Introductory Statistics for the Life and Biomedical Sciences

Derivative of
OpenIntro Statistics
Third Edition

Original Authors

David M Diez
Christopher D Barr
Mine Çetinkaya-Rundel

Contributing Authors

David Harrington
[Briefly Describe Contribution]

Julie Vu
[Briefly Describe Contribution]

Alice Zhao
[Briefly Describe Contribution]

Contents

Preface

This book provides an introduction to statistics and its applications in the life sciences, and biomedical research. It is based on the freely available *OpenIntro Statistics, Third Edition*, and, like *OpenIntro* it may be downloaded as a free PDF at **Need location**. The text adds substantial new material, revises or eliminates sections from *OpenIntro*, and re-uses some material directly. Readers need not have read *OpenIntro*, since this book is intended to be used independently. We have retained some of the exercises from *OpenIntro* that may not come directly from medicine or the life sciences but illustrate important ideas or methods that are commonly used in fields such as biology.

Introduction to Statistics for the Life and Biomedical Sciences is intended for graduate and undergraduate students interested in careers in biology or medicine, and may also be profitably read by students of public health. It covers many of the traditional introductory topics in statistics used in those fields, but also adds some newer methods being used in molecular biology. Statistics has become an integral part of research in medicine and biology, and the tools for displaying, summarizing and drawing inferences from data are essential both for understanding the outcomes of studies and for incorporating measures of uncertainty into that understanding. An introductory text in statistics for students considering careers in medicine, public health or the life sciences should be more than the usual introduction with more examples from biology or medical science. Along with the value of careful, robust analyses of experimental and observational data, it should convey some of the excitement of discovery that emerges from the interplay of science with data collection and analysis. We hope we have conveyed some of that excitement here.

We have tried to balance the sometimes competing demands of mastering the impor-

tant technical aspects of methods of analysis with gaining an understanding of important concepts. The examples and exercises include opportunities for students to build skills in conducting data analyses and to state conclusions with clear, direct language that is specific to the context of a problem. We also believe that computing is an essential part of statistics, just as mathematics was when computing was more difficult or expensive. The text includes many examples where software is used to aid in the understanding of the features of a data as well as exercises where computing is used to help illustrate the notions of randomness and variability. Because they are freely available, we use the R statistical language with the *R Studio* interface. Information on downloading R and *R Studio* is may be found in the Labs at **openintro.org**. Nearly all examples and exercises can be adapted to either SAS, Stata or other software, but we have not done that.

Textbook overview

The chapters of this book are as follows:

1. **Introduction to data.** Data structures, variables, summaries, graphics, and basic data collection techniques.

2. **Probability (special topic).** The basic principles of probability. An understanding of this chapter is not required for the main content in Chapters **??**-**??**.

3. **Distributions of random variables.** Introduction to the normal model and other key distributions.

4. **Foundations for inference.** General ideas for statistical inference in the context of estimating the population mean.

5. **Inference for numerical data.** Inference for one or two sample means using the normal model and t distribution, and also comparisons of many means using ANOVA.

6. **Inference for categorical data.** Inference for proportions using the normal and chi-square distributions, as well as simulation and randomization techniques.

7. **Introduction to linear regression.** An introduction to regression with two variables. Most of this chapter could be covered after Chapter **??**.

8. **Multiple and logistic regression.** An introduction to multiple regression and logistic regression for an accelerated course.

The remainder of this section requires revision

OpenIntro Statistics was written to allow flexibility in choosing and ordering course topics. The material is divided into two pieces: main text and special topics. The main text has been structured to bring statistical inference and modeling closer to the front of a course. Special topics, labeled in the table of contents and in section titles, may be added to a course as they arise naturally in the curriculum.

Examples, exercises, and appendices

Examples and within-chapter exercises throughout the textbook may be identified by their distinctive bullets:

● **Example 0.1** Large filled bullets signal the start of an example.

Full solutions to examples are provided and often include an accompanying table or figure.

⊙ **Guided Practice 0.2** Large empty bullets signal to readers that an exercise has been inserted into the text for additional practice and guidance. Students may find it useful to fill in the bullet after understanding or successfully completing the exercise. Solutions are provided for all within-chapter exercises in footnotes.[1]

There are exercises at the end of each chapter that are useful for practice or homework assignments. Many of these questions have multiple parts, and odd-numbered questions include solutions in Appendix **??**.

Probability tables for the normal, t, and chi-square distributions are in Appendix **??**, and PDF copies of these tables are also available from **openintro.org** for anyone to download, print, share, or modify.

[1] Full solutions are located down here in the footnote!

OpenIntro, online resources, and getting involved

OpenIntro is an organization focused on developing free and affordable education materials. *OpenIntro Statistics*, our first project, is intended for introductory statistics courses at the high school through university levels.

We encourage anyone learning or teaching statistics to visit **openintro.org** and get involved. We also provide many free online resources, including free course software. Data sets for this textbook are available on the website and through a companion R package.[2] All of these resources are free, and we want to be clear that anyone is welcome to use these online tools and resources with or without this textbook as a companion.

We value your feedback. If there is a particular component of the project you especially like or think needs improvement, we want to hear from you. You may find our contact information on the title page of this book or on the About section of **openintro.org**.

Acknowledgements

This project would not be possible without the dedication and volunteer hours of all those involved. No one has received any monetary compensation from this project, and we hope you will join us in extending a *thank you* to all those volunteers below.

The authors would like to thank Andrew Bray, Meenal Patel, Yongtao Guan, Filipp Brunshteyn, Rob Gould, and Chris Pope for their involvement and contributions. We are also very grateful to Dalene Stangl, Dave Harrington, Jan de Leeuw, Kevin Rader, and Philippe Rigollet for providing us with valuable feedback.

[2] Diez DM, Barr CD, Çetinkaya-Rundel M. 2012. `openintro`: OpenIntro data sets and supplement functions. http://cran.r-project.org/web/packages/openintro.

Chapter 4

Foundations for inference

Imagine the United States Center for Disease Control and Prevention (CDC) influencing policy makers in curbing the national obesity problem. The members of the CDC's Division of Nutrition, Physical Activity, and Obesity (DNPAO) "focus on policy and environmental strategies to make healthy eating and active living accessible and affordable for everyone." [1] Before they give policy suggestions, however, the DNPAO must first diagnose this problem of obesity in the United States. One metric that these scientists would consider could be a person's Body Mass Index, also known as BMI. A person's BMI has been a helpful tool to capture both a person's height and weight within one measurement and can be used as a measure of body fat. A high BMI can be an indicator for high body fat. In medicine, BMI can be used to categorize a person as "underweight," "overweight" or "obese."

These policy makers can be interested in a couple of questions. What is the average population BMI for all adults in the United States? Instead of providing a single number, what is a plausible range values for the average BMI in the United States? Finally using the categorization of BMI values and ranges that the World Health Organization (WHO) provides, noting that it does not consider muscularity, is the average BMI in the United States considered an average healthy BMI? Below is a categorization of BMI values and ranges [2]

Category	BMI range
Underweight	< 18.50
Normal (healthy weight)	18.5-24.99
Overweight	≥ 25
Obese	≥ 30

These questions encompass the broader idea of statistical inference in Chapter 4. Inference is a set of tools used to estimate properties or parameters about a population after observing a sample from this population. Inference also allows for different levels of quality or confidence of these parameter estimates. Once we have a parameter estimate, the average BMI in the United States for example, we can ask ourselves how confident we are that this estimate is representative of the greater population of the US. For example, a classic inferential question is, "How sure are we that the estimated mean, $\bar{x}$, is near the population mean, μ?" Statistical inference includes asking these questions but also determining which estimates to use.

[1] http://www.cdc.gov/obesity/
[2] http://apps.who.int/bmi/index.jsp?introPage=intro_3.html

Chapter 4 provides the groundwork for inference on a larger population from observing one sample, and later chapters will cover inference comparing two or more distinct populations. While the equations and details change depending on the setting, the foundations and general procedures for inference are the same throughout statistics. Understanding the foundation with point estimates in this chapter will provide familiarity for upcoming chapters.

After looking at the data in Section 4.1 that the CDC could use in making these inferences, section 4.2 will give us an introduction to point estimates, the sampling distribution that these estimates are drawn from and the variability of these estimates. Section 4.3 will give us tools to incorporate this variability. Rather than a single value, these policy makers can provide, instead, a confidence interval or a range of values that they believe are likely estimates. However with comparison, researchers might still want to compare point values against each other instead of ranges of values. Hypothesis testing in Section 4.4 allows the researchers to infer using point estimates while still encompassing the variability that the point estimates in section 4.2 did not. The hypothesis testing framework gives us structure to do so and uses the same moving parts as a confidence interval.

4.1 BRFSS Data

The Behavioral Risk Factor Surveillance System (BRFSS) by the CDC was started in 1984 and is the world's largest on-going telephone health survey system. This survey is nationwide and aims to "monitor state-level prevalence of major behavioral risks among adults associated with premature morbidity and mortality." [3] Topics like smoking, alcohol use, diet and exercise are included in this questionnaire. The annual survey data from 2000, `BRFSS`, includes records on 289 variables and could be used to estimate the average BMI of the US population. The variables that we are particularly interested in with calculating BMI are listed in Table ??.

variable	description
`sex`	Male or Female where 1 is Male and 2 is Female
`age`	In years
`height`	In feet and inches where, for example, 5' 5" is listed as 505
`weight`	In pounds

Table 4.1: Variables of interest and their descriptions for the `BRFSS` data set.

The calculation of a BMI index from weight and height using both Metric and Imperial is

$$BMI = \frac{\text{weight}_{\text{kg}}}{\text{height}_{\text{m}}{}^2} = \frac{\text{weight}_{\text{lb}}}{\text{height}_{\text{in}}{}^2} \cdot 703$$

where $\text{weight}_{\text{kg}}$ and height_{m} is the weight and height measured in kilograms and meters while $\text{weight}_{\text{lb}}$ and $\text{height}_{\text{in}}$ is the weight and height in pounds and inches.

Therefore by using the `BRFSS` data, we can use our sample of heights and weights to calculate BMI values for each observation. The `BRFSS` data comprises of 170,000 observations, and the CDC is hoping to infer the characteristics of adults in the United States,

[3] http://www.cdc.gov/brfss/about/about_brfss.htm

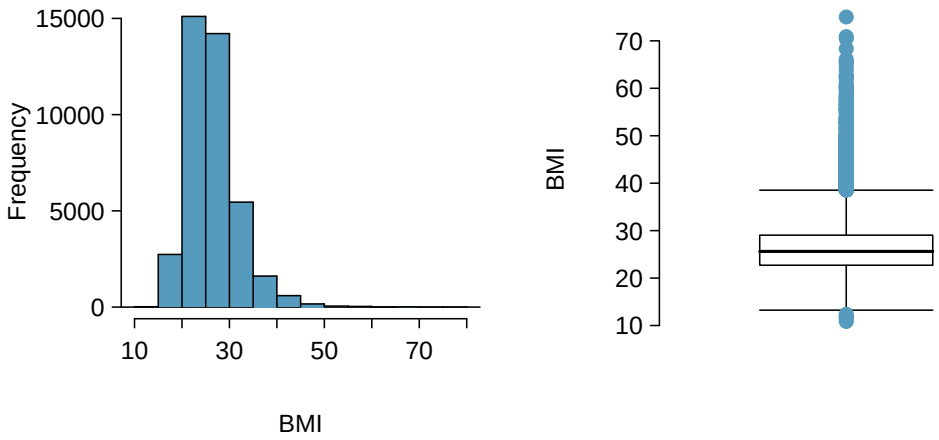


Figure 4.3: Histogram and boxplot of BMI for the `BRFSS BMI` data. The data is skewed right by both the histogram and the box plot. We can see from the box plot that it has many outliers

our target population. A target population is the group that the statistician is interested in and wants to draw conclusions about.

We take a sample of 40,000 adults from the `BRFSS` data to use as our observed sample. We will refer to this random sample of 40,000 adults as `BRFSS BMI` from now on. Part of this dataset with the BMI calculation is shown in 4.2.

	sex	age	height	weight	bmi
1	2	60	508	200	30.41
2	2	25	506	145	23.40
3	1	40	511	180	25.10
4	1	53	511	210	29.29
5	2	80	504	170	29.18
6	2	71	501	108	20.40

Table 4.2: Six observations from the BRFSS BMI dataset

This simple random sample from `BRFSS` will be used to draw conclusions about the target population of US adults. This is the practice of statistical inference in the broadest sense. Let's explore the data with the tools from Chapter 1 in Figure 4.3 before we estimate the average BMI in the US.

The data from `BRFSS BMI` is special because in order to do statistical inference, the dataset needs to be representative of the population of interest. In this case, because the size of `BRFSS BMI` is so large and drawn randomly from `BRFSS`, we can assume that our data is representative and our estimates will be close to the population parameters. Now that we have a general idea of what the data looks like, we can begin with statistical inference.

4.2 Variability in estimates

If members at the CDC, after observing the sample of 40,000 BMI values, were asked to give their best guess for the average BMI in the US, what would it be? Here they would

employ point estimation. A **point estimate** is a single value derived from sample data that serves as the "best guess" for that population parameter. Section 4.2 will touch upon point estimates as well as the variability inherent in using this single number as the best guess. [4]

4.2.1 Point estimates

A likely choice to estimate the **population mean** from our sample is to simply take the **sample mean**. That is, use the average BMI of all 40,000 survey respondents in our sample as our estimate for the average BMI among US adults.

For notation, use $\text{bmi}_1, \text{bmi}_2, \ldots, \text{bmi}_{40,000}$ to represent the BMI for each survey respondent in our sample `BRFSS BMI`. The sample mean for BMI using our 40,000 observations is

$$\bar{\text{bmi}} = \frac{30.40 + 23.40 + 25.10 + \ldots}{40,000} = 26.356$$

and is the **point estimate** of the population mean[5].

Suppose from the original respondents in `BRFSS`, we take a new sample of 40,000 people and recompute the mean; we will probably not get the same answer that we got using the `BRFSS BMI` data set. Estimates generally vary from one sample to another even with the same sample size, and low **sampling variation** can suggest our estimate may be close, but not exactly equal to the parameter. A larger sample size can ensure a closer estimate to the population parameter. In 2000, the US population was 282.2 million, [6] 7,000 times our sample size of 40,000.

What about generating point estimates of other **population parameters**, such as the population median or population standard deviation? We can estimate parameters based on sample statistics. For example, we estimate the population standard deviation for BMI using the sample standard deviation, and the population median using the sample median. Table 4.4 provides the point estimates to other population parameters relating to BMI.

BMI	estimates
mean	26.356
median	25.620
std. dev.	5.288

Table 4.4: Point estimates for the `bmi` variable

⊙ **Guided Practice 4.1** Suppose we want to estimate the average age for men and women in the US. If $\bar{\text{age}}_{\text{women}} = 47.35$ years and $\bar{\text{age}}_{\text{men}} = 45.67$ years, what would be a good point estimate for the population age difference? [7]

[4] While we focus on the the mean of BMIs in this chapter, questions regarding variation are often just as important in practice. For instance, potential action regarding obesity could change if the standard deviation of a person's BMI was 5 versus if it was 15.

[5] If we were interested in the values of another variable, `weight` denoted $w_1, \ldots w_{40,000}$ instead, we would denote the sample mean as $\bar{w}$

[6] http://www.census.gov/prod/2002pubs/c2kprof00-us.pdf

[7] We could take the difference of the two sample means: $47.35 - 45.67 = 1.69$. Women are on average older than men by 1.69 years.

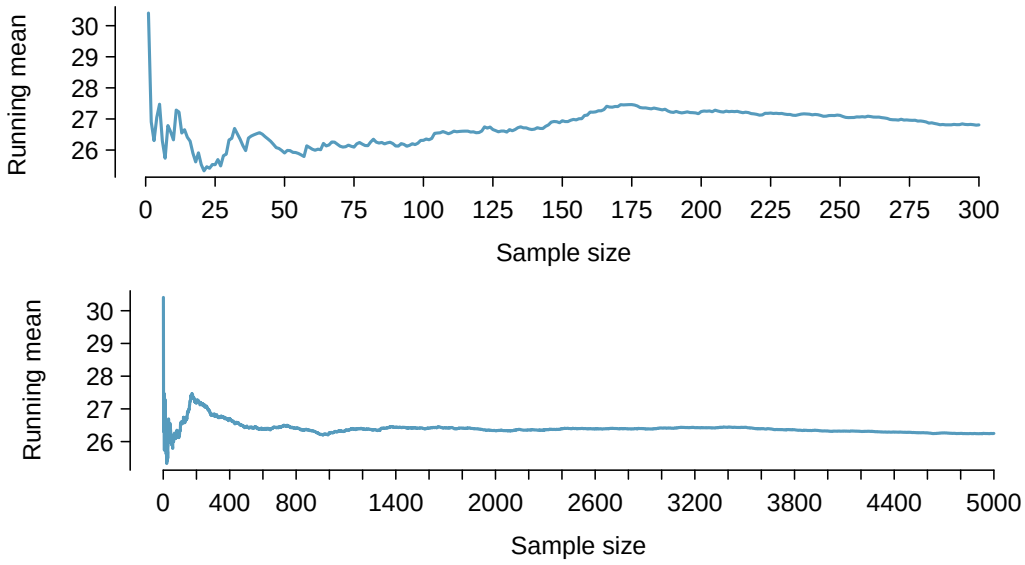


Figure 4.5: The running means from the `BRFSS BMI` sample of 40,000 observations. The mean stabilizes and approaches the mean of the entire sample $\bar{x} = 26.36$ as the number of observations increases

⊙ **Guided Practice 4.2** If you had to provide a point estimate of the population IQR for the BMI of participants, how might you make such an estimate using a sample?[8]

The sample mean calculated from this `BRFSS BMI` sample of 40,000 will likely be different from the sample mean of a different set of 40,000 respondents from `BRFSS` data. Using *R*, we take another random sample from the `BRFSS` data of 40,000 and see that that the new sample mean for the BMI is 26.344. We note that estimates will differ across samples through sampling variation but the accuracy of the point estimate will get improve once more data becomes available and sample sizes increase.

Consider a running mean from the `BRFSS BMI` data to explore increases in sample size. A **running mean** is a sequence of means, where each mean uses one more observation in its calculation than the mean directly before it in the sequence. In this case, the second mean is the average of the first two observations, $\text{bmi}_1, \text{bmi}_2$. The third number in the running mean sequence is the average of $\text{bmi}_1, \text{bmi}_2$, and bmi_3.

The running mean for `bmi` in the `BRFSS BMI` dataset is shown in Figure 4.5. We look at a running mean of 300 and 5000 observations. We note that as more values get included, the running mean converges closer to the sample mean of 26.36. Similarly if the sample size increases from 40,000 to 100,000, the sample mean from 100,000 observations will be closer to the average US population BMI than the sample mean of 40,000 observations.

Sampling variation, however, is across samples of the same size. Figure 4.6 displays the running means of two samples and of 20 samples. We see at each observation that the sample mean is not the same. There exists some sampling variation. Even more interesting, the variation decreases as the number of observations increases. We will explore this concept more in Section 4.2.2.

[8] To obtain a point estimate of the IQR for the population, we could take the IQR of the sample.

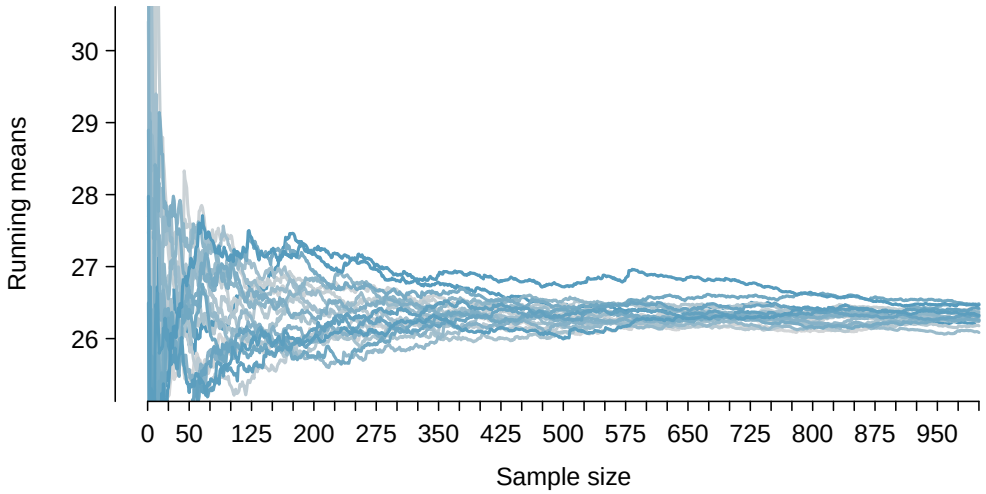


Figure 4.6: Many running means from different samples of 40,000 observations from BRFSS. The sampling variation decreases as the number of observations gets larger.

4.2.2 Accuracy and Precision of Point Estimates

Accuracy and precision have colloquially become interchangeable. In science, however, they both have very distinct meanings. Accuracy is a characteristic of how close the measurements are to their true value. Precision is the characteristic of how close the measurements are to themselves.

Within inference and point estimates, the sample mean is always accurate. The sample mean does not contain any systematic error. The sample mean is only not equal to the population mean because of random sampling error. While the sample mean may not always be equal to the population mean, in expectation, the sample mean and population parameter are equivalent. We observe the accuracy of the sample mean in practice with 20 running means from Figure 4.6. The center of these running means is, in expectation, the population average BMI.

While the sample mean is consistently accurate, it may not always be precise. As the sample size, n, increases, the randomness in the sample mean decreases. We look to sampling variation at different sample sizes as evidence. Figure 4.7 shows two histograms of sample means. The left histogram has sample means with a sample size of 5 and the right has a sample size of 50. All of these samples are randomly drawn from the BRFSS data, the sample mean calculated and plotted as an observation on the histogram. The histogram with $n = 50$ has noticeably smaller variance than the histogram with $n = 5$. The histogram with $n = 5$ is also slightly skewed. As such, with larger sample sizes, the sample variation decreases, making the sample mean more precise across samples. We look to Section 4.2.3 to quantify and measure sampling variation as more data becomes available.

4.2.3 Sampling Distribution and the Standard Error of the Mean

From the random sample represented in BRFSS BMI, we estimated the average BMI of an adult in the United States to be 26.356. Suppose we take another random sample of 40,000

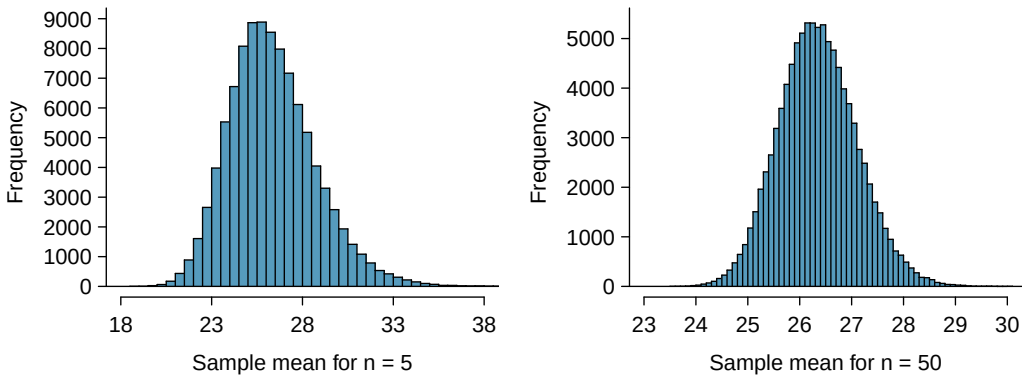


Figure 4.7: Sample means of size $n = 5$ and $n = 50$.

individuals and take its mean. In Section 4.2.1, we then get 26.359. Suppose we took another (26.350) and another (26.349), and continue to do this many many times – which we can do only because we have access to the larger `BRFSS` dataset. [9] We can then build up a **sampling distribution** for the sample mean when the sample size is 40,000, shown in Figure 4.8.

Sampling distribution

The **sampling distribution** of a point estimate represents the distribution of the point estimate based on samples of a fixed size from a certain population. There is a unique sampling distribution that exists that is inherent to the point estimator you are measuring. Every time that you are calculating your point estimate from a particular sample of said size, your point estimate s one sample from the sampling distribution. Understanding the concept of a sampling distribution is central to understanding statistical inference.

Figure 4.8 is an approximation of the sampling distribution. To truly get the sampling distribution, one would need to sample every possible unique combination of 40,000 respondents from the entire US adult population (and not just the `BRFSS` data set). However we note that just as the running mean becomes a better approximation of the population average as more data becomes available, the approximation of the sampling distribution also resembles more closely the sampling distribution as we take more and more samples. We note again that precision increases significantly as the sample size $n = 40,000$ with the sample means of $n = 40,000$ ranging from 26.25 to 26.45, a much smaller range than for $n = 5$ or $n = 50$ in Figure 4.7. We can create an approximation of the sampling distribution of the sampling mean with the following pseudocode [10]:

```
(1) Have a place to store all the sample means that we will calculate
```

[9] The sampling distribution depends on the underlying distribution of the target population. In this case, while `BRFSS` is not quite the target population of all US adults, it is large enough illustrate the concept of sampling distribution and acts as a representative substitute to the US population. If we had complete data from the target population, there would be no need to take a sample mean measurement. In practice, we generally aren't even capable of taking another sample of 40,000 from `BRFSS`!

[10] Refer to the appendix for R code

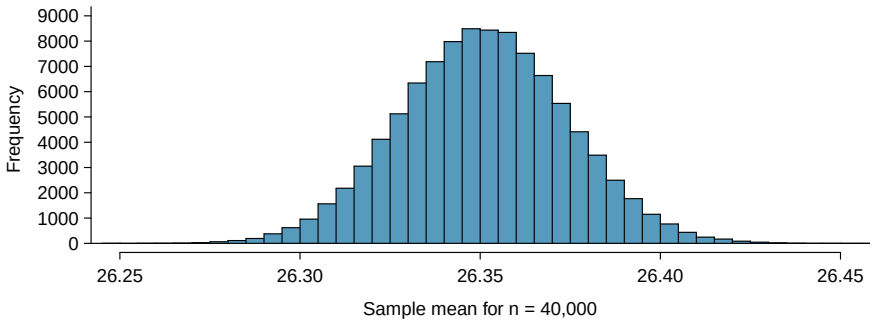


Figure 4.8: A histogram of 100,000 sample means for BMI, where the samples are of size $n = 40,000$.

```
(2) Take a sample from the BRFSS dataset of 40,000
(3) Calculate the sample mean from this specific sample and store it in (1)
(4) Repeat (2) and (3) many many times
(5) Plot all the sample means you have stored in (1) as a histogram
```

The sampling distribution, in this case, is likely to be unimodal and approximately symmetric. The sampling distribution is also centered exactly at the `BRFSS` population mean: $\mu = 26.351$. Intuitively, this makes sense. The sample means should tend to "fall around" the mean that we are drawing from.

By the sampling distribution, we also see that the point estimator will have some variability. We see the concept of sampling distribution introduced in Section 4.2.2 and note that the precision increases as the sample size gets larger. Point estimates, however, still vary sample by sample. There needs to be some metric to quantify the variability of the sample mean around the population mean.

TIP: More data means less variability

In sampling, the larger the sample size the better. The precision of the sample mean increases as you observe more data in your sample.

Standard deviation is the most obvious method to quantify variability. We use the standard deviation of the sampling distribution denoted $\sigma_{\bar{x}}$ or s in some contexts to measure sampling variation of the sample mean. [11]

Just as with the definition of standard deviation in **Chapter 1**, the standard deviation of the sample mean, tells us how far the typical estimate is away from the actual population mean. It also is a very good metric for the typical **error** of the point estimate, and for this reason we usually call this version of standard deviation the **standard error (SE)** of the estimate. [12]

SE
standard error

[11] Caution: The standard deviation of the sample mean is not equivalent to the estimate for the standard deviation of the population. Those are measuring two separate quantities.

[12] In general, standard error is the standard deviation of samples and estimates whereas we use the term standard deviation for populations or distributions. Look AT **SOME REFERENCE** for a clearer distinction of standard error versus standard deviation.

Standard error of an estimate

The standard deviation associated with an estimate is called the *standard error*. It describes the typical error or uncertainty associated with the estimate.

⊙ **Guided Practice 4.3** (a) Would you rather use a small sample or a large sample when estimating a parameter? Why? (b) Using your reasoning from (a), would you expect a point estimate based on a small sample to have smaller or larger standard error than a point estimate based on a larger sample?[13]

The standard error could be calculated if statisticians knew the sampling distribution. It would simply be the standard deviation of that sampling distribution. However when considering the case of the point estimate $\bar{x}$, there is one problem: most often, scientists only observe one sample, and there is no obvious way to estimate its standard error from a single sample. Computation methods and statistical theory, instead, provide helpful tools to address this issue.

Instead of only observing one sample from `BRFSS`, imagine if we could repeatedly sample from `BRFSS` as with creating the approximation to the sampling distribution. After many iterations, the standard deviation of the sample means becomes a fairly reasonable estimate of the standard error of the sample mean. In pseudocode:

```
(1) Have a place to store all the sample means that we will calculate
(2) Take a sample from the BRFSS dataset of 40,000
(3) Calculate the sample mean from this specific sample and store it in (1)
(4) Repeat (2) and (3) many many times
(5) Calculate the standard deviation of the values in (1). This is an
estimation of your standard error
```

Again what we are doing in *R* is creating an approximate sampling distribution and then using the standard deviation from this approximation to be our estimate for the standard error. We do this with the following code:[14]

```
> sample.means<-matrix(data=NA,nrow=1000, ncol=1) #to store the sample means
> for(i in 1:1000){
+   sample<-sample(x=brfss.df$bmi, size=40000, replace=FALSE)
+   sample.means[i]<-mean(sample)
+ }
> sd(sample.means)
[1] 0.02359317
```

[13] (a) Consider two random samples: one of size 10 and one of size 1000. Individual observations in the small sample are highly influential on the estimate while in larger samples these individual observations would more often average each other out. The larger sample would tend to provide a more accurate estimate. (b) If we think an estimate is better, we probably mean it typically has less error. Based on (a), our intuition suggests that a larger sample size corresponds to a smaller standard error.

[14] This computing experience samples without replacement to simulate experimenting in the real world. Theory states however that individual BMI values need to be independent. A reliable method to ensure sample observations are independent is to guarantee that the sample from the population is a simple random sample with a size that is less than 10% of the population. This 10% rule, remember, is used as a rule of thumb. By sampling without replacement within a finite population, we will see that `sd(sample.means)` is not exactly the theoretical standard error but quite close.

This method, however, has one main issue. Practitioners rarely are able to repeatedly sample from their larger population. Instead the standard error of the sample mean can be calculated through the following equation:

Computing SE for the sample mean

Given n independent observations from a population with standard deviation σ, the standard error of the sample mean is equal to

$$SE_{\text{sample mean}} = \frac{\sigma}{\sqrt{n}} \tag{4.4}$$

A reliable method to ensure sample observations are independent is to guarantee that the sample you have from the population is a simple random sample with a size that is less than 10% of the population.

There is one subtle issue of Equation (4.4) that you might have realized: the population standard deviation is typically unknown. To resolve this problem, we can use the point estimate of the standard deviation from the sample. This estimate tends to be sufficiently good when the sample size is at least 30 and the population distribution is not strongly skewed. Thus, we often just use the sample standard deviation denoted s instead of σ for the population standard deviation. When the sample size is smaller than 30, we will need to use a method to account for extra uncertainty in the standard error. If the skew condition is not met, a larger sample is needed to compensate for the extra skew. These topics are further discussed in Section 4.5.

With the `BRFSS BMI` sample of 40,000, the standard error of the sample mean is calculated as

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{5.288}{\sqrt{40000}} = 0.026$$

where s is the standard deviation of the sample and n is the number of observations in the sample. We see that the standard error calculated (0.026) is similar to the empirical standard deviation of the sampling distribution (0.024).

⊙ **Guided Practice 4.5** In another sample of 40,000 US adults, the standard deviation of BMI is $s_y = 5.34$. Because the sample is a simple random sample and consists of less than 10% of the United States population, the observations are independent. (a) What is the standard error of the sample mean, $\bar{y} = 26.36$? (b) Would you be surprised if someone told you the average BMI of all US adults was actually 30? What about 26? [15]

⊙ **Guided Practice 4.6** (a) Would you be more trusting of a sample that has 100 observations or 400 observations? (b) We want to show mathematically that our estimate tends to be better when the sample size is larger. If the standard deviation

[15] (a) Use Equation (4.4) with the sample standard deviation to compute the standard error: $SE_{\bar{y}} = 5.34/\sqrt{40000} = 0.0267$. (b) It would be surprising if the true average BMI was 30. A BMI of 30 is many standard deviations away from the sample mean of 26.36. In other words, a BMI of 30 seems implausible given that our sample mean (26.36) is far from the "true mean" using the standard error of 0.0267 to identify what is close and what is not close. Even a BMI of 26 in this situation would be surprising given that is it more than one standard deviation away from the sample mean (standard error of 0.0267).

of the individual observations is 10, what is our estimate of the standard error when the sample size is 100? What about when it is 400? (c) Explain how your answer to (b) mathematically justifies your intuition in part (a).[16]

4.2.4 Basic properties of point estimates

We achieved three goals in this section. First, we determined that point estimates from a sample may be used to estimate population parameters. We also determined that these point estimates are not exact, and there exists some sampling variation. The sample mean is an example of a point estimate that is always accurate but not necessarily always precise. The precision of a point estimate can be represented through sampling variation and visualized through a sampling distribution, and the point estimate that we observe is a single observation in the estimate's entire sampling distribution. Lastly, we quantified this sampling variation and the uncertainty of the sample mean using what we call the standard error. The standard error of the sample mean can be mathematically represented in Equation (4.4) or through computation with *R*. While we could also quantify the standard error for other estimates – such as the median, standard deviation, or any other number of statistics – we will postpone these extensions until later chapters and courses.

4.3 Confidence intervals

A point estimate, we saw in Section 4.2.1 provides a single plausible value for a parameter. However, a point estimate is rarely perfect and exact; usually there is some error in the estimate. We know that there exists sampling variation but a single point estimate does not convey how large this sampling variation is without including the point estimate's standard error. Instead of supplying just a point estimate, the next logical step would be to provide a plausible *range of values* to estimate the true value of the parameter.

In this section and in Section 4.4, we will emphasize the special case where the point estimate is a sample mean and the parameter that we are interested in is the population mean. In Section 4.6, we generalize these methods for a variety of point estimates and population parameters that we will encounter in Chapter **??** and beyond.

4.3.1 Capturing the population parameter

A plausible range of values for the population parameter is called a **confidence interval**. The width of an interval provides a gauge of how large the sampling variation is. For the same confidence level, the larger the interval indicates the larger sampling variation and standard error.

Using only a point estimate is like fishing in a murky lake with a spear, and using a confidence interval is like fishing in the same lake with a net. We can throw a spear where we see 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.

[16] (a) Look back to Section 4.2.2 on accuracy and precision. Extra observations are usually helpful in understanding the population, so a point estimate with 400 observations seems more trustworthy. (b) The standard error when the sample size is 100 is given by $SE_{100} = 10/\sqrt{100} = 1$. For 400: $SE_{400} = 10/\sqrt{400} = 0.5$. The larger sample has a smaller standard error. (c) The standard error of the sample with 400 observations is lower than that of the sample with 100 observations. The standard error describes the typical error, and since it is lower for the larger sample, this mathematically shows the estimate from the larger sample tends to be better – though it does not guarantee that every large sample will provide a better estimate than a particular small sample.

If we report a point estimate, we probably will not hit the exact population parameter. There is likely to be some error associated with this estimate. On the other hand, if we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter within our range. As with fishing, the goal of the confidence interval is to include the population parameter within it.

⊙ **Guided Practice 4.7** If we want to be very certain we capture the population parameter, should we use a wider interval or a smaller interval?[17]

⊙ **Guided Practice 4.8** Suppose we have a confidence interval that is 10 units wide and that we are 50% confident that the range encompasses the population parameter. If we had another interval that was instead 5 units wide centered at the same value as our original interval, are we now more or less confident than 50% that the range will include the population parameter? [18]

4.3.2 Confidence levels

The size of our fishing net depends on how confident we want to be in catching a fish. Similarly, the size or width of our confidence intervals depends on how confident we want to be in estimating the true value of the parameter in question. Before we even jump into the calculation of the confidence interval itself, let's first understand what it means to be "confident."

Before scientists embark on most inference processes, they first much choose a confidence level. For a confidence interval, the confidence level is a percent that affects how wide the interval you calculate is. Confidence levels, in general, are associated with a level of uncertainty and how much you are allowing your test to commit a Type I Error or a false positive. Section 4.4.4 goes into more depth on Type 1 and Type 2 Errors.

For example if you wanted to say that we are 75% confident that the population mean BMI is between two values, 75% would be our measure of uncertainty and 25% would be the probability of committing a Type 1 Error. In the context of confidence intervals, there is a 25% chance that the confidence interval does not include the population parameter when in fact, it should. The Type I error is also known as α.

● **Example 4.9** Consider extreme confidence levels. What are the implications of a 100% confidence level confidence interval? How about a 0.001% confidence level?

A 100% confidence level is equivalent to $\alpha = 0\%$. The confidence interval we create will *always* capture the population parameter. Therefore in order to guarantee this, the confidence interval will be $[-\infty, \infty]$. Consider a confidence level of 0.001%. This extremely low confidence level results in a very high Type 1 Error. In many cases when building a 0.001% confidence interval, the confidence interval will not capture the population parameter. Therefore we can foresee this interval being extremely narrow.

[17] 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. The more values we include in our range, the more likely it is that this range contains the true value since the interval contains simply *more* values. However just capturing the parameter with the widest interval is not always the best when constructing a confidence interval. We can always capture the parameter with an interval going from $-\infty$ to $+\infty$ but does range does not increase our understanding of the population parameter.

[18] We are less confident than 50% that the smaller interval includes the population parameter simply because it is smaller and contains fewer values. Using a smaller net, we are less confident that we have captured the true value.

Statisticians generally use a confidence level of 95% per tradition but Section 4.4.6 demonstrates that any confidence level is allowed given varying inference goals in mind. But what does "95% confident" truly mean? Suppose we took many samples and built a confidence interval based on each sample. Then to be 95% confident, we would see approximately 95% of those intervals would contain the actual mean, the population parameter, μ. In this case, if we took 100 independent samples and built 100 confidence intervals, 95 of these confidence intervals would contain the average US BMI and 5 of these would not.

Figure 4.9 shows this process with 25 samples, where 24 of the resulting confidence intervals contain the average BMI for the population and one does not.

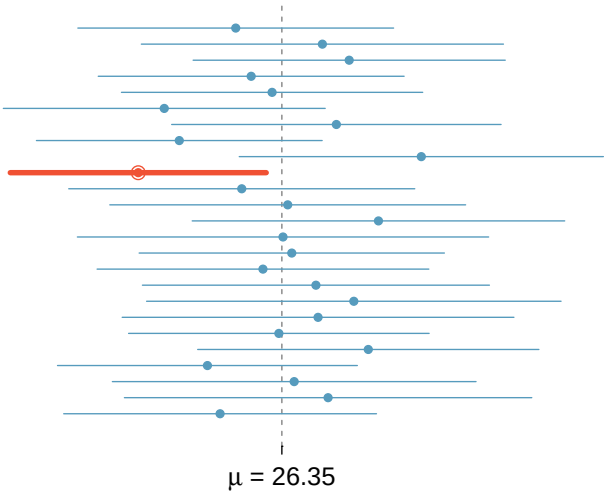


Figure 4.9: Twenty-five samples of size $n = 100$ were taken from the `BRFSS` data set. For each sample, a confidence interval was created to try to capture the average BMI for the population. Only 1 of these 25 intervals did not capture the true mean.

4.3.3 Confidence intervals through computation

Figure 4.9 should give you inspiration on how to achieve an estimate of a 95% confidence interval through computation. Our goal, again, is to find a range of values that hopefully contains the population parameter. If we have the ability to independently resample from our population, the natural method to estimate a confidence interval is through the sampling distribution like in Figure 4.8. We observed there that the mean of the sampling distribution was extremely close to the mean of the population we were sampling from. Thus a reasonable estimation for a 95% confidence interval would be to take the middle 95% of the sampling distribution. Below we have pseudocode that implements this procedure [19]

```
(1) Have a place to store all the sample means that we will calculate
(2) Take a sample from the BRFSS dataset of 40,000
(3) Calculate the sample mean from this specific sample and store it in (1)
```

[19] This highly resembles the pseudocode for approximating a sampling distribution from Section 4.2.3

```
(4) Repeat (2) and (3) many many times
(5) Use the middle 95% of values as the 95% confidence interval.
```

We can build off the method of approximation the sampling distribution to compute a 95% confidence interval. Instead of plotting the values of the sample means (stored as `sample.means` within the *R* code) to create the sampling distribution, we can find the values to get the middle 95% of sample means.

The confidence interval itself is the BMI value c_1 such that 2.5% of the `sample.mean` values is below c_1 and another BMI value c_2 such that 2.5% of the distribution values is greater than c_2. In order to find these values, recall from the distributions unit **??** that we need to use the `quantile()` function in *R*. Particularly using `sample.means` as the vector that that stores the sample means.

```
confidence.interval<-quantile(x=sample.means,c(0.025,0.975))
> confidence.interval
    2.5%    97.5%
26.30493 26.39697
```

We see that the interval (26.30,26.40) is an estimation for a 95% confidence interval using the sampling distribution of sample means. Therefore we can say that we are 95% confident that the population mean BMI is between 26.30 and 26.40. Similarly we can also say that after calculating many confidence intervals from many different observed samples, 95% of all the confidence intervals that we calculated will contain the population mean.

⊙ **Guided Practice 4.10** Say we were interested in creating a 90% confidence interval and a 50% confidence interval. (a) how do you think the widths of the confidence intervals would compare? (b) How would we use the `quantile()` function to find the 90% and 50% confidence intervals from using the array `sample.means`? [20]

4.3.4 Calculating an approximate 95% confidence interval

Computing the confidence interval from resampling is straightforward and serves as a great estimate for any confidence interval for the population. However we ask ourselves, is this realistic beyond simulation? In general do we have the ability to resample the US population independently 100,000 times? Generally no. Most times we cannot observe 100,000 sample means of BMI values let alone 100,000 complete samples from the US population. More often than not, researchers only view 1 sample and thus calculate only 1 sample mean. How then do we calculate a confidence level from only observing one sample mean?

We see from the computation in Section 4.2.3 and Figure 4.8 that the sampling distribution is centered around the population mean and that a confidence interval is derived from that sampling distribution. Therefore it becomes intuitive to build the confidence interval around the observed sample mean as the point estimate is the most plausible value for the population parameter. The width of the interval should encompass the confidence level as well as the uncertainty associated with the point estimate. The standard error becomes a natural measurement of uncertainty for building the interval.

[20] (a) We would expect the 50% confidence interval would have the smaller width. In general the more confident we are, the larger the confidence interval width will be.(b) Remember we want the middle percent of observed sample means. Therefore for the 90% confidence interval using `sample.means` again to store the sample means, the *R* code would be `quantile(x=sample.means,c(0.05,0.95))`. For a 50% confidence interval, the code would be `quantile(x=sample.means,c(0.25,0.75))`

Roughly 95% of the time, the estimate that is observed from resampling will be within approximately 2 standard errors[21] of the population parameter. Therefore we can create a 95% confidence interval that is 2 standard errors from the point estimate on either side of the sample mean. We then can be roughly **95% confident** that we have captured the population parameter with the confidence interval calculated by Equation 4.11 from the sample that we observe:

$$\text{point estimate} \ \pm \ 2 \times SE^{22} \tag{4.11}$$

Using the BMI sample from 4.2.3 with an observed sample mean of 26.36 and sample standard deviation of 5.29, we calculate the 95% confidence interval.

$$\begin{aligned} \text{point estimate} \ &\pm \ 2 \times SE \\ 26.36 &\pm 2 \times \frac{5.29}{\sqrt{40000}} \\ 26.36 &\pm 0.053 \\ (26.31, &\, 26.41) \end{aligned}$$

While not exact, we see that the confidence interval simulated through R achieves a very similar confidence interval as the one above through calculation. The difference is due to randomness within the samples since both the point estimate and the standard error vary among samples.

⊙ **Guided Practice 4.12** In Figure 4.9, one interval does not contain a BMI value of 26.36. Does this imply that the average population BMI cannot be 26.36? [23]

We forewarn that "about 95% of observations are within 2 standard deviations of the mean" is only approximately true. This rule of thumb holds very well for the normal distribution. As we will soon see in Section 4.5, the sample mean tends to be normally distributed when the sample size is sufficiently large.

● **Example 4.13** We are curious about how the average heights of men and women differ and create 95% confidence intervals for the average male height and the average female height. The `BRFSS BMI` data can be divided using the `sex` variable. Among the 40,000 individuals within the `BRFSS BMI`, there are 16,843 men and 23,157 women. The average male height is 70.22 inches and the average female height is 64.38. The sample standard deviations for males and females are 3.00 and 2.80 respectively. What are the 95% confidence intervals for the average male and female height in the US?

We calculate both 95% confidence intervals using the formula

$$\text{point estimate} \ \pm \ 2 \times SE$$

[21] 1.96 to be more precise if the sampling distribution resembles a Normal Distribution. Details coming up in Section 4.3.5

[23] Just as some observations occur more than 2 standard deviations from the mean, some point estimates will be more than 2 standard errors from the parameter. A confidence interval only provides a plausible range of values for a parameter. While we might say other values are implausible based on the data, this does not mean they are absolutely impossible.

using the information given above:

$$\begin{aligned}\text{men: point estimate } &\pm\ 2 \times SE \\ 70.22 &\pm 2 \times \frac{3.00}{\sqrt{16843}} \\ 70.22 &\pm 0.05 \\ (70.17, &\ 70.27)\end{aligned}$$

$$\begin{aligned}\text{women: point estimate } &\pm\ 2 \times SE \\ 64.38 &\pm 2 \times \frac{2.80}{\sqrt{23157}} \\ 64.38 &\pm 0.04 \\ (64.34, &\ 64.42)\end{aligned}$$

The confidence intervals for average height by gender are different. With different centers and different widths, the underlying distributions for male and female heights are different.

The creation of the 95% confidence interval depends on the center and the standard error. In Section 4.3.6, we will see how the multiplier changes beyond 2 standard deviations as confidence levels change.

⊙ **Guided Practice 4.14** The sample data `BRFSS BMI` suggest the average adult's age is about 46.64 years with a standard error of 0.09 years (estimated using the sample standard deviation, 17.35). What is an approximate 95% confidence interval for the average age of US adults?[24]

4.3.5 The sample size for a sampling distribution

In Section 4.2.3, we introduced a sampling distribution for $\bar{x}$, the average BMI value for samples of size 5 and 50. We examined this distribution earlier in Figure 4.8. We see with larger sample sizes like $n = 40,000$, the sampling variation decreases significantly than if $n = 5$ or $n = 50$. In Figure 4.7, the sampling distribution for $n = 5$ was slightly skewed but the sampling distribution for $n = 50$ looks more symmetric. We show in Figure ?refsampDistNormal a histogram of the sample means for 100,000 different random samples of size $n = 50$ with a normal probability plot of those sample means.

Does this distribution look familiar (think back to Chapter 2 of probability distributions)? Hopefully so! The distribution of sample means closely resembles the normal distribution (see Section **??**). A normal probability plot of these sample means is shown in the right panel of Figure 4.10. Because all of the points closely fall around a straight line, we can conclude the distribution of sample means is nearly normal. This result can be explained by the Central Limit Theorem[25].

[24] Again apply Equation (4.11): $46.64 \pm 2 \times 0.09 \rightarrow (46.46, 46.82)$. We interpret this interval as follows: We are about 95% confident the average age of US adults was between 46.46 and 46.82 years. Looking at the entire dataset `BRFSS` that represents the US population more closely (normally we do not have this luxury!) we see that the average age is 46.72 which is indeed within our confidence interval.

[25] A more formal definition coming soon.

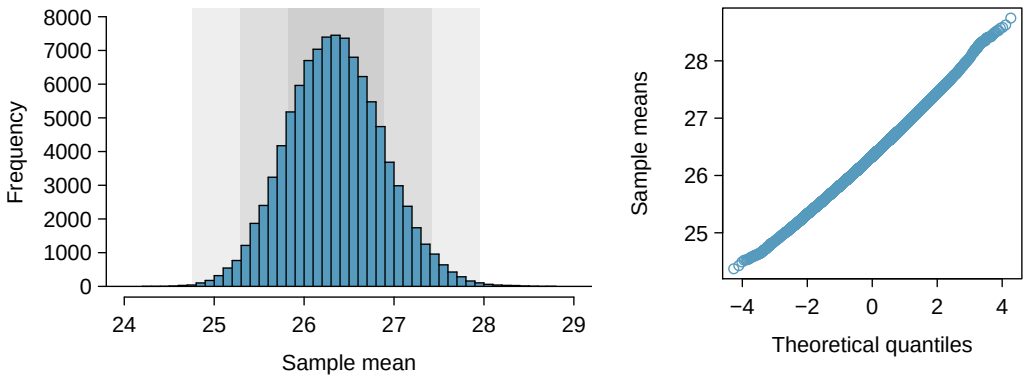


Figure 4.10: The left panel shows a histogram of the sample means for 100,000 different random samples of size $n = 100$. The right panel shows a normal probability plot of those sample means.

Central Limit Theorem, informal description

If a sample consists of at least 30 independent observations and the data are not strongly skewed, then the distribution of the sample mean is well approximated by a normal model.

Why 30?

We introduce the Central Limit Theorem uses this cutoff at 30 in this text but this cutoff varies from book to book As a quick exercise both in statistical exploration but also more practice in algorithmic thinking, think about how you would visually test if 30 independent observations is a sufficient number of observations to approximate the distribution to a normal model. Let's use the `BRFSS` data to sample from like before.

Again remember we are testing if the normal model is a good approximation for the the sampling distribution with a sample size of 30. Creating a sampling distribution for sample sizes of $n = 5, 10, 20, 30$ and overlaying a normal approximation on the histogram is a great guide. [26]

In Figure ?? we see the sampling distributions of the sample mean for sample sizes of 5, 10, 20 and 30. The curve on top is a normal density curve with the normal distribution $\mathcal{N}(0, 1)$ where μ is the mean of the sample means and σ is the standard deviation of the sample means.

We will apply this informal version of the Central Limit Theorem for now, and discuss its details further in Section 4.5.

The choice of using 2 standard errors in Equation (4.11) was based on our general guideline that roughly 95% of the time, observations are within two standard deviations of the mean. Under the normal model, with a sufficient number of samples, we can make this

[26] Use the same code for creating a sampling distribution but vary the sample size. Then use the code: `hist(sample.means, freq=FALSE )`
`curve(dnorm(x,mean=mean(sample.means), sd=sqrt(var(sample.means))), add = TRUE)` where the function `curve()` adds the normal curve on top of the histogram.

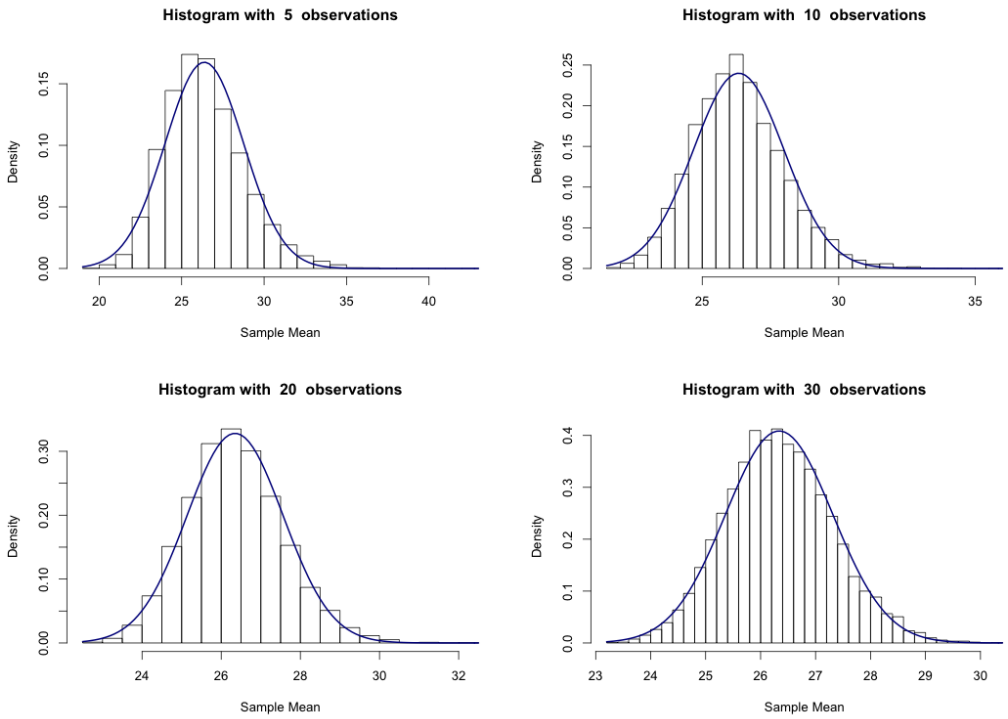


Figure 4.11: The sampling distribution of sample means with different sample sizes. With a normal density curve on top, we see that for $n = 30$, a normal model is a fitting approximation confirming the cutoff for the Central Limit Theorem.

more accurate by using 1.96 in place of 2.

$$\text{point estimate} \ \pm \ 1.96 \times SE \tag{4.15}$$

If a point estimate, such as $\bar{x}$, is associated with a normal model with standard error SE, then we use this more precise 1.96 to create a 95% confidence interval.

4.3.6 Changing the confidence level

Suppose we want to consider confidence intervals where the confidence level is somewhat higher than 95%. Perhaps we would like a confidence level of 99% or even lower like 90%. 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 new confidence levels. Below is a general 95% confidence interval for a point estimate where the point estimate follows a nearly normal distribution.

$$\text{point estimate} \ \pm \ 1.96 \times SE \tag{4.16}$$

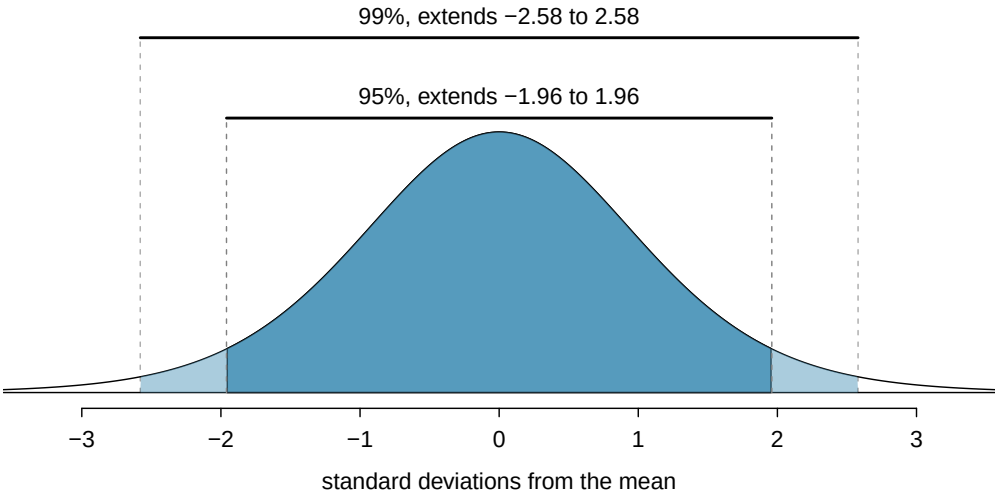


Figure 4.12: The area between -$z^\star$ and $z^\star$ increases as $|z^\star|$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

There are three components to this interval: the point estimate, the "1.96", and the standard error. The $1.96 \times SE$ value affects the confidence interval width, and the point estimate affects where the confidence interval will be centered. Since we know from a normal distribution's Z-score that approximately 95% of data that is normally distributed falls within 1.96 standard deviations of the mean, $1.96 \times SE$ represents the width required to "capture that 95%" of the sampling distribution as seen in Figure ??.

⊙ **Guided Practice 4.17** If X is a normally distributed random variable, how often will X be within 2.58 standard deviations of the mean?[27]

To 99% confident, change 1.96 in the 95% confidence interval formula to be 2.58 for a 99% confidence interval. Exercise 4.17 highlights that 99% of the time a normal random variable will be within 2.58 standard deviations of the mean. This approach – using the Z scores in the normal model to compute confidence levels – is appropriate when $\bar{x}$ is associated with a normal distribution with mean μ and standard deviation $SE_{\bar{x}}$. Thus, the formula for a 99% confidence interval is

$$\bar{x} \pm 2.58 \times SE_{\bar{x}} \tag{4.18}$$

The normal approximation is crucial to the precision of these confidence intervals. Section 4.5 provides a more detailed discussion about when the normal model can safely be applied. When the normal model is not a good fit, we will use alternative distributions

[27] 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 ??.) To determine this probability, look up -2.58 and 2.58 in the normal probability table (0.0049 and 0.9951). Thus, there is a $0.9951 - 0.0049 \approx 0.99$ probability that the unobserved random variable X will be within 2.58 standard deviations of μ.

that better characterize the sampling distribution. Below however is a good checklist to determine whether or not the Central Limit Theorem can be informally applied to the distribution of sampling mean.

Conditions for $\bar{x}$ being nearly normal and SE being accurate

Important conditions to help ensure the sampling distribution of $\bar{x}$ is nearly normal and the estimate of SE sufficiently accurate:

- The sample observations are independent.
- The sample size is large: $n \geq 30$ is a good rule of thumb.
- The population distribution is not strongly skewed. (We check this using the sample distribution as an estimate of the population distribution.)

Additionally, the larger the sample size, the more lenient we can be with the sample's skew.

These three conditions help ensure that $\bar{x}$ is both distributed normally and representative of the target population. If the distribution of $\bar{x}$ is nearly normal, choosing a precise "1.96" or "2.58" becomes much easier for calculating confidence intervals. More importantly, however, the representativeness of the sample is imperative in our ability to infer about the target population. Randomness, independence and a large sample size safeguard against an extreme observation from skewing the conclusions from our sample. These conditions ensure the ability to accurately infer and generalize about the population of interest.

Verifying independence is often the most difficult of the conditions to check, and the way to check for independence varies from one situation to another. However, we can provide simple rules for the most common scenarios.

TIP: How to verify sample observations are independent

Observations in a simple random sample consisting of less than 10% of the population are independent.

Caution: Independence for random processes and experiments

If a sample is from a random process or experiment, it is important to verify the observations from the process or subjects in the experiment are nearly independent and maintain their independence throughout the process or experiment. Usually subjects are considered independent if they undergo random assignment in an experiment or are selected randomly for some process.

⊙ **Guided Practice 4.19** Create a 99% confidence interval for the average weight of men from the `brfss.sample` sample. The point estimate is $\bar{w} = 189.4$ and the standard error is $SE_{\bar{y}} = 0.178$. Refer to Figure 4.13 for guidance on skewness. [28]

[28] The observations are independent (simple random sample, $< 10\%$ of the population), the sample size is at least 30 ($n = 100$), and the distribution is only slightly skewed (Figure 4.13); the normal approximation and estimate of SE should be reasonable. Apply the 99% confidence interval formula: $\bar{y} \pm 2.58 \times SE_{\bar{y}} \rightarrow$

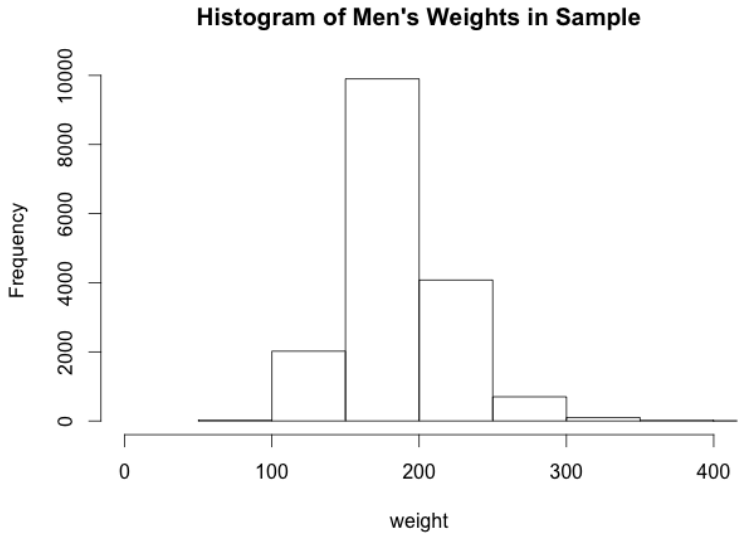


Figure 4.13: We draw a histogram of the men's weights in the `brfss.sample` and note that it is only slightly skewed. With 40,000 observations however, its skewness is more negligible because of its large sample size.

Now that we know how to calculate a 95% and 99% confidence interval given a nearly normally distributed $\bar{x}$, we can generalize this setup to any confidence level we choose. Remember while it has become tradition to use the 95% confidence level, any confidence level is allowed and vary by statistician and by goal.

Confidence interval for any confidence level (nearly normal model)

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

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

where $z^\star$ corresponds to the confidence level selected. The coefficient on the standard error, $z^\star$, is also known as the critical value. Remember that $z^\star$ is only used when the point estimate resembles a normal model [a]

[a] $z^\star$ is also used when the population standard deviation is known. However since we previously mentioned that this is rarely ever the case in practice, we have disregarded this situation completely

(188.94, 189.87). We are 99% confident that the average weight of all males is between 188.94 and 189.87 pounds.

Margin of error

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

Figure ?? provides a picture of how to identify $z^\star$ based on a confidence level. We select $z^\star$ so that the area between -$z^\star$ and $z^\star$ in the normal model corresponds to the confidence level. We note from Figure ?? that the $z^\star$ value comes from a $\mathcal{N}(0, 1)$. Therefore we can either use R or a Z-table [29] (**FOUND IN THE BACK OF THE BOOK HERE**) to find the critical value associated with some confidence level. In R, we use the `qnorm()` function. `qnorm()` takes in a probability p and outputs the quantile value z such that $P(Z \leq z) = p$. For a 95% confidence interval, $p = 0.025$ since we are looking for the *middle* 95%. Therefore in R

```
> qnorm(0.025)
[1] -1.959964
```

and we show that $z^\star = 1.96$ is the critical value for 95%.

⊙ **Guided Practice 4.20** What is the critical value associated with (a) 90%, (b) 75% and (c) 50%? [30]

⊙ **Guided Practice 4.21** Use the data in Exercise 4.19 to create a 90% confidence interval for the average weight of men in the United States.[31]

4.3.7 Interpreting confidence intervals

A careful eye might have observed the somewhat awkward language used to describe confidence intervals. Correct interpretation:

> We are XX% confident that the population parameter is between...

Looking back to **Section 4.2.2**, this means that if we took a random sample from our population 100 times and calculated a confidence interval around our point estimate each time, 95 confidence intervals would contain the true population parameter.

It is interesting to note, however, that researchers in practice would almost never be able to resample 100 times and generate 100 confidence intervals. The meaning of being "95% confident" has traditionally been one grounded in theory and less in practice. "Confidence" relates more to the reliability of the process of creating such a range and less so in the probability that the value is within the range.

Incorrect language might try to describe the confidence interval as capturing the population parameter with a certain probability. This is one of the most common errors: 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.

[29] also known as a Normal table

[30] Remember we want the *middle* and for any given confidence level C, we type into R, `qnorm(0.5·(1−C))`. Therefore (a) `qnorm(0.05)= -1.644854` so $z^\star = 1.65$ for a 90% confidence level (b) 1.15 (c) 0.67

[31] We first find $z^\star$ such that 90% of the distribution falls between -$z^\star$ and $z^\star$ in the standard normal model, $N(\mu = 0, \sigma = 1)$. We can look up -$z^\star$ in the normal probability table by looking for a lower tail of 5% (the other 5% is in the upper tail), thus $z^\star = 1.65$. The 90% confidence interval can then be computed as $\bar{y} \pm 1.65 \times SE_{\bar{y}} \rightarrow (189.11, 189.69)$. (We had already verified conditions for normality and the standard error in the previous exercise.) That is, we are 90% confident the average weight of males is between 189.11 and 189.69 pounds. Also note that because we are at a 90% confidence level, our confidence interval width is smaller than in Exercise 4.19.

Another especially important consideration of confidence intervals is that they *only try to capture the population parameter*. Our intervals say nothing about the confidence of capturing individual observations, a proportion of the observations, a percent of all the data or just the sampled data. A confidence interval also says nothing about capturing point estimates since the confidence interval is always centered at the observed point estimate. Confidence intervals only attempt to capture population parameters as statistical inference's goal is to infer about such population parameters.

Some incorrect interpretations of a 95% confidence interval include:

95% of the observed data is between ...
95% of the population distribution is contained in the confidence interval.

Remember, a confidence interval is not a range of plausible values for the sample mean, though it may be understood as an estimate of plausible values for the population parameter. A particular confidence interval of 95% calculated from an experiment does not mean that there is a 95% probability of a sample mean from a repeat of the experiment falling within this interval.[13]

While the differences in correct and incorrect interpretations are extremely nuanced, the goal of this book is to provide the tools and mechanisms of calculating and computing a confidence interval from data and less so about the wording which, in practice, has become almost meaningless and obsolete.

4.3.8 Nearly normal population with known SD (special topic)

In rare circumstances we know important characteristics of a population. For instance, we might already know a population is nearly normal and we may also know its parameter values. Even so, we may still like to study characteristics of a random sample from the population. Consider the conditions required for modeling a sample mean using the normal distribution:

(1) The observations are independent.

(2) The sample size n is at least 30.

(3) The data distribution is not strongly skewed.

These conditions are required so we can adequately estimate the standard deviation of the population from our sample and so we can ensure the distribution of sample means is nearly normal. However, if the population is known to be nearly normal, we know that the sample mean is always nearly normal (this is a special case of the Central Limit Theorem). If the standard deviation for the population is also known, then conditions (2) and (3) are not necessary for those data.

We would like to heavily emphasize however that while, in practice, the population mean is more likely to be known, the population standard deviation is rarely know. While a known population standard deviation will rarely occur in practice, the Central Limit Theorem allows us to describe the distribution of the sampling distribution more specifically.

Example 4.22 The heights of male seniors in high school closely follow a normal distribution $N(\mu = 70.43, \sigma = 2.73)$, where the units are inches.[32] If we randomly sampled the heights of five male seniors, what distribution should the sample mean follow?

[32] These values were computed using the USDA Food Commodity Intake Database.

The population is nearly normal, the population standard deviation is known, and the heights represent a random sample from a much larger population, satisfying the independence condition. Therefore the sample mean of the heights will follow a nearly normal distribution with mean $\mu = 70.43$ inches and standard error $SE = \sigma/\sqrt{n} = 2.73/\sqrt{5} = 1.22$ inches.

Alternative conditions for applying the normal distribution to model the sample mean

If the population of cases is known to be nearly normal and the population standard deviation σ is known, then the sample mean $\bar{x}$ will follow a nearly normal distribution $N(\mu, \sigma/\sqrt{n})$ if the sampled observations are also independent.

Sometimes the mean changes over time but the standard deviation remains the same. In such cases, a sample mean of small but nearly normal observations paired with a known standard deviation can be used to produce a confidence interval for the current population mean using the normal distribution.

TIP: Relaxing the nearly normal condition

As the sample size becomes larger, it is reasonable to *slowly* relax the nearly normal assumption on the data when dealing with small samples. By the time the sample size reaches 30, the data must show strong skew for us to be concerned about the normality of the sampling distribution.

4.4 Hypothesis testing

Is the average US adult satisfied with his or her weight? We consider this question in the context of the `BRFSS` dataset comparing US adults' current weight and their desired weight (we will call this "weight difference"). While media pressures women to maintain a slim figure, the same media urges men to work out more and become stronger and more fit. These opposing viewpoints and many others all are components that influence satisfaction with weight and the desire to lose or gain weight.

In addition to considering weight in this section, we consider a topic near and dear to most students: sleep. A recent study found that college students average about 7 hours of sleep per night.[33] However, researchers at a rural college are interested in showing that their students sleep longer than seven hours on average. We investigate this topic in Section 4.4.2.

Many questions, given the correct data, can be answered through Hypothesis Testing. **Hypothesis testing** is a method in statistics that evaluates whether or not a population parameter has a hypothesized value with an associated probability of error. It is, most obviously, determining the probability that a given hypothesis is true or not.

Hypotheses are often simple questions that have a yes or no answer. Consider some hypotheses below:

[33] http://theloquitur.com/?p=1161

Is the mean body temperature really 98.6F?
Has consumption of soda changed across the US overtime?
Do MCAT classes improve MCAT scores?

The hypothesis testing process consists of generally 5 steps. Going through the **Hypothesis testing framework** allows for statisticians to answer these yes/no questions with a certain degree of confidence after observing a related sample. We begin by testing a hypothesis about a population mean from observing one sample. Remember, we can do hypothesis testing on any population parameter. It can be the population mean, population standard deviation or even the population IQR if desired.

4.4.1 Hypothesis testing framework

The average weight difference that adults want to experience that we observe from our sample of the `brfss.sample` data is 15.01 lbs. We want to determine if this sample provides enough evidence that adults are satisfied with their weight versus the alternative – that they are not.[34] We use desired weight difference as a proxy for weight satisfaction and simplify this question into two **hypotheses**

H_0: US adults are satisfied with their current weight. The average desired weight difference for US adults is 0 lbs.

H_A: The average adult's desired weight difference is not 0 lbs i.e. Average adults are not satisfied with their current weights and would like to change.

Step 1: Formulating Hypotheses

The first step within the hypothesis testing framework is setting up the hypotheses. As shown above, we generally have two hypotheses, a null and an alternative.

We call H_0 the null hypothesis and H_A the alternative hypothesis.

H_0
null hypothesis

H_A
alternative hypothesis

Null and alternative hypotheses

The **null hypothesis** (H_0) often represents either 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.

The null hypothesis often represents a skeptical position. The null hypothesis is generally denoted as "no difference" or what one would observe if there is no change. The alternative hypothesis often represents a new perspective, such as the possibility that there has been a change. If the null hypothesis is true, any difference between the observed sample is due only to chance variation.

TIP: Hypothesis testing framework

The logic of hypothesis testing is that we will not reject the null hypothesis (H_0), unless the evidence in favor of the alternative hypothesis (H_A) is so strong that we must reject H_0 in favor of H_A.

[34] While we could answer this question by examining the entire population data (`BRFSS`), we only consider the sample data (`brfss.sample`), which is more realistic since statisticians rarely have access to population data.

The first step within 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. We believe our null hypothesis H_0. However, if there is sufficient evidence that we observe that supports the claim, we set aside our skepticism and reject the null hypothesis in favor of the alternative.

⊙ **Guided Practice 4.23** A new study would like to be published in a scientific journal. The board that determines the validity of the study considers two possible claims about this study: either the study is valid or pseudoscience. If we set these claims up in a hypothesis framework, which would be the null hypothesis and which the alternative? [35]

Those scientists who sit on the board of publication journals look at the study, previous literature and other evidence to see whether it convincingly supports that the science is valid. Even if these scientists leave unconvinced that the study is publishable, this does not mean that these board members believe the study is complete fabrication. 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.

TIP: 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.

In the example with the `BRFSS` data, the null hypothesis represents no change in desired weight difference. The alternative hypothesis represents something new or more interesting: there was a difference, either a desire to gain or lose weight on average. These hypotheses can be described in mathematical notation using μ_{wd} as the average weight difference for US adults.

$$H_0 : \mu_{wd} = 0 \qquad H_A : \mu_{wd} \neq 0$$

where 0 represents a desired weight difference of 0 lbs or that these US adults on average do not care to change their weight. Using this mathematical notation, the hypotheses can now be evaluated using statistical tools. We call 0 the **null value** since it represents the value of the parameter if the null hypothesis is true. We will use the `brfss.sample` data set to evaluate the hypothesis test.

Note it is important to remember that we are not testing whether or not the average weight difference observed from the `brfss.sample` is 0 or not. We don't need to test that since we have observed all of `brfss.sample` and can simply calculate it. Rather we are testing the *population parameter* or the true average value of all US adults' weight differences is 0 or not.

[35] The board considers whether the study's evidence, results and reproducibility is so convincing (strong) that there the study must be valid. In this case the board rejects the null hypothesis (the study is pseudoscience) and concludes that the study is valid an should be published (alternative hypothesis).

TIP: Null and Alternative Hypothesis Setup
The null hypothesis is generally written as $H_0 : \mu = \mu_0$ where μ is the population mean and μ_0 is the hypothesized value that we believe to be true.
The alternative hypothesis, on the other hand, can be many things.
If we have no prior belief to influence our alternative hypothesis and the researchers are interested in showing any difference –an increase or decrease– then the safest one would be $\mu \neq \mu_0$ or a two-sided alternative. If we have a prior belief of how μ and μ_0 compare or are interested in only showing an increase or decrease, but not both, we can do a one-sided alternative, $\mu \geq \mu_0$ or $\mu \leq \mu_0$. We will go into more detail on one-side versus two sided in Section 4.4.2.

Step 2: Specifying a Significance Level α

Once we have completed Step 1 and have a null and alternative hypothesis, we need to specify a **significance level**. The significance level α is the acceptable error probability of the test. In this case, the error probability is the probability of concluding the alternative hypothesis is true when it is not true. This error is called a Type I error, and α is the probability of a Type I error. We will go into more detail on error types in Section 4.4.4.

Typically, α is taken to be 0.05, 0.01, or some other small value. α plays the same role as the error probability in confidence intervals, and is a measure of uncertainty. If $\alpha = 0.05$, we are testing at a 95% confidence level for our hypothesis tests. We will see a clearer connection between hypothesis testing and confidence intervals in Section 4.4.3.

Step 3: Calculating the Test Statistic

The third step is calculating a test statistic from the data we observe. This statistic will be the value that the conclusions will be based on and measures the difference between the observed data and what is expected if the null hypothesis is true. This test statistic answers the question: "How many standard deviations from the hypothesized value is the observed sample value?" Thinking back to **THIS SECTION ??** by standardizing a normal, the test statistic follows a similar construction. When testing hypotheses about a mean, the test statistic for the population mean from one sample will always be

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation and n is the number of observations in the sample. *Note:* In general we see that test statistics follow $\frac{\text{observed}-\text{hypothesized}}{\text{standarderror}}$ to see how many standard deviations the observed value is from the hypothesized value. This T-statistic follows a t-distribution [36] and will have $n - 1$ degrees of freedom.

[36] from **Chapter 3**

Test statistic

A *test statistic* is a special summary statistic that is particularly useful for evaluating a hypothesis test or identifying the p-value. The test-statistic is a particular data summary that summarizes how many standard deviations from the hypothesized null value is the observed sample value. In general the T-statistic follows a t-distribution with $n-1$ degrees of freedom. [a]

[a] When a point estimate is nearly normal, we use the Z score of the point estimate as the test statistic. In later chapters we encounter situations where other test statistics are helpful.

Step 4: Calculating the p-value

Once we calculate a test statistic from the observed data, we know how many standard deviations our observation is from the hypothesized value if the null hypothesis were true. Now we need to tie this T-statistic value to a probability of such an observation happening. We do this through the **p-value**. Assuming the null hypothesis is true, the p-value is the probability of observing our sample or a more extreme sample. Formally the p-value is a conditional probability.

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 is true. We typically use a summary statistic of the data, in this chapter the sample mean, to help compute the p-value and evaluate the hypotheses.

How do we get this probability? Section 4.4.2 will go into more detail from using *R*, Z and t- tables.

Step 5: Making your conclusion

The final step within the hypothesis testing framework is to make a conclusion from the p-value we calculated in Step 4. Using the definition of p-value, if we observe something extreme, the probability associated with our observation will be small. Thus if our observation is rare, the T-statistic and p-value provide evidence that the hypothesized value is unlikely. Therefore if our p-value is low, we should reject our null hypothesis, and the smaller the p-value, the stronger the evidence we have against the null hypothesis.

How small is small? This is where Step 2 and our significance level comes in. If the p-value is small or smaller than the pre-specified α level (usually 0.01 or 0.05), we reject the null hypothesis and say the result that we observe is statistically significant at the α level.

If the p-value is α or greater, we simply do not have enough evidence to reject the null hypothesis. The subtle but important point is that not rejecting H_0 is not equivalent to accepting H_0 (refer back to Example 4.23). In practice, however, not rejecting H_0 is equivalent to accepting H_0 when making decisions and acting on conclusions. Most importantly, it is key that students state the conclusion in the context of the original problem, using the language and units of that problem. Most students forget this but is absolutely necessary in both theory and practice.

4.4.2 Calculating p-values

Calculating p-values can be the most difficult part of the hypothesis testing framework. The p-value depends on many moving parts, including the sample mean, the sample size and the alternative hypothesis but always remember that the p-value is the probability of observing data as extreme or more if we assume the null hypothesis is true with the data at least as favorable to the alternative hypothesis. If the p-value is small, then our sample indicates that we just observed something rare, so rare that we should probably reject the null hypothesis as true. Figure 4.14 shows the distribution of the sample mean where the p-value is the shaded area for a one sided alternative $\mu > \mu_0$.

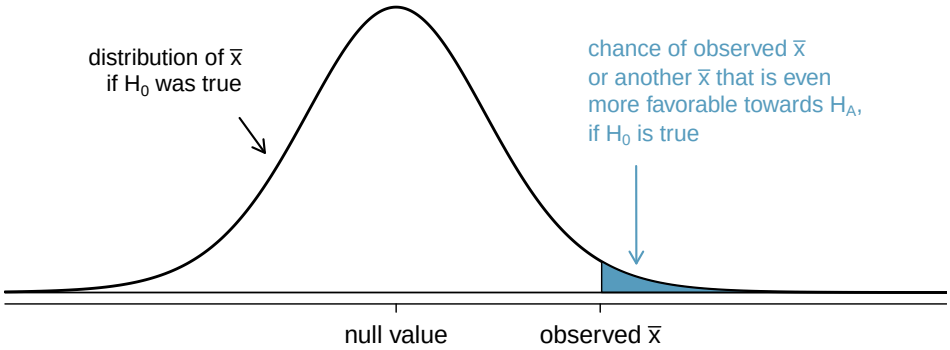


Figure 4.14: To identify the p-value, the distribution of the sample mean is considered as if the null hypothesis was true. Then the p-value is defined and computed as the probability of observing the observed $\bar{x}$ or an $\bar{x}$ even more extreme and thus favorable to follow H_A under this distribution.

If the alternative is one sided and has the form $\mu > \mu_0$, then the p-value would be represented by the upper tail (Figure 4.14). If the alternative is one sided but has the form $\mu < \mu_0$, then the p-value would be the shaded area in the lower tail. In a two-sided test, *we shade two tails* since evidence in either direction is favorable to H_A (Figure 4.19).

Now that we know what the p-value represents, how do we actually get this shaded area to be a number? Here is where the T-statistic comes into play. Before we get to the

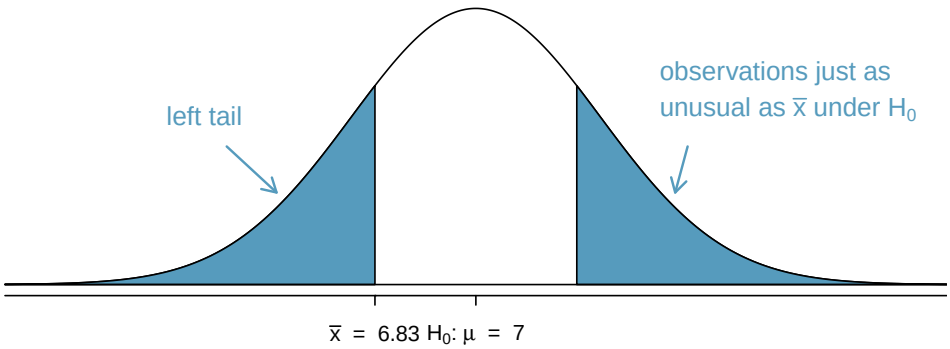


Figure 4.15: H_A is two-sided, so *both* tails must be counted for the p-value.

nitty gritty, let's look back to the `BRFSS` data.

Recall that the researchers for the `BRFSS` data are interested if US adults are satisfied with their current weight. They believe that the desired weight difference is a good proxy to measure satisfaction and have the following null and alternative hypotheses where μ_{wd} denotes the average desired weight difference in the US:

H_0: $\mu_{wd} = 0$
H_A: $\mu_{wd} \neq 0$

Instead of 40,000 within our sample, let's say that we observed a sample of 100 people and calculated a sample mean of weight differences of 0.5 pounds and standard deviation of 5 pounds. Given this information we can first calculate a T-statistic[37].

$$t = \frac{0.5 - 0}{25/\sqrt{100}} = 0.2$$

The T-statistic can be thought of as a Z-score (standard score) that indicates how many standard deviations the observed sample mean is from the null value. This standardization becomes a great way to unify all the moving parts in order to calculate the p-value.

With the T-statistic and the alternative hypothesis, we calculate the p-value from either a t-distribution or a normal distribution. The sample size determines which distribution to model our point estimate from, either a t-distribution or a normal distribution. If $n \geq 30$ **from this part**, the sample mean can be thought of coming from a normal distribution. If $n < 30$, model the sampling distribution from a t-distribution.

With $\alpha = 0.05$, students can either use a table to calculate the p-value or use *R*. First assuming a t-distribution, use the t-table given **on some page** to find the row with the correct degrees of freedom (in a one sample test, $df = n - 1$). Students then should look across that row to find the T-statistic value that they calculated. Note that the table won't have every single value listed but once they find the approximate T-statistic, look at the top of the column to get p-value using either a one sided or two sided (one tail or two tail) alternative. The normal table (Z-table) is very similar but be wary of that the Z-table only lists the areas left of the Z-score. This simply means that these probabilities coincide with a one-sided alternative. However because the normal distribution is symmetric, finding the p-value for a two sided alternative is just those values from the table times two!

If available, *R* is also a handy tool. Use the `pt()` or the `pnorm()` function to calculate the area left of the T-statistic. Students then can take the value and subtract from one or multiply by two depending on the alternative hypothesis. If students have the ability to use *R*, the $n \geq 30$ threshold can be loosened since modeling after the t-distribution becomes easier and more accurate compared to the tables. However we note again that once $n \geq 30$, both distributions become almost equal.

We use a normal table for calculating the p-value for our sample from the `BRFSS` data because $n = 100$ in our sample. A score of 0.2 corresponds to a shaded area of 0.579 to the left. Therefore in the tail we have

$$\begin{aligned} p &= Pr(T \leq -0.2) + Pr(T \geq 0.2) \\ &= Pr(|T| \geq 0.2) \\ &= 2Pr(T \geq 0.2) \\ &= 2 \cdot (1 - 0.579) \\ &= 0.842 \end{aligned}$$

[37] calculating the T-statistic using actual data is an exercise in the book

Using R, we use both `pnorm()` and `pt()` to check.

```
> 2*(1-pnorm(0.2))
[1] 0.8414806
> 2*(1-pt(0.2, 99))
[1] 0.8418908
```

We see that both output p-values that are extremely similar and agree with the p-value from the normal table as well. Now that we calculated the p-value we can conclude that this p-value $> \alpha = 0.05$ so we cannot reject H_0. To put it into context: from observing a sample mean of 0.5 for weight differences, we observed a p-value of 0.84 and cannot conclude that weight difference is nonzero. From our sample it appears that US adults are satisfied with their weight.

⊙ **Guided Practice 4.24** If the null hypothesis is true, how often should the p-value be less than 0.05?[38]

TIP: Concluding on Critical Values

Conclusions are made from the p-value but if $\alpha = 0.05$ or some other common value, we can take a quick shortcut using the critical value. We learned the critical value as the coefficient on the standard error to calculate the confidence interval. However the critical value is also the point on the test distribution that can be compared to the T-statistic in hypothesis testing. Since we know that the critical value is associated with some confidence level, this critical value is also associated with α. If the absolute value of the T-statistic is greater than the critical value (more extreme), the p-value is less than α and you can reject the null hypothesis.

Caution: Critical value $\neq$ test statistic

Many times students get confused between the critical value and the test statistic. The critical value is associated with some α and does not change. For a specific α there is only one critical value. The T-statistic can change depending on the sample that you observed. Students are comparing their T-statistic to the critical value using the critical value as a benchmark.

⊙ **Guided Practice 4.25** A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. Researchers at a rural school are interested in showing that students at their school sleep longer than seven hours on average, and they would like to demonstrate this using a sample of students. What would be an appropriate skeptical position for this research?[39]

We can set up the null hypothesis for this test as a skeptical perspective: the students at this school average 7 hours of sleep per night. The alternative hypothesis takes a new form reflecting the interests of the research: the students average more than 7 hours of sleep. We can write these hypotheses as

$$H_0 : \mu = 7 \qquad H_A : \mu \geq 7$$

[38] About 5% of the time. If the null hypothesis is true, then the data only has a 5% chance of being in the 5% of data most favorable to H_A.

[39] A skeptic would have no reason to believe that sleep patterns at this school are different than the sleep patterns at another school.

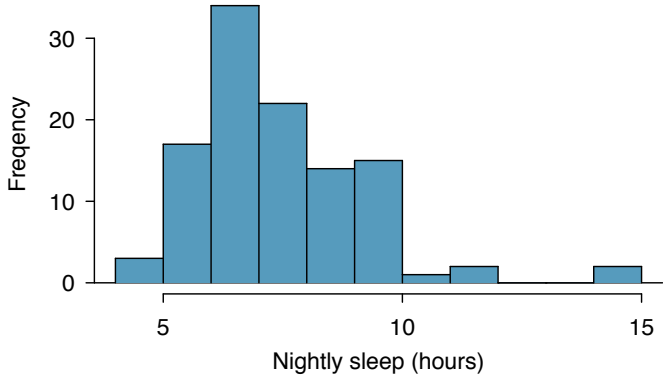


Figure 4.16: Distribution of a night of sleep for 110 college students. These data are moderately skewed.

Using $\mu \geq 7$ as the alternative is an example of a **one-sided** hypothesis test mentioned previously. In this investigation, there is no apparent interest in learning whether the mean is less than 7 hours.[40] Earlier we encountered a **two-sided** hypothesis where we looked for any clear difference, greater than or less than the null value.

Always use a two-sided test unless it was made clear prior to data collection that the test should be one-sided. Switching a two-sided test to a one-sided test after observing the data is dangerous because it can inflate the chance of an incorrect conclusion.

TIP: One-sided and two-sided tests
If the researchers are only interested in showing an increase or a decrease, but not both, use a one-sided test. If the researchers would be interested in any difference from the null value – an increase or decrease – then the test should be two-sided.

TIP: Always write the null hypothesis as an equality
We will find it most useful if we always list the null hypothesis as an equality (e.g. $\mu = 7$) while the alternative always uses an inequality (e.g. $\mu \neq 7$, $\mu \geq 7$, or $\mu \leq 7$).

The researchers at the rural school conducted a simple random sample of $n = 110$ students on campus. They found that these students averaged 7.42 hours of sleep and the standard deviation of the amount of sleep for the students was 1.75 hours. A histogram of the sample is shown in Figure 4.16.

Before we can use a normal model for the sample mean or compute the standard error of the sample mean, we must verify conditions. (1) Because this is a simple random sample from less than 10% of the student body, the observations are independent. (2) The sample size in the sleep study is sufficiently large since it is greater than 30. (3) The data show moderate skew in Figure 4.16 and the presence of a couple of outliers. This skew and the

[40] This is entirely based on the interests of the researchers. Had they been only interested in the opposite case – showing that their students were actually averaging fewer than seven hours of sleep but not interested in showing more than 7 hours – then our setup would have set the alternative as $\mu \leq 7$.

outliers (which are not too extreme) are acceptable for a sample size of $n = 110$. With these conditions verified, the normal model can be safely applied to $\bar{x}$ and the estimated standard error will be very accurate.

⊙ **Guided Practice 4.26** What is the standard deviation associated with $\bar{x}$? That is, estimate the standard error of $\bar{x}$.[41]

The hypothesis test will be evaluated using a significance level of $\alpha = 0.05$. We want to consider the data under the scenario that the null hypothesis is true. In this case, the sample mean is from a distribution that is nearly normal and has mean 7 and standard deviation of about 0.17. Such a distribution is shown in Figure 4.17.

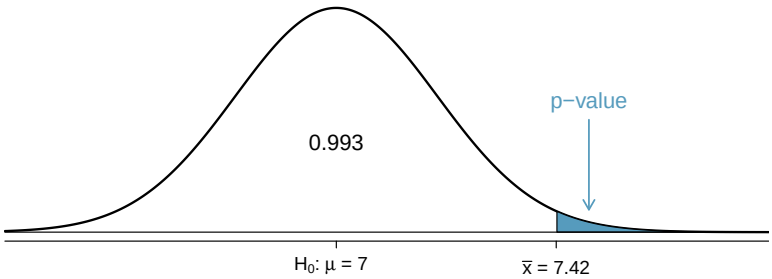


Figure 4.17: If the null hypothesis is true, then the sample mean $\bar{x}$ came from this nearly normal distribution. The right tail describes the probability of observing such a large sample mean if the null hypothesis is true.

Remember the shaded tail in Figure 4.17 is the p-value and so we shade all means larger than our sample mean, $\bar{x} = 7.42$, because they are more favorable to the alternative hypothesis than the observed mean. We compute the p-value by first computing the T-statistic for the sample mean, $\bar{x} = 7.42$:

$$T = \frac{\bar{x} - \text{null value}}{SE_{\bar{x}}} = \frac{7.42 - 7}{0.17} = 2.47$$

Using the normal probability table, the lower unshaded area is found to be 0.993. Thus the shaded area is $1 - 0.993 = 0.007$. Using R we have

```
> 1-pnorm(2.47)
[1] 0.006755653
```

If the null hypothesis is true, the probability of observing such a large sample mean for a sample of 110 students is only 0.007. That is, if the null hypothesis is true, we would not often see such a large mean.

We evaluate the hypotheses by comparing the p-value to the significance level. Because the p-value is less than the significance level (p-value $= 0.007 < 0.05 = \alpha$), we reject the null hypothesis.[42] What we observed is so unusual with respect to the null hypothesis that it casts serious doubt on H_0 and provides strong evidence favoring H_A.

[41] The standard error can be estimated from the sample standard deviation and the sample size: $SE_{\bar{x}} = \frac{s_{\bar{x}}}{\sqrt{n}} = \frac{1.75}{\sqrt{110}} = 0.17$.

[42] Using critical values instead, we know that for $\alpha = 0.05$ and a one sided alternative, the critical value is 1.65. Since our T-statistic is greater than 1.65, we know to reject H_0 without calculating the actual p-value

p-value as a tool in hypothesis testing

The p-value quantifies how strongly the data favor H_A over H_0. A small p-value (usually < 0.05) corresponds to sufficient evidence to reject H_0 in favor of H_A.

TIP: It is useful to first draw a picture to find the p-value

It is useful to draw a picture of the distribution of $\bar{x}$ as though H_0 was true (i.e. μ equals the null value), and shade the region (or regions) of sample means that are at least as favorable to the alternative hypothesis. These shaded regions represent the p-value.

⊙ **Guided Practice 4.27** Suppose we had used a significance level of 0.01 in the sleep study. Would the evidence have been strong enough to reject the null hypothesis? (The p-value was 0.007.) What if the significance level was $\alpha = 0.001$? [43]

4.4.3 Testing hypotheses using confidence intervals

While confidence intervals may appear separate from hypothesis testing, these two concepts arrive as the same conclusions. Consider a sample of 100 people from the `BRFSS` data to test if the average age of adults is 36.8 years [44]. The hypothesis setup would be $H_0 : \mu_{\text{age}} = 36.8$ and $H_A : \mu_{\text{age}} \neq 36.8$. We learned in Section 4.2 that there is fluctuation from one sample to another, and it is very unlikely that the sample mean will be exactly equal to our parameter; we should not expect $\bar{x}_{\text{ages}}$ to exactly equal μ_{ages} and the difference could be due to *sampling variation*, i.e. the variability associated with the point estimate when we take a random sample.

In Section 4.3, confidence intervals were introduced as a way to find a range of plausible values for the population mean. From `BRFSS`, the sample has a mean of 46.48 and a standard deviation of 16.83. Therefore the 95% confidence interval is

$$46.48 \pm 1.96 \cdot \frac{16.83}{\sqrt{100}} = (43.1796, 49.7804)$$

Because 36.8 years does not fall in the range of plausible values, we can say the null hypothesis is implausible. That is, we failed to reject the null hypothesis, H_0.

⊙ **Guided Practice 4.28** An investigator is studying the results of standardized IQ tests in adolescents who suffered from severe asthma during childhood. She claims that those who had childhood asthma perform worse. For the standardized test she will use, the population mean score is 100. What are the null and alternative hypotheses to test whether this claim is accurate? [45]

● **Example 4.29** In her sample of 100 children, she found a sample mean $\bar{x} = 96.7$ and standard deviation $s = 10$. Construct a 95% confidence interval for the population mean and evaluate the hypotheses of Exercise 4.28.

[43] We reject the null hypothesis whenever *p-value* $< \alpha$. Thus, we would still reject the null hypothesis if $\alpha = 0.01$ but not if the significance level had been $\alpha = 0.001$.

[44] as calculated by the US Census in 2009

[45] H_0: The average score is 100, $\mu = 100$. H_A: The average score is lower, $\mu \leq 100$.

$$SE = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$$

The normal model may be applied to the sample mean because the conditions are met: The data are a simple random sample and we assume that there are more than 1,000 adolescents who have suffered from asthma. The observations are independent and the sample size is also sufficiently large (n=100). We don't know about existing outliers but the sample size mitigates potential effects of outliers. This ensures a 95% confidence interval may be accurately constructed:

$$\bar{x} \pm z^\star SE \quad \rightarrow \quad 96.7 \pm 1.96 \times 1 \quad \rightarrow \quad (94.74, 98.66)$$

Because the null value 100 is not in the confidence interval, a true mean of 100 is implausible and we reject the null hypothesis. The data provide statistically significant evidence that adolescents who suffered from severe asthma during childhood do perform worse on standardized IQ tests.

4.4.4 Decision errors

Hypothesis tests are not flawless. Just think of the court system: innocent people are sometimes wrongly convicted and the guilty sometimes walk free. Similarly, we can make a wrong decision in statistical hypothesis tests. However, the difference is that we have the tools necessary to quantify how often we make such errors.

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 in a hypothesis test, which are summarized in Table 4.18.

		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

Table 4.18: 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 4.30** 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? Table 4.18 may be useful.[46]

⊙ **Guided Practice 4.31** How could we reduce the Type 1 Error rate in US courts? What influence would this have on the Type 2 Error rate?[47]

[46] If the court makes a Type 1 Error, this means the defendant is innocent (H_0 true) but wrongly convicted. 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).

[47] 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 4.32** How could we reduce the Type 2 Error rate in US courts? What influence would this have on the Type 1 Error rate?[48]

⊙ **Guided Practice 4.33** Consider a person getting tested for HIV. What does a Type 1 and Type 2 Error represent in this context? [49]

Exercises 4.30-4.32 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 which is the same significance level from hypothesis testing and confidence intervals. We often write the significance level using α where $\alpha = 0.05$. We discuss the appropriateness of different significance levels in Section 4.4.6.

α
significance level of a hypothesis test

If we use a 95% confidence interval to test a hypothesis where the null hypothesis is 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 $\alpha = 0.01$.

4.4.5 Two-sided versus One-sided hypothesis testing: Dos and Don'ts

Determining an alternative hypothesis can get tricky, and the choice between a one-sided and two sided test can be controversial. In this book, the examples and exercises will be obvious enough to decide a correct alternative hypothesis. In practice with real world data, however, can be less straightforward. If the sidedness is uncertain, many scientists opt to use a two-sided alternative because it is more *conservative*. What does conservative in this context mean? Let's first consider the differences.

It is never okay to change two-sided tests to one-sided tests after observing the data. In this example we explore the consequences of ignoring this advice. Using $\alpha = 0.05$, we show that freely switching from two-sided tests to one-sided tests will cause us to make twice as many Type 1 Errors as intended. [50]

⊙ **Guided Practice 4.34** Earlier we talked about a research group investigating whether the students at their school slept longer than 7 hours each night. Let's consider a second group of researchers who want to evaluate whether the students at their college differ from the norm of 7 hours. Write the null and alternative hypotheses for this investigation.[51]

[48] 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.

[49] Type 1 Error is if this person does not have HIV but was tested positive for HIV. Type 2 Error would be failing to detect HIV when the patient actually has HIV.

[50] hence to be conservative and safe, we opt to minimize the Type 1 Errors and use the two sided alternative

[51] Because the researchers are interested in any difference, they should use a two-sided setup: $H_0 : \mu = 7$, $H_A : \mu \neq 7$.

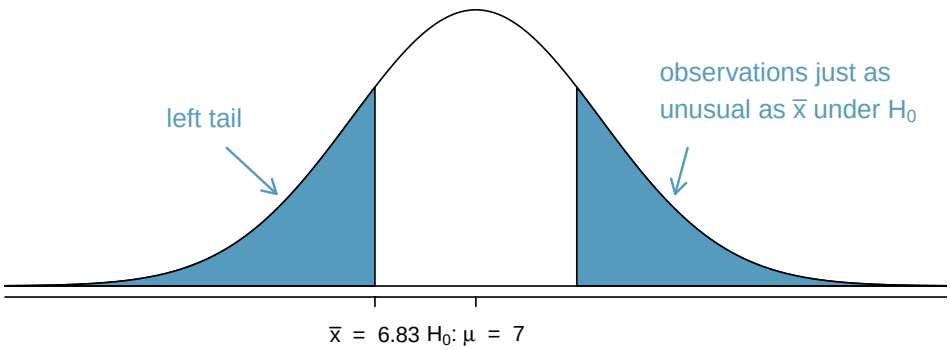


Figure 4.19: H_A is two-sided, so *both* tails must be counted for the p-value.

Example 4.35 The second college randomly samples 72 students and finds a mean of $\bar{x} = 6.83$ hours and a standard deviation of $s = 1.8$ hours. Does this provide strong evidence against H_0 in Exercise 4.34? Use a significance level of $\alpha = 0.05$.

First, we must verify assumptions. (1) A simple random sample of less than 10% of the student body means the observations are independent. (2) The sample size is 72, which is greater than 30. (3) Based on the earlier distribution and what we already know about college student sleep habits, the distribution is probably not strongly skewed.

Next we can compute the standard error ($SE_{\bar{x}} = \frac{s}{\sqrt{n}} = 0.21$) of the estimate and create a picture to represent the p-value, shown in Figure 4.19. Both tails are shaded. An estimate of 7.17 (6.83+1.65*cdot*0.21) or more provides at least as strong of evidence against the null hypothesis and in favor of the alternative as the observed estimate, $\bar{x} = 6.83$.

We can calculate the tail areas by first finding the lower tail corresponding to $\bar{x}$:

$$T = \frac{6.83 - 7.00}{0.21} = -0.81 \quad \overset{table}{\rightarrow} \quad \text{left tail} = 0.2090$$

Because the normal model is symmetric, the right tail will have the same area as the left tail. The p-value is found as the sum of the two shaded tails:

$$\text{p-value} = \text{left tail} + \text{right tail} = 2 \times (\text{left tail}) = 0.4180$$

This p-value is relatively large (larger than $\alpha = 0.05$), so we should not reject H_0. That is, if H_0 is true, it would not be very unusual to see a sample mean this far from 7 hours simply due to sampling variation. Thus, we do not have sufficient evidence to conclude that the mean is different than 7 hours.

Example 4.36 Let's consider two cases: (1) The sample mean was larger than the null value and (2) the sample mean as smaller than the null value.

Suppose the sample mean was larger than the null value, μ_0 (e.g. μ_0 would represent 7 if H_0: $\mu = 7$). Then if we can flip to a one-sided test instead of a two-sided test, we would use H_A: $\mu > \mu_0$. Now if we obtain any observation with a T-statistic greater than 1.65, we would reject H_0. If the null hypothesis is true, we incorrectly reject the

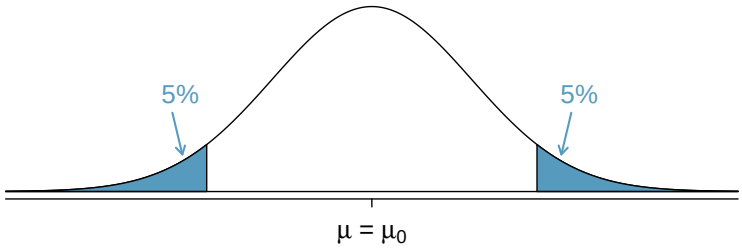


Figure 4.20: The shaded regions represent areas where we would reject H_0 under the bad practices considered in Example 4.4.5 when $\alpha = 0.05$.

null hypothesis about 5% of the time when the sample mean is above the null value, as shown in Figure 4.20.

Suppose the sample mean was smaller than the null value. Then if we change to a one-sided test, we would use H_A: $\mu < \mu_0$. If $\bar{x}$ had a T-statistic smaller than -1.65, we would reject H_0. If the null hypothesis is true, then we would observe such a case about 5% of the time.

By examining these two scenarios, we can determine that we will make a Type 1 Error $5\% + 5\% = 10\%$ of the time if we are allowed to swap to the "best" one-sided test for the data. This is twice the error rate we prescribed with our significance level: $\alpha = 0.05$!

Caution: One-sided hypotheses are allowed only *before* seeing data

After observing data, it is tempting to turn a two-sided test into a one-sided test. Avoid this temptation. Remember, the direction of a one-sided test must be made a priori, not after peeking at the data since the results could be statistically significant with a one-sided test, but not significant with a two-sided test. Hypotheses must be set up *before* observing the data. If they are not, the test must be two-sided.

4.4.6 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 is often 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. smaller than 0.05). 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. Many would use $\alpha = 0.01$ in this situation.

If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false. We will discuss this particular case in greater detail in Section 4.7.

Significance levels should reflect consequences of errors
The significance level selected for a test should reflect the consequences associated with Type 1 and Type 2 Errors.

Example 4.37 A medical machine manufacturer is considering a higher quality but more expensive supplier for parts in making an MRI. They sample a number of parts from their current supplier and also parts from the new supplier. They decide that if the high quality parts will last more than 12% longer, it makes financial sense to switch to this more expensive supplier. Is there good reason to modify the significance level in such a hypothesis test?

The null hypothesis is that the more expensive parts last no more than 12% longer while the alternative is that they do last more than 12% longer. This decision is just one of the many regular factors that have a marginal impact on the MRI and the company financial health. 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 since the machine's accuracy won't be affected.

Example 4.38 Now consider that the same MRI manufacturer is considering a slightly more expensive supplier for parts related to safety not longevity. If the durability of the machine's 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 and equally accurate in detection. Because safety is involved, the MRI machine company should be eager to switch to the slightly more expensive manufacturer (reject H_0) even if the evidence of increased safety and effectiveness is only moderately strong. A slightly larger significance level, such as $\alpha = 0.10$, might be appropriate.

Guided Practice 4.39 A part inside of a machine is very expensive to replace. However, the machine usually functions properly even if this part is broken and still detects the most common injuries at the same level with a fixed part. 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.[52]

4.5 Examining the Central Limit Theorem Closer (Special Topic)

Looking back to 4.3.5, we discovered that the normal model for the sample mean tends to be very good when the sample consists of at least 30 independent observations and the

[52] 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 $\alpha = 0.01$.

population data are not strongly skewed. The Central Limit Theorem provides the theory that allows us to make this assumption.

Central Limit Theorem, informal definition

The distribution of $\bar{x}$ is approximately normal. The approximation can be poor if the sample size is small, but it improves with larger sample sizes.

The Central Limit Theorem states that when the sample size is small, the normal approximation may not be very good. However, as the sample size becomes large, the normal approximation improves. We will investigate three theoretical cases to see roughly when the approximation is reasonable.

We consider three data sets: one from a *uniform* distribution, one from an *exponential* distribution, and the other from a *log-normal* distribution. Recall the properties of these distributions from Chapter **??**. These distributions are shown in the top panels of Figure 4.21. The uniform distribution is symmetric, the exponential distribution may be considered as having moderate skew since its right tail is relatively short (few outliers), and the log-normal distribution is strongly skewed and will tend to produce more apparent outliers.

The left panel in the $n = 2$ row represents the sampling distribution of $\bar{x}$ if it is the sample mean of two observations from the uniform distribution shown. The dashed line represents the closest approximation of the normal distribution. Similarly, the center and right panels of the $n = 2$ row represent the respective distributions of $\bar{x}$ for data from exponential and log-normal distributions.

⊙ **Guided Practice 4.40** Examine the distributions in each row of Figure 4.21. What do you notice about the normal approximation for each sampling distribution as the sample size becomes larger?[53]

● **Example 4.41** Would the normal approximation be good in all applications where the sample size is at least 30?

Not necessarily. For example, the normal approximation for the log-normal example is questionable for a sample size of 30. Generally, the more skewed a population distribution or the more common the frequency of outliers, the larger the sample required to guarantee the distribution of the sample mean is nearly normal.

TIP: With larger n, the sampling distribution of $\bar{x}$ becomes more normal

As the sample size increases, the normal model for $\bar{x}$ becomes more reasonable. We can also relax our condition on skew when the sample size is very large.

We discussed in Section 4.2.3 that the sample standard deviation, s, could be used as a substitute of the population standard deviation, σ, when computing the standard error. This estimate tends to be reasonable when $n \geq 30$. We will encounter alternative distributions for smaller sample sizes in Chapters **??** and **??**.

● **Example 4.42** Figure 4.22 shows a histogram of 50 observations. These represent the number of patient visits in a hospital for 50 consecutive days relative to their

[53] The normal approximation becomes better as larger samples are used.

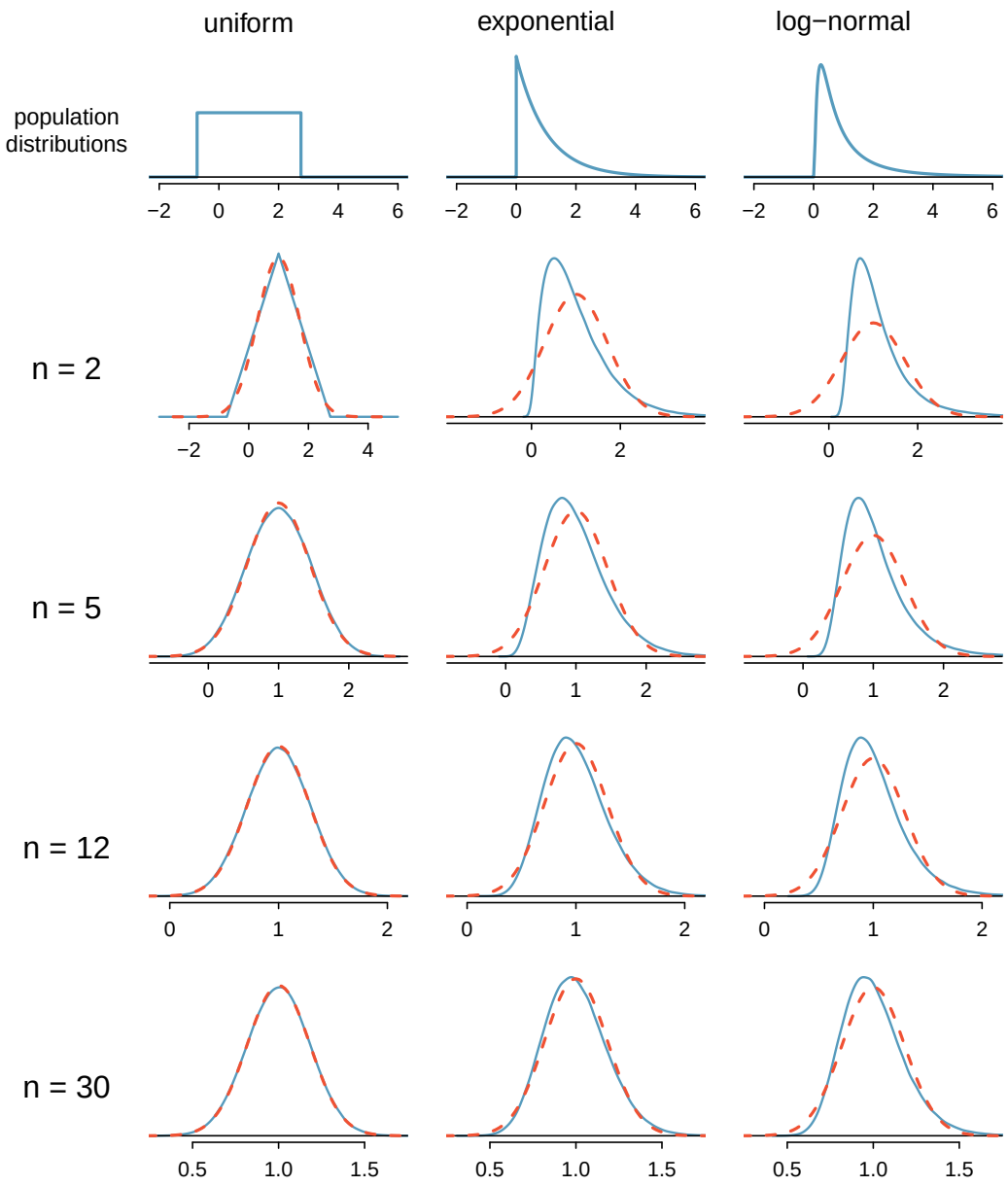


Figure 4.21: Sampling distributions for the mean at different sample sizes and for three different distributions. The dashed red lines show normal distributions.

average rate of 5000 patient visits. Can the normal approximation be applied to the sample mean, 90.69?

We should consider each of the required conditions.

(1) These are referred to as **time series data**, because the data arrived in a particular sequence. Time series data generally deals with, you guessed it, time! If there are a lot of patients in the hospital one day, it may influence how many patients there are the day after. During the flu season, patient visits might be at an all time high since many people are sick but also the time per visit is also extremely low. To make the assumption of independence we should perform careful checks on such data. While the supporting analysis is not shown, no evidence was found to indicate the observations are not independent on a whole.
(2) The sample size is 50, satisfying the sample size condition.
(3) There are two outliers, one very extreme, which suggests the data are very strongly skewed or very distant outliers may be common for this type of data. Outliers can play an important role and affect the distribution of the sample mean and the estimate of the standard error.

Since we should be skeptical of the independence of observations and the very extreme upper outlier poses a challenge, we should not use the normal model for the sample mean of these 50 observations. If we can obtain a much larger sample, perhaps several hundred observations over a longer period of time, then the concerns about skew and outliers would no longer apply.

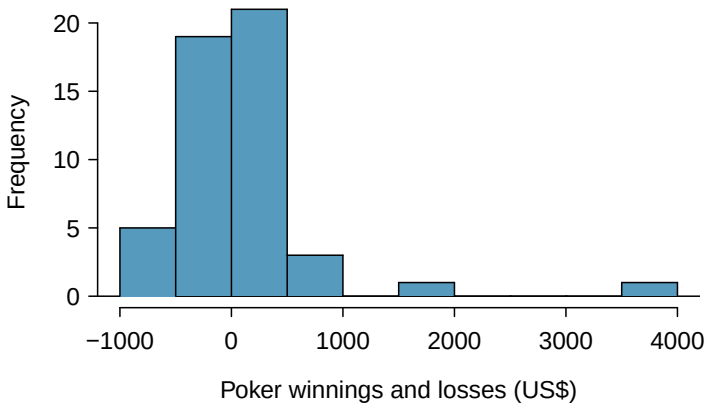


Figure 4.22: Sample distribution of total patient visits net of 5,000 visits. These data include some very clear outliers. These are problematic when considering the normality of the sample mean. For example, outliers are often an indicator of very strong skew.

Caution: Examine data structure when considering independence

Some data sets are collected in such a way that they have a natural underlying structure between observations, e.g. when observations occur consecutively. Be especially cautious about independence assumptions regarding such data sets.

Caution: Watch out for strong skew and outliers

Strong skew is often identified by the presence of clear outliers. If a data set has prominent outliers, or such observations are somewhat common for the type of data under study, then it is useful to collect a sample with many more than 30 observations if the normal model will be used for $\bar{x}$. There are no simple guidelines for what sample size is big enough for all situations, so proceed with caution when working in the presence of strong skew or more extreme outliers.

4.6 Inference for other estimators

The sample mean is not the only point estimate for which the sampling distribution is nearly normal. For example, the sampling distribution of sample proportions closely resembles the normal distribution when the sample size is sufficiently large. In this section, we introduce a number of examples where the normal approximation is reasonable for the point estimate. Chapters **??** and **??** will revisit each of the point estimates you see in this section along with some other new statistics.

We make another important assumption about each point estimate encountered in this section: the estimate is unbiased. A point estimate is **unbiased** if the sampling distribution of the estimate is centered at the parameter it estimates. A biased point estimate on the other hand can always be too high or estimates always too low. That is, an unbiased estimate does not naturally over or underestimate the parameter. Rather, it tends to provide a "good" estimate. The sample mean is an example of an unbiased point estimate, as are each of the examples we introduce in this section.

Finally, we will discuss the general case where a point estimate may follow some distribution other than the normal distribution. We also provide guidance about how to handle scenarios where the statistical techniques you are familiar with are insufficient for the problem at hand.

4.6.1 Confidence intervals for nearly normal point estimates

In Section 4.3, we used the point estimate $\bar{x}$ with a standard error $SE_{\bar{x}}$ to create a 95% confidence interval for the population mean:

$$\bar{x} \pm 1.96 \times SE_{\bar{x}} \tag{4.43}$$

We constructed this interval by noting that the sample mean is within 1.96 standard errors of the actual mean about 95% of the time. This same logic generalizes to any unbiased point estimate that is nearly normal. We may also generalize the confidence level by using a place-holder $z^\star$.

General confidence interval for the normal sampling distribution case

For any unbiased point estimate, the confidence interval for a nearly normal point estimate is

$$\text{point estimate} \pm z^\star SE \tag{4.44}$$

We see that it is of the same form as the generalized confidence interval for the sample mean where $z^\star$ is selected to correspond to the confidence level, and SE represents the standard error. Remember from previously that the value $z^\star SE$ is called the *margin of error*.

Generally the standard error for a point estimate is estimated from the data and computed using a formula. For example, the standard error for the sample mean is

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

In this section, we provide the computed standard error for each example and exercise without detailing where the values came from. In future chapters, you will learn to fill in these and other details for each situation.

Example 4.45 Using the `brfss.sample` data, we computed a point estimate for the average difference in weights between me and women: $\bar{x}_{\text{men}} - \bar{x}_{\text{women}} = 36.61162$ pounds. This point estimate is associated with a nearly normal distribution with SE = 0.35 pounds. What is a reasonable 95% confidence interval for the difference in gender weights?

The normal approximation is said to be valid, so we apply Equation (4.44):

$$\text{point estimate} \pm z^\star SE \quad \to \quad 36.61 \pm 1.96 \times 0.35 \quad \to \quad (35.91, 37.31)$$

Thus, we are 95% confident that the men were, on average, between 35.91 to 37.31 pounds heavier than women. That is, the actual average difference is plausibly between 35.91 and 37.31 pounds with 95% confidence.

Example 4.46 Does Example 4.45 guarantee that if a husband and wife both weighted themselves, the husband would weigh between 35.91 and 37.31 pounds more than the wife?

Our confidence interval says absolutely nothing about individual observations. It only makes a statement about a plausible range of values for the *average* difference between all men and women in the US.

Guided Practice 4.47 The proportion of men in the `brfss.sample` sample is $\hat{p} = 0.42$. This sample meets certain conditions that ensure $\hat{p}$ will be nearly normal, and the standard error of the estimate is $SE_{\hat{p}} = 0.05$. Create a 90% confidence interval for the proportion of participants in the BRFSS study and thus in the US who are men.[54]

[54] We use $z^\star = 1.65$, and apply the general confidence interval formula:

$$\hat{p} \pm z^\star SE_{\hat{p}} \quad \to \quad 0.42 \pm 1.65 \times 0.05 \quad \to \quad (0.3375, 0.5025)$$

Thus, we are 90% confident that between 34% and 50% are men.

4.6.2 Hypothesis testing for nearly normal point estimates

Just as the confidence interval method works with many other point estimates and we see the obvious connection between confidence intervals and hypothesis testing, it is unsurprising that we can generalize our hypothesis testing methods to new point estimates that are unbiased. Here we only consider the p-value approach, introduced in Section 4.4.2. Remember the Hypothesis testing framework from 4.4.1.

Hypothesis testing framework using the normal model

1. First write the hypotheses in plain language, then set them up in mathematical notation using the appropriate point estimate and parameter of interest.
2. State a significance level α. We generally use $\alpha = 0.05$.
3. Compute the test-statistic using the point estimate and standard error estimate.
4. Calculate the p-value by drawing a picture of the sampling distribution under H_0. Know which area you are shading to represent the correct p-value.
5. Use the p-value to evaluate your hypotheses. Write a conclusion within the context of the problem.

For point estimates other than the sampling mean which we know to be unbiased and nearly normal for $n > 30$, students need to verify conditions to ensure that the point estimate is nearly normal and unbiased so that the standard error estimate is also reasonable. This step can be done before computing the test-statistic.

⊙ **Guided Practice 4.48** A drug called sulphinpyrazone was under consideration for use in reducing the death rate in heart attack patients. To determine whether the drug was effective, a set of 1,475 patients were recruited into an experiment and randomly split into two groups: a control group that received a placebo and a treatment group that received the new drug. What would be an appropriate null hypothesis? And the alternative?[55]

We can formalize the hypotheses from Exercise 4.48 by letting $p_{control}$ and $p_{treatment}$ represent the proportion of patients who died in the control and treatment groups, respectively. Then the hypotheses can be written as

$$H_0 : p_{control} = p_{treatment} \quad \text{(the drug doesn't work)}$$
$$H_A : p_{control} > p_{treatment} \quad \text{(the drug works)}$$

or equivalently,

$$H_0 : p_{control} - p_{treatment} = 0 \quad \text{(the drug doesn't work)}$$
$$H_A : p_{control} - p_{treatment} > 0 \quad \text{(the drug works)}$$

Strong evidence against the null hypothesis and in favor of the alternative would correspond to an observed difference in death rates,

$$\text{point estimate} = \hat{p}_{control} - \hat{p}_{treatment}$$

[55] The skeptic's perspective is that the drug does not work at reducing deaths in heart attack patients (H_0), while the alternative is that the drug does work (H_A).

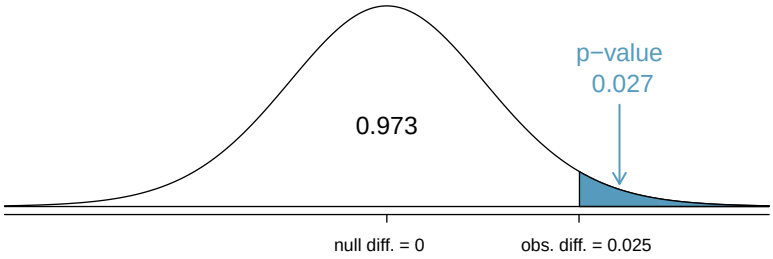


Figure 4.23: The distribution of the sample difference if the null hypothesis is true.

being larger than we would expect from chance alone. This difference in sample proportions represents a point estimate that is useful in evaluating the hypotheses.

Example 4.49 We want to evaluate the hypothesis setup from Exericse 4.48 using data from the actual study.[56] In the control group, 60 of 742 patients died. In the treatment group, 41 of 733 patients died. The sample difference in death rates can be summarized as

$$\text{point estimate} = \hat{p}_{control} - \hat{p}_{treatment} = \frac{60}{742} - \frac{41}{733} = 0.025$$

This point estimate is nearly normal and is an unbiased estimate of the actual difference in death rates. The standard error of this sample difference is $SE = 0.013$. Evaluate the hypothesis test at a 5% significance level: $\alpha = 0.05$.

We would like to identify the p-value to evaluate the hypotheses. If the null hypothesis is true, then the point estimate would have come from a nearly normal distribution, like the one shown in Figure 4.23. The distribution is centered at zero since $p_{control} - p_{treatment} = 0$ under the null hypothesis. Because a large positive difference provides evidence against the null hypothesis and in favor of the alternative, the upper tail has been shaded to represent the p-value. We need not shade the lower tail since this is a one-sided test: an observation in the lower tail does not support the alternative hypothesis.

The p-value can be computed by using the Z score of the point estimate and the normal probability table.

$$Z = \frac{\text{point estimate} - \text{null value}}{SE_{\text{point estimate}}} = \frac{0.025 - 0}{0.013} = 1.92 \tag{4.50}$$

Examining Z in the normal probability table, we find that the lower unshaded tail is about 0.973. Thus, the upper shaded tail representing the p-value is

$$\text{p-value} = 1 - 0.973 = 0.027$$

Because the p-value is less than the significance level ($\alpha = 0.05$), we say the null hypothesis is implausible. That is, we reject the null hypothesis in favor of the alternative and conclude that the drug is effective at reducing deaths in heart attack patients.

[56] 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.

4.6.3 Non-normal point estimates

We may apply the ideas of confidence intervals and hypothesis testing to cases where the point estimate or test statistic is not necessarily normal. There are many reasons why such a situation may arise:

- the sample size is too small for the normal approximation to be valid;
- the standard error estimate may be poor; or
- the point estimate tends towards some distribution that is not the normal distribution.

For each case where the normal approximation is not valid, our first task is always to understand and characterize the sampling distribution of the point estimate or test statistic. Next, we can apply the general frameworks for confidence intervals and hypothesis testing to these alternative distributions.

4.6.4 When to retreat

Statistical tools rely on conditions. When the conditions are not met, these tools are unreliable and drawing conclusions from them is treacherous. The conditions for these tools typically come in two forms.

- **The individual observations must be independent.** A random sample from less than 10% of the population ensures the observations are independent. In experiments, we generally require that subjects are randomized into groups. If independence fails, then advanced techniques must be used, and in some such cases, inference may not be possible.
- **Other conditions focus on sample size and skew.** For example, if the sample size is too small, the skew too strong, or extreme outliers are present, then the normal model for the sample mean will fail.

Verification of conditions for statistical tools is always necessary. Whenever conditions are not satisfied for a statistical technique, there are three options. The first is to learn new methods that are appropriate for the data. The second route is to consult a statistician.[57] The third route is to ignore the failure of conditions. This last option effectively invalidates any analysis and may discredit novel and interesting findings.

Finally, we caution that there may be no inference tools helpful when considering data that include unknown biases, such as convenience samples. For this reason, there are books, courses, and researchers devoted to the techniques of sampling and experimental design. See Sections **??**-**??** for basic principles of data collection.

4.7 Sample size and power (special topic)

The Type 2 Error rate and the magnitude of the error for a point estimate are controlled by the sample size [58]. Real differences from the null value, even large ones, may be difficult to detect with small samples. If we take a very large sample, we might find a statistically significant difference but the magnitude might be so small that it is of no practical value. In this section we describe techniques for selecting an appropriate sample size based on these considerations.

[57] If you work at a university, then there may be campus consulting services to assist you. Alternatively, there are many private consulting firms that are also available for hire.

[58] Remember the margin of error comes from the confidence interval (point estimate $\pm$ margin of error where the margin of error $= q^\star \cdot SE$ for a certain confidence level)

4.7.1 Finding a sample size for a certain margin of error

Many companies are concerned about rising healthcare costs. A company may estimate certain health characteristics of its employees, such as blood pressure, to project its future cost obligations. However, it might be too expensive to measure the blood pressure of every employee at a large company, and the company may choose to take a sample instead.

Example 4.51 Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg. How large of a sample is necessary to estimate the average systolic blood pressure with a margin of error of 4 mmHg using a 95% confidence level?

First, we frame the problem carefully. Recall that the margin of error is the part we add and subtract from the point estimate when computing a confidence interval. Here we assume that the company has more than 30 employees and thus we can use 1.96 as the critical value for this nearly normal point estimate [59] The margin of error for a 95% confidence interval estimating a mean can be written as

$$ME_{95\%} = 1.96 \times SE = 1.96 \times \frac{\sigma_{employee}}{\sqrt{n}}$$

The challenge in this case is to find the sample size n so that this margin of error is less than or equal to 4, which we write as an inequality:

$$1.96 \times \frac{\sigma_{employee}}{\sqrt{n}} \leq 4$$

In the above equation we wish to solve for the appropriate value of n, but we need a value for $\sigma_{employee}$ before we can proceed. However, we haven't yet collected any data, so we have no direct estimate! Instead, we use the best estimate available to us: the approximate standard deviation for the U.S. population, 25. To proceed and solve for n, we substitute 25 for $\sigma_{employee}$:

$$\begin{aligned} 1.96 \times \frac{\sigma_{employee}}{\sqrt{n}} \approx 1.96 \times \frac{25}{\sqrt{n}} &\leq 4 \\ 1.96 \times \frac{25}{4} &\leq \sqrt{n} \\ \left(1.96 \times \frac{25}{4}\right)^2 &\leq n \\ 150.06 &\leq n \end{aligned}$$

This suggests we should choose a sample size of at least 151 employees. We round up because the sample size must be *greater than or equal to 150.06* to ensure a margin of error of 4.

A potentially controversial part of Example 4.51 is the use of the U.S. standard deviation for the employee standard deviation. Usually the standard deviation for the sample is not known since we haven't taken the sample just yet! In such cases, many practicing

[59] Students should verify the other assumptions as well: independence etc.

statisticians review scientific literature or market research to make an educated guess about the standard deviation to calculate the standard error.

Identify a sample size for a particular margin of error

To estimate the necessary sample size for a maximum margin of error m, we set up an equation to represent this relationship:

$$m \geq ME = q^\star \frac{\sigma}{\sqrt{n}}$$

where $z^\star$ is chosen to correspond to the desired confidence level for a nearly normal point estimate, and σ is the standard deviation associated with the population. Solve for the sample size, n.
If we believed the point estimate not to be nearly normal, use $q^\star$ from the T-distribution instead. However in practice, a nearly normal point estimate is used more often than not.

Sample size computations are helpful in planning data collection, and they require careful forethought. Next we consider another topic important in planning data collection and setting a sample size: the Type 2 Error rate.

4.7.2 Power and the Type 2 Error rate

Consider the following two hypotheses:

H_0: The average blood pressure of employees is the same as the national average, $\mu = 130$.

H_A: The average blood pressure of employees is different than the national average, $\mu \neq 130$.

Suppose the alternative hypothesis is actually true. Then we might like to know, what is the chance we make a Type 2 Error? That is, what is the chance we will fail to reject the null hypothesis even though we should reject it? The answer is not obvious! If the average blood pressure of the employees is 132 (just 2 mmHg from the null value), it might be very difficult to detect the difference unless we use a large sample size. On the other hand, it would be easier to detect a difference if the real average of employees was 140.

Example 4.52 Suppose the actual employee average is 132 and we take a sample of 100 individuals. Then the true sampling distribution of $\bar{x}$ is approximately $N(132, 2.5)$ (since $SE = \frac{25}{\sqrt{100}} = 2.5$). What is the probability of successfully rejecting the null hypothesis?

This problem can be divided into two normal probability questions. First, we identify what values of $\bar{x}$ would represent sufficiently strong evidence to reject H_0. Second, we use the hypothetical sampling distribution with center $\mu = 132$ to find the probability of observing sample means in the areas we found in the first step.

Step 1. The null distribution could be represented by $N(130, 2.5)$, the same standard deviation as the true distribution but with the null value as its center. Then we can find the two tail areas by identifying the T-statistic corresponding to the 2.5% tails

(± 1.96), and solving for x in the T-statistic equation:

$$-1.96 = T_1 = \frac{x_1 - 130}{2.5} \qquad +1.96 = T_2 = \frac{x_2 - 130}{2.5}$$
$$x_1 = 125.1 \qquad x_2 = 134.9$$

(An equally valid approach is to recognize that x_1 is $1.96 \times SE$ below the mean and x_2 is $1.96 \times SE$ above the mean to compute the values.) Figure 4.24 shows the null distribution on the left with these two dotted cutoffs.

Step 2. Next, we compute the probability of rejecting H_0 if $\bar{x}$ actually came from $N(132, 2.5)$. This is the same as finding the two shaded tails for the second distribution in Figure 4.24. We again use the T-statistic method:

$$T_{left} = \frac{125.1 - 132}{2.5} = -2.76 \qquad T_{right} = \frac{134.9 - 132}{2.5} = 1.16$$
$$area_{left} = 0.003 \qquad area_{right} = 0.123$$

The probability of rejecting the null mean, if the true mean is 132, is the sum of these areas: $0.003 + 0.123 = 0.126$.

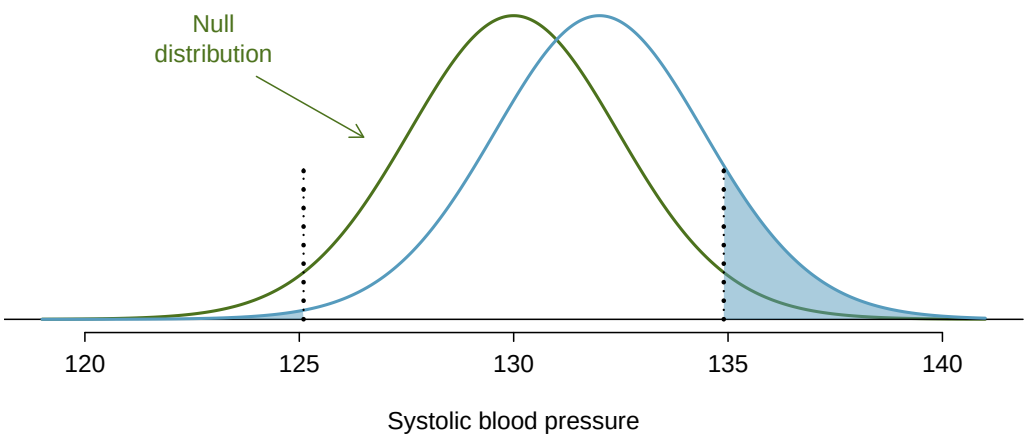


Figure 4.24: The sampling distribution of $\bar{x}$ under two scenarios. Left: $N(130, 2.5)$. Right: $N(132, 2.5)$, and the shaded areas in this distribution represent the power of the test.

The probability of rejecting the null hypothesis is called the **power**. The power varies depending on what we suppose the truth might be. In Example 4.52, the difference between the null value and the supposed true mean was relatively small, so the power was also small: only 0.126. However, when the truth is far from the null value, where we use the standard error as a measure of what is far, the power tends to increase.

⊙ **Guided Practice 4.53** Suppose the true sampling distribution of $\bar{x}$ is centered at 140. That is, $\bar{x}$ comes from $N(140, 2.5)$. What would the power be under this scenario? It may be helpful to draw $N(140, 2.5)$ and shade the area representing power on Figure 4.24; use the same cutoff values identified in Example 4.52.[60]

[60] Draw the distribution $N(140, 2.5)$, then find the area below 125.1 (about zero area) and above 134.9 (about 0.979). If the true mean is 140, the power is about 0.979.

⊙ **Guided Practice 4.54** If the power of a test is 0.979 for a particular mean, what is the Type 2 Error rate for this mean?[61]

⊙ **Guided Practice 4.55** Provide an intuitive explanation for why we are more likely to reject H_0 when the true mean is further from the null value.[62]

4.7.3 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. While we still say that difference is **statistically significant**, it might not be **practically significant**.

Statistically significant differences are sometimes so minor that they are not practically relevant. This is especially important to research: if we conduct a study, we want to focus on finding a meaningful result. We don't want to spend lots of money finding results that hold no practical and applicable value.

The role of a statistician in conducting a study often includes planning the size of the study and determining the value of α. Statisticians might first consult experts or scientific literature to learn what would be the smallest meaningful difference from the null value. They also would obtain some reasonable estimate for the standard deviation. With these important pieces of information, a sufficiently large sample size would be chosen so that the power for the meaningful difference is perhaps 80% or 90%. While larger sample sizes may still be used, statisticians in practice might advise against using them in some cases, especially in sensitive areas of research. While we note the statistical rigor in our hypothesis testing, we must also note that many of these tests must also stand up to practical significance in the real world.

---

[61] The Type 2 Error rate represents the probability of failing to reject the null hypothesis. Since the power is the probability we do reject, the Type 2 Error rate will be $1 - 0.979 = 0.021$.

[62] Answers may vary a little. When the truth is far from the null value, the point estimate also tends to be far from the null value, making it easier to detect the difference and reject H_0.

4.8 Exercises

4.8.1 Variability in estimates

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

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

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

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

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

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

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

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

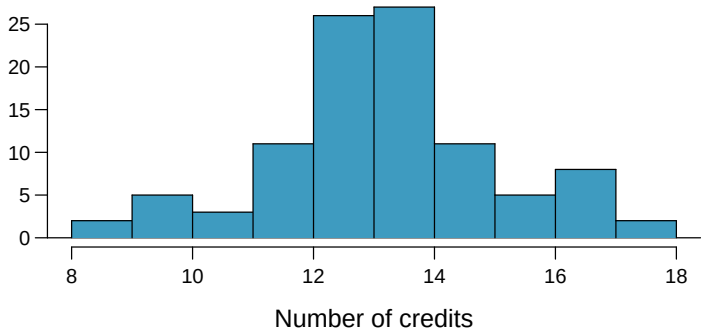
(b) A survey reports that local TV news has shown a 17% increase in revenue between 2009 and 2011 while newspaper revenues decreased by 6.4% during this time period.

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

(d) In a survey, internet users are asked whether or not they purchased any Groupon coupons.

(e) In a survey, internet users are asked how many Groupon coupons they purchased over the last year.

4.3 College credits. A college counselor is interested in estimating how many credits a student typically enrolls in each semester. The counselor decides to randomly sample 100 students by using the registrar's database of students. The histogram below shows the distribution of the number of credits taken by these students. Sample statistics for this distribution are also provided.



Min	8
Q1	13
Median	14
Mean	13.65
SD	1.91
Q3	15
Max	18

(a) What is the point estimate for the average number of credits taken per semester by students at this college? What about the median?

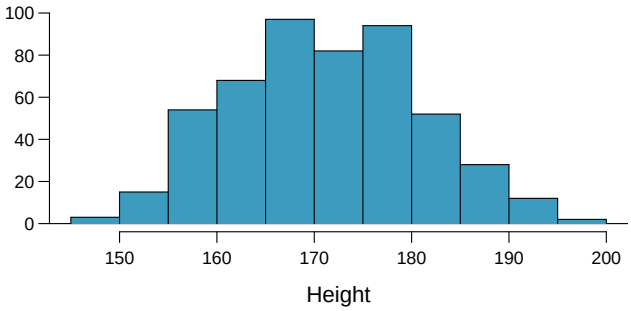
(b) What is the point estimate for the standard deviation of the number of credits taken per semester by students at this college? What about the IQR?

(c) Is a load of 16 credits unusually high for this college? What about 18 credits? Explain your reasoning. *Hint:* Observations farther than two standard deviations from the mean are usually considered to be unusual.

(d) The college counselor takes another random sample of 100 students and this time finds a sample mean of 14.02 units. Should she be surprised that this sample statistic is slightly different than the one from the original sample? Explain your reasoning.

(e) The sample means given above are point estimates for the mean number of credits taken by all students at that college. What measures do we use to quantify the variability of this estimate? Compute this quantity using the data from the original sample.

4.4 Heights of adults. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram below shows the sample distribution of heights in centimeters.[63]



Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
SD	9.4
Q3	177.8
Max	198.1

(a) What is the point estimate for the average height of active individuals? What about the median?

(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

(c) Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

(d) The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

(e) The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

4.5 Wireless routers. John is shopping for wireless routers and is overwhelmed by the number of available options. In order to get a feel for the average price, he takes a random sample of 75 routers and finds that the average price for this sample is $75 and the standard deviation is $25.

(a) Based on this information, how much variability should he expect to see in the mean prices of repeated samples, each containing 75 randomly selected wireless routers?

(b) A consumer website claims that the average price of routers is $80. Is a true average of $80 consistent with John's sample?

4.6 Chocolate chip cookies. Students are asked to count the number of chocolate chips in 22 cookies for a class activity. They found that the cookies on average had 14.77 chocolate chips with a standard deviation of 4.37 chocolate chips.

(a) Based on this information, about how much variability should they expect to see in the mean number of chocolate chips in random samples of 22 chocolate chip cookies?

[63] **Heinz:2003**.

(b) The packaging for these cookies claims that there are at least 20 chocolate chips per cookie. One student thinks this number is unreasonably high since the average they found is much lower. Another student claims the difference might be due to chance. What do you think?

4.8.2 Confidence intervals

4.7 Relaxing after work. The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. In 2010, the survey collected responses from 1,154 US residents. The survey is conducted face-to-face with an in-person interview of a randomly-selected sample of adults. One of the questions on the survey is "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" A 95% confidence interval from the 2010 GSS survey is 3.53 to 3.83 hours.[64]

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

(b) What does a 95% confidence level mean in this context?

(c) Suppose the researchers think a 90% confidence level would be more appropriate for this interval. Will this new interval be smaller or larger than the 95% confidence interval? Assume the standard deviation has remained constant since 2010.

4.8 Mental health. Another question on the General Social Survey introduced in Exercise 4.7 is "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.

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

(b) What does a 95% confidence level mean in this context?

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

(d) If a new survey asking the same questions was to be done with 500 Americans, would the standard error of the estimate be larger, smaller, or about the same. Assume the standard deviation has remained constant since 2010.

4.9 Width of a confidence interval. Earlier in Chapter 4, we calculated the 99% confidence interval for the average age of runners in the 2012 Cherry Blossom Run as (32.7, 37.4) based on a sample of 100 runners. How could we decrease the width of this interval without losing confidence?

4.10 Confidence levels. If a higher confidence level means that we are more confident about the number we are reporting, why don't we always report a confidence interval with the highest possible confidence level?

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

(a) This confidence interval is not valid since we do not know if the population distribution of the ER wait times is nearly normal.

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

[64] **data:gss:2010**.

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

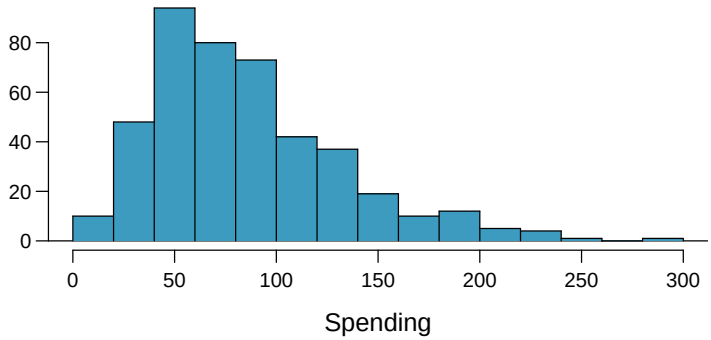
(d) 95% of such random samples would have a sample mean between 128 and 147 minutes.

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

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

(g) In order to decrease the margin of error of a 95% confidence interval to half of what it is now, we would need to double the sample size.

4.12 Thanksgiving spending, Part I. The 2009 holiday retail season, which kicked off on November 27, 2009 (the day after Thanksgiving), had been marked by somewhat lower self-reported consumer spending than was seen during the comparable period in 2008. To get an estimate of consumer spending, 436 randomly sampled American adults were surveyed. Daily consumer spending for the six-day period after Thanksgiving, spanning the Black Friday weekend and Cyber Monday, averaged $84.71. A 95% confidence interval based on this sample is ($80.31, $89.11). Determine whether the following statements are true or false, and explain your reasoning.



(a) We are 95% confident that the average spending of these 436 American adults is between $80.31 and $89.11.

(b) This confidence interval is not valid since the distribution of spending in the sample is right skewed.

(c) 95% of such random samples would have a sample mean between $80.31 and $89.11.

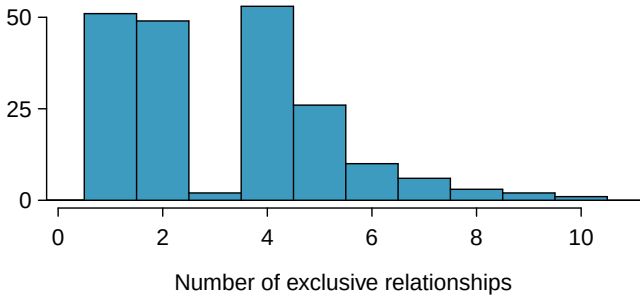
(d) We are 95% confident that the average spending of all American adults is between $80.31 and $89.11.

(e) A 90% confidence interval would be narrower than the 95% confidence interval.

(f) In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger.

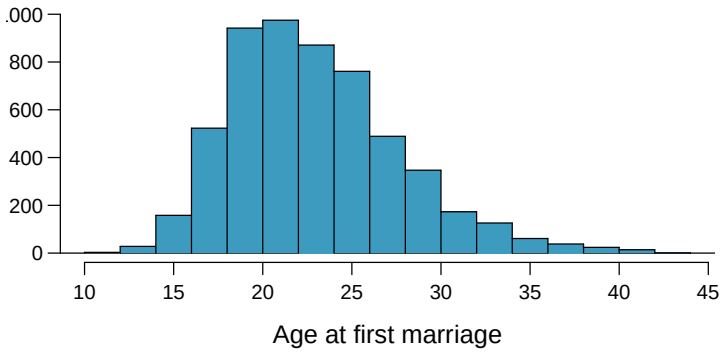
(g) The margin of error for the reported interval is 4.4.

4.13 Exclusive relationships. A survey was conducted on 203 undergraduates from Duke University who took an introductory statistics course in Spring 2012. Among many other questions, this survey asked them about the number of exclusive relationships they have been in. The histogram below shows the distribution of the data from this sample. The sample average is 3.2 with a standard deviation of 1.97.



Estimate the average number of exclusive relationships Duke students have been in using a 90% confidence interval and interpret this interval in context. Check any conditions required for inference, and note any assumptions you must make as you proceed with your calculations and conclusions.

4.14 Age at first marriage, Part I. The National Survey of Family Growth conducted by the Centers for Disease Control gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. One of the variables collected on this survey is the age at first marriage. The histogram below shows the distribution of ages at first marriage of 5,534 randomly sampled women between 2006 and 2010. The average age at first marriage among these women is 23.44 with a standard deviation of 4.72.[65]



Estimate the average age at first marriage of women using a 95% confidence interval, and interpret this interval in context. Discuss any relevant assumptions.

4.8.3 Hypothesis testing

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

(a) New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Do these data provide convincing evidence that New Yorkers on average sleep less than 8 hours a night?

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

[65] **data:nsfg:2010**.

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

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

(b) Based on the performance of those who took the GRE exam between July 1, 2004 and June 30, 2007, the average Verbal Reasoning score was calculated to be 462. In 2011 the average verbal score was slightly higher. Do these data provide convincing evidence that the average GRE Verbal Reasoning score has changed since 2004?[66]

4.17 Online communication. A study suggests that the average college student spends 2 hours per week communicating with others online. You believe that this is an underestimate and decide to collect your own sample for a hypothesis test. You randomly sample 60 students from your dorm and find that on average they spent 3.5 hours a week communicating with others online. A friend of yours, who offers to help you with the hypothesis test, comes up with the following set of hypotheses. Indicate any errors you see.

$$H_0 : \bar{x} < 2 \; hours$$
$$H_A : \bar{x} > 3.5 \; hours$$

4.18 Age at first marriage, Part II. Exercise 4.14 presents the results of a 2006 - 2010 survey showing that the average age of women at first marriage is 23.44. Suppose a researcher believes that this value has increased in 2012, but he would also be interested if he found a decrease. Below is how he set up his hypotheses. Indicate any errors you see.

$$H_0 : \bar{x} = 23.44 \; years \; old$$
$$H_A : \bar{x} > 23.44 \; years \; old$$

4.19 Waiting at an ER, Part II. Exercise 4.11 provides a 95% confidence interval for the mean waiting time at an emergency room (ER) of (128 minutes, 147 minutes).

(a) A local newspaper claims that the average waiting time at this ER exceeds 3 hours. What do you think of this claim?

(b) The Dean of Medicine at this hospital claims the average wait time is 2.2 hours. What do you think of this claim?

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

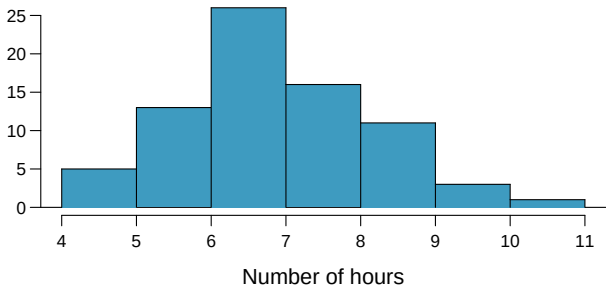
4.20 Thanksgiving spending, Part II. Exercise 4.12 provides a 95% confidence interval for the average spending by American adults during the six-day period after Thanksgiving 2009: ($80.31, $89.11).

(a) A local news anchor claims that the average spending during this period in 2009 was $100. What do you think of this claim?

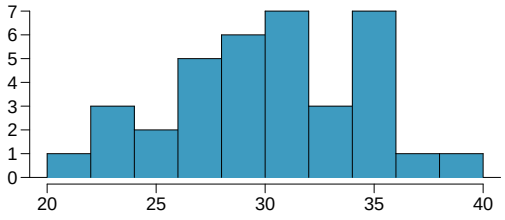
(b) Would the news anchor's claim be considered reasonable based on a 90% confidence interval? Why or why not?

4.21 Ball bearings. A manufacturer claims that bearings produced by their machine last 7 hours on average under harsh conditions. A factory worker randomly samples 75 ball bearings, and records their lifespans under harsh conditions. He calculates a sample mean of 6.85 hours, and the standard deviation of the data is 1.25 working hours. The following histogram shows the distribution of the lifespans of the ball bearings in this sample. Conduct a formal hypothesis test of this claim. Make sure to check that relevant conditions are satisfied.

[66] **webpage:GRE**.



4.22 Gifted children, Part I. Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. The following histogram shows the distribution of the ages (in months) at which these children first counted to 10 successfully. Also provided are some sample statistics.[67]



n	36
min	21
mean	30.69
sd	4.31
max	39

(a) Are conditions for inference satisfied?

(b) Suppose you read on a parenting website that children first count to 10 successfully when they are 32 months old, on average. Perform a hypothesis test to evaluate if these data provide convincing evidence that the average age at which gifted children first count to 10 successfully is different than the general average of 32 months. Use a significance level of 0.10.

(c) Interpret the p-value in context of the hypothesis test and the data.

(d) Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.

(e) Do your results from the hypothesis test and the confidence interval agree? Explain.

4.23 Waiting at an ER, Part III. The hospital administrator mentioned in Exercise 4.11 randomly selected 64 patients and measured the time (in minutes) between when they checked in to the ER and the time they were first seen by a doctor. The average time is 137.5 minutes and the standard deviation is 39 minutes. He is getting grief from his supervisor on the basis that the wait times in the ER increased greatly from last year's average of 127 minutes. However, the administrator claims that the increase is probably just due to chance.

(a) Are conditions for inference met? Note any assumptions you must make to proceed.

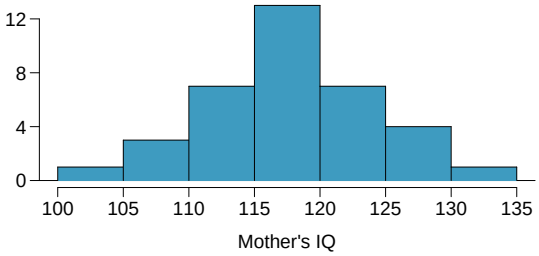
(b) Using a significance level of $\alpha = 0.05$, is the change in wait times statistically significant? Use a two-sided test since it seems the supervisor had to inspect the data before he suggested an increase occurred.

(c) Would the conclusion of the hypothesis test change if the significance level was changed to $\alpha = 0.01$?

[67] **Graybill:1994**.

4.24 Gifted children, Part II. Exercise 4.22 describes a study on gifted children. In this study, along with variables on the children, the researchers also collected data on the mother's and father's IQ of the 36 randomly sampled gifted children. The histogram below shows the distribution of mother's IQ. Also provided are some sample statistics.

n	36
min	101
mean	118.2
sd	6.5
max	131



(a) Perform a hypothesis test to evaluate if these data provide convincing evidence that the average IQ of mothers of gifted children is different than the average IQ for the population at large, which is 100. Use a significance level of 0.10.

(b) Calculate a 90% confidence interval for the average IQ of mothers of gifted children.

(c) Do your results from the hypothesis test and the confidence interval agree? Explain.

4.25 Nutrition labels. The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a sample mean of 134 calories with a standard deviation of 17 calories. Is there evidence that the nutrition label does not provide an accurate measure of calories in the bags of potato chips? We have verified the independence, sample size, and skew conditions are satisfied.

4.26 Find the sample mean. You are given the following hypotheses: H_0: $\mu = 34$, H_A: $\mu > 34$. We know that the sample standard deviation is 10 and the sample size is 65. For what sample mean would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

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

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

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

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

(d) How would these errors affect the patient?

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

(a) Write the hypotheses in words.

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

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

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

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

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

4.29 Errors in drug testing. Suppose regulators monitored 403 drugs last year, each for a particular adverse response. For each drug they conducted a single hypothesis test with a significance level of 5% to determine if the adverse effect was higher in those taking the drug than those who did not take the drug; the regulators ultimately rejected the null hypothesis for 42 drugs.

(a) Describe the error the regulators might have made for a drug where the null hypothesis was rejected.

(b) Describe the error regulators might have made for a drug where the null hypothesis was not rejected.

(c) Suppose the vast majority of the 403 drugs do not have adverse effects. Then, if you picked one of the 42 suspect drugs at random, about how sure would you be that the drug really has an adverse effect?

(d) Can you also say how sure you are that a particular drug from the 361 where the null hypothesis was not rejected does not have the corresponding adverse response?

4.30 Car insurance savings, Part I. A car insurance company advertises that customers switching to their insurance save, on average, $432 on their yearly premiums. A market researcher at a competing insurance discounter is interested in showing that this value is an overestimate so he can provide evidence to government regulators that the company is falsely advertising their prices. He randomly samples 82 customers who recently switched to this insurance and finds an average savings of $395, with a standard deviation of $102.

(a) Are conditions for inference satisfied?

(b) Perform a hypothesis test and state your conclusion.

(c) Do you agree with the market researcher that the amount of savings advertised is an overestimate? Explain your reasoning.

(d) Calculate a 90% confidence interval for the average amount of savings of all customers who switch their insurance.

(e) Do your results from the hypothesis test and the confidence interval agree? Explain.

4.31 Happy hour. A restaurant owner is considering extending the happy hour at his restaurant since he would like to see if it increases revenue. If it does, he will permanently extend happy hour. He estimates that the current average revenue per customer is $18 during happy hour. He runs the extended happy hour for a week and finds an average revenue of $19.25 with a standard deviation $3.02 based on a simple random sample of 70 customers.

(a) Are conditions for inference satisfied?

(b) Perform a hypothesis test. Suppose the customers and their buying habits this week were no different than in any other week for this particular bar. (This may not always be a reasonable assumption.)

(c) Calculate a 90% confidence interval for the average revenue per customer.

(d) Do your results from the hypothesis test and the confidence interval agree? Explain.

(e) If your hypothesis test and confidence interval suggest a significant increase in revenue per customer, why might you still not recommend that the restaurant owner extend the happy hour based on this criterion? What may be a better measure to consider?

4.32 Speed reading, Part I. A company offering online speed reading courses claims that students who take their courses show a 5 times (500%) increase in the number of words they can read in a minute without losing comprehension. A random sample of 100 students yielded an average increase of 415% with a standard deviation of 220%. Is there evidence that the company's claim is false?

(a) Are conditions for inference satisfied?

(b) Perform a hypothesis test evaluating if the company's claim is reasonable or if the true average improvement is less than 500%. Make sure to interpret your response in context of the hypothesis test and the data. Use $\alpha = 0.025$.

(c) Calculate a 95% confidence interval for the average increase in the number of words students can read in a minute without losing comprehension.

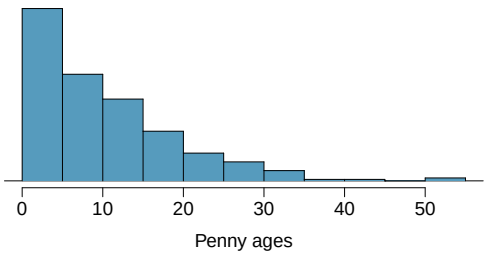
(d) Do your results from the hypothesis test and the confidence interval agree? Explain.

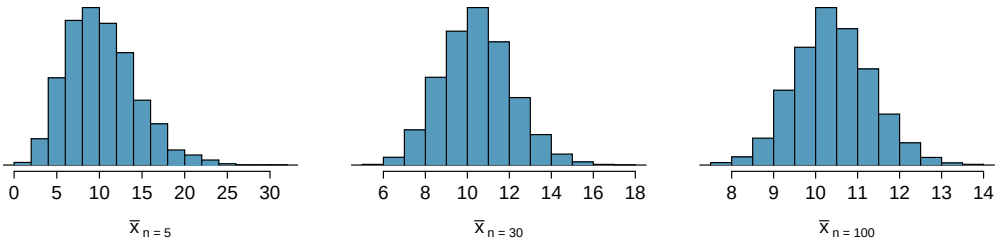
4.8.4 Examining the Central Limit Theorem

4.33 Ages of pennies, Part I. The histogram below shows the distribution of ages of pennies at a bank.

(a) Describe the distribution.

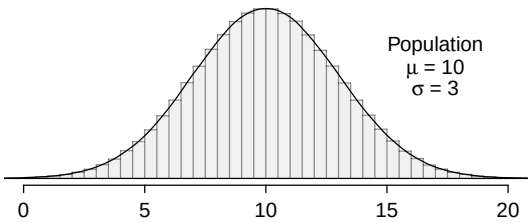
(b) Sampling distributions for means from simple random samples of 5, 30, and 100 pennies is shown in the histograms below. Describe the shapes of these distributions and comment on whether they look like what you would expect to see based on the Central Limit Theorem.

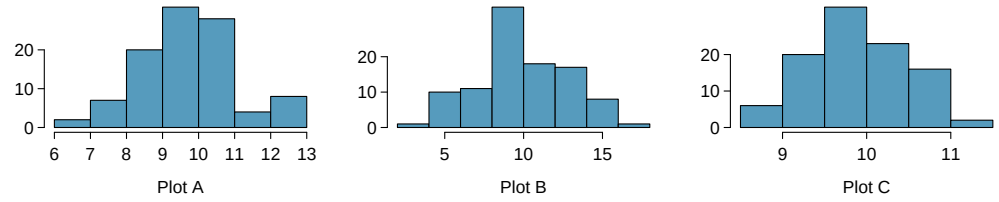




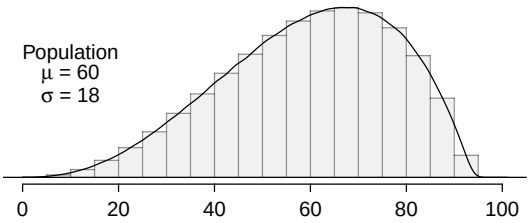
4.34 Ages of pennies, Part II. The mean age of the pennies from Exercise 4.33 is 10.44 years with a standard deviation of 9.2 years. Using the Central Limit Theorem, calculate the means and standard deviations of the distribution of the mean from random samples of size 5, 30, and 100. Comment on whether the sampling distributions shown in Exercise 4.33 agree with the values you compute.

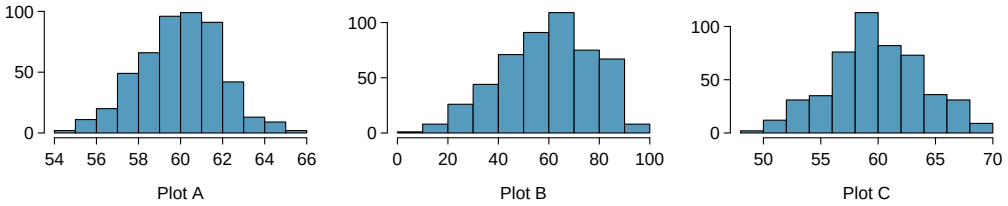
4.35 Identify distributions, Part I. Four plots are presented below. The plot at the top is a distribution for a population. The mean is 10 and the standard deviation is 3. Also shown below is a distribution of (1) a single random sample of 100 values from this population, (2) a distribution of 100 sample means from random samples with size 5, and (3) a distribution of 100 sample means from random samples with size 25. Determine which plot (A, B, or C) is which and explain your reasoning.





4.36 Identify distributions, Part II. Four plots are presented below. The plot at the top is a distribution for a population. The mean is 60 and the standard deviation is 18. Also shown below is a distribution of (1) a single random sample of 500 values from this population, (2) a distribution of 500 sample means from random samples of each size 18, and (3) a distribution of 500 sample means from random samples of each size 81. Determine which plot (A, B, or C) is which and explain your reasoning.





4.37 Housing prices, Part I. A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

(a) Is the distribution of housing prices in Topanga symmetric, right skewed, or left skewed? *Hint:* Sketch the distribution.

(b) Would you expect most houses in Topanga to cost more or less than \$1.3 million?

(c) Can we estimate the probability that a randomly chosen house in Topanga costs more than \$1.4 million using the normal distribution?

(d) What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

(e) How would doubling the sample size affect the standard error of the mean?

4.38 Stats final scores. Each year about 1500 students take the introductory statistics course at a large university. This year scores on the final exam are distributed with a median of 74 points, a mean of 70 points, and a standard deviation of 10 points. There are no students who scored above 100 (the maximum score attainable on the final) but a few students scored below 20 points.

(a) Is the distribution of scores on this final exam symmetric, right skewed, or left skewed?

(b) Would you expect most students to have scored above or below 70 points?

(c) Can we calculate the probability that a randomly chosen student scored above 75 using the normal distribution?

(d) What is the probability that the average score for a random sample of 40 students is above 75?

(e) How would cutting the sample size in half affect the standard error of the mean?

4.39 Weights of pennies. The distribution of weights of US pennies is approximately normal with a mean of 2.5 grams and a standard deviation of 0.03 grams.

(a) What is the probability that a randomly chosen penny weighs less than 2.4 grams?

(b) Describe the sampling distribution of the mean weight of 10 randomly chosen pennies.

(c) What is the probability that the mean weight of 10 pennies is less than 2.4 grams?

(d) Sketch the two distributions (population and sampling) on the same scale.

(e) Could you estimate the probabilities from (a) and (c) if the weights of pennies had a skewed distribution?

4.40 CFLs. A manufacturer of compact fluorescent light bulbs advertises that the distribution of the lifespans of these light bulbs is nearly normal with a mean of 9,000 hours and a standard deviation of 1,000 hours.

(a) What is the probability that a randomly chosen light bulb lasts more than 10,500 hours?

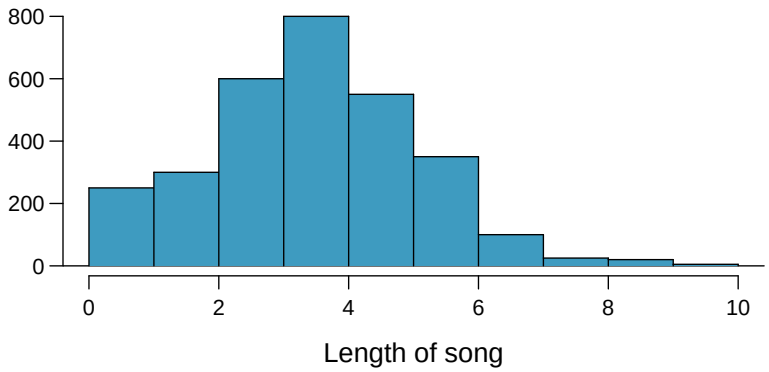
(b) Describe the distribution of the mean lifespan of 15 light bulbs.

(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?

(d) Sketch the two distributions (population and sampling) on the same scale.

(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution?

4.41 Songs on an iPod. Suppose an iPod has 3,000 songs. The histogram below shows the distribution of the lengths of these songs. We also know that, for this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes.



(a) Calculate the probability that a randomly selected song lasts more than 5 minutes.

(b) You are about to go for an hour run and you make a random playlist of 15 songs. What is the probability that your playlist lasts for the entire duration of your run? *Hint:* If you want the playlist to last 60 minutes, what should be the minimum average length of a song?

(c) You are about to take a trip to visit your parents and the drive is 6 hours. You make a random playlist of 100 songs. What is the probability that your playlist lasts the entire drive?

4.42 Spray paint. Suppose the area that can be painted using a single can of spray paint is slightly variable and follows a nearly normal distribution with a mean of 25 square feet and a standard deviation of 3 square feet.

(a) What is the probability that the area covered by a can of spray paint is more than 27 square feet?

(b) Suppose you want to spray paint an area of 540 square feet using 20 cans of spray paint. On average, how many square feet must each can be able to cover to spray paint all 540 square feet?

(c) What is the probability that you can cover a 540 square feet area using 20 cans of spray paint?

(d) If the area covered by a can of spray paint had a slightly skewed distribution, could you still calculate the probabilities in parts (a) and (c) using the normal distribution?

4.8.5 Inference for other estimators

4.43 Spam mail, Part I. The 2004 National Technology Readiness Survey sponsored by the Smith School of Business at the University of Maryland surveyed 418 randomly sampled Americans, asking them how many spam emails they receive per day. The survey was repeated on a new random sample of 499 Americans in 2009.[68]

(a) What are the hypotheses for evaluating if the average spam emails per day has changed from 2004 to 2009.

(b) In 2004 the mean was 18.5 spam emails per day, and in 2009 this value was 14.9 emails per day. What is the point estimate for the difference between the two population means?

(c) A report on the survey states that the observed difference between the sample means is not statistically significant. Explain what this means in context of the hypothesis test and the data.

(d) Would you expect a confidence interval for the difference between the two population means to contain 0? Explain your reasoning.

4.44 Nearsightedness. It is believed that nearsightedness affects about 8% of all children. In a random sample of 194 children, 21 are nearsighted.

(a) Construct hypotheses appropriate for the following question: do these data provide evidence that the 8% value is inaccurate?

(b) What proportion of children in this sample are nearsighted?

(c) Given that the standard error of the sample proportion is 0.0195 and the point estimate follows a nearly normal distribution, calculate the test statistic (the Z statistic).

(d) What is the p-value for this hypothesis test?

(e) What is the conclusion of the hypothesis test?

4.45 Spam mail, Part II. The National Technology Readiness Survey from Exercise 4.43 also asked Americans how often they delete spam emails. 23% of the respondents in 2004 said they delete their spam mail once a month or less, and in 2009 this value was 16%.

(a) What are the hypotheses for evaluating if the proportion of those who delete their email once a month or less (or never) has changed from 2004 to 2009?

(b) What is the point estimate for the difference between the two population proportions?

(c) A report on the survey states that the observed decrease from 2004 to 2009 is statistically significant. Explain what this means in context of the hypothesis test and the data.

(d) Would you expect a confidence interval for the difference between the two population proportions to contain 0? Explain your reasoning.

4.46 Unemployment and relationship problems. A USA Today/Gallup poll conducted between 2010 and 2011 asked a group of unemployed and underemployed Americans if they have had major problems in their relationships with their spouse or another close family member as a result of not having a job (if unemployed) or not having a full-time job (if underemployed). 27%

[68] **webpage:spam**.

of the 1,145 unemployed respondents and 25% of the 675 underemployed respondents said they had major problems in relationships as a result of their employment status.

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

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

4.8.6 Sample size and power

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

(a) The standard error of $\bar{x}$ when $s = 120$ and (I) n = 25 or (II) n = 125.

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

(c) The p-value for a Z statistic of 2.5 when (I) n = 500 or (II) n = 1000.

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

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

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

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

(c) Suppose the null hypothesis is $\mu = 5$ and we fail to reject H_0. Under this scenario, the true population mean is 5.

(d) If the alternative hypothesis is true, then the probability of making a Type 2 error and the power of a test add up to 1.

(e) With large sample sizes, even small differences between the null value and the true value of the parameter, a difference often called the effect size, will be identified as statistically significant.

(f) A cutoff of $\alpha = 0.05$ is the ideal value for all hypothesis tests.

4.49 Car insurance savings, Part II. The market researcher from Exercise 4.30 collected data about the savings of 82 customers at a competing car insurance company. The mean and standard deviation of this sample are \$395 and \$102, respectively. He would like to conduct another survey but have a margin of error of no more than \$10 at a 99% confidence level. How large of a sample should he collect?

4.50 Speed reading, Part II. A random sample of 100 students who took online speed reading courses from the company described in Exercise 4.32 yielded an average increase in reading speed of 415% and a standard deviation of 220%. We would like to calculate a 95% confidence interval for the average increase in reading speed with a margin of error of no more than 15%. How many students should we sample?

4.51 Waiting at the ER, Part IV. Exercise 4.23 introduced us to a hospital where ER wait times were being analyzed. The previous year's average was 128 minutes. Suppose that this year's average wait time is 135 minutes.

(a) Provide the hypotheses for this situation in plain language.

(b) If we plan to collect a sample size of $n = 64$, what values could $\bar{x}$ take so that we reject H_0? Suppose the sample standard deviation from the earlier exercise (39 minutes) is the population standard deviation. You may assume that the conditions for the nearly normal model for $\bar{x}$ are satisfied.

(c) Calculate the probability of a Type 2 error.