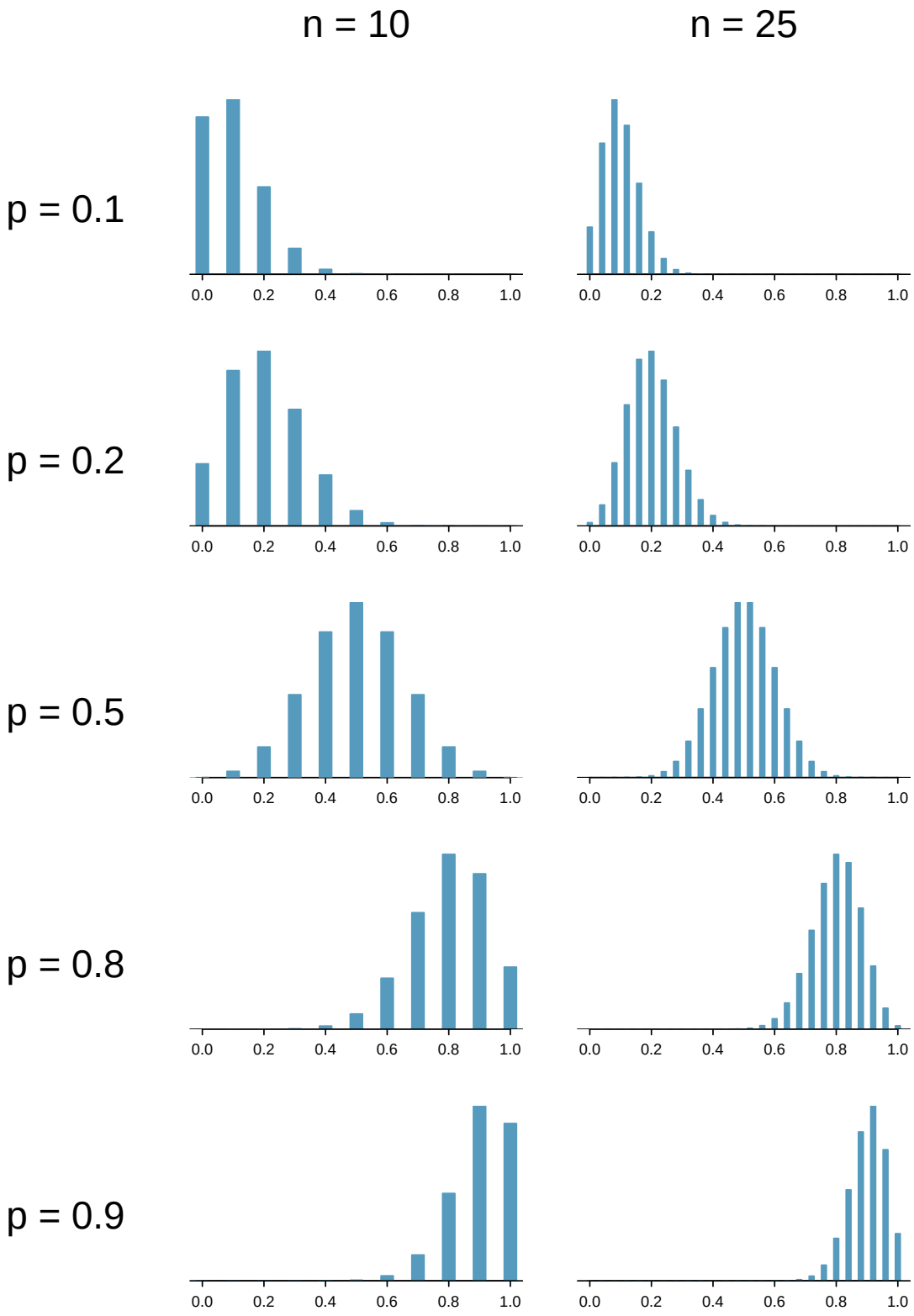


Figure 5.4: Sampling distributions for several scenarios of p and n.
Rows: $p = 0.10$, $p = 0.20$, $p = 0.50$, $p = 0.80$, and $p = 0.90$.
Columns: $n = 10$ and $n = 25$.

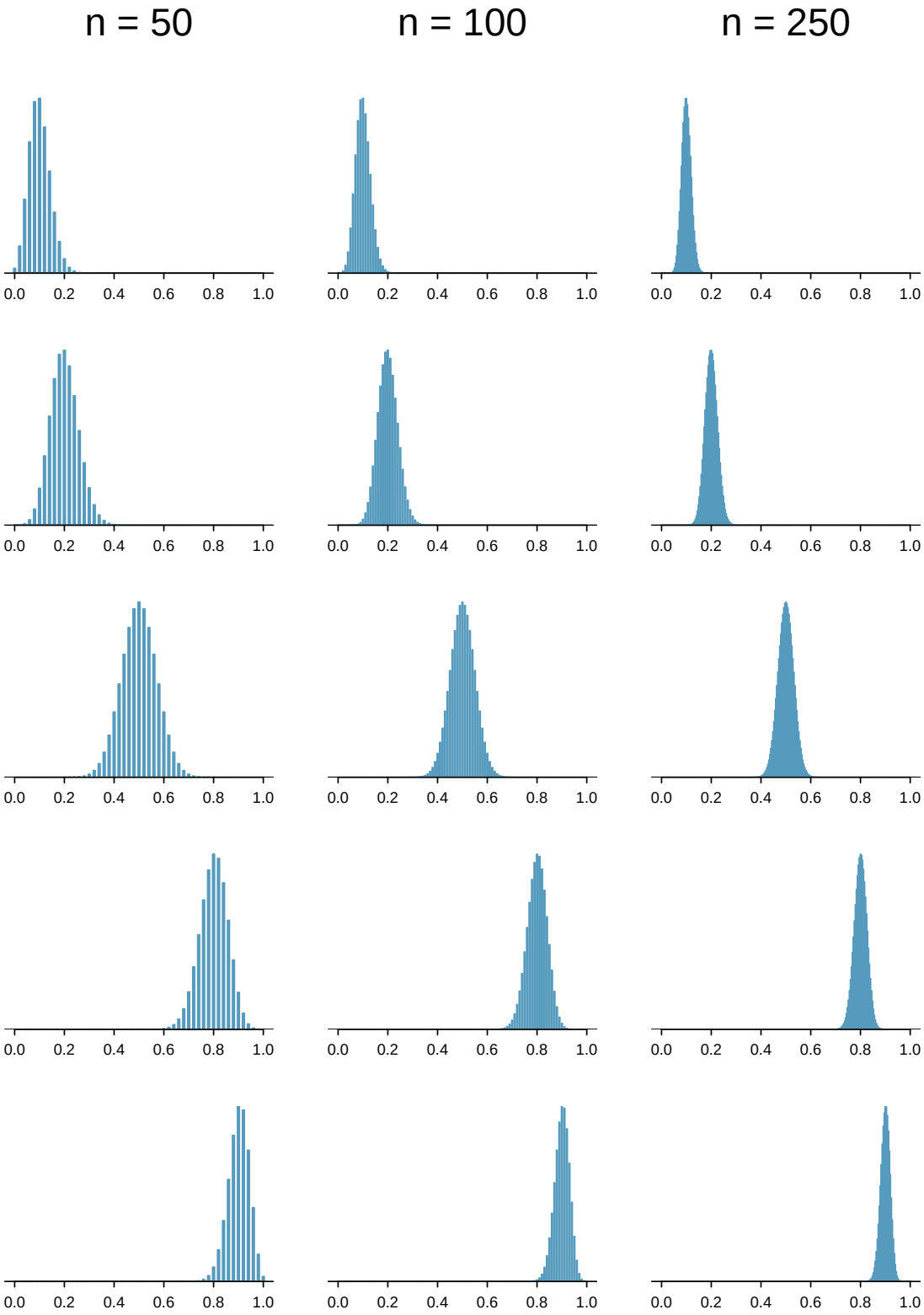


Figure 5.5: Sampling distributions for several scenarios of p and n.
Rows: $p = 0.10$, $p = 0.20$, $p = 0.50$, $p = 0.80$, and $p = 0.90$.
Columns: $n = 50$, $n = 100$, and $n = 250$.

So far we've only focused on the skew and discreteness of the distributions. We haven't considered how the mean and standard error of the distributions change. Take a moment to look back at the graphs, and pay attention to three things:

1. The centers of the distribution are always at the population proportion, p, that was used to generate the simulation. Because the sampling distribution of $\hat{p}$ is always centered at the population parameter p, it means the sample proportion $\hat{p}$ is **unbiased** when the data are independent and drawn from such a population.

2. For a particular population proportion p, the variability in the sampling distribution decreases as the sample size n becomes larger. This will likely align with your intuition: an estimate based on a larger sample size will tend to be more accurate.

3. For a particular sample size, the variability will be largest when $p = 0.5$. The differences may be a little subtle, so take a close look. This reflects the role of the proportion p in the standard error formula: $SE = \sqrt{\frac{p(1-p)}{n}}$. The standard error is largest when $p = 0.5$.

At no point will the distribution of $\hat{p}$ look *perfectly* normal, since $\hat{p}$ will always take discrete values (x/n). It is always a matter of degree, and we will use the standard success-failure condition with minimums of 10 for np and $n(1-p)$ as our guideline within this book.

5.1.6 Extending the framework for other statistics

The strategy of using a sample statistic to estimate a parameter is quite common, and it's a strategy that we can apply to other statistics besides a proportion. For instance, if we want to estimate the average salary for graduates from a particular college, we could survey a random sample of recent graduates; in that example, we'd be using a sample mean $\bar{x}$ to estimate the population mean μ for all graduates. As another example, if we want to estimate the difference in product prices for two websites, we might take a random sample of products available on both sites, check the prices on each, and then compute the average difference; this strategy certainly would give us some idea of the actual difference through a point estimate.

While this chapter emphasizes a single proportion context, we'll encounter many different contexts throughout this book where these methods will be applied. The principles and general ideas are the same, even if the details change a little. We've also sprinkled some other contexts into the exercises to help you start thinking about how the ideas generalize.

Exercises

5.1 Identify the parameter, Part I. For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

(a) In a survey, one hundred college students are asked how many hours per week they spend on the Internet.

(b) In a survey, one hundred college students are asked: "What percentage of the time you spend on the Internet is part of your course work?"

(c) In a survey, one hundred college students are asked whether or not they cited information from Wikipedia in their papers.

(d) In a survey, one hundred college students are asked what percentage of their total weekly spending is on alcoholic beverages.

(e) In a sample of one hundred recent college graduates, it is found that 85 percent expect to get a job within one year of their graduation date.

5.2 Identify the parameter, Part II. For each of the following situations, state whether the parameter of interest is a mean or a proportion.

(a) A poll shows that 64% of Americans personally worry a great deal about federal spending and the budget deficit.

(b) A survey reports that local TV news has shown a 17% increase in revenue within a two year period while newspaper revenues decreased by 6.4% during this time period.

(c) In a survey, high school and college students are asked whether or not they use geolocation services on their smart phones.

(d) In a survey, smart phone users are asked whether or not they use a web-based taxi service.

(e) In a survey, smart phone users are asked how many times they used a web-based taxi service over the last year.

5.3 Quality control. As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(a) What population is under consideration in the data set?

(b) What parameter is being estimated?

(c) What is the point estimate for the parameter?

(d) What is the name of the statistic we use to measure the uncertainty of the point estimate?

(e) Compute the value from part (d) for this context.

(f) The historical rate of defects is 10%. Should the engineer be surprised by the observed rate of defects during the current week?

(g) Suppose the true population value was found to be 10%. If we use this proportion to recompute the value in part (e) using $p = 0.1$ instead of $\hat{p}$, does the resulting value change much?

5.4 Unexpected expense. In a random sample 765 adults in the United States, 322 say they could not cover a $400 unexpected expense without borrowing money or going into debt.

(a) What population is under consideration in the data set?

(b) What parameter is being estimated?

(c) What is the point estimate for the parameter?

(d) What is the name of the statistic we use to measure the uncertainty of the point estimate?

(e) Compute the value from part (d) for this context.

(f) A cable news pundit thinks the value is actually 50%. Should she be surprised by the data?

(g) Suppose the true population value was found to be 40%. If we use this proportion to recompute the value in part (e) using $p = 0.4$ instead of $\hat{p}$, does the resulting value change much?

5.5 Repeated water samples. A nonprofit wants to understand the fraction of households that have elevated levels of lead in their drinking water. They expect at least 5% of homes will have elevated levels of lead, but not more than about 30%. They randomly sample 800 homes and work with the owners to retrieve water samples, and they compute the fraction of these homes with elevated lead levels. They repeat this 1,000 times and build a distribution of sample proportions.

(a) What is this distribution called?

(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

(c) If the proportions are distributed around 8%, what is the variability of the distribution?

(d) What is the formal name of the value you computed in (c)?

(e) Suppose the researchers' budget is reduced, and they are only able to collect 250 observations per sample, but they can still collect 1,000 samples. They build a new distribution of sample proportions. How will the variability of this new distribution compare to the variability of the distribution when each sample contained 800 observations?

5.6 Repeated student samples. Of all freshman at a large college, 16% made the dean's list in the current year. As part of a class project, students randomly sample 40 students and check if those students made the list. They repeat this 1,000 times and build a distribution of sample proportions.

(a) What is this distribution called?

(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

(c) Calculate the variability of this distribution.

(d) What is the formal name of the value you computed in (c)?

(e) Suppose the students decide to sample again, this time collecting 90 students per sample, and they again collect 1,000 samples. They build a new distribution of sample proportions. How will the variability of this new distribution compare to the variability of the distribution when each sample contained 40 observations?

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.

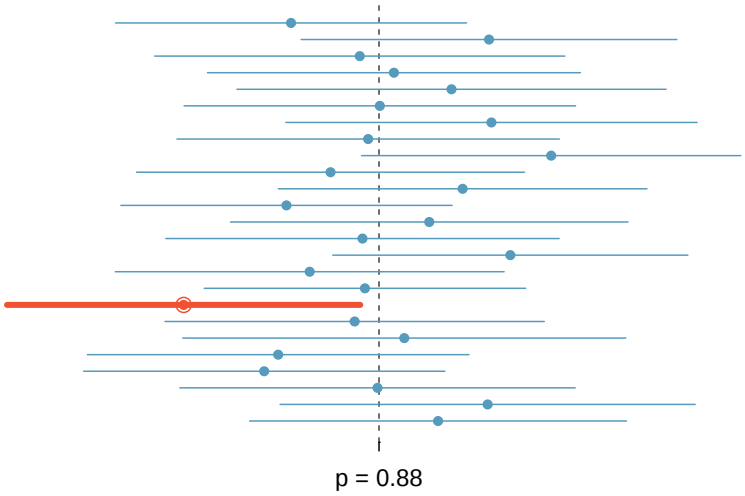


Figure 5.6: Twenty-five point estimates and confidence intervals from the simulations in Section 5.1.2. These intervals are shown relative to the population proportion $p = 0.88$. Only 1 of these 25 intervals did not capture the population proportion, and this interval has been bolded.

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

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

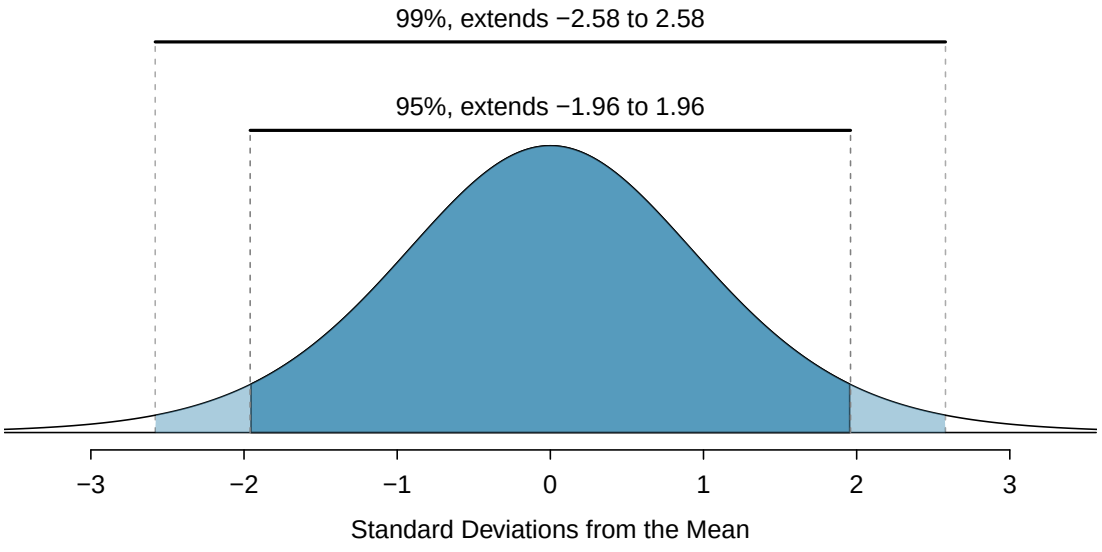


Figure 5.7: The area between -$z^\star$ and $z^\star$ increases as $z^\star$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of a normal normal distribution is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star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

[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

$$\hat{p} \pm 1.6449 \times SE_{\hat{p}} \quad \rightarrow \quad 0.887 \pm 1.65 \times 0.0100 \quad \rightarrow \quad (0.8705, 0.9034)$$

That is, we are 90% confident that 87.1% to 90.3% 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.

Exercises

5.7 Chronic illness, Part I. In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions".[11] However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used in this setting. Create a 95% confidence interval for the proportion of U.S. adults who live with one or more chronic conditions. Also interpret the confidence interval in the context of the study.

5.8 Twitter users and news, Part I. A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter.[12]. The standard error for this estimate was 2.4%, and a normal distribution may be used to model the sample proportion. Construct a 99% confidence interval for the fraction of U.S. adult Twitter users who get some news on Twitter, and interpret the confidence interval in context.

5.9 Chronic illness, Part II. In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions", and the standard error for this estimate is 1.2%. Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

(a) We can say with certainty that the confidence interval from Exercise 5.7 contains the true percentage of U.S. adults who suffer from a chronic illness.

(b) If we repeated this study 1,000 times and constructed a 95% confidence interval for each study, then approximately 950 of those confidence intervals would contain the true fraction of U.S. adults who suffer from chronic illnesses.

(c) The poll provides statistically significant evidence (at the $\alpha = 0.05$ level) that the percentage of U.S. adults who suffer from chronic illnesses is below 50%.

(d) Since the standard error is 1.2%, only 1.2% of people in the study communicated uncertainty about their answer.

5.10 Twitter users and news, Part II. A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter, and the standard error for this estimate was 2.4%. Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

(a) The data provide statistically significant evidence that more than half of U.S. adult Twitter users get some news through Twitter. Use a significance level of $\alpha = 0.01$. (This part uses concepts from Section 5.3 and will be corrected in a future edition.)

(b) Since the standard error is 2.4%, we can conclude that 97.6% of all U.S. adult Twitter users were included in the study.

(c) If we want to reduce the standard error of the estimate, we should collect less data.

(d) If we construct a 90% confidence interval for the percentage of U.S. adults Twitter users who get some news through Twitter, this confidence interval will be wider than a corresponding 99% confidence interval.

[11] Pew Research Center, Washington, D.C. The Diagnosis Difference, November 26, 2013.

[12] Pew Research Center, Washington, D.C. Twitter News Consumers: Young, Mobile and Educated, November 4, 2013.

5.11 Waiting at an ER, Part I. A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. She collects a simple random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. A 95% confidence interval based on this sample is (128 minutes, 147 minutes), which is based on the normal model for the mean. Determine whether the following statements are true or false, and explain your reasoning.

(a) We are 95% confident that the average waiting time of these 64 emergency room patients is between 128 and 147 minutes.

(b) We are 95% confident that the average waiting time of all patients at this hospital's emergency room is between 128 and 147 minutes.

(c) 95% of random samples have a sample mean between 128 and 147 minutes.

(d) A 99% confidence interval would be narrower than the 95% confidence interval since we need to be more sure of our estimate.

(e) The margin of error is 9.5 and the sample mean is 137.5.

(f) In order to decrease the margin of error of a 95% confidence interval to half of what it is now, we would need to double the sample size. (Hint: the margin of error for a mean scales in the same way with sample size as the margin of error for a proportion.)

5.12 Mental health. The General Social Survey asked the question: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.

(a) Interpret this interval in context of the data.

(b) What does "95% confident" mean? Explain in the context of the application.

(c) Suppose the researchers think a 99% confidence level would be more appropriate for this interval. Will this new interval be smaller or wider than the 95% confidence interval?

(d) If a new survey were to be done with 500 Americans, do you think the standard error of the estimate be larger, smaller, or about the same.

5.13 Website registration. A website is trying to increase registration for first-time visitors, exposing 1% of these visitors to a new site design. Of 752 randomly sampled visitors over a month who saw the new design, 64 registered.

(a) Check any conditions required for constructing a confidence interval.

(b) Compute the standard error.

(c) Construct and interpret a 90% confidence interval for the fraction of first-time visitors of the site who would register under the new design (assuming stable behaviors by new visitors over time).

5.14 Coupons driving visits. A store randomly samples 603 shoppers over the course of a year and finds that 142 of them made their visit because of a coupon they'd received in the mail. Construct a 95% confidence interval for the fraction of all shoppers during the year whose visit was because of a coupon they'd received in the mail.

5.3 Hypothesis testing for a proportion

The following question comes from a book written by Hans Rosling, Anna Rosling Rönnlund, and Ola Rosling called *Factfulness*:

> *How many of the world's 1 year old children today have been vaccinated against some disease:*
>
> *a. 20%*
> *b. 50%*
> *c. 80%*

Write down what your answer (or guess), and when you're ready, find the answer in the footnote.[13]

In this section, we'll be exploring how people with a 4-year college degree perform on this and other world health questions as we learn about hypothesis tests, which are a framework used to rigorously evaluate competing ideas and claims.

5.3.1 Hypothesis testing framework

We're interested in understanding how much people know about world health and development. If we take a multiple choice world health question, then we might like to understand if

H_0: People never learn these particular topics and their responses are simply equivalent to random guesses.

H_A: People have knowledge that helps them do better than random guessing, or perhaps, they have false knowledge that leads them to actually do worse than random guessing.

These competing ideas are called **hypotheses**. We call H_0 the null hypothesis and H_A the alternative hypothesis. When there is a subscript 0 like in H_0, data scientists pronounce it as "nought" (e.g. H_0 is pronounced "H-nought").

NULL AND ALTERNATIVE HYPOTHESES

The **null hypothesis** (H_0) often represents a skeptical perspective or a claim to be tested. The **alternative hypothesis** (H_A) represents an alternative claim under consideration and is often represented by a range of possible parameter values.

Our job as data scientists is to play the role of a skeptic: before we buy into the alternative hypothesis, we need to see strong supporting evidence.

The null hypothesis often represents a skeptical position or a perspective of "no difference". In our first example, we'll consider whether the typical person does any different than random guessing on Roslings' question about infant vaccinations.

The alternative hypothesis generally represents a new or stronger perspective. In the case of the question about infant vaccinations, it would certainly be interesting to learn whether people do better than random guessing, since that would mean that the typical person knows something about world health statistics. It would also be very interesting if we learned that people do *worse* than random guessing, which would suggest people believe incorrect information about world health.

The hypothesis testing framework is a very general tool, and we often use it without a second thought. If a person makes a somewhat unbelievable claim, we are initially skeptical. However, if there is sufficient evidence that supports the claim, we set aside our skepticism and reject the null hypothesis in favor of the alternative. The hallmarks of hypothesis testing are also found in the US court system.

[13] The correct answer is (c): 80% of the world's 1 year olds have been vaccinated against some disease.

GUIDED PRACTICE 5.17

G A US court considers two possible claims about a defendant: she is either innocent or guilty. If we set these claims up in a hypothesis framework, which would be the null hypothesis and which the alternative?[14]

Jurors examine the evidence to see whether it convincingly shows a defendant is guilty. Even if the jurors leave unconvinced of guilt beyond a reasonable doubt, this does not mean they believe the defendant is innocent. This is also the case with hypothesis testing: *even if we fail to reject the null hypothesis, we typically do not accept the null hypothesis as true.* Failing to find strong evidence for the alternative hypothesis is not equivalent to accepting the null hypothesis.

When considering Roslings' question about infant vaccination, the null hypothesis represents the notion that the people we will be considering – college-educated adults – are as accurate as random guessing. That is, the proportion p of respondents who pick the correct answer, that 80% of 1 year olds have been vaccinated against some disease, is about 33.3% (or 1-in-3 if wanting to be perfectly precise). The alternative hypothesis is that this proportion is something other than 33.3%. While it's helpful to write these hypotheses in words, it can be useful to write them using mathematical notation:

H_0**:** $p = 0.333$

H_A**:** $p \neq 0.333$

In this hypothesis setup, we want to make a conclusion about the population parameter p. The value we are comparing the parameter to is called the **null value**, which in this case is 0.333. It's common to label the null value with the same symbol as the parameter but with a subscript '0'. That is, in this case, the null value is $p_0 = 0.333$ (pronounced "p-nought equals 0.333").

EXAMPLE 5.18

It may seem impossible that the proportion of people who get the correct answer is *exactly* 33.3%. If we don't believe the null hypothesis, should we simply reject it?

E No. While we may not buy into the notion that the proportion is exactly 33.3%, the hypothesis testing framework requires that there be strong evidence before we reject the null hypothesis and conclude something more interesting.

After all, even if we don't believe the proportion is *exactly* 33.3%, that doesn't really tell us anything useful! We would still be stuck with the original question: do people do better or worse than random guessing on Roslings' question? Without data that strongly points in one direction or the other, it is both uninteresting and pointless to reject H_0.

GUIDED PRACTICE 5.19

G Another example of a real-world hypothesis testing situation is evaluating whether a new drug is better or worse than an existing drug at treating a particular disease. What should we use for the null and alternative hypotheses in this case?[15]

[14] The jury considers whether the evidence is so convincing (strong) that there is no reasonable doubt regarding the person's guilt; in such a case, the jury rejects innocence (the null hypothesis) and concludes the defendant is guilty (alternative hypothesis).

[15] The null hypothesis (H_0) in this case is the declaration of *no difference*: the drugs are equally effective. The alternative hypothesis (H_A) is that the new drug performs differently than the original, i.e. it could perform better or worse.

5.3.2 Testing hypotheses using confidence intervals

We will use the `rosling_responses` data set to evaluate the hypothesis test evaluating whether college-educated adults who get the question about infant vaccination correct is different from 33.3%. This data set summarizes the answers of 50 college-educated adults. Of these 50 adults, 24% of respondents got the question correct that 80% of 1 year olds have been vaccinated against some disease.

Up until now, our discussion has been philosophical. However, now that we have data, we might ask ourselves: does the data provide strong evidence that the proportion of all college-educated adults who would answer this question correctly is different than 33.3%?

We learned in Section 5.1 that there is fluctuation from one sample to another, and it is unlikely that our sample proportion, $\hat{p}$, will exactly equal p, but we want to make a conclusion about p. We have a nagging concern: is this deviation of 24% from 33.3% simply due to chance, or does the data provide strong evidence that the population proportion is different from 33.3%?

In Section 5.2, we learned how to quantify the uncertainty in our estimate using confidence intervals. The same method for measuring variability can be useful for the hypothesis test.

EXAMPLE 5.20

Check whether it is reasonable to construct a confidence interval for p using the sample data, and if so, construct a 95% confidence interval.

The conditions are met for $\hat{p}$ to be approximately normal: the data come from a simple random sample (satisfies independence), and $n\hat{p} = 12$ and $n(1-\hat{p}) = 38$ are both at least 10 (success-failure condition).

(E) To construct the confidence interval, we will need to identify the point estimate ($\hat{p} = 0.24$), the critical value for the 95% confidence level ($z^\star = 1.96$), and the standard error of $\hat{p}$ ($SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n} = 0.060$). With those pieces, the confidence interval for p can be constructed:

$$\hat{p} \pm z^\star \times SE_{\hat{p}}$$
$$0.24 \pm 1.96 \times 0.060$$
$$(0.122, 0.358)$$

We are 95% confident that the proportion of all college-educated adults to correctly answer this particular question about infant vaccination is between 12.2% and 35.8%.

Because the null value in the hypothesis test is $p_0 = 0.333$, which falls within the range of plausible values from the confidence interval, we cannot say the null value is implausible.[16] That is, the data do not provide sufficient evidence to reject the notion that the performance of college-educated adults was different than random guessing, and we do not reject the null hypothesis, H_0.

EXAMPLE 5.21

Explain why we cannot conclude that college-educated adults simply guessed on the infant vaccination question.

(E) While we failed to reject H_0, that does not necessarily mean the null hypothesis is true. Perhaps there was an actual difference, but we were not able to detect it with the relatively small sample of 50.

DOUBLE NEGATIVES CAN SOMETIMES BE USED IN STATISTICS

In many statistical explanations, we use double negatives. For instance, we might say that the null hypothesis is *not implausible* or we *failed to reject* the null hypothesis. Double negatives are used to communicate that while we are not rejecting a position, we are also not saying it is correct.

[16] Arguably this method is slightly imprecise. As we'll see in a few pages, the standard error is often computed slightly differently in the context of a hypothesis test for a proportion.

GUIDED PRACTICE 5.22

Let's move onto a second question posed by the Roslings:

> *There are 2 billion children in the world today aged 0-15 years old, how many children will there be in year 2100 according to the United Nations?*
>
> *a. 4 billion.*
> *b. 3 billion.*
> *c. 2 billion.*

Set up appropriate hypotheses to evaluate whether college-educated adults are better than random guessing on this question. Also, see if you can guess the correct answer before checking the answer in the footnote![17]

GUIDED PRACTICE 5.23

This time we took a larger sample of 228 college-educated adults, 34 (14.9%) selected the correct answer to the question in Guided Practice 5.22: 2 billion. Can we model the sample proportion using a normal distribution and construct a confidence interval?[18]

EXAMPLE 5.24

Compute a 95% confidence interval for the fraction of college-educated adults who answered the children-in-2100 question correctly, and evaluate the hypotheses in Guided Practice 5.22.

To compute the standard error, we'll again use $\hat{p}$ in place of p for the calculation:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.149(1-0.149)}{228}} = 0.024$$

In Guided Practice 5.23, we found that $\hat{p}$ can be modeled using a normal distribution, which ensures a 95% confidence interval may be accurately constructed as

$$\hat{p} \pm z^\star \times SE \quad \rightarrow \quad 0.149 \pm 1.96 \times 0.024 \quad \rightarrow \quad (0.103, 0.195)$$

Because the null value, $p_0 = 0.333$, is not in the confidence interval, a population proportion of 0.333 is implausible and we reject the null hypothesis. That is, the data provide statistically significant evidence that the actual proportion of college adults who get the children-in-2100 question correct is different from random guessing. Because the entire 95% confidence interval is below 0.333, we can conclude college-educated adults do *worse* than random guessing on this question.

One subtle consideration is that we used a 95% confidence interval. What if we had used a 99% confidence level? Or even a 99.9% confidence level? It's possible to come to a different conclusion if using a different confidence level. Therefore, when we make a conclusion based on confidence interval, we should also be sure it is clear what confidence level we used.

The worse-than-random performance on this last question is not a fluke: there are many such world health questions where people do worse than random guessing. In general, the answers suggest that people tend to be more pessimistic about progress than reality suggests. This topic is discussed in much greater detail in the Roslings' book, *Factfulness*.

---

[17] The appropriate hypotheses are:
H_0: the proportion who get the answer correct is the same as random guessing: 1-in-3, or $p = 0.333$.
H_A: the proportion who get the answer correct is different than random guessing, $p \neq 0.333$.
The correct answer to the question is 2 billion. While the world population is projected to increase, the average age is also expected to rise. That is, the majority of the population growth will happen in older age groups, meaning people are projected to live longer in the future across much of the world.

[18] We check both conditions, which are satisfied, so it is reasonable to use a normal distribution for $\hat{p}$:
Independence. Since the data are from a simple random sample, the observations are independent.
Success-failure. We'll use $\hat{p}$ in place of p to check: $n\hat{p} = 34$ and $n(1-\hat{p}) = 194$. Both are greater than 10, so the success-failure condition is satisfied.

5.3.3 Decision errors

Hypothesis tests are not flawless: we can make an incorrect decision in a statistical hypothesis test based on the data. For example, in the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free. One key distinction with statistical hypothesis tests is that we have the tools necessary to probabilistically quantify how often we make errors in our conclusions.

Recall that there are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a statement about which one might be true, but we might choose incorrectly. There are four possible scenarios, which are summarized in Figure 5.8.

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	**Type 1 Error**
	H_A true	**Type 2 Error**	okay

Figure 5.8: Four different scenarios for hypothesis tests.

A **Type 1 Error** is rejecting the null hypothesis when H_0 is actually true. A **Type 2 Error** is failing to reject the null hypothesis when the alternative is actually true.

GUIDED PRACTICE 5.25

G In a US court, the defendant is either innocent (H_0) or guilty (H_A). What does a Type 1 Error represent in this context? What does a Type 2 Error represent? Figure 5.8 may be useful.[19]

EXAMPLE 5.26

How could we reduce the Type 1 Error rate in US courts? What influence would this have on the Type 2 Error rate?

E To lower the Type 1 Error rate, we might raise our standard for conviction from "beyond a reasonable doubt" to "beyond a conceivable doubt" so fewer people would be wrongly convicted. However, this would also make it more difficult to convict the people who are actually guilty, so we would make more Type 2 Errors.

GUIDED PRACTICE 5.27

G How could we reduce the Type 2 Error rate in US courts? What influence would this have on the Type 1 Error rate?[20]

Exercises 5.25-5.27 provide an important lesson: if we reduce how often we make one type of error, we generally make more of the other type.

Hypothesis testing is built around rejecting or failing to reject the null hypothesis. That is, we do not reject H_0 unless we have strong evidence. But what precisely does *strong evidence* mean? As a general rule of thumb, for those cases where the null hypothesis is actually true, we do not want to incorrectly reject H_0 more than 5% of the time. This corresponds to a **significance level** of 0.05. That is, if the null hypothesis is true, the significance level indicates how often the data lead us to incorrectly reject H_0. We often write the significance level using α (the Greek letter *alpha*): $\alpha = 0.05$. We discuss the appropriateness of different significance levels in Section 5.3.5.

[19] If the court makes a Type 1 Error, this means the defendant is innocent (H_0 true) but wrongly convicted. Note that a Type 1 Error is only possible if we've rejected the null hypothesis.

A Type 2 Error means the court failed to reject H_0 (i.e. failed to convict the person) when she was in fact guilty (H_A true). Note that a Type 2 Error is only possible if we have failed to reject the null hypothesis.

[20] To lower the Type 2 Error rate, we want to convict more guilty people. We could lower the standards for conviction from "beyond a reasonable doubt" to "beyond a little doubt". Lowering the bar for guilt will also result in more wrongful convictions, raising the Type 1 Error rate.

If we use a 95% confidence interval to evaluate a hypothesis test and the null hypothesis happens to be true, we will make an error whenever the point estimate is at least 1.96 standard errors away from the population parameter. This happens about 5% of the time (2.5% in each tail). Similarly, using a 99% confidence interval to evaluate a hypothesis is equivalent to a significance level of $\alpha = 0.01$.

A confidence interval is very helpful in determining whether or not to reject the null hypothesis. However, the confidence interval approach isn't always sustainable. In several sections, we will encounter situations where a confidence interval cannot be constructed. For example, if we wanted to evaluate the hypothesis that several proportions are equal, it isn't clear how to construct and compare many confidence intervals altogether.

Next we will introduce a statistic called the *p-value* to help us expand our statistical toolkit, which will enable us to both better understand the strength of evidence and work in more complex data scenarios in later sections.

5.3.4 Formal testing using p-values

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative hypothesis. Statistical hypothesis testing typically uses the p-value method rather than making a decision based on confidence intervals.

P-VALUE

The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true. We typically use a summary statistic of the data, in this section the sample proportion, to help compute the p-value and evaluate the hypotheses.

EXAMPLE 5.28

Pew Research asked a random sample of 1000 American adults whether they supported the increased usage of coal to produce energy. Set up hypotheses to evaluate whether a majority of American adults support or oppose the increased usage of coal.

(E) The uninteresting result is that there is no majority either way: half of Americans support and the other half oppose expanding the use of coal to produce energy. The alternative hypothesis would be that there is a majority support or oppose (though we do not known which one!) expanding the use of coal. If p represents the proportion supporting, then we can write the hypotheses as

H_0**:** $p = 0.5$

H_A**:** $p \neq 0.5$

In this case, the null value is $p_0 = 0.5$.

When evaluating hypotheses for proportions using the p-value method, we will slightly modify how we check the success-failure condition and compute the standard error for the single proportion case. These changes aren't dramatic, but pay close attention to how we use the null value, p_0.

EXAMPLE 5.29

Pew Research's sample show that 37% of American adults support increased usage of coal. We now wonder, does 37% represent a real difference from the null hypothesis of 50%? What would the sampling distribution of $\hat{p}$ look like if the null hypothesis were true?

If the null hypothesis were true, the population proportion would be the null value, 0.5. We previously learned that the sampling distribution of $\hat{p}$ will be normal when two conditions are met:

Independence. The poll was based on a simple random sample, so independence is satisfied.

Success-failure. Based on the poll's sample size of $n = 1000$, the success-failure condition is met, since

$$np \overset{H_0}{=} 1000 \times 0.5 = 500 \qquad n(1-p) \overset{H_0}{=} 1000 \times (1 - 0.5) = 500$$

(E) are both at least 10. Note that the success-failure condition was checked using the null value, $p_0 = 0.5$; this is the first procedural difference from confidence intervals.

If the null hypothesis were true, the sampling distribution indicates that a sample proportion based on $n = 1000$ observations would be normally distributed. Next, we can compute the standard error, where we will again use the null value $p_0 = 0.5$ in the calculation:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \overset{H_0}{=} \sqrt{\frac{0.5 \times (1 - 0.5)}{1000}} = 0.016$$

This marks the other procedural difference from confidence intervals: since the sampling distribution is determined under the null proportion, the null value p_0 was used for the proportion in the calculation rather than $\hat{p}$.

Ultimately, if the null hypothesis were true, then the sample proportion should follow a normal distribution with mean 0.5 and a standard error of 0.016. This distribution is shown in Figure 5.9.

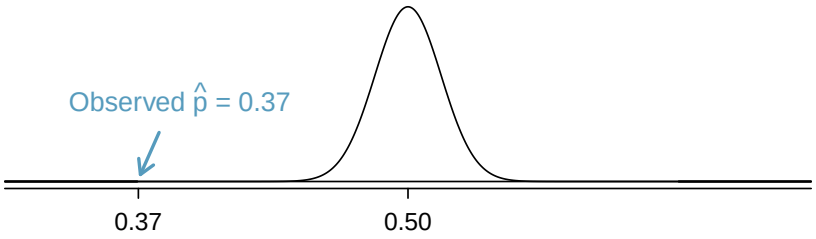


Figure 5.9: If the null hypothesis were true, this normal distribution describes the distribution of $\hat{p}$.

CHECKING SUCCESS-FAILURE AND COMPUTING $SE_{\hat{p}}$ FOR A HYPOTHESIS TEST

When using the p-value method to evaluate a hypothesis test, we check the conditions for $\hat{p}$ and construct the standard error using the null value, p_0, instead of using the sample proportion.

In a hypothesis test with a p-value, we are supposing the null hypothesis is true, which is a different mindset than when we compute a confidence interval. This is why we use p_0 instead of $\hat{p}$ when we check conditions and compute the standard error in this context.

When we identify the sampling distribution under the null hypothesis, it has a special name: the **null distribution**. The p-value represents the probability of the observed $\hat{p}$, or a $\hat{p}$ that is more extreme, if the null hypothesis were true. To find the p-value, we generally find the null distribution, and then we find a tail area in that distribution corresponding to our point estimate.

EXAMPLE 5.30

If the null hypothesis were true, determine the chance of finding $\hat{p}$ at least as far into the tails as 0.37 under the null distribution, which is a normal distribution with mean $\mu = 0.5$ and $SE = 0.016$.

This is a normal probability problem where $x = 0.37$. First, we draw a simple graph to represent the situation, similar to what is shown in Figure 5.9. Since $\hat{p}$ is so far out in the tail, we know the tail area is going to be very small. To find it, we start by computing the Z-score using the mean of 0.5 and the standard error of 0.016:

$$Z = \frac{0.37 - 0.5}{0.016} = -8.125$$

(E)

We can use software to find the tail area: 2.2×10^{-16} (0.00000000000000022). If using the normal probability table in Appendix C.1, we'd find that $Z = -8.125$ is off the table, so we would use the smallest area listed: 0.0002.

The potential $\hat{p}$'s in the upper tail beyond 0.63, which are shown in Figure 5.10, also represent observations at least as extreme as the observed value of 0.37. To account for these values that are also more extreme under the hypothesis setup, we double the lower tail to get an estimate of the p-value: 4.4×10^{-16} (or if using the table method, 0.0004).

The p-value represents the probability of observing such an extreme sample proportion by chance, if the null hypothesis were true.

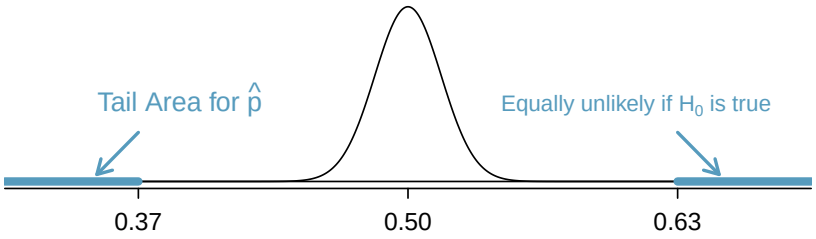


Figure 5.10: If H_0 were true, then the values above 0.63 are just as unlikely as values below 0.37.

EXAMPLE 5.31

How should we evaluate the hypotheses using the p-value of 4.4×10^{-16}? Use the standard significance level of $\alpha = 0.05$.

If the null hypothesis were true, there's only an incredibly small chance of observing such an extreme deviation of $\hat{p}$ from 0.5. This means one of the following must be true:

(E)

1. The null hypothesis is true, and we just happened to observe something so extreme that it only happens about once in every 23 quadrillion times (1 quadrillion = 1 million × 1 billion).
2. The alternative hypothesis is true, which would be consistent with observing a sample proportion far from 0.5.

The first scenario is laughably improbable, while the second scenario seems much more plausible.

Formally, when we evaluate a hypothesis test, we compare the p-value to the significance level, which in this case is $\alpha = 0.05$. Since the p-value is less than α, we reject the null hypothesis. That is, the data provide strong evidence against H_0. The data indicate the direction of the difference: a majority of Americans do not support expanding the use of coal-powered energy.

COMPARE THE P-VALUE TO α TO EVALUATE H_0

When the p-value is less than the significance level, α, reject H_0. We would report a conclusion that the data provide strong evidence supporting the alternative hypothesis.

When the p-value is greater than α, do not reject H_0, and report that we do not have sufficient evidence to reject the null hypothesis.

In either case, it is important to describe the conclusion in the context of the data.

GUIDED PRACTICE 5.32

(G) Do a majority of Americans support or oppose nuclear arms reduction? Set up hypotheses to evaluate this question.[21]

EXAMPLE 5.33

A simple random sample of 1028 US adults in March 2013 show that 56% support nuclear arms reduction. Does this provide convincing evidence that a majority of Americans supported nuclear arms reduction at the 5% significance level?

First, check conditions:

Independence. The poll was of a simple random sample of US adults, meaning the observations are independent.

Success-failure. In a one-proportion hypothesis test, this condition is checked using the null proportion, which is $p_0 = 0.5$ in this context: $np_0 = n(1 - p_0) = 1028 \times 0.5 = 514 \geq 10$.

With these conditions verified, we can model $\hat{p}$ using a normal model.

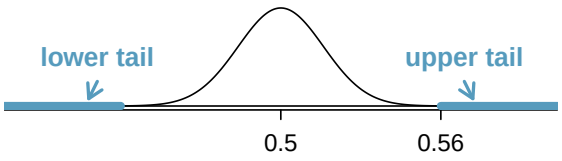
Next the standard error can be computed. The null value p_0 is used again here, because this is a hypothesis test for a single proportion.

(E)
$$SE_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{1028}} = 0.0156$$

Based on the normal model, the test statistic can be computed as the Z-score of the point estimate:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.56 - 0.50}{0.0156} = 3.85$$

It's generally helpful to draw null distribution and the tail areas of interest for computing the p-value:



The upper tail area is about 0.0001, and we double this tail area to get the p-value: 0.0002. Because the p-value is smaller than 0.05, we reject H_0. The poll provides convincing evidence that a majority of Americans supported nuclear arms reduction efforts in March 2013.

[21] We would like to understand if a majority supports or opposes, or ultimately, if there is no difference. If p is the proportion of Americans who support nuclear arms reduction, then H_0: $p = 0.50$ and H_A: $p \neq 0.50$.

HYPOTHESIS TESTING FOR A SINGLE PROPORTION

Once you've determined a one-proportion hypothesis test is the correct procedure, there are four steps to completing the test:

Prepare. Identify the parameter of interest, list hypotheses, identify the significance level, and identify $\hat{p}$ and n.

Check. Verify conditions to ensure $\hat{p}$ is nearly normal under H_0. For one-proportion hypothesis tests, use the null value to check the success-failure condition.

Calculate. If the conditions hold, compute the standard error, again using p_0, compute the Z-score, and identify the p-value.

Conclude. Evaluate the hypothesis test by comparing the p-value to α, and provide a conclusion in the context of the problem.

5.3.5 Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is $\alpha = 0.05$. However, it can be helpful to adjust the significance level based on the application. We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.

If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.

If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we might choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the alternative hypothesis is actually true.

Additionally, if the cost of collecting data is small relative to the cost of a Type 2 Error, then it may also be a good strategy to collect more data. Under this strategy, the Type 2 Error can be reduced while not affecting the Type 1 Error rate. Of course, collecting extra data is often costly, so there is typically a cost-benefit analysis to be considered.

EXAMPLE 5.34

A car manufacturer is considering switching to a new, higher quality piece of equipment that constructs vehicle door hinges. They figure that they will save money in the long run if this new machine produces hinges that have flaws less than 0.2% of the time. However, if the hinges are flawed more than 0.2% of the time, they wouldn't get a good enough return-on-investment from the new piece of equipment, and they would lose money. Is there good reason to modify the significance level in such a hypothesis test?

(E)

The null hypothesis would be that the rate of flawed hinges is 0.2%, while the alternative is that it the rate is different than 0.2%. This decision is just one of many that have a marginal impact on the car and company. A significance level of 0.05 seems reasonable since neither a Type 1 or Type 2 Error should be dangerous or (relatively) much more expensive.

EXAMPLE 5.35

The same car manufacturer is considering a slightly more expensive supplier for parts related to safety, not door hinges. If the durability of these safety components is shown to be better than the current supplier, they will switch manufacturers. Is there good reason to modify the significance level in such an evaluation?

(E)

The null hypothesis would be that the suppliers' parts are equally reliable. Because safety is involved, the car company should be eager to switch to the slightly more expensive manufacturer (reject H_0), even if the evidence of increased safety is only moderately strong. A slightly larger significance level, such as $\alpha = 0.10$, might be appropriate.

GUIDED PRACTICE 5.36

(G) A part inside of a machine is very expensive to replace. However, the machine usually functions properly even if this part is broken, so the part is replaced only if we are extremely certain it is broken based on a series of measurements. Identify appropriate hypotheses for this test (in plain language) and suggest an appropriate significance level.[22]

WHY IS 0.05 THE DEFAULT?

The $\alpha = 0.05$ threshold is most common. But why? Maybe the standard level should be smaller, or perhaps larger. If you're a little puzzled, you're reading with an extra critical eye – good job! We've made a 5-minute task to help clarify *why 0.05*:

www.openintro.org/why05

5.3.6 Statistical significance versus practical significance

When the sample size becomes larger, point estimates become more precise and any real differences in the mean and null value become easier to detect and recognize. Even a very small difference would likely be detected if we took a large enough sample. Sometimes researchers will take such large samples that even the slightest difference is detected, even differences where there is no practical value. In such cases, we still say the difference is **statistically significant**, but it is not **practically significant**. For example, an online experiment might identify that placing additional ads on a movie review website statistically significantly increases viewership of a TV show by 0.001%, but this increase might not have any practical value.

One role of a data scientist in conducting a study often includes planning the size of the study. The data scientist might first consult experts or scientific literature to learn what would be the smallest meaningful difference from the null value. She also would obtain other information, such as a very rough estimate of the true proportion p, so that she could roughly estimate the standard error. From here, she can suggest a sample size that is sufficiently large that, if there is a real difference that is meaningful, we could detect it. While larger sample sizes may still be used, these calculations are especially helpful when considering costs or potential risks, such as possible health impacts to volunteers in a medical study.

[22] Here the null hypothesis is that the part is not broken, and the alternative is that it is broken. If we don't have sufficient evidence to reject H_0, we would not replace the part. It sounds like failing to fix the part if it is broken (H_0 false, H_A true) is not very problematic, and replacing the part is expensive. Thus, we should require very strong evidence against H_0 before we replace the part. Choose a small significance level, such as $\alpha = 0.01$.

5.3.7 One-sided hypothesis tests (special topic)

So far we've only considered what are called **two-sided hypothesis tests**, where we care about detecting whether p is either above or below some null value p_0. There is a second type of hypothesis test called a **one-sided hypothesis test**. For a one-sided hypothesis test, the hypotheses take one of the following forms:

1. There's only value in detecting if the population parameter is *less than* some value p_0. In this case, the alternative hypothesis is written as $p < p_0$ for some null value p_0.

2. There's only value in detecting if the population parameter is *more than* some value p_0: In this case, the alternative hypothesis is written as $p > p_0$.

While we adjust the form of the alternative hypothesis, we continue to write the null hypothesis using an equals-sign in the one-sided hypothesis test case.

In the entire hypothesis testing procedure, there is only one difference in evaluating a one-sided hypothesis test vs a two-sided hypothesis test: how to compute the p-value. In a one-sided hypothesis test, we compute the p-value as the tail area in the *direction of the alternative hypothesis only*, meaning it is represented by a single tail area. Herein lies the reason why one-sided tests are sometimes interesting: if we don't have to double the tail area to get the p-value, then the p-value is smaller and the level of evidence required to identify an interesting finding in the direction of the alternative hypothesis goes down. However, one-sided tests aren't all sunshine and rainbows: the heavy price paid is that any interesting findings in the opposite direction must be disregarded.

EXAMPLE 5.37

In Section 1.1, we encountered an example where doctors were interested in determining whether stents would help people who had a high risk of stroke. The researchers believed the stents would help. Unfortunately, the data showed the opposite: patients who received stents actually did worse. Why was using a two-sided test so important in this context?

(E)

Before the study, researchers had reason to believe that stents would help patients since existing research suggested stents helped in patients with heart attacks. It would surely have been tempting to use a one-sided test in this situation, and had they done this, they would have limited their ability to identify potential harm to patients.

Example 5.37 highlights that using a one-sided hypothesis creates a risk of overlooking data supporting the opposite conclusion. We could have made a similar error when reviewing the Roslings' question data this section; if we had a pre-conceived notion that college-educated people wouldn't do worse than random guessing and so used a one-sided test, we would have missed the really interesting finding that many people have incorrect knowledge about global public health.

When might a one-sided test be appropriate to use? *Very rarely.* Should you ever find yourself considering using a one-sided test, carefully answer the following question:

> *What would I, or others, conclude if the data happens to go clearly in the opposite direction than my alternative hypothesis?*

If you or others would find any value in making a conclusion about the data that goes in the opposite direction of a one-sided test, then a two-sided hypothesis test should actually be used. These considerations can be subtle, so exercise caution. We will only apply two-sided tests in the rest of this book.

EXAMPLE 5.38

Why can't we simply run a one-sided test that goes in the direction of the data?

We've been building a careful framework that controls for the Type 1 Error, which is the significance level α in a hypothesis test. We'll use the $\alpha = 0.05$ below to keep things simple.

Imagine we could pick the one-sided test after we saw the data. What will go wrong?

(E)

- If $\hat{p}$ is *smaller* than the null value, then a one-sided test where $p < p_0$ would mean that any observation in the *lower* 5% tail of the null distribution would lead to us rejecting H_0.
- If $\hat{p}$ is *larger* than the null value, then a one-sided test where $p > p_0$ would mean that any observation in the *upper* 5% tail of the null distribution would lead to us rejecting H_0.

Then if H_0 were true, there's a 10% chance of being in one of the two tails, so our testing error is actually $\alpha = 0.10$, not 0.05. That is, not being careful about when to use one-sided tests effectively undermines the methods we're working so hard to develop and utilize.

Exercises

5.15 Identify hypotheses, Part I. Write the null and alternative hypotheses in words and then symbols for each of the following situations.

(a) A tutoring company would like to understand if most students tend to improve their grades (or not) after they use their services. They sample 200 of the students who used their service in the past year and ask them if their grades have improved or declined from the previous year.

(b) Employers at a firm are worried about the effect of March Madness, a basketball championship held each spring in the US, on employee productivity. They estimate that on a regular business day employees spend on average 15 minutes of company time checking personal email, making personal phone calls, etc. They also collect data on how much company time employees spend on such non-business activities during March Madness. They want to determine if these data provide convincing evidence that employee productivity changed during March Madness.

5.16 Identify hypotheses, Part II. Write the null and alternative hypotheses in words and using symbols for each of the following situations.

(a) Since 2008, chain restaurants in California have been required to display calorie counts of each menu item. Prior to menus displaying calorie counts, the average calorie intake of diners at a restaurant was 1100 calories. After calorie counts started to be displayed on menus, a nutritionist collected data on the number of calories consumed at this restaurant from a random sample of diners. Do these data provide convincing evidence of a difference in the average calorie intake of a diners at this restaurant?

(b) The state of Wisconsin would like to understand the fraction of its adult residents that consumed alcohol in the last year, specifically if the rate is different from the national rate of 70%. To help them answer this question, they conduct a random sample of 852 residents and ask them about their alcohol consumption.

5.17 Online communication. A study suggests that 60% of college student spend 10 or more hours per week communicating with others online. You believe that this is incorrect and decide to collect your own sample for a hypothesis test. You randomly sample 160 students from your dorm and find that 70% spent 10 or more hours a week communicating with others online. A friend of yours, who offers to help you with the hypothesis test, comes up with the following set of hypotheses. Indicate any errors you see.

$$H_0 : \hat{p} < 0.6$$
$$H_A : \hat{p} > 0.7$$

5.18 Married at 25. A study suggests that the 25% of 25 year olds have gotten married. You believe that this is incorrect and decide to collect your own sample for a hypothesis test. From a random sample of 25 year olds in census data with size 776, you find that 24% of them are married. A friend of yours offers to help you with setting up the hypothesis test and comes up with the following hypotheses. Indicate any errors you see.

$$H_0 : \hat{p} = 0.24$$
$$H_A : \hat{p} \neq 0.24$$

5.19 Cyberbullying rates. Teens were surveyed about cyberbullying, and 54% to 64% reported experiencing cyberbullying (95% confidence interval).[23] Answer the following questions based on this interval.

(a) A newspaper claims that a majority of teens have experienced cyberbullying. Is this claim supported by the confidence interval? Explain your reasoning.

(b) A researcher conjectured that 70% of teens have experienced cyberbullying. Is this claim supported by the confidence interval? Explain your reasoning.

(c) Without actually calculating the interval, determine if the claim of the researcher from part (b) would be supported based on a 90% confidence interval?

[23] Pew Research Center, A Majority of Teens Have Experienced Some Form of Cyberbullying. September 27, 2018.

5.20 Waiting at an ER, Part II. Exercise 5.11 provides a 95% confidence interval for the mean waiting time at an emergency room (ER) of (128 minutes, 147 minutes). Answer the following questions based on this interval.

(a) A local newspaper claims that the average waiting time at this ER exceeds 3 hours. Is this claim supported by the confidence interval? Explain your reasoning.

(b) The Dean of Medicine at this hospital claims the average wait time is 2.2 hours. Is this claim supported by the confidence interval? Explain your reasoning.

(c) Without actually calculating the interval, determine if the claim of the Dean from part (b) would be supported based on a 99% confidence interval?

5.21 Minimum wage, Part I. Do a majority of US adults believe raising the minimum wage will help the economy, or is there a majority who do not believe this? A Rasmussen Reports survey of a random sample of 1,000 US adults found that 42% believe it will help the economy.[24] Conduct an appropriate hypothesis test to help answer the research question.

5.22 Getting enough sleep. 400 students were randomly sampled from a large university, and 289 said they did not get enough sleep. Conduct a hypothesis test to check whether this represents a statistically significant difference from 50%, and use a significance level of 0.01.

5.23 Working backwards, Part I. You are given the following hypotheses:

$$H_0 : p = 0.3$$
$$H_A : p \neq 0.3$$

We know the sample size is 90. For what sample proportion would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

5.24 Working backwards, Part II. You are given the following hypotheses:

$$H_0 : p = 0.9$$
$$H_A : p \neq 0.9$$

We know that the sample size is 1,429. For what sample proportion would the p-value be equal to 0.01? Assume that all conditions necessary for inference are satisfied.

5.25 Testing for Fibromyalgia. A patient named Diana was diagnosed with Fibromyalgia, a long-term syndrome of body pain, and was prescribed anti-depressants. Being the skeptic that she is, Diana didn't initially believe that anti-depressants would help her symptoms. However after a couple months of being on the medication she decides that the anti-depressants are working, because she feels like her symptoms are in fact getting better.

(a) Write the hypotheses in words for Diana's skeptical position when she started taking the anti-depressants.

(b) What is a Type 1 Error in this context?

(c) What is a Type 2 Error in this context?

5.26 Which is higher? In each part below, there is a value of interest and two scenarios (I and II). For each part, report if the value of interest is larger under scenario I, scenario II, or whether the value is equal under the scenarios.

(a) The standard error of $\hat{p}$ when (I) $n = 125$ or (II) $n = 500$.

(b) The margin of error of a confidence interval when the confidence level is (I) 90% or (II) 80%.

(c) The p-value for a Z-statistic of 2.5 calculated based on a (I) sample with $n = 500$ or based on a (II) sample with $n = 1000$.

(d) The probability of making a Type 2 Error when the alternative hypothesis is true and the significance level is (I) 0.05 or (II) 0.10.

[24] Rasmussen Reports survey, Most Favor Minimum Wage of $10.50 Or Higher, April 16, 2019.

Chapter exercises

5.27 Relaxing after work. The General Social Survey asked the question: "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" to a random sample of 1,155 Americans.[25] A 95% confidence interval for the mean number of hours spent relaxing or pursuing activities they enjoy was (1.38, 1.92).

(a) Interpret this interval in context of the data.

(b) Suppose another set of researchers reported a confidence interval with a larger margin of error based on the same sample of 1,155 Americans. How does their confidence level compare to the confidence level of the interval stated above?

(c) Suppose next year a new survey asking the same question is conducted, and this time the sample size is 2,500. Assuming that the population characteristics, with respect to how much time people spend relaxing after work, have not changed much within a year. How will the margin of error of the 95% confidence interval constructed based on data from the new survey compare to the margin of error of the interval stated above?

5.28 Minimum wage, Part II. In Exercise 5.21, we learned that a Rasmussen Reports survey of 1,000 US adults found that 42% believe raising the minimum wage will help the economy. Construct a 99% confidence interval for the true proportion of US adults who believe this.

5.29 Testing for food safety. A food safety inspector is called upon to investigate a restaurant with a few customer reports of poor sanitation practices. The food safety inspector uses a hypothesis testing framework to evaluate whether regulations are not being met. If he decides the restaurant is in gross violation, its license to serve food will be revoked.

(a) Write the hypotheses in words.

(b) What is a Type 1 Error in this context?

(c) What is a Type 2 Error in this context?

(d) Which error is more problematic for the restaurant owner? Why?

(e) Which error is more problematic for the diners? Why?

(f) As a diner, would you prefer that the food safety inspector requires strong evidence or very strong evidence of health concerns before revoking a restaurant's license? Explain your reasoning.

5.30 True or false. Determine if the following statements are true or false, and explain your reasoning. If false, state how it could be corrected.

(a) If a given value (for example, the null hypothesized value of a parameter) is within a 95% confidence interval, it will also be within a 99% confidence interval.

(b) Decreasing the significance level (α) will increase the probability of making a Type 1 Error.

(c) Suppose the null hypothesis is $p = 0.5$ and we fail to reject H_0. Under this scenario, the true population proportion is 0.5.

(d) With large sample sizes, even small differences between the null value and the observed point estimate, a difference often called the effect size, will be identified as statistically significant.

5.31 Unemployment and relationship problems. A USA Today/Gallup poll asked a group of unemployed and underemployed Americans if they have had major problems in their relationships with their spouse or another close family member as a result of not having a job (if unemployed) or not having a full-time job (if underemployed). 27% of the 1,145 unemployed respondents and 25% of the 675 underemployed respondents said they had major problems in relationships as a result of their employment status.

(a) What are the hypotheses for evaluating if the proportions of unemployed and underemployed people who had relationship problems were different?

(b) The p-value for this hypothesis test is approximately 0.35. Explain what this means in context of the hypothesis test and the data.

[25] National Opinion Research Center, General Social Survey, 2018.

5.32 Nearsighted. It is believed that nearsightedness affects about 8% of all children. In a random sample of 194 children, 21 are nearsighted. Conduct a hypothesis test for the following question: do these data provide evidence that the 8% value is inaccurate?

5.33 Nutrition labels. The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a confidence interval for the number of calories per bag of 128.2 to 139.8 calories. Is there evidence that the nutrition label does not provide an accurate measure of calories in the bags of potato chips?

5.34 CLT for proportions. Define the term "sampling distribution" of the sample proportion, and describe how the shape, center, and spread of the sampling distribution change as the sample size increases when $p = 0.1$.

5.35 Practical vs. statistical significance. Determine whether the following statement is true or false, and explain your reasoning: "With large sample sizes, even small differences between the null value and the observed point estimate can be statistically significant."

5.36 Same observation, different sample size. Suppose you conduct a hypothesis test based on a sample where the sample size is $n = 50$, and arrive at a p-value of 0.08. You then refer back to your notes and discover that you made a careless mistake, the sample size should have been $n = 500$. Will your p-value increase, decrease, or stay the same? Explain.

5.37 Gender pay gap in medicine. A study examined the average pay for men and women entering the workforce as doctors for 21 different positions.[26]

(a) If each gender was equally paid, then we would expect about half of those positions to have men paid more than women and women would be paid more than men in the other half of positions. Write appropriate hypotheses to test this scenario.

(b) Men were, on average, paid more in 19 of those 21 positions. Supposing these 21 positions represent a simple random sample, complete a hypothesis test using your hypotheses from part (a).

[26] Lo Sasso AT et al. "The $16,819 Pay Gap For Newly Trained Physicians: The Unexplained Trend Of Men Earning More Than Women". In: *Health Affairs* 30.2 (2011).