OpenIntro Statistics

Fourth Edition

David Diez

Data Scientist, OpenIntro

Mine Çetinkaya-Rundel

Data Scientist & Professional Educator, RStudio and Duke University

Christopher D Barr

Investment Analyst

Contents

Preface

OpenIntro Statistics covers a first course in statistics. This book was written with the undergraduate level in mind, but it has also been popular in high schools and graduate courses. This diverse usage highlights the book's goals: provide a rigorous introduction to applied statistics that is clear, concise, and accessible.

We hope readers will take away three ideas from this book in addition to forming a foundation of statistical thinking and methods.

- Statistics is an applied field with a wide range of practical applications.
- You don't have to be a math guru to learn from real, interesting data.
- Data are messy, and statistical tools are imperfect. But, when you understand the strengths and weaknesses of these tools, you can use them to learn about the world.

Textbook overview

The chapters of this book are as follows:

1. **Introduction to data.** Data structures, variables, and basic data collection techniques.
2. **Summarizing data.** Data summaries, graphics, and a teaser of inference using randomization.
3. **Probability.** Basic principles of probability.
4. **Distributions of random variables.** The normal model and other key distributions.
5. **Foundations for inference.** General ideas for statistical inference in the context of estimating the population proportion.
6. **Inference for categorical data.** Inference for proportions and tables using the normal and chi-square distributions.
7. **Inference for numerical data.** Inference for one or two sample means using the t-distribution, statistical power for comparing two groups, and also comparisons of many means using ANOVA.
8. **Introduction to linear regression.** Regression for a numerical outcome with one predictor variable. Most of this chapter could be covered after Chapter **??**.
9. **Multiple and logistic regression.** Regression for numerical and categorical data using many predictors.

OpenIntro Statistics supports flexibility in choosing and ordering topics. If the main goal is to reach multiple regression (Chapter **??**) as quickly as possible, then the following are the ideal prerequisites:

- Chapter **??**, Sections **??**, and Section **??** for a solid introduction to data structures and statistical summaries that are used throughout the book.
- Section 4.1 for a solid understanding of the normal distribution.
- Chapter 5 to establish the core set of inference tools.
- Section 7.1 to give a foundation for the t-distribution
- Chapter **??** for establishing ideas and principles for single predictor regression.

Examples and exercises

Examples are provided to establish an understanding of how to apply methods

EXAMPLE 0.1

E

This is an example. When a question is asked here, where can the answer be found?

The answer can be found here, in the solution section of the example!

When we think the reader should be ready to try determining the solution to an example, we frame it as Guided Practice.

GUIDED PRACTICE 0.2

G

The reader may check or learn the answer to any Guided Practice problem by reviewing the full solution in a footnote.[1]

A large number of exercises are also provided at the end of each [section or chapter?] for practice or homework assignments. Solutions are given for odd-numbered exercises in Appendix **??**.

Additional resources

Video overviews, slides, statistical software labs, data sets used in the textbook, and much more are readily available at

openintro.org/os

We also have improved the ability to access data in this book through the addition of Appendix A, which provides additional information for each of the data sets used in the main text and is new in the Fourth Edition. Online guides to each of these data sets are also provided at **openintro.org/data**.

We appreciate all feedback as well as reports of any typos through the website. A short-link to report a new typo or review known typos is **openintro.org/os/typos**.

For those focused on statistics at the high school level, consider [ensure no name change] *Advanced High School Statistics*, which is a version of *OpenIntro Statistics* that has been heavily customized by Leah Dorazio for high school courses and AP® Statistics.

Acknowledgements

This project would not be possible without the passion and dedication of many more people beyond those on the author list. The authors would like to thank the OpenIntro Staff for their involvement and ongoing contributions. We are also very grateful to the hundreds of students and instructors who have provided us with valuable feedback since we first started posting book content in 2009.

We also want to thank the many teachers who helped review this edition, including Laura Acion, Matthew E. Aiello-Lammens, Jonathan Akin, Stacey C. Behrensmeyer, Jo Hardin, Nicholas Horton, Danish Khan, Peter H.M. Klaren, Jesse Mostipak, Jon C. New, Mario Orsi, and David Rockoff. [Ensure this list is complete.] We appreciate all of their feedback, which helped us tune the text in significant ways for the better.

[1] Guided Practice problems are intended to stretch your thinking, and you can check yourself by reviewing the footnote solution for any Guided Practice.

Chapter 4

Distributions of random variables

In this chapter, we discuss statistical distributions that frequently arise in the context of data analysis or statistical inference. We start with the normal distribution in the first section, which is used frequently in later chapters of this book. The remaining sections will occasionally be referenced but may be considered optional for the content in this book.

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

4.1 Normal distribution

Among all the distributions we see in practice, one is overwhelmingly the most common. The symmetric, unimodal, bell curve is ubiquitous throughout statistics. Indeed it is so common, that people often know it as the **normal curve** or **normal distribution**,[1] shown in Figure 4.1. Variables such as SAT scores and heights of US adult males closely follow the normal distribution.

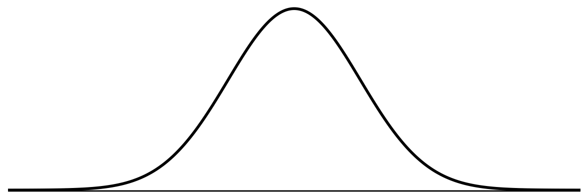


Figure 4.1: A normal curve.

NORMAL DISTRIBUTION FACTS

Many variables are nearly normal, but none are exactly normal. Thus the normal distribution, while not perfect for any single problem, is very useful for a variety of problems. We will use it in data exploration and to solve important problems in statistics.

4.1.1 Normal distribution model

The **normal distribution** always describes a symmetric, unimodal, bell-shaped curve. However, these curves can look different depending on the details of the model. Specifically, the normal distribution model can be adjusted using two parameters: mean and standard deviation. As you can probably guess, changing the mean shifts the bell curve to the left or right, while changing the standard deviation stretches or constricts the curve. Figure 4.2 shows the normal distribution with mean 0 and standard deviation 1 in the left panel and the normal distributions with mean 19 and standard deviation 4 in the right panel. Figure 4.3 shows these distributions on the same axis.

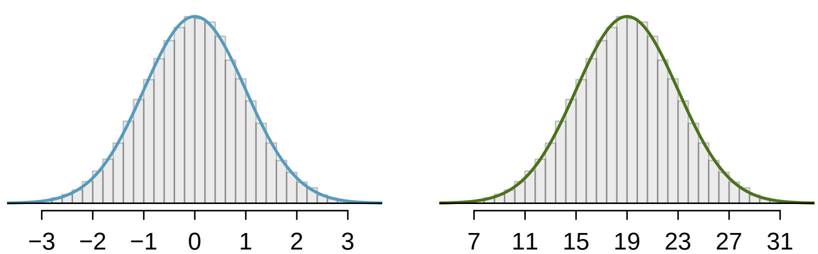


Figure 4.2: Both curves represent the normal distribution. However, they differ in their center and spread.

If a normal distribution has mean μ and standard deviation σ, we may write the distribution as $N(\mu, \sigma)$. The two distributions in Figure 4.3 may be written as

$$N(\mu = 0, \sigma = 1) \quad \text{and} \quad N(\mu = 19, \sigma = 4)$$

Because the mean and standard deviation describe a normal distribution exactly, they are called the distribution's **parameters**. The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the **standard normal distribution**.

[1]It is also introduced as the Gaussian distribution after Frederic Gauss, the first person to formalize its mathematical expression.

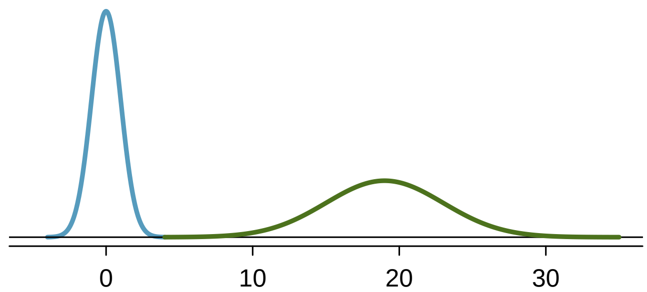


Figure 4.3: The normal distributions shown in Figure 4.2 but plotted together and on the same scale.

G

GUIDED PRACTICE 4.1

Write down the short-hand for a normal distribution with[2] (a) mean 5 and standard deviation 3,
(b) mean -100 and standard deviation 10, and
(c) mean 2 and standard deviation 9.

4.1.2 Standardizing with Z-scores

We often want to put data onto a standardized scale, which can make comparisons more reasonable.

E

EXAMPLE 4.2

Table 4.4 shows the mean and standard deviation for total scores on the SAT and ACT. The distribution of SAT and ACT scores are both nearly normal. Suppose Ann scored 1300 on her SAT and Tom scored 24 on his ACT. Who performed better?

We use the standard deviation as a guide. Ann is 1 standard deviation above average on the SAT: $1100 + 200 = 1300$. Tom is 0.5 standard deviations above the mean on the ACT: $21 + 0.5 \times 6 = 24$. In Figure 4.5, we can see that Ann tends to do better with respect to everyone else than Tom did, so her score was better.

	SAT	ACT
Mean	1100	21
SD	200	6

Figure 4.4: Mean and standard deviation for the SAT and ACT.

Example 4.2 used a standardization technique called a Z-score, a method most commonly employed for nearly normal observations but that may be used with any distribution. The **Z-score** of an observation is defined as the number of standard deviations it falls above or below the mean. If the observation is one standard deviation above the mean, its Z-score is 1. If it is 1.5 standard deviations *below* the mean, then its Z-score is -1.5. If x is an observation from a distribution $N(\mu, \sigma)$, we define the Z-score mathematically as

$$Z = \frac{x - \mu}{\sigma}$$

Using $\mu_{SAT} = 1100$, $\sigma_{SAT} = 200$, and $x_{\text{Ann}} = 1300$, we find Ann's Z-score:

$$Z_{\text{Ann}} = \frac{x_{\text{Ann}} - \mu_{\text{SAT}}}{\sigma_{\text{SAT}}} = \frac{1300 - 1100}{200} = 1$$

[2](a) $N(\mu = 5, \sigma = 3)$. (b) $N(\mu = -100, \sigma = 10)$. (c) $N(\mu = 2, \sigma = 9)$.

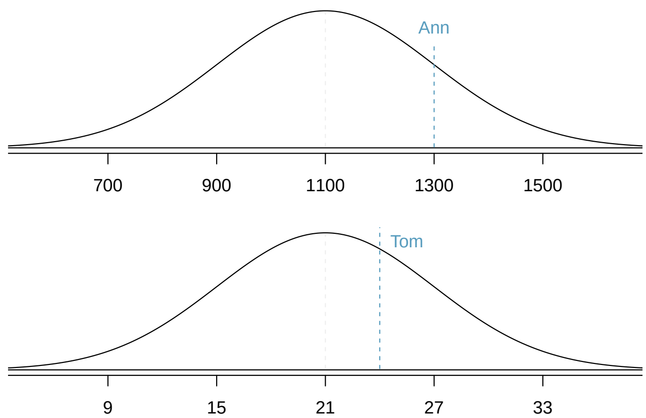


Figure 4.5: Ann's and Tom's scores shown with the distributions of SAT and ACT scores.

THE Z-SCORE

The Z-score of an observation is the number of standard deviations it falls above or below the mean. We compute the Z-score for an observation x that follows a distribution with mean μ and standard deviation σ using

$$Z = \frac{x - \mu}{\sigma}$$

GUIDED PRACTICE 4.3

Use Tom's ACT score, 24, along with the ACT mean and standard deviation to compute his Z-score.[3]

Observations above the mean always have positive Z-scores, while those below the mean always have negative Z-scores. If an observation is equal to the mean, such as an SAT score of 1100, then the Z-score is 0.

GUIDED PRACTICE 4.4

Let X represent a random variable from $N(\mu = 3, \sigma = 2)$, and suppose we observe $x = 5.19$.
(a) Find the Z-score of x.
(b) Use the Z-score to determine how many standard deviations above or below the mean x falls.[4]

GUIDED PRACTICE 4.5

Head lengths of brushtail possums follow a nearly normal distribution with mean 92.6 mm and standard deviation 3.6 mm. Compute the Z-scores for possums with head lengths of 95.4 mm and 85.8 mm.[5]

We can use Z-scores to roughly identify which observations are more unusual than others. An observation x_1 is said to be more unusual than another observation x_2 if the absolute value of its Z-score is larger than the absolute value of the other observation's Z-score: $|Z_1| > |Z_2|$. This technique is especially insightful when a distribution is symmetric.

GUIDED PRACTICE 4.6

Which of the observations in Guided Practice 4.5 is more unusual?[6]

[3] $Z_{Tom} = \frac{x_{\text{Tom}} - \mu_{\text{ACT}}}{\sigma_{\text{ACT}}} = \frac{24-21}{6} = 0.5$

[4] (a) Its Z-score is given by $Z = \frac{x-\mu}{\sigma} = \frac{5.19-3}{2} = 2.19/2 = 1.095$. (b) The observation x is 1.095 standard deviations *above* the mean. We know it must be above the mean since Z is positive.

[5] For $x_1 = 95.4$ mm: $Z_1 = \frac{x_1-\mu}{\sigma} = \frac{95.4-92.6}{3.6} = 0.78$. For $x_2 = 85.8$ mm: $Z_2 = \frac{85.8-92.6}{3.6} = -1.89$.

4.1.3 Finding tail areas

It's very useful in statistics to be able to identify tail areas of distributions. For instance, how many people have an SAT score below Ann's score of 1300? This is the same as Ann's **percentile**, which is the fraction of cases that have lower scores than Ann. We can visualize such a tail area like the curve and shading shown in Figure 4.6.

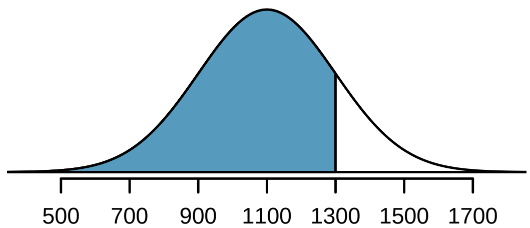


Figure 4.6: The area to the left of Z represents the percentile of the observation.

There are many techniques for doing this, and we'll discuss three of the options.

1. The most common approach in practice is to use statistical software. For example, in the program **R**, we could find the area shown in Figure 4.6 using the following command, which takes in the Z-score and returns the lower tail area:
   ```
   > pnorm(1)
   [1] 0.8413447
   ```
   According to this calculation, the region shaded that is below 1300 represents the proportion 0.841 (84.1%) of SAT test takers who had Z-scores below $Z = 1$. More generally, we can also specify the cutoff explicitly if we also note the mean and standard deviation:
   ```
   > pnorm(1300, mean = 1100, sd = 200)
   [1] 0.8413447
   ```
   There are many other software options, such as Python or SAS; even spreadsheet programs such as Excel and Google Sheets support these calculations.

2. A common strategy in classrooms is to use a graphing calculator, such as a TI or Casio calculator. These calculators require a series of button presses that are less concisely described. You can find instructions on using these calculators for finding tail areas of a normal distribution in the OpenIntro video library:

   www.openintro.org/videos

3. The last option for finding tail areas is to use what's called a **probability table**; these are occasionally used in classrooms but rarely in practice. Appendix B.1 contains such a table and a guide for how to use it.

We will solve normal distribution problems in this section by always first finding the Z-score. The reason is that we will encounter close parallels called test statistics beginning in Chapter 5; these are, in many instances, an equivalent of a Z-score.

4.1.4 Normal probability examples

Cumulative SAT scores are approximated well by a normal model, $N(\mu = 1100, \sigma = 200)$.

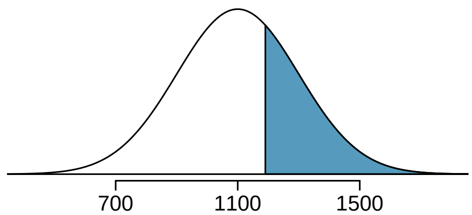
[6] Because the *absolute value* of Z-score for the second observation is larger than that of the first, the second observation has a more unusual head length.

E

EXAMPLE 4.7

Shannon is a randomly selected SAT taker, and nothing is known about Shannon's SAT aptitude. What is the probability Shannon scores at least 1190 on her SATs?

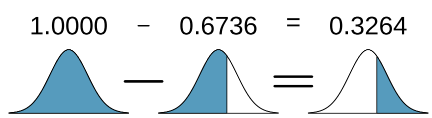
First, always draw and label a picture of the normal distribution. (Drawings need not be exact to be useful.) We are interested in the chance she scores above 1190, so we shade this upper tail:



The picture shows the mean and the values at 2 standard deviations above and below the mean. The simplest way to find the shaded area under the curve makes use of the Z-score of the cutoff value. With $\mu = 1100$, $\sigma = 200$, and the cutoff value $x = 1190$, the Z-score is computed as

$$Z = \frac{x - \mu}{\sigma} = \frac{1190 - 1100}{200} = \frac{90}{200} = 0.45$$

Using statistical software (or another preferred method), we can area left of $Z = 0.45$ as 0.6736. To find the area *above* $Z = 0.45$, we compute one minus the area of the lower tail:



The probability Shannon scores at least 1190 on the SAT is 0.3264.

ALWAYS DRAW A PICTURE FIRST, AND FIND THE Z-SCORE SECOND

For any normal probability situation, *always always always* draw and label the normal curve and shade the area of interest first. The picture will provide an estimate of the probability. After drawing a figure to represent the situation, identify the Z-score for the value of interest.

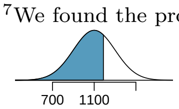
G

GUIDED PRACTICE 4.8

If the probability of Shannon scoring at least 1190 is 0.3264, then what is the probability she scores less than 1190? Draw the normal curve representing this exercise, shading the lower region instead of the upper one.[7]

---

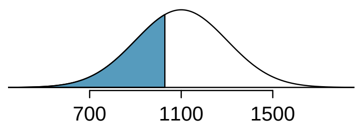
[7]We found the probability in Example 4.7: 0.6736.



EXAMPLE 4.9

Edward earned a 1030 on his SAT. What is his percentile?

First, a picture is needed. Edward's percentile is the proportion of people who do not get as high as a 1030. These are the scores to the left of 1030.



Identifying the mean $\mu = 1100$, the standard deviation $\sigma = 200$, and the cutoff for the tail area $x = 1030$ makes it easy to compute the Z-score:

$$Z = \frac{x - \mu}{\sigma} = \frac{1030 - 1100}{200} = -0.35$$

Using statistical software, we get a tail area of 0.3632. Edward is at the 36^{th} percentile.

GUIDED PRACTICE 4.10

Use the results of Example 4.9 to compute the proportion of SAT takers who did better than Edward. Also draw a new picture.[8]

FINDING AREAS TO THE RIGHT

Many software programs return the area to the left when given a Z-score. If you would like the area to the right, first find the area to the left and then subtract this amount from one.

GUIDED PRACTICE 4.11

Stuart earned an SAT score of 1500. Draw a picture for each part.
(a) What is his percentile?
(b) What percent of SAT takers did better than Stuart?[9]

Based on a sample of 100 men, the heights of male adults in the US is nearly normal with mean 70.0" and standard deviation 3.3".

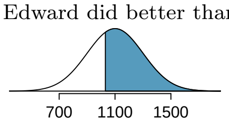
GUIDED PRACTICE 4.12

Mike is 5'7" and Jose is 6'4", and they both live in the US.
(a) What is Mike's height percentile?
(b) What is Jose's height percentile?
Also draw one picture for each part.[10]

The last several problems have focused on finding the percentile (or upper tail) for a particular observation. What if you would like to know the observation corresponding to a particular percentile?

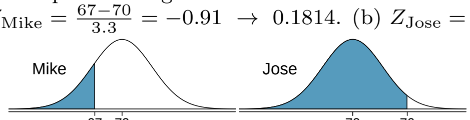
---

[8]If Edward did better than 36% of SAT takers, then about 64% must have done better than him.



[9]We leave the drawings to you. (a) $Z = \frac{1500-1100}{200} = 2 \rightarrow 0.9772$. (b) $1 - 0.9772 = 0.0228$.

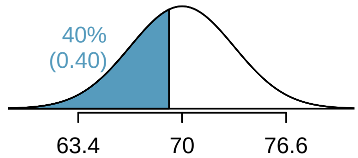
[10]First put the heights into inches: 67 and 76 inches. Figures are shown below.
(a) $Z_{\text{Mike}} = \frac{67-70}{3.3} = -0.91 \ \rightarrow\ 0.1814$. (b) $Z_{\text{Jose}} = \frac{76-70}{3.3} = 1.82 \ \rightarrow\ 0.9656$.



EXAMPLE 4.13

Erik's height is at the 40^{th} percentile. How tall is he?

As always, first draw the picture.



E

In this case, the lower tail probability is known (0.40), which can be shaded on the diagram. We want to find the observation that corresponds to this value. As a first step in this direction, we determine the Z-score associated with the 40^{th} percentile. Using software, we can obtain the corresponding Z-score of about -0.25.

Knowing $Z_{Erik} = -0.25$ and the population parameters $\mu = 70$ and $\sigma = 3.3$ inches, the Z-score formula can be set up to determine Erik's unknown height, labeled x_{Erik}:

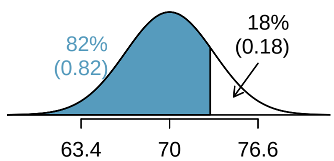
$$-0.25 = Z_{\text{Erik}} = \frac{x_{\text{Erik}} - \mu}{\sigma} = \frac{x_{\text{Erik}} - 70}{3.3}$$

Solving for x_{Erik} yields a height of 69.18 inches. That is, Erik is about 5'9".

EXAMPLE 4.14

What is the adult male height at the 82^{nd} percentile?

Again, we draw the figure first.



E

Next, we want to find the Z-score at the 82^{nd} percentile, which will be a positive value and can be found using software as $Z = 0.92$. Finally, the height x is found using the Z-score formula with the known mean μ, standard deviation σ, and Z-score $Z = 0.92$:

$$0.92 = Z = \frac{x - \mu}{\sigma} = \frac{x - 70}{3.3}$$

This yields 73.04 inches or about 6'1" as the height at the 82^{nd} percentile.

GUIDED PRACTICE 4.15

G

The SAT scores follow $N(1100, 200)$.[11]
(a) What is the 95^{th} percentile for SAT scores?
(b) What is the 97.5^{th} percentile for SAT scores?

GUIDED PRACTICE 4.16

G

Adult male heights follow $N(70.0\text{''}, 3.3\text{''})$.[12]
(a) What is the probability that a randomly selected male adult is at least 6'2" (74 inches)?
(b) What is the probability that a male adult is shorter than 5'9" (69 inches)?

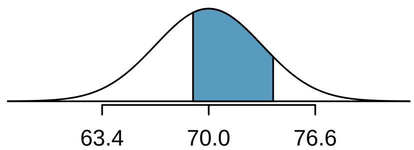
---

[11]Short answers: (a) $Z_{95} = 1.65 \to 1430$ SAT score. (b) $Z_{97.5} = 1.96 \to 1492$ SAT score.
[12]Short answers: (a) $Z = 1.21 \to 0.8869$, then subtract this value from 1 to get 0.1131. (b) $Z = -0.30 \to 0.3821$.

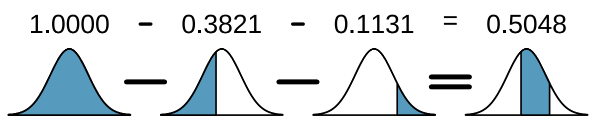
EXAMPLE 4.17

What is the probability that a random adult male is between 5'9" and 6'2"?

These heights correspond to 69 inches and 74 inches. First, draw the figure. The area of interest is no longer an upper or lower tail.



The total area under the curve is 1. If we find the area of the two tails that are not shaded (from Guided Practice 4.16, these areas are 0.3821 and 0.1131), then we can find the middle area:



That is, the probability of being between 5'9" and 6'2" is 0.5048.

GUIDED PRACTICE 4.18

SAT scores follow $N(1100, 200)$. What percent of SAT takers get between 1100 and 1400?[13]

GUIDED PRACTICE 4.19

Adult male heights follow $N(70.0", 3.3")$. What percent of adult males are between 5'5" and 5'7"?[14]

4.1.5 68-95-99.7 rule

Here, we present a useful rule of thumb for the probability of falling within 1, 2, and 3 standard deviations of the mean in the normal distribution. This will be useful in a wide range of practical settings, especially when trying to make a quick estimate without a calculator or Z-table.

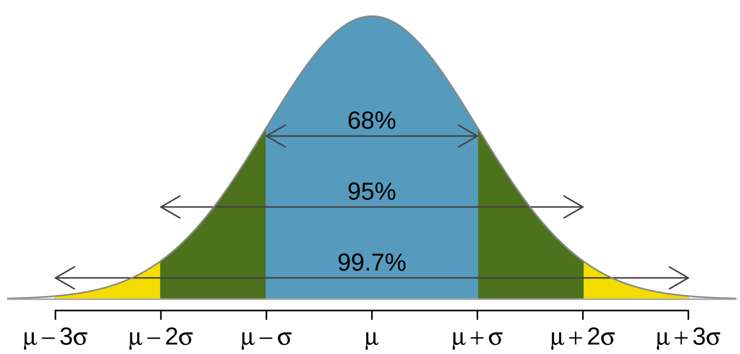


Figure 4.7: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

---

[13] This is an abbreviated solution. (Be sure to draw a figure!) First find the percent who get below 1100 and the percent that get above 1400: $Z_{1100} = 0.00 \rightarrow 0.5000$ (area below), $Z_{1400} = 1.5 \rightarrow 0.0668$ (area above). Final answer: $1.0000 - 0.5000 - 0.0668 = 0.4332$.

[14] 5'5" is 65 inches ($Z = -1.52$). 5'7" is 67 inches ($Z = -0.91$). Numerical solution: $1.000 - 0.0643 - 0.8186 = 0.1171$, i.e. 11.71%.

G

GUIDED PRACTICE 4.20

Use software, a calculator, or a probability table to confirm that about 68%, 95%, and 99.7% of observations fall within 1, 2, and 3, standard deviations of the mean in the normal distribution, respectively. For instance, first find the area that falls between $Z = -1$ and $Z = 1$, which should have an area of about 0.68. Similarly there should be an area of about 0.95 between $Z = -2$ and $Z = 2$.[15]

It is possible for a normal random variable to fall 4, 5, or even more standard deviations from the mean. However, these occurrences are very rare if the data are nearly normal. The probability of being further than 4 standard deviations from the mean is about 1-in-15,000. For 5 and 6 standard deviations, it is about 1-in-2 million and 1-in-500 million, respectively.

G

GUIDED PRACTICE 4.21

SAT scores closely follow the normal model with mean $\mu = 1100$ and standard deviation $\sigma = 200$.[16]
(a) About what percent of test takers score 700 to 1500?
(b) What percent score between 1100 and 1500?

[15] First draw the pictures. Using software, we get 0.6827 within 1 standard deviation, 0.9545 within 2 standard deviations, and 0.9973 within 3 standard deviations.

[16] (a) 700 and 1500 represent two standard deviations below and above the mean, which means about 95% of test takers will score between 700 and 1500. (b) We found that 700 to 1500 represents about 95% of test takers. These test takers would be evenly split by the center of the distribution, 1100, so $\frac{95\%}{2} = 47.5\%$ of all test takers score between 1100 and 1500.

4.2 Geometric distribution

How long should we expect to flip a coin until it turns up `heads`? Or how many times should we expect to roll a die until we get a `1`? These questions can be answered using the geometric distribution. We first formalize each trial – such as a single coin flip or die toss – using the Bernoulli distribution, and then we combine these with our tools from probability (Chapter **??**) to construct the geometric distribution.

4.2.1 Bernoulli distribution

Many health insurance plans in the United States have a deductible, where the insured individual is responsible for costs up to the deductible, and then the costs above the deductible are shared between the individual and insurance company for the remainder of the year.

Suppose a health insurance company found that 70% of the people they insure stay below their deductible in any given year. Each of these people can be thought of as a **trial**. We label a person a **success** if her healthcare costs do not exceed the deductible. We label a person a **failure** if she does exceed her deductible in the year. Because 80% of the individuals will not hit their deductible, we denote the **probability of a success** as $p = 0.7$. The probability of a failure is sometimes denoted with $q = 1 - p$, which would be 0.3 in for the insurance example.

When an individual trial only has two possible outcomes, often labeled as `success` or `failure`, it is called a **Bernoulli random variable**. We chose to label a person who does not hit her deductible as a "success" and all others as "failures". However, we could just as easily have reversed these labels. The mathematical framework we will build does not depend on which outcome is labeled a success and which a failure, as long as we are consistent.

Bernoulli random variables are often denoted as `1` for a success and `0` for a failure. In addition to being convenient in entering data, it is also mathematically handy. Suppose we observe ten trials:

`1 1 1 0 1 0 0 1 1 0`

Then the **sample proportion**, $\hat{p}$, is the sample mean of these observations:

$$\hat{p} = \frac{\text{\# of successes}}{\text{\# of trials}} = \frac{1+1+1+0+1+0+0+1+1+0}{10} = 0.6$$

This mathematical inquiry of Bernoulli random variables can be extended even further. Because `0` and `1` are numerical outcomes, we can define the mean and standard deviation of a Bernoulli random variable.[17]

BERNOULLI RANDOM VARIABLE

If X is a random variable that takes value 1 with probability of success p and 0 with probability $1 - p$, then X is a Bernoulli random variable with mean and standard deviation

$$\mu = p \qquad\qquad \sigma = \sqrt{p(1-p)}$$

In general, it is useful to think about a Bernoulli random variable as a random process with only two outcomes: a success or failure. Then we build our mathematical framework using the numerical labels `1` and `0` for successes and failures, respectively.

[17] If p is the true probability of a success, then the mean of a Bernoulli random variable X is given by

$$\begin{aligned}\mu = E[X] &= P(X=0)\times 0 + P(X=1)\times 1\\ &= (1-p)\times 0 + p\times 1 = 0 + p = p\end{aligned}$$

Similarly, the variance of X can be computed:

$$\begin{aligned}\sigma^2 &= P(X=0)(0-p)^2 + P(X=1)(1-p)^2\\ &= (1-p)p^2 + p(1-p)^2 = p(1-p)\end{aligned}$$

The standard deviation is $\sigma = \sqrt{p(1-p)}$.

4.2.2 Geometric distribution

The **geometric distribution** is used to describe how many trials it takes to observe a success. Let's first look at an example.

EXAMPLE 4.22

Suppose we are working at the insurance company and need to find a case where the person did not exceed her (or his) deductible as a case study. If the probability a person will not exceed her deductible is 0.7 and we are drawing people at random, what are the chances that the first person will not have exceeded her deductible, i.e. be a success? The second person? The third? What about we pull $n-1$ cases before we find the first success, i.e. the first success is the n^{th} person? (If the first success is the fifth person, then we say $n = 5$.)

E

The probability of stopping after the first person is just the chance the first person will not hit her (or his) deductible: 0.7. The probability it will be the second person is

$$\begin{aligned} &P(\text{second person is the first to not administer the worst shock}) \\ &= P(\text{the first will, the second won't}) = (0.3)(0.7) = 0.21 \end{aligned}$$

Likewise, the probability it will be the third case is $(0.3)(0.3)(0.7) = 0.063$.

If the first success is on the n^{th} person, then there are $n - 1$ failures and finally 1 success, which corresponds to the probability $(0.3)^{n-1}(0.7)$. This is the same as $(1 - 0.7)^{n-1}(0.7)$.

Example 4.22 illustrates what the **geometric distribution**, which describes the waiting time until a success for **independent and identically distributed (iid)** Bernoulli random variables. In this case, the *independence* aspect just means the individuals in the example don't affect each other, and *identical* means they each have the same probability of success.

The geometric distribution from Example 4.22 is shown in Figure 4.8. In general, the probabilities for a geometric distribution decrease **exponentially** fast.

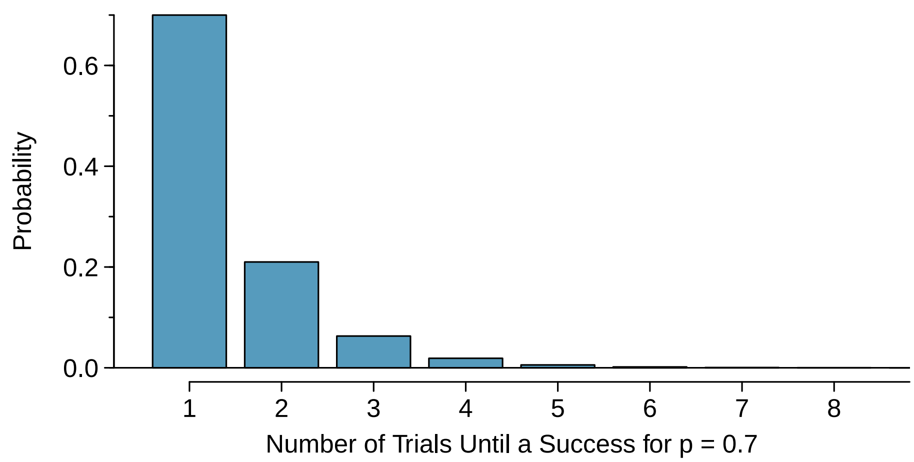


Figure 4.8: The geometric distribution when the probability of success is $p = 0.7$.

While this text will not derive the formulas for the mean (expected) number of trials needed to find the first success or the standard deviation or variance of this distribution, we present general formulas for each.

GEOMETRIC DISTRIBUTION

If the probability of a success in one trial is p and the probability of a failure is $1 - p$, then the probability of finding the first success in the n^{th} trial is given by

$$(1-p)^{n-1}p$$

The mean (i.e. expected value), variance, and standard deviation of this wait time are given by

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

It is no accident that we use the symbol μ for both the mean and expected value. The mean and the expected value are one and the same.

It takes, on average, $1/p$ trials to get a success under the geometric distribution. This mathematical result is consistent with what we would expect intuitively. If the probability of a success is high (e.g. 0.8), then we don't usually wait very long for a success: $1/0.8 = 1.25$ trials on average. If the probability of a success is low (e.g. 0.1), then we would expect to view many trials before we see a success: $1/0.1 = 10$ trials.

G

GUIDED PRACTICE 4.23

The probability that a particular case would not exceed their deductible is said to be 0.7. If we were to examine cases until we found one that where the person did not hit her deductible, how many cases should we expect to check?[18]

E

EXAMPLE 4.24

What is the chance that we would find the first success within the first 3 cases?

This is the chance it is the first ($n = 1$), second ($n = 2$), or third ($n = 3$) case is the first success, which are three disjoint outcomes. Because the individuals in the sample are randomly sampled from a large population, they are independent. We compute the probability of each case and add the separate results:

$$\begin{aligned} &P(n = 1, 2, \text{ or } 3) \\ &\quad = P(n = 1) + P(n = 2) + P(n = 3) \\ &\quad = (0.3)^{1-1}(0.7) + (0.3)^{2-1}(0.7) + (0.3)^{3-1}(0.7) \\ &\quad = 0.973 \end{aligned}$$

There is a probability of 0.973 that we would find a successful case within 3 cases.

G

GUIDED PRACTICE 4.25

Determine a more clever way to solve Example 4.24. Show that you get the same result.[19]

---

[18]We would expect to see about $1/0.7 \approx 1.43$ individuals to find the first success.

[19]First find the probability of the complement: $P(\text{no success in first 3 trials}) = 0.3^3 = 0.027$. Next, compute one minus this probability: $1 - P(\text{no success in 3 trials}) = 1 - 0.027 = 0.973$.

EXAMPLE 4.26

Suppose a car insurer has determined that 88% of its drivers will not exceed their deductible in a given year. If someone at the company were to randomly draw driver files until they found one that had not exceeded their deductible, what is the expected number of drivers the insurance employee must check? What is the standard deviation of the number of driver files that must be drawn?

In this example, a success is again when someone will not exceed the insurance deductible, which has probability $p = 0.88$. The expected number of people to be checked is $1/p = 1/0.88 = 1.14$ and the standard deviation is $\sqrt{(1-p)/p^2} = 0.39$.

GUIDED PRACTICE 4.27

Using the results from Example 4.26, $\mu = 1.14$ and $\sigma = 0.39$, would it be appropriate to use the normal model to find what proportion of experiments would end in 3 or fewer trials?[20]

The independence assumption is crucial to the geometric distribution's accurate description of a scenario. Mathematically, we can see that to construct the probability of the success on the n^{th} trial, we had to use the Multiplication Rule for Independent Processes. It is no simple task to generalize the geometric model for dependent trials.

[20]No. The geometric distribution is always right skewed and can never be well-approximated by the normal model.

4.3 Binomial distribution

The **binomial distribution** is used to describe the number of successes in a fixed number of trials. This is different from the geometric distribution, which described the number of trials we must wait before we observe a success.

4.3.1 The binomial distribution

Let's again imagine ourselves back at the insurance agency where 70% of individuals do not exceed their deductible.

EXAMPLE 4.28

Suppose the insurance agency is considering a random sample of four individuals they insure. What is the chance exactly one of them will exceed the deductible and the other four will not? Let's call the four people Ariana (A), Brittany (B), Carlton (C), and Damian (D) for convenience.

E

Let's consider a scenario where one person exceeds the deductible:

$$\begin{aligned}
&P(A = \texttt{exceed},\ B = \texttt{not},\ C = \texttt{not},\ D = \texttt{not}) \\
&\quad = P(A = \texttt{exceed})\ P(B = \texttt{not})\ P(C = \texttt{not})\ P(D = \texttt{not}) \\
&\quad = (0.3)(0.7)(0.7)(0.7) \\
&\quad = (0.7)^3(0.3)^1 \\
&\quad = 0.103
\end{aligned}$$

But there are three other scenarios: Brittany, Carlton, or Damian could have been the one to exceed the deductible. In each of these cases, the probability is again $(0.7)^3(0.3)^1$. These four scenarios exhaust all the possible ways that exactly one of these four people could have exceeded the deductible, so the total probability is $4 \times (0.7)^3(0.3)^1 = 0.412$.

GUIDED PRACTICE 4.29

G

Verify that the scenario where Brittany is the only one exceed the deductible has probability $(0.7)^3(0.3)^1$. [21]

The scenario outlined in Example 4.28 is an example of a binomial distribution scenario. The **binomial distribution** describes the probability of having exactly k successes in n independent Bernoulli trials with probability of a success p (in Example 4.28, $n = 4$, $k = 3$, $p = 0.7$). We would like to determine the probabilities associated with the binomial distribution more generally, i.e. we want a formula where we can use n, k, and p to obtain the probability. To do this, we reexamine each part of Example 4.28.

There were four individuals who could have been the one to exceed the deductible, and each of these four scenarios had the same probability. Thus, we could identify the final probability as

$$[\#\text{ of scenarios}] \times P(\text{single scenario})$$

The first component of this equation is the number of ways to arrange the $k = 3$ successes among the $n = 4$ trials. The second component is the probability of any of the four (equally probable) scenarios.

Consider $P(\text{single scenario})$ under the general case of k successes and $n - k$ failures in the n trials. In any such scenario, we apply the Multiplication Rule for independent events:

$$p^k(1-p)^{n-k}$$

This is our general formula for $P(\text{single scenario})$.

[21] $P(A = \texttt{not},\ B = \texttt{exceed},\ C = \texttt{not},\ D = \texttt{not}) = (0.7)(0.3)(0.7)(0.7) = (0.7)^3(0.3)^1$.

Secondly, we introduce a general formula for the number of ways to choose k successes in n trials, i.e. arrange k successes and $n-k$ failures:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The quantity $\binom{n}{k}$ is read **n choose k**.[22] The exclamation point notation (e.g. $k!$) denotes a **factorial** expression.

$$\begin{aligned}
0! &= 1 \\
1! &= 1 \\
2! &= 2 \times 1 = 2 \\
3! &= 3 \times 2 \times 1 = 6 \\
4! &= 4 \times 3 \times 2 \times 1 = 24 \\
&\vdots \\
n! &= n \times (n-1) \times ... \times 3 \times 2 \times 1
\end{aligned}$$

Using the formula, we can compute the number of ways to choose $k = 3$ successes in $n = 4$ trials:

$$\binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4!}{3!1!} = \frac{4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(1)} = 4$$

This result is exactly what we found by carefully thinking of each possible scenario in Example 4.28.

Substituting n choose k for the number of scenarios and $p^k(1-p)^{n-k}$ for the single scenario probability yields the general binomial formula.

BINOMIAL DISTRIBUTION

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

The mean, variance, and standard deviation of the number of observed successes are

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \sigma = \sqrt{np(1-p)}$$

IS IT BINOMIAL? FOUR CONDITIONS TO CHECK.

(1) The trials are independent.
(2) The number of trials, n, is fixed.
(3) Each trial outcome can be classified as a *success* or *failure*.
(4) The probability of a success, p, is the same for each trial.

---

[22] Other notation for n choose k includes ${}_nC_k$, C_n^k, and $C(n,k)$.

EXAMPLE 4.30

What is the probability that 3 of 8 randomly selected individuals will have exceeded the insurance deductible, i.e. that 5 of 8 will not exceed the deductible? Recall that 70% of individuals will not exceed the deductible.

We would like to apply the binomial model, so we check the conditions. The number of trials is fixed ($n = 8$) (condition 2) and each trial outcome can be classified as a success or failure (condition 3). Because the sample is random, the trials are independent (condition 1) and the probability of a success is the same for each trial (condition 4).

In the outcome of interest, there are $k = 5$ successes in $n = 8$ trials (recall that a success is an individual who does *not* exceed the deductible, and the probability of a success is $p = 0.7$. So the probability that 5 of 8 will not exceed the deductible and 3 will exceed the deductible is given by

$$\binom{8}{5}(0.7)^5(1-0.7)^{8-5} = \frac{8!}{5!(5-3)!}(0.7)^5(1-0.7)^{8-5}$$
$$= \frac{8!}{5!3!}(0.7)^5(0.3)^3$$

Dealing with the factorial part:

$$\frac{8!}{5!3!} = \frac{8\times 7\times 6\times 5\times 4\times 3\times 2\times 1}{(5\times 4\times 3\times 2\times 1)(3\times 2\times 1)} = \frac{8\times 7\times 6}{3\times 2\times 1} = 56$$

Using $(0.7)^5(0.3)^3 \approx 0.00454$, the final probability is about $56 * 0.00454 \approx 0.254$.

COMPUTING BINOMIAL PROBABILITIES

The first step in using the binomial model is to check that the model is appropriate. The second step is to identify n, p, and k. As the last stage use software or the formulas to determine the probability, then interpret the results.

If you must do calculations by hand, it's often useful to cancel out as many terms as possible in the top and bottom of the binomial coefficient.

GUIDED PRACTICE 4.31

If we randomly sampled 40 case files from the insurance agency discussed earlier, how many of the cases would you expect to not have exceeded the deductible in a given year? What is the standard deviation of the number that would not have exceeded the deductible?[23]

GUIDED PRACTICE 4.32

The probability that a random smoker will develop a severe lung condition in his or her lifetime is about 0.3. If you have 4 friends who smoke, are the conditions for the binomial model satisfied?[24]

---

[23] We are asked to determine the expected number (the mean) and the standard deviation, both of which can be directly computed from the formulas: $\mu = np = 40 \times 0.7 = 28$ and $\sigma = \sqrt{np(1-p)} = \sqrt{40 \times 0.7 \times 0.3} = 2.9$. Because very roughly 95% of observations fall within 2 standard deviations of the mean (see Section **??**), we would probably observe at least 22 but fewer than 34 individuals in our sample who would not exceed the deductible.

[24] One possible answer: if the friends know each other, then the independence assumption is probably not satisfied. For example, acquaintances may have similar smoking habits, or those friends might make a pact to quit together.

GUIDED PRACTICE 4.33

Suppose these four friends do not know each other and we can treat them as if they were a random sample from the population. Is the binomial model appropriate? What is the probability that[25]

(a) None of them will develop a severe lung condition?

(b) One will develop a severe lung condition?

(c) That no more than one will develop a severe lung condition?

GUIDED PRACTICE 4.34

What is the probability that at least 2 of your 4 smoking friends will develop a severe lung condition in their lifetimes?[26]

GUIDED PRACTICE 4.35

Suppose you have 7 friends who are smokers and they can be treated as a random sample of smokers.[27]

(a) How many would you expect to develop a severe lung condition, i.e. what is the mean?

(b) What is the probability that at most 2 of your 7 friends will develop a severe lung condition.

Next we consider the first term in the binomial probability, n choose k under some special scenarios.

GUIDED PRACTICE 4.36

Why is it true that $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$ for any number n?[28]

GUIDED PRACTICE 4.37

How many ways can you arrange one success and $n-1$ failures in n trials? How many ways can you arrange $n-1$ successes and one failure in n trials?[29]

---

[25] To check if the binomial model is appropriate, we must verify the conditions. (i) Since we are supposing we can treat the friends as a random sample, they are independent. (ii) We have a fixed number of trials ($n = 4$). (iii) Each outcome is a success or failure. (iv) The probability of a success is the same for each trials since the individuals are like a random sample ($p = 0.3$ if we say a "success" is someone getting a lung condition, a morbid choice). Compute parts (a) and (b) using the binomial formula: $P(0) = \binom{4}{0}(0.3)^0(0.7)^4 = 1 \times 1 \times 0.7^4 = 0.2401$, $P(1) = \binom{4}{1}(0.3)^1(0.7)^3 = 0.4116$. Note: $0! = 1$. Part (c) can be computed as the sum of parts (a) and (b): $P(0) + P(1) = 0.2401 + 0.4116 = 0.6517$. That is, there is about a 65% chance that no more than one of your four smoking friends will develop a severe lung condition.

[26] The complement (no more than one will develop a severe lung condition) as computed in Guided Practice 4.33 as 0.6517, so we compute one minus this value: 0.3483.

[27] (a) $\mu = 0.3 \times 7 = 2.1$. (b) $P(0, 1, \text{ or } 2 \text{ develop severe lung condition}) = P(k = 0) + P(k = 1) + P(k = 2) = 0.6471$.

[28] Frame these expressions into words. How many different ways are there to arrange 0 successes and n failures in n trials? (1 way.) How many different ways are there to arrange n successes and 0 failures in n trials? (1 way.)

[29] One success and $n-1$ failures: there are exactly n unique places we can put the success, so there are n ways to arrange one success and $n-1$ failures. A similar argument is used for the second question. Mathematically, we show these results by verifying the following two equations:

$$\binom{n}{1} = n, \qquad \binom{n}{n-1} = n$$

4.3.2 Normal approximation to the binomial distribution

The binomial formula is cumbersome when the sample size (n) is large, particularly when we consider a range of observations. In some cases we may use the normal distribution as an easier and faster way to estimate binomial probabilities.

EXAMPLE 4.38

Approximately 15% of the US population smokes cigarettes. A local government believed their community had a lower smoker rate and commissioned a survey of 400 randomly selected individuals. The survey found that only 42 of the 400 participants smoke cigarettes. If the true proportion of smokers in the community was really 15%, what is the probability of observing 42 or fewer smokers in a sample of 400 people?

We leave the usual verification that the four conditions for the binomial model are valid as an exercise.

E

The question posed is equivalent to asking, what is the probability of observing $k = 0$, 1, 2, ..., or 42 smokers in a sample of $n = 400$ when $p = 0.15$? We can compute these 43 different probabilities and add them together to find the answer:

$$\begin{aligned} &P(k = 0 \text{ or } k = 1 \text{ or } \cdots \text{ or } k = 42) \\ &\quad = P(k = 0) + P(k = 1) + \cdots + P(k = 42) \\ &\quad = 0.0054 \end{aligned}$$

If the true proportion of smokers in the community is $p = 0.15$, then the probability of observing 42 or fewer smokers in a sample of $n = 400$ is 0.0054.

The computations in Example 4.38 are tedious and long. In general, we should avoid such work if an alternative method exists that is faster, easier, and still accurate. Recall that calculating probabilities of a range of values is much easier in the normal model. We might wonder, is it reasonable to use the normal model in place of the binomial distribution? Surprisingly, yes, if certain conditions are met.

GUIDED PRACTICE 4.39

G

Here we consider the binomial model when the probability of a success is $p = 0.10$. Figure 4.9 shows four hollow histograms for simulated samples from the binomial distribution using four different sample sizes: $n = 10, 30, 100, 300$. What happens to the shape of the distributions as the sample size increases? What distribution does the last hollow histogram resemble?[30]

NORMAL APPROXIMATION OF THE BINOMIAL DISTRIBUTION

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1 - p)$ are both at least 10. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \qquad\qquad \sigma = \sqrt{np(1-p)}$$

The normal approximation may be used when computing the range of many possible successes. For instance, we may apply the normal distribution to the setting of Example 4.38.

[30] The distribution is transformed from a blocky and skewed distribution into one that rather resembles the normal distribution in last hollow histogram.

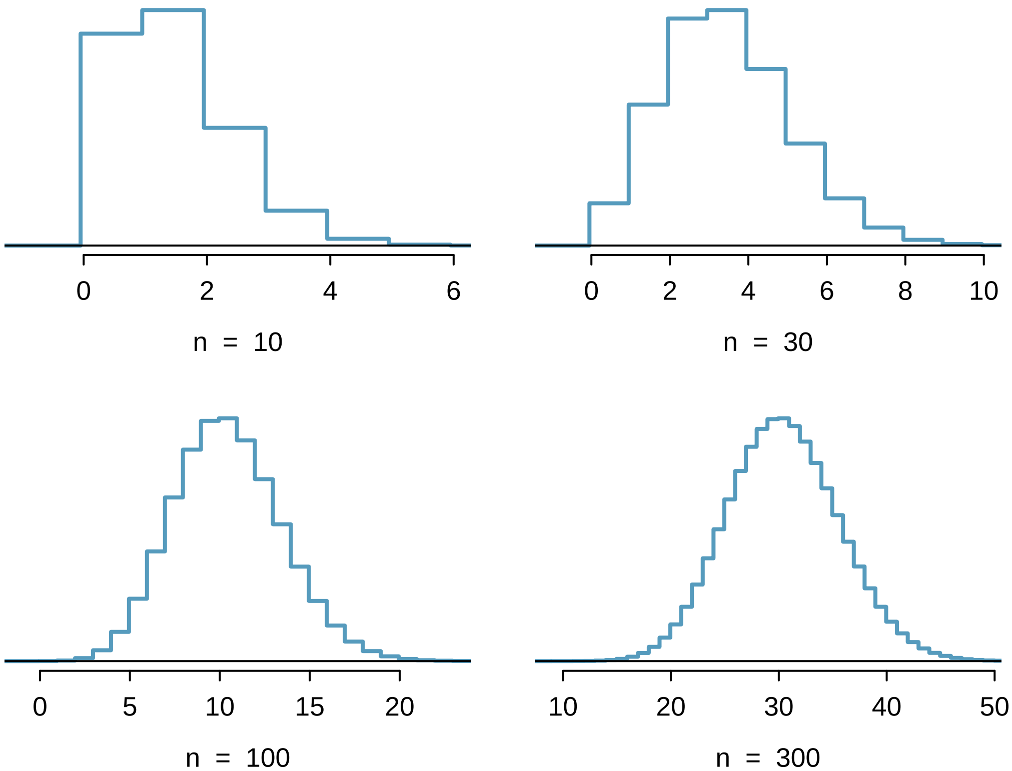


Figure 4.9: Hollow histograms of samples from the binomial model when $p = 0.10$. The sample sizes for the four plots are $n = 10$, 30, 100, and 300, respectively.

EXAMPLE 4.40

How can we use the normal approximation to estimate the probability of observing 42 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.15$?

Showing that the binomial model is reasonable was a suggested exercise in Example 4.38. We also verify that both np and $n(1-p)$ are at least 10:

$$np = 400 \times 0.15 = 60 \qquad n(1-p) = 400 \times 0.85 = 340$$

With these conditions checked, we may use the normal approximation in place of the binomial distribution using the mean and standard deviation from the binomial model:

$$\mu = np = 60 \qquad \sigma = \sqrt{np(1-p)} = 7.14$$

We want to find the probability of observing 42 or fewer smokers using this model.

GUIDED PRACTICE 4.41

Use the normal model $N(\mu = 60, \sigma = 7.14)$ to estimate the probability of observing 42 or fewer smokers. Your answer should be approximately equal to the solution of Example 4.38: 0.0054. [31]

---

[31] Compute the Z-score first: $Z = \frac{42-60}{7.14} = -2.52$. The corresponding left tail area is 0.0059.

4.3.3 The normal approximation breaks down on small intervals

The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts, even when the conditions are met.

Suppose we wanted to compute the probability of observing 49, 50, or 51 smokers in 400 when $p = 0.15$. With such a large sample, we might be tempted to apply the normal approximation and use the range 49 to 51. However, we would find that the binomial solution and the normal approximation notably differ:

Binomial: 0.0649 Normal: 0.0421

We can identify the cause of this discrepancy using Figure 4.10, which shows the areas representing the binomial probability (outlined) and normal approximation (shaded). Notice that the width of the area under the normal distribution is 0.5 units too slim on both sides of the interval.

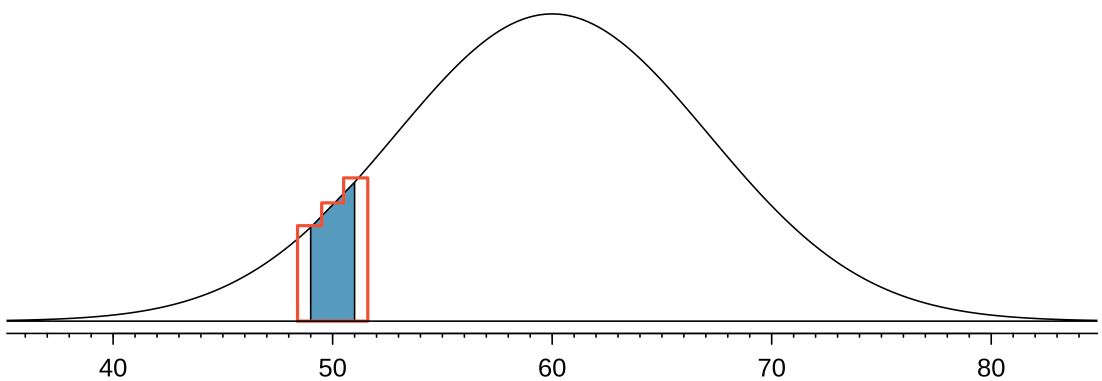


Figure 4.10: A normal curve with the area between 49 and 51 shaded. The outlined area represents the exact binomial probability.

IMPROVING THE NORMAL APPROXIMATION FOR THE BINOMIAL DISTRIBUTION

The normal approximation to the binomial distribution for intervals of values is usually improved if cutoff values are modified slightly. The cutoff values for the lower end of a shaded region should be reduced by 0.5, and the cutoff value for the upper end should be increased by 0.5.

The tip to add extra area when applying the normal approximation is most often useful when examining a range of observations. In the example above, the revised normal distribution estimate is 0.0633, much closer to the exact value of 0.0649. While it is possible to also apply this correction when computing a tail area, the benefit of the modification usually disappears since the total interval is typically quite wide.

4.4 Negative binomial distribution

The geometric distribution describes the probability of observing the first success on the n^{th} trial. The **negative binomial distribution** is more general: it describes the probability of observing the k^{th} success on the n^{th} trial.

E

EXAMPLE 4.42

Each day a high school football coach tells his star kicker, Brian, that he can go home after he successfully kicks four 35 yard field goals. Suppose we say each kick has a probability p of being successful. If p is small – e.g. close to 0.1 – would we expect Brian to need many attempts before he successfully kicks his fourth field goal?

We are waiting for the fourth success ($k = 4$). If the probability of a success (p) is small, then the number of attempts (n) will probably be large. This means that Brian is more likely to need many attempts before he gets $k = 4$ successes. To put this another way, the probability of n being small is low.

To identify a negative binomial case, we check 4 conditions. The first three are common to the binomial distribution.

IS IT NEGATIVE BINOMIAL? FOUR CONDITIONS TO CHECK

(1) The trials are independent.
(2) Each trial outcome can be classified as a success or failure.
(3) The probability of a success (p) is the same for each trial.
(4) The last trial must be a success.

G

GUIDED PRACTICE 4.43

Suppose Brian is very diligent in his attempts and he makes each 35 yard field goal with probability $p = 0.8$. Take a guess at how many attempts he would need before making his fourth kick.[32]

E

EXAMPLE 4.44

In yesterday's practice, it took Brian only 6 tries to get his fourth field goal. Write out each of the possible sequence of kicks.

Because it took Brian six tries to get the fourth success, we know the last kick must have been a success. That leaves three successful kicks and two unsuccessful kicks (we label these as failures) that make up the first five attempts. There are ten possible sequences of these first five kicks, which are shown in Figure 4.11. If Brian achieved his fourth success ($k = 4$) on his sixth attempt ($n = 6$), then his order of successes and failures must be one of these ten possible sequences.

G

GUIDED PRACTICE 4.45

Each sequence in Figure 4.11 has exactly two failures and four successes with the last attempt always being a success. If the probability of a success is $p = 0.8$, find the probability of the first sequence.[33]

If the probability Brian kicks a 35 yard field goal is $p = 0.8$, what is the probability it takes

[32]One possible answer: since he is likely to make each field goal attempt, it will take him at least 4 attempts but probably not more than 6 or 7.

[33]The first sequence: $0.2 \times 0.2 \times 0.8 \times 0.8 \times 0.8 \times 0.8 = 0.0164$.

	Kick Attempt					
	1	2	3	4	5	6
1	F	F	$\overset{1}{S}$	$\overset{2}{S}$	$\overset{3}{S}$	$\overset{4}{S}$
2	F	$\overset{1}{S}$	F	$\overset{2}{S}$	$\overset{3}{S}$	$\overset{4}{S}$
3	F	$\overset{1}{S}$	$\overset{2}{S}$	F	$\overset{3}{S}$	$\overset{4}{S}$
4	F	$\overset{1}{S}$	$\overset{2}{S}$	$\overset{3}{S}$	F	$\overset{4}{S}$
5	$\overset{1}{S}$	F	F	$\overset{2}{S}$	$\overset{3}{S}$	$\overset{4}{S}$
6	$\overset{1}{S}$	F	$\overset{2}{S}$	F	$\overset{3}{S}$	$\overset{4}{S}$
7	$\overset{1}{S}$	F	$\overset{2}{S}$	$\overset{3}{S}$	F	$\overset{4}{S}$
8	$\overset{1}{S}$	$\overset{2}{S}$	F	F	$\overset{3}{S}$	$\overset{4}{S}$
9	$\overset{1}{S}$	$\overset{2}{S}$	F	$\overset{3}{S}$	F	$\overset{4}{S}$
10	$\overset{1}{S}$	$\overset{2}{S}$	$\overset{3}{S}$	F	F	$\overset{4}{S}$

Figure 4.11: The ten possible sequences when the fourth successful kick is on the sixth attempt.

Brian exactly six tries to get his fourth successful kick? We can write this as

$$\begin{aligned}&P(\text{it takes Brian six tries to make four field goals})\\&\quad = P(\text{Brian makes three of his first five field goals, and he makes the sixth one})\\&\quad = P(1^{st}\text{ sequence OR }2^{nd}\text{ sequence OR ... OR }10^{th}\text{ sequence})\end{aligned}$$

where the sequences are from Figure 4.11. We can break down this last probability into the sum of ten disjoint possibilities:

$$\begin{aligned}&P(1^{st}\text{ sequence OR }2^{nd}\text{ sequence OR ... OR }10^{th}\text{ sequence})\\&\quad = P(1^{st}\text{ sequence}) + P(2^{nd}\text{ sequence}) + \cdots + P(10^{th}\text{ sequence})\end{aligned}$$

The probability of the first sequence was identified in Guided Practice 4.45 as 0.0164, and each of the other sequences have the same probability. Since each of the ten sequence has the same probability, the total probability is ten times that of any individual sequence.

The way to compute this negative binomial probability is similar to how the binomial problems were solved in Section 4.3. The probability is broken into two pieces:

$$\begin{aligned}&P(\text{it takes Brian six tries to make four field goals})\\&\quad = [\text{Number of possible sequences}] \times P(\text{Single sequence})\end{aligned}$$

Each part is examined separately, then we multiply to get the final result.

We first identify the probability of a single sequence. One particular case is to first observe all the failures ($n - k$ of them) followed by the k successes:

$$\begin{aligned}&P(\text{Single sequence})\\&\quad = P(n-k\text{ failures and then }k\text{ successes})\\&\quad = (1-p)^{n-k}p^k\end{aligned}$$

We must also identify the number of sequences for the general case. Above, ten sequences were identified where the fourth success came on the sixth attempt. These sequences were identified by fixing the last observation as a success and looking for all the ways to arrange the other observations. In other words, how many ways could we arrange $k-1$ successes in $n-1$ trials? This can be found using the n choose k coefficient but for $n-1$ and $k-1$ instead:

$$\binom{n-1}{k-1} = \frac{(n-1)!}{(k-1)!\,((n-1)-(k-1))!} = \frac{(n-1)!}{(k-1)!\,(n-k)!}$$

This is the number of different ways we can order $k-1$ successes and $n-k$ failures in $n-1$ trials. If the factorial notation (the exclamation point) is unfamiliar, see page **??**.

NEGATIVE BINOMIAL DISTRIBUTION

The negative binomial distribution describes the probability of observing the k^{th} success on the n^{th} trial, where all trials are independent:

$$P(\text{the } k^{th} \text{ success on the } n^{th} \text{ trial}) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

The value p represents the probability that an individual trial is a success.

EXAMPLE 4.46

E

Show using the formula for the negative binomial distribution that the probability Brian kicks his fourth successful field goal on the sixth attempt is 0.164.

The probability of a single success is $p = 0.8$, the number of successes is $k = 4$, and the number of necessary attempts under this scenario is $n = 6$.

$$\binom{n-1}{k-1} p^k (1-p)^{n-k} \;=\; \frac{5!}{3!2!}(0.8)^4(0.2)^2 \;=\; 10 \times 0.0164 \;=\; 0.164$$

GUIDED PRACTICE 4.47

G

The negative binomial distribution requires that each kick attempt by Brian is independent. Do you think it is reasonable to suggest that each of Brian's kick attempts are independent?[34]

GUIDED PRACTICE 4.48

G

Assume Brian's kick attempts are independent. What is the probability that Brian will kick his fourth field goal within 5 attempts?[35]

BINOMIAL VERSUS NEGATIVE BINOMIAL

In the binomial case, we typically have a fixed number of trials and instead consider the number of successes. In the negative binomial case, we examine how many trials it takes to observe a fixed number of successes and require that the last observation be a success.

---

[34] Answers may vary. We cannot conclusively say they are or are not independent. However, many statistical reviews of athletic performance suggests such attempts are very nearly independent.

[35] If his fourth field goal ($k = 4$) is within five attempts, it either took him four or five tries ($n = 4$ or $n = 5$). We have $p = 0.8$ from earlier. Use the negative binomial distribution to compute the probability of $n = 4$ tries and $n = 5$ tries, then add those probabilities together:

$$\begin{aligned} &P(n = 4 \text{ OR } n = 5) = P(n = 4) + P(n = 5) \\ &\quad = \binom{4-1}{4-1} 0.8^4 + \binom{5-1}{4-1}(0.8)^4(1-0.8) = 1 \times 0.41 + 4 \times 0.082 = 0.41 + 0.33 = 0.74 \end{aligned}$$

GUIDED PRACTICE 4.49

On 70% of days, a hospital admits at least one heart attack patient. On 30% of the days, no heart attack patients are admitted. Identify each case below as a binomial or negative binomial case, and compute the probability.[36]

G

(a) What is the probability the hospital will admit a heart attack patient on exactly three days this week?

(b) What is the probability the second day with a heart attack patient will be the fourth day of the week?

(c) What is the probability the fifth day of next month will be the first day with a heart attack patient?

[36] In each part, $p = 0.7$. (a) The number of days is fixed, so this is binomial. The parameters are $k = 3$ and $n = 7$: 0.097. (b) The last "success" (admitting a heart attack patient) is fixed to the last day, so we should apply the negative binomial distribution. The parameters are $k = 2$, $n = 4$: 0.132. (c) This problem is negative binomial with $k = 1$ and $n = 5$: 0.006. Note that the negative binomial case when $k = 1$ is the same as using the geometric distribution.

4.5 Poisson distribution

EXAMPLE 4.50

There are about 8 million individuals in New York City. How many individuals might we expect to be hospitalized for acute myocardial infarction (AMI), i.e. a heart attack, each day? According to historical records, the average number is about 4.4 individuals. However, we would also like to know the approximate distribution of counts. What would a histogram of the number of AMI occurrences each day look like if we recorded the daily counts over an entire year?

A histogram of the number of occurrences of AMI on 365 days for NYC is shown in Figure 4.12.[37] The sample mean (4.38) is similar to the historical average of 4.4. The sample standard deviation is about 2, and the histogram indicates that about 70% of the data fall between 2.4 and 6.4. The distribution's shape is unimodal and skewed to the right.

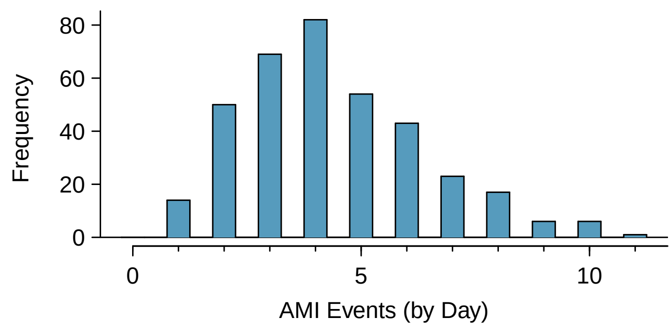


Figure 4.12: A histogram of the number of occurrences of AMI on 365 separate days in NYC.

The **Poisson distribution** is often useful for estimating the number of events in a large population over a unit of time. For instance, consider each of the following events:

- having a heart attack,
- getting married, and
- getting struck by lightning.

The Poisson distribution helps us describe the number of such events that will occur in a day for a fixed population if the individuals within the population are independent. The Poisson distribution could also be used over another unit of time, such as an hour or a week.

The histogram in Figure 4.12 approximates a Poisson distribution with rate equal to 4.4. The **rate** for a Poisson distribution is the average number of occurrences in a mostly-fixed population per unit of time. In Example 4.50, the time unit is a day, the population is all New York City residents, and the historical rate is 4.4. The parameter in the Poisson distribution is the rate – or how many events we expect to observe – and it is typically denoted by λ (the Greek letter *lambda*) or μ. Using the rate, we can describe the probability of observing exactly k events in a single unit of time.

[37] These data are simulated. In practice, we should check for an association between successive days.

POISSON DISTRIBUTION

Suppose we are watching for events and the number of observed events follows a Poisson distribution with rate λ. Then

$$P(\text{observe } k \text{ events}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where k may take a value 0, 1, 2, and so on, and $k!$ represents k-factorial, as described on page **??**. The letter $e \approx 2.718$ is the base of the natural logarithm. The mean and standard deviation of this distribution are λ and $\sqrt{\lambda}$, respectively.

We will leave a rigorous set of conditions for the Poisson distribution to a later course. However, we offer a few simple guidelines that can be used for an initial evaluation of whether the Poisson model would be appropriate.

A random variable may follow a Poisson distribution if we are looking for the number of events, the population that generates such events is large, and the events occur independently of each other.

Even when events are not really independent – for instance, Saturdays and Sundays are especially popular for weddings – a Poisson model may sometimes still be reasonable if we allow it to have a different rate for different times. In the wedding example, the rate would be modeled as higher on weekends than on weekdays. The idea of modeling rates for a Poisson distribution against a second variable such as `dayOfTheWeek` forms the foundation of some more advanced methods that fall in the realm of **generalized linear models**. In Chapters **??** and **??**, we will discuss a foundation of linear models.

Chapter 5

Foundations for inference

Statistical inference is primarily concerned with understanding 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 values where the true population value is likely to lie. Finally, we introduce a *hypothesis testing framework*, which allows us to formally evaluate claims about the population, such as whether a survey shows a candidate has a majority of support of the voting population (whether the proportion that supports is greater than 0.5).

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, and 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]). [Joke in the footnote okay?] Unless we collect responses from every individual in the sample, 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 **??** 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 expanding solar energy to be 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 code; we've written a short program in Figure 5.1 in case you are curious what the computer code

[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] Note: 88% written as a proportion would be 0.88. It is common to switch between proportion and percent.

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.

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

E

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 $n = 1000$?

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.

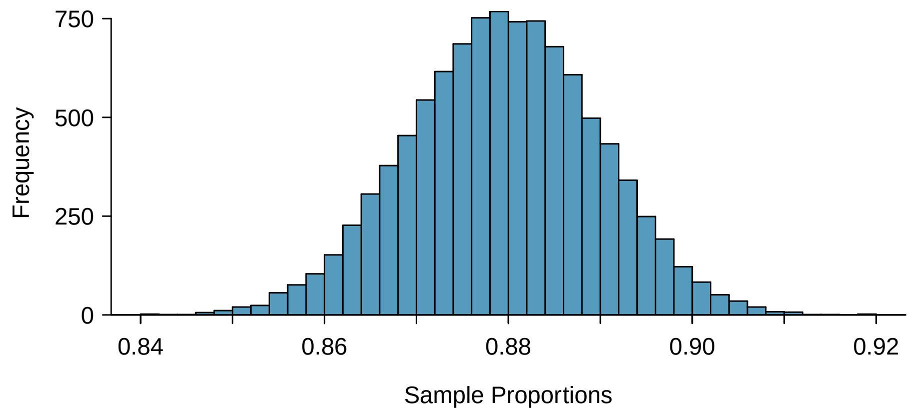


Figure 5.2: A histogram of 10,000 sample proportions, where each sample is taken from a population where the population proportion is 0.88 and the sample size is $n = 1000$.

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:[4]

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

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 $n(1-p) \geq 10$.

[4] Technically this formula is the standard deviation of $\hat{p}$, and the standard error is the estimated version using $\hat{p}$. However, to keep terminology simpler – since in every case of inference we do not know the value of p – we use the same name.

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 distribution is approximately normal.

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:

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

Both of the independence and success-failure conditions are satisfied, so the Central Limit Theorem applies and a normal distribution are reasonable in this context.

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.[5] 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.

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:

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

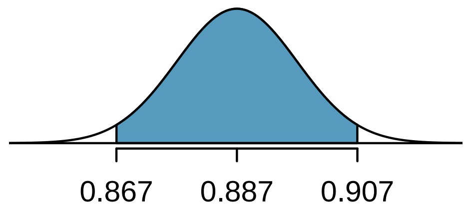
[5] For example, we could use what's called the **finite population correction factor**: if the sample is of size n and the population size is N, then we can multiple the typical standard error formula by $\sqrt{\frac{N-n}{N-1}}$ to obtain a smaller, more precise estimate of the actual standard error.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

E

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 $N(\mu_{\hat{p}} = 0.88, SE_{\hat{p}} = 0.010)$.

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 $\mu_{\hat{p}} = 0.887$ and $SE_{\hat{p}} = 0.010$, we can compute the Z-score for both the left and right cutoffs:

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

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, $p = 0.88$.

G

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 $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.[6]

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 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:

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

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

[6]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.

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.

5.1.5 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.

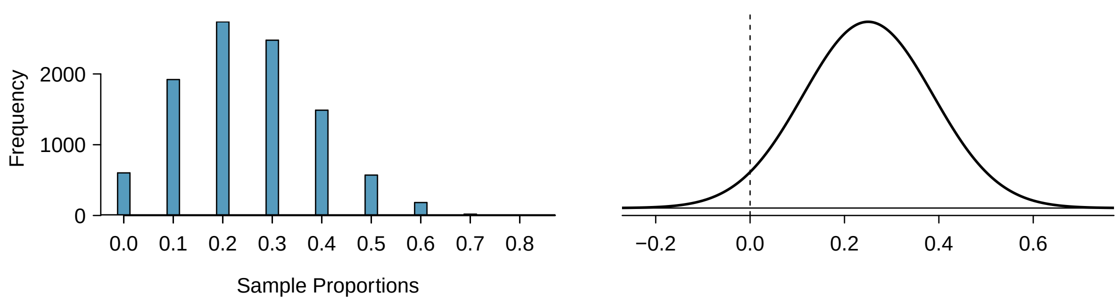


Figure 5.3: Left: simulations of $\hat{p}$ when the sample size is $n = 10$ and the population proportion is $p = 0.25$. Right: a normal distribution with the same mean (0.25) and standard deviation (0.137).

	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.

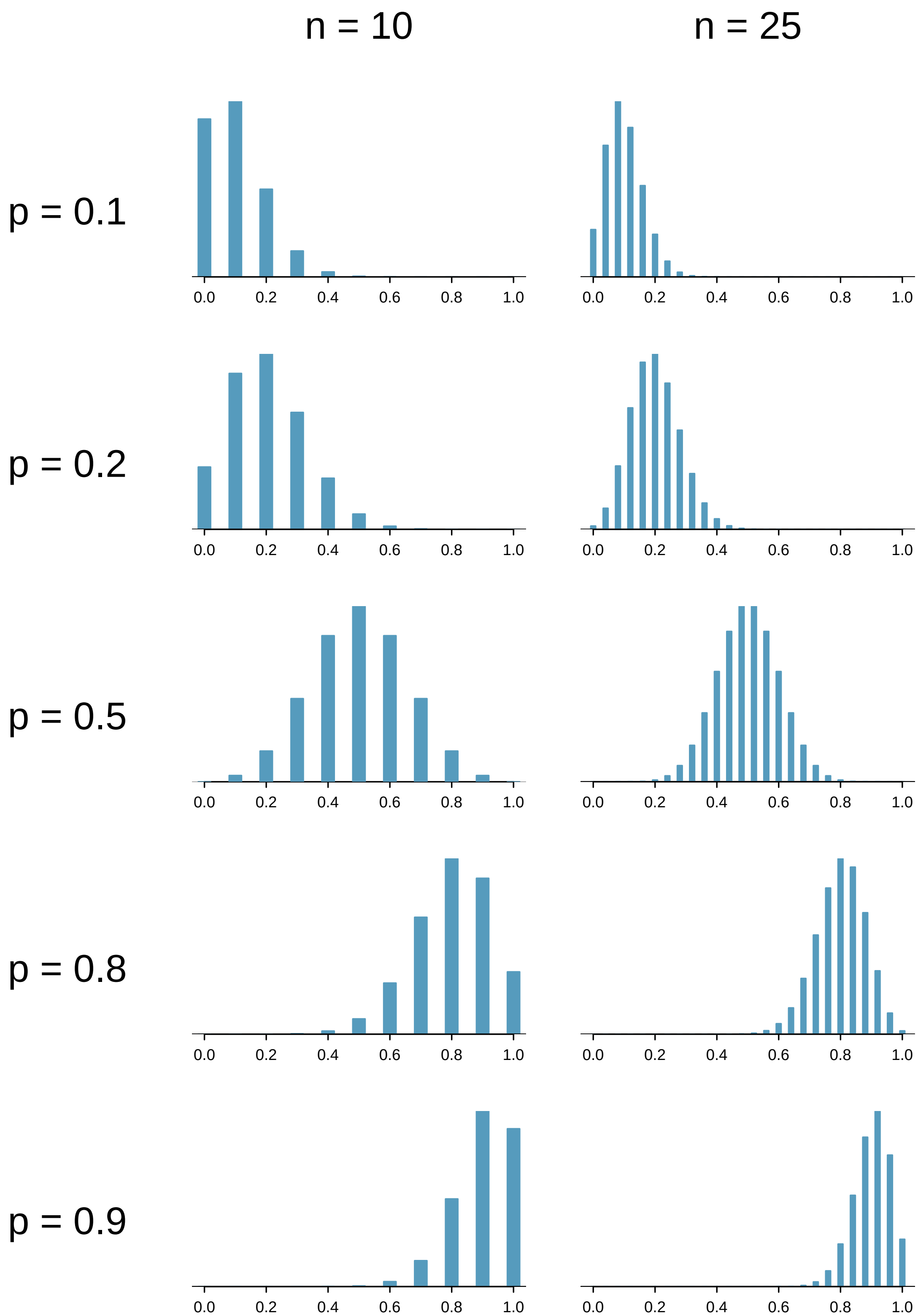


Figure 5.4: Sampling distributions for several scenarios of p and n.
Rows: $p = 0.10$, $p = 0.20$, $p = 0.50$, $p = 0.80$, and $p = 0.90$.
Columns: $n = 10$ and $n = 25$.

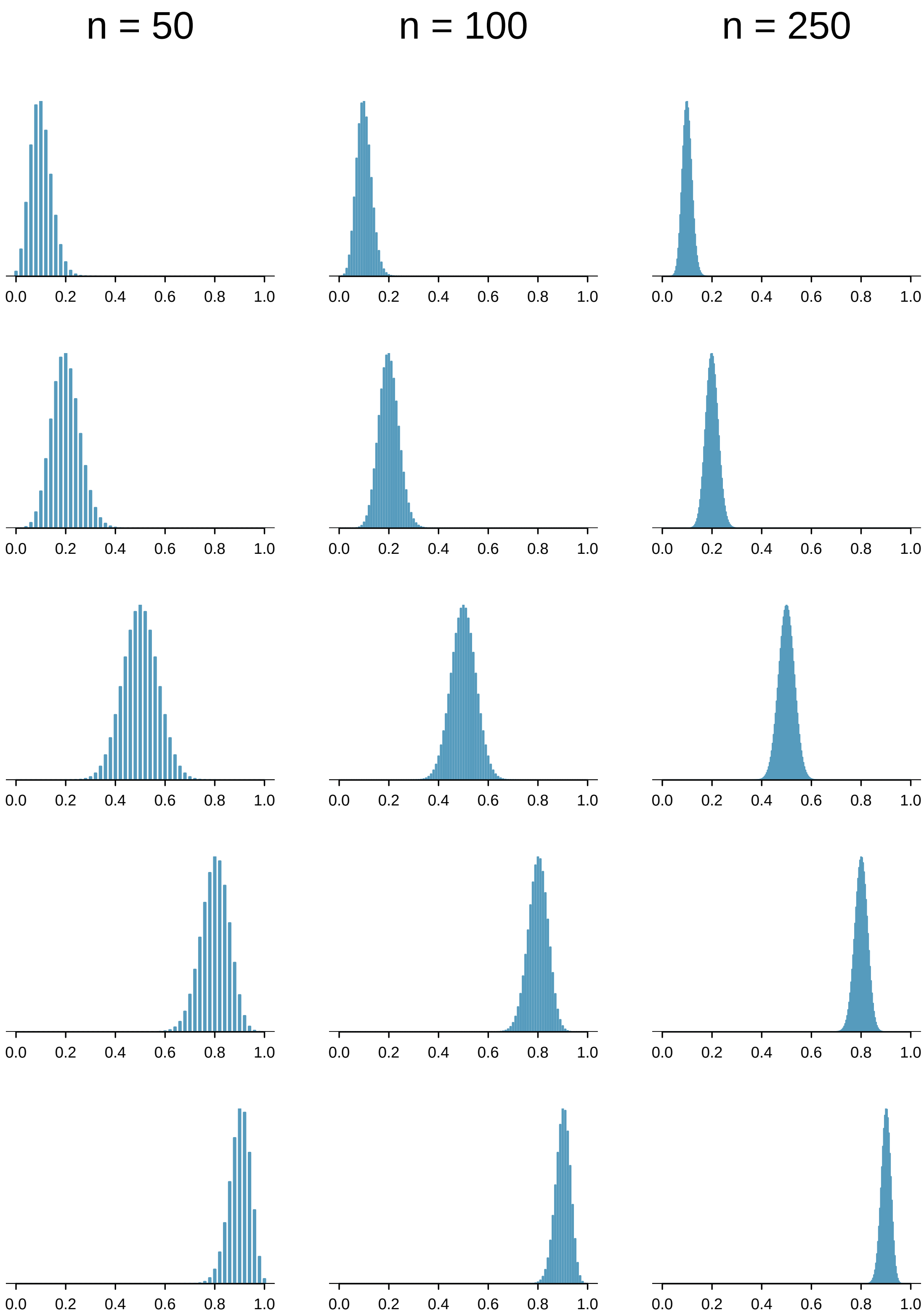


Figure 5.5: Sampling distributions for several scenarios of p and n.
Rows: $p = 0.10$, $p = 0.20$, $p = 0.50$, $p = 0.80$, and $p = 0.90$.
Columns: $n = 50$, $n = 100$, and $n = 250$.

We can complete several additional simulations, shown in Figures 5.4 and 5.5. 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.

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 be 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.

5.2 Confidence intervals for a sample 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. Instead of supplying just this point estimate of the population proportion, a next logical step would be to provide a plausible *range of values* for the population proportion.

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?[7]

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, we also know that 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:

$$\begin{aligned} \text{point estimate} &\pm 1.96 \times SE \\ \hat{p} &\pm 1.96 \times SE_{\hat{p}} \end{aligned}$$

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.

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.

---

[7] 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.

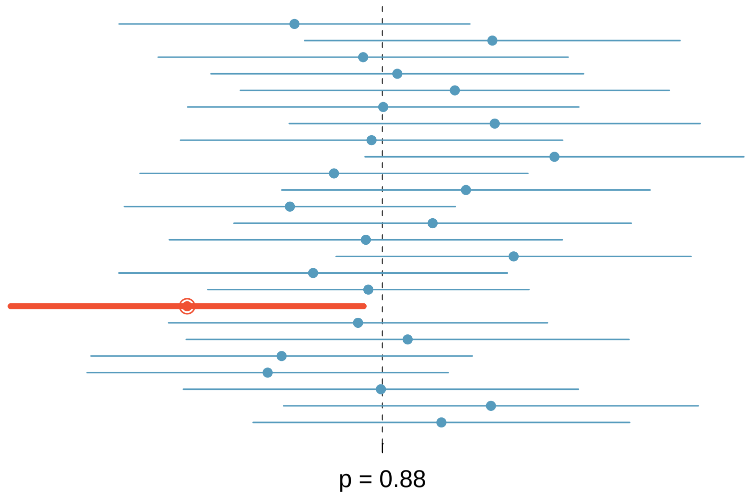


Figure 5.6: Twenty-five point estimates and confidence intervals from the simulations in Section 5.1.2. These intervals are shown relative to the population proportion $p = 0.88$. Only 1 of these 25 intervals did not capture the population proportion, and this interval has been bolded.

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 make our original 95% interval slightly slimmer.

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

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

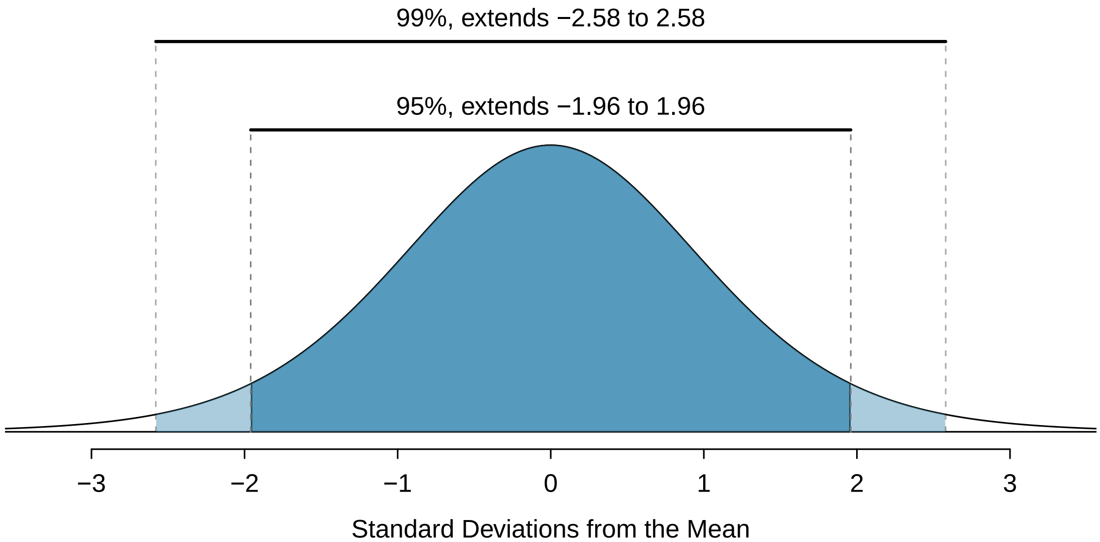


Figure 5.7: The area between -$z^\star$ and $z^\star$ increases as $z^\star$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of a normal normal distribution is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star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

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.

G

GUIDED PRACTICE 5.9

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?[8]

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

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

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

[8]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.

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.

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

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

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

CONFIDENCE INTERVAL FOR A SINGLE PROPORTION

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

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

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

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

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

5.2.4 More case studies

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

EXAMPLE 5.11

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

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.

E

EXAMPLE 5.12

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

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

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

E

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.

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.

G

GUIDED PRACTICE 5.14

Answer the following two questions about the confidence interval from Example 5.13:[9]

(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?

G

GUIDED PRACTICE 5.15

In the poll by Pew Research asking about solar energy, the researchers also inquired about other forms of energy. Support for expanding wind turbines received support from 84.8% of the 1000 respondents.[10]

(a) Is a normal approximation reasonable for the proportion of US adults who support expanding wind turbines?

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

---

[9](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.

[10](a) We check independence, which is okay since this survey was a simple random sample, and also the success-failure condition ($1000 \times 0.848 = 848$ and $1000 \times 0.152 = 152$ are both at least 10). Since both conditions are satisfied, $\hat{p} = 0.848$ can be modeled using a normal distribution.
(b) Guided Practice 5.15 confirmed that 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%.

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 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 result in a bias that tends 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.

G

GUIDED PRACTICE 5.16

Consider the 90% confidence interval from the solar energy survey, which was 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%?[11]

[11] No, a confidence interval only provides a range of plausible values for a population parameter. It does not describe what we might observe for future point estimates.

5.3 Hypothesis testing for a proportion

[Update OpenIntro's site to handle all new redirect links, many of which are currently broken in this edition.]

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.[12]

In this section, we'll be exploring how people with a 4-year college degree perform on this and other world health questions.

5.3.1 Hypothesis testing framework

We're interested in understanding whether people know much 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: Folks 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.

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

GUIDED PRACTICE 5.17

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?[13]

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?

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

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?[14]

5.3.2 Testing hypotheses using confidence intervals

[Build the `rosling_responses` data set and put it in the R package.]

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%.

---

[13] 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).

[14] 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.

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 college-educated adults 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).

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:

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

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.[15] 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.

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.

[15] 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.

G

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![16]

G

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?[17]

E

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

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

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*.

---

[16] 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.

[17] 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.[18]

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?[19]

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.

[18] 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.

[19] 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

[Jon suggested we use a simulation to introduce the p-value. Seems like it could ease some of the concepts.]

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 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.

E

EXAMPLE 5.29

Pew Research's sample found 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

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

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.

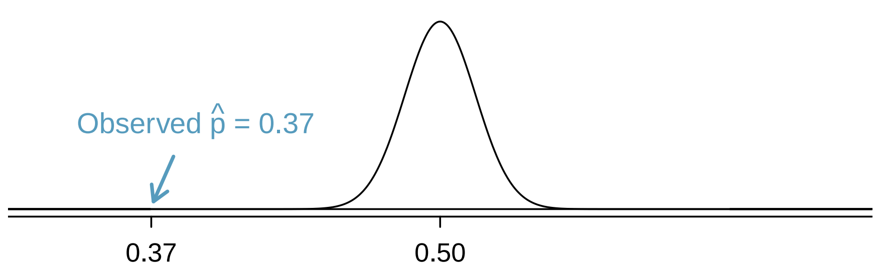


Figure 5.9: If the null hypothesis were true, this normal distribution describes the distribution of $\hat{p}$.

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:

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

We can use software to find the tail area: 2.2×10^{-16} (0.00000000000000022). If using the normal probability table in Appendix B.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.

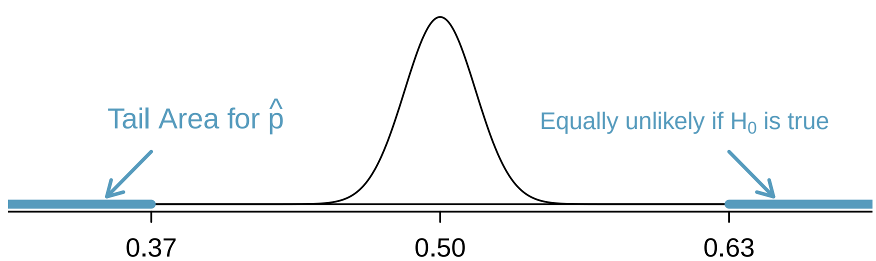


Figure 5.10: If H_0 were true, then the values above 0.63 are just as unlikely as values below 0.37.

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:

1. The null hypothesis is true, and we just happened to get observe something so extreme that only happens about once in every 23 quadrillion times (1 quadrillion = 1 million $\times$ 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.

G

GUIDED PRACTICE 5.32

Do a majority of Americans support or oppose nuclear arms reduction? Set up hypotheses to evaluate this question.[20]

EXAMPLE 5.33

A simple random sample of 1028 US adults in March 2013 found 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.

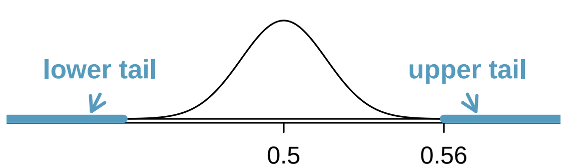
E

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

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

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.56 - 0.50}{0.0156} = 3.75$$

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



The upper tail area is 0.0002 or less, and we double this tail area to get the p-value: 0.0004. 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.

---

[20] 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 out 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

E

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?

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.

E

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?

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.

G

GUIDED PRACTICE 5.36

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.[21]

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.

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

[21] 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 $\alpha = 0.01$.

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

E

In Section **??**, 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?

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 folks 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. [If planning to add some EOCEs that touch on this topic, then will want to tweak the text here.]

[The current plan is to cut this example. @Mine, what do you think about making this example into an odd-numbered EOCE?]

E

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?

- 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.

Chapter 6

Inference for categorical data

In this chapter, we apply the methods and ideas from Chapter 5 in several contexts for categorical data. We'll start by revisiting what we learned for a single proportion, where the normal distribution can be used to model the uncertainty in the sample proportion. Next, we apply these same ideas to analyze the difference of two proportions using the normal model. Later in the chapter, we apply inference techniques to contingency tables; while we will use a different distribution in this context, the core ideas of hypothesis testing remain the same.

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

6.1 Inference for a single proportion

We encountered inference methods for a single proportion in Chapter 5, exploring point estimates, confidence intervals, and hypothesis tests. In this section, we'll do a review of these topics and also explore how to perform sample size calculations for data collection purposes in the context of a single proportion.

6.1.1 Identifying when the sample proportion is nearly normal

A sample proportion $\hat{p}$ will tend to be well-modeled using a normal distribution when the observations in the sample are independent and the sample size is sufficiently large.

SAMPLING DISTRIBUTION OF $\hat{P}$

The sampling distribution for $\hat{p}$, taken from a sample of size n from a population with a true proportion p, is nearly normal when:

1. The sample observations are independent, e.g. are from a simple random sample.
2. We expected to see at least 10 successes and 10 failures in the sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$. This is called the **success-failure condition**.

When these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Typically we don't know the true proportion, p, so we substitute some value to check conditions and estimate the standard error. For confidence intervals, usually the sample proportion $\hat{p}$ is used to check the success-failure condition and compute the standard error. For hypothesis tests, typically the null value – that is, the proportion claimed in the null hypothesis – is used in place of p.

6.1.2 Confidence intervals for a proportion

Often times we want to understand what a range of plausible values for the parameter, p. When this is our goal, a confidence interval is useful, which for a proportion usually takes the form of

$$\hat{p} \pm z^{\star} \times SE_{\hat{p}}$$

A core requirement is that $\hat{p}$ can be reasonably modeled with a normal distribution, and we check this requirement using conditions.

EXAMPLE 6.1

A simple random sample of 826 payday loan borrowers was surveyed to better understand their interests around regulation and costs. 70% of the responses supported new regulations on payday lenders. Is it reasonable to model $\hat{p} = 0.70$ using a normal distribution?

E

This is a random sample, so the observations are independent and representative of the population of interest.

We also must check the success-failure condition. In the example, we are defining a *success* as a respondent who supported new regulations and a *failure* as someone who did not. We check this condition using $\hat{p}$ in place of p when computing a confidence interval:

Successes: $np \approx 826 * 0.70 = 578$ Failures: $n(1-p) \approx 826 * (1-0.70) = 248$

Since both values are at least 10, we can use the normal distribution to model $\hat{p}$.

GUIDED PRACTICE 6.2

Estimate the standard error of $\hat{p} = 0.70$. Because p is unknown and the standard error is for a confidence interval, use $\hat{p}$ in place of p in the formula.[1]

EXAMPLE 6.3

Construct a 95% confidence interval for p, the proportion of payday borrowers who support increased regulation for payday lenders.

Using the standard error $SE = 0.016$ from Guided Practice 6.2, the point estimate 0.70, and $z^\star = 1.96$ for a 95% confidence interval, the confidence interval is

$$\text{point estimate} \ \pm \ z^\star \times SE \quad \rightarrow \quad 0.70 \ \pm \ 1.96 \times 0.016 \quad \rightarrow \quad (0.669, 0.731)$$

We are 95% confident that the true proportion of payday borrowers who supported regulation at the time of the poll was between 0.669 and 0.731.

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.

For additional one-proportion confidence interval examples, see Section 5.2.

6.1.3 Hypothesis testing for a proportion

One possible regulation for payday lenders is that they would be required to do a credit check and evaluate debt payments against the borrower's finances. We would like to know: would borrowers support this form of regulation?

GUIDED PRACTICE 6.4

Set up hypotheses to evaluate whether borrowers have a majority support or majority opposition for this type of regulation.[2]

To apply the normal distribution framework in the context of a hypothesis test for a proportion, the independence and success-failure conditions must be satisfied. In a hypothesis test, the success-failure condition is checked using the null proportion: we verify np_0 and $n(1 - p_0)$ are at least 10, where p_0 is the null value.

GUIDED PRACTICE 6.5

We consider another question asked of the payday loan borrowers: Do you support a regulation that would require lenders to pull your credit report and evaluate your debt payments? Of the 826 borrowers, 51% said they supported such a regulation. Is it reasonable to model $\hat{p} = 0.51$ for a hypothesis test here?[3]

---

[1] $SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.70(1-0.70)}{826}} = 0.016$.

[2] H_0: $p = 0.50$. H_A: $p \neq 0.50$.

EXAMPLE 6.6

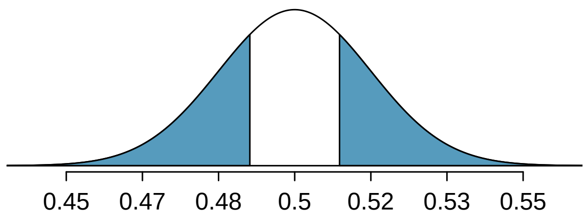
Using the hypotheses and data from Guided Practice 6.4 and 6.5, evaluate whether the poll provides convincing evidence that a majority of payday loan borrowers support a new regulation that would require lenders to pull credit reports and evaluate debt payments.

With hypotheses already set up and conditions checked, we can move onto calculations. The standard error in the context of a one-proportion hypothesis test is computed using the null value, p_0:

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{826}} = 0.017$$

A picture of the normal model is shown below with the p-value represented by the shaded region.

E



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

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.51 - 0.50}{0.017} = 0.59$$

The single tail area is 0.2776, and the p-value, represented by both tail areas together, is 0.5552. Because the p-value is larger than 0.05, we do not reject H_0. The poll does not provide convincing evidence that a majority of payday loan borrowers support or oppose regulations around credit checks and evaluation of debt payments.

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 out 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.

For additional one-proportion hypothesis test examples, see Section 5.3.

[3]Independence holds since the poll is based on a random sample. The success failure also holds, which we check using the null proportion, $p_0 = 0.5$ from the null hypothesis: $np \approx 826 \times 0.5 = 413$, $n(1-p) \approx 826 \times 0.5 = 413$.

6.1.4 When one or more conditions aren't met

We've spent a lot of time discussing conditions for when $\hat{p}$ can be reasonably modeled by a normal distribution. What happens when the success-failure condition fails? What about when the independence condition fails? In either case, the general ideas of confidence intervals and hypothesis tests remain the same, but the strategy or technique used to generate the interval or p-value would change.

When the success-failure condition isn't met for a hypothesis test, we can simulate the null distribution of $\hat{p}$ using the null value, p_0. The simulation concept is similar to the ideas used in the malaria case study presented in Section **??**, and an online section outlines this strategy:

www.openintro.org/r?go=stat_sim_prop_ht

[Port over the deleted OS3 section to be an online extra.] For a confidence interval when the success-failure condition isn't met, we can use what's called the **Clopper-Pearson interval**, where the details of this method live an internet search away, even if those details are beyond the scope of this book.

The independence condition is a more nuanced requirement. When it isn't met, it is important to understand how and why it isn't met. For example, if we took a cluster sample (see Section **??**), suitable statistical methods are available but would be beyond the scope of even most second or third courses in statistics. On the other hand, we'd be stretched to find any method that we could confidently apply to correct the inherent biases of data from a convenience sample.

While this book is scoped to well-constrained statistical problems, do remember that this is just the first book in what is a large library of statistical methods that are suitable for a very wide range of data and contexts.

6.1.5 Choosing a sample size when estimating a proportion

When collecting data, we choose a sample size suitable for the purpose of the study. Often times this means choosing a sample size large enough that the **margin of error** – which is the part we add and subtract from the point estimate in a confidence interval – is sufficiently small that the sample is useful. More explicitly, our task is to find a sample size n so that the sample proportion is within some margin of error m of the actual proportion with a certain level of confidence.

EXAMPLE 6.7

A university newspaper is conducting a survey to determine what fraction of students support a \$200 per year increase in fees to pay for a new football stadium. How big of a sample is required to ensure the margin of error is smaller than 0.04 using a 95% confidence level?

The margin of error for a sample proportion is

$$z^\star \sqrt{\frac{p(1-p)}{n}}$$

Our goal is to find the smallest sample size n so that this margin of error is smaller than $m = 0.04$. For a 95% confidence level, the value $z^\star$ corresponds to 1.96:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \ < \ 0.04$$

E

There are two unknowns in the equation: p and n. If we have an estimate of p, perhaps from a similar survey, we could enter in that value and solve for n. If we have no such estimate, we must use some other value for p. It turns out that the margin of error is largest when p is 0.5, so we typically use this *worst case value* if no estimate of the proportion is available:

$$\begin{aligned}
1.96 \times \sqrt{\frac{0.5(1-0.5)}{n}} &< 0.04 \\
1.96^2 \times \frac{0.5(1-0.5)}{n} &< 0.04^2 \\
1.96^2 \times \frac{0.5(1-0.5)}{0.04^2} &< n \\
600.25 &< n
\end{aligned}$$

We would need over 600.25 participants, which means we need 601 participants or more, to ensure the sample proportion is within 0.04 of the true proportion with 95% confidence.

When an estimate of the proportion is available, we use it in place of the worst case proportion value, 0.5.

GUIDED PRACTICE 6.8

G

A manager is about to oversee the mass production of a new tire model in her factory, and she would like to estimate what proportion of these tires will be rejected through quality control. The quality control team has monitored the last three tire models produced by the factory, failing 1.7% of tires in the first model, 6.2% of the second model, and 1.3% of the third model. The manager would like to examine enough tires to estimate the failure rate of the new tire model to within about 1% with a 90% confidence level. There are three different failure rates to choose from. Perform the sample size computation for each separately, and identify three sample sizes to consider.[4]

[4]For a 90% confidence interval, $z^\star = 1.65$, and since an estimate of the proportion 0.017 is available, we'll use it in the margin of error formula:

$$1.65 \times \sqrt{\frac{0.017(1-0.017)}{n}} < 0.01 \quad \to \quad \frac{0.017(1-0.017)}{n} < \left(\frac{0.01}{1.65}\right)^2 \quad \to \quad 454.96 < n$$

E

EXAMPLE 6.9

The sample sizes vary widely in Guided Practice 6.8. Which of the three would you suggest using? What would influence your choice?

We could examine which of the old models is most like the new model, then choose the corresponding sample size. Or if two of the previous estimates are based on small samples while the other is based on a larger sample, we should consider the value corresponding to the larger sample. There are also other reasonable approaches.

Also observe that the success-failure condition would need to be checked in the final sample. For instance, if we sampled $n = 1584$ tires and found a failure rate of 0.5%, the normal approximation would not be reasonable, and we would require more advanced statistical methods for creating the confidence interval.

G

GUIDED PRACTICE 6.10

Suppose we want to continually track the support of payday borrowers for regulation on lenders, where we would conduct a new poll every month. Running such frequent polls is expensive, so we decide a wider margin of error of 5% for each individual survey would be acceptable. Based on the original sample of borrowers where 70% supported some form of regulation, how big should our monthly sample be for a margin of error of 0.04 with 95% confidence?[5]

For sample size calculations, we always round up, so the first tire model suggests 455 tires would be sufficient.

A similar computation can be accomplished using 0.062 and 0.013 for p, and you should verify that using these proportions results in minimum sample sizes of 1584 and 350 tires, respectively.

[5] We complete the same computations as before, except now we use 0.70 instead of 0.5 for p:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.96 \times \sqrt{\frac{0.70(1-0.70)}{n}} \leq 0.05 \qquad \to \qquad n \geq 322.7$$

A sample size of 323 or more would be reasonable. (Reminder: always round up for sample size calculations!) Given that we plan to track this poll over time, we also may want to periodically repeat these calculations to ensure that we're being thoughtful in our sample size recommendations.

6.2 Difference of two proportions

We would like to make conclusions about the difference in two population proportions: $p_1 - p_2$. We consider three examples. In the first, we compare the utility of a blood thinner for heart attack patients. In the second application, we examine the efficacy of mammograms in reducing deaths from breast cancer. In the last example, a quadcopter company weighs whether to switch to a higher quality manufacturer of rotor blades.

In our investigations, we first identify a reasonable point estimate of $p_1 - p_2$ based on the sample. You may have already guessed its form: $\hat{p}_1 - \hat{p}_2$. Next, in each example we verify that the point estimate follows a normal distribution by checking certain conditions. Finally, we compute the estimate's standard error and apply our inferential framework.

6.2.1 Sample distribution of the difference of two proportions

We must check two conditions before modeling $\hat{p}_1 - \hat{p}_2$ using a normal distribution. First, the sampling distribution for each sample proportion must be nearly normal, and secondly, the samples must be independent. Under these two conditions, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ may be well approximated using a normal distribution.

CONDITIONS FOR THE SAMPLING DISTRIBUTION OF $\hat{P}_1 - \hat{P}_2$ TO BE NORMAL

The difference $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when

- each proportion separately follows a normal model, and
- the two samples are independent of each other.

The standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE^2_{\hat{p}_1} + SE^2_{\hat{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

where p_1 and p_2 represent the population proportions, and n_1 and n_2 represent the sample sizes.

6.2.2 Confidence intervals for $p_1 - p_2$

In the setting of confidence intervals for a difference of two proportions, the two sample proportions are used to verify the success-failure condition and also compute the standard error, just as was the case with a single proportion. The generic formula of a confidence interval will apply in this case, just as it did in the one-proportion case:

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

As with a single proportion context, we can follow the same Prepare, Check, Calculate, Conclude steps for computing a confidence interval or completing a hypothesis test for a difference of two proportions. The details change a little, but the general approach remain the same. Think about these steps when you apply statistical methods.

EXAMPLE 6.11

We consider an experiment for patients who underwent CPR for a heart attack and were subsequently admitted to a hospital. These patients were randomly divided into a treatment group where they received a blood thinner or the control group where they did not receive a blood thinner. The outcome variable of interest was whether the patients survived for at least 24 hours. The results are shown in Figure 6.1. Create and interpret a 90% confidence interval of the difference in the survival rates.

First the conditions must be verified. Because the patients were randomized to their groups and one heart attack patient is unlikely to influence the next that was in the study, the observations are independent, both within the samples and between the samples. The success-failure condition also holds for each group, where we have at least 10 successes and 10 failures in each experiment arm. With these conditions met, a normal distribution can be used for the point estimate of the difference in survival rates, where we'll use p_t corresponds to the survival rate in the treatment group and p_c for the control group:

E

$$\hat{p}_t - \hat{p}_c = \frac{14}{40} - \frac{11}{50} = 0.35 - 0.22 = 0.13$$

The standard error may be computed from We use the standard error formula provided on page 72. As with the one-sample proportion case, we use our best estimates of each proportion in the formula:

$$SE \approx \sqrt{\frac{0.35(1-0.35)}{40} + \frac{0.22(1-0.22)}{50}} = 0.095$$

For a 90% confidence interval, we use $z^\star = 1.65$:

$$\text{point estimate} \ \pm \ z^\star \times SE \quad \to \quad 0.13 \ \pm \ 1.65 \times 0.095 \quad \to \quad (-0.027, 0.287)$$

We are 90% confident that that blood thinners have a difference of -2.7% to +28.7% percentage point impact on survival rate for patients like those in the study. That is, we do not have enough information to say with confidence whether blood thinners help or harm heart attack patients who have been admitted after they have undergone CPR.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Figure 6.1: Results for the CPR study. Patients in the treatment group were given a blood thinner, and patients in the control group were not.

GUIDED PRACTICE 6.12

A 5-year experiment was conducted to evaluate the effectiveness of fish oils on reducing cardiovascular events, where each subject was randomized into one of two treatment groups. We'll consider heart attack outcomes in these patients:

G

	heart_attack	no_event	Total
fish_oil	145	12788	12933
placebo	200	12738	12938

Create a 95% confidence interval for the effect of fish oils on heart attacks for patients who are well-represented by those in the study. Also interpret the interval in the context of the study.[6]

---

[6] The experiment randomized, subjects are unlikely to interact with each other, the observations are independent. The success-failure condition is also met for both groups as all counts are at least 10. Compute the sample proportions ($\hat{p}_{\text{fish oil}} = 0.0112$, $\hat{p}_{\text{placebo}} = 0.0155$), point estimate of the difference ($0.0112 - 0.0155 = -0.0043$), and standard error

6.2.3 Hypothesis tests for $p_1 - p_2$

A mammogram is an X-ray procedure used to check for breast cancer. Whether mammograms should be used is part of a controversial discussion, and it's the topic of our next example where we examine 2-proportion hypothesis test when H_0 is $p_1 - p_2 = 0$ (or equivalently, $p_1 = p_2$).

A 30-year study was conducted with nearly 90,000 female participants. During a 5-year screening period, each woman was randomized to one of two groups: in the first group, women received regular mammograms to screen for breast cancer, and in the second group, women received regular non-mammogram breast cancer exams. No intervention was made during the following 25 years of the study, and we'll consider death resulting from breast cancer over the full 30-year period. Results from the study are summarized in Figure 6.2.

If mammograms are much more effective than non-mammogram breast cancer exams, then we would expect to see additional deaths from breast cancer in the control group. On the other hand, if mammograms are not as effective as regular breast cancer exams, we would expect to see an increase in breast cancer deaths in the mammogram group.

	Death from breast cancer?	
	Yes	No
Mammogram	500	44,425
Control	505	44,405

Figure 6.2: Summary results for breast cancer study.

GUIDED PRACTICE 6.13

Is this study an experiment or an observational study?[7]

GUIDED PRACTICE 6.14

Set up hypotheses to test whether there was a difference in breast cancer deaths in the mammogram and control groups.[8]

In Example **??**, we will check the conditions for using a normal distribution to analyze the results of the study. The details are very similar to that of confidence intervals. However, this time we use a special proportion called the **pooled proportion** to check the success-failure condition:

$$\begin{aligned}\hat{p}_{pooled} &= \frac{\text{\# of patients who died from breast cancer in the entire study}}{\text{\# of patients in the entire study}} \\ &= \frac{500 + 505}{500 + 44{,}425 + 505 + 44{,}405} \\ &= 0.0112\end{aligned}$$

This proportion is an estimate of the breast cancer death rate across the entire study, and it's our best estimate of the proportions p_{mgm} and p_{ctrl} *if the null hypothesis is true that* $p_{mgm} = p_{ctrl}$. We will also use this pooled proportion when computing the standard error.

($SE = \sqrt{\frac{0.0112 \times 0.9888}{12933} + \frac{0.0155 \times 0.9845}{12938}} = 0.00145$). Next, plug the values into the general formula for a confidence interval, where we'll use a 95% confidence level with $z^\star = 1.96$:

$$-0.0043 \pm 1.96 \times 0.00145 \quad \to \quad (-0.0071, -0.0015)$$

We are 95% confident that fish oils decreases heart attacks by 0.15 to 0.71 percentage points (off of a baseline of about 1.55%) over a 5-year period for subjects who are similar to those in the study. Because the interval is entirely below 0, the data provide strong evidence that fish oil supplements reduce heart attacks in patients like those in the study.

[7] This is an experiment. Patients were randomized to receive mammograms or a standard breast cancer exam. We will be able to make causal conclusions based on this study.

[8] H_0: the breast cancer death rate for patients screened using mammograms is the same as the breast cancer death rate for patients in the control, $p_{mgm} - p_{ctrl} = 0$.
H_A: the breast cancer death rate for patients screened using mammograms is different than the breast cancer death rate for patients in the control, $p_{mgm} - p_{ctrl} \neq 0$.

EXAMPLE 6.15

Can we use a normal model to analyze this study?

Because the patients are randomized, they can be treated as independent.

We also must check the success-failure condition for each group. Under the null hypothesis, the proportions p_{mgm} and p_{ctrl} are equal, so we check the success-failure condition with our best estimate of these values under H_0, the pooled proportion from the two samples, $\hat{p}_{pooled} = 0.0112$:

$$\hat{p}_{pooled} \times n_{mgm} = 0.0112 \times 44{,}925 = 503 \qquad (1 - \hat{p}_{pooled}) \times n_{mgm} = 0.9888 \times 44{,}925 = 44{,}422$$
$$\hat{p}_{pooled} \times n_{ctrl} = 0.0112 \times 44{,}910 = 503 \qquad (1 - \hat{p}_{pooled}) \times n_{ctrl} = 0.9888 \times 44{,}910 = 44{,}407$$

The success-failure condition is satisfied since all values are at least 10, and we can safely use a normal distribution.

USE THE POOLED PROPORTION ESTIMATE WHEN H_0 IS $P_1 - P_2 = 0$

When the null hypothesis is that the proportions are equal, use the pooled proportion ($\hat{p}_{pooled}$) to verify the success-failure condition and estimate the standard error:

$$\hat{p}_{pooled} = \frac{\text{number of “successes”}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

Here $\hat{p}_1 n_1$ represents the number of successes in sample 1 since

$$\hat{p}_1 = \frac{\text{number of successes in sample 1}}{n_1}$$

Similarly, $\hat{p}_2 n_2$ represents the number of successes in sample 2.

In Example **??**, the pooled proportion was used to check the success-failure condition.[9] In the next example, we see the second place where the pooled proportion comes into play: the standard error calculation.

EXAMPLE 6.16

Compute the point estimate of the difference in breast cancer death rates in the two groups, and use the pooled proportion $\hat{p}_{pooled} = 0.0112$ to calculate the standard error.

The point estimate of the difference in breast cancer death rates is

$$\begin{aligned}\hat{p}_{mgm} - \hat{p}_{ctrl} &= \frac{500}{500 + 44,425} - \frac{505}{505 + 44,405} \\ &= 0.01113 - 0.01125 \\ &= -0.00012\end{aligned}$$

The breast cancer death rate in the mammogram group was 0.012% less than in the control group. Next, the standard error is calculated *using the pooled proportion*, $\hat{p}_{pooled}$:

$$SE = \sqrt{\frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_{mgm}} + \frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_{ctrl}}} = 0.00070$$

[9] For an example of a two-proportion hypothesis test that does not require the the success-failure condition to be met, see Section **??**.

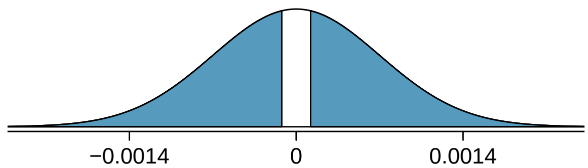
E

EXAMPLE 6.17

Using the point estimate $\hat{p}_{mgm} - \hat{p}_{ctrl} = -0.00012$ and standard error $SE = 0.00070$, calculate a p-value for the hypothesis test and write a conclusion.

Just like in past tests, we first compute a test statistic and draw a picture:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{-0.00012 - 0}{0.00070} = -0.17$$



The lower tail area is 0.4325, which we double to get the p-value: 0.8650. Because this p-value is larger than 0.05, we do not reject the null hypothesis. That is, the difference in breast cancer death rates is reasonably explained by chance, and we do not observe benefits or harm from mammograms relative to a regular breast exam.

Can we conclude that mammograms have no benefits or harm? Here are a few important considerations to keep in mind when reviewing the mammogram study as well as any other medical study:

- If mammograms are helpful or harmful, the data suggest the effect isn't very large. So while we do not accept the null hypothesis, we also don't have sufficient evidence to conclude that mammograms reduce or increase breast cancer deaths.
- Are mammograms more or less expensive than a non-mammogram breast exam? If one option is much more expensive than the other and doesn't offer clear benefits, then we should lean towards the less expensive option.
- The study's authors also found that mammograms led to overdiagnosis of breast cancer, which means some breast cancers were found (or thought to be found) but that these cancers would not cause symptoms during patients' lifetimes. That is, something else would kill the patient before breast cancer symptoms appeared. This means some patients may have been treated for breast cancer unnecessarily, and this treatment is another cost to consider. It is also important to recognize that overdiagnosis can cause unnecessary physical or emotional harm to patients.

These considerations highlight the complexity around medical care and treatment recommendations. Experts and medical boards who study medical treatments use considerations like those above to provide their best recommendation based on the current evidence.

6.2.4 More on 2-proportion hypothesis tests (special topic)

When we conduct a 2-proportion hypothesis test, usually H_0 is $p_1 - p_2 = 0$. However, there are rare situations where we want to check for some difference in p_1 and p_2 that is some value other than 0. For example, maybe we care about checking a null hypothesis where $p_1 - p_2 = 0.1$.[10] In contexts like these, we generally use $\hat{p}_1$ and $\hat{p}_2$ to check the success-failure condition and construct the standard error.

G

GUIDED PRACTICE 6.18

A quadcopter company is considering a new manufacturer for rotor blades. The new manufacturer would be more expensive but their higher-quality blades are more reliable, resulting in happier customers and fewer warranty claims. However, management must be convinced that the more expensive blades are worth the conversion before they approve the switch. If there is strong evidence of a more than 3% improvement in the percent of blades that pass inspection, management says they will switch suppliers, otherwise they will maintain the current supplier. Set up appropriate hypotheses for the test.[11]

Figure 6.3: A Phantom quadcopter.

[10] We can also encounter a similar situation with a difference of two means, though no such example is given in Chapter 7 since the methods remain exactly the same in the context of sample means. On the other hand, the success-failure condition and the calculation of the standard error vary slightly in different proportion contexts.

[11] H_0: The higher-quality blades will pass inspection just 3% more frequently than the standard-quality blades. $p_{highQ} - p_{standard} = 0.03$. H_A: The higher-quality blades will pass inspection >3% more often than the standard-quality blades. $p_{highQ} - p_{standard} > 0.03$.

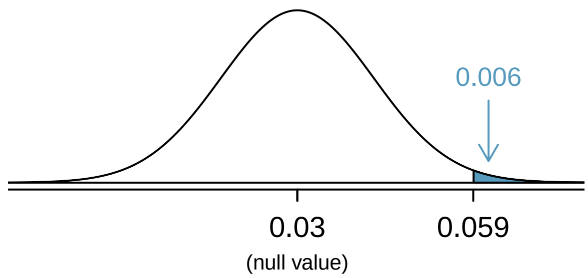


Figure 6.4: Distribution of the test statistic if the null hypothesis was true. The p-value is represented by the shaded area.

EXAMPLE 6.19

The quality control engineer from Guided Practice 6.18 collects a sample of blades, examining 1000 blades from each company and finds that 899 blades pass inspection from the current supplier and 958 pass inspection from the prospective supplier. Using these data, evaluate the hypothesis setup of Guided Practice 6.18 with a significance level of 5%.

First, we check the conditions. The sample is not necessarily random, so to proceed we must assume the blades are all independent; for this sample we will suppose this assumption is reasonable, but the engineer would be more knowledgeable as to whether this assumption is appropriate. The success-failure condition also holds for each sample. Thus, the difference in sample proportions, $0.958 - 0.899 = 0.059$, can be said to come from a nearly normal distribution.

The standard error is computed using the two sample proportions since we do not use a pooled proportion for this context:

$$SE = \sqrt{\frac{0.958(1-0.958)}{1000} + \frac{0.899(1-0.899)}{1000}} = 0.0114$$

In this hypothesis test, because the null is that $p_1 - p_2 = 0.03$, the sample proportions were used for the standard error calculation rather than a pooled proportion.

Next, we compute the test statistic and use it to find the p-value, which is depicted in Figure 6.4.

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.059 - 0.03}{0.0114} = 2.54$$

Using a standard normal distribution for this test statistic, we identify the right tail area as 0.006. Since this is a one-sided test, this single tail area is also the p-value, and we reject the null hypothesis because 0.006 is less than 0.05. That is, we have statistically significant evidence that the higher-quality blades actually do pass inspection more than 3% as often as the currently used blades. Based on these results, management will approve the switch to the new supplier.

6.2.5 Examining the standard error formula (special topic)

The formula for the standard error of the difference in two proportions is similar to the formula for other standard errors. Recall that the standard error of a single proportion, $\hat{p}_1$, is

$$SE_{\hat{p}_1} = \sqrt{\frac{p_1(1-p_1)}{n_1}}$$

The standard error of the difference of two sample proportions can be constructed from the standard errors of the separate sample proportions:

$$SE_{\hat{p}_1-\hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

This special relationship follows from probability theory.

G

GUIDED PRACTICE 6.20

Prerequisite: Section **??**. We can rewrite the equation above in a different way:

$$SE_{\hat{p}_1-\hat{p}_2}^2 = SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2$$

Explain where this formula comes from using the ideas of probability theory.[12]

[12]The standard error squared represents the variance of the estimate. If X and Y are two random variables with variances σ_x^2 and σ_y^2, then the variance of $X - Y$ is $\sigma_x^2 + \sigma_y^2$. Likewise, the variance corresponding to $\hat{p}_1 - \hat{p}_2$ is $\sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2$. Because $\sigma_{\hat{p}_1}^2$ and $\sigma_{\hat{p}_2}^2$ are just another way of writing $SE_{\hat{p}_1}^2$ and $SE_{\hat{p}_2}^2$, the variance associated with $\hat{p}_1 - \hat{p}_2$ may be written as $SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2$.

6.3 Testing for goodness of fit using chi-square

In this section, we develop a method for assessing a null model when the data are binned. This technique is commonly used in two circumstances:

- Given a sample of cases that can be classified into several groups, determine if the sample is representative of the general population.
- Evaluate whether data resemble a particular distribution, such as a normal distribution or a geometric distribution.

Each of these scenarios can be addressed using the same statistical test: a chi-square test.

In the first case, we consider data from a random sample of 275 jurors in a small county. Jurors identified their racial group, as shown in Figure 6.5, and we would like to determine if these jurors are racially representative of the population. If the jury is representative of the population, then the proportions in the sample should roughly reflect the population of eligible jurors, i.e. registered voters.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

Figure 6.5: Representation by race in a city's juries and population.

While the proportions in the juries do not precisely represent the population proportions, it is unclear whether these data provide convincing evidence that the sample is not representative. If the jurors really were randomly sampled from the registered voters, we might expect small differences due to chance. However, unusually large differences may provide convincing evidence that the juries were not representative.

A second application, assessing the fit of a distribution, is presented at the end of this section. Daily stock returns from the S&P500 for 25 years are used to assess whether stock activity each day is independent of the stock's behavior on previous days.

In these problems, we would like to examine all bins simultaneously, not simply compare one or two bins at a time, which will require us to develop a new test statistic.

6.3.1 Creating a test statistic for one-way tables

E

EXAMPLE 6.21

Of the people in the city, 275 served on a jury. If the individuals are randomly selected to serve on a jury, about how many of the 275 people would we expect to be white? How many would we expect to be black?

About 72% of the population is white, so we would expect about 72% of the jurors to be white: $0.72 \times 275 = 198$.

Similarly, we would expect about 7% of the jurors to be black, which would correspond to about $0.07 \times 275 = 19.25$ black jurors.

G

GUIDED PRACTICE 6.22

Twelve percent of the population is Hispanic and 9% represent other races. How many of the 275 jurors would we expect to be Hispanic or from another race? Answers can be found in Figure 6.6.

The sample proportion represented from each race among the 275 jurors was not a precise match for any ethnic group. While some sampling variation is expected, we would expect the sample proportions to be fairly similar to the population proportions if there is no bias on juries.

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

Figure 6.6: Actual and expected make-up of the jurors.

We need to test whether the differences are strong enough to provide convincing evidence that the jurors are not a random sample. These ideas can be organized into hypotheses:

H_0: The jurors are a random sample, i.e. there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.

H_A: The jurors are not randomly sampled, i.e. there is racial bias in juror selection.

To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts. Strong evidence for the alternative hypothesis would come in the form of unusually large deviations in the groups from what would be expected based on sampling variation alone.

6.3.2 The chi-square test statistic

In previous hypothesis tests, we constructed a test statistic of the following form:

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

This construction was based on (1) identifying the difference between a point estimate and an expected value if the null hypothesis was true, and (2) standardizing that difference using the standard error of the point estimate. These two ideas will help in the construction of an appropriate test statistic for count data.

Our strategy will be to first compute the difference between the observed counts and the counts we would expect if the null hypothesis was true, then we will standardize the difference:

$$Z_1 = \frac{\text{observed white count} - \text{null white count}}{\text{SE of observed white count}}$$

The standard error for the point estimate of the count in binned data is the square root of the count under the null.[13] Therefore:

$$Z_1 = \frac{205 - 198}{\sqrt{198}} = 0.50$$

The fraction is very similar to previous test statistics: first compute a difference, then standardize it. These computations should also be completed for the black, Hispanic, and other groups:

Black

$$Z_2 = \frac{26 - 19.25}{\sqrt{19.25}} = 1.54$$

Hispanic

$$Z_3 = \frac{25 - 33}{\sqrt{33}} = -1.39$$

Other

$$Z_4 = \frac{19 - 24.75}{\sqrt{24.75}} = -1.16$$

We would like to use a single test statistic to determine if these four standardized differences are irregularly far from zero. That is, Z_1, Z_2, Z_3, and Z_4 must be combined somehow to help determine if they – as a group – tend to be unusually far from zero. A first thought might be to take the absolute value of these four standardized differences and add them up:

$$|Z_1| + |Z_2| + |Z_3| + |Z_4| = 4.58$$

[13]Using some of the rules learned in earlier chapters, we might think that the standard error would be $np(1 - p)$, where n is the sample size and p is the proportion in the population. This would be correct if we were looking only at one count. However, we are computing many standardized differences and adding them together. It can be shown – though not here – that the square root of the count is a better way to standardize the count differences.

Indeed, this does give one number summarizing how far the actual counts are from what was expected. However, it is more common to add the squared values:

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 5.89$$

Squaring each standardized difference before adding them together does two things:

- Any standardized difference that is squared will now be positive.
- Differences that already look unusual – e.g. a standardized difference of 2.5 – will become much larger after being squared.

The test statistic X^2, which is the sum of the Z^2 values, is generally used for these reasons. We can also write an equation for X^2 using the observed counts and null counts:

$$X^2 = \frac{(\text{observed count}_1 - \text{null count}_1)^2}{\text{null count}_1} + \dots + \frac{(\text{observed count}_4 - \text{null count}_4)^2}{\text{null count}_4}$$

The final number X^2 summarizes how strongly the observed counts tend to deviate from the null counts. In Section 6.3.4, we will see that if the null hypothesis is true, then X^2 follows a new distribution called a *chi-square distribution*. Using this distribution, we will be able to obtain a p-value to evaluate the hypotheses.

6.3.3 The chi-square distribution and finding areas

The **chi-square distribution** is sometimes used to characterize data sets and statistics that are always positive and typically right skewed. Recall a normal distribution had two parameters – mean and standard deviation – that could be used to describe its exact characteristics. The chi-square distribution has just one parameter called **degrees of freedom (df)**, which influences the shape, center, and spread of the distribution.

GUIDED PRACTICE 6.23

G

Figure 6.7 shows three chi-square distributions. (a) How does the center of the distribution change when the degrees of freedom is larger? (b) What about the variability (spread)? (c) How does the shape change?[14]

Figure 6.7 and Guided Practice 6.23 demonstrate three general properties of chi-square distributions as the degrees of freedom increases: the distribution becomes more symmetric, the center moves to the right, and the variability inflates.

Our principal interest in the chi-square distribution is the calculation of p-values, which (as we have seen before) is related to finding the relevant area in the tail of a distribution. The most common ways to do this are using computer software, using a graphing calculator, or using a table. For folks wanting to use the table option, we provide an outline of how to read the chi-square table in Appendix B.3, which is also where you may find the table. [If giving some **R** in the text, then put **R** code in the examples / exercises below.] For the examples below, use your preferred approach to confirm you get the same answers.

EXAMPLE 6.24

E

Figure 6.8(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Find the shaded area.

Using statistical software or a graphing calculator, we can find that the upper tail area for a chi-square distribution with 3 degrees of freedom (*df*) and a cutoff of 6.25 is 0.1001. That is, the shaded upper tail of Figure 6.8(a) has area 0.1.

[14] (a) The center becomes larger. If took a careful look, we could see that the mean of each distribution is equal to the distribution's degrees of freedom. (b) The variability increases as the degrees of freedom increases. (c) The distribution is very strongly skewed for $df = 2$, and then the distributions become more symmetric for the larger degrees of freedom $df = 4$ and $df = 9$. We would see this trend continue if we examined distributions with even more larger degrees of freedom.

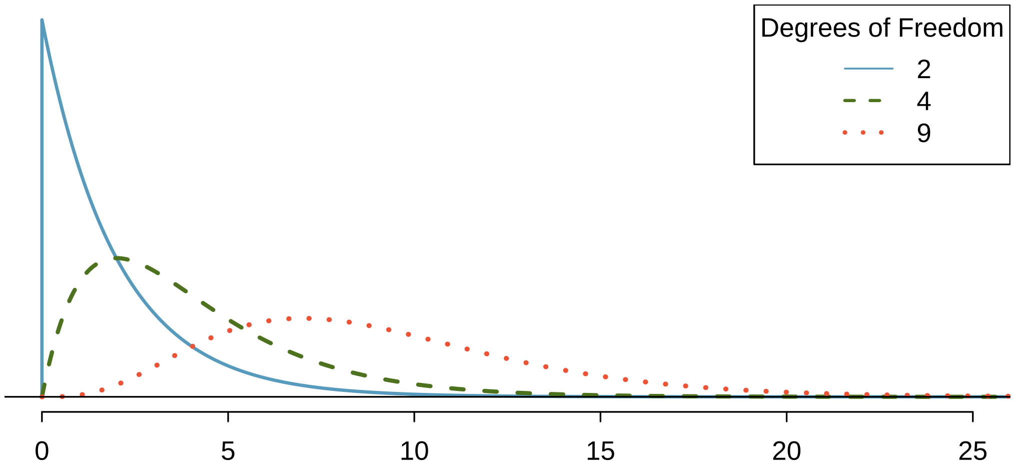


Figure 6.7: Three chi-square distributions with varying degrees of freedom.

E

EXAMPLE 6.25

Figure 6.8(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The bound for this upper tail is at 4.3. Find the tail area.

Using software, we can find that the tail area shaded in Figure 6.8(b) to be 0.1165. If using a table, we would only be able to find a range of values for the tail area: between 0.1 and 0.2.

E

EXAMPLE 6.26

Figure 6.8(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area.

Using software, we would obtain a tail area of 0.4038. If using the table in Appendix B.3, we would have identified that the tail area is larger than 0.3 but not be able to give the precise value.

G

GUIDED PRACTICE 6.27

Figure 6.8(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.[15]

G

GUIDED PRACTICE 6.28

Figure 6.8(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.[16]

G

GUIDED PRACTICE 6.29

Figure 6.8(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.[17]

---

[15] The area is 0.1109. If using a table, we would identify that it falls between 0.1 and 0.2.
[16] Precise value: 0.0404. If using the table: between 0.02 and 0.05.
[17] Precise value: 0.0266. If using the table: between 0.02 and 0.05.

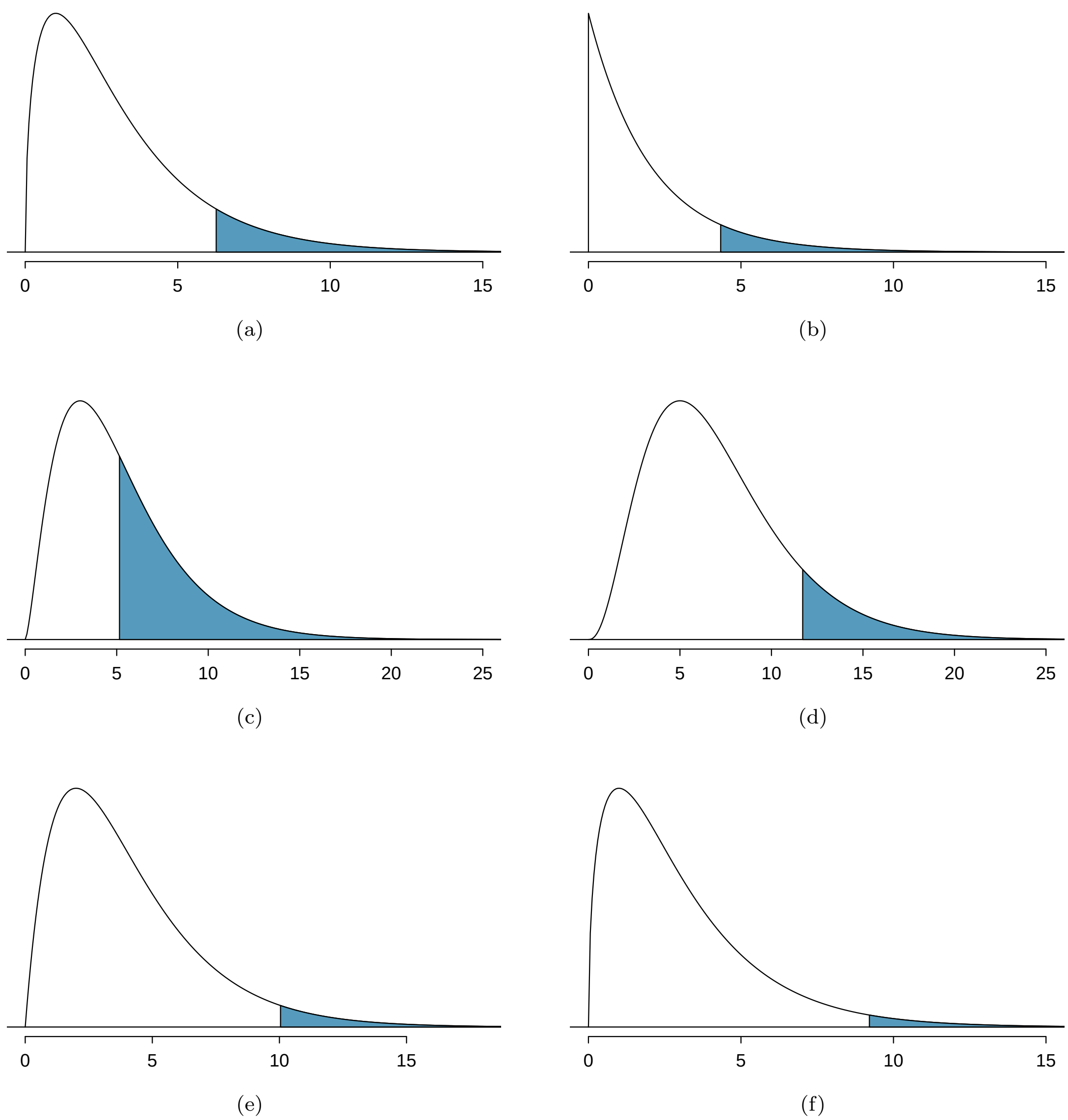


Figure 6.8: **(a)** Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. **(b)** 2 degrees of freedom, area above 4.3 shaded. **(c)** 5 degrees of freedom, area above 5.1 shaded. **(d)** 7 degrees of freedom, area above 11.7 shaded. **(e)** 4 degrees of freedom, area above 10 shaded. **(f)** 3 degrees of freedom, area above 9.21 shaded.

6.3.4 Finding a p-value for a chi-square distribution

In Section 6.3.2, we identified a new test statistic (X^2) within the context of assessing whether there was evidence of racial bias in how jurors were sampled. The null hypothesis represented the claim that jurors were randomly sampled and there was no racial bias. The alternative hypothesis was that there was racial bias in how the jurors were sampled.

We determined that a large X^2 value would suggest strong evidence favoring the alternative hypothesis: that there was racial bias. However, we could not quantify what the chance was of observing such a large test statistic ($X^2 = 5.89$) if the null hypothesis actually was true. This is where the chi-square distribution becomes useful. If the null hypothesis was true and there was no racial bias, then X^2 would follow a chi-square distribution, with three degrees of freedom in this case. Under certain conditions, the statistic X^2 follows a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of bins.

EXAMPLE 6.30

How many categories were there in the juror example? How many degrees of freedom should be associated with the chi-square distribution used for X^2?

E

In the jurors example, there were $k = 4$ categories: white, black, Hispanic, and other. According to the rule above, the test statistic X^2 should then follow a chi-square distribution with $k - 1 = 3$ degrees of freedom if H_0 is true.

Just like we checked sample size conditions to use a normal distribution in earlier sections, we must also check a sample size condition to safely apply the chi-square distribution for X^2. Each expected count must be at least 5. In the juror example, the expected counts were 198, 19.25, 33, and 24.75, all easily above 5, so we can apply the chi-square model to the test statistic, $X^2 = 5.89$.

EXAMPLE 6.31

If the null hypothesis is true, the test statistic $X^2 = 5.89$ would be closely associated with a chi-square distribution with three degrees of freedom. Using this distribution and test statistic, identify the p-value.

E

The chi-square distribution and p-value are shown in Figure 6.9. Because larger chi-square values correspond to stronger evidence against the null hypothesis, we shade the upper tail to represent the p-value. Using statistical software (or the table in Appendix B.3), we can determine that the area is 0.1171. Generally we do not reject the null hypothesis with such a large p-value. In other words, the data do not provide convincing evidence of racial bias in the juror selection.

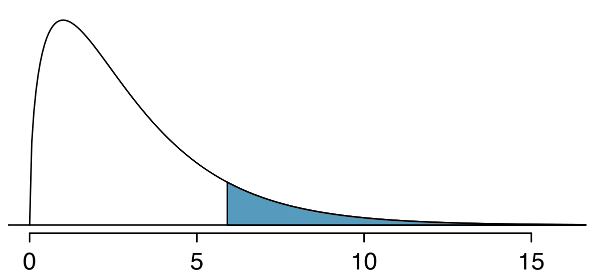


Figure 6.9: The p-value for the juror hypothesis test is shaded in the chi-square distribution with $df = 3$.

CHI-SQUARE TEST FOR ONE-WAY TABLE

Suppose we are to evaluate whether there is convincing evidence that a set of observed counts O_1, O_2, ..., O_k in k categories are unusually different from what might be expected under a null hypothesis. Call the *expected counts* that are based on the null hypothesis E_1, E_2, ..., E_k. If each expected count is at least 5 and the null hypothesis is true, then the test statistic below follows a chi-square distribution with $k - 1$ degrees of freedom:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

The p-value for this test statistic is found by looking at the upper tail of this chi-square distribution. We consider the upper tail because larger values of X^2 would provide greater evidence against the null hypothesis.

CONDITIONS FOR THE CHI-SQUARE TEST

There are two conditions that must be checked before performing a chi-square test:

Independence. Each case that contributes a count to the table must be independent of all the other cases in the table.

Sample size / distribution. Each particular scenario (i.e. cell count) must have at least 5 expected cases.

Failing to check conditions may affect the test's error rates.

When examining a table with just two bins, pick a single bin and use the one-proportion methods introduced in Section 6.1.

6.3.5 Evaluating goodness of fit for a distribution

Section 4.2 would be useful background reading for this example, but it is not a prerequisite.

We can apply the chi-square testing framework to the second problem in this section: evaluating whether a certain statistical model fits a data set. Daily stock returns from the S&P500 for 10 can be used to assess whether stock activity each day is independent of the stock's behavior on previous days. This sounds like a very complex question, and it is, but a chi-square test can be used to study the problem. We will label each day as `Up` or `Down` (`D`) depending on whether the market was up or down that day. For example, consider the following changes in price, their new labels of up and down, and then the number of days that must be observed before each `Up` day:

Change in price	2.52	-1.46	0.51	-4.07	3.36	1.10	-5.46	-1.03	-2.99	1.71
Outcome	Up	D	Up	D	Up	Up	D	D	D	Up
Days to Up	1	-	2	-	2	1	-	-	-	4

If the days really are independent, then the number of days until a positive trading day should follow a geometric distribution. The geometric distribution describes the probability of waiting for the k^{th} trial to observe the first success. Here each up day (Up) represents a success, and down (D) days represent failures. In the data above, it took only one day until the market was up, so the first wait time was 1 day. It took two more days before we observed our next `Up` trading day, and two more for the third `Up` day. We would like to determine if these counts (1, 2, 2, 1, 4, and so on) follow the geometric distribution. Figure 6.10 shows the number of waiting days for a positive trading day during 10 years for the S&P500.

Days	1	2	3	4	5	6	7+	Total
Observed	717	369	155	69	28	14	10	1362

Figure 6.10: Observed distribution of the waiting time until a positive trading day for the S&P500.

We consider how many days one must wait until observing an `Up` day on the S&P500 stock index. If the stock activity was independent from one day to the next and the probability of a positive trading day was constant, then we would expect this waiting time to follow a *geometric distribution*. We can organize this into a hypothesis framework:

H_0: The stock market being up or down on a given day is independent from all other days. We will consider the number of days that pass until an `Up` day is observed. Under this hypothesis, the number of days until an `Up` day should follow a geometric distribution.

H_A: The stock market being up or down on a given day is not independent from all other days. Since we know the number of days until an `Up` day would follow a geometric distribution under the null, we look for deviations from the geometric distribution, which would support the alternative hypothesis.

There are important implications in our result for stock traders: if information from past trading days is useful in telling what will happen today, that information may provide an advantage over other traders.

We consider data for the S&P500 and summarize the waiting times in Figure 6.11 and Figure 6.12. The S&P500 was positive on 54.5% of those days.

Because applying the chi-square framework requires expected counts to be at least 5, we have *binned* together all the cases where the waiting time was at least 7 days to ensure each expected count is well above this minimum. The actual data, shown in the *Observed* row in Figure 6.11, can be compared to the expected counts from the *Geometric Model* row. The method for computing expected counts is discussed in Figure 6.11. In general, the expected counts are determined by (1) identifying the null proportion associated with each bin, then (2) multiplying each null proportion by the total count to obtain the expected counts. That is, this strategy identifies what proportion of the total count we would expect to be in each bin.

Days	1	2	3	4	5	6	7+	Total
Observed	717	369	155	69	28	14	10	1362
Geometric Model	743	338	154	70	32	14	12	1362

Figure 6.11: Distribution of the waiting time until a positive trading day. The expected counts based on the geometric model are shown in the last row. To find each expected count, we identify the probability of waiting D days based on the geometric model ($P(D) = (1 - 0.545)^{D-1}(0.545)$) and multiply by the total number of streaks, 1362. For example, waiting for three days occurs under the geometric model about $0.455^2 \times 0.545 = 11.28\%$ of the time, which corresponds to $0.1128 \times 1362 = 154$ streaks.

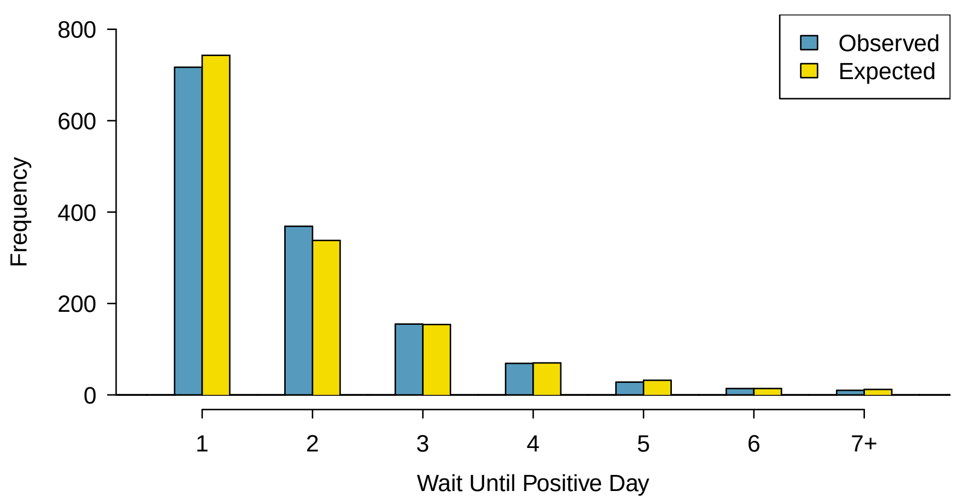


Figure 6.12: Side-by-side bar plot of the observed and expected counts for each waiting time.

EXAMPLE 6.32

Do you notice any unusually large deviations in the graph? Can you tell if these deviations are due to chance just by looking?

It is not obvious whether differences in the observed counts and the expected counts from the geometric distribution are significantly different. That is, it is not clear whether these deviations might be due to chance or whether they are so strong that the data provide convincing evidence against the null hypothesis. However, we can perform a chi-square test using the counts in Figure 6.11.

GUIDED PRACTICE 6.33

Figure 6.11 provides a set of count data for waiting times ($O_1 = 717$, $O_2 = 369$, ...) and expected counts under the geometric distribution ($E_1 = 743$, $E_2 = 338$, ...). Compute the chi-square test statistic, X^2.[18]

GUIDED PRACTICE 6.34

Because the expected counts are all at least 5, we can safely apply the chi-square distribution to X^2. However, how many degrees of freedom should we use?[19]

EXAMPLE 6.35

If the observed counts follow the geometric model, then the chi-square test statistic $X^2 = 4.61$ would closely follow a chi-square distribution with $df = 6$. Using this information, compute a p-value.

Figure 6.13 shows the chi-square distribution, cutoff, and the shaded p-value. Using software, we can find the p-value: 0.5951. Ultimately, we do not have sufficient evidence to reject the notion that the wait times follow a geometric distribution for the last 10 years of data for the S&P500, i.e. we cannot reject the notion that trading days are independent.

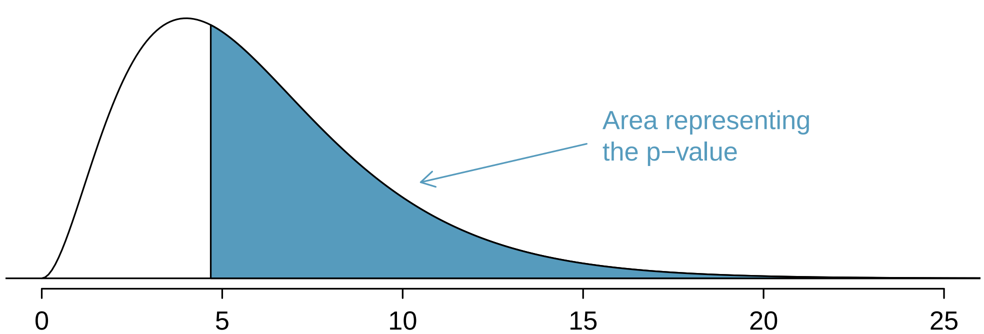


Figure 6.13: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

---

[18] $X^2 = \frac{(717-743)^2}{743} + \frac{(369-338)^2}{338} + \cdots + \frac{(10-12)^2}{12} = 4.61$

[19] There are $k = 7$ groups, so we use $df = k - 1 = 6$.

EXAMPLE 6.36

In Example 6.35, we did not reject the null hypothesis that the trading days are independent during the last 10 of data. Why is this so important?

It may be tempting to think the market is “due” for an `Up` day if there have been several consecutive days where it has been down. However, we haven’t found strong evidence that there’s any such property where the market is “due” for a correction. At the very least, the analysis suggests any dependence between days is very weak.

6.4 Testing for independence in two-way tables

We all buy used products – cars, computers, textbooks, and so on – and we sometimes assume the sellers of those products will be forthright about any underlying problems with what they're selling. This is not something we should take for granted. Researchers recruited 219 participants in a study where they would sell a used iPod[20] that was known to have frozen twice in the past. The participants were incentivized to get as much money as they could for the iPod since they would receive a 5% cut of the sale on top of $10 for participating. The researchers wanted to understand what types of questions would elicit the seller to disclose the freezing issue.

Unbeknownst to the participants who were the sellers in the study, the buyers were collaborating with the researchers to evaluate the influence of different questions on the likelihood of getting the sellers to disclose the past issues with the iPod. The scripted buyers started with "Okay, I guess I'm supposed to go first. So you've had the iPod for 2 years ..." and ended with one of three questions:

- General: What can you tell me about it?
- Positive Assumption: It doesn't have any problems, does it?
- Negative Assumption: What problems does it have?

The question is the treatment given to the sellers, and the response is whether the question prompted them to disclose the freezing issue with the iPod. The results are shown in Figure 6.14, and the data suggest that asking the, *What problems does it have?*, was the most effective at getting the seller to disclose the past freezing issues. However, you should also be asking yourself: could we see these results due to chance alone, or is this in fact evidence that some questions are more effective for getting at the truth?

	General	Positive Assumption	Negative Assumption	Total
Disclose Problem	2	23	36	61
Hide Problem	71	50	37	158
Total	73	73	73	219

Figure 6.14: Summary of the iPod study, where a question was posed to the study participant who acted

DIFFERENCES OF ONE-WAY TABLES VS TWO-WAY TABLES

A one-way table describes counts for each outcome in a single variable. A two-way table describes counts for *combinations* of outcomes for two variables. When we consider a two-way table, we often would like to know, are these variables related in any way? That is, are they dependent (versus independent)?

The hypothesis test for the iPod experiment is really about assessing whether there is statistically significant evidence that the success each question had on getting the participant to disclose the problem with the iPod. In other words, the goal is to check whether the buyer's question was independent of whether the seller disclosed a problem.

[20]For readers not as old as the authors, an iPod is basically an iPhone without any cellular service, assuming it was one of the later generations. Earlier generations were more basic.

6.4.1 Expected counts in two-way tables

Like with one-way tables, we will need to compute estimated counts for each cell in a two-way table.

EXAMPLE 6.37

From the experiment, we can compute the proportion of all sellers who disclosed the freezing problem as $61/219 = 0.2785$. If there really is no difference among the questions and 27.85% of sellers were going to disclose the freezing problem no matter the question that was put to them, how many of the 73 people in the `General` group would we have expected to disclose the freezing problem?

We would predict that $0.2785 \times 73 = 20.33$ sellers would disclose the problem. Obviously we observed fewer than this, though it is not yet clear if that is due to chance variation or whether that is because the questions vary in how effective they are at getting to the truth.

GUIDED PRACTICE 6.38

If the questions were actually equally effective, meaning about 27.85% of respondents would disclose the freezing issue regardless of what question they were asked, about how many sellers would we expect to *hide* the freezing problem from the Positive Assumption group?[21]

We can compute the expected number of sellers who we would expect to disclose or hide the freezing issue for all groups, if the questions had no impact on what they disclosed, using the same strategy employed in Example 6.37 and Guided Practice 6.38. These expected counts were used to construct Figure 6.15, which is the same as Figure 6.14, except now the expected counts have been added in parentheses.

	General	Positive Assumption	Negative Assumption	Total
Disclose Problem	2 **(20.33)**	23 **(20.33)**	36 **(20.33)**	61
Hide Problem	71 **(52.67)**	50 **(52.67)**	37 **(52.67)**	158
Total	73	73	73	219

Figure 6.15: The observed counts and the **(expected counts)**.

The examples and exercises above provided some help in computing expected counts. In general, expected counts for a two-way table may be computed using the row totals, column totals, and the table total. For instance, if there was no difference between the groups, then about 27.85% of each column should be in the first row:

$$0.2785 \times (\text{column 1 total}) = 20.33$$
$$0.2785 \times (\text{column 2 total}) = 20.33$$
$$0.2785 \times (\text{column 3 total}) = 20.33$$

Looking back to how 0.2785 was computed – as the fraction of sellers who disclosed the freezing issue (158/219) – these three expected counts could have been computed as

$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 1 total}) = 20.33$$
$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 2 total}) = 20.33$$
$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 3 total}) = 20.33$$

This leads us to a general formula for computing expected counts in a two-way table when we would like to test whether there is strong evidence of an association between the column variable and row variable.

[21] We would expect $(1 - 0.2785) \times 73 = 52.67$. It is okay that this result, like the result from Example 6.37, is a fraction.

COMPUTING EXPECTED COUNTS IN A TWO-WAY TABLE

To identify the expected count for the i^{th} row and j^{th} column, compute

$$\text{Expected Count}_{\text{row } i, \text{ col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

6.4.2 The chi-square test for two-way tables

The chi-square test statistic for a two-way table is found the same way it is found for a one-way table. For each table count, compute

$$\begin{array}{ll}
\text{General formula} & \dfrac{(\text{observed count } - \text{expected count})^2}{\text{expected count}} \\
\text{Row 1, Col 1} & \dfrac{(2 - 20.33)^2}{20.33} = 16.53 \\
\text{Row 1, Col 2} & \dfrac{(23 - 20.33)^2}{20.33} = 0.35 \\
\vdots & \vdots \\
\text{Row 2, Col 3} & \dfrac{(37 - 52.67)^2}{52.67} = 4.66
\end{array}$$

Adding the computed value for each cell gives the chi-square test statistic X^2:

$$X^2 = 16.53 + 0.35 + \cdots + 4.66 = 40.13$$

Just like before, this test statistic follows a chi-square distribution. However, the degrees of freedom are computed a little differently for a two-way table.[22] For two way tables, the degrees of freedom is equal to

$$df = (\text{number of rows minus 1}) \times (\text{number of columns minus 1})$$

In our example, the degrees of freedom parameter is

$$df = (2 - 1) \times (3 - 1) = 2$$

If the null hypothesis is true (i.e. the questions had no impact on the sellers in the experiment), then the test statistic $X^2 = 40.13$ closely follows a chi-square distribution with 2 degrees of freedom. Using this information, we can compute the p-value for the test, which is depicted in Figure 6.16.

COMPUTING DEGREES OF FREEDOM FOR A TWO-WAY TABLE

When applying the chi-square test to a two-way table, we use

$$df = (R - 1) \times (C - 1)$$

where R is the number of rows in the table and C is the number of columns.

When analyzing 2-by-2 contingency tables, one guideline is to use the two-proportion methods introduced in Section 6.2.

[22]Recall: in the one-way table, the degrees of freedom was the number of cells minus 1.

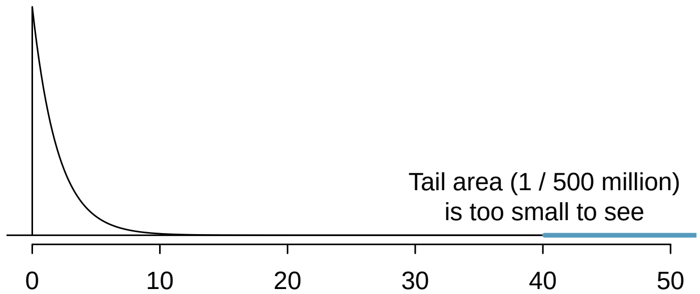


Figure 6.16: Visualization of the p-value for $X^2 = 40.13$ when $df = 2$.

EXAMPLE 6.39

Find the p-value and draw a conclusion about whether the question affects the sellers likelihood of reporting the freezing problem.

Using a computer, we can compute a very precise value for the tail area above $X^2 = 40.13$ for a chi-square distribution with 2 degrees of freedom: 0.000000002. (If using the table in Appendix B.3, we would identify the p-value is smaller than 0.001.) Using a significance level of $\alpha = 0.05$, the null hypothesis is rejected since the p-value is smaller. That is, the data provide convincing evidence that the question asked did affect a seller's likelihood to tell the truth about problems with the iPod.

EXAMPLE 6.40

Figure 6.17 summarizes the results of an experiment evaluating three treatments for Type 2 Diabetes in patients aged 10-17 who were being treated with metformin. The three treatments considered were continued treatment with metformin (`met`), treatment with metformin combined with rosiglitazone (`rosi`), or a lifestyle intervention program. Each patient had a primary outcome, which was either lacked glycemic control (failure) or did not lack that control (success). What are appropriate hypotheses for this test?

H_0: There is no difference in the effectiveness of the three treatments.

H_A: There is some difference in effectiveness between the three treatments, e.g. perhaps the `rosi` treatment performed better than `lifestyle`.

	Failure	Success	Total
`lifestyle`	109	125	234
`met`	120	112	232
`rosi`	90	143	233
Total	319	380	699

Figure 6.17: Results for the Type 2 Diabetes study.

G

GUIDED PRACTICE 6.41

A chi-square test for a two-way table may be used to test the hypotheses in Example 6.40. As a first step, compute the expected values for each of the six table cells.[23]

G

GUIDED PRACTICE 6.42

Compute the chi-square test statistic for the data in Figure 6.17.[24]

G

GUIDED PRACTICE 6.43

Because there are 3 rows and 2 columns, the degrees of freedom for the test is $df = (3-1)\times(2-1) = 2$. Use $X^2 = 8.16$, $df = 2$, evaluate whether to reject the null hypothesis using a significance level of 0.05.[25]

[23]The expected count for row one / column one is found by multiplying the row one total (234) and column one total (319), then dividing by the table total (699): $\frac{234\times 319}{699} = 106.8$. Similarly for the second column and the first row: $\frac{234\times 380}{699} = 127.2$. Row 2: 105.9 and 126.1. Row 3: 106.3 and 126.7.

[24]For each cell, compute $\frac{(\text{obs}-\text{exp})^2}{exp}$. For instance, the first row and first column: $\frac{(109-106.8)^2}{106.8} = 0.05$. Adding the results of each cell gives the chi-square test statistic: $X^2 = 0.05 + \cdots + 2.11 = 8.16$.

[25] If using a computer, we can identify the p-value as 0.017. That is, we reject the null hypothesis because the p-value is less than 0.05, and we conclude that at least one of the treatments is more or less effective than the others at treating Type 2 Diabetes for glycemic control.

Chapter 7

Inference for numerical data

Chapters 5 introduced a framework for statistical inference based on confidence intervals and hypotheses using the normal distribution for sample proportions. In this chapter, we encounter several new point estimates and a couple new distributions. In each case, the inference ideas remain the same: determine which point estimate or test statistic is useful, identify an appropriate distribution for the point estimate or test statistic, and apply the ideas from Chapter 5.

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

7.1 One-sample means with the t-distribution

Similar to how we can model the behavior of the sample proportion $\hat{p}$ using a normal distribution, the sample mean $\bar{x}$ can also be modeled using a normal distribution when certain conditions are met. However, we'll soon learn that a new distribution, called the t-distribution, tends to be more useful when working with the sample mean.

7.1.1 The sampling distribution of $\bar{x}$

The sample mean tends to follow a normal distribution centered at the the population mean, μ, when certain conditions are met. Additionally, we can compute a standard error for the sample mean using the population standard deviation σ and the sample size n.

CENTRAL LIMIT THEOREM FOR THE SAMPLE MEAN

When we collect a sufficiently large sample of n independent observations from a population with standard deviation σ, the sampling distribution of $\bar{x}$ will be nearly normal with

$$\text{Mean} = \mu \qquad\qquad \text{Standard Error } (SE) = \frac{\sigma}{\sqrt{n}}$$

Before diving into confidence intervals and hypothesis tests using $\bar{x}$, we first need to cover two topics:

- When we modeled $\hat{p}$ using the normal distribution, certain conditions had to be satisfied. The conditions for working with $\bar{x}$ are a bit more complex, and we'll spend Section 7.1.2 discussing rules of thumb.
- The standard error is dependent on the population standard deviation, σ. However, we generally do not know σ perfectly. A new distribution, called the t-distribution, will play a critical role in fixing this problem and is discussed in Section 7.1.3.

7.1.2 Evaluating the two conditions required for modeling $\bar{x}$

Two conditions are nestled into the Central Limit Theorem for the sample mean:

Independence. The sample observations must be independent, The most common way to satisfy this condition is when the sample is a simple random sample from the population. If the data come from a random process, analogous to rolling a die, this would also satisfy the independence condition.

Normality. We also require that the sample observations come from a normally distributed population, where we can relax this condition for larger sample sizes. This condition is vague and very difficult to evaluate for small samples (when it is most important), so we next provide a rule of thumb for checking this condition.

RULES OF THUMB: HOW TO PERFORM THE NORMALITY CHECK

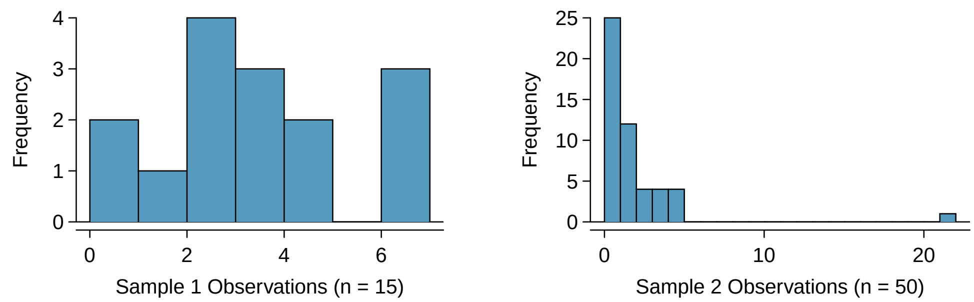
There is no perfect way to check the normality condition, so instead we use two rules of thumb:

- If the sample size n is less than 30 and there are no clear outliers in the data, then the normality condition is satisfied.
- If the sample size n is at least 30 and there are no *particularly extreme* outliers, then the normality condition is satisfied.

In this first course in statistics, you aren't expected to develop perfect judgement on the normality condition. However, you are expected to be able to handle clear cut cases based on the rules of thumb.[1]

EXAMPLE 7.1

Consider the following two plots that come from simple random samples from different populations. Their sample sizes are $n_1 = 15$ and $n_2 = 50$.



Is it reasonable to model $\bar{x}$ using a normal distribution in each of these cases?

Both samples are said to be from a simple random sample on their respective populations, so the independence condition is satisfied. Let's next check the normality condition for each using the rule of thumb.

The first sample has fewer than 30 observations, so we are watching for any clear outliers. None are present, so the normality condition is reasonably met and it is reasonable to model $\bar{x}_1$ using a normal distribution.

The second sample has a sample size greater than 30 and includes an outlier that appears to be 5 times further from the center of the distribution than the next furthest observation. This is an example of a particularly extreme outlier, so the normality condition would not be satisfied.

In practice, it's typical to also do a mental check to evaluate whether we have reason to believe the underlying population would have moderate skew (if $n < 30$) or have particularly extreme outliers ($n \geq 30$) beyond what we observe in the data. For example, consider the distribution of the number of followers for anyone who has ever posted on Twitter. Most such individuals will have built up relatively few followers, while others will have amassed tens of millions of followers.

7.1.3 Introducing the t-distribution

In practice, we cannot directly calculate the standard error for $\bar{x}$ since we do not know the population standard deviation, σ. We encountered a similar issue when computing the standard error for a sample proportion, which relied on the population proportion, p; our solution in that case was to use sample value in place of the population value when computing the standard error. We'll employ a similar strategy for computing the standard error of $\bar{x}$, using the sample standard deviation s in place of σ:

$$SE = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

This strategy tends to work well when we have a lot of data and can estimate σ using s accurately. However, the estimate is less precise with smaller samples, and this leads to problems when using the normal distribution to model $\bar{x}$.

We'll find it useful to use a new distribution for inference calculations called the **t-distribution**. A t-distribution, shown as a solid line in Figure 7.1, has a bell shape. However, its tails are thicker than the normal distribution's, meaning observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution. The extra thick tails of the t-distribution are exactly the correction needed to resolve the problem of using s in place of σ in the SE calculation.

[1] More nuanced guidelines would consider further relaxing the *particularly extreme outlier* check when the sample size is very large. However, we'll leave further discussion here to a future course.

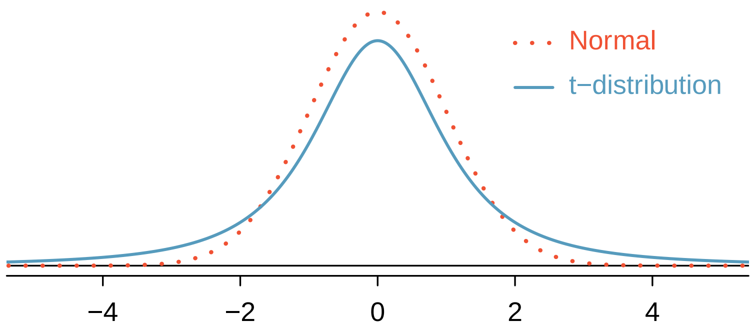


Figure 7.1: Comparison of a t-distribution and a normal distribution.

The t-distribution is always centered at zero and has a single parameter: degrees of freedom. The **degrees of freedom (df)** describes the precise form of the bell-shaped t-distribution. Several t-distributions are shown in Figure 7.2 in comparison to the normal distribution.

In general, we'll use a t-distribution with $df = n - 1$ to model the sample mean when the sample size is n. That is, when we have more observations, the degrees of freedom will be larger and the t-distribution will look more like the standard normal distribution; when the degrees of freedom is about 30 or more, the t-distribution is nearly indistinguishable from the normal distribution.

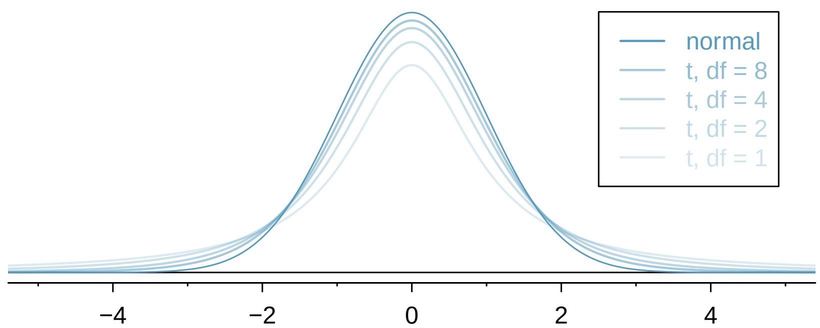


Figure 7.2: The larger the degrees of freedom, the more closely the t-distribution resembles the standard normal model.

DEGREES OF FREEDOM (DF)

The degrees of freedom describe the shape of the t-distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.

When modeling $\bar{x}$ using the t-distribution, use $df = n - 1$.

The t-distribution allows us greater flexibility than the normal distribution when analyzing numerical data. In practice, it's common to use statistical software, such as R, Python, or SAS for these analyses. Alternatively, a graphing calculator or a **t-table** may be used; the t-table is similar to the normal distribution table, and it may be found in Appendix B.2 for those who wish to use this option. No matter the approach you choose, apply your method using the examples below to confirm your working understanding of the t-distribution.

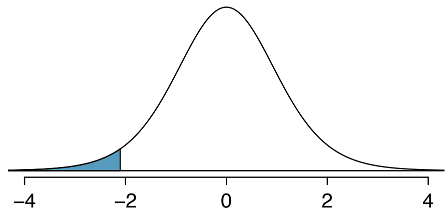


Figure 7.3: The t-distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

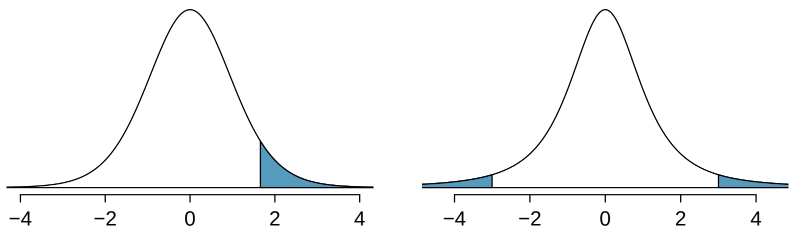


Figure 7.4: Left: The t-distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t-distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

EXAMPLE 7.2

What proportion of the t-distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw the picture in Figure 7.3 and shade the area below -2.10. The t-distribution has thicker tails than the normal distribution, and using statistical software, we can obtain a precise value: 0.0250.

EXAMPLE 7.3

A t-distribution with 20 degrees of freedom is shown in the left panel of Figure 7.4. Estimate the proportion of the distribution falling above 1.65.

With a normal distribution, this would correspond to about 0.05, so we should expect the t-distribution to give us a value in this neighborhood. Using statistical software, we can obtain the value as 0.0573.

EXAMPLE 7.4

A t-distribution with 2 degrees of freedom is shown in the right panel of Figure 7.4. Estimate the proportion of the distribution falling more than 3 units from the mean (above or below).

With so few degrees of freedom, the t-distribution will give a more notably different value than the normal distribution. Under a normal distribution, the area would be about 0.003 using the 68-95-99.7 rule. For a t-distribution with $df = 2$, the area in both tails beyond 3 units totals 0.0955. This area is dramatically different than what we obtain from the normal distribution.

GUIDED PRACTICE 7.5

What proportion of the t-distribution with 19 degrees of freedom falls above -1.79 units? Use your preferred method for finding tail areas.[2]

[2] We want to find the shaded area *above* -1.79 (we leave the picture to you). The lower tail area has an area of 0.0447, so the upper area would have an area of $1 - 0.0447 = 0.9553$.

7.1.4 One sample t-confidence intervals

Let's get our first taste of applying the t-distribution in the context of an example about the mercury content of dolphin muscle. Elevated mercury concentrations are an important problem for both dolphins and other animals, like humans, who occasionally eat them.

Figure 7.5: A Risso's dolphin.

We will identify a confidence interval for the average mercury content in dolphin muscle using a sample of 19 Risso's dolphins from the Taiji area in Japan. The data are summarized in Figure 7.6. The minimum and maximum observed values can be used to evaluate whether or not there are clear outliers.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Figure 7.6: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in micrograms of mercury per wet gram of muscle (μg/wet g).

EXAMPLE 7.6

Are the independence and normality conditions satisfied for this data set?

E

The observations are a simple random sample, therefore independence is reasonable. The summary statistics in Figure 7.6 do not suggest any skew or outliers; all observations are within 2.5 standard deviations of the mean. Based on this evidence, the normality condition seems reasonable.

In the normal model, we used $z^\star$ and the standard error to determine the width of a confidence interval. We revise the confidence interval formula slightly when using the t-distribution:

$$\bar{x} \ \pm \ t^\star_{df} \times SE$$

The sample mean is the point estimate of interest. The standard error is computed using $SE = s/\sqrt{n}$. The value $t^\star_{df}$ is a cutoff we obtain based on the confidence level and the t-distribution with df degrees of freedom.

EXAMPLE 7.7

Using the summary statistics in Figure 7.6, compute the standard error for the average mercury content in the $n = 19$ dolphins.

We plug in s and n into the formula: $SE = s/\sqrt{n} = 2.3/\sqrt{19} = 0.528$.

EXAMPLE 7.8

When $n = 19$, what is the appropriate degrees of freedom? Find $t^{\star}_{df}$ for this degrees of freedom and the confidence level of 95%

The degrees of freedom is easy to calculate: $df = n - 1 = 18$.

Using statistical software or a t-table, we can find that $t^{\star}_{18} = 2.10$. That is, 95% of the t-distribution with $df = 18$ lies within 2.10 units of 0.

EXAMPLE 7.9

Compute and interpret the 95% confidence interval for the average mercury content in Risso's dolphins.

We can construct the confidence interval as

$$\bar{x} \ \pm \ t^{\star}_{18} \times SE \quad \to \quad 4.4 \ \pm \ 2.10 \times 0.528 \quad \to \quad (3.29, 5.51)$$

We are 95% confident the average mercury content of muscles in Risso's dolphins is between 3.29 and 5.51 μg/wet gram, which is considered extremely high.

FINDING A T-CONFIDENCE INTERVAL FOR THE MEAN

Based on a sample of n independent and nearly normal observations, a confidence interval for the population mean is

$$\bar{x} \ \pm \ t^{\star}_{df} \times SE$$

where $\bar{x}$ is the sample mean, $t^{\star}_{df}$ corresponds to the confidence level and degrees of freedom, and SE is the standard error as estimated by the sample.

GUIDED PRACTICE 7.10

The FDA's webpage provides some data on mercury content of fish. Based on a sample of 15 croaker white fish (Pacific), a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. We will assume these observations are independent. Based on the summary statistics of the data, do you have any objections to the normality condition of the individual observations?[3]

---

[3] The sample size is under 30, so we check for obvious outliers: since all observations are within 2 standard deviations of the mean, there are no such clear outliers.

EXAMPLE 7.11

Estimate the standard error of $\bar{x} = 0.287$ ppm using the data summaries in Guided Practice 7.10. If we are to use the t-distribution to create a 90% confidence interval for the actual mean of the mercury content, identify the degrees of freedom we should use and also find $t^\star_{df}$.

The standard error: $SE = \frac{0.069}{\sqrt{15}} = 0.0178$.

Degrees of freedom: $df = n - 1 = 14$.

Since the goal is a 90% confidence interval, we choose $t^\star_{14}$ so that the two-tail area is 0.1: $t^\star_{14} = 1.76$.

CONFIDENCE INTERVAL FOR A SINGLE MEAN

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

Prepare. Identify $\bar{x}$, s, n, and determine what confidence level you wish to use.

Check. Verify the conditions to ensure $\bar{x}$ is nearly normal.

Calculate. If the conditions hold, compute SE, find $t^\star_{df}$, and construct the interval.

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

GUIDED PRACTICE 7.12

Using the results of Guided Practice 7.10 and Example 7.11, compute a 90% confidence interval for the average mercury content of croaker white fish (Pacific).[4]

GUIDED PRACTICE 7.13

The 90% confidence interval from Guided Practice 7.12 is 0.256 ppm to 0.318 ppm. Can we say that 90% of croaker white fish (Pacific) have mercury levels between 0.256 and 0.318 ppm?[5]

7.1.5 One sample t-tests

Is the typical US runner getting faster or slower over time? We consider this question in the context of the Cherry Blossom Race, which is a 10-mile race in Washington, DC each spring.

The average time for all runners who finished the Cherry Blossom Race in 2006 was 93.29 minutes (93 minutes and about 17 seconds). We want to determine using data from 100 participants in the 2017 Cherry Blossom Race whether runners in this race are getting faster or slower, versus the other possibility that there has been no change.

GUIDED PRACTICE 7.14

What are appropriate hypotheses for this context?[6]

---

[4] $\bar{x} \pm t^\star_{14} \times SE \rightarrow 0.287 \pm 1.76 \times 0.0178 \rightarrow (0.256, 0.318)$. We are 90% confident that the average mercury content of croaker white fish (Pacific) is between 0.256 and 0.318 ppm.

[5] No, a confidence interval only provides a range of plausible values for a population parameter, in this case the population mean. It does not describe what we might observe for individual observations.

[6] H_0: The average 10 mile run time was the same for 2006 and 2017. $\mu = 93.29$ minutes. H_A: The average 10 mile run time for 2017 was *different* than that of 2006. $\mu \neq 93.29$ minutes.

GUIDED PRACTICE 7.15

The data come from a simple random sample of all participants, so the observations are independent. However, should we be worried about the normality condition? See Figure 7.7 for a histogram of the differences.[7]

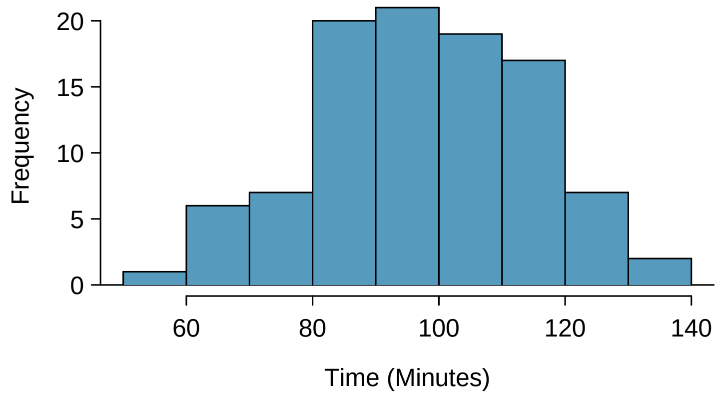


Figure 7.7: A histogram of `time` for the sample Cherry Blossom Race data.

When completing a hypothesis test for the one-sample mean, the process is nearly identical to completing a hypothesis test for a single proportion. First, we find the Z-score using the observed value, null value, and standard error; however, we call it a **T-score** since we use a t-distribution for calculating the tail area. Then we finding the p-value using the same ideas we used previously: find the one-tail area under the sampling distribution, and double it.

GUIDED PRACTICE 7.16

With both the independence and normality conditions satisfied, we can proceed with a hypothesis test using the t-distribution. The sample mean and sample standard deviation of the sample of 100 runners from the 2017 Cherry Blossom Race are 97.32 and 16.98 minutes, respectively. Recall that the sample size is 100 and the average run time in 2006 was 93.29 minutes. Find the test statistic and p-value. What is your conclusion?[8]

HYPOTHESIS TESTING FOR A SINGLE MEAN

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

Prepare. Identify the parameter of interest, list out hypotheses, identify the significance level, and identify $\bar{x}$, s, and n.

Check. Verify conditions to ensure $\bar{x}$ is nearly normal.

Calculate. If the conditions hold, compute SE, compute the T-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.

[7]With a sample of 100, we should only be concerned if there is are particularly extreme outliers. The histogram of the data doesn't show any outliers of concern (and arguably, no outliers at all).

[8]To find the test statistic (T-score), we first must determine the standard error: $SE = 16.98/\sqrt{100} = 1.70$. Now we can compute the *T-score* using the sample mean (97.32), null value (98.29), and SE: $T = \frac{97.32-93.29}{1.70} = 2.37$. For $df = 100 - 1 = 99$, we would find a one-tail area of 0.01 (using statistical software or t-table), which we double to get the p-value: 0.02. Because the p-value is smaller than 0.05, we reject the null hypothesis. That is, the data provide strong evidence that the average run time for the Cherry Blossom Run in 2017 is different than the 2006 average. Since the observed value is above the null value and we have rejected the null hypothesis, we would conclude that runners in the race were slower on average in 2017 than in 2006.

7.2 Paired data

In an earlier edition of this textbook, we found that Amazon prices were, on average, lower than those of the UCLA Bookstore for UCLA courses in 2010. It's been several years, and many stores have adapted to the online market, so we wondered, how is the UCLA Bookstore doing today?

We sampled 201 UCLA courses. Of those, 68 required books that could be found on Amazon. A portion of the data set from these courses is shown in Figure 7.8, where prices are in US dollars.

	subject	course_number	bookstore	amazon	price_difference
1	American Indian Studies	M10	47.97	47.45	0.52
2	Anthropology	2	14.26	13.55	0.71
3	Arts and Architecture	10	13.50	12.53	0.97
⋮	⋮	⋮	⋮	⋮	⋮
67	Korean	1	24.96	23.79	1.17
68	Jewish Studies	M10	35.96	32.40	3.56

Figure 7.8: Five cases of the `textbooks` data set.

7.2.1 Paired observations

Each textbook has two corresponding prices in the data set: one for the UCLA Bookstore and one for Amazon. Therefore, each textbook price from the UCLA bookstore has a natural correspondence with a textbook price from Amazon. When two sets of observations have this special correspondence, they are said to be **paired**.

PAIRED DATA

Two sets of observations are *paired* if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations. In the `textbook` data set, we look at the differences in prices, which is represented as the `diff` variable in the `textbooks` data. Here the differences are taken as

$$\text{UCLA Bookstore price} - \text{Amazon price}$$

It is important that we always subtract using a consistent order; here Amazon prices are always subtracted from UCLA prices. A histogram of these differences is shown in Figure 7.9. Using differences between paired observations is a common and useful way to analyze paired data.

GUIDED PRACTICE 7.17

G

The first difference shown in Figure 7.8 is computed as $47.97 - 47.45 = 0.52$. Verify the differences are calculated correctly for observations 2 and 3.[9]

[9] Observation 2: $14.26 - 13.55 = 0.71$. Observation 3: $13.50 - 12.53 = 0.97$.

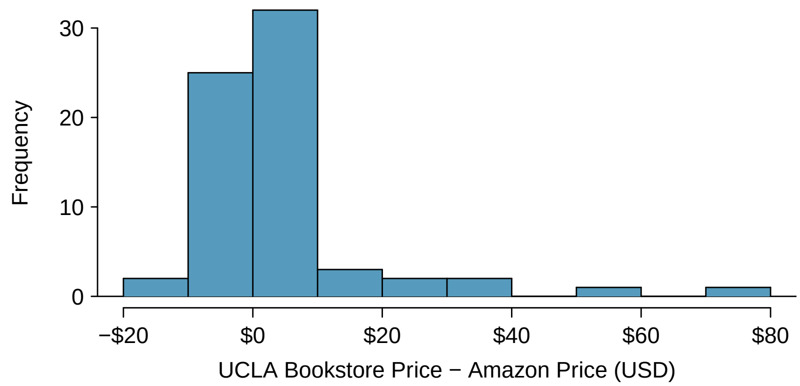


Figure 7.9: Histogram of the difference in price for each book sampled. These data are very strongly skewed.

7.2.2 Inference for paired data

To analyze a paired data set, we simply analyze the differences. We can use the same t-distribution techniques we applied in the last section.

n_{diff}	$\bar{x}_{diff}$	s_{diff}
68	3.58	13.42

Figure 7.10: Summary statistics for the 68 price differences.

EXAMPLE 7.18

Set up a hypothesis test to determine whether, on average, there is a difference between Amazon's price for a book and the UCLA bookstore's price. Also, check the conditions for whether we can move forward with the test using the t-distribution.

We are considering two scenarios: there is no difference or there is some difference in average prices.

H_0: $\mu_{diff} = 0$. There is no difference in the average textbook price.

H_A: $\mu_{diff} \neq 0$. There is a difference in average prices.

Next we check the independence and normality conditions. The observations are based on a simple random sample, so independence is reasonable. While there are some outliers, $n = 68$ and none of the outliers are particularly extreme, so normality is satisfied. With these conditions satisfied, we can move forward with the t-distribution.

EXAMPLE 7.19

Compute the standard error and test statistic from the hypothesis test started in Example 7.18.

To compute the test compute the standard error associated with $\bar{x}_{diff}$ using the standard deviation of the differences ($s_{diff} = 13.42$) and the number of differences ($n_{diff} = 68$):

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{13.42}{\sqrt{68}} = 1.63$$

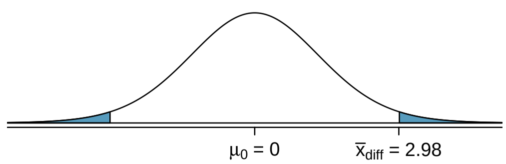
The test statistic is the T-score of $\bar{x}_{diff}$ under the null condition that the actual mean difference is 0:

$$T = \frac{\bar{x}_{diff} - 0}{SE_{x_{diff}}} = \frac{3.58 - 0}{1.63} = 2.20$$

EXAMPLE 7.20

Continue the hypothesis test from Examples 7.18-7.19, computing the p-value and making a conclusion in the context of the data.

To visualize the p-value, the sampling distribution of $\bar{x}_{diff}$ is drawn as though H_0 is true, and the p-value is represented by the two shaded tails:



The degrees of freedom is $df = 68 - 1 = 67$. Using statistical software, we find the one-tail area of 0.0156. Doubling this area gives the p-value: 0.0312. Because the p-value is less than 0.05, we reject the null hypothesis. Amazon prices are, on average, lower than the UCLA Bookstore prices for UCLA courses.

GUIDED PRACTICE 7.21

Create a 95% confidence interval for the average price difference between books at the UCLA bookstore and books on Amazon.[10]

GUIDED PRACTICE 7.22

We have strong evidence that Amazon is, on average, less expensive. How should this conclusion affect UCLA student buying habits? Should UCLA students always buy their books on Amazon?[11]

---

[10]Conditions have already verified and the standard error computed in Example 7.18. To find the interval, identify $t^\star_{67}$ using statistical software or the t-table ($t^\star_{67} = 2.00$), and plug it, the point estimate, and the standard error into the confidence interval formula:

$$\text{point estimate} \ \pm\ z^\star \times SE \quad \to \quad 3.58 \ \pm\ 2.00 \times 1.63 \quad \to \quad (0.326.84)$$

We are 95% confident that Amazon is, on average, between \$0.32 and \$6.84 less expensive than the UCLA Bookstore for UCLA course books.

[11]The average price difference is only mildly useful for this question. Examine the distribution shown in Figure 7.9. There are certainly a handful of cases where Amazon prices are much below the UCLA Bookstore's, which suggests it is worth checking Amazon or other online sites before purchasing. However, in many cases the Amazon price is above what the UCLA Bookstore charges, and most of the time the price isn't that different. Ultimately, if getting a book immediately from the bookstore is notably more convenient, e.g. to get started on homework, it's likely a good idea to go with the UCLA Bookstore unless the price difference on a specific book happens to be quite large.

For reference, this is a very different result from what we (the authors) had seen in a similar data set from 2010. At that time, Amazon prices were almost uniformly lower than those of the UCLA Bookstore's and by a large margin, making the case to use Amazon over the UCLA Bookstore quite compelling at that time.

7.3 Difference of two means

In this section we consider a difference in two population means, $\mu_1 - \mu_2$, under the condition that the data are not paired. Just as with a single sample, we identify conditions to ensure we can use the t-distribution with a point estimate of the difference, $\bar{x}_1 - \bar{x}_2$. We also will use a new standard error formula. Other than these two differences, the details are almost identical to the one-mean procedures.

We apply these methods in three contexts: determining whether stem cells can improve heart function, exploring the relationship between pregnant womens' smoking habits and birth weights of newborns, and exploring whether there is statistically significant evidence that one variation of an exam is harder than another variation. This section is motivated by questions like "Is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?"

7.3.1 Confidence interval for a difference of means

Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? Figure 7.11 contains summary statistics for an experiment to test ESCs in sheep that had a heart attack. Each of these sheep was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured in the study. Figure 7.12 provides histograms of the two data sets. A positive value corresponds to increased pumping capacity, which generally suggests a stronger recovery. Our goal will be to identify a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity relative to the control group.

	n	$\bar{x}$	s
ESCs	9	3.50	5.17
control	9	-4.33	2.76

Figure 7.11: Summary statistics of the embryonic stem cell study.

The point estimate of the difference in the heart pumping variable is straightforward to find: it is the difference in the sample means.

$$\bar{x}_{esc} - \bar{x}_{control} = 3.50 - (-4.33) = 7.83$$

For the question of whether we can model this difference using a t-distribution, we'll need to check new conditions.

USING THE T-DISTRIBUTION FOR A DIFFERENCE IN MEANS

The t-distribution can be used for inference when working with the standardized difference of two means if

1. each sample meets the conditions for using the t-distribution and
2. the samples are independent.

E

EXAMPLE 7.23

Can the t-distribution be used to make inference using the point estimate, $\bar{x}_{esc} - \bar{x}_{control} = 7.83$?

We check the two required conditions:

1. The fact that the sheep were randomized into the two groups will satisfy the independence consideration. Additionally, no observations in either group in Figure 7.12 appears to be a clear outlier (even if the ESC group has more variability). Therefore, each sample mean could itself be modeled using a t-distribution.
2. Here again, the random assignment of the sheep to the groups also ensures the groups are independent.

With both conditions met, we can use the t-distribution to model the difference of sample means.

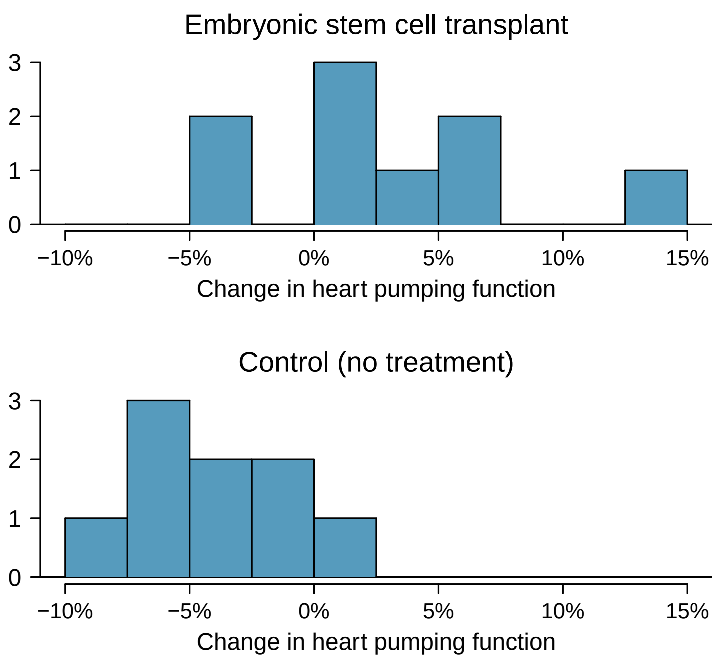


Figure 7.12: Histograms for both the embryonic stem cell group and the control group. Higher values are associated with greater improvement. We don't see any evidence of skew in these data; however, it is worth noting that skew would be difficult to detect with such a small sample.

In addition to new conditions, we also will need an updated formula for the standard error for the difference of two means.

DISTRIBUTION OF A DIFFERENCE OF SAMPLE MEANS

The sample difference of two means, $\bar{x}_1 - \bar{x}_2$, can be modeled using the t-distribution and the standard error

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

when each sample mean can itself be modeled using a t-distribution and the samples are independent. The official standard formula for the standard error is quite complex and is generally computed using software. However, if necessary, you may use the smaller of $n_1 - 1$ and $n_2 - 1$ instead for the degrees of freedom.

As with the one-sample case, we always compute the standard error using sample standard deviations rather than population standard deviations:

$$SE = \sqrt{\frac{s_{esc}^2}{n_{esc}} + \frac{s_{control}^2}{n_{control}}} = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Generally, we use statistical software to find the appropriate degrees of freedom. If software isn't available, we can use the smaller of $n_1 - 1$ and $n_2 - 1$ for the degrees of freedom, e.g. if using a t-table to find tail areas. For transparency in the Examples and Guided Practice, we'll use the latter approach for finding df; in the case of the ESC example, this means we'll use $df = 8$.

EXAMPLE 7.24

Calculate a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity of sheep after they've suffered a heart attack.

We will use the sample difference and the standard error that we computed earlier calculations:

$$\bar{x}_{esc} - \bar{x}_{control} = 7.83 \qquad SE = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

E

Using $df = 8$, we can identify the critical value of $t_8^\star = 2.31$ for a 95% confidence interval. Finally, we can enter the values into the confidence interval formula:

$$\text{point estimate} \ \pm \ t^\star \times SE \quad \rightarrow \quad 7.83 \ \pm \ 2.31 \times 1.95 \quad \rightarrow \quad (3.32, 12.34)$$

We are 95% confident that embryonic stem cells improve the heart's pumping function in sheep that have suffered a heart attack by 3.32% to 12.34%.

As with past statistical inference applications, there is a well-trodden procedure.

1. Prepare: retrieve critical contextual information, and if appropriate, set up hypotheses.
2. Check: ensure the required conditions are reasonably satisfied.
3. Calculate: find the standard error, and then construct a confidence interval or find a test statistic and p-value.
4. Conclude: interpret the result in the context of the application.

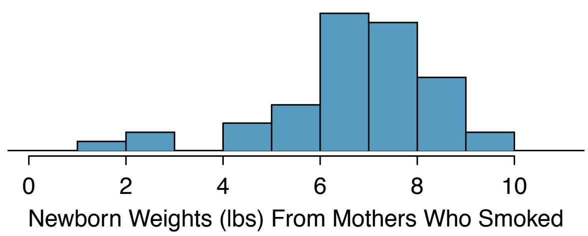
The details change a little from one setting to the next, but this general approach remain the same.

7.3.2 Hypothesis tests based on a difference in means

A data set called `ncbirths` represents a random sample of 150 cases of mothers and their newborns in North Carolina over a year. Four cases from this data set are represented in Figure 7.13. We are particularly interested in two variables: `weight` and `smoke`. The `weight` variable represents the weights of the newborns and the `smoke` variable describes which mothers smoked during pregnancy. We would like to know, is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke? We will use the North Carolina sample to try to answer this question. The smoking group includes 50 cases and the nonsmoking group contains 100 cases, represented in Figure 7.14.

	fAge	mAge	weeks	weight	sexBaby	smoke
1	NA	13	37	5.00	female	nonsmoker
2	NA	14	36	5.88	female	nonsmoker
3	19	15	41	8.13	male	smoker
⋮	⋮	⋮	⋮	⋮	⋮	
150	45	50	36	9.25	female	nonsmoker

Figure 7.13: Four cases from the `ncbirths` data set. The value "NA", shown for the first two entries of the first variable, indicates that piece of data is missing.



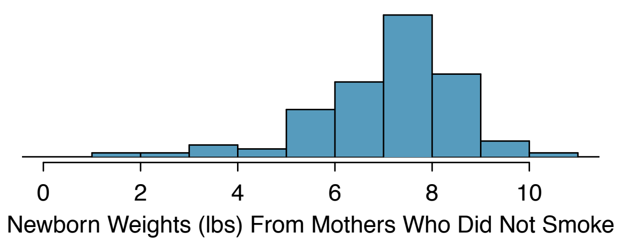


Figure 7.14: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke. The distributions exhibit moderate-to-strong and strong skew, respectively.

EXAMPLE 7.25

Set up appropriate hypotheses to evaluate whether there is a relationship between a mother smoking and average birth weight.

The null hypothesis represents the case of no difference between the groups.

H_0: There is no difference in average birth weight for newborns from mothers who did and did not smoke. In statistical notation: $\mu_n - \mu_s = 0$, where μ_n represents non-smoking mothers and μ_s represents mothers who smoked.

H_A: There is some difference in average newborn weights from mothers who did and did not smoke ($\mu_n - \mu_s \neq 0$).

We check the two conditions necessary to model the difference in sample means using the t-distribution. (1) Because the data come from a simple random sample, the observations are independent. Additionally, with both data sets over 30 observations and no particularly extreme outliers evident, normality of each sample mean is satisfied. (2) The independence reasoning applied in (1) also ensures the observations in each sample are independent. Since both conditions are satisfied, the difference in sample means may be modeled using a t-distribution.

	`smoker`	`nonsmoker`
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

Figure 7.15: Summary statistics for the `ncbirths` data set.

G

GUIDED PRACTICE 7.26

The summary statistics in Figure 7.15 may be useful for this exercise.

(a) What is the point estimate of the population difference, $\mu_n - \mu_s$?

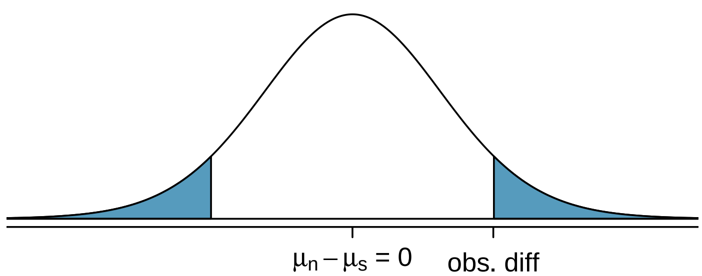
(b) Compute the standard error of the point estimate from part (a).[12]

E

EXAMPLE 7.27

Draw a picture to represent the p-value for the hypothesis test from Example 7.25.

To depict the p-value, we draw the distribution of the point estimate as though H_0 were true and shade areas representing at least as much evidence against H_0 as what was observed. Both tails are shaded because it is a two-sided test.



[12](a) The difference in sample means is an appropriate point estimate: $\bar{x}_n - \bar{x}_s = 0.40$. (b) The standard error of the estimate can be estimated using the standard error formula:

$$SE = \sqrt{\frac{\sigma_n^2}{n_n} + \frac{\sigma_s^2}{n_s}} \approx \sqrt{\frac{s_n^2}{n_n} + \frac{s_s^2}{n_s}} = \sqrt{\frac{1.60^2}{100} + \frac{1.43^2}{50}} = 0.26$$

EXAMPLE 7.28

E

Compute the p-value of the hypothesis test using the figure in Example 7.27, and evaluate the hypotheses using a significance level of $\alpha = 0.05$.

We start by computing the T-score:

$$T = \frac{0.40 - 0}{0.26} = 1.54$$

Next, we find the single tail area using software (or the table in Appendix B.2 on page 142). We'll use the smaller of $n_n - 1 = 99$ and $n_s - 1 = 49$ as the degrees of freedom: $df = 49$. The one tail area is 0.065; doubling this value gives the two-tail area and p-value, 0.135. This p-value is larger than the significance value, 0.05, so we fail to reject the null hypothesis. There is insufficient evidence to say there is a difference in average birth weight of newborns from North Carolina mothers who did smoke during pregnancy and newborns from North Carolina mothers who did not smoke during pregnancy.

GUIDED PRACTICE 7.29

G

We've seen much research suggesting smoking is harmful during pregnancy, so how could we fail to reject the null hypothesis in Example 7.28? [13]

GUIDED PRACTICE 7.30

G

If we made a Type 2 Error and there is a difference, what could we have done differently in data collection to be more likely to detect the difference?[14]

Public service announcement: while we have used this relatively small data set as an example, larger data sets show that women who smoke tend to have smaller newborns. In fact, some in the tobacco industry actually had the audacity to tout that as a *benefit* of smoking:

> *It's true. The babies born from women who smoke are smaller, but they're just as healthy as the babies born from women who do not smoke. And some women would prefer having smaller babies.*
>
> - Joseph Cullman, Philip Morris' Chairman of the Board
> on CBS' *Face the Nation*, Jan 3, 1971

Fact check: the babies from women who smoke are not actually as healthy as the babies from women who do not smoke.[15]

---

[13] It is possible that there is some difference but we did not detect it. If there is a difference, we made a Type 2 Error. Notice: we also don't have enough information to, if there is an actual difference, confidently say which direction that difference would be in.

[14] We could have collected more data. If the sample sizes are larger, we tend to have a better shot at finding a difference if one exists. In fact, this is exactly what we would find if we examined a larger data set!

[15] You can watch an episode of John Oliver on *Last Week Tonight* to explore the present day offenses of the tobacco industry. Please be aware that there is some adult language: youtu.be/6UsHHOCH4q8.

7.3.3 Case study: two versions of a course exam

An instructor decided to run two slight variations of the same exam. Prior to passing out the exams, she shuffled the exams together to ensure each student received a random version. Summary statistics for how students performed on these two exams are shown in Figure 7.16. Anticipating complaints from students who took Version B, she would like to evaluate whether the difference observed in the groups is so large that it provides convincing evidence that Version B was more difficult (on average) than Version A.

Version	n	$\bar{x}$	s	min	max
A	30	79.4	14	45	100
B	27	74.1	20	32	100

Figure 7.16: Summary statistics of scores for each exam version.

GUIDED PRACTICE 7.31

G

Construct a hypotheses to evaluate whether the observed difference in sample means, $\bar{x}_A - \bar{x}_B = 5.3$, is due to chance. We will later evaluate these hypotheses using $\alpha = 0.01$.[16]

GUIDED PRACTICE 7.32

To evaluate the hypotheses in Guided Practice 7.31 using the t-distribution, we must first verify conditions.[17]

G

(a) Does it seem reasonable that the scores are independent within each group?

(b) What about the normality condition for each group?

(c) Do you think scores from the two groups would be independent of each other, i.e. the two samples are independent?

After verifying the conditions for each sample and confirming the samples are independent of each other, we are ready to conduct the test using the t-distribution. In this case, we are estimating the true difference in average test scores using the sample data, so the point estimate is $\bar{x}_A - \bar{x}_B = 5.3$. The standard error of the estimate can be calculated as

$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{14^2}{30} + \frac{20^2}{27}} = 4.62$$

Finally, we construct the test statistic:

$$T = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{(79.4 - 74.1) - 0}{4.62} = 1.15$$

If we have a computer handy, we can identify the degrees of freedom as 45.97. Otherwise we use the smaller of $n_1 - 1$ and $n_2 - 1$: $df = 26$.

[16] Because the teacher did not expect one exam to be more difficult prior to examining the test results, she should use a two-sided hypothesis test. H_0: the exams are equally difficult, on average. $\mu_A - \mu_B = 0$. H_A: one exam was more difficult than the other, on average. $\mu_A - \mu_B \neq 0$.

[17] (a) It is probably reasonable to conclude the scores are independent, provided there was no cheating. (b) The summary statistics suggest the data are roughly symmetric about the mean, and the min/max values don't suggest outliers of concern. (c) It seems reasonable to suppose that the samples are independent since the exams were handed out randomly.

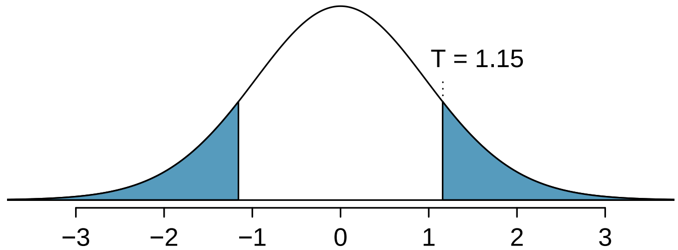


Figure 7.17: The t-distribution with 26 degrees of freedom. The shaded right tail represents values with $T \geq 1.15$. Because it is a two-sided test, we also shade the corresponding lower tail.

EXAMPLE 7.33

Identify the p-value using $df = 26$ and provide a conclusion in the context of the case study.

Using software, we can find the one-tail area (0.13) and then double this value to get the two-tail area, which is the p-value: 0.26. (Alternatively, we could use the t-table in Appendix B.2.) We examine row $df = 26$ in the t-table. In Guided Practice 7.31, we specified that we would use $\alpha = 0.01$. Since the p-value is larger than α, we do not reject the null hypothesis. That is, the data do not convincingly show that one exam version is more difficult than the other, and the teacher should not be convinced that she should add points to the Version B exam scores.

7.3.4 Pooled standard deviation estimate (special topic)

Occasionally, two populations will have standard deviations that are so similar that they can be treated as identical. For example, historical data or a well-understood biological mechanism may justify this strong assumption. In such cases, we can make the t-distribution approach slightly more precise by using a pooled standard deviation.

The **pooled standard deviation** of two groups is a way to use data from both samples to better estimate the standard deviation and standard error. If s_1 and s_2 are the standard deviations of groups 1 and 2 and there are good reasons to believe that the population standard deviations are equal, then we can obtain an improved estimate of the group variances by pooling their data:

$$s^2_{pooled} = \frac{s_1^2 \times (n_1 - 1) + s_2^2 \times (n_2 - 1)}{n_1 + n_2 - 2}$$

where n_1 and n_2 are the sample sizes, as before. To use this new statistic, we substitute s^2_{pooled} in place of s_1^2 and s_2^2 in the standard error formula, and we use an updated formula for the degrees of freedom:

$$df = n_1 + n_2 - 2$$

The benefits of pooling the standard deviation are realized through obtaining a better estimate of the standard deviation for each group and using a larger degrees of freedom parameter for the t-distribution. Both of these changes may permit a more accurate model of the sampling distribution of $\bar{x}_1 - \bar{x}_2$, if the standard deviations of the two groups are equal.

POOL STANDARD DEVIATIONS ONLY AFTER CAREFUL CONSIDERATION

A pooled standard deviation is only appropriate when background research indicates the population standard deviations are nearly equal. When the sample size is large and the condition may be adequately checked with data, the benefits of pooling the standard deviations greatly diminishes.

7.4 Power calculations for a difference of means

Often times in experiment planning, there are two competing considerations:

- We want to collect enough data that we can detect important effects.
- Collecting data can be expensive, and in experiments involving people, there may be some risk to patients.

In this section, we focus on the context of a clinical trial, which is a health-related experiment where the subject are people, and we will determine an appropriate sample size where we can be 80% sure that we would detect any practically important effects.[18]

7.4.1 Going through the motions of a test

We're going to go through the motions of a hypothesis test. This will help us frame our calculations for determining an appropriate sample size for the study.

EXAMPLE 7.34

Suppose a pharmaceutical company has developed a new drug for lowering blood pressure, and they are preparing a clinical trial (experiment) to test the drug's effectiveness. They recruit people who are taking a particular standard blood pressure medication. People in the control group will continue to take their current medication through generic-looking pills to ensure blinding. Write down the hypotheses for a two-sided hypothesis test in this context.

E

Generally, clinical trials use a two-sided alternative hypothesis, so below are suitable hypotheses for this context:

H_0**:** The new drug performs exactly as well as the standard medication.
$\mu_{trmt} - \mu_{ctrl} = 0$.

H_A**:** The new drug's performance differs from the standard medication.
$\mu_{trmt} - \mu_{ctrl} \neq 0$.

EXAMPLE 7.35

The researchers would like to run the clinical trial on patients with systolic blood pressures between 140 and 180 mmHg. Suppose previously published studies suggest that the standard deviation of the patients' blood pressures will be about 12 mmHg and the distribution of patient blood pressures will be approximately symmetric.[19] If we had 100 patients per group, what would be the approximate standard error for $\bar{x}_{trmt} - \bar{x}_{ctrl}$?

E

The standard error is calculated as follows:

$$SE_{\bar{x}_{trmt} - \bar{x}_{ctrl}} = \sqrt{\frac{s_{trmt}^2}{n_{trmt}} + \frac{s_{ctrl}^2}{n_{ctrl}}} = \sqrt{\frac{12^2}{100} + \frac{12^2}{100}} = 1.70$$

This may be an imperfect estimate of $SE_{\bar{x}_{trmt} - \bar{x}_{ctrl}}$, since the standard deviation estimate we used may not be perfectly correct for this group of patients. However, it is sufficient for our purposes.

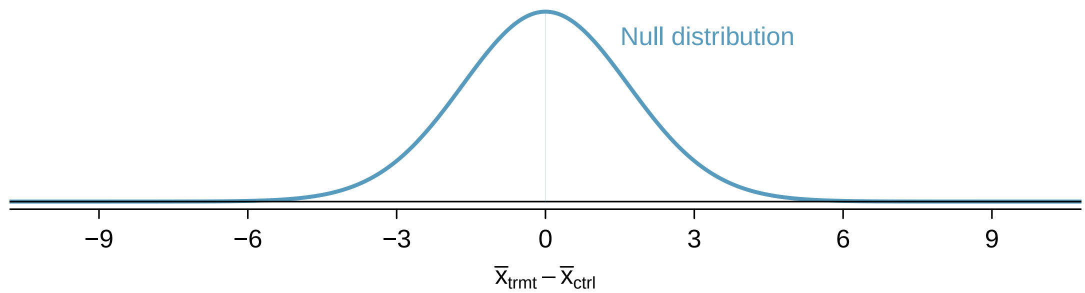
[18] Even though we don't cover it explicitly, similar sample size planning is also helpful for observational studies.

[19] In this particular study, we'd generally measure each patient's blood pressure at the beginning and end of the study, and then the outcome measurement for the study would be the average change in blood pressure. That is, both μ_{trmt} and μ_{ctrl} would represent average differences. This is what you might think of as a 2-sample paired testing structure, and we'd analyze it exactly just like a hypothesis test for a difference in the average change for patients. In the calculations we perform here, we'll suppose that 12 mmHg is the predicted standard deviation of a patient's blood pressure difference over the course of the study.

EXAMPLE 7.36

What does the null distribution of $\bar{x}_{trmt} - \bar{x}_{ctrl}$ look like?

The degrees of freedom are greater than 30, so the distribution of $\bar{x}_{trmt} - \bar{x}_{ctrl}$ will be approximately normal. The standard deviation of this distribution (the standard error) would be about 1.70, and under the null hypothesis, its mean would be 0.



EXAMPLE 7.37

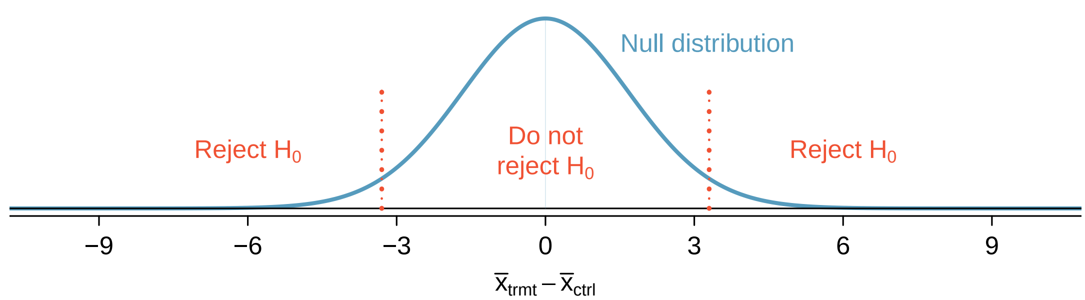
For what values of $\bar{x}_{trmt} - \bar{x}_{ctrl}$ would we reject the null hypothesis?

For $\alpha = 0.05$, we would reject H_0 if the difference is in the lower 2.5% or upper 2.5% tail:

Lower 2.5%: For the normal model, this is 1.96 standard errors below 0, so any difference smaller than $-1.96 \times 1.70 = -3.332$ mmHg.

Upper 2.5%: For the normal model, this is 1.96 standard errors above 0, so any difference larger than $1.96 \times 1.70 = 3.332$ mmHg.

The boundaries of these **rejection regions** are shown below:



Next, we'll perform some hypothetical calculations to determine the probability we reject the null hypothesis, if the alternative hypothesis were actually true.

7.4.2 Computing the power for a 2-sample test

When planning a study, we want to know how likely we are to detect an effect we care about. In other words, if there is a real effect, and that effect is large enough that it has practical value, then what's the probability that we detect that effect? This probability is called the **power**, and we can compute it for different sample sizes or for different *effect sizes*.

We first determine what is a practically significant result. Suppose that the company researchers care about finding any effect on blood pressure that is 3 mmHg or larger vs the standard medication. Here, 3 mmHg is the minimum **effect size** of interest, and we want to know how likely we are to detect this size of an effect in the study.

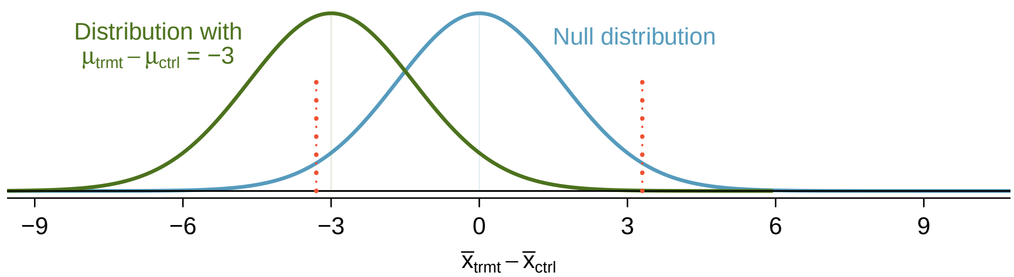
EXAMPLE 7.38

Suppose we decided to move forward with 100 patients per treatment group and the new drug reduces blood pressure by an additional 3 mmHg relative to the standard medication. What is the probability that we detect a drop?

Before we even do any calculations, notice that if $\bar{x}_{trmt} - \bar{x}_{ctrl} = -3$ mmHg, there wouldn't even be sufficient evidence to reject H_0. That's not a good sign.

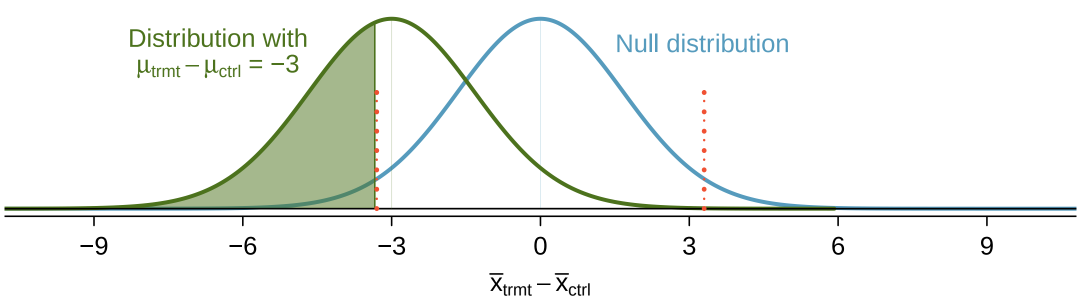
To calculate the probability that we will reject H_0, we need to determine a few things:

- The sampling distribution for $\bar{x}_{trmt} - \bar{x}_{ctrl}$ when the true difference is -3 mmHg. This is the same as the null distribution, except it is shifted to the left by 3:



- The rejection regions, which are outside of the dotted lines above.
- The fraction of the distribution that falls in the rejection region.

In short, we need to calculate the probability that $x < -3.332$ for a normal distribution with mean -3 and standard deviation 1.7. To do so, we first shade the area we want to calculate:



We'll use a normal approximation, which is good approximation when the degrees of freedom is about 30 or more. We'll start by calculating the Z-score and find the tail area using either statistical software or the probability table:

$$Z = \frac{-3.332 - (-3)}{1.7} = -0.20 \qquad \rightarrow \qquad 0.42$$

The power for the test is about 42% when $\mu_{trmt} - \mu_{ctrl} = -3$ and each group has a sample size of 100.

In Example 7.38, we ignored the upper rejection region in the calculation, which was in the opposite direction of the hypothetical truth, i.e. -3. The reasoning? There wouldn't be any value in rejecting the null hypothesis and concluding there was an increase when in fact there was a decrease.

We've also used a normal distribution instead of the t-distribution. This is a convenience, and if the sample size is too small, we'd need to revert back to using the t-distribution. We'll discuss this a bit further at the end of this section.

7.4.3 Determining a proper sample size

In the last example, we found that if we have a sample size of 100 in each group, we can only detect an effect size of 3 mmHg with a probability of about 0.42. Suppose the researchers moved forward and only used 100 patients per group, and the data did not support the alternative hypothesis, i.e. the researchers did not reject H_0. This is a very bad situation to be in for a few reasons:

- In the back of the researchers' minds, they'd all be wondering, *maybe there is a real and meaningful difference, but we weren't able to detect it with such a small sample.*
- The company probably invested hundreds of millions of dollars in developing the new drug, so now they are left with great uncertainty about its potential since the experiment didn't have a great shot at detecting effects that could still be important.
- Patients were subjected to the drug, and we can't even say with much certainty that the drug doesn't help (or harm) patients.
- Another clinical trial may need to be run to get a more conclusive answer as to whether the drug does hold any practical value, and conducting a second clinical trial may take years and many millions of dollars.

We want to avoid this situation, so we need to determine an appropriate sample size to ensure we can be pretty confident that we'll detect any effects that are practically important. As mentioned earlier, a change of 3 mmHg was deemed to be the minimum difference that was practically important. As a first step, we could calculate power for several different sample sizes. For instance, let's try 500 patients per group.

G

GUIDED PRACTICE 7.39

Calculate the power to detect a change of -3 mmHg when using a sample size of 500 per group.[20]

(a) Determine the standard error (recall that the standard deviation for patients was expected to be about 12 mmHg).

(b) Identify the null distribution and rejection regions.

(c) Identify the alternative distribution when $\mu_{trmt} - \mu_{ctrl} = -3$.

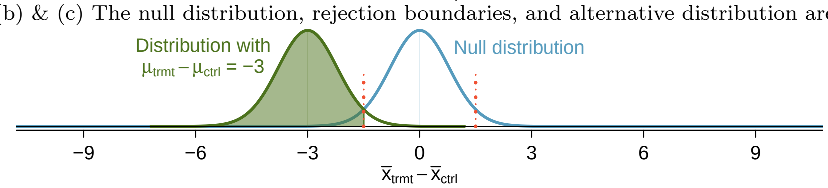
(d) Compute the probability we reject the null hypothesis.

The researchers decided 3 mmHg was the minimum difference that was practically important, and with a sample size of 500, we can be very certain (97.7% or better) that we will detect any such difference. We now have moved to another extreme where we are exposing an unnecessary number of patients to the new drug in the clinical trial. Not only is this ethically questionable, but it would also cost a lot more money than is necessary to be quite sure we'd detect any important effects.

The most common practice is to identify the sample size where the power is around 80%, and sometimes 90%. Other values may be reasonable for a specific context, but 80% and 90% are most commonly targeted as a good balance between high power and not exposing too many patients to a new treatment (or wasting too much money).

We could compute the power of the test at several other possible sample sizes until we find one that's close to 80%, but there's a better way. We should solve the problem backwards.

[20](a) The standard error is given as $SE = \sqrt{\frac{12^2}{500} + \frac{12^2}{500}} = 0.76$.
(b) & (c) The null distribution, rejection boundaries, and alternative distribution are shown below:



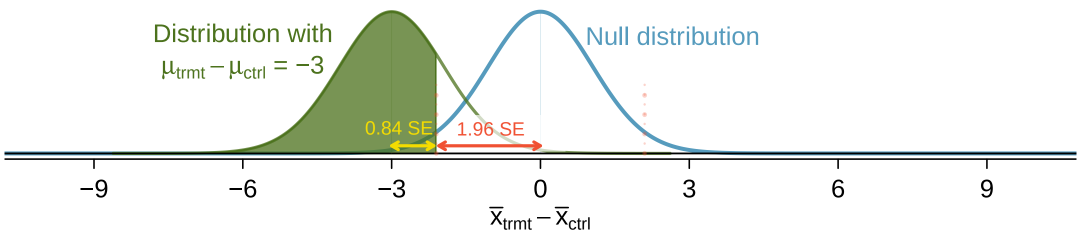
The rejection regions are the areas on the outside of the two dotted lines and are at $\pm 0.76 \times 1.96 = \pm 1.49$.
(d) The area of the alternative distribution where $\mu_{trmt} - \mu_{ctrl} = -3$ has been shaded. We compute the Z-score and find the tail area: $Z = \frac{-1.49-(-3)}{0.76} = 1.99 \rightarrow 0.977$. With 500 patients per group, we would be about 97.7% sure (or more) that we'd detect any effects that are at least 3 mmHg in size.

EXAMPLE 7.40

What sample size will lead to a power of 80%?

We'll assume we have a large enough sample that the normal distribution is a good approximation for the test statistic, since the normal distribution and the t-distribution look almost identical when the degrees of freedom are moderately large (e.g. $df \geq 30$). If that doesn't turn out to be true, then we'd need to make a correction.

We start by identifying the Z-score that would give us a lower tail of 80%. For a moderately large sample size per group, the Z-score for a lower tail of 80% would be about $Z = 0.84$.



E

Additionally, the rejection region extends $1.96 \times SE$ from the center of the null distribution for $\alpha = 0.05$. This allows us to calculate the target distance between the center of the null and alternative distributions in terms of the standard error:

$$0.84 \times SE + 1.96 \times SE = 2.8 \times SE$$

In our example, we want the distance between the null and alternative distributions' centers to equal the minimum effect size of interest, 3 mmHg, which allows us to set up an equation between this difference and the standard error:

$$3 = 2.8 \times SE$$

$$3 = 2.8 \times \sqrt{\frac{12^2}{n} + \frac{12^2}{n}}$$

$$n = \frac{2.8^2}{3^2} \times \left(12^2 + 12^2\right) = 250.88$$

We should target 251 patients per group in order to achieve 80% power at the 0.05 significance level for this context.

The standard error difference of $2.8 \times SE$ is specific to a context where the targeted power is 80% and the significance level is $\alpha = 0.05$. If the targeted power is 90% or if we use a different significance level, then we'll use something a little different than $2.8 \times SE$.

Had the suggested sample size been relatively small – roughly 30 or smaller – it would have been a good idea to rework the calculations using the degrees of fredom for the smaller sample size under that initial sample size. That is, we would have revised the 0.84 and 1.96 values based on degrees of freedom implied by the initial sample size. The revised sample size target would generally have then been a little larger.

GUIDED PRACTICE 7.41

G

Suppose the targeted power was 90% and we were using $\alpha = 0.01$. How many standard errors should separate the centers of the null and alternative distribution, where the alternative distribution is centered at the minimum effect size of interest?[21]

---

[21]First, find the Z-score such that 90% of the distribution is below it: $Z = 1.28$. Next, find the cutoffs for the rejection regions: ± 2.58. Then the difference in centers should be about $1.28 \times SE + 2.58 \times SE = 3.86 \times SE$.

GUIDED PRACTICE 7.42

What are some considerations that are important in determining what the power should be for an experiment?[22]

Figure 7.18 shows the power for sample sizes from 20 patients to 5,000 patients when $\alpha = 0.05$ and the true difference is -3. This curve was constructed by writing a program to compute the power for many different sample sizes.

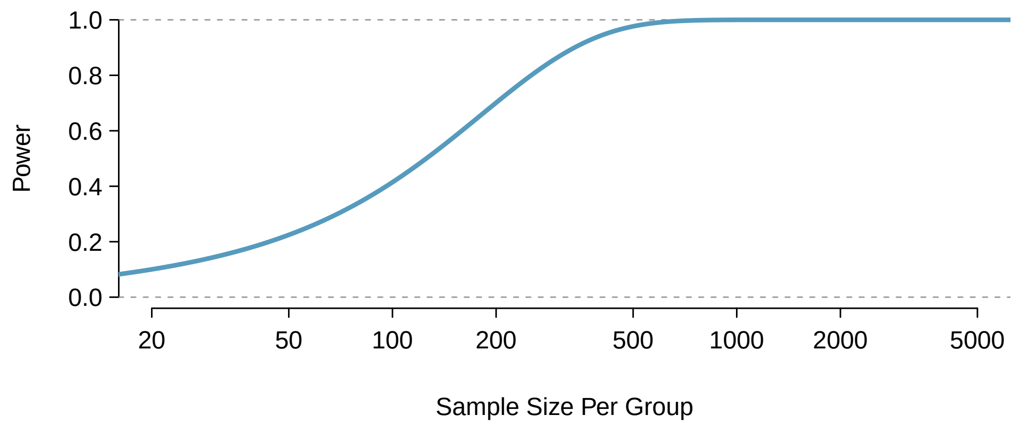


Figure 7.18: The curve shows the power for different sample sizes in the context of the blood pressure example when the true difference is -3. Having more than about 250 to 350 observations doesn't provide much additional value in detecting an effect when $\alpha = 0.05$.

Power calculations for expensive or risky experiments are critical. However, what about experiments that are inexpensive and where the ethical considerations are minimal? For example, if we are doing final testing on a new feature on a popular website, how would our sample size considerations change? As before, we'd want to make sure the sample is big enough. However, suppose the feature has undergone some testing and is known to perform well (e.g. the website's users seem to enjoy the feature). Then it may be reasonable to run a larger experiment if there's value from having a more precise estimate of the feature's effect, such as helping guide the development of the next useful feature.

[22]Answers will vary, but here are a few important considerations:

- Whether there is any risk to patients in the study.
- The cost of enrolling more patients.
- The potential downside of not detecting an effect of interest.

7.5 Comparing many means with ANOVA

Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons; for example, if there were three groups, we might be tempted to compare the first mean with the second, then with the third, and then finally compare the second and third means for a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually find a difference just by chance, even if there is no difference in the populations.

In this section, we will learn a new method called **analysis of variance (ANOVA)** and a new test statistic called F. ANOVA uses a single hypothesis test to check whether the means across many groups are equal:

H_0: The mean outcome is the same across all groups. In statistical notation, $\mu_1 = \mu_2 = \cdots = \mu_k$ where μ_i represents the mean of the outcome for observations in category i.

H_A: At least one mean is different.

Generally we must check three conditions on the data before performing ANOVA:

- the observations are independent within and across groups,
- the data within each group are nearly normal, and
- the variability across the groups is about equal.

When these three conditions are met, we may perform an ANOVA to determine whether the data provide strong evidence against the null hypothesis that all the μ_i are equal.

EXAMPLE 7.43

College departments commonly run multiple lectures of the same introductory course each semester because of high demand. Consider a statistics department that runs three lectures of an introductory statistics course. We might like to determine whether there are statistically significant differences in first exam scores in these three classes (A, B, and C). Describe appropriate hypotheses to determine whether there are any differences between the three classes.

E

The hypotheses may be written in the following form:

H_0: The average score is identical in all lectures. Any observed difference is due to chance. Notationally, we write $\mu_A = \mu_B = \mu_C$.

H_A: The average score varies by class. We would reject the null hypothesis in favor of the alternative hypothesis if there were larger differences among the class averages than what we might expect from chance alone.

Strong evidence favoring the alternative hypothesis in ANOVA is described by unusually large differences among the group means. We will soon learn that assessing the variability of the group means relative to the variability among individual observations within each group is key to ANOVA's success.

EXAMPLE 7.44

Examine Figure 7.19. Compare groups I, II, and III. Can you visually determine if the differences in the group centers is due to chance or not? Now compare groups IV, V, and VI. Do these differences appear to be due to chance?

E

Any real difference in the means of groups I, II, and III is difficult to discern, because the data within each group are very volatile relative to any differences in the average outcome. On the other hand, it appears there are differences in the centers of groups IV, V, and VI. For instance, group V appears to have a higher mean than that of the other two groups. Investigating groups IV, V, and VI, we see the differences in the groups' centers are noticeable because those differences are large *relative to the variability in the individual observations within each group*.

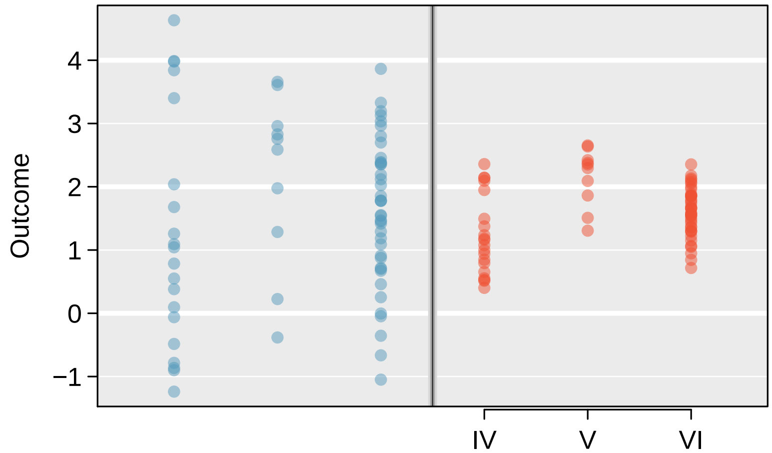


Figure 7.19: Side-by-side dot plot for the outcomes for six groups.

7.5.1 Is batting performance related to player position in MLB?

We would like to discern whether there are real differences between the batting performance of baseball players according to their position: outfielder (`OF`), infielder (`IF`), and catcher (`C`). We will use a data set called `bat18`, which includes batting records of 429 Major League Baseball (MLB) players from the 2018 season who had at least 100 at bats. Six of the 429 cases represented in `bat18` are shown in Figure 7.20, and descriptions for each variable are provided in Figure 7.21. The measure we will use for the player batting performance (the outcome variable) is on-base percentage (`OBP`). The on-base percentage roughly represents the fraction of the time a player successfully gets on base or hits a home run.

	name	team	position	AB	H	HR	RBI	AVG	OBP
1	Abreu, J	CWS	IF	499	132	22	78	0.265	0.325
2	Acuna Jr., R	ATL	OF	433	127	26	64	0.293	0.366
3	Adames, W	TB	IF	288	80	10	34	0.278	0.348
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
427	Zimmerman, R	WSH	IF	288	76	13	51	0.264	0.337
428	Zobrist, B	CHC	IF	455	139	9	58	0.305	0.378
429	Zunino, M	SEA	C	373	75	20	44	0.201	0.259

Figure 7.20: Six cases from the `bat18` data matrix.

variable	**description**
`name`	Player name
`team`	The abbreviated name of the player's team
`position`	The player's primary field position (`OF`, `IF`, `C`)
`AB`	Number of opportunities at bat
`H`	Number of hits
`HR`	Number of home runs
`RBI`	Number of runs batted in
`AVG`	Batting average, which is equal to `H/AB`
`OBP`	On-base percentage, which is roughly equal to the fraction of times a player gets on base or hits a home run

Figure 7.21: Variables and their descriptions for the `bat18` data set.

GUIDED PRACTICE 7.45

The null hypothesis under consideration is the following: $\mu_{\texttt{OF}} = \mu_{\texttt{IF}} = \mu_{\texttt{C}}$. Write the null and corresponding alternative hypotheses in plain language.[23]

EXAMPLE 7.46

The player positions have been divided into four groups: outfield (`OF`), infield (`IF`), and catcher (`C`). What would be an appropriate point estimate of the on-base percentage by outfielders, $\mu_{\texttt{OF}}$?

A good estimate of the on-base percentage by outfielders would be the sample average of `OBP` for just those players whose position is outfield: $\bar{x}_{OF} = 0.320$.

Figure 7.22 provides summary statistics for each group. A side-by-side box plot for the on-base percentage is shown in Figure 7.23. Notice that the variability appears to be approximately constant across groups; nearly constant variance across groups is an important assumption that must be satisfied before we consider the ANOVA approach.

	OF	IF	C
Sample size (n_i)	160	205	64
Sample mean ($\bar{x}_i$)	0.320	0.318	0.302
Sample SD (s_i)	0.043	0.038	0.038

Figure 7.22: Summary statistics of on-base percentage, split by player position.

EXAMPLE 7.47

The largest difference between the sample means is between the designated hitter and the outfielder positions. Consider again the original hypotheses:

H_0: $\mu_{\texttt{OF}} = \mu_{\texttt{IF}} = \mu_{\texttt{C}}$

H_A: The average on-base percentage (μ_i) varies across some (or all) groups.

Why might it be inappropriate to run the test by simply estimating whether the difference of $\mu_{\texttt{C}}$ and $\mu_{\texttt{OF}}$ is statistically significant at a 0.05 significance level?

The primary issue here is that we are inspecting the data before picking the groups that will be compared. It is inappropriate to examine all data by eye (informal testing) and only afterwards decide which parts to formally test. This is called **data snooping** or **data fishing**. Naturally we would pick the groups with the large differences for the formal test, leading to an inflation in the Type 1 Error rate. To understand this better, let's consider a slightly different problem.

Suppose we are to measure the aptitude for students in 20 classes in a large elementary school at the beginning of the year. In this school, all students are randomly assigned to classrooms, so any differences we observe between the classes at the start of the year are completely due to chance. However, with so many groups, we will probably observe a few groups that look rather different from each other. If we select only these classes that look so different, we will probably make the wrong conclusion that the assignment wasn't random. While we might only formally test differences for a few pairs of classes, we informally evaluated the other classes by eye before choosing the most extreme cases for a comparison.

For additional information on the ideas expressed in Example 7.47, we recommend reading about the **prosecutor's fallacy**.[24]

In the next section we will learn how to use the F statistic and ANOVA to test whether observed differences in sample means could have happened just by chance even if there was no difference in the respective population means.

---

[23] H_0: The average on-base percentage is equal across the four positions. H_A: The average on-base percentage varies across some (or all) groups.

[24] See, for example, andrewgelman.com/2007/05/18/the_prosecutors.

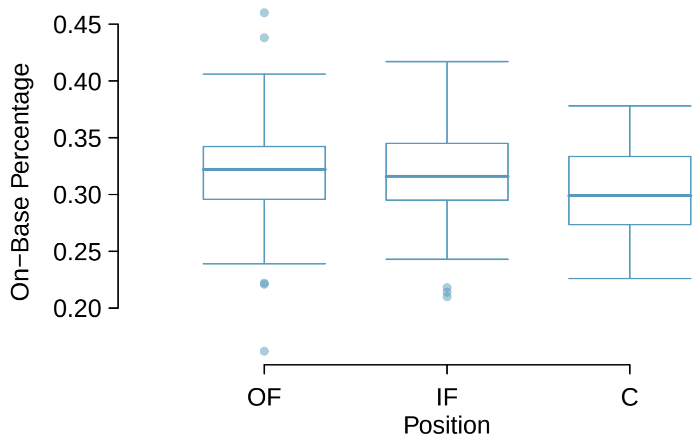


Figure 7.23: Side-by-side box plot of the on-base percentage for 429 players across four groups. There is one prominent outlier visible in the infield group, but with 154 observations in the infield group, this outlier is not a concern.

7.5.2 Analysis of variance (ANOVA) and the F test

The method of analysis of variance in this context focuses on answering one question: is the variability in the sample means so large that it seems unlikely to be from chance alone? This question is different from earlier testing procedures since we will *simultaneously* consider many groups, and evaluate whether their sample means differ more than we would expect from natural variation. We call this variability the **mean square between groups (MSG)**, and it has an associated degrees of freedom, $df_G = k - 1$ when there are k groups. The MSG can be thought of as a scaled variance formula for means. If the null hypothesis is true, any variation in the sample means is due to chance and shouldn't be too large. Details of MSG calculations are provided in the footnote.[25] However, we typically use software for these computations.

The mean square between the groups is, on its own, quite useless in a hypothesis test. We need a benchmark value for how much variability should be expected among the sample means if the null hypothesis is true. To this end, we compute a pooled variance estimate, often abbreviated as the **mean square error (MSE)**, which has an associated degrees of freedom value $df_E = n - k$. It is helpful to think of MSE as a measure of the variability within the groups. Details of the computations of the MSE and a link to an extra online section for ANOVA calculations are provided in the footnote[26] for interested readers.

When the null hypothesis is true, any differences among the sample means are only due to chance, and the MSG and MSE should be about equal. As a test statistic for ANOVA, we examine

[25]Let $\bar{x}$ represent the mean of outcomes across all groups. Then the mean square between groups is computed as

$$MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}\right)^2$$

where SSG is called the **sum of squares between groups** and n_i is the sample size of group i.

[26]Let $\bar{x}$ represent the mean of outcomes across all groups. Then the **sum of squares total (SST)** is computed as

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where the sum is over all observations in the data set. Then we compute the **sum of squared errors (SSE)** in one of two equivalent ways:

$$\begin{aligned} SSE &= SST - SSG \\ &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 \end{aligned}$$

where s_i^2 is the sample variance (square of the standard deviation) of the residuals in group i. Then the MSE is the standardized form of SSE: $MSE = \frac{1}{df_E} SSE$.

For additional details on ANOVA calculations, see www.openintro.org/d?file=stat_extra_anova_calculations the fraction of MSG and MSE:

$$F = \frac{MSG}{MSE}$$

The MSG represents a measure of the between-group variability, and MSE measures the variability within each of the groups.

GUIDED PRACTICE 7.48

For the baseball data, $MSG = 0.00803$ and $MSE = 0.00158$. Identify the degrees of freedom associated with MSG and MSE and verify the F statistic is approximately 5.077.[27]

We can use the F statistic to evaluate the hypotheses in what is called an **F test**. A p-value can be computed from the F statistic using an F distribution, which has two associated parameters: df_1 and df_2. For the F statistic in ANOVA, $df_1 = df_G$ and $df_2 = df_E$. An F distribution with 2 and 426 degrees of freedom, corresponding to the F statistic for the baseball hypothesis test, is shown in Figure 7.24.

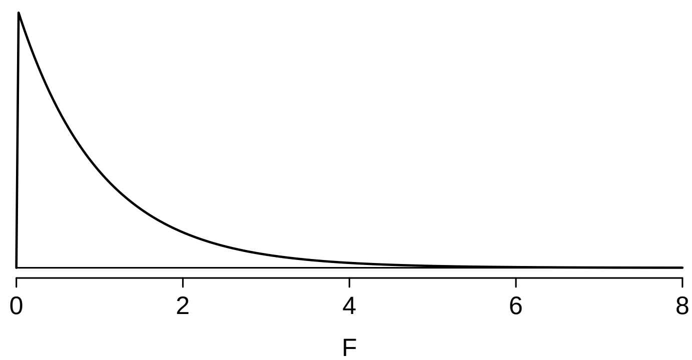


Figure 7.24: An F distribution with $df_1 = 3$ and $df_2 = 323$.

The larger the observed variability in the sample means (MSG) relative to the within-group observations (MSE), the larger F will be and the stronger the evidence against the null hypothesis. Because larger values of F represent stronger evidence against the null hypothesis, we use the upper tail of the distribution to compute a p-value.

THE F STATISTIC AND THE F TEST

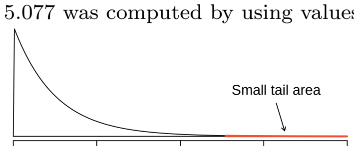
Analysis of variance (ANOVA) is used to test whether the mean outcome differs across 2 or more groups. ANOVA uses a test statistic F, which represents a standardized ratio of variability in the sample means relative to the variability within the groups. If H_0 is true and the model conditions are satisfied, the statistic F follows an F distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$. The upper tail of the F distribution is used to represent the p-value.

GUIDED PRACTICE 7.49

The test statistic for the baseball example is $F = 5.077$. Shade the area corresponding to the p-value in Figure 7.24. [28]

[27] There are $k = 3$ groups, so $df_G = k - 1 = 2$. There are $n = n_1 + n_2 + n_3 = 429$ total observations, so $df_E = n - k = 426$. Then the F statistic is computed as the ratio of MSG and MSE: $F = \frac{MSG}{MSE} = \frac{0.00803}{0.00158} = 5.082 \approx 5.077$. ($F = 5.077$ was computed by using values for MSG and MSE that were not rounded.)

[28]



EXAMPLE 7.50

The p-value corresponding to the shaded area in the solution of Guided Practice 7.49 is equal to about 0.0066. Does this provide strong evidence against the null hypothesis?

E

The p-value is smaller than 0.05, indicating the evidence is strong enough to reject the null hypothesis at a significance level of 0.05. That is, the data provide strong evidence that the average on-base percentage varies by player's primary field position.

7.5.3 Reading an ANOVA table from software

The calculations required to perform an ANOVA by hand are tedious and prone to human error. For these reasons, it is common to use statistical software to calculate the F statistic and p-value.

An ANOVA can be summarized in a table very similar to that of a regression summary, which we will see in Chapters **??** and **??**. Figure 7.25 shows an ANOVA summary to test whether the mean of on-base percentage varies by player positions in the MLB. Many of these values should look familiar; in particular, the F test statistic and p-value can be retrieved from the last columns.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
position	2	0.0161	0.0080	5.0766	0.0066
Residuals	426	0.6740	0.0016		
			$s_{pooled} = 0.040$ on $df = 423$		

Figure 7.25: ANOVA summary for testing whether the average on-base percentage differs across player positions.

7.5.4 Graphical diagnostics for an ANOVA analysis

There are three conditions we must check for an ANOVA analysis: all observations must be independent, the data in each group must be nearly normal, and the variance within each group must be approximately equal.

Independence. If the data are a simple random sample, this condition is satisfied. For processes and experiments, carefully consider whether the data may be independent (e.g. no pairing). For example, in the MLB data, the data were not sampled. However, there are not obvious reasons why independence would not hold for most or all observations.

Approximately normal. As with one- and two-sample testing for means, the normality assumption is especially important when the sample size is quite small when it is ironically difficult to check for non-normality. A histogram of the observations from each group is shown in Figure 7.26. Since each of the groups we're considering have relatively large sample sizes, what we're looking for are major outliers. None are apparent, so this conditions is reasonably met.

Constant variance. The last assumption is that the variance in the groups is about equal from one group to the next. This assumption can be checked by examining a side-by-side box plot of the outcomes across the groups, as in Figure 7.23 on page 126. In this case, the variability is similar in the four groups but not identical. We see in Table 7.22 on page 125 that the standard deviation doesn't vary much from one group to the next.

DIAGNOSTICS FOR AN ANOVA ANALYSIS

Independence is always important to an ANOVA analysis. The normality condition is very important when the sample sizes for each group are relatively small. The constant variance condition is especially important when the sample sizes differ between groups.

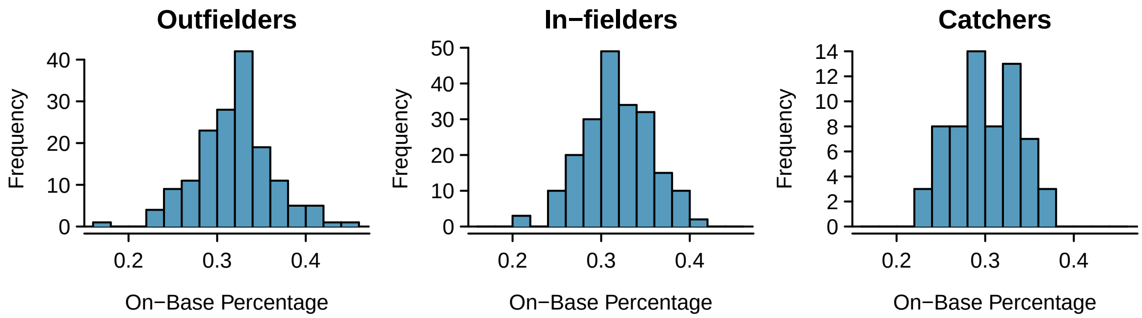


Figure 7.26: Histograms of OBP for each field position.

Class i	A	B	C
n_i	58	55	51
$\bar{x}_i$	75.1	72.0	78.9
s_i	13.9	13.8	13.1

Figure 7.27: Summary statistics for the first midterm scores in three different lectures of the same course.

7.5.5 Multiple comparisons and controlling Type 1 Error rate

When we reject the null hypothesis in an ANOVA analysis, we might wonder, which of these groups have different means? To answer this question, we compare the means of each possible pair of groups. For instance, if there are three groups and there is strong evidence that there are some differences in the group means, there are three comparisons to make: group 1 to group 2, group 1 to group 3, and group 2 to group 3. These comparisons can be accomplished using a two-sample t-test, but we use a modified significance level and a pooled estimate of the standard deviation across groups. Usually this pooled standard deviation can be found in the ANOVA table, e.g. along the bottom of Figure 7.25.

EXAMPLE 7.51

Example 7.43 on page 123 discussed three statistics lectures, all taught during the same semester. Figure 7.27 shows summary statistics for these three courses, and a side-by-side box plot of the data is shown in Figure 7.28. We would like to conduct an ANOVA for these data. Do you see any deviations from the three conditions for ANOVA?

E

In this case (like many others) it is difficult to check independence in a rigorous way. Instead, the best we can do is use common sense to consider reasons the assumption of independence may not hold. For instance, the independence assumption may not be reasonable if there is a star teaching assistant that only half of the students may access; such a scenario would divide a class into two subgroups. No such situations were evident for these particular data, and we believe that independence is acceptable.

The distributions in the side-by-side box plot appear to be roughly symmetric and show no noticeable outliers.

The box plots show approximately equal variability, which can be verified in Figure 7.27, supporting the constant variance assumption.

GUIDED PRACTICE 7.52

G

An ANOVA was conducted for the midterm data, and summary results are shown in Figure 7.29. What should we conclude?[29]

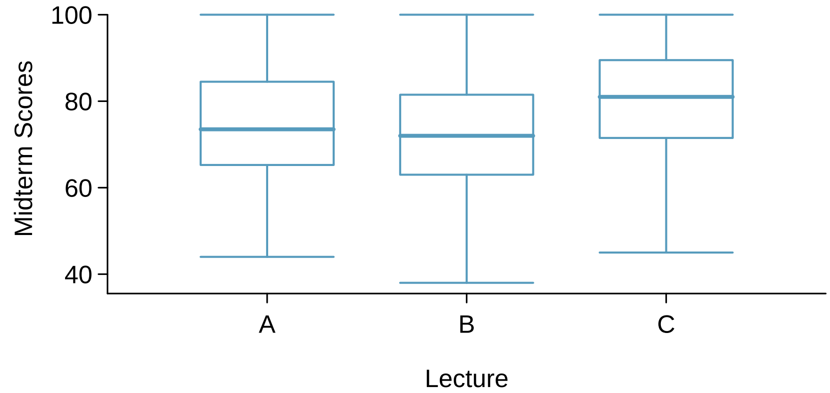


Figure 7.28: Side-by-side box plot for the first midterm scores in three different lectures of the same course.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lecture	2	1290.11	645.06	3.48	0.0330
Residuals	161	29810.13	185.16		
			$s_{pooled} = 13.61$ on $df = 161$		

Figure 7.29: ANOVA summary table for the midterm data.

There is strong evidence that the different means in each of the three classes is not simply due to chance. We might wonder, which of the classes are actually different? As discussed in earlier chapters, a two-sample t-test could be used to test for differences in each possible pair of groups. However, one pitfall was discussed in Example 7.47 on page 125: when we run so many tests, the Type 1 Error rate increases. This issue is resolved by using a modified significance level.

MULTIPLE COMPARISONS AND THE BONFERRONI CORRECTION FOR α

The scenario of testing many pairs of groups is called **multiple comparisons**. The **Bonferroni correction** suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^* = \alpha / K$$

where K is the number of comparisons being considered (formally or informally). If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$.

[29] The p-value of the test is 0.0330, less than the default significance level of 0.05. Therefore, we reject the null hypothesis and conclude that the difference in the average midterm scores are not due to chance.

EXAMPLE 7.53

In Guided Practice 7.52, you found strong evidence of differences in the average midterm grades between the three lectures. Complete the three possible pairwise comparisons using the Bonferroni correction and report any differences.

We use a modified significance level of $\alpha^* = 0.05/3 = 0.0167$. Additionally, we use the pooled estimate of the standard deviation: $s_{pooled} = 13.61$ on $df = 161$, which is provided in the ANOVA summary table.

Lecture A versus Lecture B: The estimated difference and standard error are, respectively,

$$\bar{x}_A - \bar{x}_B = 75.1 - 72 = 3.1 \qquad\qquad SE = \sqrt{\frac{13.61^2}{58} + \frac{13.61^2}{55}} = 2.56$$

(See Section 7.3.4 on page 116 for additional details.) This results in a T-score of 1.21 on $df = 161$ (we use the df associated with s_{pooled}). Statistical software was used to precisely identify the two-sided p-value since the modified significance level of 0.0167 is not found in the t-table. The p-value (0.228) is larger than $\alpha^* = 0.0167$, so there is not strong evidence of a difference in the means of lectures A and B.

Lecture A versus Lecture C: The estimated difference and standard error are 3.8 and 2.61, respectively. This results in a T score of 1.46 on $df = 161$ and a two-sided p-value of 0.1462. This p-value is larger than α^*, so there is not strong evidence of a difference in the means of lectures A and C.

Lecture B versus Lecture C: The estimated difference and standard error are 6.9 and 2.65, respectively. This results in a T score of 2.60 on $df = 161$ and a two-sided p-value of 0.0102. This p-value is smaller than α^*. Here we find strong evidence of a difference in the means of lectures B and C.

We might summarize the findings of the analysis from Example 7.53 using the following notation:

$$\mu_A \stackrel{?}{=} \mu_B \qquad\qquad \mu_A \stackrel{?}{=} \mu_C \qquad\qquad \mu_B \neq \mu_C$$

The midterm mean in lecture A is not statistically distinguishable from those of lectures B or C. However, there is strong evidence that lectures B and C are different. In the first two pairwise comparisons, we did not have sufficient evidence to reject the null hypothesis. Recall that failing to reject H_0 does not imply H_0 is true.

REJECT H_0 WITH ANOVA BUT FIND NO GROUP DIFFERENCES

It is possible to reject the null hypothesis using ANOVA and then to not subsequently identify differences in the pairwise comparisons. However, *this does not invalidate the ANOVA conclusion.* It only means we have not been able to successfully identify which specific groups differ in their means.

The ANOVA procedure examines the big picture: it considers all groups simultaneously to decipher whether there is evidence that some difference exists. Even if the test indicates that there is strong evidence of differences in group means, identifying with high confidence a specific difference as statistically significant is more difficult.

Consider the following analogy: we observe a Wall Street firm that makes large quantities of money based on predicting mergers. Mergers are generally difficult to predict, and if the prediction success rate is extremely high, that may be considered sufficiently strong evidence to warrant investigation by the Securities and Exchange Commission (SEC). While the SEC may be quite certain that there is insider trading taking place at the firm, the evidence against any single trader may not be very strong. It is only when the SEC considers all the data that they identify the pattern. This is effectively the strategy of ANOVA: stand back and consider all the groups simultaneously.

Appendix A

Data sets within the text

Each data set within the text is described in this appendix, and there is a corresponding page for each of these data sets at **openintro.org/data**. This page also includes additional data sets that can be used for honing your skills. Each data set has its own page with the following information:

- Description of each data set.
- Detailed overview of each data set's variables.
- CSV download.
- R object file download.

Over time we will also expand the information available on these pages.

[Redirects must be created for each link in this appendix.]

A.1 ??

?? [`stent30`, `stent365`] The stent data is split across two data sets, one for the 0-30 day and one for the 0-365 day results.
Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. New England Journal of Medicine 365:993-1003. www.nejm.org/doi/full/10.1056/NEJMoa1105335.
NY Times article: www.nytimes.com/2011/09/08/health/research/08stent.html.

?? [`loan50`, `loans_full_schema`] This data comes from Lending Club (lendingclub.com), which provides a large set of data on the people who received loans through their platform. The data used in the textbook comes from a sample of the loans made in Q1 (Jan, Feb, March) 2018.

?? [`county`, `county_complete`] These data come from several government sources. For those variables included in the county data set, only the most recent data is reported, as of what was available in late 2018. Data prior to 2011 is all from census.gov, where the specific Quick Facts page providing the data is no longer available. The more recent data comes from USDA (ers.usda.gov), Bureau of Labor Statistics (bls.gov/lau), SAIPE (census.gov/did/www/saipe), and American Community Survey (census.gov/programs-surveys/acs).

?? No data sets were described in this section.

?? The Nurses' Health Study was mentioned. For more information on this data set, see www.channing.harvard.edu/nhs

?? The study we had in mind when discussing the simple randomization (no blocking) study was Anturane Reinfarction Trial Research Group. 1980. *Sulfinpyrazone in the prevention of sudden death after myocardial infarction.* New England Journal of Medicine 302(5):250-256.

A.2 ??

?? [loan50, county] These data sets are described in Data Appendix A.1.

?? [loan50, county] These data sets are described in Data Appendix A.1.

?? [malaria] Lyke et al. 2017. PfSPZ vaccine induces strain-transcending T cells and durable protection against heterologous controlled human malaria infection. PNAS 114(10):2711-2716. http://www.pnas.org/content/114/10/2711

A.3 ??

?? [loan50, county] These data sets are described in Data Appendix A.1.

?? [playing_cards] A table describing the 52 cards in a standard deck.

?? [family_college] A simulated data set based on real population summaries at nces.ed.gov/pubs2001/2001126.pdf.

?? [smallpox] Fenner F. 1988. Smallpox and Its Eradication (History of International Public Health, No. 6). Geneva: World Health Organization. ISBN 92-4-156110-6.

?? [Mammogram screening, probabilities.] The probabilities reported were obtained using studies reported at www.breastcancer.org and www.ncbi.nlm.nih.gov/pmc/articles/PMC1173421.

?? [Jose campus visits, probabilities, no data link] This example was entirely made up.

?? No data sets were described in this section.

?? [Course material purchases, probabilities, no data link] The probabilities for course materials purchases were created as hypotheticals. [Should this even be mentioned?]

?? [Auctions for TV and toaster, no data link] The statistics used were created as hypotheticals. [Should this even be mentioned?]

?? [stocks_18] Monthly returns for Caterpillar, Exxon Mobil Corp, and Google for November 2015 to October 2018.

?? [fcid] This sample can be considered a simple random sample from the US population. It relies on the USDA Food Commodity Intake Database.

A.4 Distributions of random variables

4.1 [SAT and ACT score distributions] The SAT score data comes from the 2018 distribution, which is provided at
reports.collegeboard.org/pdf/2018-total-group-sat-suite-assessments-annual-report.pdf
The ACT score data is available at
act.org/content/dam/act/unsecured/documents/cccr2018/P_99_999999_N_S_N00_ACT-GCPR_National.pdf
We also acknowledge that the actual ACT score distribution is *not* nearly normal. However, since the topic is very accessible, we decided to keep the context and examples.

4.1 [Male heights] The distribution is based on the USDA Food Commodity Intake Database.

4.1 [possum] The distribution parameters are based on a sample of possums from Australia and New Guinea. The original source of this data is as follows. Lindenmayer DB, et al. 1995. *Morphological variation among columns of the mountain brushtail possum, Trichosurus caninus Ogilby (Phalangeridae: Marsupiala).* Australian Journal of Zoology 43: 449-458.

4.2 [Exceeding insurance deductible] These statistics were made up but are possible values one might observe for low-deductible plans.

4.3 [Exceeding insurance deductible] These statistics were made up but are possible values one might observe for low-deductible plans.

4.3 [Smoking friends] Unfortunately, we don't currently have additional information on the source for the 30% statistic, so don't consider this one as fact since we cannot verify it was from a reputable source.

4.3 [US smoking rate] The 15% smoking rate in the US figure is close to the value from the Centers for Disease Control and Prevention website, which reports a value of 14% as of the 2017 estimate:
cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm

4.4 [Football kicker] This example was made up.

4.4 [Heart attack admissions] This example was made up, though the heart attack admissions are realistic for some hospitals.

4.5 [ami_occurrences] This is a simulated data set but resembles actual AMI data for New York City based on typical AMI incidence rates.

A.5 Foundations for inference

5.1 [pew_energy_2018] The actual data has more observations than were referenced in this chapter. That is, we used a subsample since it helped smooth some of the examples to have a bit more variability. The `pew_energy_2018` data set represents the full data set for each of the different energy source questions, which covers solar, wind, offshore drilling, hydrolic fracturing, and nuclear energy. The statistics used to construct the data are from the following page:

www.pewinternet.org/2018/05/14/majorities-see-government-efforts-to-protect-the-environment-as-insufficient/

5.2 [pew_energy_2018] See the details for this data set above in the Section 5.1 data section.

5.2 [ebola_survey] 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. Poll ID NY141026 on maristpoll.marist.edu.

5.3 [pew_energy_2018] See the details for this data set above in the Section 5.1 data section.

5.3 [Rosling questions] We noted much smaller samples than the Roslings' describe in their book, Factfulness, The samples we describe are similar but not the same as the actual rates. The approximate rates for the correct answers for the two questions for (sometimes different) populations discussed in the book, as reported in Factfulness, are

- 80% of the world's 1 year olds have been vaccinated against some disease: 13% get this correct (17% in the US). gapm.io/q9
- Number of children in the world in 2100: 9% correct. gapm.io/q5

Here are a few more questions and a rough percent of people who get them correct:

- In all low-income countries across the world today, how many girls finish primary school: 20%, 40%, or 60%? Answer: 60%. About 7% of people get this question correct. gapm.io/q1
- What is the life expectancy of the world today: 50 years, 60 years, or 70 years? Answer: 70 years. In the US, about 43% of people get this question correct. gapm.io/q4
- In 1996, tigers, giant pandas, and black rhinos were all listed as endangered. How many of these three species are more critically endangered today: two of them, one of them, none of them? Answer: none of them. About 7% of people get this question correct. gapm.io/q11

- How many people in the world have some access to electricity? 20%, 50%, 80%. Answer: 80%. About 22% of people get this correct. gapm.io/q12

For more information, check out the book, Factfulness.

5.3 [nuclear_survey] A simple random sample of 1,028 US adults in March 2013 found that 56% of US adults support nuclear arms reduction. www.gallup.com/poll/161198/favor-russian-nuclear-arms-reductions.aspx

5.3 [stent30, stent365] This data is described in Data Appendix A.1.

A.6 Inference for categorical data

6.1 [Payday loans] The statistics come from the following source:
pewtrusts.org/-/media/assets/2017/04/payday-loan-customers-want-more-protections-methodology.pdf

6.1 [Tire factory] The statistics were created for the purposes of performing sample size calculations.

6.2 [cpr] Böttiger et al. *Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial.* The Lancet, 2001.

6.2 [fish_oil_18] Manson JE, et al. 2018. Marine n-3 Fatty Acids and Prevention of Cardiovascular Disease and Cancer. NEJMoa1811403.

6.2 [mammogram] Miller AB. 2014. *Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial.* BMJ 2014;348:g366.

6.2 [drone_blades] The quality control data set for quadcopter drone blades is a made-up data set for an example. We provide the simulated data in the `drone_blades` data set.

6.3 [jury] The jury data set for examining discrimination is a made-up data set an example. We provide the simulated data in the `jury` data set.

6.3 [sp500_1950_2018] Data is sourced from finance.yahoo.com.

6.4 [ask] Minson JA, Ruedy NE, Schweitzer ME. *There is such a thing as a stupid question: Question disclosure in strategic communication.*
opim.wharton.upenn.edu/DPlab/papers/workingPapers/
Minson_working_Ask%20(the%20Right%20Way)%20and%20You%20Shall%20Receive.pdf

6.4 [diabetes2] Zeitler P, et al. 2012. A Clinical Trial to Maintain Glycemic Control in Youth with Type 2 Diabetes. N Engl J Med.

A.7 Inference for numerical data

7.1 [Risso's dolphins] Endo T and Haraguchi K. 2009. High mercury levels in hair samples from residents of Taiji, a Japanese whaling town. Marine Pollution Bulletin 60(5):743-747.

Taiji was featured in the movie *The Cove*, and it is a significant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually, and we will assume these 19 dolphins represent a simple random sample from those dolphins.

7.1 [Croaker white fish] www.fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

7.1 [run17] www.cherryblossom.org

7.2 [textbooks, ucla_textbooks_f18] Data were collected by OpenIntro staff in 2010 and again in 2018. For the 2018 sample, we sampled 201 UCLA courses. Of those, 68 required books that could be found on Amazon. The websites where information was retrieved:
sa.ucla.edu/ro/public/soc, ucla.verbacompare.com, and amazon.com.

7.3 [stem_cells] Menard C, et al. 2005. Transplantation of cardiac-committed mouse embryonic stem cells to infarcted sheep myocardium: a preclinical study. The Lancet: 366:9490, p1005-1012.

7.3 [ncbirths] Birth records released by North Carolina in 2004. Unfortunately, we don't currently have additional information on the source for this data set.

7.3 [Exam versions] This is a made-up data set for an example. There isn't any explicit data set to analyze, only summary statistics.

7.4 [Blood pressure statistics] The blood pressure standard deviation for patients with blood pressure ranging from from 140 to 180 mmHg is guessed and may be a little (but likely not dramatically) imprecise from what we'd observe in actual data.

7.5 [toy_anova] Data used for Figure 7.19, which was fake data.

7.5 [mlb_players_18] Data were retrieved from mlb.mlb.com/stats. Only players with at least 100 at bats were considered during the analysis.

7.5 [class_data] These data are simulated. [Data set is currently called `classData` in the package.]

A.8 ??

?? [simulated_scatter] Fake data used for the first three plots. The perfect linear plot uses group 4 data, where `group` variable in the data set (Figure **??**). The group of 3 imperfect linear plots use groups 1-3 (Figure **??**). The sinusoidal curve uses group 5 data (Figure **??**). The curved plot of data with the curved band uses group 30 data (right panel of Figure **??**). The group of 3 scatterplots with residual plots use groups 6-8 (Figure **??**). The correlation plots uses groups 9-19 data (Figures **??** and **??**).

?? [possum] This data is described in Data Appendix A.4.

?? [elmhurst] These data were sampled from a table of data for all freshman from the 2011 class at Elmhurst College that accompanied an article titled *What Students Really Pay to Go to College* published online by *The Chronicle of Higher Education*: chronicle.com/article/What-Students-Really-Pay-to-Go/131435.

?? [simulated_scatter] The plots for things that can go wrong uses groups 20-23 (Figure **??**).

?? [mario_kart] [Confirm data set name change.] Auction data from Ebay (ebay.com) for the game Mario Kart for the Nintendo Wii. This data set was collected in early October, 2009.

?? [simulated_scatter] The plots for types of outliers uses groups 24-29 (Figure **??**).

?? [midterms_house] Data was retrieved from Wikipedia.

A.9 ??

?? [loans_full_schema] This data is described in Data Appendix A.1.

?? [loans_full_schema] This data is described in Data Appendix A.1.

?? [loans_full_schema] This data is described in Data Appendix A.1.

?? [mario_kart] This data is describe in Data Appendix A.8.

?? [resume] Bertrand M, Mullainathan S. 2004. *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.* The American Economic Review 94:4 (991-1013). www.nber.org/papers/w9873

We did omit discussion of some structure in the data for the analysis presented: the experiment design included blocking, where typically four resumes were sent to each job: one for each inferred race/sex combination (as inferred based on the first name). We did not worry about this blocking aspect, since accounting for the blocking would *reduce* the standard error without notably changing the point estimates for the `race` and `sex` variables versus the analysis performed in the section. That is, the most interesting conclusions in the study would be unaffected even with a more sophisticated analysis.

Appendix B

Distribution tables

B.1 Normal Probability Table

A **normal probability table** may be used to find percentiles of a normal distribution using a Z-score, or vice-versa. Such a table lists Z-scores and the corresponding percentiles. An abbreviated probability table is provided in Figure B.1 that we'll use for the examples in this appendix. A full table may be found on page 140.

	Second decimal place of Z									
Z	*0.00*	0.01	0.02	0.03	**0.04**	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	**0.7995**	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	*0.8413*	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure B.1: A section of the normal probability table. The percentile for a normal random variable with $Z = 1.00$ has been *highlighted*, and the percentile closest to 0.8000 has also been **highlighted**.

When using a normal probability table to find a percentile for Z (rounded to two decimals), identify the proper row in the normal probability table up through the first decimal, and then determine the column representing the second decimal value. The intersection of this row and column is the percentile of the observation. For instance, the percentile of $Z = 0.45$ is shown in row 0.4 and column 0.05 in Figure B.1: 0.6736, or the 67.36^{th} percentile.

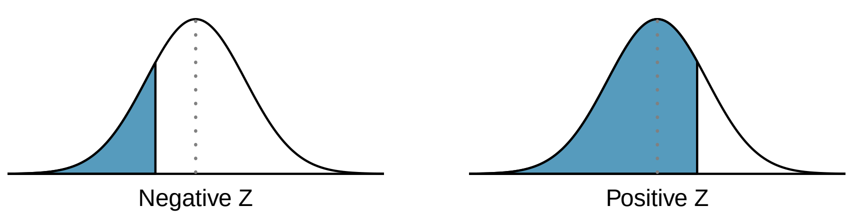
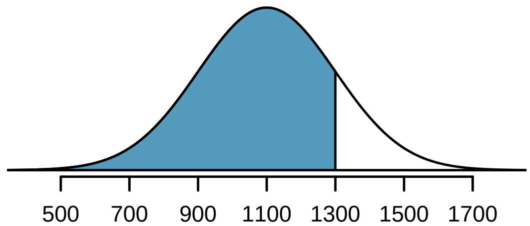


Figure B.2: The area to the left of Z represents the percentile of the observation.

EXAMPLE B.1

SAT scores follow a normal distribution, $N(1100, 200)$. Ann earned a score of 1300 on her SAT with a corresponding Z-score of $Z = 1$. She would like to know what percentile she falls in among all SAT test-takers.

Ann's **percentile** is the percentage of people who earned a lower SAT score than her. We shade the area representing those individuals in the following graph:



The total area under the normal curve is always equal to 1, and the proportion of people who scored below Ann on the SAT is equal to the *area* shaded in the graph. We find this area by looking in row 1.0 and column 0.00 in the normal probability table: 0.8413. In other words, Ann is in the 84^{th} percentile of SAT takers.

EXAMPLE B.2

How do we find an upper tail area?

The normal probability table *always* gives the area to the left. This means that if we want the area to the right, we first find the lower tail and then subtract it from 1. For instance, 84.13% of SAT takers scored below Ann, which means 15.87% of test takers scored higher than Ann.

We can also find the Z-score associated with a percentile. For example, to identify Z for the 80^{th} percentile, we look for the value closest to 0.8000 in the middle portion of the table: 0.7995. We determine the Z-score for the 80^{th} percentile by combining the row and column Z values: 0.84.

EXAMPLE B.3

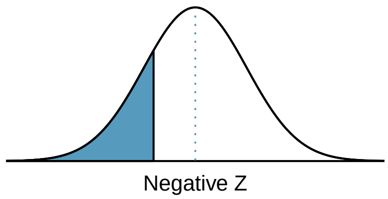
Find the SAT score for the 80^{th} percentile.

We look for the are to the value in the table closest to 0.8000. The closest value is 0.7995, which corresponds to $Z = 0.84$, where 0.8 comes from the row value and 0.04 comes from the column value. Next, we set up the equation for the Z-score and the unknown value x as follows, and then we solve for x:

$$Z = 0.84 = \frac{x - 1100}{200} \quad \rightarrow \quad x = 1268$$

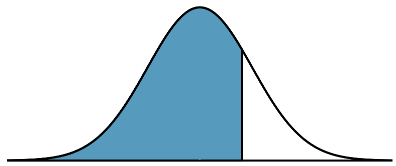
The College Board scales scores to increments of 10, so the 80^{th} percentile is 1270. (Reporting 1268 would have been perfectly okay for our purposes.)

For additional details about working with the normal distribution and the normal probability table, see Section 4.1, which starts on page 8.



Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5
0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	−1.4
0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	−1.3
0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151	−1.2
0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	−1.1
0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	−1.0
0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	−0.9
0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	−0.8
0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420	−0.7
0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743	−0.6
0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	−0.5
0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	−0.4
0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	−0.3
0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207	−0.2
0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	−0.1
0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000	−0.0

*For $Z \leq -3.50$, the probability is less than or equal to 0.0002.



Positive Z

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

*For $Z \geq 3.50$, the probability is greater than or equal to 0.9998.

B.2 t-Probability Table

A **t-probability table** may be used to find tail areas of a t-distribution using a T-score, or vice-versa. Such a table lists T-scores and the corresponding percentiles. A partial **t-table** is shown in Figure B.3, and the complete table starts on page 144. Each row in the t-table represents a t-distribution with different degrees of freedom. The columns correspond to tail probabilities. For instance, if we know we are working with the t-distribution with $df = 18$, we can examine row 18, which is highlighted in Figure B.3. If we want the value in this row that identifies the T-score (cutoff) for an upper tail of 10%, we can look in the column where *one tail* is 0.100. This cutoff is 1.33. If we had wanted the cutoff for the lower 10%, we would use -1.33. Just like the normal distribution, all t-distributions are symmetric.

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
⋮	⋮	⋮	⋮	⋮	
17	1.33	1.74	2.11	2.57	2.90
18	**1.33**	**1.73**	**2.10**	**2.55**	**2.88**
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
⋮	⋮	⋮	⋮	⋮	
400	1.28	1.65	1.97	2.34	2.59
500	1.28	1.65	1.96	2.33	2.59
∞	1.28	1.64	1.96	2.33	2.58

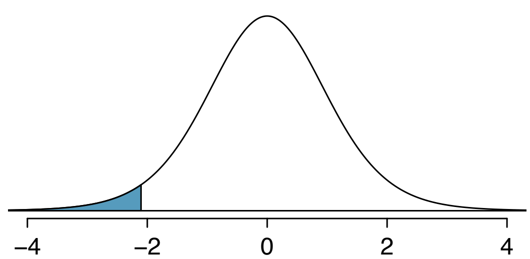
Figure B.3: An abbreviated look at the t-table. Each row represents a different t-distribution. The columns describe the cutoffs for specific tail areas. The row with $df = 18$ has been **highlighted**.

E

EXAMPLE B.4

What proportion of the t-distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw the picture and shade the area below -2.10:



To find this area, we first identify the appropriate row: $df = 18$. Then we identify the column containing the absolute value of -2.10; it is the third column. Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 0.025. That is, 2.5% of the distribution falls below -2.10.

In the next example we encounter a case where the exact T-score is not listed in the table.

EXAMPLE B.5

A t-distribution with 20 degrees of freedom is shown in the left panel of Figure B.4. Estimate the proportion of the distribution falling above 1.65.

E

We identify the row in the t-table using the degrees of freedom: $df = 20$. Then we look for 1.65; it is not listed. It falls between the first and second columns. Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65. We identify the one tail area of the first and second columns, 0.050 and 0.10, and we conclude that between 5% and 10% of the distribution is more than 1.65 standard deviations above the mean. If we like, we can identify the precise area using statistical software: 0.0573.

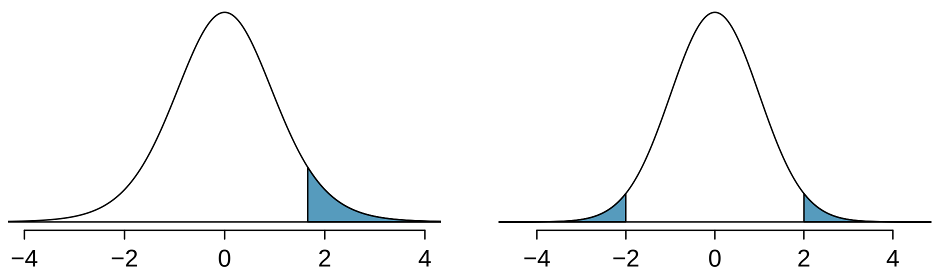


Figure B.4: Left: The t-distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t-distribution with 475 degrees of freedom, with the area further than 2 units from 0 shaded.

EXAMPLE B.6

A t-distribution with 475 degrees of freedom is shown in the right panel of Figure B.4. Estimate the proportion of the distribution falling more than 2 units from the mean (above or below).

E

As before, first identify the appropriate row: $df = 475$. This row does not exist! When this happens, we use the next smaller row, which in this case is $df = 400$. Next, find the columns that capture 2.00; because $1.97 < 3 < 2.34$, we use the third and fourth columns. Finally, we find bounds for the tail areas by looking at the two tail values: 0.02 and 0.05. We use the two tail values because we are looking for two symmetric tails in the t-distribution.

G

GUIDED PRACTICE B.7

What proportion of the t-distribution with 19 degrees of freedom falls above -1.79 units?[1]

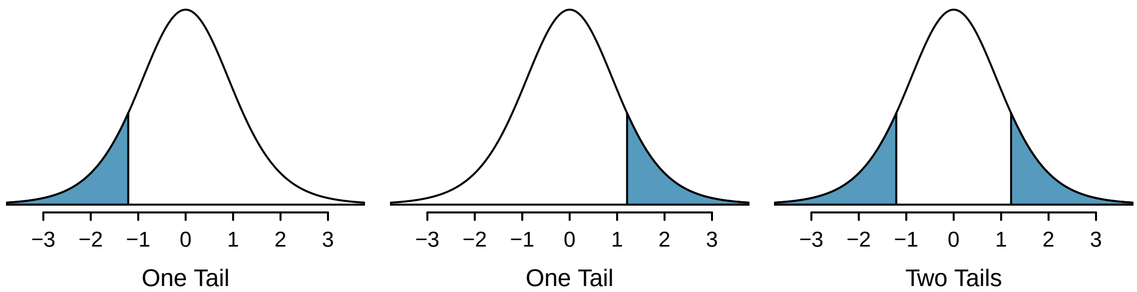
EXAMPLE B.8

Find the value of $t^\star_{18}$ using the t-table, where $t^\star_{18}$ is the cutoff for the t-distribution with 18 degrees of freedom where 95% of the distribution lies between -$t^\star_{18}$ and +$t^\star_{18}$.

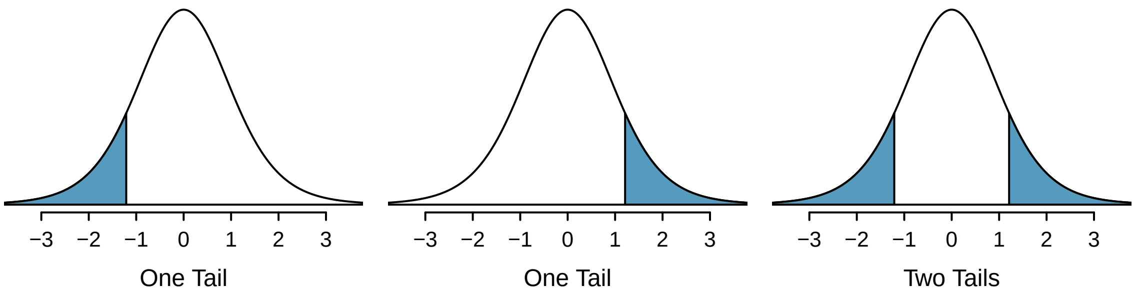
E

For a 95% confidence interval, we want to find the cutoff $t^\star_{18}$ such that 95% of the t-distribution is between -$t^\star_{18}$ and $t^\star_{18}$; this is the same as where the two tails have a total area of 0.05. We look in the t-table on page 142, find the column with area totaling 0.05 in the two tails (third column), and then the row with 18 degrees of freedom: $t^\star_{18} = 2.10$.

[1]We find the shaded area *above* -1.79 (we leave the picture to you). The small left tail is between 0.025 and 0.05, so the larger upper region must have an area between 0.95 and 0.975.



one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11
	12	1.36	1.78	2.18	2.68	3.05
	13	1.35	1.77	2.16	2.65	3.01
	14	1.35	1.76	2.14	2.62	2.98
	15	1.34	1.75	2.13	2.60	2.95
	16	1.34	1.75	2.12	2.58	2.92
	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79
	26	1.31	1.71	2.06	2.48	2.78
	27	1.31	1.70	2.05	2.47	2.77
	28	1.31	1.70	2.05	2.47	2.76
	29	1.31	1.70	2.05	2.46	2.76
	30	1.31	1.70	2.04	2.46	2.75



one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	31	1.31	1.70	2.04	2.45	2.74
	32	1.31	1.69	2.04	2.45	2.74
	33	1.31	1.69	2.03	2.44	2.73
	34	1.31	1.69	2.03	2.44	2.73
	35	1.31	1.69	2.03	2.44	2.72
	36	1.31	1.69	2.03	2.43	2.72
	37	1.30	1.69	2.03	2.43	2.72
	38	1.30	1.69	2.02	2.43	2.71
	39	1.30	1.68	2.02	2.43	2.71
	40	1.30	1.68	2.02	2.42	2.70
	41	1.30	1.68	2.02	2.42	2.70
	42	1.30	1.68	2.02	2.42	2.70
	43	1.30	1.68	2.02	2.42	2.70
	44	1.30	1.68	2.02	2.41	2.69
	45	1.30	1.68	2.01	2.41	2.69
	46	1.30	1.68	2.01	2.41	2.69
	47	1.30	1.68	2.01	2.41	2.68
	48	1.30	1.68	2.01	2.41	2.68
	49	1.30	1.68	2.01	2.40	2.68
	50	1.30	1.68	2.01	2.40	2.68
	60	1.30	1.67	2.00	2.39	2.66
	70	1.29	1.67	1.99	2.38	2.65
	80	1.29	1.66	1.99	2.37	2.64
	90	1.29	1.66	1.99	2.37	2.63
	100	1.29	1.66	1.98	2.36	2.63
	150	1.29	1.66	1.98	2.35	2.61
	200	1.29	1.65	1.97	2.35	2.60
	300	1.28	1.65	1.97	2.34	2.59
	400	1.28	1.65	1.97	2.34	2.59
	500	1.28	1.65	1.96	2.33	2.59
	∞	1.28	1.65	1.96	2.33	2.58

B.3 Chi-Square Probability Table

A **chi-square probability table** may be used to find tail areas of a chi-square distribution. The **chi-square table** is partially shown in Figure B.5, and the complete table may be found on page 147. When using a chi-square table, we examine a particular row for distributions with different degrees of freedom, and we identify a range for the area (e.g. 0.025 to 0.05). Note that the chi-square table provides upper tail values, which is different than the normal and t-distribution tables.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	2	2.41	**3.22**	**4.61**	5.99	7.82	9.21	10.60	13.82
	3	*3.66*	*4.64*	6.25	*7.81*	*9.84*	*11.34*	*12.84*	*16.27*
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Figure B.5: A section of the chi-square table. A complete table is in Appendix B.3.

EXAMPLE B.9

Figure B.6(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Figure B.5 to estimate the shaded area.

E

This distribution has three degrees of freedom, so only the row with 3 degrees of freedom (df) is relevant. This row has been italicized in the table. Next, we see that the value – 6.25 – falls in the column with upper tail area 0.1. That is, the shaded upper tail of Figure B.6(a) has area 0.1.

This example was unusual, in that we observed the *exact* value in the table. In the next examples, we encounter situations where we cannot precisely estimate the tail area and must instead provide a range of values.

EXAMPLE B.10

Figure B.6(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The area above value 4.3 has been shaded; find this tail area.

E

The cutoff 4.3 falls between the second and third columns in the 2 degrees of freedom row. Because these columns correspond to tail areas of 0.2 and 0.1, we can be certain that the area shaded in Figure B.6(b) is between 0.1 and 0.2.

EXAMPLE B.11

Figure B.6(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area.

E

Looking in the row with 5 df, 5.1 falls below the smallest cutoff for this row (6.06). That means we can only say that the area is *greater than* 0.3.

EXAMPLE B.12

Figure B.6(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.

E

The value 11.7 falls between 9.80 and 12.02 in the 7 df row. Thus, the area is between 0.1 and 0.2.

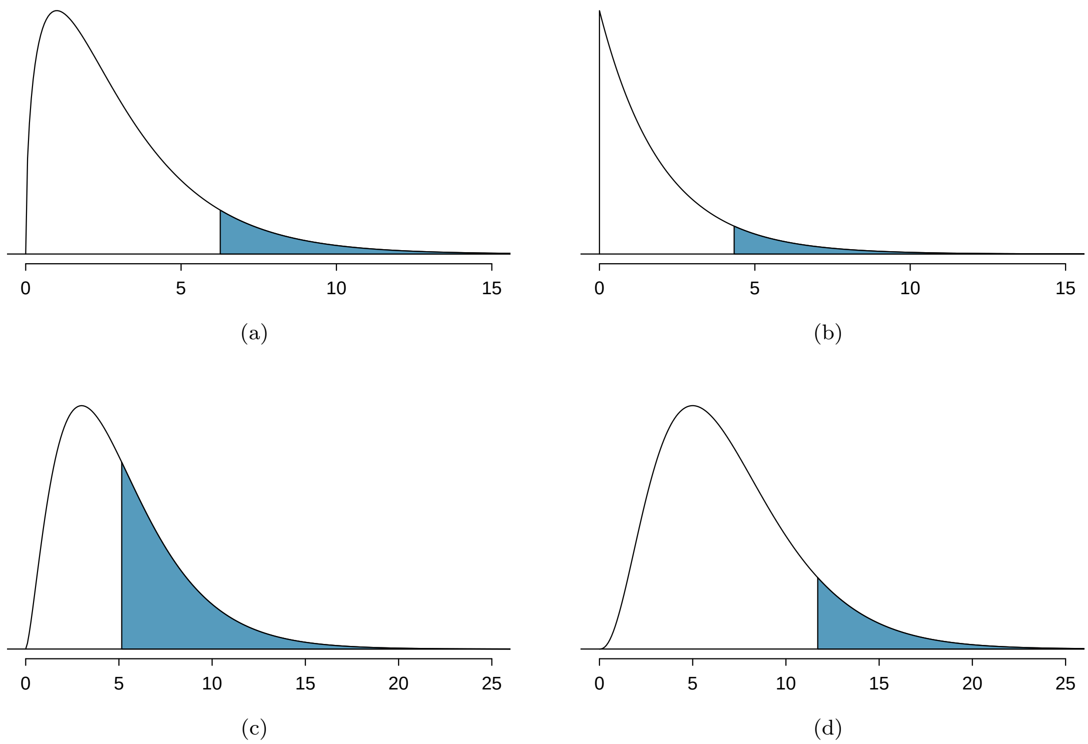


Figure B.6: (a) Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. (b) 2 degrees of freedom, area above 4.3 shaded. (c) 5 degrees of freedom, area above 5.1 shaded. (d) 7 degrees of freedom, area above 11.7 shaded.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26
	12	14.01	15.81	18.55	21.03	24.05	26.22	28.30	32.91
	13	15.12	16.98	19.81	22.36	25.47	27.69	29.82	34.53
	14	16.22	18.15	21.06	23.68	26.87	29.14	31.32	36.12
	15	17.32	19.31	22.31	25.00	28.26	30.58	32.80	37.70
	16	18.42	20.47	23.54	26.30	29.63	32.00	34.27	39.25
	17	19.51	21.61	24.77	27.59	31.00	33.41	35.72	40.79
	18	20.60	22.76	25.99	28.87	32.35	34.81	37.16	42.31
	19	21.69	23.90	27.20	30.14	33.69	36.19	38.58	43.82
	20	22.77	25.04	28.41	31.41	35.02	37.57	40.00	45.31
	25	28.17	30.68	34.38	37.65	41.57	44.31	46.93	52.62
	30	33.53	36.25	40.26	43.77	47.96	50.89	53.67	59.70
	40	44.16	47.27	51.81	55.76	60.44	63.69	66.77	73.40
	50	54.72	58.16	63.17	67.50	72.61	76.15	79.49	86.66