Introductory Statistics for the Life and Biomedical Sciences

Derivative of
OpenIntro Statistics
Third Edition

Original Authors

David M Diez
Christopher D Barr
Mine Çetinkaya-Rundel

Contributing Authors

David Harrington
[Briefly Describe Contribution]

Julie Vu
[Briefly Describe Contribution]

Alice Zhao
[Briefly Describe Contribution]

Contents

Preface

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

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

We have tried to balance the sometimes competing demands of mastering the 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.

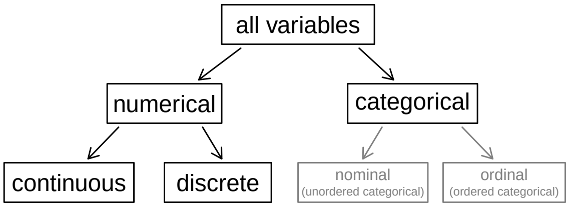


Figure 1.7: Breakdown of variables into their respective types.

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

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

[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?

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

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

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

[13] http://www.cdc.gov/brfss/
[14] http://www.who.int/healthinfo/survey/en/
[15] https://www.ncjrs.gov/pdffiles1/nij/182369.pdf

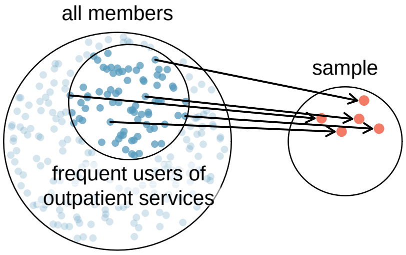


Figure 1.8: Instead of sampling from all members equally, approaching members visiting a clinic during a particular week would disproportionately select members who typically use outpatient services.

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.

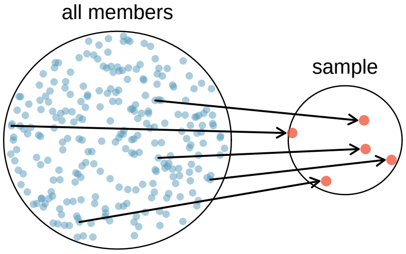


Figure 1.9: In this graphic, five members are randomly selected from the population to be interviewed.

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.

⊙ 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

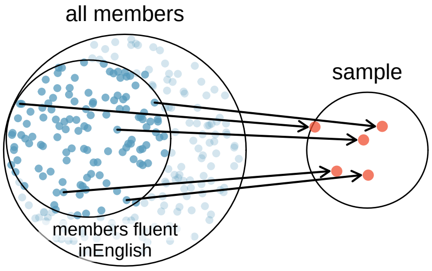


Figure 1.10: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. For example, a survey written in English may only result in responses from health plan members fluent in English.

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 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 individu-

[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.

als. 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** 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

[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**.

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

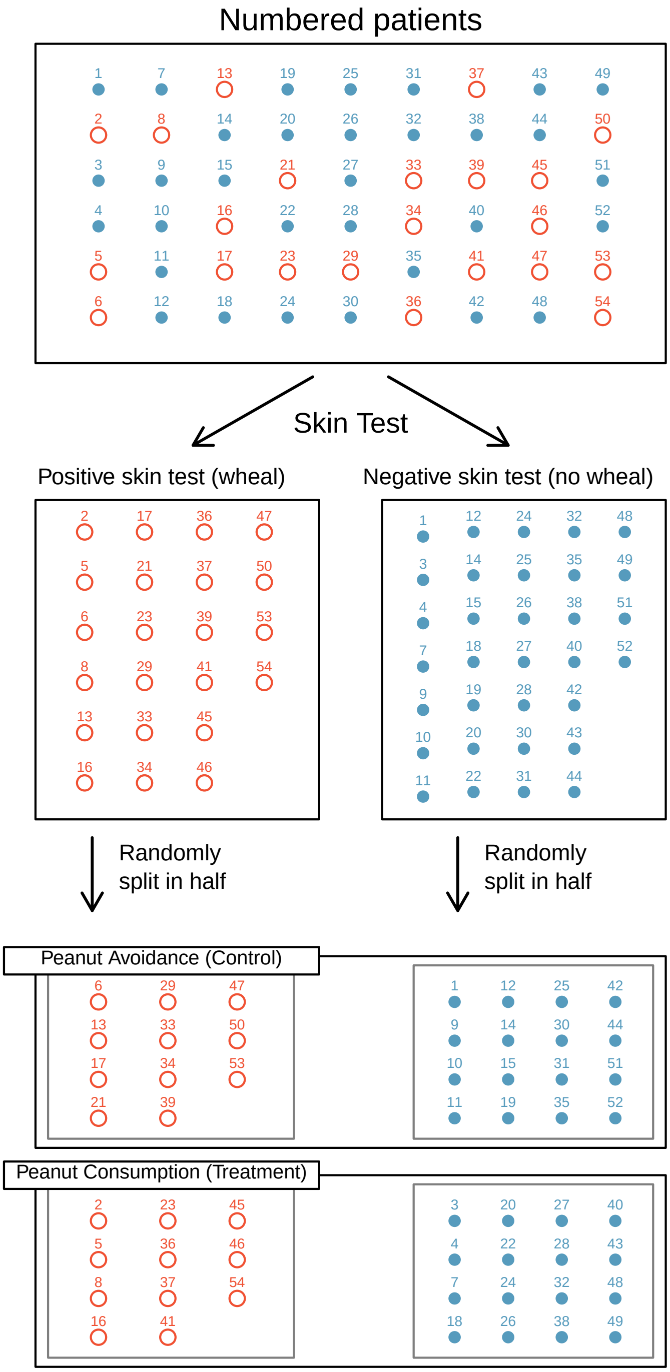


Figure 1.11: A simplified schematic of the blocking scheme used in the LEAP study, depicting 54 patients that underwent randomization. Patients are first divided into groups based on response to the initial skin test, then each block is evenly separated into the treatment groups using randomization.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

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.

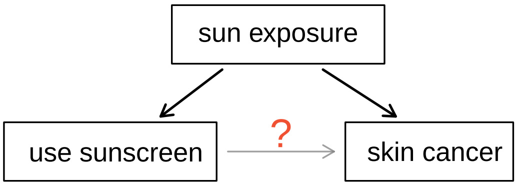
DH: This is another instance where absolute statements are risky. The data on smoking and lung cancer are all observational, or were until a short time ago. I wonder if we should mention that. JV: I have adjusted the language in the first paragraph. Scroll down past the sunscreen example to see where 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**:

[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.

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. researchers 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

[22] Also called a **lurking variable**, **confounding factor**, or a **confounder**.
[23] www.channing.harvard.edu/nhs

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.

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

$\bar{x}$
sample
mean

The sample mean is often labeled $\bar{x}$. 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:

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

where $x_1, x_2, \ldots, x_n$ 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

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

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 $x_1, x_2, \dots, x_n$ 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?

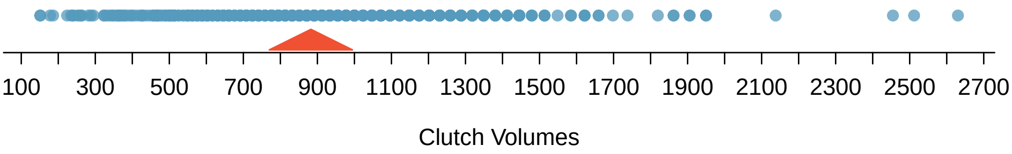


Figure 1.12: Dot plot of the clutch volume variable in the `frog` data.

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 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.

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.

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

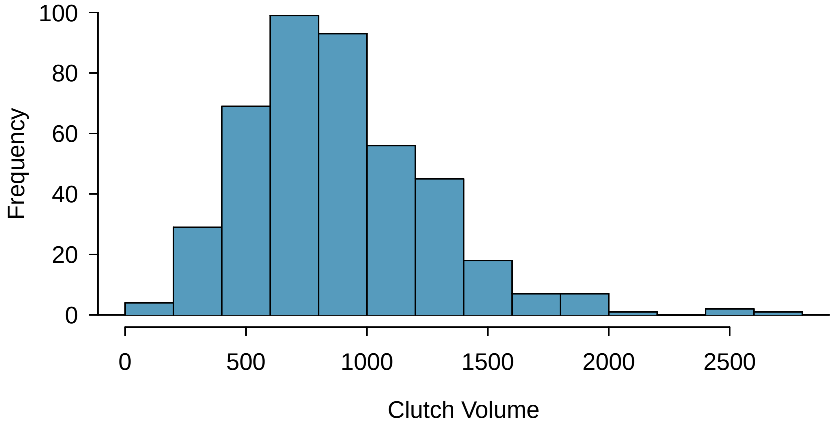


Figure 1.15: A histogram of `clutch.volume`.

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 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?*

[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.

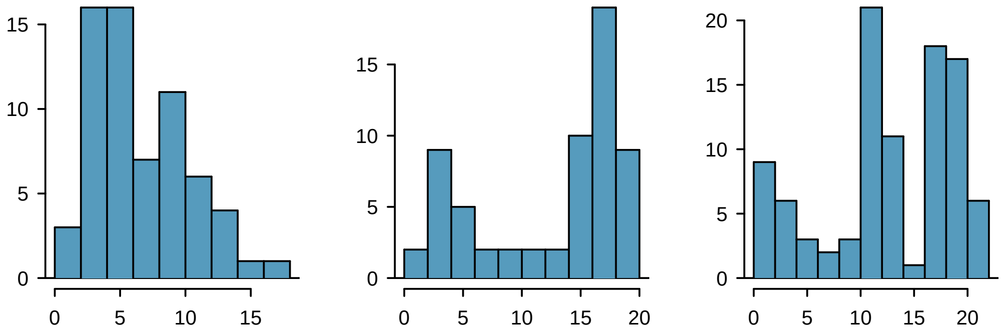


Figure 1.16: From left to right: unimodal, bimodal, and multimodal distributions.

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 $Q_3 + 1.5 \times IQR$.

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

[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.

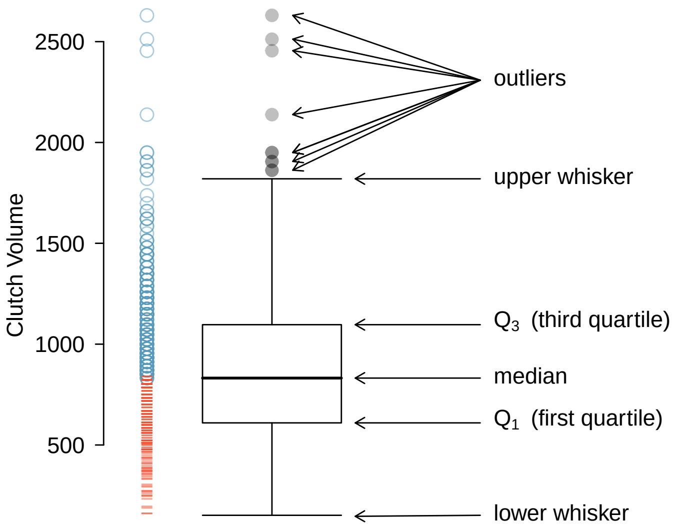


Figure 1.17: A vertical dot plot next to a labeled boxplot for the volumes of 431 egg clutches. The median (831.8 mm^3), splits the data into the bottom 50% and the top 50%, marked in the dot plot by horizontal dashes and open circles, respectively.

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 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 pro-

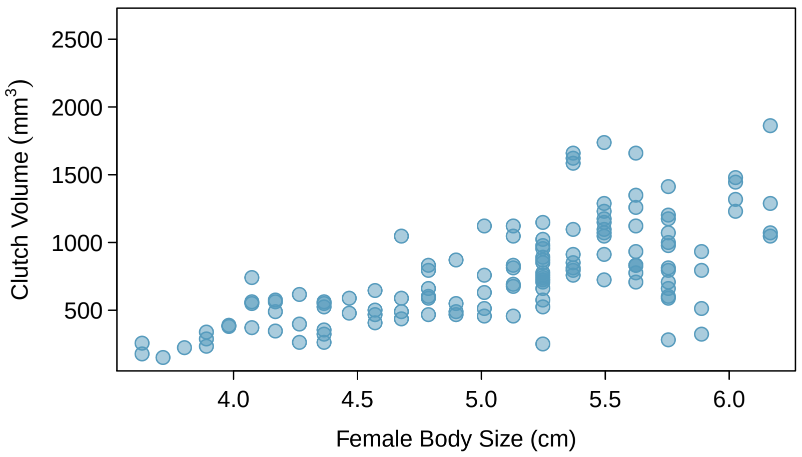


Figure 1.18: A scatterplot showing `clutch.volume` (horizontal axis) vs. `body.size` (vertical axis).

duce 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 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

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

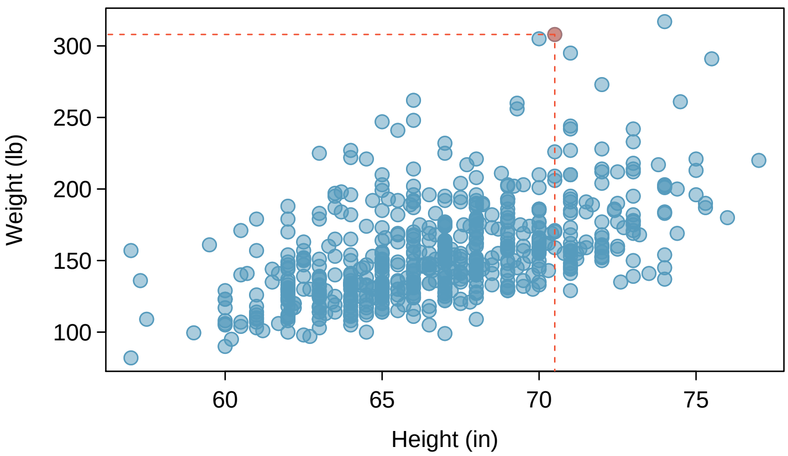


Figure 1.19: A scatterplot showing `height` (horizontal axis) vs. `weight` (vertical axis). One participant 70.5 inches tall and weighing 308 pounds is highlighted.

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 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

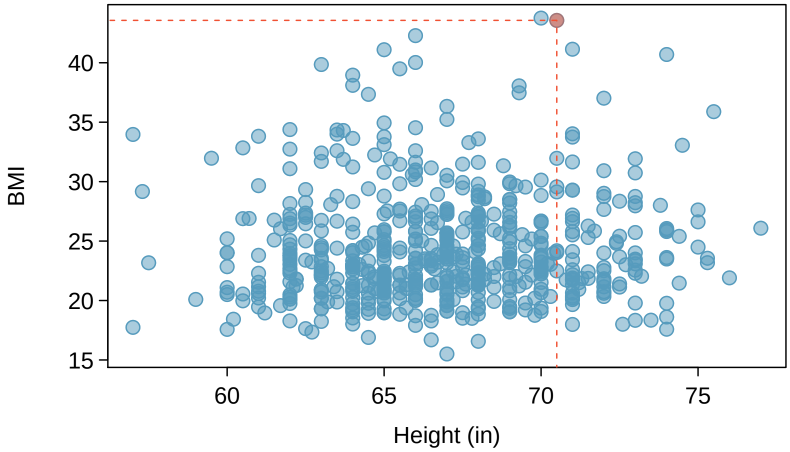
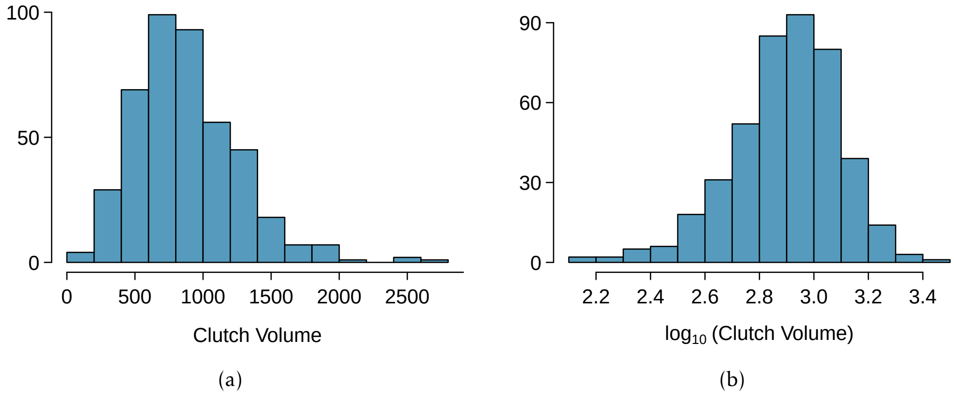


Figure 1.20: A scatterplot showing `height` (horizontal axis) vs. `bmi` (vertical axis). The same individual highlighted in Figure 1.19 is marked here, with BMI 43.56.

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.



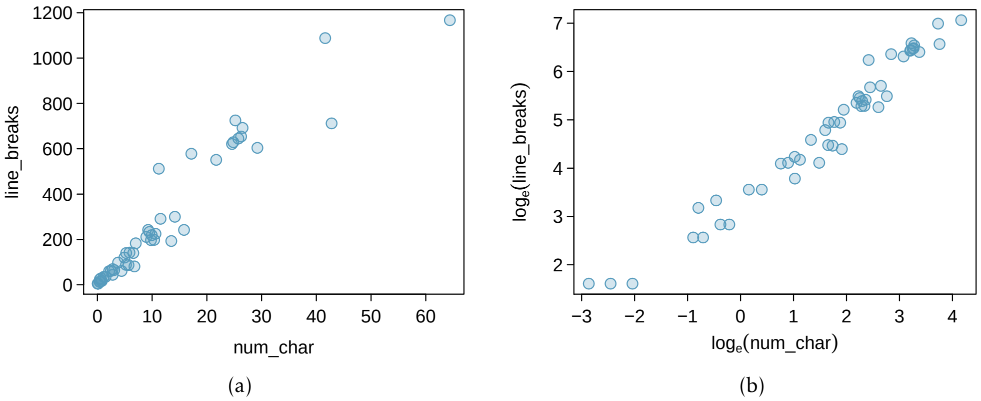
(a)

(b)

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

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



(a)

(b)

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.

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 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.

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

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).

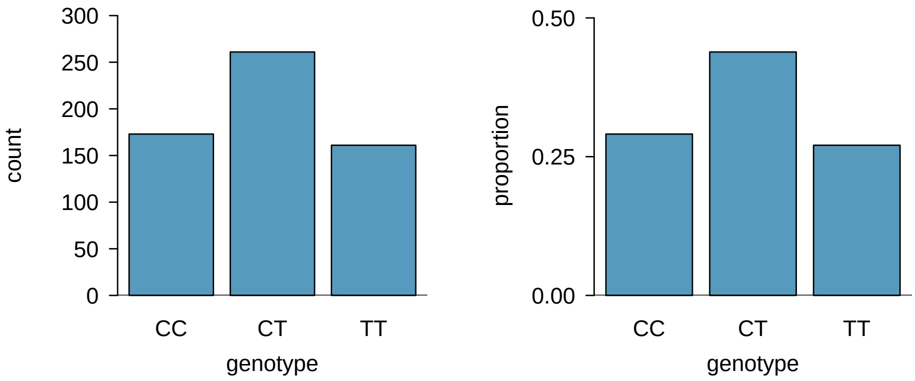


Figure 1.27: Two bar plots of `actn3.r577x`. The left panel shows the counts, and the right panel shows the proportions for each genotype.

[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`.

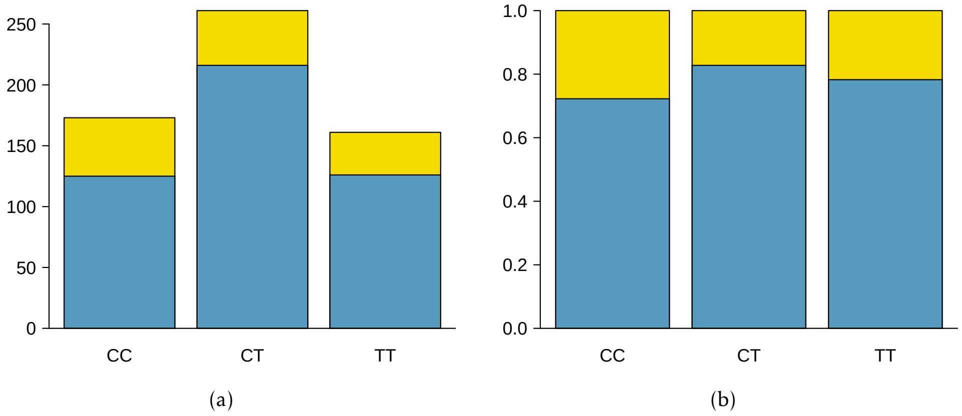


Figure 1.28: (a) Segmented bar plot for individuals by genotype, in which the counts have been further broken down by Caucasian (blue) and not-Caucasian (yellow). (b) Standardized version of Figure (a).

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

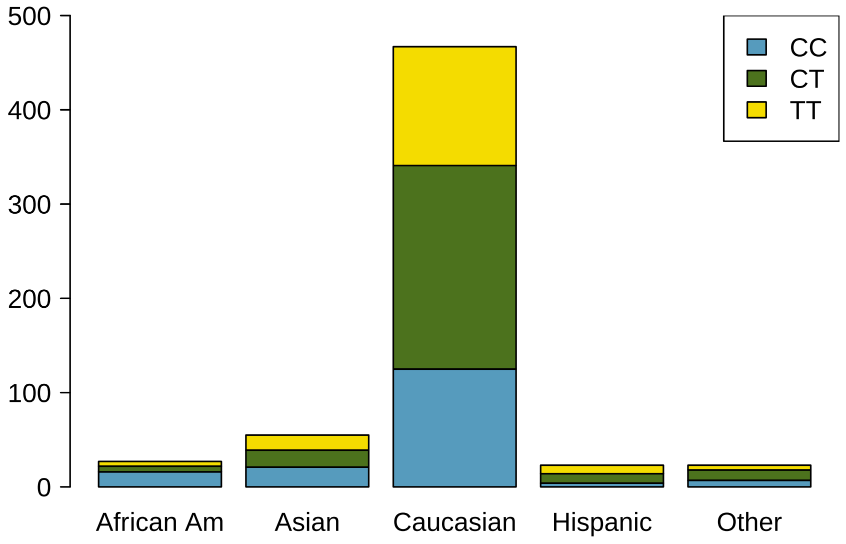


Figure 1.29: Segmented bar plot for individuals by race, where the counts have been further broken down by genotype.

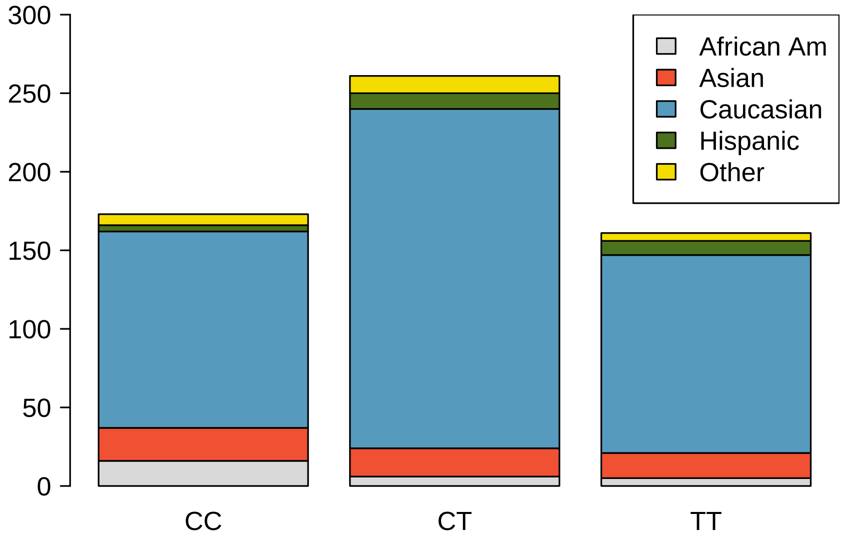


Figure 1.30: Segmented bar plot for individuals by genotype, where the counts have been further broken down by race.

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.

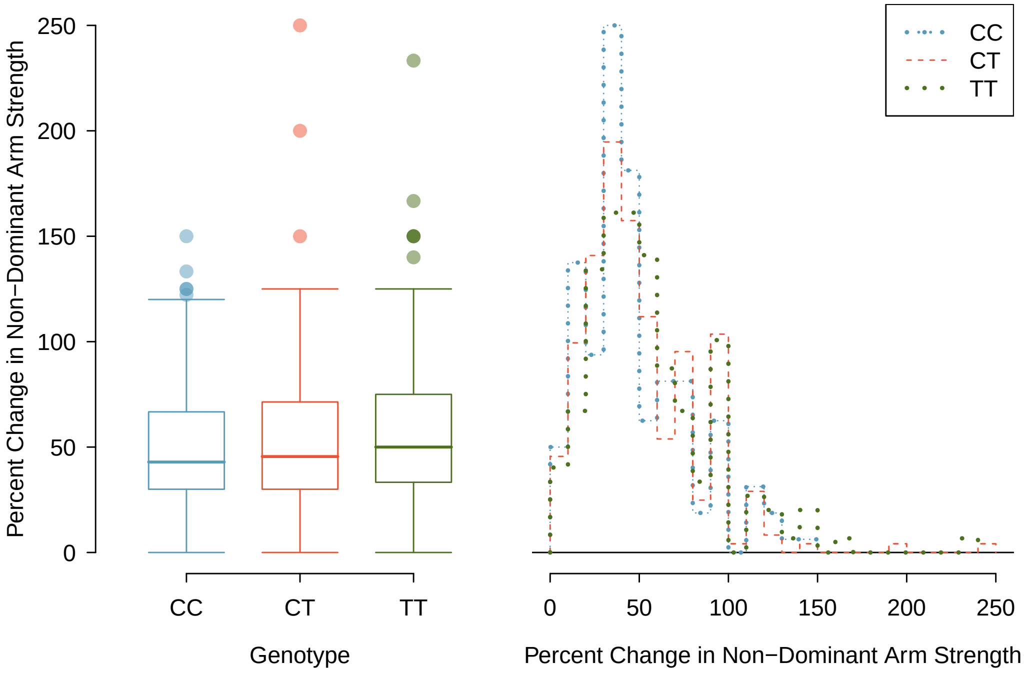


Figure 1.31: Side-by-side boxplot (left panel) and hollow histograms (right panel) for `ndrm.ch`, split by ACTN3 genotype.

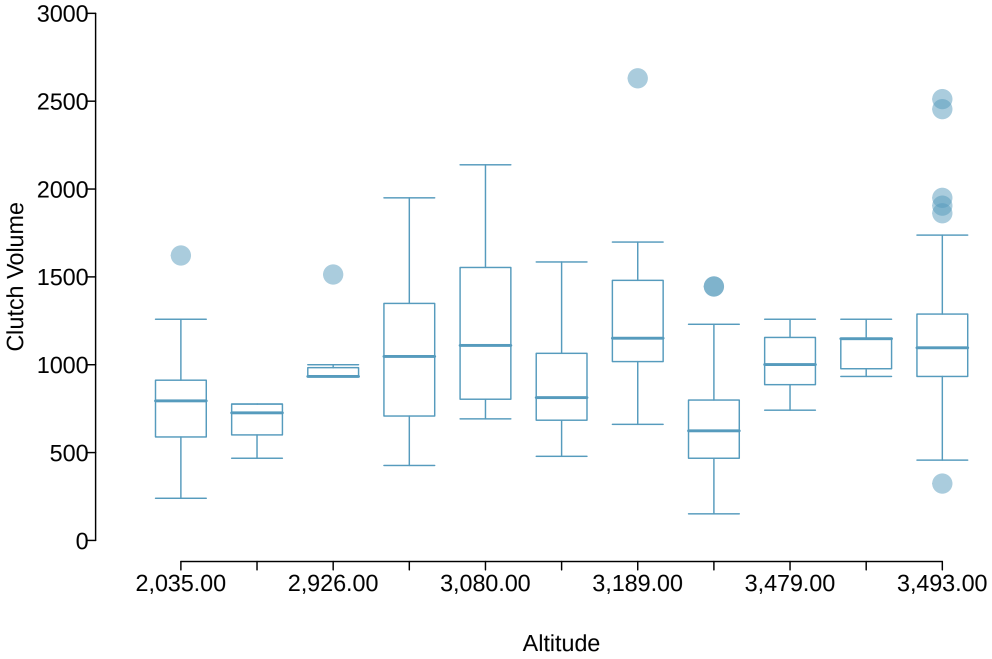


Figure 1.32: Side-by-side boxplot comparing the distribution of `clutch.volume` for different altitudes.

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.

Index