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

1 **Introduction to data** **8**

1.1 Case study 9
1.2 Data basics 12
1.3 Data collection principles 17
1.4 Examining numerical data 28
1.5 Considering categorical data 41
1.6 Exercises 48

2 **Probability** **58**

2.1 Defining probability 58
2.2 Conditional probability 75
2.3 Exercises 83

4 **Foundations for inference** **94**

4.1 BRFSS Data 95
4.2 Variability in estimates 97
4.3 Confidence intervals 105
4.4 Hypothesis testing 118
4.5 Examining the Central Limit Theorem Closer (Special Topic) 133
4.6 Inference for other estimators 136
4.7 Sample size and power (special topic) 141
4.8 Exercises 146

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 important 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 1.

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 1

Introduction to data

Scientists seek to answer questions using rigorous methods and careful observations. These observations – collected from the likes of field notes, surveys, and experiments – form the backbone of a statistical investigation and are called **data**. Statistics is the study of how to best collect, analyze, and draw conclusions from data. It is helpful to place statistics in the context of a general process of investigation:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Statistics as a subject focuses on making stages 2-4 objective, rigorous, and efficient. That is, statistics has three primary components: How can data best be collected? How should it be analyzed? What can be inferred from the analysis?

This chapter provides a brief introduction to the basic principles of data collection techniques and analytic tools, and illustrates the important role statistics plays in medicine and biology.

JV: Make reference to how this chapter can be used by either someone new to statistics or someone who has had the material before?

JV: Something to address later – where in the formatting does it specify that the first paragraph in a section does not get indented, but the others do...

1.1 Case study: preventing peanut allergies

Section 1.1 introduces an important problem in medicine: evaluating the effect of an intervention. *JV: Intervention should be defined, could work in a dependent clause.* Terms in this section, and indeed much of this chapter, will all be revisited later in more detail.

The proportion of young children in Western countries with peanut allergies has doubled in the last 10 years. Previous research suggests that exposing infants to peanut-based foods, rather than excluding such foods from their diets, may be an effective strategy for preventing the development of peanut allergies. This section describes an experiment (a clinical trial, in the terminology of medical research) designed to address the following research question: Does early exposure to peanut products reduce the probability that a child will develop peanut allergies?

The "Learning Early about Peanut Allergy" (LEAP) study was reported in the New England Journal of Medicine in 2015.[1] The study team enrolled children in the United Kingdom between 2006 and 2009, selecting 640 infants with eczema, egg allergy, or both. Each child was randomly assigned to the treatment group (peanut consumption) or the control group (peanut avoidance); children in the treatment group were fed at least 6 grams of peanut protein until 5 years of age, while children in the control group were to avoid consuming peanut protein until 5 years of age. *In this study, the control group provides a reference point for estimating the effect of peanut exposure in the treatment group. JV: This last sentence comes off as a bit technical/vague, perhaps due to "estimating the effect" – could be omitted, I think.*

At age 5, each child was tested for peanut allergy using an oral food challenge (OFC): 5 grams of peanut protein in a single dose. Children had been previously been tested for peanut allergy through a skin test, conducted at the time of study entry; the main analysis presented in the paper was based on the 530 children with a negative skin test

[1] Du Toit, George, et al. Randomized trial of peanut consumption in infants at risk for peanut allergy. New England Journal of Medicine 372.9 (2015): 803-813.

result. Of these children, 263 were assigned to "Peanut Avoidance" and 267 to "Peanut Consumption." The outcome at 5 years was reported as either "Fail OFC" (allergic reaction) or "Pass OFC" (no allergic reaction).

Table 1.1 shows the participant study ID number, treatment assignment, and OFC outcome for 5 children. All five of these children passed the food challenge.

JV: The last two row labels should be corrected to 529 and 530; not sure how to fix that in the R code. We should be careful later on to clarify that in the file, the row.names column corresponds to the dataset with all 640 children. Alternatively, that column could be omitted and a separate dataset made with the children with a positive skin test, which could be analyzed in an end-of-chapter exercise.

JV: Just curious...so 267 for the peanut consumption group includes the one child with a positive skin test?

	participant.ID	treatment.group	overall.V60.outcome
1	LEAP_100522	Peanut Consumption	PASS OFC
2	LEAP_103358	Peanut Consumption	PASS OFC
3	LEAP_105069	Peanut Avoidance	PASS OFC
	⋮	⋮	⋮
639	LEAP_994047	Peanut Avoidance	PASS OFC
640	LEAP_997608	Peanut Consumption	PASS OFC

Table 1.1: Results for five children from the peanut study.

Summary tables are generally more helpful than individual participant listings when looking for patterns in data. Table 1.2 shows outcomes grouped by treatment group and the result of the OFC test. From this table, it is possible to compute some simple summary statistics.

	FAIL OFC	PASS OFC	Sum
Peanut Avoidance	36	227	263
Peanut Consumption	5	262	267
Sum	41	489	530

Table 1.2: LEAP Study Results

A **summary statistic** is a single number summarizing a large amount of data.[2] In

[2] Formally, a summary statistic is a value computed from the data. Some summary statistics are more useful than others.

the Peanut Avoidance group, the proportion of participants failing the food challenge at 5 years of age is 36/263 = 0.137 (13.7%); in the Peanut Consumption intervention, the proportion failing is 5/267 = 0.019 (1.9%). The difference between these two proportions, 11.8%, is a single summary statistic describing the extent to which these two proportions differ. A second summary statistic, the ratio of the two proportions, 0.137/0.019 = 7.31, indicates that the proportion failing in the Avoidance group is more than 7 times that of the Consumption group. This ratio is called a **relative risk**.

JV: Interpretation of RR should be more clearly stated here, or else RR should be omitted entirely.

The summary statistics for the LEAP study highlight an important point – the results of a study can sometimes be surprising. A parent of a child already known to be allergic to eggs might be justifiably skeptical about feeding peanut butter to their child. The LEAP study suggests that, at least for children similar to those enrolled in the study, the benefits of early exposure might be substantial.

There are important aspects of the study to be cautious about. This study was conducted in the United Kingdom at a single site of pediatric care; it is not at all clear that results in children from that site can be generalized to other countries or cultures. Furthermore, the results also raise an important statistical issue: does the study provide definitive evidence that peanut consumption is beneficial? In other words, is the 11.8% difference between the two groups larger than one would expect by chance variation alone?

Suppose a coin is flipped 100 times. While the chance a coin lands heads in any given coin flip is 50%, observing exactly 50 heads is unlikely; instead, the coin may land heads 43 times, 51 times, 59 times, etc. This type of fluctuation is part of almost any experiment or study. It may well be possible that the 11.8% difference in the peanut allergy study is only due to this natural variation, and that the two interventions are actually equally effective. However, the larger the difference observed (for a particular study size), the less credible it is that the difference is due to chance alone. If out of 100 flips, a coin landed heads only 5 times, it would be reasonable to doubt that the outcome was due to chance; perhaps the coin is weighted so that tails are more likely to occur.

For the LEAP study, the 11.8% difference is indeed larger than that expected by chance alone, suggesting that peanut consumption is the more effective intervention for preventing subsequent allergies. The material on hypothesis testing in later chapters will provide the statistical tools to examine this issue.

1.2 Data basics

Effective presentation and description of data is a first step in most analyses. This section introduces one structure for organizing data as well as some terminology that will be used throughout this book.

1.2.1 Observations, variables, and data matrices

This section describes data used in a study published in the *Journal of Evolutionary Biology* about maternal investment at differing altitudes, conducted in a frog species endemic to the Tibetan Plateau (*Rana kukunoris*).[3] Reproduction is a costly process for females, necessitating a trade-off between individual egg size and total number of eggs produced. Researchers collected measurements on egg clutches found at breeding ponds across 11 study sites; for 5 sites, they also collected data on individual female frogs.

	altitude	latitude	egg.size	clutch.size	clutch.volume	body.size
1	3,462.00	34.82	1.95	181.97	177.83	3.63
2	3,462.00	34.82	1.95	269.15	257.04	3.63
3	3,462.00	34.82	1.95	158.49	151.36	3.72
150	2,597.00	34.05	2.24	537.03	776.25	NA

Table 1.3: Frog Study Data Matrix

Table 1.3 displays rows 1, 2, 3, and 150 of the data from the 431 clutches. The complete set of observations will be referred to as the `frog` dataset. Each row in the table corresponds to a single clutch, indicating where the clutch was collected (`altitude` and `latitude`), `egg.size`, `clutch.size`, `clutch.volume`, and `body.size` of the mother when available. "NA" corresponds to a missing value; information on individual females was not collected for that particular site. The columns represent characteristics, called **variables**,

[3] Chen, W., et al. Maternal investment increases with altitude in a frog on the Tibetan Plateau. Journal of evolutionary biology 26.12 (2013): 2710-2715.

variable	description
`altitude`	Altitude of the study site in meters above sea level
`latitude`	Latitude of the study site measured in degrees
`egg.size`	Average diameter of an individual egg to the 0.01 mm
`clutch.size`	Estimated number of eggs in clutch
`clutch.volume`	Volume of egg clutch in mm^3
`body.size`	Length of mother frog in cm

Table 1.4: Variables and their descriptions for the `frog` dataset.

for each clutch.

For example, the first row represents a clutch collected at altitude 3,462 meters above sea level, latitude 34.82 degrees; the clutch contained an estimated 182 eggs, with individual eggs averaging 1.95 mm in diameter, for a total volume of 177.8 mm^3. The eggs were laid by a female measuring 3.63 cm long. It is important to understand the definitions of variables, as they are not always obvious. For example, why has `clutch.size` not been recorded as whole numbers? This has to do with how the observations were collected. In a given clutch, researchers counted approximately 5 grams' worth of eggs and then estimated the total number of eggs based on the mass of the entire clutch. Definitions of the variables are given in Table 1.4.[4]

The data in Table 1.3 are organized as a **data matrix**. Each row of a data matrix corresponds to a unique observational unit, and each column corresponds to a variable. A data matrix for the LEAP study introduced in Section 1.1 is shown in Table 1.1 on page 10, in which the cases were patients and three variables were recorded for each patient. Data matrices are a convenient way to record and store data. If the data are collected for another individual, another row can easily be added; similarly, another column can be added for a new variable.

1.2.2 Types of variables

The Functional polymorphisms Associated with Human Muscle Size and Strength study (FAMuSS), funded by the National Institutes of Health (NIH), measured a variety of demographic, phenotypic, and genetic characteristics for about 1,300 participants.[5] Data

[4] The data discussed here are in the original scale; in the published paper, values have been log-transformed.

[5] Thompson PD, Moyna M, Seip, R, et al., 2004. Functional Polymorphisms Associated with Human Muscle Size and Strength. Medicine and Science in Sports and Exercise 36:1132 - 1139

from the study has been used in many subsequent studies[6], such as one examining the relationship between muscle strength and genotype at a location on the ACTN3 gene.[7] Four rows of the `famuss` dataset are shown in Table 1.5, and the variables are summarized in Table 1.6.[8]

	sex	age	race	height	weight	actn3.r577x	ndrm.ch
1	Female	27	Caucasian	65.0	199.0	CC	40.0
2	Male	36	Caucasian	71.7	189.0	CT	25.0
3	Female	24	Caucasian	65.0	134.0	CT	40.0
	⋮	⋮	⋮	⋮	⋮	⋮	
595	Female	30	Caucasian	64.0	134.0	CC	43.8

Table 1.5: Four rows from the `famuss` data matrix.

The variables `age`, `height`, `weight`, and `ndrm.ch` are **numerical** variables. They can take on a wide range of numerical values, and it is possible to add, subtract, or take averages with these values. On the other hand, a variable reporting telephone numbers would not be classified as numerical, since averages, sums, and differences in this context would have no meaning. Age measured in years is said to be **discrete**, since it can only take numerical values with jumps. On the other hand, percent change in strength in the non-dominant arm (`ndrm.ch`) is said to be **continuous**.

The variables `sex`, `race`, and `actn3.r577x` are **categorical** variables, and the possible values are called the variable's **levels**.[9] For example, the levels of `actn3.r577x` are the three possible genotypes at this particular locus: CC, CT, or TT. Categorical variables with levels that have a natural ordering can be more specifically referred to as **ordered categorical** variables. There are no ordered categorical variables in the `famuss` data, but it would be easy to create one; age of the participants grouped into 5-year intervals (15-20, 21-25, 26-30, etc.) would be an ordered categorical variable. Statistical software such as R calls categorical variables **factors**, and the possible values of factors are called **levels**.

In the `frog` data, the variables `egg.size`, `clutch.size`, `clutch.volume`, and `body.size` are all continuous variables. *DH: JV, agree about latitude? JV: I don't think either altitude or*

[6] Pescatello L, et al. Highlights from the functional single nucleotide polymorphisms associated with human muscle size and strength or FAMuSS study, BioMed Research International 2013.

[7] Clarkson P, et al., Journal of Applied Physiology 99: 154-163, 2005.

[8] Data freely available at `http://people.umass.edu/foulkes/asg/data.html`

[9] Categorical variables are sometimes called **nominal** variables.

variable	description
sex	Sex of the participant
age	Age in years
race	Recorded as African Am (African American), Caucasian, Asian, Hispanic and Other
height	Height in inches
weight	Weight in lbs
actn3.r577x	Genotype at the location r577x in the ACTN3 gene.
ndrm.ch	Percent change in strength in the non-dominant arm, comparing strength after to before training

Table 1.6: Variables and their descriptions for the famuss data set.

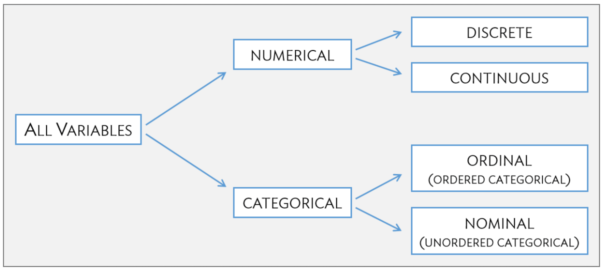


Figure 1.7: Breakdown of variables into their respective types.

latitude qualifies, because they refer to 11 specific locations; both function more like categorical variables in this context.

● **Example 1.1** Suppose a research assistant collected data on the first 20 individuals to visit one of the new walk-in clinics being offered by major commercial pharmacies. In addition to other variables, the research assistant collected age (measured as less than 21, 21 - 65, and older than 65 years of age), sex, height, weight, and reason for the visit. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

Height and weight are continuous numerical variables. Age as measured by the research assistant is ordered categorical. Sex and the reason for the visit are nominal categorical variables; sex has two categories, while reason for the visit will have many possible values.

⊙ **Guided Practice 1.2** Characterize the variables participant.ID, treatment.group, and overall.V60.outcome from the LEAP study (discussed in Section 1.1). [10]

[10] These variables measure non-numerical quantities, and thus are categorical variables. The variables treat-

1.2.3 Relationships between variables

Many studies are motivated by a researcher examining a possible relationship between two or more variables. Statistical relationships between two variables occur when they tend to vary in a related way. A **response variable** measures an outcome of interest, while an **explanatory variable** may be useful in predicting or understanding the response variable. There may be several possible explanatory variables for a single response variable in a given study.

Researchers were interested in using the `famuss` data to answer several questions, including: is ACTN3 genotype associated with variation in muscle function? The ACTN3 gene codes for a protein involved in muscle function. A common mutation (polymorphism) at residue 577 in the ACTN3 gene changes C to T; TT individuals are unable to produce any ACTN3 protein in their muscle. Thus, researchers hypothesized that ACTN3 genotype might influence muscle function in humans. The response variable in this study is `ndrm.ch`, the change in non-dominant arm strength, with strength gain being used as a way to measure muscle function. The explanatory variable of interest is `actn3.r557x`, ACTN3 genotype at residue 577.

Both numerical and graphical ways to examine possible relationships between two variables will be covered later in the text.

● **Example 1.3** In the study conducted on Tibetan frogs, researchers collected measurements on egg clutches and female frogs at 11 study sites, located at differing altitudes. Identify the explanatory and response variables in the study.

The explanatory variable examined in the study is `altitude`. The variables `egg.size`, `clutch.size`, `clutch.volume`, and `body.size` are response variables measuring the level of maternal investment.

⊙ **Guided Practice 1.4** Refer to the variables from the `famuss` data set described in Table 1.6 to formulate two questions about the relationships between these variables

`ment.group` and `outcome.V60.overall` have two values or levels, while `participant.ID` has many possible values.

that differ from the one addressed by the research team.[11]

1.3 Data collection principles

The first step in conducting research is to identify questions to investigate. A clearly articulated research question is essential for selecting subjects to be studied, identifying relevant variables, and determining how data should be measured. In order to obtain reliable data, it is also important to consider *how* data are collected.

1.3.1 Populations and samples

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. If an infant seems predisposed to a peanut allergy, is it better to introduce or to avoid peanut products during the first 6 months of the infant's life?
3. What proportion of female college students experience sexual victimization?

Each of these questions refers to a target **population**. In the first question, the target population is all swordfish in the Atlantic ocean, and each fish represents a case. Almost always, it is either too expensive or logistically impossible to collect data for every case in a population, so nearly all research is based on samples from populations. A **sample** represents a subset of the cases and is often a small fraction of the population. For instance, 60 swordfish (or some other number) in the population might be selected, and this sample data may be used (with some assumptions) to provide an estimate of the population average and answer the research question.

DH: Removed the exercise question on identifying samples for three reasons: the stent example is gone; it refers to the stent example as if it was a drug (it isn't); and I want us to be both realistic and clear about samples. They are almost never random samples from the ideal target population. If asked, I think almost any student, even at the high school level would say that drawing a random sample from the population of people with a disease is fundamentally

[11] Two sample questions: (1) Do participants appear respond differently to training according to race? (2) Do male participants appear to respond differently to training than females?

impossible. We can replace the stent example by LEAP, but as in other trials, the validity comes from the randomization among the recruited subjects, not the assumption of it being a RS.

1.3.2 Anecdotal evidence

Anecdotal evidence is typically composed of unusual observations that are easily recalled based on their striking characteristics. Physicians are sometimes more likely to remember the characteristics of a single patient with an unusually good response to a drug as opposed to the many patients who did not respond. The dangers of drawing general conclusions from anecdotal information are obvious; no single observation can be used to draw conclusions about a population. Often, the anecdotal case may not have been remembered correctly or may have involved errors in measurements. To learn about the characteristics of a population, it is necessary to examine a sample of many cases drawn randomly from the population.

Thomas Jefferson, the second president of the United States, recognized the pitfall of drawing conclusions from a single observation. In a letter to his nephew, he wrote "The patient, treated on the fashionable theory, sometimes gets well in spite of the medicine. The medicine therefore restored him, and the young doctor receives new courage to proceed in his bold experiments on the lives of his fellow creatures." [12]

JV: I find the reference to TJ rather funny, but I suspect the story only appeals to people with a very specific sense of humor...Would it break your heart to omit this? Doing so may also improve the flow here.

While it is incorrect to generalize from individual observations, scientists know that unusual observations can sometimes be valuable; such observations may be a reason to question previously held assumptions or to design a study to examine an unconventional perspective. Cures for certain diseases have been discovered through research inspired by a patient with a disease thought be be incurable responding to a new drug. *DH: Insert references sent by D Longo and D. Spriggs here. JV: This last sentence is too vague/unwieldy, adjust once references are included.*

An anecdotal observation can never be the basis for a conclusion, but it may well lead

[12] Jefferson, T.(1985). Letters, 1760âĂŞ1826. Ed. Merrill D. Peterson. New York: Viking.

to the design of a more systematic study that could be definitive.

1.3.3 Sampling from a population

Sampling from a population is a useful tool in population-based research in the health sciences. When done carefully, it provides reliable information about the health characteristics of a large population without having to directly measure those characteristics for each member, which is often an impossible task. The US Centers for Disease Control (US CDC) conducts many such surveys, including the Behavioral Risk Factors Surveillance System (BRFSS)[13]. The BRFSS conducts approximately 400,000 telephone interviews annually to ask U.S. residents questions regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. The CDC conducts similar surveys for diabetes, health care access, and immunization. Likewise, the World Health Organization (WHO) conducts the World Health Survey in partnership with approximately 70 countries to learn about the health of adult populations and the health systems in those countries.[14] In 2000, the US Department of Justice released the *The Sexual Victimization of College Women*, based on a survey conducted in 1996 of 4,446 undergraduate women.[15]

DH: we should check to see if the data are available at DoJ website. JV: The page for the sexual victimization survey only references NCVS data, which is freely available online (scroll down to "Codebooks and Datasets" under Documentation section), but I couldn't find data for questions specific to that survey. Perhaps better to switch to a survey for which data are available online? For example, Gender and Violent Victimization, see http://doi.org/10.3886/ICPSR27082.v1

DH: placeholder for the Harvard survey, despite its flaws

Sampling from a population is easier when the population is relatively small and members of the population are easy to identify and contact. For instance, the quality care team at an integrated health care system, such as Kaiser Permanente or Harvard Pilgrim Health Care, might like to learn about how members of the system perceive quality of care. Since health plans have contact information for each of their members, a selected subset can be contacted (with their consent) for participation in an interview or mailed

[13] http://www.cdc.gov/brfss/

[14] http://www.who.int/healthinfo/survey/en/

[15] https://www.ncjrs.gov/pdffiles1/nij/182369.pdf

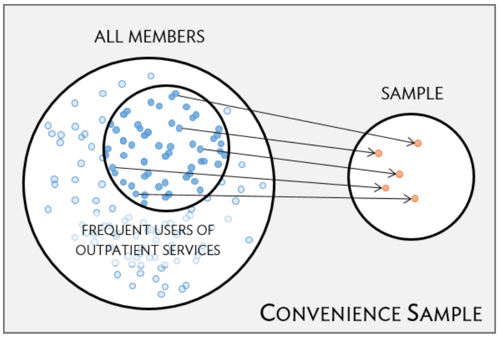


Figure 1.8: Instead of sampling from all members equally, approaching members visiting a clinic during a particular week would disproportionately select members who typically use outpatient services.

survey. More complex methods are required for other surveys, such as the study on sexual victimization of college women.

One common downfall in conducting a sample is to use a **convenience sample**, in which individuals who are easily accessible are more likely to be included in the sample. For instance, the quality control team in the healthcare plan might ask interviewers to approach plan members visiting an outpatient clinic during a particular week. The sample would fail to enroll generally healthy members who typically do not use outpatient services or schedule routine physical examinations. Similarly, the Department of Justice could have only sampled women in colleges or universities in or near the District of Columbia.

The general principle of sampling is straightforward; a sample from a population is useful for learning about a population only when the sample matches, on average, the characteristics of the population. Random sampling is the best way to ensure that a sample reflects a population, because random samples do not reflect the conscious or unconscious bias of the team gathering the sample. However, even a well-defined sampling strategy can lead to an unrepresentative sample if there are substantial barriers to subject participation, such as questions that assume participants are fluent in English or calls to potential participants that do not account for working hours or time-zone differences.

The easiest random samples to analyze are those in which each member of a population has the same chance of being sampled. In a **simple random sample**, each member of the population is directly chosen at random for the sample, with probability the size of the sample divided by the size of the population. Simple random samples are essentially equivalent to how raffles are conducted. For example, if there are 5 prizes available and 100 people each have a single ticket, each person has a 5% chance (5/100) of being called. In the health plan example, a subset of members might be chosen randomly from the plan membership roster for an interview. *OpenIntro*, third edition, Section 1.4.2 describes the four most commonly used sampling strategies.

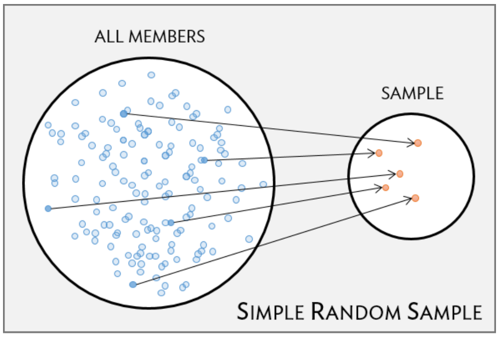


Figure 1.9: In this graphic, five members are randomly selected from the population to be interviewed.

Sometimes a simple random sample is difficult to implement and an alternative method is helpful. One such substitute is a **systematic sample**, in which one case is sampled after letting a fixed number of other cases pass by. Since this approach uses a mechanism that is not easily subject to personal biases, it often yields a reasonably representative sample. This book will focus on random samples since the use of systematic samples is uncommon and requires additional considerations of the context. *JV: It may be possible to omit this paragraph. If not, needs an example.*

The act of taking a simple random sample helps minimize bias, but bias can crop up in other ways. Even when people are picked at random, caution must be exercised if the **non-response** is high. For instance, if only 30% of the people randomly sampled for a survey actually respond, then it is unclear whether the results are truly **representative** of the entire population. Such **non-response bias** can skew results; it is important to minimize barriers that might discourage subject participation in order to collect reliable data.

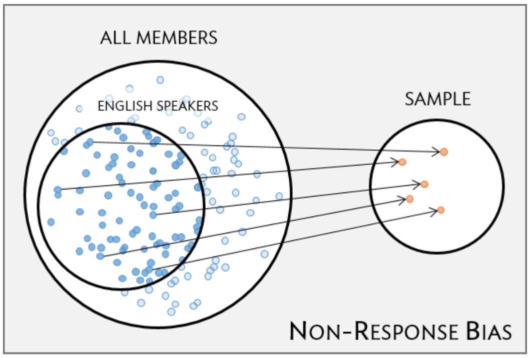


Figure 1.10: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. For example, a survey written in English may only result in responses from health plan members fluent in English.

Guided Practice 1.5

DH: replace this with a better example We can easily access ratings for products, sellers, and companies through websites. These ratings are based only on those people who go out of their way to provide a rating. If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?[16]

1.3.4 Introducing experiments and observational studies

Experiments and observational studies are the two primary types of study designs used to collect data.

When researchers want to investigate the possibility of a causal connection, they conduct an **experiment**. For instance, it might be hypothesized that administering a certain drug will reduce mortality in heart attack patients. To find evidence for a causal connection between the explanatory and response variables, researchers will collect a sample of individuals and randomly assign them into one of two groups. The first group, called a control group, may receive either a **placebo** (an inert substance with the appearance of

[16] Answers will vary. From our own anecdotal experiences, we believe people tend to rant more about products that fell below expectations than rave about those that perform as expected. For this reason, we suspect there is a negative bias in product ratings on sites like Amazon. However, since our experiences may not be representative, we also keep an open mind.

the study drug) or a commonly used drug known to have some effect; the second group (the experimental group) receives the new drug.

Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise. For instance, to study why certain diseases develop, researchers may collect information through conducting surveys, reviewing medical or company records, or following a **cohort** of many similar individuals. In each of these situations, researchers merely observe the data that arise. Observational studies can provide evidence of an association between variables, but they cannot by themselves show a causal connection. In general, causation can only be inferred from a randomized experiment.

1.3.5 Experiments

Studies in which researchers assign treatments to cases are called **experiments**. Randomized experiments are generally built on three principles.

Controlling. Researchers assign treatments to cases, doing their best to **control** for any other differences in the groups. For example, all infants enrolled in the LEAP study were required to be between 4 and 11 months of age, with severe eczema and/or allergies to eggs.

Randomization. Researchers randomize patients into treatment groups to account for variables that cannot be controlled. For example, some infants may have been more susceptible to peanut allergies because of an unmeasured genetic condition. Randomly assigning patients to the treatment or control group helps even out such differences. In situations where researchers suspect that variables other than the treatment may influence the response, they may first group individuals into **blocks** and then, within each block, randomize cases to treatment groups; this technique is referred to as **blocking** or **stratification**. In the LEAP study, infants were stratified into two cohorts based on whether or not the child developed a red, swollen mark (a wheal) after a skin test at the time of enrollment. The main analysis of the study analyzed data collected for infants without a wheal after the skin test. Figure 1.11

illustrates the blocking scheme used in the study. General methods for analyzing blocked data are relatively complicated and will not be covered in this book.

Replication. The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, **replication** is accomplished by collecting a sufficiently large sample. The LEAP study randomized a total of 640 infants; 542 infants were in the block without the wheal response.

It is important to incorporate the three experimental design principles into any study; this book describes applicable methods for analyzing data from such experiments. Blocking is a slightly more advanced technique, and statistical methods in this book may be extended to analyze data collected using blocking.

DH: Update the following figure for the LEAP data. JV: I changed the labels, but I'm getting lost in how to change the number in individuals in each group so that they are not equal.

1.3.6 Reducing bias in human experiments

Randomized experiments are the gold standard for data collection, but they do not automatically ensure an unbiased perspective in all cases. Human studies are perfect examples where bias can arise unintentionally.

In 1980, researchers reported the results of a study assessing the efficacy of a new drug used to treat heart attack patients.[17] Researchers wanted to know whether the drug reduced deaths in patients; in order to draw causal conclusions about the effect of the drug, they designed a randomized experiment in which study volunteers were randomly assigned to one of two study groups.[18] The **treatment group** received the drug; the other group, called the **control group**, did not receive any drug treatment.

Typically, researchers do not want patients to know which group they are in. The emotional response of a patient who knows they are either receiving or not receiving a potentially helpful new drug may cause different behavior between the two groups. In order to eliminate this source of bias from the study, researchers conduct a **blinded**

[17] 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.

[18] Human subjects are often called **patients**, **volunteers**, or **study participants**.

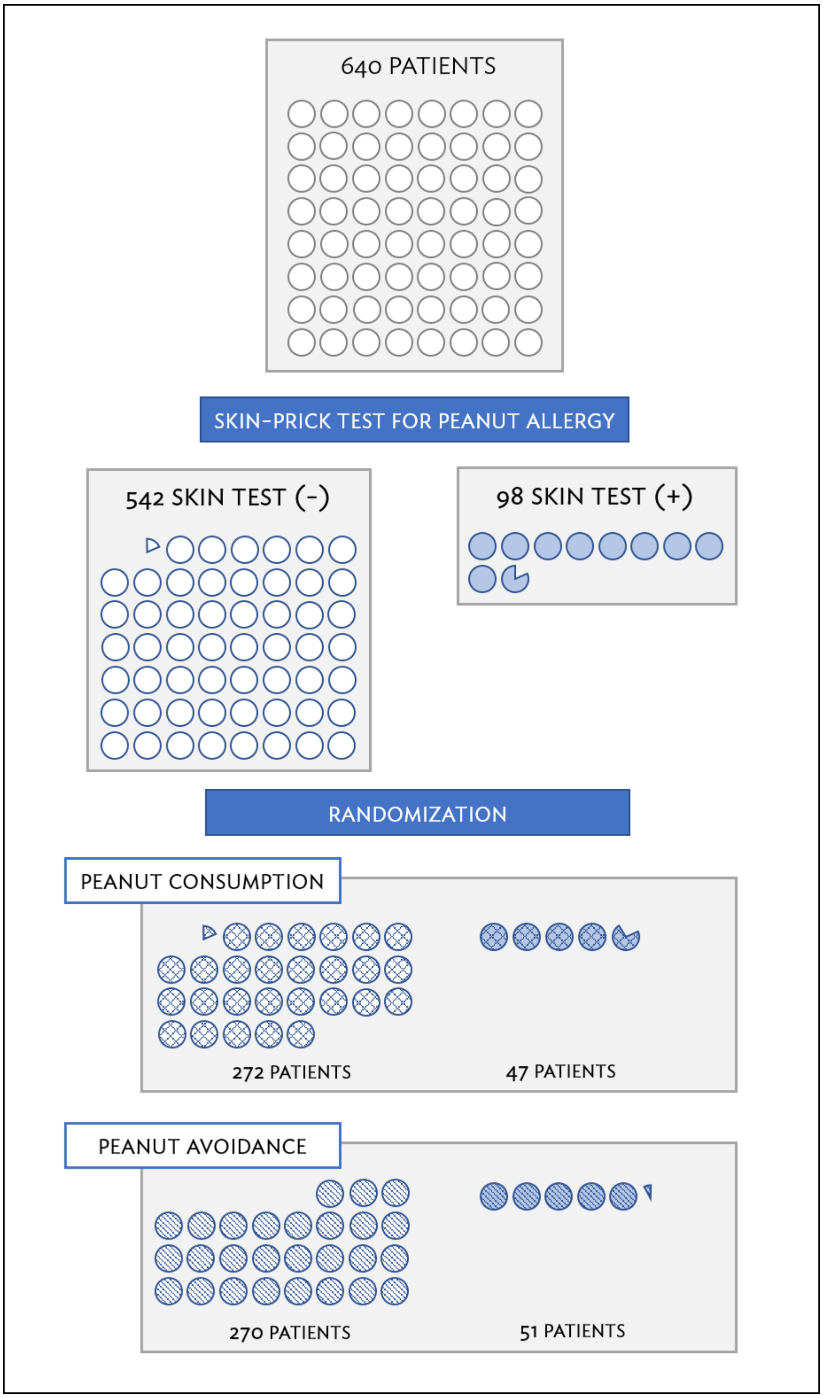


Figure 1.11: A simplified schematic of the blocking scheme used in the LEAP study, depicting 54 patients that underwent randomization. Patients are first divided into groups based on response to the initial skin test, then each block is evenly separated into the treatment groups using randomization. This strategy ensures an even representation of patients in each treatment group from both the skin test positive and skin test negative groups.

study in which patients are kept uninformed about their treatment. Patients in the control group, instead of being given a drug, are given an inert substance called a **placebo**. An effective placebo is the key to making a study truly blind. A placebo may often result in a slight but real improvement in patients; this effect is referred to as the **placebo effect**.[19]

The patients are not the only ones who should be blinded: doctors and researchers can accidentally bias a study. For example, out of concern for potential side effects of a new drug, a doctor might inadvertently give a patient in the treatment group more attention and care than they would to a patient known to be taking a placebo. To guard against this bias, which has also been found to have a measurable effect in some instances, most modern studies employ a **double-blind** setup in which doctors who interact with patients are, just like the patients, unaware of who is or is not receiving the experimental treatment.[20]

⊙ **Guided Practice 1.6** Look back to the study in Section 1.1 in which researchers were testing whether peanut product consumption was effective at reducing the likelihood of peanut allergies in children at-risk for these allergies. Is this an experiment? Was the study blinded? Was it double-blinded? [21]

1.3.7 Observational studies

Generally, data in observational studies are collected only by monitoring what occurs, while experiments require researchers to assign the primary explanatory variable in a study for each subject *JV: This sentence should be reworded for clarity, not sure how.* Making causal conclusions based on experiments is often reasonable; however, making the same causal conclusions solely based on observational data should be avoided.

DH: This is another instance where absolute statements are risky. The data on smoking and lung cancer are all observational, or were until a short time ago. I wonder if we should

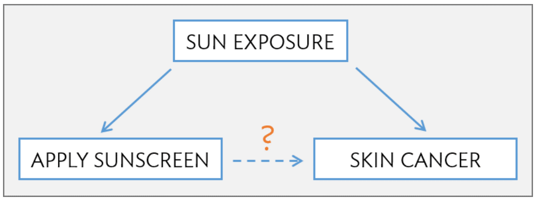
[19] Kaptchuk, TJ and Miller, FG. 2015.Placebo effects in medicine, New England Journal of Medicine, 373(1):8-9.

[20] There are always some researchers involved in the study who do know which patients are receiving which treatment. However, they do not directly interact with patients and do not tell the blinded health care professionals who is receiving which treatment.

[21] The researchers assigned the patients into their treatment groups, so this study was an experiment. However, the patients (and their parents) could distinguish which treatment they received, so this study was not blind. The study could not be double-blind since it was not blind.

mention that. JV: I have adjusted the language in the first paragraph. Scroll down past the sunscreen example to see where I think the mention of smoking and lung cancer can come in.

Suppose an observational study tracked sunscreen use and skin cancer, finding that the more sunscreen a person uses, the more likely they are to have skin cancer. However, this does not mean that sunscreen *causes* skin cancer. One important piece of missing information is sun exposure – if someone is out in the sun all day, they are both more likely to use sunscreen and to get skin cancer. Sun exposure is a **confounding variable**: a variable correlated with both the explanatory and response variables.[22] There is no guarantee that all confounding variables can be examined or measured; as a result, it is difficult to justify making causal conclusions from observational studies.



Observational studies are useful in that they can reveal interesting patterns or associations, providing researchers with the information necessary to design follow-up experiments. For example...

JV: Smoking and lung cancer can come in here, or anything else that illustrates the point, which I think is an important one. There might also be an example here that asks for a description of a reasonable follow-up experiment from some observational data – frog and famuss both seem too difficult, though.

Observational studies come in two forms: prospective and retrospective studies. A **prospective study** identifies individuals and collects information as events unfold. For instance, medical researchers may identify and follow a group of similar individuals over many years to assess the possible influences of behavior on cancer risk. One example of such a study is The Nurses' Health Study, started in 1976 and expanded in 1989.[23] This prospective study recruits registered nurses and then collects data from them using questionnaires. **Retrospective studies** collect data after events have taken place, e.g. re-

[22] Also called a **lurking variable**, **confounding factor**, or a **confounder**.

[23] `www.channing.harvard.edu/nhs`

searchers may review past events in medical records. Some datasets may contain both prospectively- and retrospectively-collected variables. *DH: need an example of this. famuss does not qualify, I think. The flanders dental study would qualify but has not been introduced at this point. JV: The dental study can be introduced here in the same way as the Nurses' Health study, then.*

DH: I have commented out the section on sampling methods, though I like it. It will take some work to find example to make the sampling methods concrete. I have left the OpenIntro source for this topic in our source.

1.4 Examining numerical data

This section introduces techniques for exploring and summarizing numerical variables, using the `frog` data from Section 1.2.

1.4.1 Measures of center: mean and median

The **mean**, sometimes called the average, is a common way to measure the center of a **distribution** of data. To find the average clutch volume for the observed egg clutches, we add up all the clutch volumes and divide by the total number of clutches. For computational convenience, the volumes are rounded to the first decimal.

$$\overline{x} = \frac{177.8 + 257.0 + \cdots + 933.3}{431} = 882.5 \text{ mm}^3 \tag{1.7}$$

$\overline{x}$
sample mean

The sample mean is often labeled $\overline{x}$. The letter x is being used as a generic placeholder for the variable of interest, `clutch.volume`, and the bar over the x communicates that the average volume of the 431 clutches is 882.5mm^3. It is useful to think of the mean as the balancing point of the distribution. *JV: This last sentence needs some further clarification if is to be included – I do not think the meaning is obvious to someone new to statistics.*

Mean

The sample mean of a numerical variable is computed as the sum of all of the observations divided by the number of observations:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1.8}$$

where $x_1, x_2, \dots, x_n$ represent the n observed values.

Another measure of center is the **median**, which is the middle number in a distribution after the values have been ordered from smallest to largest. If the distribution contains an even number of observations, the median is the average of the middle two observations. There are 431 clutches in the dataset, so the median is the clutch volume of the 216^{th} observation in the sorted values of `clutch.volume`: 831.8 mm^3.

DH: We have removed the concept of a weighted mean here; we do not have a context for it. But it is an important idea that can be profitably used later, perhaps with the brfss data. Perhaps we can re-insert it later. JV: Recommend introducing weighted mean in probability with calculating expected value.

1.4.2 Measures of spread: standard deviation and interquartile range

DH: Don't like the verbal description of the sd here, but have not replaced it yet. Note also that I have changed some bar to overline. Let me know which you think is better in the pdf; I like overline. JV: I also like overline, since I prefer reading at relatively small magnification. I will change the rest to overline.

The standard deviation measures approximately the distance between a typical observation and the mean. The distance of an observation from its mean its **deviation**. Below are the deviations for the 1^{st}, 2^{nd}, 3^{rd}, and 431^{th} observations in the `clutch.volume` variable. For computational convenience, clutch volume is rounded to the first decimal.

$$\begin{aligned} x_1 - \overline{x} &= 177.8 - 882.5 = -704.7 \\ x_2 - \overline{x} &= 257.0 - 882.5 = -625.5 \\ x_3 - \overline{x} &= 151.4 - 882.5 = -731.1 \\ &\vdots \\ x_{431} - \overline{x} &= 933.2 - 882.5 = 50.7 \end{aligned}$$

If we square these deviations and then take an average, the result is the sample **variance**, denoted by s^2:

s^2
sample
variance

$$\begin{aligned} s^2 &= \frac{(-704.7)^2 + (-625.5)^2 + (-731.1)^2 + \cdots + (50.7)^2}{431 - 1} \\ &= \frac{496,602.09 + 391,250.25 + 534,507.21 + \cdots + 2570.49}{430} \\ &= 143,680.9 \end{aligned}$$

The denominator is $n - 1$ rather than n when computing the variance; this mathematical nuance comes from statistical theory and the reason for doing so is not covered in this text.

The **standard deviation** is the square root of the variance:

$$s = \sqrt{143,680.9} = 379.05$$

s
sample
standard
deviation

The standard deviation of clutch volume for the egg clutches observed is about 380 mm^3. *DH: excellent place to give an interpretation of sd, referring back to verbal definition, or perhaps to use it to mention the empirical rule, which is in the caption to one of the Open Intro plots. JV: Return to this once verbal definition of SD is refined. The empirical rule is best introduced with a picture – I considered introducing it in an example after histograms, but that would interrupt the flow. May work in the transforming data subsection.*

Formulas and methods used to compute the variance and standard deviation for a population are similar to those used for a sample.[24] However, like the mean, the population values have special symbols: σ^2 for the variance and σ for the standard deviation. The symbol σ is the Greek letter *sigma*.

σ^2
population variance

σ
population standard deviation

Standard Deviation

The sample standard deviation of a numerical variable is computed as the square root of the variance, which is the sum of squared deviations divided by the number of observations minus 1.

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}} \qquad (1.9)$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values.

Variability can also be measured using the **interquartile range** (IQR). To calculate the IQR, find the **first quartile** (the 25^{th} percentile, i.e. 25% of the data fall below this value) and the **third quartile** (the 75^{th} percentile). These are often labeled Q_1 and Q_3, respectively. The IQR is the difference: $Q_3 - Q_1$.

The IQR for `clutch.volume` is $1096.0 - 609.6 = 486.4$ mm^3. The middle 50% of the values for `clutch.volume` lie between 609.6 mm^3 and 1096.0 mm^3.

1.4.3 Robust statistics

In the `frog` data, there are four observed clutch volumes larger than 2,000 mm^3 (2138.0, 2630.3, 2454.7, 2511.9). These values can be clearly identified by plotting the data as points on a single axis, as shown in Figure 1.12; this basic graphical display is known as a **dot plot**. How do these extreme values affect the summary statistics for the clutch volume variable in the `frog` data?

The sample statistics are computed under each of two scenarios in Table 1.13, one with and one without the four largest observations. The median and IQR are referred to

[24] The only difference is that the population variance has a division by n instead of $n-1$.

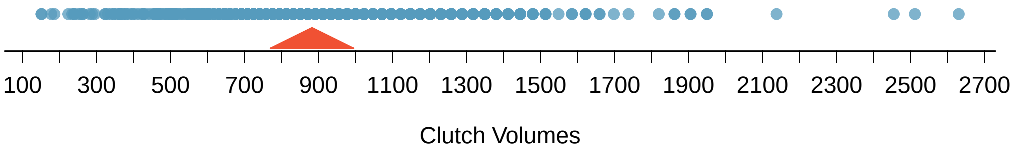


Figure 1.12: Dot plot of the clutch volume variable in the `frog` data.

as **robust estimates** because extreme observations have little effect on their values. For these data, the median does not change, while the IQR differs by only about 6 mm^3. In contrast, the mean and standard deviation are much more affected, particularly the standard deviation; since standard deviation depends on the squared distances from the mean, its change in the presence of large observations is more noticeable. Typically, extreme observations have a greater effect on the standard deviation than on the mean.

	robust		**not robust**	
scenario	median	IQR	$\bar{x}$	s
original `frog` data	831.8	486.9	882.5	379.1
drop four largest observations	831.8	493.92	867.9	349.2

Table 1.13: A comparison of how the median, IQR, mean ($\bar{x}$), and standard deviation (s) change when extreme observations are present.

JV: Agree that famuss data can be used in end-of-chapter exercise.

1.4.4 Visualizing distributions of data: histograms and boxplots

Visualizing how data are distributed can reveal characteristics of the data that are not obvious from summary statistics. Graphical summaries, such as histograms and boxplots, complement the information provided by numerical summaries.

Dot plots show the exact value of each observation; while this is useful for small datasets, dot plots are not ideal for larger samples. Instead, observations can be grouped into bins and plotted as bars to form a **histogram**. Table 1.14 shows the number of clutches with volume between 0 and 200 mm^3, 200 and 400 mm^3, etc. up until 2,600 and 2,800 mm^3. These binned counts are plotted in Figure 1.15.

Histograms provide a view of the **data density**. Higher bars indicate more common observations, while lower bars represent relatively rare observations. For instance, there

Clutch volumes	0-200	200-400	400-600	600-800	···	2400-2600	2600-2800
Count	4	29	69	99	···	2	1

Table 1.14: The counts for the binned `clutch.volume` data.

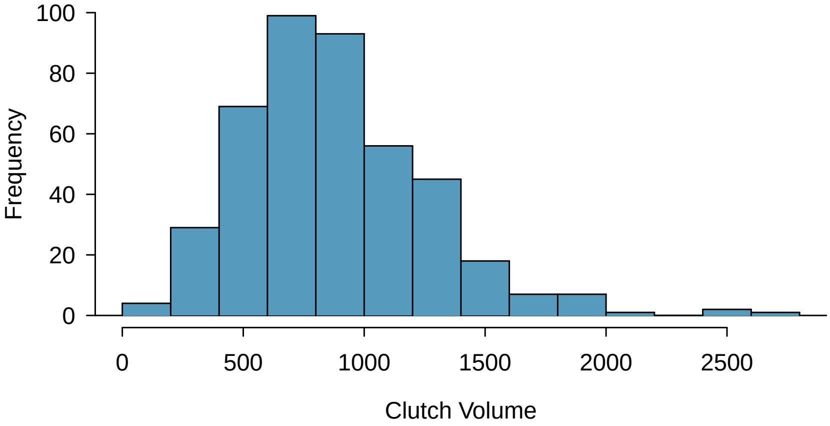


Figure 1.15: A histogram of `clutch.volume`.

are many more egg clutches with volumes smaller than 1,500 mm^3 than clutches with larger volumes. The bars make it easy to see how the density of the data changes relative to clutch volume.

Histograms are especially convenient for describing the shape of the data distribution. Figure 1.15 shows that most clutches have a relatively small volume, while fewer clutches are very large. When data trail off to the right, with a long right tail, the data are said to be **right skewed**.[25] Data with the reverse characteristic – a long, thin tail to the left – are said to be **left skewed**. The term **symmetric** is used to describe data that show roughly equal trailing off in both directions.

A **mode** is represented by a prominent peak in the distribution.[26] Figure 1.16 shows histograms that have one, two, or three prominent peaks. Such distributions are called **unimodal**, **bimodal**, and **multimodal**, respectively. Any distribution with more than two prominent peaks is called multimodal. Notice that there was one prominent peak in the

[25] Other ways to describe data that are skewed to the right: **skewed to the right**, **skewed to the high end**, or **skewed to the positive end**.

[26] Another definition of mode, which is not typically used in statistics, is the value with the most occurrences. It is common to have *no* observations with the same value in a data set, which makes this other definition useless for many real datasets.

unimodal distribution with a second less prominent peak that was not counted since it only differs from its neighboring bins by a few observations. *JV: Last sentence should be refined – thoughts on how to communicate that "prominent" is a subjective term?*

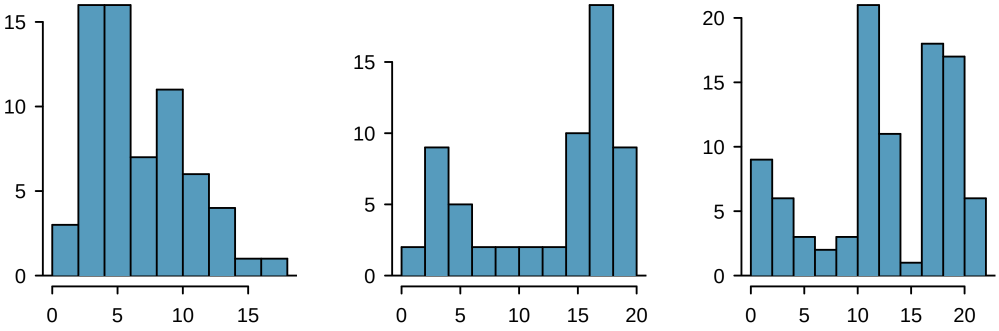


Figure 1.16: From left to right: unimodal, bimodal, and multimodal distributions.

⊙ **Guided Practice 1.10** Describe the distribution of `clutch.volume` using the histogram in Figure 1.15. Are the data skewed? Is it a unimodal, bimodal, or multimodal distribution? [27]

A **boxplot** summarizes a dataset using five statistics while also plotting unusual observations.[28] Figure 1.17 provides a vertical dot plot alongside a boxplot of the `clutch.volume` variable from the `frog` dataset.

In a boxplot, a rectangle extending from the first quartile to the third quartile represents the middle 50% of the data (the IQR); the rectangle is split in half by the **median**. Extending outwards from the box, the **whiskers** capture the data that fall between $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$.[29] Note that the whiskers must end at data points; the values given by adding or subtracting $1.5 \times IQR$ define the maximum reach of the whiskers. For example, for the `clutch.volume` variable: $Q_3 + 1.5 \times IQR = 1,096.5 - 1.5 \times 486.4 = 1,826.1$ mm^3. However, there was no clutch with volume 1,826.1 mm^3; thus, the upper whisker extends to 1,819.7 mm^3, the largest observation that is smaller than $Q_3 + 1.5 \times IQR$.

[27] The data is strongly skewed to the right; while many counts fall in the 600-1,000 mm^3 range, there are a few clutches with volume greater than 1,500 mm^3. The distribution is unimodal, with only one prominent peak.

[28] Boxplots are also known as box-and-whisker plots.

[29] While the choice of exactly 1.5 is arbitrary, it is a commonly used value for drawing boxplots.

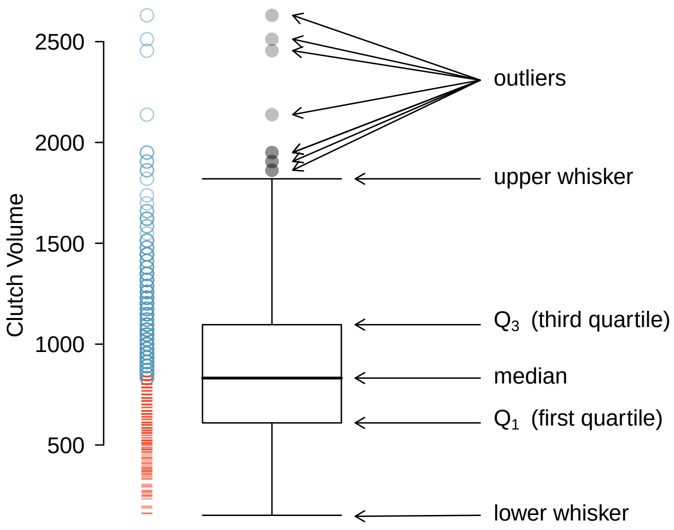


Figure 1.17: A vertical dot plot next to a labeled boxplot for the volumes of 431 egg clutches. The median (831.8 mm^3), splits the data into the bottom 50% and the top 50%, marked in the dot plot by horizontal dashes and open circles, respectively.

Any observation that lies beyond the whiskers is labeled with a dot; these observations are called outliers. An **outlier** is a value that appears extreme relative to the rest of the data. For the `clutch.volume` variable, there are several large outliers and no small outliers, indicating the presence of some unusually large egg clutches. Outliers can potentially provide insight into interesting properties of the data.

1.4.5 Scatterplots

A **scatterplot** provides a case-by-case view of data for two numerical variables. In the `frog` data, `clutch.volume` and `body.size` are two numerical variables of interest; previous research has reported that larger body size allows females to produce larger egg clutches. The relationship between clutch volume and female body size is examined via scatterplot in Figure 1.18. In any scatterplot, each point represents a single case. Since body size was measured for 129 frogs, there are 129 points in Figure 1.18.

The variables `clutch.volume` and `body.size` are said to be **associated** because the plot shows a discernible pattern. Since the points tend to lie in a straight line, the two variables are **linearly associated**. Two variables are **positively associated** if increasing

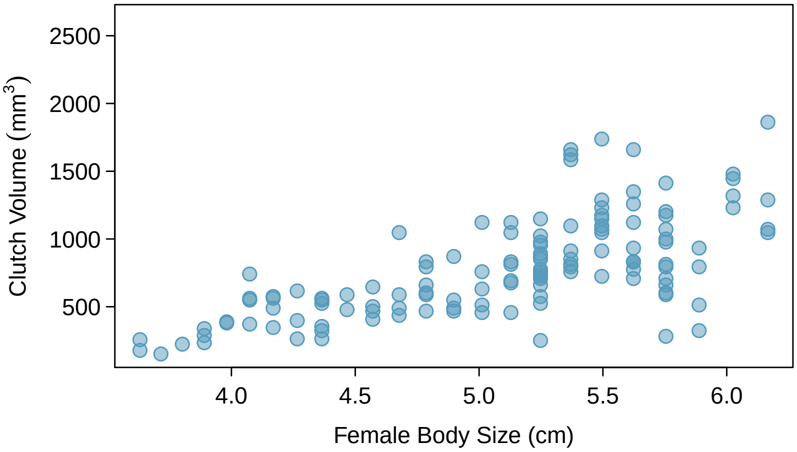


Figure 1.18: A scatterplot showing `clutch.volume` (horizontal axis) vs. `body.size` (vertical axis).

values of one tend to occur with increasing values of the other; similarly, variables are **negatively associated** if increasing values of one variable occurs with decreasing values of the other. Figure 1.18 shows an upward trend – as expected, larger frogs tend to produce egg clutches with larger volumes. Frog embryos are surrounded by a gelatinous matrix that may protect developing embryos from temperature fluctuation or ultraviolet radiation; these observations suggest that larger females are indeed capable of producing greater quantities of this material.

Figure 1.19 shows the relationship between `height` and `weight` for participants in the FAMuSS study. Each point on the plot represents a participant. As expected, taller participants tend to be heavier, so the variables `height` and `weight` are positively associated.

Taller people naturally tend to be heavier; as a consequence, weight itself is not a good measure of whether someone is overweight. Body mass index (BMI) is a measure of weight that is less affected by a person's height. A BMI of 30 or above is considered overweight.[30] In the metric system, BMI is calculated as weight in kilograms (kg) divided by height squared (m^2). If height and weight are measured in inches and pounds, as in the `famuss` data, then BMI is weight in pounds (lb) divided by height squared (in^2), then multiplied by 703. The `famuss` data includes the variable `bmi` for each participant, and

[30] http://www.nhlbi.nih.gov/health/educational/lose_wt/risk.htm

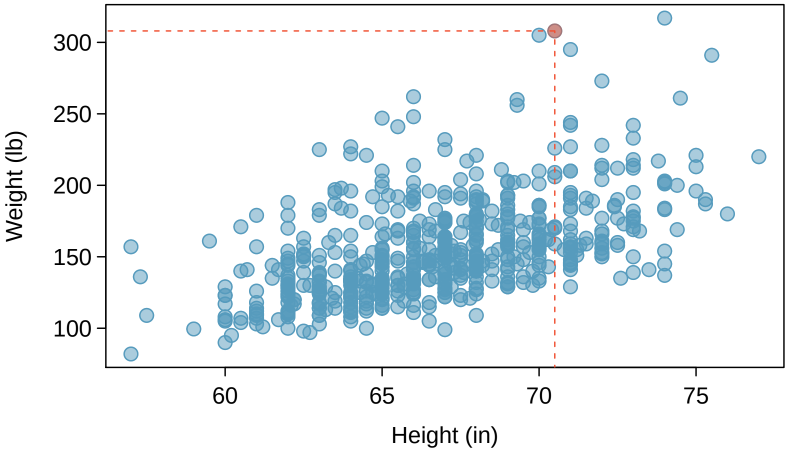


Figure 1.19: A scatterplot showing `height` (horizontal axis) vs. `weight` (vertical axis). One participant 70.5 inches tall and weighing 308 pounds is highlighted.

Figure 1.20 shows the relationship between `height` and `bmi`. The strong upward trend in Figure 1.19 is no longer evident, indicating that `height` and `bmi` have a much weaker association. For this reason, health agencies such as the US NIH and the World Health Organization (WHO) use BMI as a measure of obesity.

If two variables are not associated, then they are said to be **independent**. That is, two variables are independent if there is no evident relationship between the two. Generally, it is not easy to determine definitively whether two variables are independent from looking at a scatterplot, even in Figure 1.20.

JV: Need an example here showing a nonlinear relationship (along the lines of the car price vs weight example). Anything in famuss or LEAP? Might even work to quickly introduce caries data.

1.4.6 Transforming data (special topic)

DH: We should include a section on transforming data, esp since the frog data has been transformed. JV: Agree. I replaced the MLB histograms with ones for the clutch volume data; left the scatterplots as a placeholder, as I'm not sure the clutch volume vs body size plot is a good candidate for that example. This would be a good place to introduce the empirical rule, since the transformation on frogs makes the data look more normal; empirical rule will be discussed

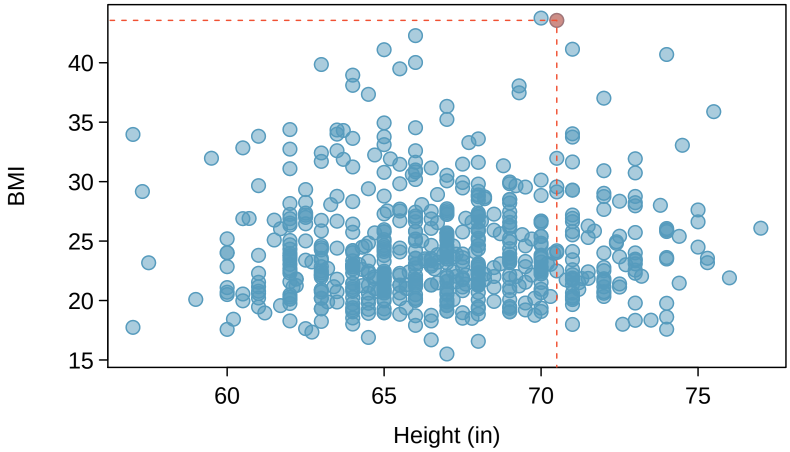


Figure 1.20: A scatterplot showing `height` (horizontal axis) vs. `bmi` (vertical axis). The same individual highlighted in Figure 1.19 is marked here, with BMI 43.56.

in more detail once normal distribution is introduced, anyways (so seems fine that it would first be mentioned as part of a special topic).

Scientists may choose to transform strongly skewed data in order to make them easier to model and analyze. A **transformation** is a rescaling of the data using a function. Consider the histogram of egg clutch volumes from the `frog` data, shown in Figure 1.21(a). In the published paper, researchers used a $\log_{10}$ transformation on the data before conducting analyses. Figure 1.21(b) shows a plot of the $\log_{10}$ of clutch volumes.

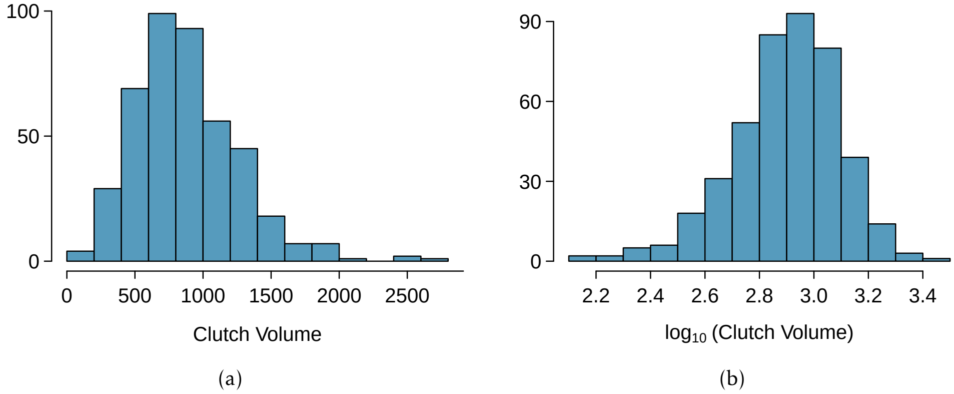


Figure 1.21: (a) Histogram of egg clutch volumes. (b) Histogram of the log-transformed egg clutch volumes.

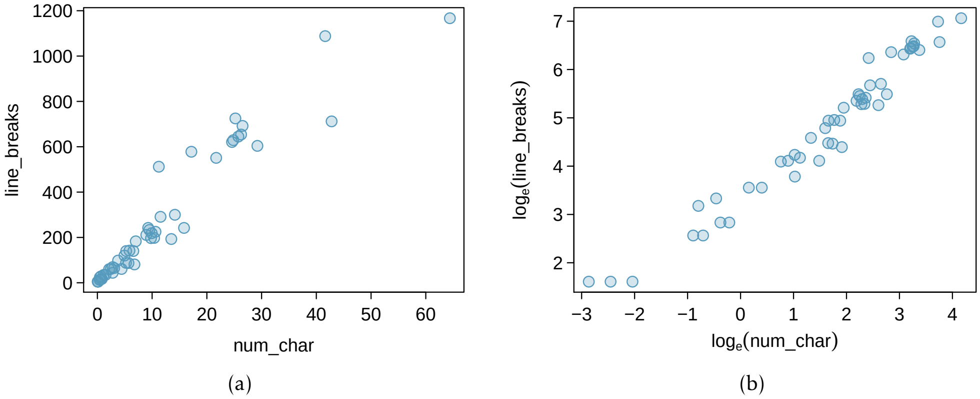


Figure 1.22: (a) Scatterplot of `line_breaks` against `num_char` for 50 emails. (b) A scatterplot of the same data but where each variable has been log-transformed.

JV: Text after this point (in this section) needs re-writing.

There are some standard transformations that are often applied when much of the data cluster near zero (relative to the larger values in the data set) and all observations are positive. A **transformation** is a rescaling of the data using a function. For instance, a plot of the natural logarithm[31] of player salaries results in a new histogram in Figure **??**. Transformed data are sometimes easier to work with when applying statistical models because the transformed data are much less skewed and outliers are usually less extreme.

Transformations can also be applied to one or both variables in a scatterplot. A scatterplot of the `line_breaks` and `num_char` variables is shown in Figure 1.22(a), which was earlier shown in Figure **??**. We can see a positive association between the variables and that many observations are clustered near zero. In Chapter **??**, we might want to use a straight line to model the data. However, we'll find that the data in their current state cannot be modeled very well. Figure 1.22(b) shows a scatterplot where both the `line_breaks` and `num_char` variables have been transformed using a log (base e) transformation. While there is a positive association in each plot, the transformed data show a steadier trend, which is easier to model than the untransformed data.

Transformations other than the logarithm can be useful, too. For instance, the square

[31] Statisticians often write the natural logarithm as log. You might be more familiar with it being written as ln.

root ($\sqrt{\text{original observation}}$) and inverse ($\frac{1}{\text{original observation}}$) are used by statisticians. Common goals in transforming data are to see the data structure differently, reduce skew, assist in modeling, or straighten a nonlinear relationship in a scatterplot.

1.5 Considering categorical data

Like numerical data, categorical data can also be organized and analyzed; however, numerical calculations cannot be done with categorical data. In this section, we will introduce tables and other basic tools for categorical data, using the `famuss` dataset introduced in Section 1.2.2.

1.5.1 Contingency tables

A table for a single variable is called a **frequency table**. Table 1.23 is a frequency table for the `actn3.r577x` variable. Recall that `actn3.r577x` is a categorical variable describing genotype at a location on the ACTN3 gene: CC, CT, or TT. If we replaced the counts with percentages or proportions, the table would be called a **relative frequency table**.

	CC	CT	TT	Sum
Counts	173	261	161	595

Table 1.23: A frequency table for the `actn3.r577x` variable.

Table 1.24 summarizes two variables: `race` and `actn3.r577x`. A table that summarizes data for two categorical variables in this way is called a **contingency table**.[32] Each value in the table represents the number of times a particular combination of variable outcomes occurred. For example, the first row of the table shows that of the African-American individuals, 16 are CC, 6 are CT, and 5 are TT.

Row and column totals, known collectively as **marginal totals**, are also included. The **row totals** provide the total counts across each row; **column totals** are the total counts down each column.

JV: Not sure how they added the variable labels to their table. I have left the OI tables and code in the source.

Table 1.25 shows the row proportions for Table 1.24. The **row proportions** are computed as the counts divided by their row totals. The value 16 at the intersection of `African American` and `CC` is replaced by 16/27 = 0.593; i.e., 16 divided by the row total, 27. The

[32] Contingency tables are also known as **two-way tables**.

	CC	CT	TT	Sum
African Am	16	6	5	27
Asian	21	18	16	55
Caucasian	125	216	126	467
Hispanic	4	10	9	23
Other	7	11	5	23
Sum	173	261	161	595

Table 1.24: A contingency table for `race` and `actn3.r577x`.

value 0.593 corresponds to the proportion of African-Americans in the study with genotype CC.

	CC	CT	TT	Sum
African Am	16/27 = 0.593	6/27 = 0.222	5/27 = 0.185	27/27 = 1.00
Asian	21/55 = 0.382	18/55 = 0.327	16/55 = 0.291	55/44 = 1.00
Caucasian	125/467 = 0.267	216/467 = 0.463	126/467 = 0.270	467/467 = 1.00
Hispanic	4/23 = 0.174	10/23 = 0.435	9/23 = 0.391	23/23 = 1.00
Other	7/23 = 0.304	11/23 = 0.478	5/23 = 0.217	23/23 = 1.00
Sum	173/595 = 0.291	261/595 = 0.438	161/595 = 0.271	595/595 = 1.00

Table 1.25: A contingency table with row proportions for the `race` and `actn3.r577x` variables.

● **Example 1.11** What does Table 1.25 highlight about the distribution of genotypes between different populations?

Ggenotype distributions vary between populations. For the Caucasian individuals sampled in the study, CT is the most common genotype at 46.3%. In contrast, over half (59.3%) of African Americans sampled are CC. CC is also the most common genotype for Asians, but in this population, genotypes are more evenly distributed: 38.2% of Asians sampled are CC, 32.7% are CT, and 29.1% are TT.

A contingency table of the column proportions is computed in a similar way, in which each **column proportion** is computed as the count divided by the corresponding column total. Table 1.26 shows such a table, and here the value 0.092 indicates that 9.2% of CC individuals in the study are African-American.

JV: Not sure if the following exercise is very clear, but there should be something here to warn against misinterpretation of the column proportions in this context. Important to point

	CC	CT	TT	Sum
African Am	16/173 = 0.092	6/261 = 0.037	5/161 = 0.191	27/595 = 0.045
Asian	21/173 = 0.080	18/261 = 0.104	16/161 = 0.993	55/595 = 0.092
Caucasian	125/173 = 0.776	216/261 = 0.828	126/161 = 0.728	467/595 = 0.785
Hispanic	4/173 = 0.023	10/261 = 0.062	9/161 = 0.034	23/595 = 0.038
Other	7/173 = 0.027	11/261 = 0.063	5/161 = 0.031	23/595 = 0.038
Sum	173/173 = 1.000	261/261 = 1.000	161/161 = 1.000	595/595 = 1.000

Table 1.26: A contingency table with column proportions for the `race` and `actn3.r577x` variables.

out that one limitation of the data is uneven representation between groups.

Guided Practice 1.12 As computed in Table 1.26, 77.6% of CC individuals in the study are Caucasian. Does this data suggest that in the general population, people of CC genotype are highly likely to be Caucasian?[33]

1.5.2 Bar plots

A bar plot is a common way to display a single categorical variable. The left panel of Figure 1.27 shows a **bar plot** for the `actn3.r577x` variable. In the right panel, the counts are converted into proportions (e.g. 173/595 = 0.291 for the `CC` genotype), showing the proportion of observations that are in each level (i.e. in each category).

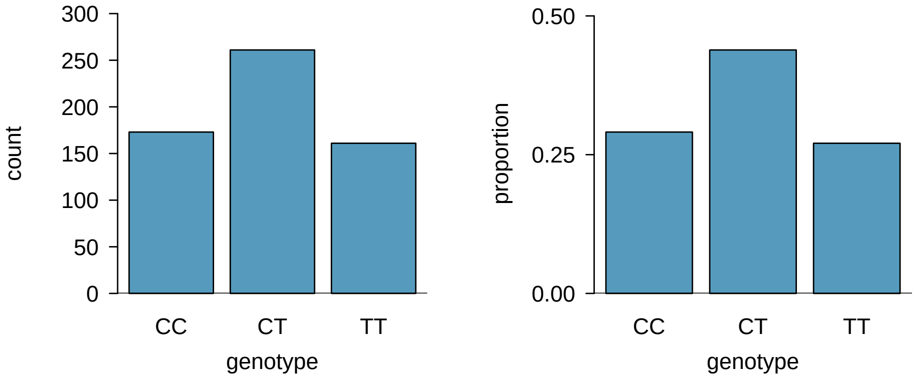


Figure 1.27: Two bar plots of `actn3.r577x`. The left panel shows the counts, and the right panel shows the proportions for each genotype.

[33] No, this is not a reasonable conclusion to draw from the data. The high proportion of Caucasians among CC individuals primarily reflects the large number of Caucasians sampled in the study – 78.5% of the people sampled are Caucasian. The uneven representation of different races is one limitation of the `famuss` data.

JV: After poking around in the R code for the segmented plots, I found out that they are made by overlaying a second graph on top of the other – a method that only works well for one category mapped onto the total; I tested out this method to make Figure 1.28, showing Caucasians as a part of the total, but it's not a particularly interesting/relevant plot.

JV: With the help of Google, I figured out how to make segmented bar plots using a different method – plots follow, one using genotype for the bars and one using race for the bars. The following paragraph needs to be re-written based on which plots we ultimately decide to include, and introduce the context for examining race and genotype together (mutant allele known to exist at different frequencies in various human populations). I did not find a way to make the standardized segmented plots.

Segmented bar plots provide a way to visualize the information in contingency tables. A **segmented bar plot** is a graphical display of contingency table information. For example, a segmented bar plot using data from Table 1.24 is shown in Figure 1.28, where a bar plot was created using the `actn3.r577x` variable, with each group divided by the levels of `race`. The column proportions of Table 1.26 have been translated into a standardized segmented bar plot in Figure 1.28(b), which is a helpful visualization of the races represented in each level of `actn3.r577x`.

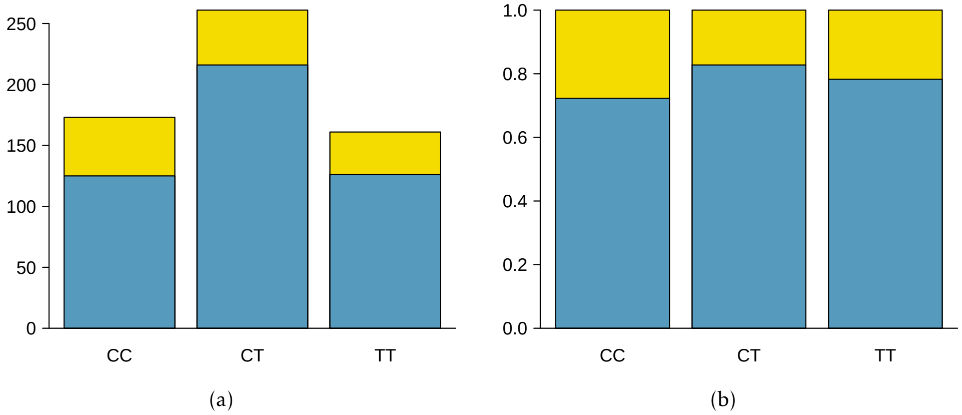


Figure 1.28: (a) Segmented bar plot for individuals by genotype, in which the counts have been further broken down by Caucasian (blue) and not-Caucasian (yellow). (b) Standardized version of Figure (a).

JV: I don't particularly like the mosaic plots, and they may not be the best choice for dis-

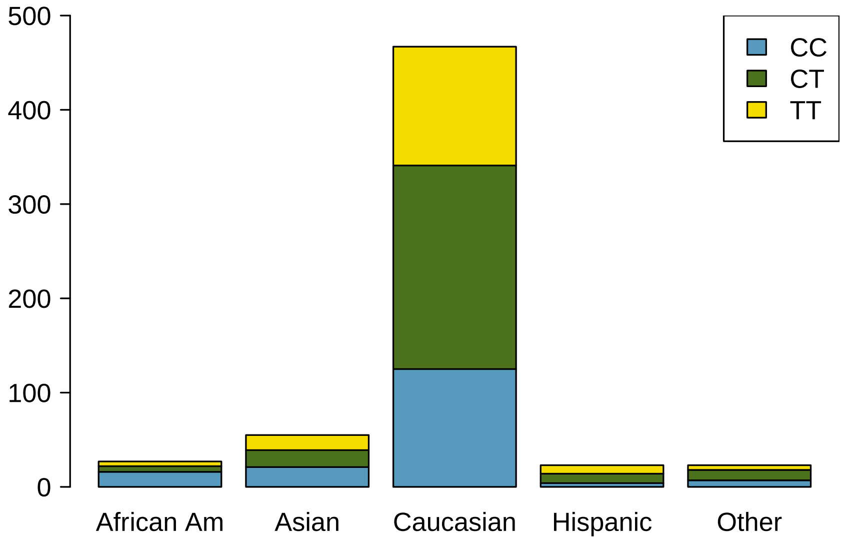


Figure 1.29: Segmented bar plot for individuals by race, where the counts have been further broken down by genotype.

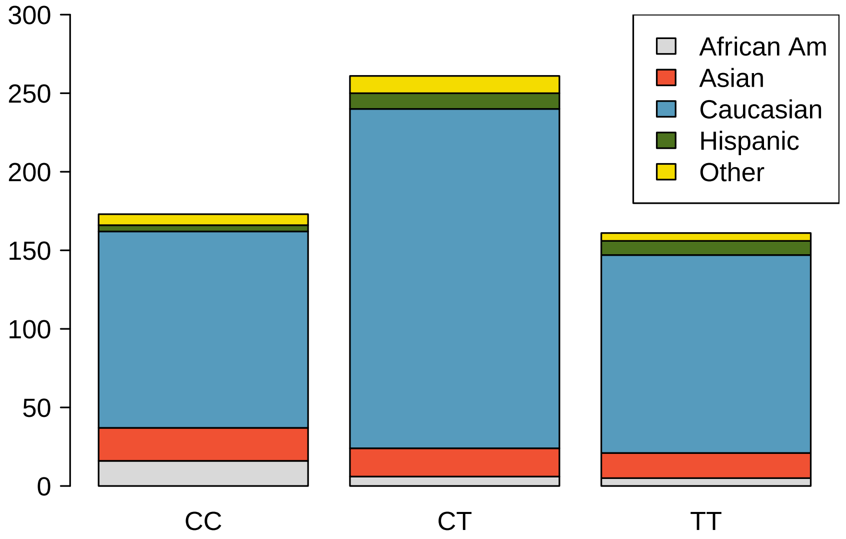


Figure 1.30: Segmented bar plot for individuals by genotype, where the counts have been further broken down by race.

playing the `famuss` *data, anyways. I also think the pie chart section doesn't necessarily need to be included.*

1.5.3 Comparing numerical data across groups

In this section, two convenient methods for examining numerical data across groups are introduced: side-by-side boxplots and hollow histograms. The **side-by-side boxplot** is a traditional tool for comparing across categories. The **hollow histogram** method plots the outlines of histograms for each group onto the same axes.

Recall the question introduced in Section 1.2.3: is ACTN3 genotype associated with variation in muscle function? To explore this question, genotype and variation in muscle function (measured by `ndrm.ch`) can be compared using side-by-side boxplots and hollow histograms, as shown in Figure 1.31. The histograms are useful for seeing distribution shape, skew, and groups of anomalies, while the side-by-side boxplots are especially useful for comparing centers and spreads. Comparison of median change in non-dominant arm strength between the two groups reveals that the TT genotype is associated with a greater increase in strength than CC or TT. In other words, the T allele appears to be associated with greater muscle function.

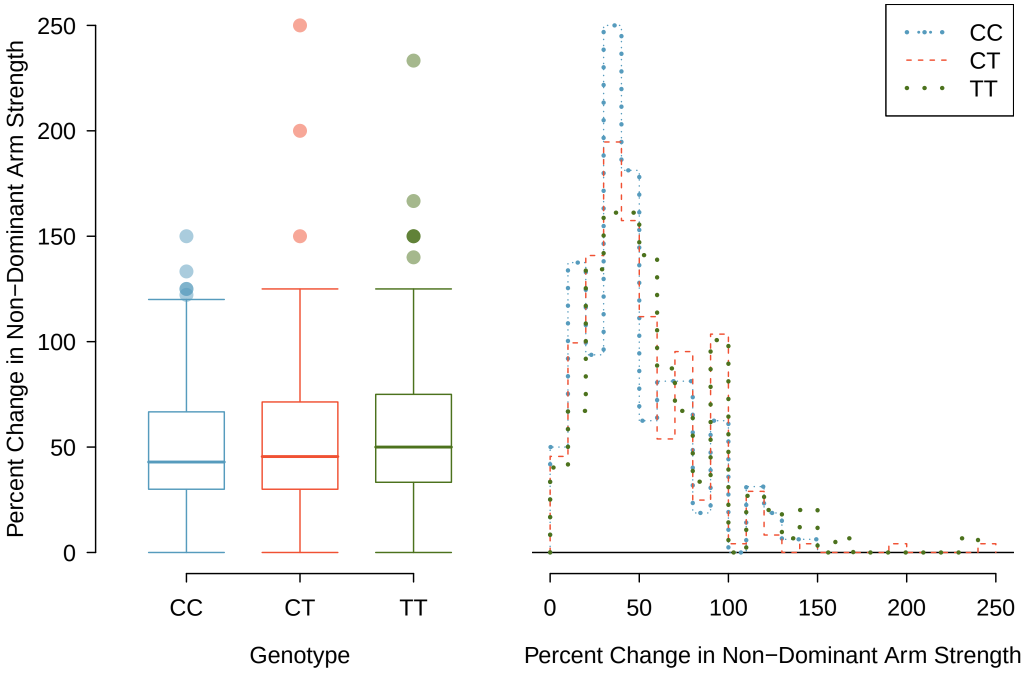


Figure 1.31: Side-by-side boxplot (left panel) and hollow histograms (right panel) for `ndrm.ch`, split by ACTN3 genotype.

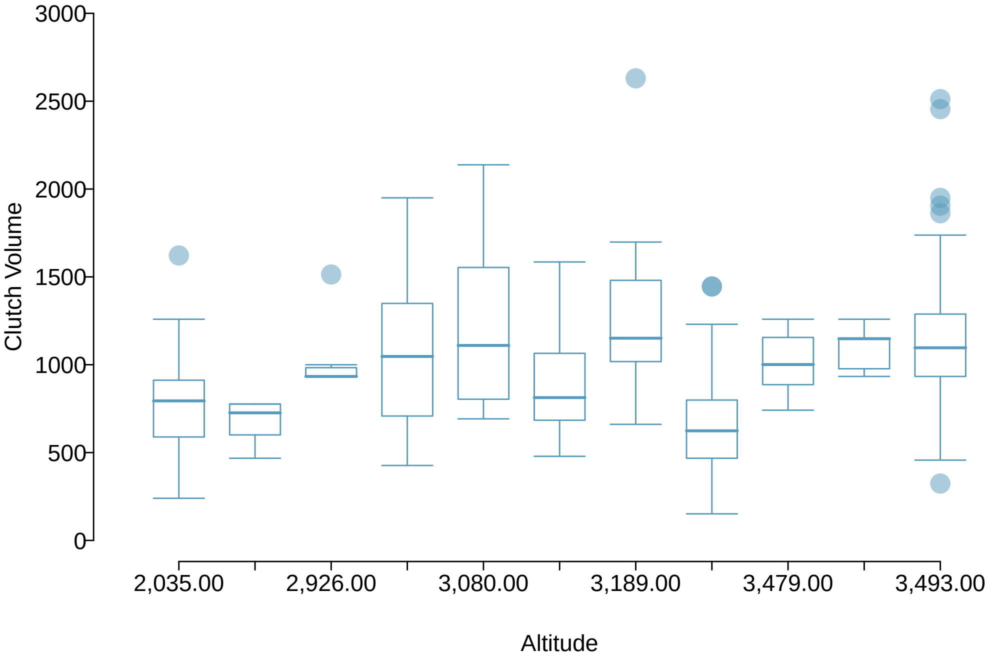


Figure 1.32: Side-by-side boxplot comparing the distribution of `clutch.volume` for different altitudes.

Not all data will show such apparent trends. For example, consider the question of interest in the `frog` dataset: how does maternal investment vary with altitude? Researchers collected data at 11 altitudes from 2,035 to 3,495 m above sea level, measuring attributes of egg clutches such as clutch volume. A side-by-side boxplot comparing clutch volume across altitudes is shown in Figure 1.32. It seems that as a general rule, clutches found at higher altitudes have greater volume. However, more advanced statistical methods, such as those used in the published study, are required to thoroughly investigate the potential association between altitude and clutch size.

1.6 Exercises

JV: Edited on 03Sept2015. Deleted exercises that seemed too easy or covered eliminated topics (ex. mapping data). Notes about replacing exercises visible in pdf, original exercises to be modified in comments. Figure/graphic pathways edited to oi_biostat. Section names updated.

1.6.1 Case study: preventing peanut allergies

1.1 Migraine and acupuncture. A migraine is a particularly painful type of headache, which patients sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a randomized controlled study where 89 females diagnosed with migraine headaches were randomly assigned to one of two groups: treatment or control. 43 patients in the treatment group received acupuncture that is specifically designed to treat migraines. 46 patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. Results are summarized in the contingency table below.[34]

		Pain free		
		Yes	No	Total
Group	Treatment	10	33	43
	Control	2	44	46
	Total	12	77	89

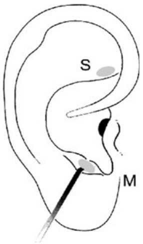


Figure from the original paper displaying the appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks.

(a) What percent of patients in the treatment group were pain free 24 hours after receiving acupuncture? What percent in the control group?

(b) At first glance, does acupuncture appear to be an effective treatment for migraines? Explain your reasoning.

(c) Do the data provide convincing evidence that there is a real pain reduction for those patients in the treatment group? Or do you think that the observed difference might just be due to chance?

1.2 Sinusitis and antibiotics. Researchers studying the effect of antibiotic treatment for acute sinusitis compared to symptomatic treatments randomly assigned 166 adults diagnosed with acute sinusitis to one of two groups: treatment or control. Study participants received either a 10-day course of amoxicillin (an antibiotic) or a placebo similar in appearance and taste. The placebo consisted of symptomatic treatments such as acetaminophen, nasal decongestants, etc. At the end of the 10-day period patients were asked if they experienced significant improvement in symptoms. The distribution of responses are summarized below.[35]

		Self-reported significant improvement in symptoms		
		Yes	No	Total
Group	Treatment	66	19	85
	Control	65	16	81
	Total	131	35	166

(a) What percent of patients in the treatment group experienced a significant improvement in symptoms? What percent in the control group?

(b) Based on your findings in part (a), which treatment appears to be more effective for sinusitis?

[34] **Allais:2011**.

[35] **Garbutt:2012**.

(c) Do the data provide convincing evidence that there is a difference in the improvement rates of sinusitis symptoms? Or do you think that the observed difference might just be due to chance?

1.6.2 Data basics

1.3 Air pollution and birth outcomes, study components. Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.[36] Identify (a) the cases, (b) the variables and their types, and (c) the main research question in this study.

1.4 Buteyko method, study components. The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.[37] Identify (a) the cases, (b) the variables and their types, and (c) the main research question in this study.

JV: Placeholder for two biology-related questions asking for cases, variables, and research question.

1.5 Fisher's irises. Sir Ronald Aylmer Fisher was an English statistician, evolutionary biologist, and geneticist who worked on a dataset that contained sepal length and width, and petal length and width from three species of iris flowers (*setosa*, *versicolor* and *virginica*). There were 50 flowers from each species in the data set.[38]

(a) How many cases were included in the data?

(b) How many numerical variables are included in the data? Indicate what they are, and if they are continuous or discrete.

(c) How many categorical variables are included in the data, and what are they? List the corresponding levels (categories).

Photo by Ryan Claussen
(http://flic.kr/p/6QTcuX)
CC BY-SA 2.0 license

1.6 Smoking habits of UK residents. A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that "£" stands for British Pounds Sterling, "cig" stands for cigarettes, and "N/A" refers to a missing component of the data.[39]

[36] **Ritz+Yu+Chapa+Fruin:2000**.

[37] **McDowan:2003**.

[38] **Fisher:1936**.

[39] **data:smoking**.

	sex	age	marital	grossIncome	smoke	amtWeekends	amtWeekdays
1	Female	42	Single	Under £2,600	Yes	12 cig/day	12 cig/day
2	Male	44	Single	£10,400 to £15,600	No	N/A	N/A
3	Male	53	Married	Above £36,400	Yes	6 cig/day	6 cig/day
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1691	Male	40	Single	£2,600 to £5,200	Yes	8 cig/day	8 cig/day

(a) What does each row of the data matrix represent?

(b) How many participants were included in the survey?

(c) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

JV: Insert exercise showing a data matrix from a bio/genetics study.

JV: Add questions involving identification of explanatory and response variables.

1.6.3 Overview of data collection principles

1.7 Air pollution and birth outcomes, scope of inference. Exercise 1.3 introduces a study where researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

JV: Placeholder for exercise referring back to previous section, replacement for cheaters exercise.

1.8 Buteyko method, scope of inference. Exercise 1.4 introduces a study on using the Buteyko shallow breathing technique to reduce asthma symptoms and improve quality of life. As part of this study 600 asthma patients aged 18-69 who relied on medication for asthma treatment were recruited and randomly assigned to two groups: one practiced the Buteyko method and the other did not. Those in the Buteyko group experienced, on average, a significant reduction in asthma symptoms and an improvement in quality of life.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

JV: Replacement for socioeconomic class + unethical behavior exercise.

JV: Can be used if reworded so that scatterplot graphic is not needed, mostly for the confounding variable question.

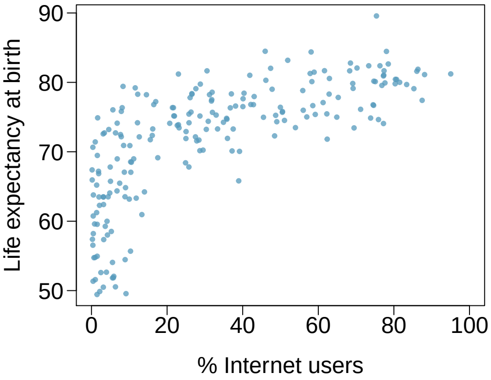
1.9 Internet use and life expectancy. The following scatterplot was created as part of a study evaluating the relationship between estimated life expectancy at birth (as of 2014) and percentage of internet users (as of 2009) in 208 countries for which such data were available.[40]

[40] **data:ciaFactbook**.

(a) Describe the relationship between life expectancy and percentage of internet users.

(b) What type of study is this?

(c) State a possible confounding variable that might explain this relationship and describe its potential effect.



1.10 Stressed out, Part I. A study that surveyed a random sample of otherwise healthy high school students found that they are more likely to get muscle cramps when they are stressed. The study also noted that students drink more coffee and sleep less when they are stressed.

(a) What type of study is this?

(b) Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

(c) State possible confounding variables that might explain the observed relationship between increased stress and muscle cramps.

JV: Modify to have a different context?

1.11 Evaluate sampling methods. A university wants to determine what fraction of its undergraduate student body support a new $25 annual fee to improve the student union. For each proposed method below, indicate whether the method is reasonable or not.

(a) Survey a simple random sample of 500 students.

(b) Stratify students by their field of study, then sample 10% of students from each stratum.

(c) Cluster students by their ages (e.g. 18 years old in one cluster, 19 years old in one cluster, etc.), then randomly sample three clusters and survey all students in those clusters.

JV: Replace with a not-psychology study?

1.12 Haters gonna hate, study confirms. A study published in the *Journal of Personality and Social Psychology* asked a group of 200 randomly sampled men and women to evaluate how they felt about various subjects, such as camping, health care, architecture, taxidermy, crossword puzzles, and Japan in order to measure their dispositional attitude towards mostly independent stimuli. Then, they presented the participants with information about a new product: a microwave oven. This microwave oven does not exist, but the participants didn't know this, and were given three positive and three negative fake reviews. People who reacted positively to the subjects on the dispositional attitude measurement also tended to react positively to the microwave oven, and those who reacted negatively also tended to react negatively to it. Researcher concluded that "some people tend to like things, whereas others tend to dislike things, and a more thorough understanding of this tendency will lead to a more thorough understanding of the psychology of attitudes."[41]

(a) What are the cases?

(b) What is (are) the response variable(s) in this study?

(c) What is (are) the explanatory variable(s) in this study?

(d) Does the study employ random sampling?

(e) Is this an observational study or an experiment? Explain your reasoning.

(f) Can we establish a causal link between the explanatory and response variables?

[41] **Hepler:2013**.

(g) Can the results of the study be generalized to the population at large?

1.13 Family size. Suppose we want to estimate household size, where a “household" is defined as people living together in the same dwelling, and sharing living accommodations. If we select students at random at an elementary school and ask them what their family size is, will this be a good measure of household size? Or will our average be biased? If so, will it overestimate or underestimate the true value?

JV: I like these.

1.14 Flawed reasoning. Identify the flaw(s) in reasoning in the following scenarios. Explain what the individuals in the study should have done differently if they wanted to make such strong conclusions.

(a) Students at an elementary school are given a questionnaire that they are asked to return after their parents have completed it. One of the questions asked is, “Do you find that your work schedule makes it difficult for you to spend time with your kids after school?" Of the parents who replied, 85% said “no". Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.

(b) A survey is conducted on a simple random sample of 1,000 women who recently gave birth, asking them about whether or not they smoked during pregnancy. A follow-up survey asking if the children have respiratory problems is conducted 3 years later, however, only 567 of these women are reached at the same address. The researcher reports that these 567 women are representative of all mothers.

(c) A orthopedist administers a questionnaire to 30 of his patients who do not have any joint problems and finds that 20 of them regularly go running. He concludes that running decreases the risk of joint problems.

1.15 City council survey. A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. Identify the sampling methods described below, and comment on whether or not you think they would be effective in this setting.

(a) Randomly sample 50 households from the city.

(b) Divide the city into neighborhoods, and sample 20 households from each neighborhood.

(c) Divide the city into neighborhoods, randomly sample 10 neighborhoods, and sample all households from those neighborhoods.

(d) Divide the city into neighborhoods, randomly sample 10 neighborhoods, and then randomly sample 20 households from those neighborhoods.

(e) Sample the 200 households closest to the city council offices.

JV: Replace w/bio one.

1.16 Reading the paper. Below are excerpts from two articles published in the *NY Times*:

(a) An article titled *Risks: Smokers Found More Prone to Dementia* states the following:[42]

> “Researchers analyzed data from 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50-60 years old. 23 years later, about 25% of the group had dementia, including 1,136 with Alzheimer’s disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up with increased smoking; 44% for one to two packs a day; and twice the risk for more than two packs."

Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.

(b) Another article titled *The School Bully Is Sleepy* states the following:[43]

[42] **news:smokingDementia**.

[43] **news:bullySleep**.

> "The University of Michigan study, collected survey data from parents on each child's sleep habits and asked both parents and teachers to assess behavioral concerns. About a third of the students studied were identified by parents or teachers as having problems with disruptive behavior or bullying. The researchers found that children who had behavioral issues and those who were identified as bullies were twice as likely to have shown symptoms of sleep disorders."

A friend of yours who read the article says, "The study shows that sleep disorders lead to bullying in school children." Is this statement justified? If not, how best can you describe the conclusion that can be drawn from this study?

JV: Merge this one with Part I.

1.17 Stressed out, Part II. In a study evaluating the relationship between stress and muscle cramps half the subjects are randomly assigned to be exposed to increased stressed by being placed into an elevator that falls rapidly and stops abruptly and the other half are left at no or baseline stress.

(a) What type of study is this?

(b) Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

1.18 Vitamin supplements. In order to assess the effectiveness of taking large doses of vitamin C in reducing the duration of the common cold, researchers recruited 400 healthy volunteers from staff and students at a university. A quarter of the patients were assigned a placebo, and the rest were evenly divided between 1g Vitamin C, 3g Vitamin C, or 3g Vitamin C plus additives to be taken at onset of a cold for the following two days. All tablets had identical appearance and packaging. The nurses who handed the prescribed pills to the patients knew which patient received which treatment, but the researchers assessing the patients when they were sick did not. No significant differences were observed in any measure of cold duration or severity between the four medication groups, and the placebo group had the shortest duration of symptoms.[44]

(a) Was this an experiment or an observational study? Why?

(b) What are the explanatory and response variables in this study?

(c) Were the patients blinded to their treatment?

(d) Was this study double-blind?

(e) Participants are ultimately able to choose whether or not to use the pills prescribed to them. We might expect that not all of them will adhere and take their pills. Does this introduce a confounding variable to the study? Explain your reasoning.

JV: Include more realistic open-ended study design questions? Or eliminate entirely?

1.19 Exercise and mental health. A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41- 55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.

(a) What type of study is this?

(b) What are the treatment and control groups in this study?

(c) Does this study make use of blocking? If so, what is the blocking variable?

(d) Does this study make use of blinding?

(e) Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.

(f) Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?

JV: Replace with biology question in similar style.

[44] **Audera:2001**.

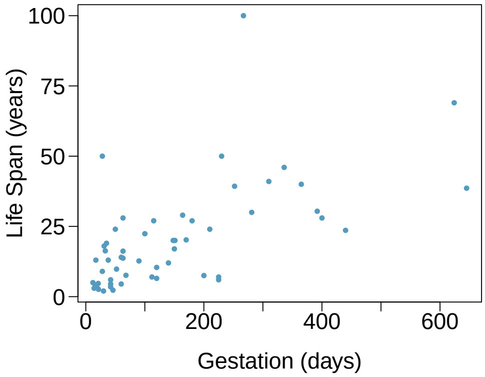
1.6.4 Examining numerical data

1.20 Mammal life spans. Data were collected on life spans (in years) and gestation lengths (in days) for 62 mammals. A scatterplot of life span versus length of gestation is shown below.[45]

(a) What type of an association is apparent between life span and length of gestation?

(b) What type of an association would you expect to see if the axes of the plot were reversed, i.e. if we plotted length of gestation versus life span?

(c) Are life span and length of gestation independent? Explain your reasoning.



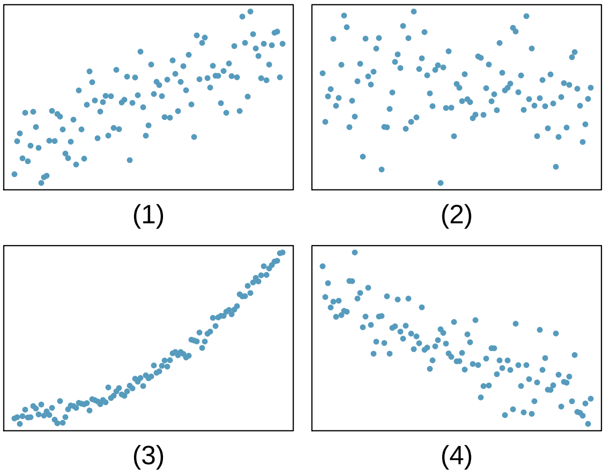
1.21 Associations. Indicate which of the plots show a

(a) positive association

(b) negative association

(c) no association

Also determine if the positive and negative associations are linear or nonlinear. Each part may refer to more than one plot.



1.22 Reproducing bacteria. Suppose that there is only sufficient space and nutrients to support one million bacterial cells in a petri dish. You place a few bacterial cells in this petri dish, allow them to reproduce freely, and record the number of bacterial cells in the dish over time. Sketch a plot representing the relationship between number of bacterial cells and time.

JV: Re-write for different contexts – identifying sample vs. population means

1.23 Sleeping in college. A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night. Identify which value represents the sample mean and which value represents the claimed population mean.

1.24 Parameters and statistics. Identify which value represents the sample mean and which value represents the claimed population mean.

(a) American households spent an average of about $52 in 2007 on Halloween merchandise such as costumes, decorations and candy. To see if this number had changed, researchers conducted a

[45] **Allison+Cicchetti:1975**.

new survey in 2008 before industry numbers were reported. The survey included 1,500 households and found that average Halloween spending was $58 per household.

(b) The average GPA of students in 2001 at a private university was 3.37. A survey on a sample of 203 students from this university yielded an average GPA of 3.59 in Spring semester of 2012.

1.25 Medians and IQRs. For each part, compare distributions (1) and (2) based on their medians and IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make sure to explain your reasoning.

(a) (1) 3, 5, 6, 7, 9
(2) 3, 5, 6, 7, 20

(b) (1) 3, 5, 6, 7, 9
(2) 3, 5, 8, 7, 9

(c) (1) 1, 2, 3, 4, 5
(2) 6, 7, 8, 9, 10

(d) (1) 0, 10, 50, 60, 100
(2) 0, 100, 500, 600, 1000

1.26 Means and SDs. For each part, compare distributions (1) and (2) based on their means and standard deviations. You do not need to calculate these statistics; simply state how the means and the standard deviations compare. Make sure to explain your reasoning. *Hint:* It may be useful to sketch dot plots of the distributions.

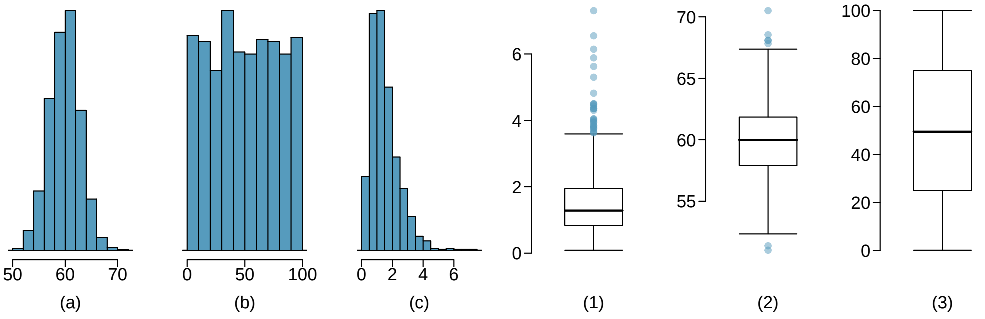
(a) (1) 3, 5, 5, 5, 8, 11, 11, 11, 13
(2) 3, 5, 5, 5, 8, 11, 11, 11, 20

(b) (1) -20, 0, 0, 0, 15, 25, 30, 30
(2) -40, 0, 0, 0, 15, 25, 30, 30

(c) (1) 0, 2, 4, 6, 8, 10
(2) 20, 22, 24, 26, 28, 30

(d) (1) 100, 200, 300, 400, 500
(2) 0, 50, 300, 550, 600

1.27 Mix-and-match. Describe the distribution in the histograms below and match them to the box plots.



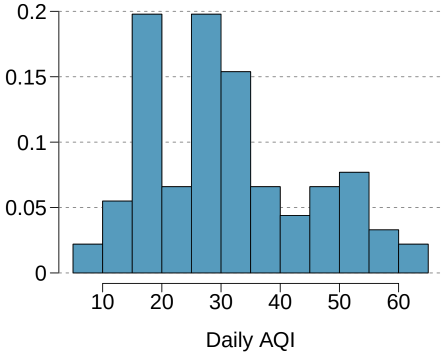
1.28 Air quality. Daily air quality is measured by the air quality index (AQI) reported by the Environmental Protection Agency. This index reports the pollution level and what associated health effects might be a concern. The index is calculated for five major air pollutants regulated by the Clean Air Act and takes values from 0 to 300, where a higher value indicates lower air quality. AQI was reported for a sample of 91 days in 2011 in Durham, NC. The relative frequency histogram below shows the distribution of the AQI values on these days.[46]

(a) Estimate the median AQI value of this sample.

(b) Would you expect the mean AQI value of this sample to be higher or lower than the median? Explain your reasoning.

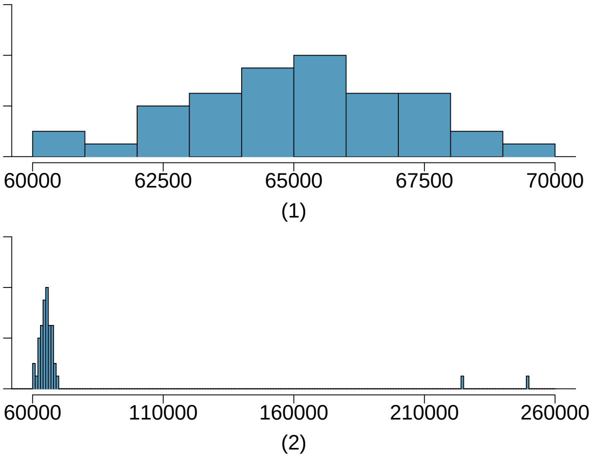
(c) Estimate Q1, Q3, and IQR for the distribution.

(d) Would any of the days in this sample be considered to have an unusually low or high AQI? Explain your reasoning.



JV: Do histogram vs. boxplot exercises for real data, incl. biological example.
JV: Adapt for more realistic examples of skew.
JV: Nice setup for robustness exercise. Adapt context?

1.29 Income at the coffee shop. The first histogram below shows the distribution of the yearly incomes of 40 patrons at a college coffee shop. Suppose two new people walk into the coffee shop: one making $225,000 and the other $250,000. The second histogram shows the new income distribution. Summary statistics are also provided.



	(1)	(2)
n	40	42
Min.	60,680	60,680
1st Qu.	63,620	63,710
Median	65,240	65,350
Mean	65,090	73,300
3rd Qu.	66,160	66,540
Max.	69,890	250,000
SD	2,122	37,321

(a) Would the mean or the median best represent what we might think of as a typical income for the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

(b) Would the standard deviation or the IQR best represent the amount of variability in the incomes of the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

1.30 Midrange. The *midrange* of a distribution is defined as the average of the maximum and the

[46] **data:durhamAQI:2011**.

minimum of that distribution. Is this statistic robust to outliers and extreme skew? Explain your reasoning

JV: Need a data transformation exercise or two.

1.6.5 Considering categorical data

JV: Use different data.

1.31 Views on immigration. 910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.[47]

		Political ideology			
		Conservative	Moderate	Liberal	Total
Response	(i) Apply for citizenship	57	120	101	278
	(ii) Guest worker	121	113	28	262
	(iii) Leave the country	179	126	45	350
	(iv) Not sure	15	4	1	20
	Total	372	363	175	910

(a) What percent of these Tampa, FL voters identify themselves as conservatives?

(b) What percent of these Tampa, FL voters are in favor of the citizenship option?

(c) What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

(d) What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

(e) Do political ideology and views on immigration appear to be independent? Explain your reasoning.

JV: Another contingency table exercise, as well as something with numerical data across groups.

[47] **survey:immigFL:2012**.

Chapter 2

Probability

What are the chances that a woman with an abnormal mammogram has breast cancer? What is the likelihood that an overweight male teenager with high blood pressure will develop cardiovascular disease by the age of 50? What is the probability that two parents who are unaffected carriers of a genetic mutation that causes cystic fibrosis will have a child that suffers from the disease. All of these questions use the language of probability, and despite how easy it is to ask these questions, answers are not always easy to come by. Probability also forms the foundation for data analysis and statistical inference, since nearly every conclusion to a study should be accompanied by a measure of uncertainty. In the publication of LEAP study discussed in Chapter 1, the manuscript included the probability that the results of the study could have been due simply to chance variation (a very small probability, as will be seen later in the text).

Like all mathematical tools, probability becomes easier to understand and work with when the important concepts and language have been formalized. With the right tools, seemingly difficult problems can be solved in a series of reliable, reproducible steps. This chapter introduces that formalization, using two types of examples. One set of examples uses familiar terms using settings most people have seen before – the outcomes of rolling dice or picking cards from a deck of playing cards. The second type of examples are drawn from medicine, biology or public health, and reflect the context and language used in those fields. The approaches to solving both types of problems are surprisingly similar, once the problem has been posed clearly.

2.1 Defining probability

2.1.1 Some examples

Some of these dice examples can be dropped, but leaving them for now in case they are reference later.

We begin with some familiar examples.

● **Example 2.1** A "die", the singular of dice, is a cube with six faces numbered 1, 2, 3, 4, 5, and 6. What is the chance of getting 1 when rolling a die?

If the die is fair, then the chance of a 1 is as good as the chance of any other number. Since there are six outcomes, the chance must be 1-in-6 or, equivalently, 1/6.

Example 2.2 What is the chance of getting a 1 or 2 in the next roll?

1 and 2 constitute two of the six equally likely possible outcomes, so the chance of getting one of these two outcomes must be $2/6 = 1/3$.

Example 2.3 What is the chance of getting either 1, 2, 3, 4, 5, or 6 on the next roll?

100%. The outcome must be one of these numbers.

Example 2.4 What is the chance of not rolling a 2?

Since the chance of rolling a 2 is 1/6 or $16.\bar{6}\%$, the chance of not rolling a 2 must be $100\% - 16.\bar{6}\% = 83.\bar{3}\%$ or 5/6.

Alternatively, we could have noticed that not rolling a 2 is the same as getting a 1, 3, 4, 5, or 6, which makes up five of the six equally likely outcomes and has probability 5/6.

Example 2.5 Consider rolling two dice. If $1/6^{th}$ of the time the first die is a 1 and $1/6^{th}$ of those times the second die is a 1, what is the chance of getting two 1s?

If $16.\bar{6}\%$ of the time the first die is a 1 and $1/6^{th}$ of *those* times the second die is also a 1, then the chance that both dice are 1 is $(1/6) \times (1/6)$ or 1/36.

Here is an example from genetics.

Example 2.6 Cystic fibrosis (CF) is a life-threatening genetic disorder characterized by the buildup of thick mucus in the lungs and pancreas, caused by mutations in the *CFTR* gene located on chromosome 7. Defective copies of *CFTR* can result in the reduced quantity and/or function of the CFTR protein, which transports sodium and chloride across cell membranes. CF is an autosomal recessive disorder – an individual only develops CF if they have inherited two affected copies of *CFTR*. Individuals with one normal (wild-type) copy and one defective (mutated) copy are known as carriers; they do not develop CF, but may pass the disease-causing mutation onto their offspring.

Suppose that both members of a couple are CF carriers. What is the probability that a child of this couple will be affected by CF? The problem sounds a bit more complicated than calculating probabilities for the outcome of rolling a die, but can be solved with the same simple methods. We show two solutions.

Solution 1: Enumerate all of the possible outcomes and exploit the fact that the outcomes are equally likely, as in **??**. During reproduction, each parent passes along one copy of the *CFTR* gene, with each copy passed along with probability 1/2. Figure 2.1 shows the four possible genotypes for a child of these parents, with the paternal chromosome in blue, the maternal chromosome in green, chromosomes with the wild-type and mutated version of CFTR marked with + and−. Each of the four outcomes (wild-type CFTR, wild-type CFTR), (wild-type CFTR, CFTR mutation) (CFTR mutation, wild-type CFTR) and (CFTR mutation, CFTR mutation), so the child will be affected with probability 1/4. It is important to recognize that the child being an unaffected carrier consists of two distinct outcomes, not one.

Solution 2: Calculate the proportion of outcomes that produce an affected child, as in 2.1. During reproduction, half of the time, the mother will pass along an affected gene.

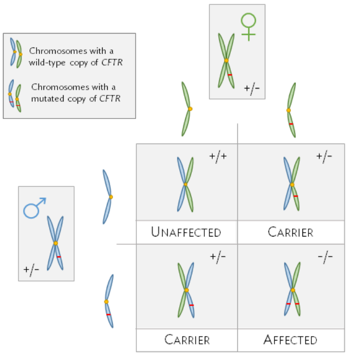

Chromosomes with a wild-type copy of *CFTR*
Chromosomes with a mutated copy of *CFTR*
♀
+/-
♂
+/-
+/+
UNAFFECTED
+/-
CARRIER
+/-
CARRIER
-/-
AFFECTED

Figure 2.1: Pattern of inheritance of a child of two unaffected carriers of CFTR

When the child receives an affected gene from the mother, about half of those times, the father will have passed along an affected gene. So the proportion of times the child will be affected is (1/2) × (1/2) = 1/4.

Guided Practice 2.7 Suppose the father is affected by CF and the mother is an unaffected carrier. What is the probability that their child will be affected by the disease?

Solution: Since the father is affected, he will always pass along a defective copy of the gene. Since the mother will pass along a defective copy half of the time, the child will be affected half of the time, or with probability 1/4.

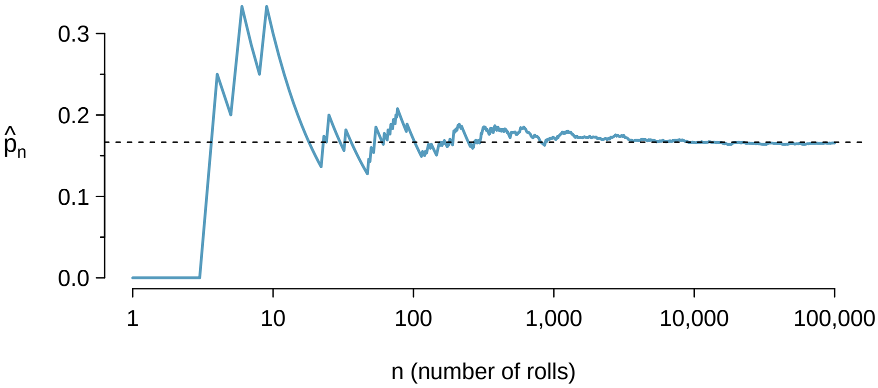


Figure 2.2: The fraction of die rolls that are 1 at each stage in a simulation. The proportion tends to get closer to the probability 1/6 ≈ 0.167 as the number of rolls increases.

2.1.2 Probability

Probability is used to assign a level of uncertainty to outcomes of phenomena that are happen randomly (rolling dice, passing along a defective gene during reproduction), or appear random because of a lack of understanding about exactly how the phenomenon occurs (an obese teenager with high blood pressure developing cardiovascular disease later in life).In either case, the interpretation is the same – the chance that some event will happen in the future – and modeling these complex phenomena as random can be useful.

Mathematicians and philosophers have struggled for centuries (literally) to arrive at a clear statement of how probability is defined, or what it means. In this text we use the most common definition, which also has the clearest interpretation.

Probability

The **probability** of an outcome is the proportion of times the outcome would occur if the random phenomenon could be observed an infinite number of times.

Probability is defined as a proportion, and it always takes values between 0 and 1 (inclusively). It may also be displayed as a percentage between 0% and 100%.

It is easy to imagine rolling dice a large number of times to observe the law of large numbers, but for examples like the CF example, the interpretation of probability is more hypothetical, since family sizes are typically small. But it is not too difficult to imagine a thought experiment in which two parents have many children. If the two parents are unaffected carriers, approximately 25% of their off spring will suffer from CF.

This definition of probability can be illustrated by rolling a die many times. Let $\hat{p}_n$ be the proportion of outcomes that are 1 after the first n rolls. As the number of rolls increases, $\hat{p}_n$ will converge to the probability of rolling a 1, $p = 1/6$. Figure **??** shows this convergence for 100,000 die rolls. The tendency of $\hat{p}_n$ to stabilize around p is described by the **Law of Large Numbers**.

The behavior shown in 2.2 matches most people's intuition about probability, but proving mathematically that the behavior is always true is surprisingly difficult and is beyond the level of this text. Mathematicians call the result *The Law of Large Numbers*, which is used to justify mathematically this intuitively appealing definition.

Law of Large Numbers

As more observations are collected, the proportion $\hat{p}_n$ of occurrences with a particular outcome converges to the probability p of that outcome.

Occasionally the proportion will veer off from the probability and appear to defy the Law of Large Numbers, as $\hat{p}_n$ does many times in Figure **??**. However, these deviations become smaller as the number of rolls increases.

The notation p is the probability of rolling a 1. We can also write this probability as

$P(A)$
Probability of outcome A

$$P(\text{rolling a 1})$$

As we become more comfortable with this notation, we will abbreviate it further. For instance, if it is clear that the process is "rolling a die", we could abbreviate $P(\text{rolling a 1})$ as $P(1)$. We also have a notation for an event itself, so the event A of rolling a 1 will be written as $A = \{\text{ rolling a 1}\}$, with associated probability $P(A)$.

2.1.3 Disjoint or mutually exclusive outcomes

Two outcomes are called **disjoint** or **mutually exclusive** if they cannot both happen. When rolling a die, the outcomes 1 and 2 are disjoint since they cannot both occur. In the cystic fibrosis example, the two outcomes of a wild-type gene from the mother and a mutated gene from the father and a mutated gene from the mother, wild-type from the father are disjoint. In the die example, the outcomes 1 and "rolling an odd number" are not disjoint since both occur if the outcome of the roll is a 1. The outcomes of a child being affected and having at least one mutated copy of CFTR and not disjoint. The terms *disjoint* and *mutually exclusive* are equivalent and interchangeable.

Calculating the probability of disjoint outcomes is easy. When rolling a die, the outcomes 1 and 2 are disjoint, and we compute the probability that one of these outcomes will occur by adding their separate probabilities:

$$P(1 \text{ or } 2) = P(1) + P(2) = 1/6 + 1/6 = 1/3$$

What about the probability of rolling a 1, 2, 3, 4, 5, or 6? Here again, all of the outcomes are disjoint so we add the probabilities:

$$\begin{aligned}
&P(1 \text{ or } 2 \text{ or } 3 \text{ or } 4 \text{ or } 5 \text{ or } 6)\\
&\quad = P(1) + P(2) + P(3) + P(4) + P(5) + P(6)\\
&\quad = 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1.
\end{aligned}$$

The probability that a child will be an unaffected carrier in the CF example is $(1/2) = (1/2) = 1/4$.

The **Addition Rule** guarantees the accuracy of this approach when the outcomes are disjoint.

Addition Rule of disjoint outcomes

If A_1 and A_2 represent two disjoint outcomes, then the probability that one of them occurs is given by

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2)$$

If there are many disjoint outcomes A_1, ..., A_k, then the probability that one of these outcomes will occur is

$$P(A_1) + P(A_2) + \cdots + P(A_k) \qquad (2.8)$$

Probability problems rarely consider individual outcomes and instead consider *sets* or *collections* of outcomes. Let A represent the event where a die roll results in 1 or 2 and B represent the event that the die roll is a 4 or a 6. We write A as the set of outcomes $\{1, 2\}$ and $B = \{4, 6\}$. These sets are commonly called **events**. Because A and B have no elements in common, they are disjoint events. A and B are represented in Figure 2.3.

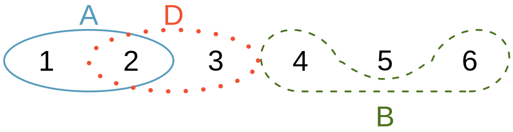


Figure 2.3: Three events, A, B, and D, consist of outcomes from rolling a die. A and B are disjoint since they do not have any outcomes in common.

The Addition Rule applies to both disjoint outcomes and disjoint events. The probability that one of the disjoint events A or B occurs is the sum of the separate probabilities:

$$P(A \text{ or } B) = P(A) + P(B) = 1/3 + 1/3 = 2/3$$

⊙ **Guided Practice 2.9** (a) Verify the probability of event A, $P(A)$, is 1/3 using the Addition Rule. (b) Do the same for event B.[1]

⊙ **Guided Practice 2.10** (a) Using Figure 2.3 as a reference, what outcomes are represented by event D? (b) Are events B and D disjoint? (c) Are events A and D disjoint?[2]

⊙ **Guided Practice 2.11** In Guided Practice 2.10, you confirmed B and D from Figure 2.3 are disjoint. Compute the probability that event B or event D occurs.[3]

should we add more genetics problems here? I have removed the email example because of the possible confusion between events involving sampling from a population vs a study sample. If we think we can make that clear, we can use examples from famuss, perhaps by posing a problem of sampling members from the study participants. Note also that this is moving more slowly than the Stat 102 notes, but we did show some of this material on the blackboard. If we use this chapter in 102, perhaps we can move quickly to more complicated examples.

[1] (a) $P(A) = P(1 \text{ or } 2) = P(1) + P(2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$. (b) Similarly, $P(B) = 1/3$.

[2] (a) Outcomes 2 and 3. (b) Yes, events B and D are disjoint because they share no outcomes. (c) The events A and D share an outcome in common, 2, and so are not disjoint.

[3] Since B and D are disjoint events, use the Addition Rule: $P(B \text{ or } D) = P(B) + P(D) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$.

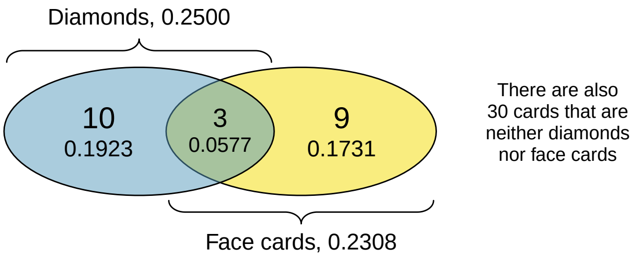


Figure 2.5: A Venn diagram for diamonds and face cards.

2.1.4 Probabilities when events are not disjoint

Let's consider calculations for two events that are not disjoint in the context of a regular deck of 52 cards, represented in Table 2.4. If you are unfamiliar with the cards in a regular deck, please see the footnote.[4]

2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣	A♣
2♢	3♢	4♢	5♢	6♢	7♢	8♢	9♢	10♢	J♢	Q♢	K♢	A♢
2♡	3♡	4♡	5♡	6♡	7♡	8♡	9♡	10♡	J♡	Q♡	K♡	A♡
2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠	A♠

Table 2.4: Representations of the 52 unique cards in a deck.

Guided Practice 2.12 (a) What is the probability that a randomly selected card is a diamond? (b) What is the probability that a randomly selected card is a face card?[5]

Venn diagrams are useful when outcomes can be categorized as "in" or "out" for two or three variables, attributes, or random processes. The Venn diagram in Figure 2.5 uses a circle to represent diamonds and another to represent face cards. If a card is both a diamond and a face card, it falls into the intersection of the circles. If it is a diamond but not a face card, it will be in part of the left circle that is not in the right circle (and so on). The total number of cards that are diamonds is given by the total number of cards in the diamonds circle: $10 + 3 = 13$. The probabilities are also shown (e.g. $10/52 = 0.1923$).

Let A represent the event that a randomly selected card is a diamond and B represent the event that it is a face card. How do we compute $P(A \text{ or } B)$? Events A and B are not disjoint – the cards $J\diamondsuit$, $Q\diamondsuit$, and $K\diamondsuit$ fall into both categories – so we cannot use the Addition Rule for disjoint events. Instead we use the Venn diagram. We start by adding the probabilities of the two events:

$$P(A) + P(B) = P(\diamondsuit) + P(\text{face card}) = 12/52 + 13/52$$

[4] The 52 cards are split into four **suits**: ♣ (club), ♢ (diamond), ♡ (heart), ♠ (spade). Each suit has its 13 cards labeled: 2, 3, ..., 10, J (jack), Q (queen), K (king), and A (ace). Thus, each card is a unique combination of a suit and a label, e.g. 4♡ and J♣. The 12 cards represented by the jacks, queens, and kings are called `face cards`. The cards that are ♢ or ♡ are typically colored red while the other two suits are typically colored black.

[5] (a) There are 52 cards and 13 diamonds. If the cards are thoroughly shuffled, each card has an equal chance of being drawn, so the probability that a randomly selected card is a diamond is $P(\diamondsuit) = \frac{13}{52} = 0.250$. (b) Likewise, there are 12 face cards, so $P(\text{face card}) = \frac{12}{52} = \frac{3}{13} = 0.231$.

However, the three cards that are in both events were counted twice, once in each probability. We must correct this double counting:

$$\begin{aligned} P(A \text{ or } B) &= P(\text{face card or } \diamondsuit) \\ &= P(\text{face card}) + P(\diamondsuit) - P(\text{face card and } \diamondsuit) \qquad (2.13) \\ &= 13/52 + 12/52 - 3/52 \\ &= 22/52 = 11/26 \end{aligned}$$

Equation (2.13) is an example of the **General Addition Rule**.

General Addition Rule

If A and B are any two events, disjoint or not, then the probability that at least one of them will occur is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \qquad (2.14)$$

where $P(A \text{ and } B)$ is the probability that both events occur.

TIP: "or" is inclusive

When we write "or" in statistics, we mean "and/or" unless we explicitly state otherwise. Thus, A or B occurs means A, B, or both A and B occur.

Guided Practice 2.15 If A and B are disjoint, describe why this implies $P(A$ and $B) = 0$. (b) Using part (a), verify that the General Addition Rule simplifies to the simpler Addition Rule for disjoint events if A and B are disjoint.[6]

Guided Practice 2.16

In areas of the developing world, the human immunodeficiency virus (HIV) and tuberculosis (TB) are infectious diseases that affect substantial proportions of the population. Individuals sometimes have both diseases (are co-infected); children of HIV-infected mothers may have HIV (be HIV$^+$) and TB can spread from one family member to another. In a mother child pair, let $A = \{$ the mother has HIV $\}$, $B = \{$ the mother has TB $\}$, $C = \{$ the child has HIV $\}$, $D = \{$ the child has TB $\}$. Write out the definitions of the events A or B, A and B, A and C, A or D.

2.1.5 Probability distributions

A **probability distribution** is a table of all disjoint outcomes and their associated probabilities. Table 2.6 shows the probability distribution for the sum of two dice.

[6] (a) If A and B are disjoint, A and B can never occur simultaneously. (b) If A and B are disjoint, then the last term of Equation (2.14) is 0 (see part (a)) and we are left with the Addition Rule for disjoint events.

Dice sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Table 2.6: Probability distribution for the sum of two dice.

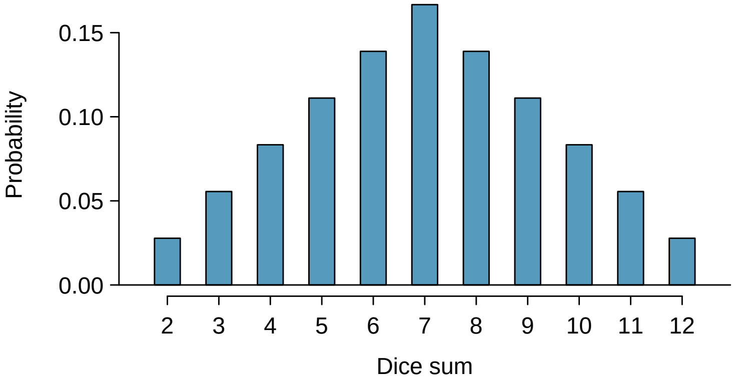


Figure 2.7: The probability distribution of the sum of two dice.

Rules for probability distributions

A probability distribution is a list of the possible outcomes with corresponding probabilities that satisfies three rules:

1. The outcomes listed must be disjoint.
2. Each probability must be between 0 and 1.
3. The probabilities must total 1.

Chapter 1 emphasized the importance of plotting data to provide quick summaries. Probability distributions can also be summarized in a bar plot. The probability distribution for the sum of two dice is shown in Table 2.6 and plotted in Figure 2.7.

In this bar plots, the bar heights represent the probabilities of outcomes. If the outcomes are numerical and discrete, it is usually (visually) convenient to make a bar plot that resembles a histogram, as in the case of the sum of two dice.

A graph of probability distribution can convey important information about a distribution quickly.

The distribution of birth weights for 3,999,386 live births in the United States in 2010 is shown in figure 2.8. The data are available as part of the US CDC National Vital Statistics System [7]. The graph of the distribution shows that most babies born weighed between 2000 and 5000 grams (2kg to 5 kg), but there were both small (less than 1000 grams) and large (greater than 5000 grams) babies. Pediatricians think of normal birthweight as between 2.5 and 5 kg.

[7] http://205.207.175.93/vitalstats/ReportFolders/reportFolders.aspx

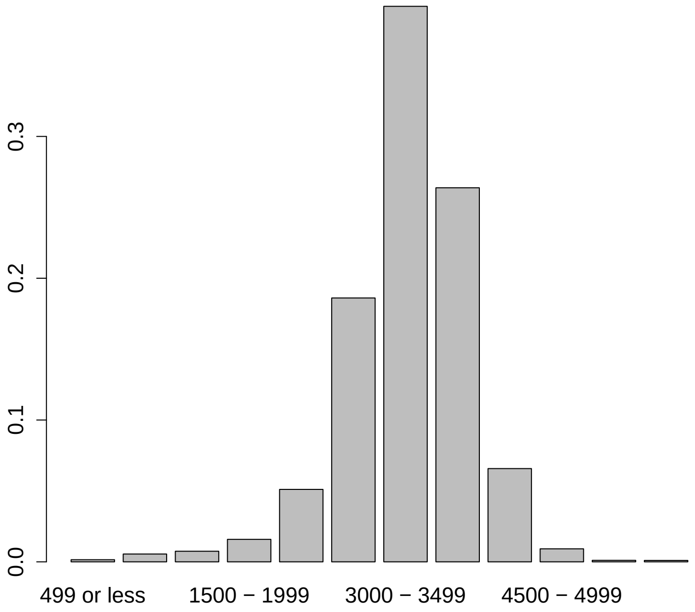


Figure 2.8: Distribution of birth weights (in grams) of babies born in the US in 2010

2.1.6 Complement of an event

S
Sample space

Rolling a die produces a value in the set {1, 2, 3, 4, 5, 6}. This set of all possible outcomes is called the **sample space** (S) for rolling a die. We often use the sample space to examine the scenario where an event does not occur.

A^c
Complement of outcome A

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is 2 or 3. Then the **complement** of D represents all outcomes in our sample space that are not in D, which is denoted by $D^c = \{1, 4, 5, 6\}$. That is, D^c is the set of all possible outcomes not already included in D. Figure 2.9 shows the relationship between D, D^c, and the sample space S.

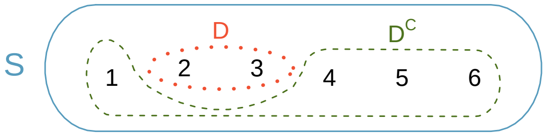


Figure 2.9: Event $D = \{2, 3\}$ and its complement, $D^c = \{1, 4, 5, 6\}$. S represents the sample space, which is the set of all possible events.

⊙ **Guided Practice 2.17** (a) Compute $P(D^c) = P(\text{rolling a 1, 4, 5, or 6})$. (b) What is $P(D) + P(D^c)$?[8]

⊙ **Guided Practice 2.18** Events $A = \{1, 2\}$ and $B = \{4, 6\}$ are shown in Figure 2.3 on page 63. (a) Write out what A^c and B^c represent. (b) Compute $P(A^c)$ and $P(B^c)$. (c) Compute $P(A) + P(A^c)$ and $P(B) + P(B^c)$.[9]

A complement of an event A is constructed to have two very important properties: (i) every possible outcome not in A is in A^c, and (ii) A and A^c are disjoint. Property (i) implies

$$P(A \text{ or } A^c) = 1 \qquad (2.19)$$

That is, if the outcome is not in A, it must be represented in A^c. We use the Addition Rule for disjoint events to apply Property (ii):

$$P(A \text{ or } A^c) = P(A) + P(A^c) \qquad (2.20)$$

Combining Equations (2.19) and (2.20) yields a very useful relationship between the probability of an event and its complement.

Complement

The complement of event A is denoted A^c, and A^c represents all outcomes not in A. A and A^c are mathematically related:

$$P(A) + P(A^c) = 1, \quad \text{i.e.} \quad P(A) = 1 - P(A^c) \qquad (2.21)$$

[8] (a) The outcomes are disjoint and each has probability 1/6, so the total probability is 4/6 = 2/3. (b) We can also see that $P(D) = \frac{1}{6} + \frac{1}{6} = 1/3$. Since D and D^c are disjoint, $P(D) + P(D^c) = 1$.

[9] Brief solutions: (a) $A^c = \{3, 4, 5, 6\}$ and $B^c = \{1, 2, 3, 5\}$. (b) Noting that each outcome is disjoint, add the individual outcome probabilities to get $P(A^c) = 2/3$ and $P(B^c) = 2/3$. (c) A and A^c are disjoint, and the same is true of B and B^c. Therefore, $P(A) + P(A^c) = 1$ and $P(B) + P(B^c) = 1$.

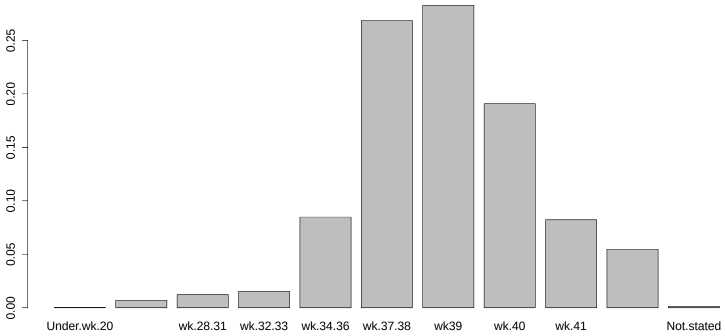


Figure 2.10: Distribution of gestational age for live births in the US in 2010, measured in weeks

In simple examples, computing A or A^c is feasible in a few steps. However, using the complement can save a lot of time as problems grow in complexity.

Guided Practice 2.22 Let A represent the event where we roll two dice and their total is less than 12. (a) What does the event A^c represent? (b) Determine $P(A^c)$ from Table 2.6 on page 66. (c) Determine $P(A)$.[10]

Guided Practice 2.23 Consider again the probabilities from Table 2.6 and rolling two dice. Find the following probabilities: (a) The sum of the dice is *not* 6. (b) The sum is at least 4. That is, determine the probability of the event $B = \{4, 5, ..., 12\}$. (c) The sum is no more than 10. That is, determine the probability of the event $D = \{2, 3, ..., 10\}$.[11]

Sometimes, information from a graph can be combined with using the complement of an event to calculate approximate probabilities. The gestational age of a newborn is the time between conception and birth. Because of the obvious difficulty of determining the exact date of conception, gestational ages are typically recorded in weeks. Figure 2.10 is the graphical representation of the distribution of gestational ages for the 3,999,386 babies born in 2010. Babies born between 38 and 42 weeks of gestational age is considered normal, but the term 'full term' is used for births between 39 and 40 weeks gestational age. The graph shows that approximately 30% of births occur at 39 weeks, and slightly less than 30% occur at 40 weeks, so approximately 50% of babies are considered full term. Instead of adding up the heights of the bars for gestational ages outside the full term range, using the complement of the event of a full term birth, it is clear that approximately 50% of births not considered full term.

The distribution of gestational age is shown in tabular form in Table ??. The table shows the exact value of the proportion of babies born at 39 or 40 weeks (0.47), but when

[10] (a) The complement of A: when the total is equal to 12. (b) $P(A^c) = 1/36$. (c) Use the probability of the complement from part (b), $P(A^c) = 1/36$, and Equation (2.21): $P(\text{less than } 12) = 1 - P(12) = 1 - 1/36 = 35/36$.

[11] (a) First find $P(6) = 5/36$, then use the complement: $P(\text{not } 6) = 1 - P(6) = 31/36$.

(b) First find the complement, which requires much less effort: $P(2 \text{ or } 3) = 1/36 + 2/36 = 1/12$. Then calculate $P(B) = 1 - P(B^c) = 1 - 1/12 = 11/12$.

(c) As before, finding the complement is the more direct way to determine $P(D)$. First find $P(D^c) = P(11 \text{ or } 12) = 2/36 + 1/36 = 1/12$. Then calculate $P(D) = 1 - P(D^c) = 11/12$.

examining the important features of a distribution, approximate values are often sufficient. In some instances, the graph is all that will be available. Since small probabilities are difficult to read accurately from the graph of a distribution, they are best read from the table. Pre-term babies are those born at less than 37 weeks gestational age. Table ?? shows that the probability of this event is $0.01 + 0.01 + 0.02 + 0.08 = 0.12$. Of course, even the table shows approximate values, since the small proportion of very premature babies born at less than 20 weeks is rounded to zero.

two problems with the example: it is too clumsy for what it accomplishes, and I have included the not stated category in the calculations of the proportions. This is negligible, but wrong.

	x
Under.wk.20	0.00
wk.20.27	0.01
wk.28.31	0.01
wk.32.33	0.02
wk.34.36	0.08
wk.37.38	0.27
wk39	0.28
wk.40	0.19
wk.41	0.08
wk.42.and.over	0.05
Not.stated	0.00

value labels should be modified

not satisfied with the way this example worked out; it should be improved or changed

2.1.7 Independence

Just as variables and observations can be independent, random phenomena can be independent, too. Two phenomena or processes are **independent** if knowing the outcome of one provides no information about the outcome of the other. For instance, flipping a coin and rolling a die are two independent processes – knowing the coin was heads does not help determine the outcome of a die roll. On the other hand, stock prices usually move up or down together, so they are not independent.

Independence was used implicitly on page 59 in the second solution to the probability that two carriers will have an affected child with cystic fibrosis. The assumption that half of the offspring who have received a mutated CF gene from the mother will receive a mutated gene from the father is essentially an independence assumption – genes are passed along from the mother and father independently.

Example 2.5 provides a basic example of two independent processes: rolling two dice. We want to determine the probability that both will be 1. Suppose one of the dice is red and the other white. If the outcome of the red die is a 1, it provides no information about the outcome of the white die. We first encountered this same question in Example 2.5 (page 59), where we calculated the probability using the following reasoning: $1/6^{th}$ of the time the red die is a 1, and $1/6^{th}$ of *those* times the white die will also be 1. This is illustrated in Figure 2.11. Because the rolls are independent, the probabilities of the corresponding outcomes can be multiplied to get the final answer: $(1/6)\times(1/6) = 1/36$. This can be generalized to many independent processes.

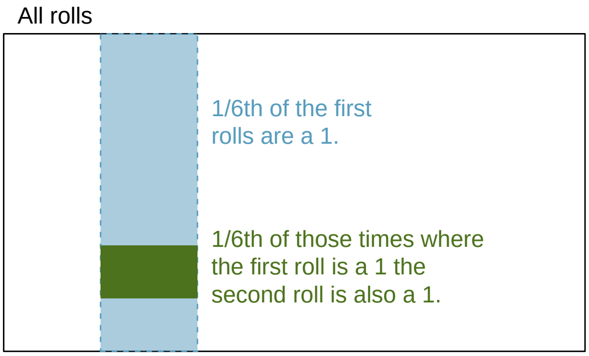

All rolls
1/6th of the first rolls are a 1.
1/6th of those times where the first roll is a 1 the second roll is also a 1.

Figure 2.11: $1/6^{th}$ of the time, the first roll is a 1. Then $1/6^{th}$ of *those* times, the second roll will also be a 1.

● **Example 2.24** What if there was also a blue die independent of the other two? What is the probability of rolling the three dice and getting all 1s?

The same logic applies from Example 2.5. If $1/36^{th}$ of the time the white and red dice are both 1, then $1/6^{th}$ of *those* times the blue die will also be 1, so multiply:

$$\begin{aligned} P(white = 1 \text{ and } red = 1 \text{ and } blue = 1) &= P(white = 1) \times P(red = 1) \times P(blue = 1) \\ &= (1/6) \times (1/6) \times (1/6) = 1/216 \end{aligned}$$

Example 2.24 illustrates what is called the Multiplication Rule for independent processes.

Multiplication Rule for independent processes

If A and B represent events from two different and independent processes, then the probability that both A and B occur can be calculated as the product of their separate probabilities:

$$P(A \text{ and } B) = P(A) \times P(B) \tag{2.25}$$

Similarly, if there are k events A_1, ..., A_k from k independent processes, then the probability they all occur is

$$P(A_1) \times P(A_2) \times \cdots \times P(A_k)$$

In applications to biology or medicine, complicated probability problems are often solved with the simple ideas used in the dice examples.

⊙ **Guided Practice 2.26** About 9% of people are left-handed. Suppose 2 people are selected at random from the U.S. population. Because the sample size of 2 is very small relative to the population, it is reasonable to assume these two people are independent. (a) What is the probability that both are left-handed? (b) What is the probability that both are right-handed?[12]

[12] (a) The probability the first person is left-handed is 0.09, which is the same for the second person. We apply

⊙ **Guided Practice 2.27** Suppose 5 people are selected at random.[13]

(a) What is the probability that all are right-handed?

(b) What is the probability that all are left-handed?

(c) What is the probability that not all of the people are right-handed?

Suppose the variables `handedness` and `gender` are independent, i.e. knowing someone's `gender` provides no useful information about their `handedness` and vice-versa. Then we can compute whether a randomly selected person is right-handed and female[14] using the Multiplication Rule:

$$\begin{aligned} P(\text{right-handed and female}) &= P(\text{right-handed}) \times P(\text{female}) \\ &= 0.91 \times 0.50 = 0.455 \end{aligned}$$

⊙ **Guided Practice 2.28** Three people are selected at random.[15]

(a) What is the probability that the first person is male and right-handed?

(b) What is the probability that the first two people are male and right-handed?.

(c) What is the probability that the third person is female and left-handed?

(d) What is the probability that the first two people are male and right-handed and the third person is female and left-handed?

● **Example 2.29** *Mandatory drug testing* Mandatory drug testing in the work place is common in professions such as air traffic controllers, transportation workers and government security agencies. A false positive in a drug screening test occurs when the test incorrectly indicates that a screened person is an illegal drug user. Suppose a mandatory drug test has a false-positive rate of 1.2% (i.e., has probability 0.012 of indicating that an employee is using illegal drugs even when that is not the case), and suppose a company uses the test to screen employees for drug use. Given 150 employees who are in reality drug free, what is the probability that at least one will (falsely) test positive if the outcome of one drug test has no affect on the other 149?

the Multiplication Rule for independent processes to determine the probability that both will be left-handed: $0.09 \times 0.09 = 0.0081$.

(b) It is reasonable to assume the proportion of people who are ambidextrous (both right and left handed) is nearly 0, which results in $P(\text{right-handed}) = 1 - 0.09 = 0.91$. Using the same reasoning as in part (a), the probability that both will be right-handed is $0.91 \times 0.91 = 0.8281$.

[13] (a) The abbreviations `RH` and `LH` are used for right-handed and left-handed, respectively. Since each are independent, we apply the Multiplication Rule for independent processes:

$$\begin{aligned} P(\text{all five are RH}) &= P(\text{first = RH, second = RH, ..., fifth = RH}) \\ &= P(\text{first = RH}) \times P(\text{second = RH}) \times \cdots \times P(\text{fifth = RH}) \\ &= 0.91 \times 0.91 \times 0.91 \times 0.91 \times 0.91 = 0.624 \end{aligned}$$

(b) Using the same reasoning as in (a), $0.09 \times 0.09 \times 0.09 \times 0.09 \times 0.09 = 0.0000059$

(c) Use the complement, $P(\text{all five are RH})$, to answer this question:

$$P(\text{not all RH}) = 1 - P(\text{all RH}) = 1 - 0.624 = 0.376$$

[14] The actual proportion of the U.S. population that is `female` is about 50%, and so we use 0.5 for the probability of sampling a woman. However, this probability does differ in other countries.

[15] Brief answers are provided. (a) This can be written in probability notation as $P(\text{a randomly selected person is male and right-handed}) = 0.455$. (b) 0.207. (c) 0.045. (d) 0.0093.

The solution uses independence (the assumption that the outcome of one test has no effect on the others) and the multiplication rule to calculate the probability of the complement of the event asked about.

$$\begin{aligned} P(\text{At least 1 "+"}) &= P(\text{1 or 2 or 3} \ldots \text{or 150 are "+"}) \\ &= 1 - P(\text{None are "+"}) \\ &= 1 - P(\text{150 "-"}) \\ P(\text{150 are "-"}) &= P(\text{1 is "-"})^{150} \\ &= (0.988)^{150} = 0.16. \end{aligned}$$

So $P(\text{At least 1 is "+"}) = 1 - P(\text{150 are "-"}) = 0.84$.

should we be more formal here in defining events? Also, this is the example that we also solved in R, two different ways. Those solutions are candidates for the R supplement

Some people find the result surprising. Even when using a test with a small probability of a false positive, the company is more than 80% likely to incorrectly claim at least one employee is an illegal drug user.

Guided Practice 2.30 Because of the high likelihood of at least one false positive in company wide drug screening programs, an individual with a positive test is almost always re-tested with a different screening test, one that is more expensive than the first but with a lower false positive probability. Suppose the second test has a false positive rate of 0.8%. What is the probability that an employee who is not using illegal drugs will test positive on both tests?

solution to be added if we keep the problem

Guided Practice 2.31

There are eight different common blood types, which are determined by the presence of certain antigens located on cell surfaces. Antigens are substances used by the immune system to recognize self versus non-self; if the immune system encounters antigens not normally found on the body's own cells, it will attack the foreign cells. When patients receive blood transfusions, it is critical that the antigens of transfused cells match those of the patient's, or else an immune system response will be triggered.

The ABO blood group system consists of four different blood groups, which describe whether an individual's red blood cells carry the A antigen, B antigen, both, or neither. The ABO gene has three alleles: I^A, I^B, and i. The i allele is recessive to both I^A and I^B, and does not produce antigens; thus, an individual with genotype I^Ai is blood group A and an individual with genotype I^Bi is blood group B. The I^A and I^B are codominant, such that individuals of I^AI^B genotype are AB. Individuals homozygous for the i allele are known as blood group O, with neither A nor B antigens.

a) *ABO, Independence.* Suppose that both members of a couple have Group AB blood.

   i. What is the probability that a child of this couple will have Group A blood?

   ii. What is the probability that they have two children with Group A blood?

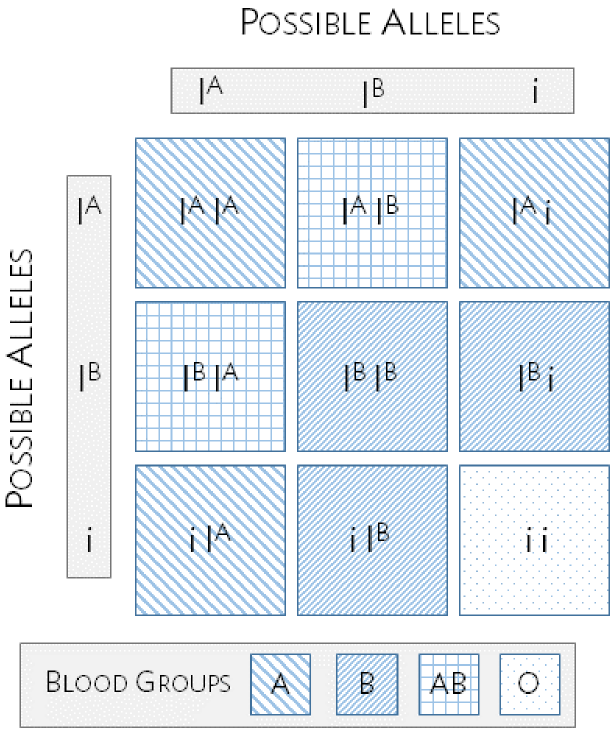


Figure 2.12: Inheritance of ABO blood groups.

solutions to be added if we keep the exercise

The examples in this section have used independence to solve probability problems. Sometimes the definition of independence can be used to check whether two events are independent – two events A and B are independent if they satisfy Equation (2.25).

● **Example 2.32** If we shuffle up a deck of cards and draw one, is the event that the card is a heart independent of the event that the card is an ace?

The probability the card is a heart is 1/4 and the probability that it is an ace is 1/13. The probability the card is the ace of hearts is 1/52. We check whether Equation 2.25 is satisfied:

$$P(\heartsuit) \times P(\text{ace}) = \frac{1}{4} \times \frac{1}{13} = \frac{1}{52} = P(\heartsuit \text{ and ace})$$

Because the equation holds, the event that the card is a heart and the event that the card is an ace are independent events.

● **Example 2.33** I

n the general population, about 15% of adults between 25 and 40 years of age are hypertensive. Suppose that among males of this age, hypertension occurs about 18% of the time. Is hypertension independent of sex?

solution to be filled in if we keep it. Emphasize in the solution how the wording here is more realistic than the playing card/dice examples.

2.2 Conditional probability

Precise estimates are difficult to come by, but the US CDC estimated that in 2012, approximately 29.1 million people have type 2 diabetes, or about 9.3% of the population. Twenty-one million of these cases of diabetes are diagnosed, while 8.1 million cases are undiagnosed (people living with diabetes but they and their physicians are unaware that they have the disease). A health care practitioner seeing a new patient and having no demographic or health information about the patient should expect a 9.3% chance that the patient might have diabetes, diagnosed or otherwise. But intake interviews usually include background information about patients, so that a health care practitioner knows a bit more about a new patient. Not surprisingly, the prevalence or probability of type 2 diabetes varies with age. Between the ages of 20 and 44, approximately 4% of the population have diabetes, but by age 65 and older, almost 27% of of that age group have diabetes. Knowing the age of a patient provides information about the chance of diabetes, so that age and diabetes status are not independent. While the probability of diabetes in randomly chosen member of the population is 0.093, the *conditional* probability of diabetes in a person known to be 65 or older is about 0.27.

Conditional probability is used to characterize how the probability of an outcome varies with the knowledge of another factor or condition, and is closely related to the concepts of marginal and joint probabilities.

2.2.1 Marginal and joint probabilities

Tables 2.13 and 2.14 provide additional information about the relationship between diabetes prevalence and age. [16] Table 2.13 is a contingency table like those discussed in Chapter 1, but for the entire US 2012 population; the values in the table are in thousands, to make the table more readable. The table shows in the first row, for instance, that in the entire population of approximately 313,320,000 approximately 200,000 individuals were in the age group less than 20 and suffered from type 2 diagnosis, or about 0.1%. The table also provides the information among the approximately 86,864,000 individuals less than 20 years of age, only 200,000 suffered from type 2 diabetes, approximately 0.2%. The distinction between the two statements is small but important – the first provides information about the size of the type 2 diabetes population relative to the entire population and the second about the size of the diabetes population less than 20 year old age group relative to the size of that age group.

Guided Practice 2.34

What fraction of the US population are 45 to 64 years of age and have diabetes? What fraction of the population age 45 to 64 have diabetes?

The counts in Table 2.13 have been converted to proportions by dividing each value in the cells of the contingency table by the total population size, 313,320,000. The entries in this table show the proportions of the population in each of the 8 categories defined by diabetes status and age. If these proportions are interpreted as probabilities for randomly chosen individuals from the population, 0.014 in row 2 implies that the probability of selecting someone at random who has diabetes and whose age is between 20 and 44 is 0.014, or 1.4%. The entries in the 8 main table cells (excluding the values in the margins)

[16] Because the CDC provides only approximate numbers for diabetes prevalence, the numbers in the table are approximations to actual population counts.

	Diabetes	No Diabetes	Sum
Less than 20 years	200	86664	86864
20 to 44 years	4300	98724	103024
45 to 64 years	13400	68526	81926
Greater than 64 years	11200	30306	41506
Sum	29100	284220	313320

Table 2.13: Contingency table showing type 2 diabetes status and age group, in thousands

are called **joint probabilities** since they specify the probability of two events happening at the same time – diabetes and a particular age group. In probability notation, $0.014 = P(\text{diabetes and age 20 to 44})$. It is common to also write this as $P(\text{diabetes, age 20 to 44})$, with a comma replacing "and".

The values in the last row and column of the table are the sums of the corresponding rows or columns. Since 0.329 is the sum of the of the probabilities of the disjoint events (diabetes and age 20 to 44) and (no diabetes and age 20 to 44), it is the probability of being in the age group 20 to 44. The row and column sums are called **marginal probabilities**; they are probabilities about only one type of event, age in the case of 0.0329. The sum of the first column (0.093) is the marginal probability of a member of the population having diabetes.

	Diabetes	No Diabetes	Sum
Less than 20 years	0.001	0.277	0.277
20 to 44 years	0.014	0.315	0.329
45 to 64 years	0.043	0.219	0.261
Greater than 64 years	0.036	0.097	0.132
Sum	0.093	0.907	1.000

Table 2.14: Probability table summarizing diabetes status and age group

Marginal and joint probabilities

If a probability is based on a single variable, it is a *marginal probability*. The probability of outcomes for two or more variables or processes is called a *joint probability*.

⊙ **Guided Practice 2.35** What is the interpretation of the value 0.907 in the last row of the table? Of the value 1.000 in the bottom right corner?

2.2.2 Defining conditional probability

The probability that a randomly selected individual from the US has diabetes is 0.093, the sum of the first column in Table 2.14. How does that probability change if we know the individual's age is 65 or greater? Table 2.13 shows that 11,200,000 of the 41,506,000 people in that age group have diabetes, so the likelihood that someone from that age has diabetes is

$$\frac{11,200,000}{41,506,000} = 0.27,$$

or 27%. The additional information about age allows a better estimate of the probability of diabetes; the conditional probability of diabetes, given the information that an individual is older than 65, is 0.27.

Since

$$\begin{aligned}\frac{11,200,000}{41,506,000} &= \frac{11,200,000/313,320,000}{41,506,000/313,320,000}\\ &= \frac{0.036}{0.132}\\ &= 0.270,\end{aligned}$$

the calculation of conditional probability could have been done using the values in Table 2.13. The conditional probability of diabetes given age 65 or greater is simply the ratio of the proportion of the population with diabetes and age 65 or greater divided by the proportion greater than age 65. This leads to the mathematical definition of conditional probability.

Conditional probability

The conditional probability of the outcome of interest A given condition B is computed as the following:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \tag{2.36}$$

⊙ **Guided Practice 2.37** (a) Write out the following statement in conditional probability notation: *"The probability a randomly selected person has diabetes, given that his or her age is between 45 and 64 "*. Notice that the condition is now based on the teenager, not the parent.

(b) Calculate the conditional probability in part (a) (c) Write out the following statement in conditional probability notation: *"The probability a randomly selected person is between 45 and 64 years old, given that the person has diabetes"*. Notice that the condition is now based on the diabetes, not the age.

(d) Calculate the probability in part (c).

2.2.3 Smallpox in Boston, 1721

I am not fond of small pox example, since it is a biased sample, but leaving it for now

The `smallpox` data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston.[17] Doctors at the time believed that inoculation, which involves exposing a person to the disease in a controlled form, could reduce the likelihood of death.

[17] Fenner F. 1988. *Smallpox and Its Eradication (History of International Public Health, No. 6)*. Geneva: World Health Organization. ISBN 92-4-156110-6.

Each case represents one person with two variables: `inoculated` and `result`. The variable `inoculated` takes two levels: `yes` or `no`, indicating whether the person was inoculated or not. The variable `result` has outcomes `lived` or `died`. These data are summarized in Tables 2.15 and 2.16.

		inoculated		
		`yes`	`no`	Total
`result`	`lived`	238	5136	5374
	`died`	6	844	850
	Total	244	5980	6224

Table 2.15: Contingency table for the `smallpox` data set.

		inoculated		
		`yes`	`no`	Total
`result`	`lived`	0.0382	0.8252	0.8634
	`died`	0.0010	0.1356	0.1366
	Total	0.0392	0.9608	1.0000

Table 2.16: Table proportions for the `smallpox` data, computed by dividing each count by the table total, 6224.

⊙ **Guided Practice 2.38** Write out, in formal notation, the probability a randomly selected person who was not inoculated died from smallpox, and find this probability.[18]

⊙ **Guided Practice 2.39** Determine the probability that an inoculated person died from smallpox. How does this result compare with the result of Guided Practice 2.38?[19]

⊙ **Guided Practice 2.40** The people of Boston self-selected whether or not to be inoculated. (a) Is this study observational or was this an experiment? (b) Can we infer any causal connection using these data? (c) What are some potential confounding variables that might influence whether someone `lived` or `died` and also affect whether that person was inoculated?[20]

2.2.4 General multiplication rule

Section 2.1.7 introduced the Multiplication Rule for independent processes. Here we provide the **General Multiplication Rule** for events that might not be independent.

[18] $P(\texttt{result} = \texttt{died} \mid \texttt{inoculated} = \texttt{no}) = \frac{P(\texttt{result} = \texttt{died and inoculated} = \texttt{no})}{P(\texttt{inoculated} = \texttt{no})} = \frac{0.1356}{0.9608} = 0.1411$.

[19] $P(\texttt{result} = \texttt{died} \mid \texttt{inoculated} = \texttt{yes}) = \frac{P(\texttt{result} = \texttt{died and inoculated} = \texttt{yes})}{P(\texttt{inoculated} = \texttt{yes})} = \frac{0.0010}{0.0392} = 0.0255$. The death rate for individuals who were inoculated is only about 1 in 40 while the death rate is about 1 in 7 for those who were not inoculated.

[20] Brief answers: (a) Observational. (b) No, we cannot infer causation from this observational study. (c) Accessibility to the latest and best medical care. There are other valid answers for part (c).

General Multiplication Rule

If A and B represent two outcomes or events, then

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

It is useful to think of A as the outcome of interest and B as the condition.

This General Multiplication Rule is simply a rearrangement of the definition for conditional probability in Equation (2.36) on page 77.

● **Example 2.41** Consider the `smallpox` data set. Suppose we are given only two pieces of information: 96.08% of residents were not inoculated, and 85.88% of the residents who were not inoculated ended up surviving. How could we compute the probability that a resident was not inoculated and lived?

We will compute our answer using the General Multiplication Rule and then verify it using Table 2.16. We want to determine

$$P(\texttt{result} = \texttt{lived} \text{ and } \texttt{inoculated} = \texttt{no})$$

and we are given that

$$P(\texttt{result} = \texttt{lived} \mid \texttt{inoculated} = \texttt{no}) = 0.8588$$
$$P(\texttt{inoculated} = \texttt{no}) = 0.9608$$

Among the 96.08% of people who were not inoculated, 85.88% survived:

$$P(\texttt{result} = \texttt{lived} \text{ and } \texttt{inoculated} = \texttt{no}) = 0.8588 \times 0.9608 = 0.8251$$

This is equivalent to the General Multiplication Rule. We can confirm this probability in Table 2.16 at the intersection of `no` and `lived` (with a small rounding error).

⊙ **Guided Practice 2.42** Use $P(\texttt{inoculated} = \texttt{yes}) = 0.0392$ and $P(\texttt{result} = \texttt{lived} \mid \texttt{inoculated} = \texttt{yes}) = 0.9754$ to determine the probability that a person was both inoculated and lived.[21]

⊙ **Guided Practice 2.43** If 97.45% of the people who were inoculated lived, what proportion of inoculated people must have died?[22]

Sum of conditional probabilities

Let A_1, ..., A_k represent all the disjoint outcomes for a variable or process. Then if B is an event, possibly for another variable or process, we have:

$$P(A_1|B) + \cdots + P(A_k|B) = 1$$

The rule for complements also holds when an event and its complement are conditioned on the same information:

$$P(A|B) = 1 - P(A^c|B)$$

[21] The answer is 0.0382, which can be verified using Table 2.16.

[22] There were only two possible outcomes: `lived` or `died`. This means that 100% - 97.45% = 2.55% of the people who were inoculated died.

⊙ **Guided Practice 2.44** Based on the probabilities computed above, does it appear that inoculation is effective at reducing the risk of death from smallpox?[23]

2.2.5 Independence and conditional probability

If two events are independent, knowing the outcome of one should provide no information about the other. That intuitively clear statement can be shown mathematically.

⊙ **Guided Practice 2.45** Let X and Y represent the outcomes of rolling two dice.[24]

(a) What is the probability that the first die, X, is 1?

(b) What is the probability that both X and Y are 1?

(c) Use the formula for conditional probability to compute $P(Y = 1 \mid X = 1)$.

(d) What is $P(Y = 1)$? Is this different from the answer from part (c)? Explain.

[23] The samples are large relative to the difference in death rates for the "inoculated" and "not inoculated" groups, so it seems there is an association between `inoculated` and `outcome`. However, as noted in the solution to Guided Practice 2.40, this is an observational study and we cannot be sure if there is a causal connection. (Further research has shown that inoculation is effective at reducing death rates.)

[24] Brief solutions: (a) 1/6. (b) 1/36. (c) $\frac{P(Y=1 \text{ and } X=1)}{P(X=1)} = \frac{1/36}{1/6} = 1/6$. (d) The probability is the same as in part (c): $P(Y = 1) = 1/6$. The probability that $Y = 1$ was unchanged by knowledge about X, which makes sense as X and Y are independent.

It is not difficult to show in Guided Practice 2.45(c) that the conditioning information has no influence by using the Multiplication Rule for independence events:

$$\begin{aligned} P(Y=1 \mid X=1) &= \frac{P(Y=1 \text{ and } X=1)}{P(X=1)} \\ &= \frac{P(Y=1)\times P(X=1)}{P(X=1)} \\ &= P(Y=1) \end{aligned}$$

⊙ **Guided Practice 2.46** Casinos often rely on gamblers not understanding independence. Suppose the last five outcomes on a roulette table were `black`. What is wrong with the reasoning that the next outcome is highly likely to be red?

There is a subtle but important point behind the last example. The probability of the next six outcomes being black is different than the probability that the sixth outcome is black when a gambler has seen the last five outcomes and knows that they are black. This is an example of an unconditional versus a conditional probability.

2.2.6 Tree diagrams

Tree diagrams are a tool to organize outcomes and probabilities around the structure of the data. They are most useful when two or more processes occur in a sequence and each process is conditioned on its predecessors.

The `smallpox` data fit this description. We see the population as split by `inoculation`: `yes` and `no`. Following this split, survival rates were observed for each group. This structure is reflected in the **tree diagram** shown in Figure 2.17. The first branch for `inoculation` is said to be the **primary** branch while the other branches are **secondary**.

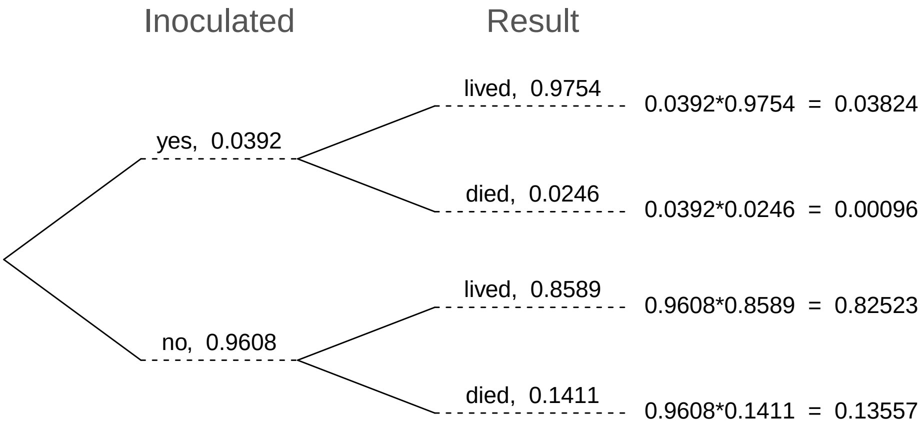


Figure 2.17: A tree diagram of the `smallpox` data set.

Tree diagrams are annotated with marginal and conditional probabilities, as shown in Figure 2.17. This tree diagram splits the smallpox data by `inoculation` into the yes and no groups with respective marginal probabilities 0.0392 and 0.9608. The secondary branches are conditioned on the first, so we assign conditional probabilities to these branches. For example, the top branch in Figure 2.17 is the probability that result = lived conditioned on the information that inoculated = yes. We may (and usually do) construct joint probabilities at the end of each branch in our tree by multiplying the numbers we come across as we move from left to right. These joint probabilities are computed using the General Multiplication Rule:

$$\begin{aligned}&P(\texttt{inoculated} = \texttt{yes and } \texttt{result} = \texttt{lived})\\ &\quad = P(\texttt{inoculated} = \texttt{yes}) \times P(\texttt{result} = \texttt{lived}|\texttt{inoculated} = \texttt{yes})\\ &\quad = 0.0392 \times 0.9754 = 0.0382\end{aligned}$$

replace the next example with the CF example: prob of affected child, no information about the carrier status of the parents. Complicated tree

In addition to being a graphical representation of how to compute a probability, tree diagrams are a useful way to organize the information in a probability problem and can often reduce a seemingly difficult problem to a series of simple steps

● **Example 2.47**

In the general population, about 1 in 28 individuals is an unaffected carrier of a mutation in the cystic fibrosis gene, *CFTR*, discussed in example 2.6. Most unaffected carriers are unaware that they harbor the mutation. Suppose that people with cystic fibrosis do not live long enough to reproduce with a partner. In the absence of any testing information, what is the probability that a child of two parent will have CF?

solution not give for now, since I did not want to take the time to draw the tree. This will be a good example to also solve algebraically, to show that the tree diagram is a representation of the theorem of total prob. Useful, since it will come up with Bayes Thm. Borrow some of the wording from the solution to the grade problem in OI that I removed here.

2.3 Exercises

2.3.1 Defining probability

2.1 True or false. Determine if the statements below are true or false, and explain your reasoning.

(a) If a fair coin is tossed many times and the last eight tosses are all heads, then the chance that the next toss will be heads is somewhat less than 50%.

(b) Drawing a face card (jack, queen, or king) and drawing a red card from a full deck of playing cards are mutually exclusive events.

(c) Drawing a face card and drawing an ace from a full deck of playing cards are mutually exclusive events.

2.2 Roulette wheel. The game of roulette involves spinning a wheel with 38 slots: 18 red, 18 black, and 2 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball.

(a) You watch a roulette wheel spin 3 consecutive times and the ball lands on a red slot each time. What is the probability that the ball will land on a red slot on the next spin?

(b) You watch a roulette wheel spin 300 consecutive times and the ball lands on a red slot each time. What is the probability that the ball will land on a red slot on the next spin?

(c) Are you equally confident of your answers to parts (a) and (b)? Why or why not?

Photo by Håkan Dahlström
(http://flic.kr/p/93fEzp)
CC BY 2.0 license

2.3 Four games, one winner. Below are four versions of the same game. Your archnemisis gets to pick the version of the game, and then you get to choose how many times to flip a coin: 10 times or 100 times. Identify how many coin flips you should choose for each version of the game. It costs $1 to play each game. Explain your reasoning.

(a) If the proportion of heads is larger than 0.60, you win $1.

(b) If the proportion of heads is larger than 0.40, you win $1.

(c) If the proportion of heads is between 0.40 and 0.60, you win $1.

(d) If the proportion of heads is smaller than 0.30, you win $1.

2.4 Backgammon. Backgammon is a board game for two players in which the playing pieces are moved according to the roll of two dice. Players win by removing all of their pieces from the board, so it is usually good to roll high numbers. You are playing backgammon with a friend and you roll two 6s in your first roll and two 6s in your second roll. Your friend rolls two 3s in his first roll and again in his second row. Your friend claims that you are cheating, because rolling double 6s twice in a row is very unlikely. Using probability, show that your rolls were just as likely as his.

2.5 Coin flips. If you flip a fair coin 10 times, what is the probability of

(a) getting all tails?

(b) getting all heads?

(c) getting at least one tails?

2.6 Dice rolls. If you roll a pair of fair dice, what is the probability of

(a) getting a sum of 1?

(b) getting a sum of 5?

(c) getting a sum of 12?

2.7 Swing voters. A 2012 Pew Research survey asked 2,373 randomly sampled registered voters their political affiliation (Republican, Democrat, or Independent) and whether or not they identify as swing voters. 35% of respondents identified as Independent, 23% identified as swing voters, and 11% identified as both.[25]

(a) Are being Independent and being a swing voter disjoint, i.e. mutually exclusive?

(b) Draw a Venn diagram summarizing the variables and their associated probabilities.

(c) What percent of voters are Independent but not swing voters?

(d) What percent of voters are Independent or swing voters?

(e) What percent of voters are neither Independent nor swing voters?

(f) Is the event that someone is a swing voter independent of the event that someone is a political Independent?

2.8 Poverty and language. The American Community Survey is an ongoing survey that provides data every year to give communities the current information they need to plan investments and services. The 2010 American Community Survey estimates that 14.6% of Americans live below the poverty line, 20.7% speak a language other than English (foreign language) at home, and 4.2% fall into both categories.[26]

(a) Are living below the poverty line and speaking a foreign language at home disjoint?

(b) Draw a Venn diagram summarizing the variables and their associated probabilities.

(c) What percent of Americans live below the poverty line and only speak English at home?

(d) What percent of Americans live below the poverty line or speak a foreign language at home?

(e) What percent of Americans live above the poverty line and only speak English at home?

(f) Is the event that someone lives below the poverty line independent of the event that the person speaks a foreign language at home?

2.9 Disjoint vs. independent. In parts (a) and (b), identify whether the events are disjoint, independent, or neither (events cannot be both disjoint and independent).

(a) You and a randomly selected student from your class both earn A's in this course.

(b) You and your class study partner both earn A's in this course.

(c) If two events can occur at the same time, must they be dependent?

2.10 Guessing on an exam. In a multiple choice exam, there are 5 questions and 4 choices for each question (a, b, c, d). Nancy has not studied for the exam at all and decides to randomly guess the answers. What is the probability that:

(a) the first question she gets right is the 5^{th} question?

(b) she gets all of the questions right?

(c) she gets at least one question right?

[25] **indepSwing**.

[26] **poorLang**.

2.11 Educational attainment of couples. The table below shows the distribution of education level attained by US residents by gender based on data collected during the 2010 American Community Survey.[27]

		Gender	
		Male	Female
Highest education attained	Less than 9th grade	0.07	0.13
	9th to 12th grade, no diploma	0.10	0.09
	HS graduate (or equivalent)	0.30	0.20
	Some college, no degree	0.22	0.24
	Associate's degree	0.06	0.08
	Bachelor's degree	0.16	0.17
	Graduate or professional degree	0.09	0.09
	Total	1.00	1.00

(a) What is the probability that a randomly chosen man has at least a Bachelor's degree?

(b) What is the probability that a randomly chosen woman has at least a Bachelor's degree?

(c) What is the probability that a man and a woman getting married both have at least a Bachelor's degree? Note any assumptions you must make to answer this question.

(d) If you made an assumption in part (c), do you think it was reasonable? If you didn't make an assumption, double check your earlier answer and then return to this part.

2.12 School absences. Data collected at elementary schools in DeKalb County, GA suggest that each year roughly 25% of students miss exactly one day of school, 15% miss 2 days, and 28% miss 3 or more days due to sickness.[28]

(a) What is the probability that a student chosen at random doesn't miss any days of school due to sickness this year?

(b) What is the probability that a student chosen at random misses no more than one day?

(c) What is the probability that a student chosen at random misses at least one day?

(d) If a parent has two kids at a DeKalb County elementary school, what is the probability that neither kid will miss any school? Note any assumption you must make to answer this question.

(e) If a parent has two kids at a DeKalb County elementary school, what is the probability that both kids will miss some school, i.e. at least one day? Note any assumption you make.

(f) If you made an assumption in part (d) or (e), do you think it was reasonable? If you didn't make any assumptions, double check your earlier answers.

2.13 Grade distributions. Each row in the table below is a proposed grade distribution for a class. Identify each as a valid or invalid probability distribution, and explain your reasoning.

			Grades		
	A	B	C	D	F
(a)	0.3	0.3	0.3	0.2	0.1
(b)	0	0	1	0	0
(c)	0.3	0.3	0.3	0	0
(d)	0.3	0.5	0.2	0.1	-0.1
(e)	0.2	0.4	0.2	0.1	0.1
(f)	0	-0.1	1.1	0	0

[27] **eduSex**.

[28] **Mizan:2011**.

2.14 Health coverage, frequencies. The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table summarizes two variables for the respondents: health status and health coverage, which describes whether each respondent had health insurance.[29]

		Health Status					
		Excellent	Very good	Good	Fair	Poor	Total
Health	No	459	727	854	385	99	2,524
Coverage	Yes	4,198	6,245	4,821	1,634	578	17,476
	Total	4,657	6,972	5,675	2,019	677	20,000

(a) If we draw one individual at random, what is the probability that the respondent has excellent health and doesn't have health coverage?

(b) If we draw one individual at random, what is the probability that the respondent has excellent health or doesn't have health coverage?

2.3.2 Conditional probability

2.15 Joint and conditional probabilities. P(A) = 0.3, P(B) = 0.7

(a) Can you compute P(A and B) if you only know P(A) and P(B)?

(b) Assuming that events A and B arise from independent random processes,

i. what is P(A and B)?

ii. what is P(A or B)?

iii. what is P(A|B)?

(c) If we are given that P(A and B) = 0.1, are the random variables giving rise to events A and B independent?

(d) If we are given that P(A and B) = 0.1, what is P(A|B)?

2.16 PB & J. Suppose 80% of people like peanut butter, 89% like jelly, and 78% like both. Given that a randomly sampled person likes peanut butter, what's the probability that he also likes jelly?

2.17 Global warming. A 2010 Pew Research poll asked 1,306 Americans "From what you've read and heard, is there solid evidence that the average temperature on earth has been getting warmer over the past few decades, or not?". The table below shows the distribution of responses by party and ideology, where the counts have been replaced with relative frequencies.[30]

		Response			
		Earth is warming	Not warming	Don't Know Refuse	Total
	Conservative Republican	0.11	0.20	0.02	0.33
Party and	Mod/Lib Republican	0.06	0.06	0.01	0.13
Ideology	Mod/Cons Democrat	0.25	0.07	0.02	0.34
	Liberal Democrat	0.18	0.01	0.01	0.20
	Total	0.60	0.34	0.06	1.00

(a) Are believing that the earth is warming and being a liberal Democrat mutually exclusive?

(b) What is the probability that a randomly chosen respondent believes the earth is warming or is a liberal Democrat? **(See the next page for parts (c)-(f).)**

(c) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a liberal Democrat?

[29] **data:BRFSS2010**.

[30] **globalWarming**.

(d) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a conservative Republican?

(e) Does it appear that whether or not a respondent believes the earth is warming is independent of their party and ideology? Explain your reasoning.

(f) What is the probability that a randomly chosen respondent is a moderate/liberal Republican given that he does not believe that the earth is warming?

2.18 Health coverage, relative frequencies. The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table displays the distribution of health status of respondents to this survey (excellent, very good, good, fair, poor) conditional on whether or not they have health insurance.

		Health Status					
		Excellent	Very good	Good	Fair	Poor	Total
Health Coverage	No	0.0230	0.0364	0.0427	0.0192	0.0050	0.1262
	Yes	0.2099	0.3123	0.2410	0.0817	0.0289	0.8738
	Total	0.2329	0.3486	0.2838	0.1009	0.0338	1.0000

(a) Are being in excellent health and having health coverage mutually exclusive?

(b) What is the probability that a randomly chosen individual has excellent health?

(c) What is the probability that a randomly chosen individual has excellent health given that he has health coverage?

(d) What is the probability that a randomly chosen individual has excellent health given that he doesn't have health coverage?

(e) Do having excellent health and having health coverage appear to be independent?

2.19 Burger preferences. A 2010 SurveyUSA poll asked 500 Los Angeles residents, "What is the best hamburger place in Southern California? Five Guys Burgers? In-N-Out Burger? Fat Burger? Tommy's Hamburgers? Umami Burger? Or somewhere else?" The distribution of responses by gender is shown below.[31]

		Gender		
		Male	Female	Total
Best hamburger place	Five Guys Burgers	5	6	11
	In-N-Out Burger	162	181	343
	Fat Burger	10	12	22
	Tommy's Hamburgers	27	27	54
	Umami Burger	5	1	6
	Other	26	20	46
	Not Sure	13	5	18
	Total	248	252	500

(a) Are being female and liking Five Guys Burgers mutually exclusive?

(b) What is the probability that a randomly chosen male likes In-N-Out the best?

(c) What is the probability that a randomly chosen female likes In-N-Out the best?

(d) What is the probability that a man and a woman who are dating both like In-N-Out the best? Note any assumption you make and evaluate whether you think that assumption is reasonable.

(e) What is the probability that a randomly chosen person likes Umami best or that person is female?

2.20 Assortative mating. Assortative mating is a nonrandom mating pattern where individuals with similar genotypes and/or phenotypes mate with one another more frequently than what would

[31] **burgers**.

be expected under a random mating pattern. Researchers studying this topic collected data on eye colors of 204 Scandinavian men and their female partners. The table below summarizes the results. For simplicity, we only include heterosexual relationships in this exercise.[32]

		Partner (female)			
		Blue	Brown	Green	Total
Self (male)	Blue	78	23	13	114
	Brown	19	23	12	54
	Green	11	9	16	36
	Total	108	55	41	204

(a) What is the probability that a randomly chosen male respondent or his partner has blue eyes?

(b) What is the probability that a randomly chosen male respondent with blue eyes has a partner with blue eyes?

(c) What is the probability that a randomly chosen male respondent with brown eyes has a partner with blue eyes? What about the probability of a randomly chosen male respondent with green eyes having a partner with blue eyes?

(d) Does it appear that the eye colors of male respondents and their partners are independent? Explain your reasoning.

2.21 Drawing box plots. After an introductory statistics course, 80% of students can successfully construct box plots. Of those who can construct box plots, 86% passed, while only 65% of those students who could not construct box plots passed.

(a) Construct a tree diagram of this scenario.

(b) Calculate the probability that a student is able to construct a box plot if it is known that he passed.

2.22 Predisposition for thrombosis. A genetic test is used to determine if people have a predisposition for *thrombosis*, which is the formation of a blood clot inside a blood vessel that obstructs the flow of blood through the circulatory system. It is believed that 3% of people actually have this predisposition. The genetic test is 99% accurate if a person actually has the predisposition, meaning that the probability of a positive test result when a person actually has the predisposition is 0.99. The test is 98% accurate if a person does not have the predisposition. What is the probability that a randomly selected person who tests positive for the predisposition by the test actually has the predisposition?

2.23 HIV in Swaziland. Swaziland has the highest HIV prevalence in the world: 25.9% of this country's population is infected with HIV.[33] The ELISA test is one of the first and most accurate tests for HIV. For those who carry HIV, the ELISA test is 99.7% accurate. For those who do not carry HIV, the test is 92.6% accurate. If an individual from Swaziland has tested positive, what is the probability that he carries HIV?

2.24 Exit poll. Edison Research gathered exit poll results from several sources for the Wisconsin recall election of Scott Walker. They found that 53% of the respondents voted in favor of Scott Walker. Additionally, they estimated that of those who did vote in favor for Scott Walker, 37% had a college degree, while 44% of those who voted against Scott Walker had a college degree. Suppose we randomly sampled a person who participated in the exit poll and found that he had a college degree. What is the probability that he voted in favor of Scott Walker?[34]

2.25 It's never lupus. Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to

[32] **Laeng:2007**.

[33] **ciaFactBookHIV:2012**.

[34] **data:scott**.

a high risk of blood clotting. It is believed that 2% of the population suffer from this disease. The test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease. There is a line from the Fox television show *House* that is often used after a patient tests positive for lupus: "It's never lupus." Do you think there is truth to this statement? Use appropriate probabilities to support your answer.

2.26 Twins. About 30% of human twins are identical, and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the probability that they are identical?

2.3.3 Sampling from a small population

2.27 Marbles in an urn. Imagine you have an urn containing 5 red, 3 blue, and 2 orange marbles in it.

(a) What is the probability that the first marble you draw is blue?

(b) Suppose you drew a blue marble in the first draw. If drawing with replacement, what is the probability of drawing a blue marble in the second draw?

(c) Suppose you instead drew an orange marble in the first draw. If drawing with replacement, what is the probability of drawing a blue marble in the second draw?

(d) If drawing with replacement, what is the probability of drawing two blue marbles in a row?

(e) When drawing with replacement, are the draws independent? Explain.

2.28 Socks in a drawer. In your sock drawer you have 4 blue, 5 gray, and 3 black socks. Half asleep one morning you grab 2 socks at random and put them on. Find the probability you end up wearing

(a) 2 blue socks

(b) no gray socks

(c) at least 1 black sock

(d) a green sock

(e) matching socks

2.29 Chips in a bag. Imagine you have a bag containing 5 red, 3 blue, and 2 orange chips.

(a) Suppose you draw a chip and it is blue. If drawing without replacement, what is the probability the next is also blue?

(b) Suppose you draw a chip and it is orange, and then you draw a second chip without replacement. What is the probability this second chip is blue?

(c) If drawing without replacement, what is the probability of drawing two blue chips in a row?

(d) When drawing without replacement, are the draws independent? Explain.

2.30 Books on a bookshelf. The table below shows the distribution of books on a bookcase based on whether they are nonfiction or fiction and hardcover or paperback.

		Format		
		Hardcover	Paperback	Total
Type	Fiction	13	59	72
	Nonfiction	15	8	23
	Total	28	67	95

(a) Find the probability of drawing a hardcover book first then a paperback fiction book second when drawing without replacement.

(b) Determine the probability of drawing a fiction book first and then a hardcover book second, when drawing without replacement.

(c) Calculate the probability of the scenario in part (b), except this time complete the calculations under the scenario where the first book is placed back on the bookcase before randomly drawing the second book.

(d) The final answers to parts (b) and (c) are very similar. Explain why this is the case.

2.31 Student outfits. In a classroom with 24 students, 7 students are wearing jeans, 4 are wearing shorts, 8 are wearing skirts, and the rest are wearing leggings. If we randomly select 3 students without replacement, what is the probability that one of the selected students is wearing leggings and the other two are wearing jeans? Note that these are mutually exclusive clothing options.

2.32 The birthday problem. Suppose we pick three people at random. For each of the following questions, ignore the special case where someone might be born on February 29th, and assume that births are evenly distributed throughout the year.

(a) What is the probability that the first two people share a birthday?

(b) What is the probability that at least two people share a birthday?

2.3.4 Random variables

2.33 College smokers. At a university, 13% of students smoke.

(a) Calculate the expected number of smokers in a random sample of 100 students from this university.

(b) The university gym opens at 9 am on Saturday mornings. One Saturday morning at 8:55 am there are 27 students outside the gym waiting for it to open. Should you use the same approach from part (a) to calculate the expected number of smokers among these 27 students?

2.34 Ace of clubs wins. Consider the following card game with a well-shuffled deck of cards. If you draw a red card, you win nothing. If you get a spade, you win \$5. For any club, you win \$10 plus an extra \$20 for the ace of clubs.

(a) Create a probability model for the amount you win at this game. Also, find the expected winnings for a single game and the standard deviation of the winnings.

(b) What is the maximum amount you would be willing to pay to play this game? Explain your reasoning.

2.35 Hearts win. In a new card game, you start with a well- shuffled full deck and draw 3 cards without replacement. If you draw 3 hearts, you win \$50. If you draw 3 black cards, you win \$25. For any other draws, you win nothing.

(a) Create a probability model for the amount you win at this game, and find the expected winnings. Also compute the standard deviation of this distribution.

(b) If the game costs \$5 to play, what would be the expected value and standard deviation of the net profit (or loss)? *(Hint: profit = winnings − cost; $X-5$)*

(c) If the game costs \$5 to play, should you play this game? Explain.

2.36 Is it worth it? Andy is always looking for ways to make money fast. Lately, he has been trying to make money by gambling. Here is the game he is considering playing: The game costs \$2 to play. He draws a card from a deck. If he gets a number card (2-10), he wins nothing. For any face card (jack, queen or king), he wins \$3. For any ace, he wins \$5, and he wins an *extra* \$20 if he draws the ace of clubs.

(a) Create a probability model and find Andy's expected profit per game.

(b) Would you recommend this game to Andy as a good way to make money? Explain.

2.37 Portfolio return. A portfolio's value increases by 18% during a financial boom and by 9% during normal times. It decreases by 12% during a recession. What is the expected return on this portfolio if each scenario is equally likely?

2.38 Baggage fees. An airline charges the following baggage fees: \$25 for the first bag and \$35 for the second. Suppose 54% of passengers have no checked luggage, 34% have one piece of checked luggage and 12% have two pieces. We suppose a negligible portion of people check more than two bags.

(a) Build a probability model, compute the average revenue per passenger, and compute the corresponding standard deviation.

(b) About how much revenue should the airline expect for a flight of 120 passengers? With what standard deviation? Note any assumptions you make and if you think they are justified.

2.39 American roulette. The game of American roulette involves spinning a wheel with 38 slots: 18 red, 18 black, and 2 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball. Gamblers can place bets on red or black. If the ball lands on their color, they double their money. If it lands on another color, they lose their money. Suppose you bet \$1 on red. What's the expected value and standard deviation of your winnings?

2.40 European roulette. The game of European roulette involves spinning a wheel with 37 slots: 18 red, 18 black, and 1 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball. Gamblers can place bets on red or black. If the ball lands on their color, they double their money. If it lands on another color, they lose their money.

(a) Suppose you play roulette and bet \$3 on a single round. What is the expected value and standard deviation of your total winnings?

(b) Suppose you bet \$1 in three different rounds. What is the expected value and standard deviation of your total winnings?

(c) How do your answers to parts (a) and (b) compare? What does this say about the riskiness of the two games?

2.41 Cost of breakfast. Sally gets a cup of coffee and a muffin every day for breakfast from one of the many coffee shops in her neighborhood. She picks a coffee shop each morning at random and independently of previous days. The average price of a cup of coffee is \$1.40 with a standard deviation of 30¢(\$0.30), the average price of a muffin is \$2.50 with a standard deviation of 15¢, and the two prices are independent of each other.

(a) What is the mean and standard deviation of the amount she spends on breakfast daily?

(b) What is the mean and standard deviation of the amount she spends on breakfast weekly (7 days)?

2.42 Scooping ice cream. Ice cream usually comes in 1. 5 quart boxes (48 fluid ounces), and ice cream scoops hold about 2 ounces. However, there is some variability in the amount of ice cream in a box as well as the amount of ice cream scooped out. We represent the amount of ice cream in the box as X and the amount scooped out as Y. Suppose these random variables have the following means, standard deviations, and variances:

	mean	SD	variance
X	48	1	1
Y	2	0.25	0.0625

(a) An entire box of ice cream, plus 3 scoops from a second box is served at a party. How much ice cream do you expect to have been served at this party? What is the standard deviation of the amount of ice cream served?

(b) How much ice cream would you expect to be left in the box after scooping out one scoop of ice cream? That is, find the expected value of $X - Y$. What is the standard deviation of the amount left in the box?

(c) Using the context of this exercise, explain why we add variances when we subtract one random variable from another.

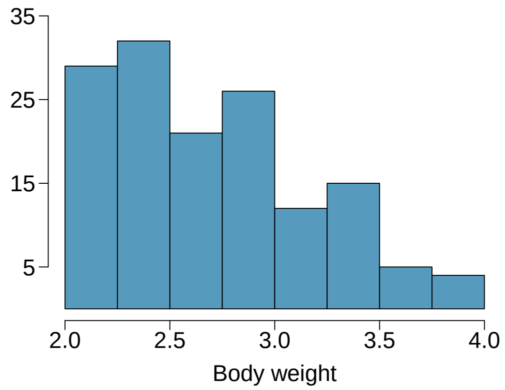
2.3.5 Continuous distributions

2.43 Cat weights. The histogram shown below represents the weights (in kg) of 47 female and 97 male cats.[35]

(a) What fraction of these cats weigh less than 2.5 kg?

(b) What fraction of these cats weigh between 2.5 and 2.75 kg?

(c) What fraction of these cats weigh between 2.75 and 3.5 kg?



[35] **cats**.

2.44 Income and gender. The relative frequency table below displays the distribution of annual total personal income (in 2009 inflation-adjusted dollars) for a representative sample of 96,420,486 Americans. These data come from the American Community Survey for 2005-2009. This sample is comprised of 59% males and 41% females.[36]

(a) Describe the distribution of total personal income.

(b) What is the probability that a randomly chosen US resident makes less than $50,000 per year?

(c) What is the probability that a randomly chosen US resident makes less than $50,000 per year and is female? Note any assumptions you make.

(d) The same data source indicates that 71.8% of females make less than $50,000 per year. Use this value to determine whether or not the assumption you made in part (c) is valid.

Income	*Total*
$1 to $9,999 or loss	2.2%
$10,000 to $14,999	4.7%
$15,000 to $24,999	15.8%
$25,000 to $34,999	18.3%
$35,000 to $49,999	21.2%
$50,000 to $64,999	13.9%
$65,000 to $74,999	5.8%
$75,000 to $99,999	8.4%
$100,000 or more	9.7%

[36] **acsIncome2005-2009**.

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

[1] http://www.cdc.gov/obesity/

[2] http://apps.who.int/bmi/index.jsp?introPage=intro_3.html

the population mean, μ?" Statistical inference includes asking these questions but also determining which estimates to use.

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 4.1.

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 respectively while $\text{weight}_{\text{lb}}$ and $\text{height}_{\text{in}}$ is the weight and height in pounds and inches respectively.

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

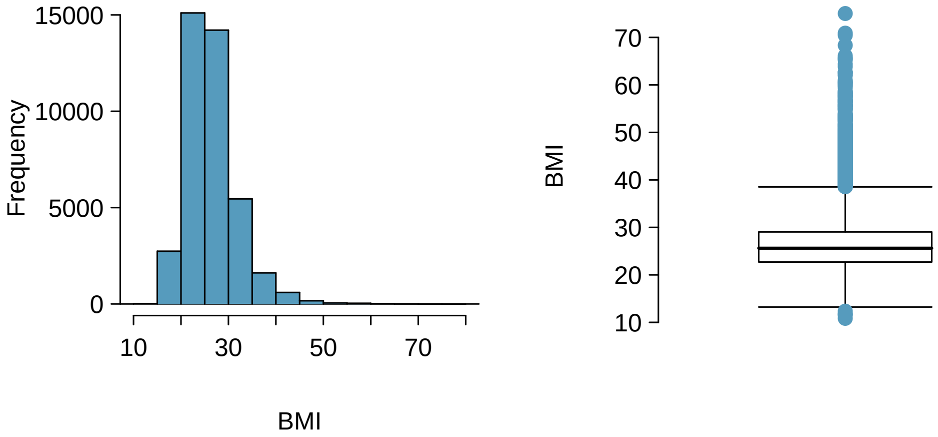


Figure 4.3: Histogram and boxplot of BMI for the BRFSS BMI data. The data is skewed right by both the histogram and the box plot.

Therefore from the BRFSS data, we can use a sample of heights and weights to calculate BMI values for each person in our sample. Thus, each person's BMI value will be considered a single observation. The BRFSS data comprises of 170,000 observations, and the CDC is hoping to infer the characteristics of adults in the United States, 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. [4] 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.

[4] We note that the conductors of the BRFSS data did not use a simple random sample to obtain the data. We instead will assume so for our purposes but sampling methodology can be found http://www.cdc.gov/brfss/annual_data/2000/pdf/overview_00.pdf.

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.[5]

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[6].

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 samples of 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, [7] 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 these parameters based on sample statistics as well. 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.

⊙ **Guided Practice 4.1** Suppose we want to estimate the average age for men and women in the US as well as the difference in ages. 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? [8]

[5] 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.

[6] 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 $\bar{w}$

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

[8] 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.

BMI	estimates
mean	26.356
median	25.620
std. dev.	5.288

Table 4.4: Point estimates for the bmi variable

⊙ **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?[9]

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 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 and increase in accuracy. 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 number of observations that the sample mean is not the same. There exists some sampling variation. Even more interesting, the variation across the different running means decreases as the number of observations increases. We will explore this concept more in Section 4.2.2.

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 values. 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 not equal to the population mean only 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.

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

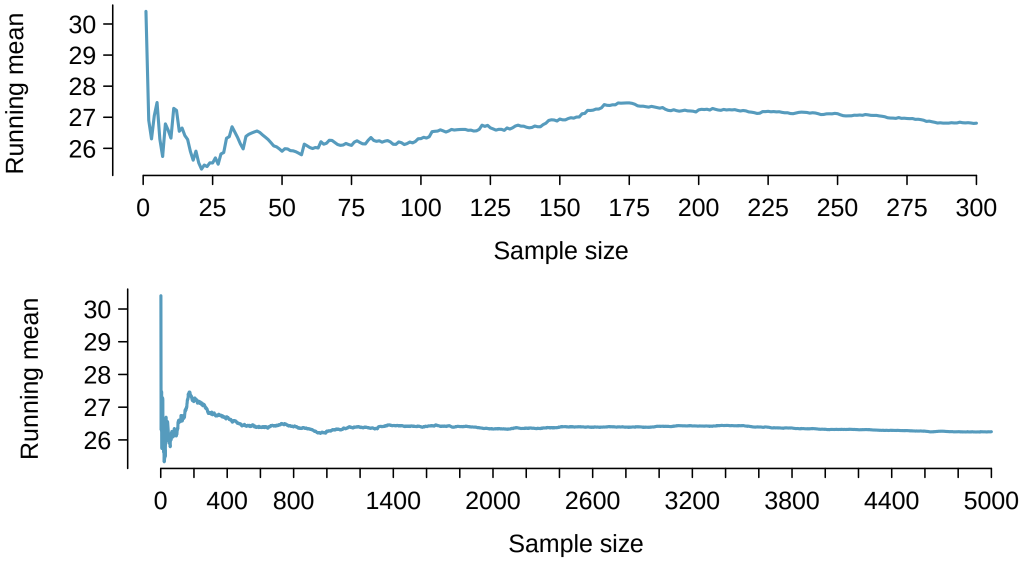


Figure 4.5: The running means from the BRFSS BMI sample of 40,000 observations. The mean stabilizes and approaches the mean of the entire sample $\bar{x} = 26.36$ as the number of observations increases

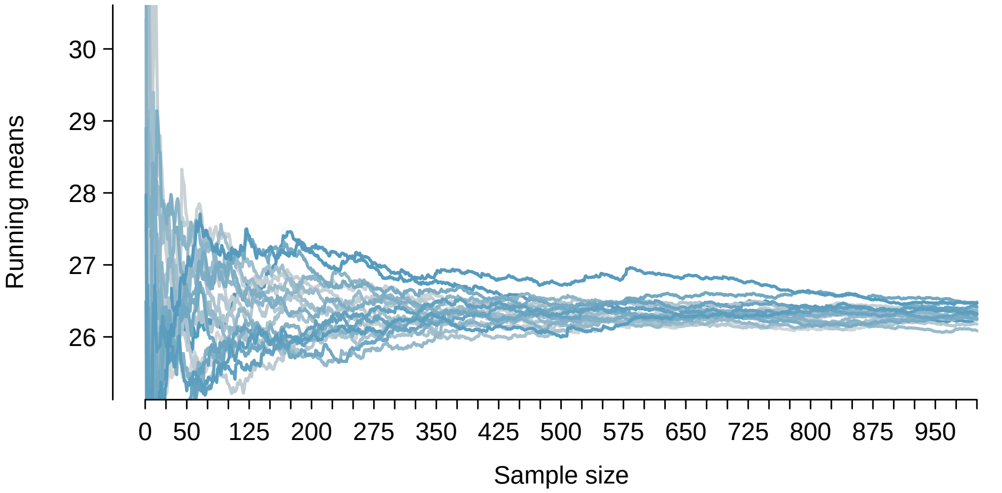


Figure 4.6: Many running means from different samples of 40,000 observations from BRFSS. The sampling variation decreases as the number of observations gets larger.

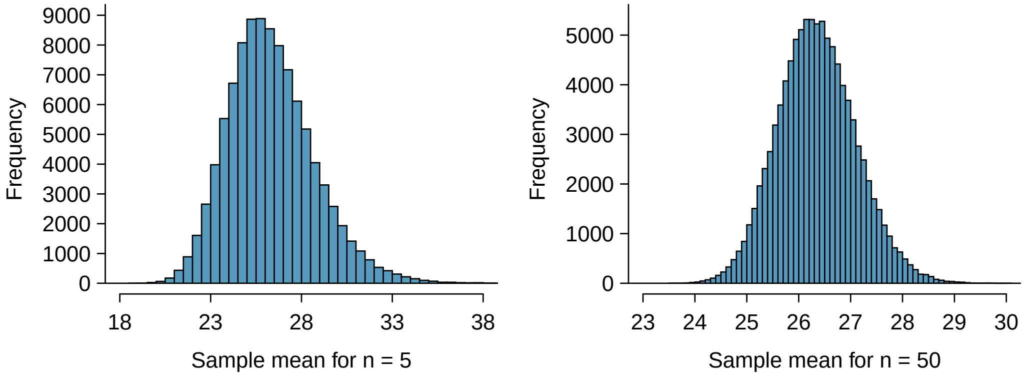


Figure 4.7: Sample means of size $n = 5$ and $n = 50$.

While the sample mean is consistently accurate, it may not always be precise. As the sample size, n, increases, the randomness and variability 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 individuals and take its mean. In Section 4.2.1, we then get 26.344. 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. [10] We can then build up a **sampling distribution** for the sample mean when the sample size is 40,000, shown in Figure 4.8.

[10] 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 to 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 are not even capable of taking another sample of 40,000 from BRFSS!

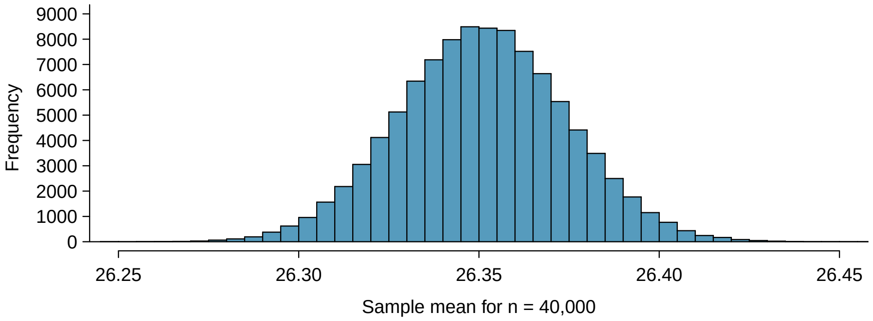


Figure 4.8: A histogram of 100,000 sample means for BMI, where the samples are of size $n = 40,000$.

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 is 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 [11]:

```
(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) 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 of the population that we are drawing from.

From the sampling distribution, we also see that the point estimator will have some variability, and from the concept of sampling distribution introduced in Section 4.2.2,

[11] Refer to the appendix for R code

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 more data is observed in the 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. [12]

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. [13]

SE
standard
error

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?[14]

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
```

[12] 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.

[13] 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.

[14] (a) Prefer a large sample. 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), intuition suggests that a larger sample size corresponds to a smaller standard error.

```
(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:[15]

```
> 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
```

This method, however, has one main issue. Practitioners rarely are able to repeatedly sample from the 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 population standard deviation from the sample collected. 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

[15] 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.

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

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? [16]

⊙ **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 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).[17]

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 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 using *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.

[16] (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 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" of 30 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 (standard error of 0.0267) away from the sample mean.

[17] (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 more trustworthy – though it does not guarantee that every large sample will provide a better estimate than a particular small sample.

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.

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?[18]

⊙ **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? [19]

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

[18] 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.

[19] 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.

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.

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.

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 [20]

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

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

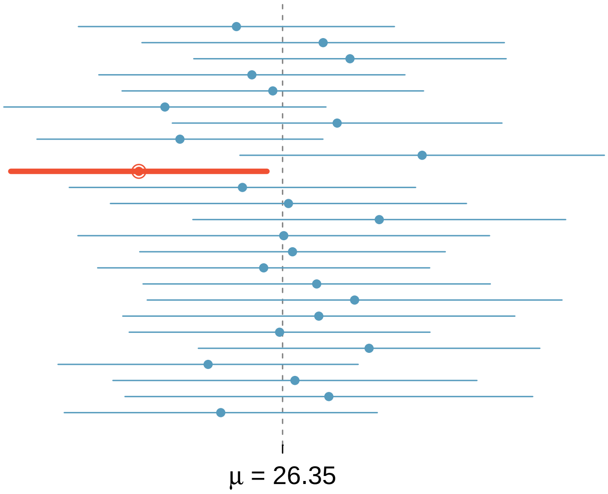


Figure 4.9: Twenty-five samples of size $n = 100$ were taken from the BRFSS data set. For each sample, a confidence interval was created to try to capture the average BMI for the population. Only 1 of these 25 intervals did not capture the true mean.

```
(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) 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`? [21]

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.

Roughly 95% of the time, the estimate that is observed from resampling will be within approximately 2 standard errors[22] 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:

$$\text{point estimate} \ \pm \ 2 \times SE^{23} \tag{4.11}$$

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.

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

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.

[21] (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))`

[22] 1.96 to be more precise if the sampling distribution resembles a Normal Distribution. Details coming up in Section 4.3.5

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? [24]

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

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

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?[25]

[24] 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.

[25] Again apply Equation (4.11): $46.64 \ \pm \ 2 \times 0.09 \rightarrow (46.46, 46.82)$. We interpret this interval as follows: We

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 ĩefsampDistNormal 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.

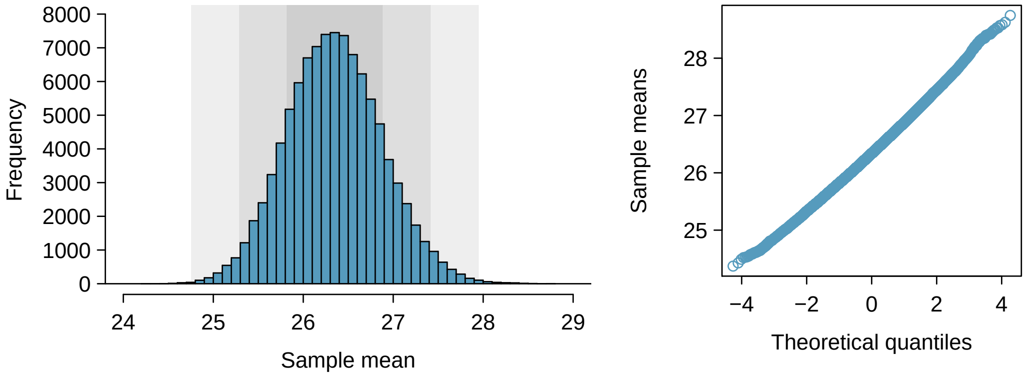


Figure 4.10: The left panel shows a histogram of the sample means for 100,000 different random samples of size $n = 100$. The right panel shows a normal probability plot of those sample means.

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[26].

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

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.

[26] A more formal definition coming soon.

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. [27]

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 $\mathcal{N}(\mu, \sigma)$ where μ is the mean of the sample means and σ is the standard deviation of the sample means. FIX THISSSSSS

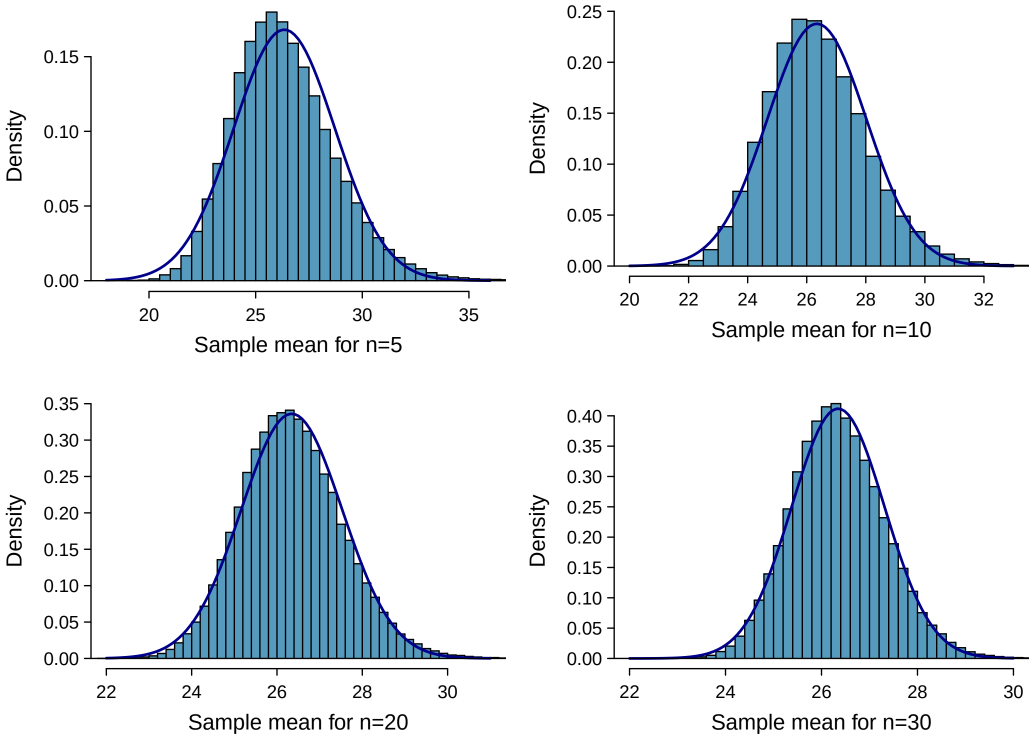


Figure 4.11: The sampling distribution of sample means with different sample sizes $n = 5, 10, 20, 30$. With a normal density curve on top, we see that for $n = 30$, a normal model is a fitting approximation confirming the Central Limit Theorem rule of thumb.

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

[27] 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.

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 ($n \geq 30$), we can make this 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

6 Changing the confidence level ? Motivation: Sometimes your problem parameters don?t need you to be 95example with life threatening tests) you need to be more confident. Solution: you can change your confidence level.

We have only discussed confidence intervals at the 95% confidence level. Some situations, however, require conclusions to be more than 95% confident. Consider a test for diagonosing a deathly disease in which doctors need to be more accurate. Perhaps we would like a confidence level of 99%. There are also some scenarios where the problem parameters only require you to be 90%

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.

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

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?[28]

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

[28] 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.

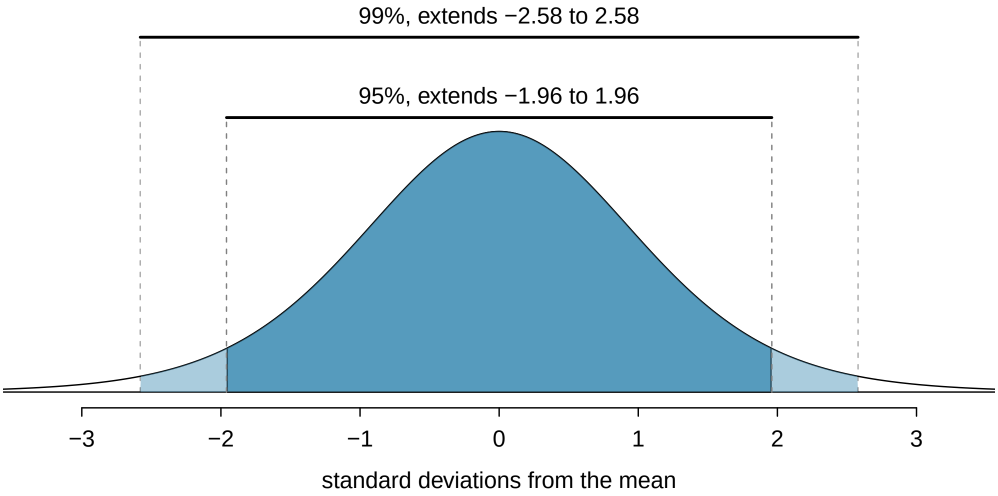


Figure 4.12: The area between -$z^\star$ and $z^\star$ increases as $|z^\star|$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star =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

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 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.

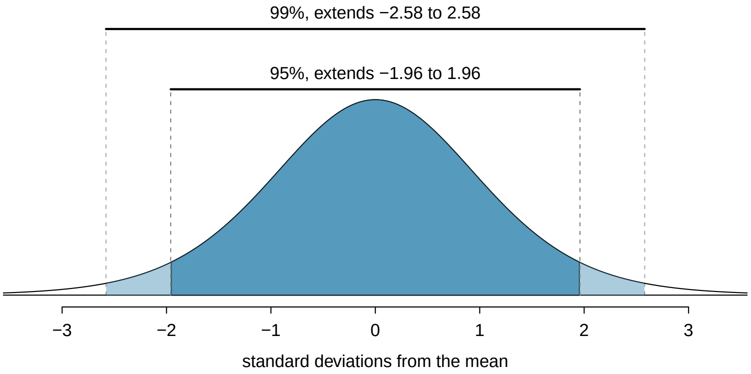


Figure 4.13: We draw a histogram of the men's weights in the `brfss.sample` and note that it is only slightly skewed. With 40,000 observations however, its skewness is more negligible because of its large sample size.

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. [29]

[29] The observations are independent (simple random sample, < 10% of the population), the sample size is at

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

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 [30] (**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%? [31]

⊙ **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.[32]

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: $\bar{y} \pm 2.58 \times SE_{\bar{y}} \rightarrow (188.94, 189.87)$. We are 99% confident that the average weight of all males is between 188.94 and 189.87 pounds.

[30] also known as a Normal table

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

[32] 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

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.

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

other 5% is in the upper tail), thus $z^\star = 1.65$. The 90% confidence interval can then be computed as $\bar{y} \pm 1.65 \times SE_{\bar{y}} \to (189.11, 189.69)$. (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.

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.[33] If we randomly sampled the heights of five male seniors, what distribution should the sample mean follow?

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.

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

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.[34] 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:

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.[35] 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.

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

[35] 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.

We call H_0 the null hypothesis and H_A the 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.

H_0
null hypothesis

H_A
alternative
hypothesis

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.

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? [36]

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.

[36] 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).

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.

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

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.

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 [37] and will have $n-1$ degrees of freedom.

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

[37] from **Chapter 3**

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 $\mu > \mu_0$.

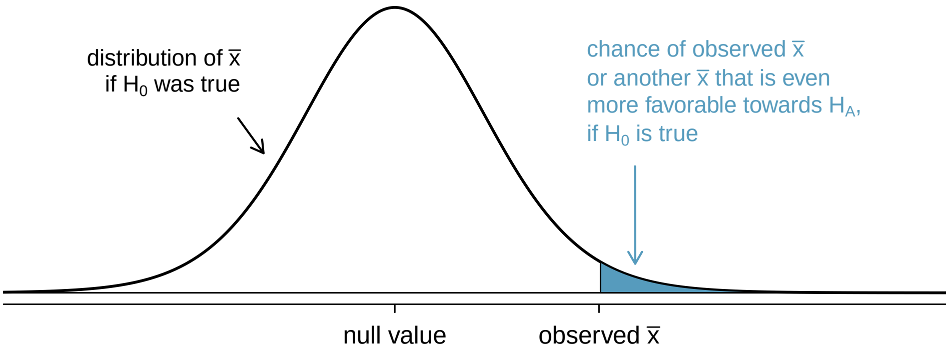


Figure 4.14: To identify the p-value, the distribution of the sample mean is considered as if the null hypothesis was true. Then the p-value is defined and computed as the probability of observing the observed $\bar{x}$ or an $\bar{x}$ even more extreme and thus favorable to follow H_A under this distribution.

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

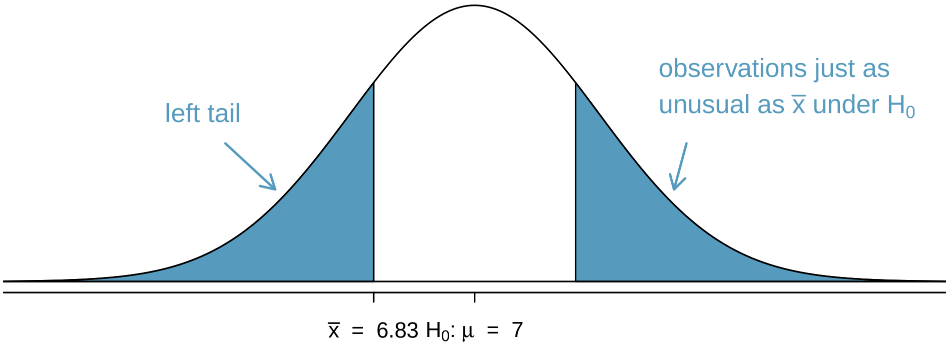


Figure 4.15: H_A is two-sided, so *both* tails must be counted for the p-value.

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[38].

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

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!

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

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

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

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?[39]

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.

[39] 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.

Caution: Critical value ≠ 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?[40]

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

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

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.[41] 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.

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

[41] 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 $\mu \leq 7$.

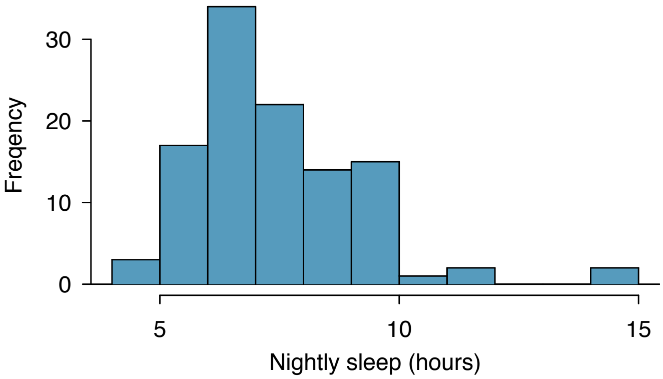


Figure 4.16: Distribution of a night of sleep for 110 college students. These data are moderately skewed.

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 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 $\bar{x}$.[42]

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.

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
```

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

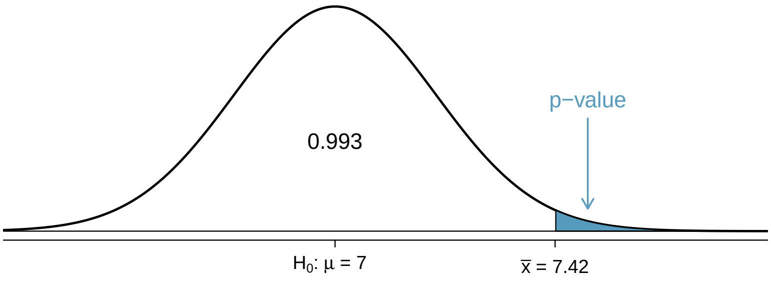


Figure 4.17: If the null hypothesis is true, then the sample mean $\bar{x}$ came from this nearly normal distribution. The right tail describes the probability of observing such a large sample mean if the null hypothesis is true.

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.[43] 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.

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 $\alpha = 0.001$? [44]

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

[43] 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

[44] 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 $\alpha = 0.001$.

data to test if the average age of adults is 36.8 years [45]. 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? [46]

● **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.

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

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.

[45] as calculated by the US Census in 2009

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

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.[47]

⊙ **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?[48]

⊙ **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?[49]

⊙ **Guided Practice 4.33** Consider a person getting tested for HIV. What does a Type 1 and Type 2 Error represent in this context? [50]

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

[47] 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).

[48] 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.

[49] 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.

[50] 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.

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. [51]

⊙ **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.[52]

● **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:

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

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

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

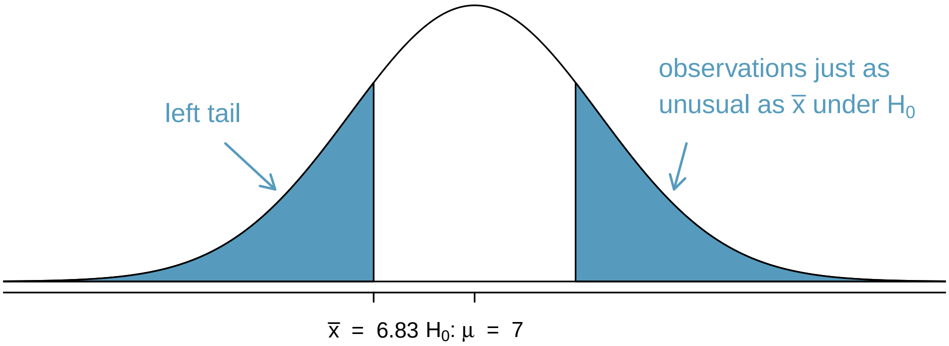


Figure 4.19: H_A is two-sided, so *both* tails must be counted for the p-value.

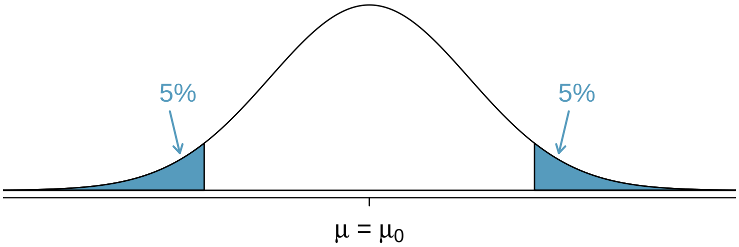


Figure 4.20: The shaded regions represent areas where we would reject H_0 under the bad practices considered in Example 4.4.5 when $\alpha = 0.05$.

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 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.[53]

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 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.

[53] 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 $\alpha = 0.01$.

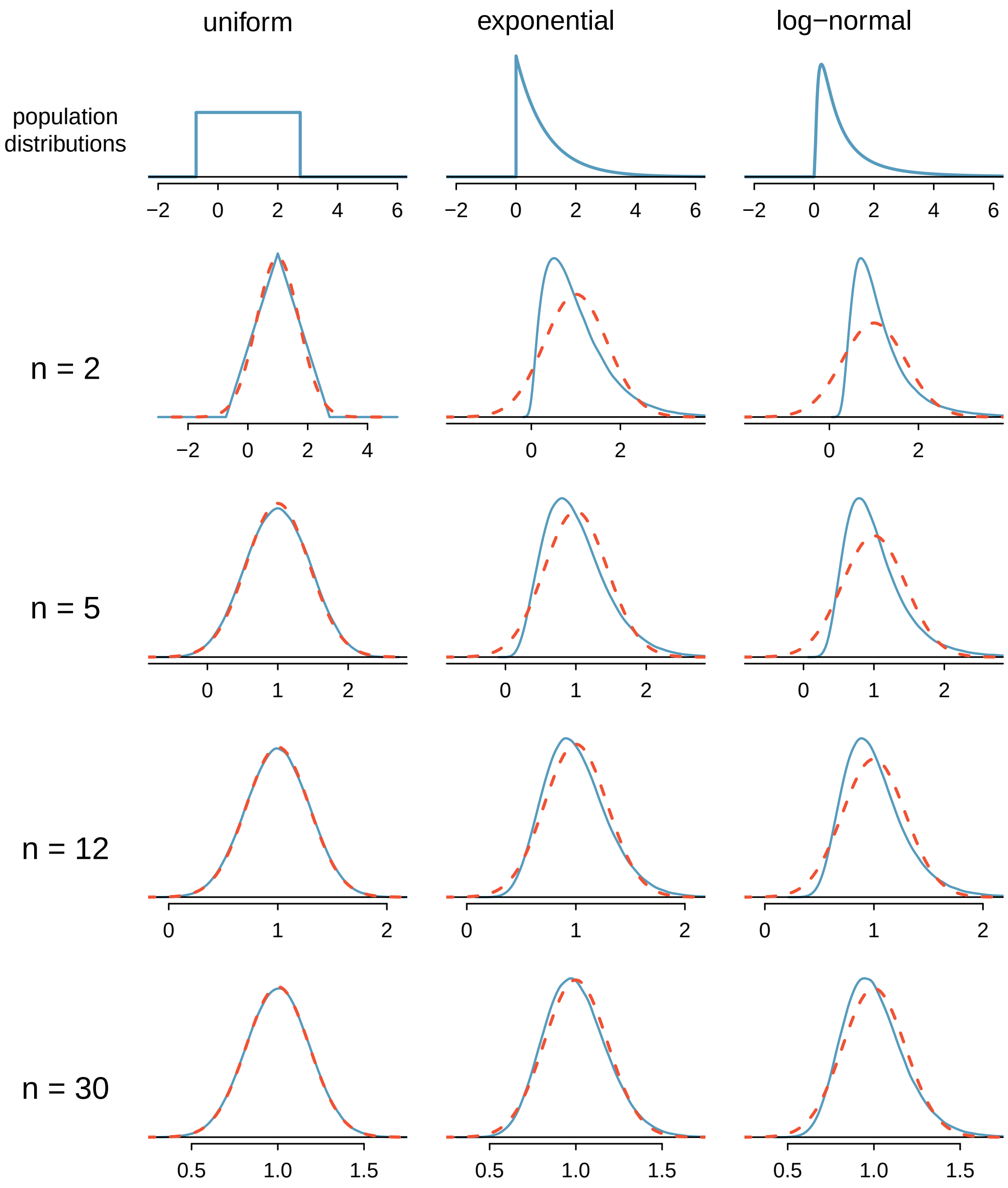


Figure 4.21: Sampling distributions for the mean at different sample sizes and for three different distributions. The dashed red lines show normal distributions.

⊙ **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?[54]

● **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 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.

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

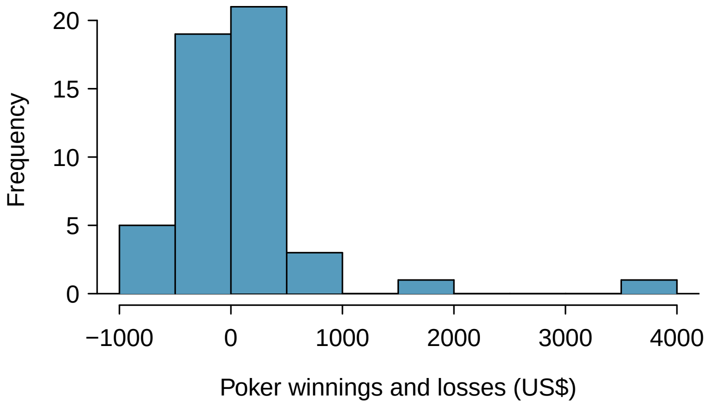


Figure 4.22: Sample distribution of total patient visits net of 5,000 visits. These data include some very clear outliers. These are problematic when considering the normality of the sample mean. For example, outliers are often an indicator of very strong skew.

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):

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

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.[55]

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

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

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

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

group that received the new drug. What would be an appropriate null hypothesis? And the alternative?[56]

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

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

or equivalently,

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

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

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

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.[57] 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.

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

[56] 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).

[57] 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.

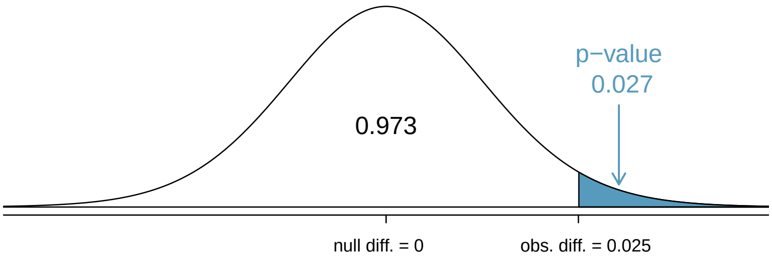


Figure 4.23: The distribution of the sample difference if the null hypothesis is true.

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.

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.[58] 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 [59]. 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.

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 [60] The margin of error for a 95% confidence interval estimating a mean can be written as

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

[58] 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.

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

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

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 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:

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

(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:

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

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

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.

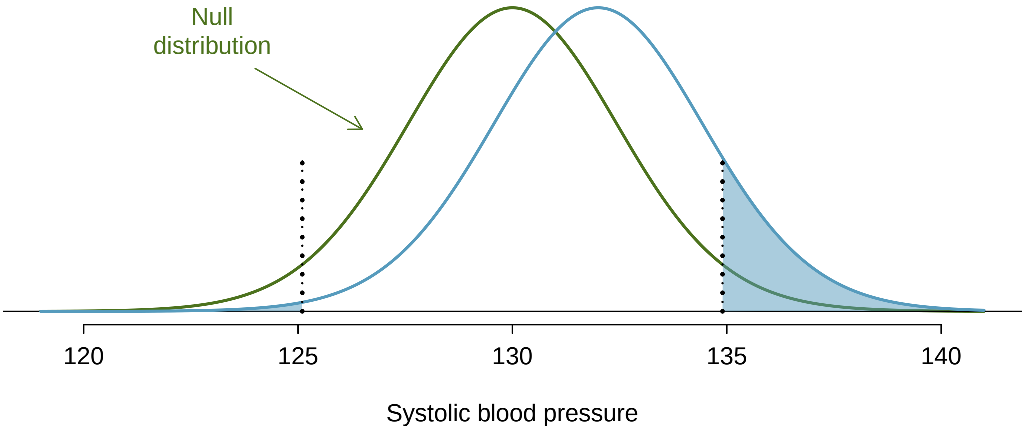


Figure 4.24: The sampling distribution of $\bar{x}$ under two scenarios. Left: $N(130, 2.5)$. Right: $N(132, 2.5)$, and the shaded areas in this distribution represent the power of the test.

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.[61]

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?[62]

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.[63]

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.

[61] 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.

[62] 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 $1 - 0.979 = 0.021$.

[63] 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.

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.

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.

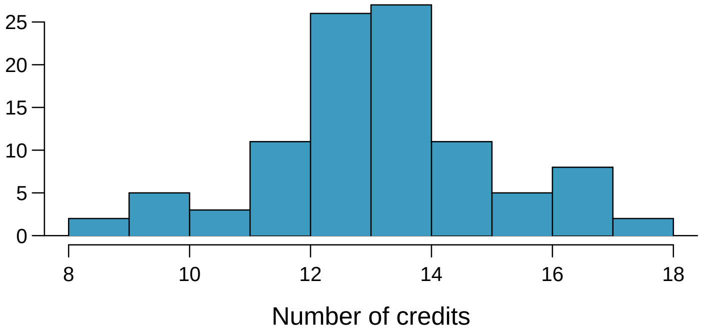
(b) A survey reports that local TV news has shown a 17% increase in revenue between 2009 and 2011 while newspaper revenues decreased by 6.4% during this time period.

(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?

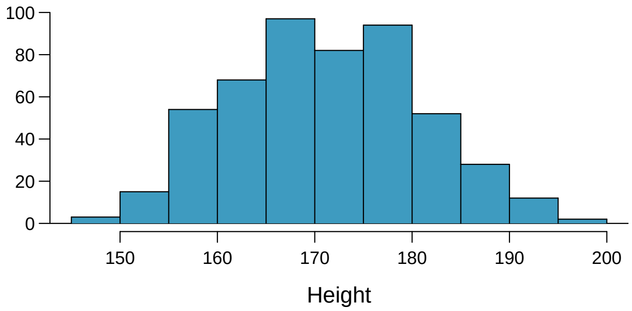
(b) What is the point estimate for the standard deviation of the number of credits taken per semester by students at this college? What about the IQR?

(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.[64]



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?

[64] **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.[65]

(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.

(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.

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

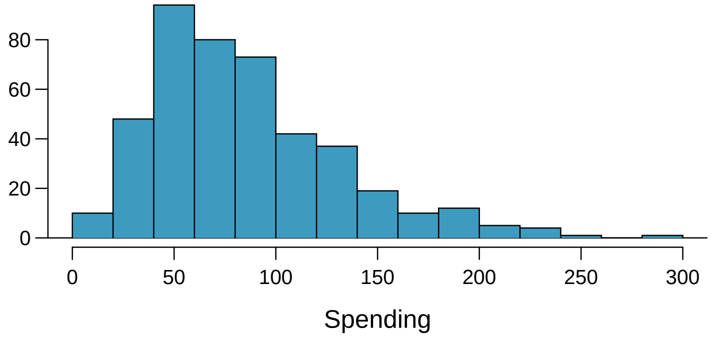
(d) 95% of such random samples would have a sample mean between 128 and 147 minutes.

(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.

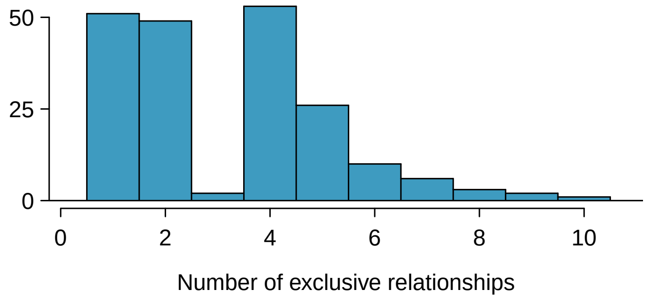
(d) We are 95% confident that the average spending of all American adults is between \$80.31 and \$89.11.

(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.

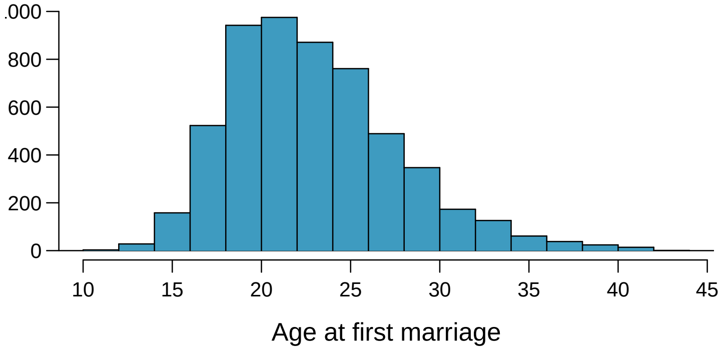
(g) The margin of error for the reported interval is 4.4.

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.[66]



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.

[66] **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?[67]

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?

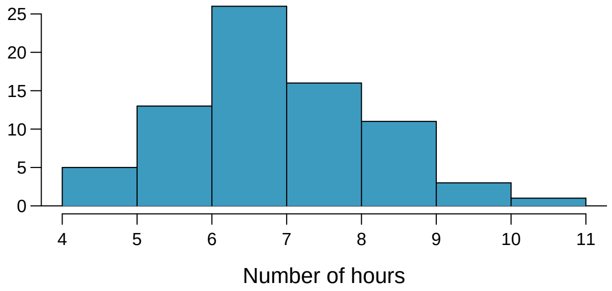
4.20 Thanksgiving spending, Part II. Exercise 4.12 provides a 95% confidence interval for the average spending by American adults during the six-day period after Thanksgiving 2009: ($80.31, $89.11).

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

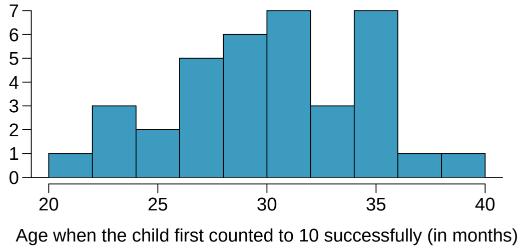
(b) Would the news anchor's claim be considered reasonable based on a 90% confidence interval? Why or why not?

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.

[67] **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.[68]



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 $\alpha = 0.01$?

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

[68] **Graybill:1994**.

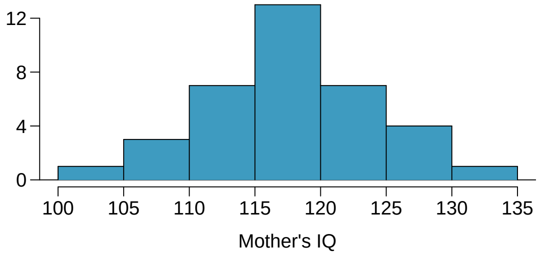
IQ of the 36 randomly sampled gifted children. The histogram below shows the distribution of mother's IQ. Also provided are some sample statistics.

(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.

n	36
min	101
mean	118.2
sd	6.5
max	131



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.

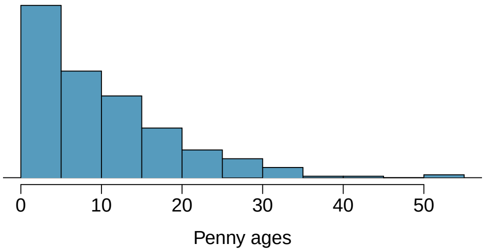
(d) Do your results from the hypothesis test and the confidence interval agree? Explain.

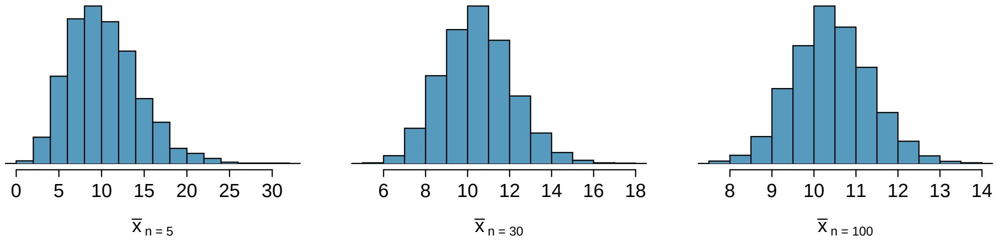
4.8.4 Examining the Central Limit Theorem

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

(a) Describe the distribution.

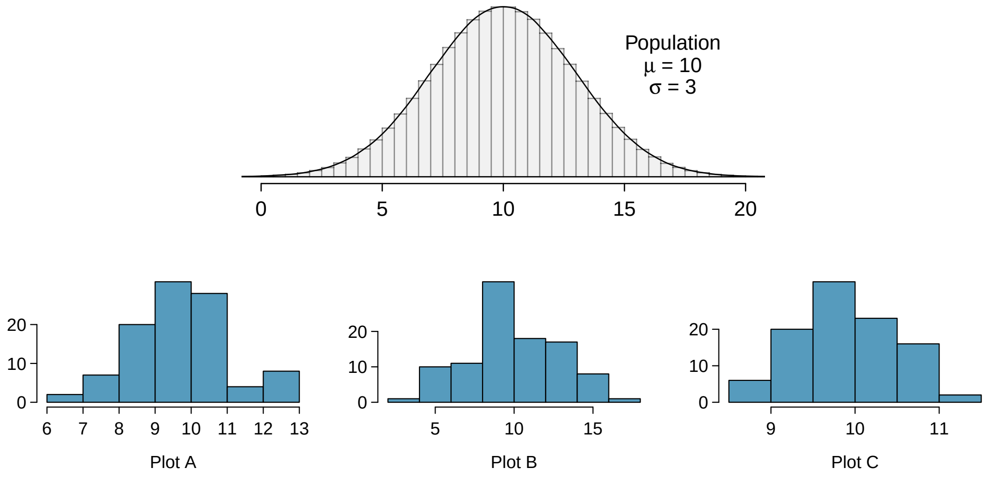
(b) Sampling distributions for means from simple random samples of 5, 30, and 100 pennies is shown in the histograms below. Describe the shapes of these distributions and comment on whether they look like what you would expect to see based on the Central Limit Theorem.



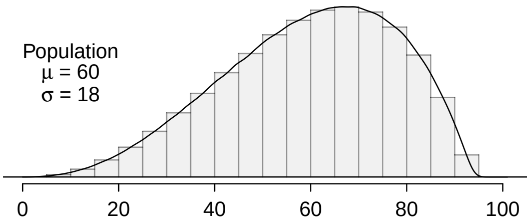


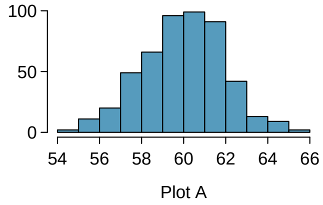
4.34 Ages of pennies, Part II. The mean age of the pennies from Exercise 4.33 is 10.44 years with a standard deviation of 9.2 years. Using the Central Limit Theorem, calculate the means and standard deviations of the distribution of the mean from random samples of size 5, 30, and 100. Comment on whether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

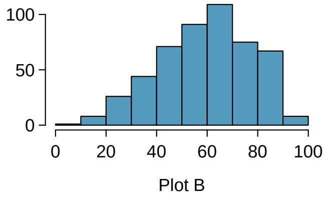
4.35 Identify distributions, Part I. Four plots are presented below. The plot at the top is a distribution for a population. The mean is 10 and the standard deviation is 3. Also shown below is a distribution of (1) a single random sample of 100 values from this population, (2) a distribution of 100 sample 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.

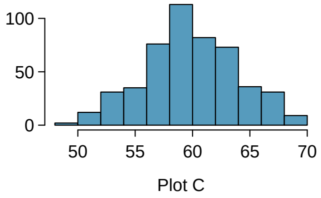


4.36 Identify distributions, Part II. Four plots are presented below. The plot at the top is a distribution for a population. The mean is 60 and the standard deviation is 18. Also shown below is a distribution of (1) a single random sample of 500 values from this population, (2) a distribution of 500 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?

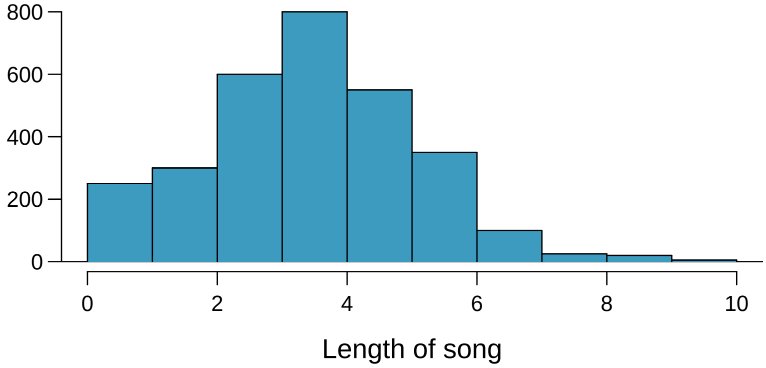
(b) Describe the distribution of the mean lifespan of 15 light bulbs.

(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.[69]

(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% 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.

[69] **webpage:spam**.

(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.