Introductory Statistics for the Life and Biomedical Sciences

Derivative of
OpenIntro Statistics
Third Edition

Original Authors

David M Diez
Christopher D Barr
Mine Çetinkaya-Rundel

Contributing Authors

David Harrington
[Briefly Describe Contribution]

Julie Vu
[Briefly Describe Contribution]

Alice Zhao
[Briefly Describe Contribution]

Contents

Preface

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

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

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

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

Textbook overview

The chapters of this book are as follows:

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

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

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

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

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

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

7. **Introduction to linear regression.** An introduction to regression with two variables. Most of this chapter could be covered after Chapter 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 best to collect, analyze, and draw conclusions from data. It is helpful to put 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 best can we collect data? How should it be analyzed? And what can we infer from the analysis?

Many scientific investigations can be conducted with a few important data collection techniques and analytic tools. This chapter provides a brief introduction to the basic principles of these areas that will be encountered later in the book, and illustrates the important role statistics plays in medicine and biology.

1.1 Case study: Preventing Peanut Allergies

Section 1.1 introduces an important problem in medicine: evaluating the effect of an intervention. Terms in this section, and indeed much of this chapter, will all be revisited later in more detail.

The proportion of young children with peanut allergies in Western countries has doubled in the last 10 years. Does the exposure to peanut products during the first 5 years of a child's life reduced the probability that a child will develop an allergy? This section describes an experiment (a clinical trial, in the terminology of medical research) designed to assess the effectiveness of exposing infants at risk for peanut allergy either to consume or avoid peanut products during the first 5 years of life. The study was called the "Learning Early about Peanut Allergy" (LEAP), enrolled children in the United Kingdom between 2006 and 2009, and was reported in the New England Journal of Medicine in 2015. .[1] Earlier research had suggested that infants predisposed to peanut allergies might develop resistance to the allergy with exposure to peanut products before the allergy appeared.

The study team selected 640 infants with either or both of excema and egg allergies and randomly assigned each child to peanut consumption (the treatment group) or avoidance (the control group) for five years. In this study, the control group provides a reference point for estimating the effect of peanut exposure in the treatment group. Each child was tested for a peanut allergy at age 5 using an oral food challenge (OFC); the main analysis was based on 530 children with a negative skin test at the time of study entry. Among these 530 children, 263 were assigned to 'Peanut Avoidance' and 267 to 'Peanut Consumption'. The outcome at 5 years was coded as either 'Fail OFC' (allergic reaction) or 'Pass OFC' (no allergic reaction. The dataset `LEAP` contains the treatment and outcome data the 530 children.

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

Summary tables are generally more helpful than individual participant listings when

[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.

	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.

looking for patterns in data. Table 1.2 shows outcomes grouped by treatment group and the result of the OFC test.

	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

The table makes it possible to compute some simple summary statistics. A **summary statistic** is a single number summarizing a large amount of data.[2] In the Peanut Avoidance intervention, the proportion of participants failing the food challenge a 5 years of age was $36/263 = 0.137$ (13.7%); in the Peanut Consumption intervention, the proportion failing was $5/267 = 0.019$ (1.9%). The difference between these two proportions 11.8% is a single summary statistic showing the gap between the two proportions. A second summary statistic, the ratio of the two proportions, $0.137/0.019 = 7.31$, indicates that the proportion failing on the Avoidance group was more than 7 times that on the Consumption group. This ratio is called a **relative risk**.

The summary statistics for the LEAP study highlight an important point – the results of a study can sometimes be surprising. Someone unaware of early preliminary results about the potential value of exposure to peanut products (perhaps a parent of a child allergic to eggs) might be justifiably skeptical about the advisability of feeding peanut butter to his or her child. The LEAP study suggests that, at least in children similar to those in the study, the benefit might be substantial.

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

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, and it is not at all clear that results in children from that site can be generalized to other countries or cultures. Even if the study can be generalized, the results also raise an important statistical issue. Peanut consumption among infants susceptible to peanut allergies should be adopted only if the study results are definitive 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%, it is unlikely for exactly 50 heads to be observed. This type of fluctuation is part of almost experiment or study. It may well be possible that the 8% difference in the stent study is due to this natural variation. However, the larger the difference we observe (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 up 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.

The material on hypothesis testing will provide the statistical tools to examine this issue. In LEAP, we will be able to show that the 11.8% difference was indeed larger that that expected by chance alone if the two interventions were equally effective at preventing subsequent allergies.

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`, 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 1.3.

JV:please create this table of defs, using the famuss table as a model. Also, we should add that the data discussed here are in the original scale, not transformed, as in the paper. Note that I have changed some variable names. Please see the R file oi_biosta_ch1.R

[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 definitions]

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) [4], funded by the National Institutes of Health (NIH), measured a variety of demographic, phenotypic, and genetic characteristics of about 1300 participants. Data from the study has been used in many subsequent populations [5], such as a study examining the relationship between muscle strength and the genotype at a location on the gene *actn3* [6] Four rows of the `FAMuSS` dataset are shown in Table 1.4, and the variables are summarized in Table 1.5. Additional

	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 1.4: 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, we would not classify a variable sex or race as numerical since their average, sum, and difference have no meaning. Age measured in

[4] 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

[5] 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.

[6] Clarkson P, et al., Journal of Applied Physiology 99: 154163, 2005.

variable	description
`sex`	Sex of the participant
`age`	Age in years
`race`	Recorded as African AM (African American), Caucasian, Hispanic and Other.
`height`	Height in inches
`weight`	Weight in lbs
`actn3.r577x`	Genotype at the location r577x in the gene actn3. The four genotypes observed were CC, CT and TT
`ndrm.ch`	Percent change in strength in the non-dominant arm, comparing strength after to before training

Table 1.5: Variables and their descriptions for the `FAMuSS` data set.

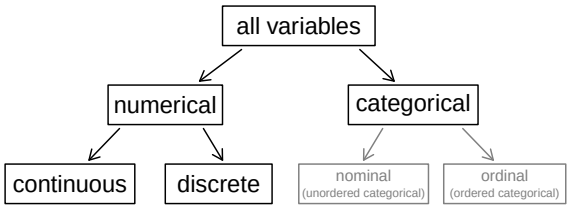


Figure 1.6: Breakdown of variables into their respective types.

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 `actn3.3577x` are **categorical** variables, [7] and the possible values are called the variable's **levels**. For example, the levels of `actn3.3577x` 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 call categorical variables **factors**, and the possible values of factors are called **levels**.

In the `frog` data, the variables `latitude`, `altitude`, `egg.size`, `clutch.size`, `clutch.volume`, and `body.size` are all continuous variables. *JV: agree about latitude?*

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 a major commercial pharmacies.

[7] sometimes called **nominal** variables.

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.The number of siblings and student height represent numerical variables.

⊙ **Guided Practice 1.2** Characterize the data types for the variables `participant.ID` `treatment.group` and `overall.V60.outcome` from the LEAP study in Section 1.1. [8]

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 in order 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 polymorphism of ACTN3 at residue 577 that changes C to T produces a stop codon; TT individuals are unable to produce any ACTN3 protein in their muscle. The TT genotype does not cause any discernible phenotype changes, which suggests that the ACTN3 protein is not critical

[8] All these variables measure non-numerical quantities, and are categorical. The variables `treatment.group` and `outcome.V60.overall` have two values or levels, while `participant.ID` has many possible values.

to muscle function. However, the ACTN3 gene is highly conserved, and may potentially influence variation in muscle function. *'conserved' is a technical term that needs a definition. Perhaps we can avoid it altogether. There are too many technical terms in this paragraph generally.*

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. Later in the text we will examine methods for characterizing a relationship numerically. *too vague*

⊙ **Guided Practice 1.3** Use the variables from the `FAMuSS` data set described in Table 1.5 to pose two questions about the relationships between these variables that are different from the question of interest to the research team. [9]

Much of this text examines both numerical and graphical ways to examine possible relationships between two variable. Scatterplots for numerical variables and mosaic plots for categorical variables are discussed later in this chapter. *need ref to section*

1.3 Data collection principles

The first step in conducting research is to identify questions to investigate. A clearly articulated research question is also essential in identifying the type of subjects to be studied, relevant variables, and 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 that peanut products during the first 6 months of the infant's life.

[9] 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.

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.

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.

We can introduce a question using LEAP

I think this is one of our biggest challenges throughout: provide a clear, honest intro to statistics as it is used, not as it is idealized in the math stat books. The idealization is useful in the theoretical books; it is downright silly at the intro level and causes us to lose good students.

1.3.2 Anecdotal evidence

Anecdotal evidence is typically composed of unusual observations that are easy recall based on their striking characteristics. Physicians are sometimes more likely to remember the characteristics of patient with an unusually good response to a drug than the features of 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 be remembered correctly or may have been measured in error. To learn about the characteristics of a population, it is

necessary to examine a sample of many cases drawn randomly 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. " [10]

While it is incorrect to generalize from individual observations, scientists know that interesting observations can be valuable. Striking observations may be a reason to question assumptions or to design a study to examine an unconventional question more closely. Cures for certain diseases have been discovered in research inspired by a response to a new drug of a patient with a disease thought be be incurable. *Insert references sent by D Longo and D. Spriggs here.*

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.

Anecdotal evidence

Be careful of data collected in a haphazard fashion. Such evidence may be true and verifiable, but it may only represent extraordinary cases.

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 measure those characteristics on every member, often an impossible task. The US Centers for Disease Control (US CDC) conducts many such surveys, including the Behavioral Risk Factors Surveillance System (BRFSS)[11]. The BRFSS conducts approximately 400,000 telephone interviews annually to ask U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of

[10] Jefferson, T. (1985). Letters, 17601826. Ed. Merrill D. Peterson. New York: Viking.
[11] http://www.cdc.gov/brfss/

preventive services. The CDC conducts similar surveys in diabetes, health care access, and immunization. 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 these countries [12]. In 2000, the US Department of Justice released the *The Sexual Victimization of College Women* [13], based on a survey of 4,446 women undergraduates conducted in 1996. *we should check to see if the data are available.*

placeholder for the Harvard Survey, despite its flaws

Sampling from a population is easier when a population is relatively small and members of the population are easy to identify and reach. 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 the perception of care by members of the system. Since health plans have contact information for each of their members, a selected subset can be contacted and, with their consent, participate in an interview or mailed survey. More complex methods are used in surveys such as the study of sexual victimization of college women. The general principle of sampling is straightforward; a sample from a population is useful in learning about a population only when the sample matches, on average, the characteristics of the population.

One common downfall in conducting a sample is to use a **convenience sample**, where individuals who are easily accessible are more likely to be included in the sample. For instance, the quality control team in the health care plan might ask interviewers to approach plan members visiting an outpatient clinic during a particular week. The sample would not enroll generally healthy members who typically do not use outpatient services or schedule routine physical examinations.

Random sampling is best way to insure that a sample reflects a population, because random samples do not reflect the conscious or unconscious bias of the team gathering the sample. 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

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

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 **simple random samples**, each member of the population is chosen directly 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. In the health plan example, a subset of members might be chosen randomly from the plan membership roster and called. More complex sampling methods are sometimes used in larger populations, and despite their complexity, are often designed to ensure that each member of a population has equal chance of being included in the sample. The Department of Justice report of sexual victimization describes the sampling strategy used to draw a random sample of women attending 2- or 4-year colleges with at least 1,000 students during the fall of 1996. *OpenIntro*, third edition, Section 1.4.2 describes the 4 most commonly used sampling strategies.

Sometimes a simple random sample is difficult to implement and an alternative method is helpful. One such substitute is a **systematic sample**, where one case is sampled after letting a fixed number of others, say 10 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.

The act of taking a simple random sample helps minimize bias, however, bias can crop up in other ways. Even when people are picked at random, e.g. for surveys, 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 **representative** of the entire population. This **non-response bias** can skew results. Since it is usually impossible to obtain reliable results from surveys with high non-response rates, it is best to minimize the barriers that might discourage subject participation, such as English only surveys in populations where some members may not use English as a first language.

adapt Open Intro graphic here? I have left them in for placeholders

⊙ Guided Practice 1.4

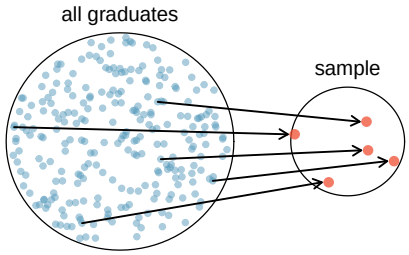


Figure 1.7: In this graphic, five graduates are randomly selected from the population to be included in the sample.

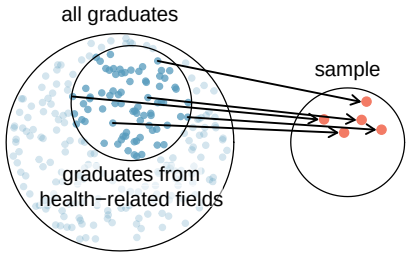


Figure 1.8: Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionately often.

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?[14]

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, we may suspect that administering a certain drug will reduce mortality in heart attack patients. To find evidence for a causal connection between

[14] 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.

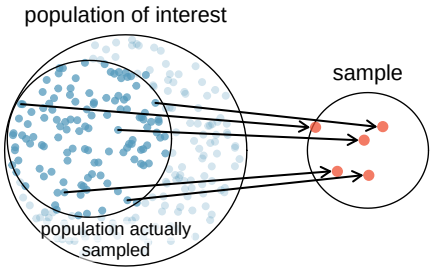


Figure 1.9: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

the explanatory and response variables, researchers will collect a sample of individuals and split them into groups. The individuals in each group are randomly assigned 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 that is known to have some effect, and 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, researchers may collect information via surveys, review medical or company records, or follow a **cohort** of many similar individuals to study why certain diseases might develop. 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.

1.3.5 Experiments

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

Controlling. Researchers assign treatments to cases, and they do their best to **control** for any other differences in the groups. For example, infants in the LEAP study had be between 4 and 11 months of age and had to have 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. Randomizing patients into 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. The data examined earlier from the LEAP study includes only the patients without a wheal after the skin test. General methods for combining strata in when 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, we **replicate** by collecting a sufficiently large sample. The LEAP study randomized a total of 640 infants, 542 in the block without the wheal response.

Blocking. Researchers sometimes know or suspect that variables, other than the treatment, influence the response. Under these circumstances, they may first group individuals based on this variable into **blocks** and then randomize cases within each block to the treatment groups. This strategy is often referred to as **blocking**. For instance, if we are looking at the effect of a drug on heart attacks, we might first split patients in the study into low-risk and high-risk blocks, then randomly assign half the patients from each block to the control group and the other half to the treatment group, as shown in Figure 1.7. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

update the following figure

It is important to incorporate at least the first three experimental design principles into any study, and this book describes applicable methods for analyzing data from such experiments. Blocking is a slightly more advanced technique, and statistical methods in

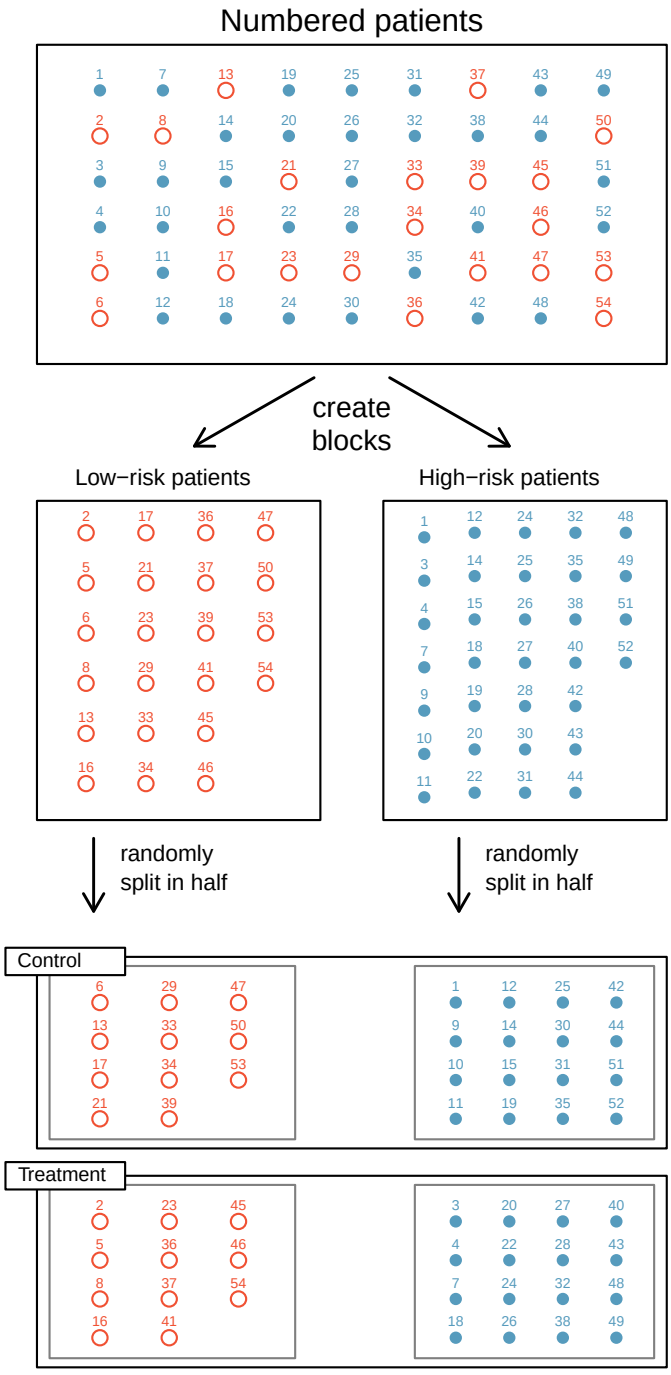


Figure 1.10: Blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly separated into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both the low-risk and high-risk categories.

this book may be extended to analyze data collected using blocking.

1.3.6 Reducing bias in human experiments

Randomized experiments are the gold standard for data collection, but they do not ensure an unbiased perspective into the cause and effect relationships in all cases. Human studies are perfect examples where bias can unintentionally arise. Here we reconsider a study where a new drug was used to treat heart attack patients.[15] In particular, researchers wanted to know if the drug reduced deaths in patients.

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[16] were randomly placed into two study groups. One group, the **treatment group**, received the drug. The other group, called the **control group**, did not receive any drug treatment.

Put yourself in the place of a person in the study. If you are in the treatment group, you are given a new drug that you anticipate will help you. On the other hand, a person in the control group doesn't receive the drug and hopes her participation doesn't increase her risk of death. These perspectives suggest there are actually two effects: the clinical effectiveness of the drug, and the potential emotional response of a patient that is difficult to quantify.

If Researchers interested in only the clinical effect, the emotional response may cause different behavior between patients in the two groups, which might bias the study. To circumvent this problem, researchers typically do not want patients to know which group they are in. When researchers keep the patients uninformed about their treatment, the study is said to be **blinded**. But there is one problem: in studies where a member of the control group does not receive any treatment, she will know she is in the control group. The solution to this problem is to give an inert substance to patients in the control group, called a **placebo**, and an effective placebo is the key to making a study truly blind. Often times, a placebo results in a slight but real improvement in patients. This effect has been

[15] 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.

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

dubbed the **placebo effect**. [17]

The patients are not the only ones who should be blinded: doctors and researchers can accidentally bias a study. When a doctor knows a patient has been given the real treatment, she might inadvertently give that patient more attention or care than a patient that she knows is on the placebo, perhaps looking for side-effects of the new drug. To guard against this bias, which again has been found to have a measurable effect in some instances, most modern studies employ a **double-blind** setup where doctors or researchers who interact with patients are, just like the patients, unaware of who is or is not receiving the experimental treatment.[18]

⊙ **Guided Practice 1.5** Look back to the study in Section 1.1 where researchers were testing whether peanut product consumption were effective at reducing the likelihood of peanut allergies children at-risk for these allergies. Is this an experiment? Was the study blinded? Was it double-blinded?[19]

1.3.7 Observational studies

Generally, data in observational studies are collected only by monitoring what occurs, while experiments require the primary explanatory variable in a study be assigned for each subject by the researchers.

Making causal conclusions based on experiments is often reasonable. However, making the same causal conclusions based on observational data is generally wrong and and is not recommended. Observational studies are generally sufficient to show only associations.

⊙ **Guided Practice 1.6** Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer?[20]

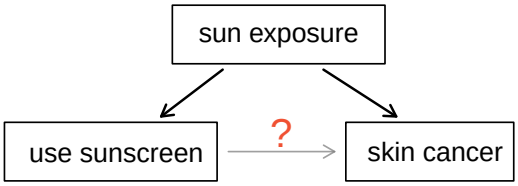
[17] Kaptchuk, TJ and Miller, FG. 2015.Placebo effects in medicine, New England Journal of Medicine, 373(1):8-9.

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

[19] The researchers assigned the patients into their treatment groups, so this study was an experiment. However, the patients could distinguish what treatment they received, so this study was not blind. The study could not be double-blind since it was not blind.

[20] No. See the paragraph following the exercise for an explanation.

Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer. One important piece of information that is absent is sun exposure. If someone is out in the sun all day, she is more likely to use sunscreen *and* more likely to get skin cancer. Exposure to the sun is unaccounted for in the simple investigation.



Sun exposure is what is called a **confounding variable**,[21] which is a variable that is correlated with both the explanatory and response variables. While one method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured.

The `famuss` data set is an observational study with confounding variables, and its data cannot easily be used to make causal conclusions.

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.[22] 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 data sets, may contain both prospectively- and retrospectively-collected variables. *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*

have commented out the section on sampling methods, though I like it. It will take

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

some work to find example to make the sampling methods concrete

1.4 Examining numerical data

This section introduces techniques for exploring and summarizing numerical variables, using the `frog_altitude` dataset 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 all 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 on the x communicates that the average volume of the 431 clutches was 882.5mm^3. It is useful to think of the mean as the balancing point of the distribution.

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.

n
sample size

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 sorted values of `clutch.volume`: 831.8mm^2.

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.

1.4.2 Measures of spread: standard deviation and interquartile range

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

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.

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

If we square these deviations and then take an average, the result is about equal to 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 it is not covered here.

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 380mm^3.

excellent place to give an interpretation of sd, referring back to verbal definition

insert tip box for SD formula

Variability can also be measured by 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.

Variance and standard deviation

The variance is roughly the average squared distance from the mean. The standard deviation is the square root of the variance. The standard deviation is useful when considering how close the data are to the mean.

Formulas and methods used to compute the variance and standard deviation for a population are similar to those used for a sample.[23] However, like the mean, the population

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

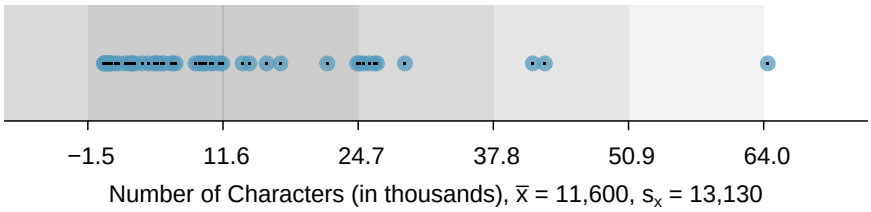


Figure 1.11: In the `num_char` data, 41 of the 50 emails (82%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% are within 2 standard deviations, though this rule of thumb is less accurate for skewed data, as shown in this example.

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

I have re-inserted the figures from Open Intro here; I like them. The will have to be redone for the frog data. The code for producing the figures is in the directory containing the figure.

TIP: standard deviation describes variability

Focus on the conceptual meaning of the standard deviation as a descriptor of variability rather than the formulas. Usually 70% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, as seen in Figures 1.8 and 1.9, these percentages are not strict rules.

Guided Practice 1.9 On page 29, the concept of shape of a distribution was introduced. A good description of the shape of a distribution should include modality and whether the distribution is symmetric or skewed to one side. Using Figure 1.9 as an example, explain why such a description is important.[24]

Example 1.10 Describe the distribution of the `num_char` variable using the his-

[24] Figure 1.9 shows three distributions that look quite different, but all have the same mean, variance, and standard deviation. Using modality, we can distinguish between the first plot (bimodal) and the last two (unimodal). Using skewness, we can distinguish between the last plot (right skewed) and the first two. While a picture, like a histogram, tells a more complete story, we can use modality and shape (symmetry/skew) to characterize basic information about a distribution.

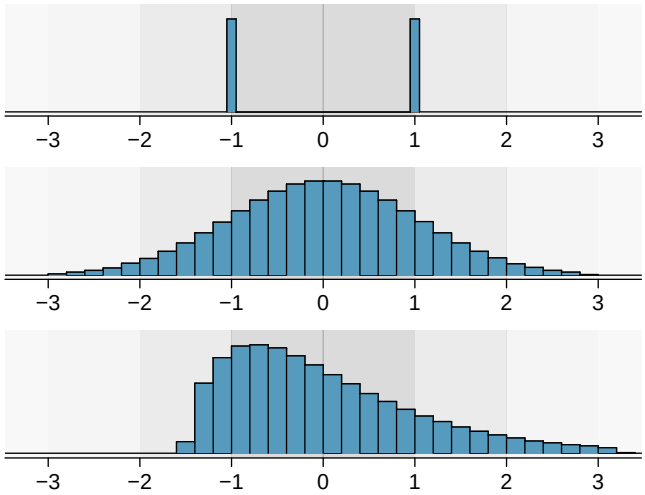


Figure 1.12: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

togram in Figure 1.15 on page 30. The description should incorporate the center, variability, and shape of the distribution, and it should also be placed in context: the number of characters in emails. Also note any especially unusual cases.

The distribution of email character counts is unimodal and very strongly skewed to the high end. Many of the counts fall near the mean at 11,600, and most fall within one standard deviation (13,130) of the mean. There is one exceptionally long email with about 65,000 characters.

In practice, the variance and standard deviation are sometimes used as a means to an end, where the "end" is being able to accurately estimate the uncertainty associated with a sample statistic. For example, in Chapter **??** we will use the variance and standard deviation to assess how close the sample mean is to the population mean.

1.4.3 Robust statistics

I wrote this using the famuss data because I thought the frogs would not be as illustrative. But when I checked, it is a pretty good example of this. In the R code for this chapter, you will find the summary statistics for the clutch.volume data with and without the 4 largest observations, along with the code for a dotPlot. dotPlot() requires that the openintro

package be loaded, so you have to install it from the web, then precede the dotplot command with library(openintro)

If you agree, please replace the example below with one using clutch.volume and revise accordingly. Because this is an important concept we could leave both examples but it may interrupt the flow.

The median and IQR are called **robust estimates** because extreme observations have little effect on their values. The mean and standard deviation are much more affected by changes in extreme observations.

In `famuss` there are six observed weights larger than 270 pounds (273, 291, 295, 305, 308 and 317); these weights are evident in Figure 1.10 Despite the size of these weights, it is unlikely that the observations are errors, since the six values are clustered above 270. How do these three weights affect the summary statistics for the weight variable in `famuss`? These large weights are evident in the plot below.

How are the sample statistics of the `weight` affected by these observations? The sample statistics are computed under each of two scenarios in Table 1.11, one with and one without these large observations. The table shows that neither the median or the interquartile range change when the six largest observations are dropped, but the mean and standard deviation both become lower. Because the standard deviation depends on the squared distances of observations from the mean, its change is more noticeable. Typically, extreme observations have a greater effect on the standard deviation than on the mean.

plot needs to be fancified using myPDF() in openintro

	robust		**not robust**	
scenario	median	IQR	$\bar{x}$	s
original `weight` data	150	42	155.65	34.59
drop six largest observations	150	42	154.2	31.58

Table 1.14: A comparison of how the median, IQR, mean ($\bar{x}$), and standard deviation (s) change when extreme observations are present.

1.4.4 Visualizing distributions of data: dot plots and histograms

Graphical summaries are useful tools for visualizing how data are distributed. A **dot plot** provides the most basic of displays, representing data as points plotted on a single axis.

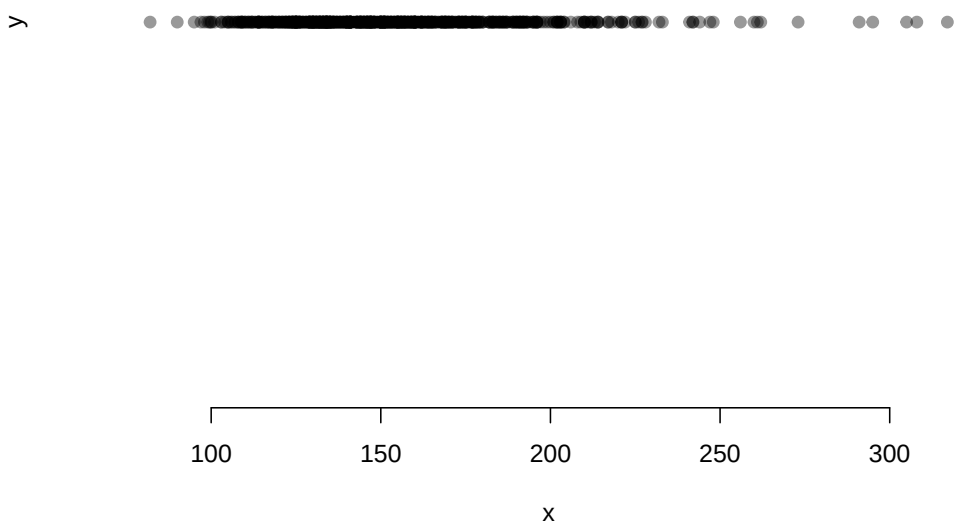


Figure 1.13: Dot plots of the weight variable in `famuss`.

An example using the number of characters from 50 emails is shown in Figure 1.12. A stacked version of this dot plot is shown in Figure 1.13.

Dot plots show the exact value for each observation. This is useful for small data sets, but they can become hard to read with larger samples. Rather than showing the value of each observation, we prefer to think of the value as belonging to a *bin*. For example, in the `email50` data set, we create a table of counts for the number of cases with character counts between 0 and 5,000, then the number of cases between 5,000 and 10,000, and so on. Observations that fall on the boundary of a bin (e.g. 5,000) are allocated to the lower

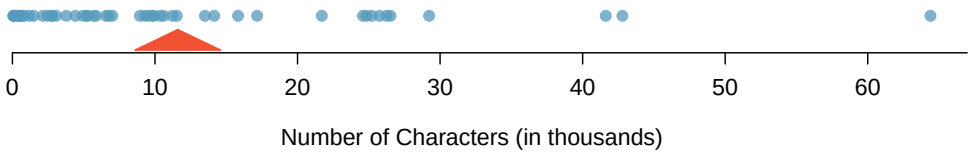


Figure 1.15: A dot plot of `num_char` for the `email50` data set.

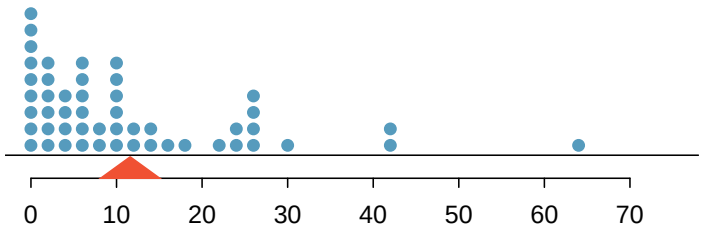


Figure 1.16: A stacked dot plot of `num_char` for the `email50` data set. The values have been rounded to the nearest 2,000 in this plot.

bin. This tabulation is shown in Table 1.14. These binned counts are plotted as bars in Figure 1.15 into what is called a **histogram**, which resembles the stacked dot plot shown in Figure 1.13.

Characters (in thousands)	0-5	5-10	10-15	15-20	20-25	25-30	···	55-60	60-65
Count	19	12	6	2	3	5	···	0	1

Table 1.17: The counts for the binned `num_char` data.

Histograms provide a view of the **data density**. Higher bars represent where the data are relatively more common. For instance, there are many more emails with fewer than 20,000 characters than emails with at least 20,000 in the data set. The bars make it easy to see how the density of the data changes relative to the number of characters.

Histograms are especially convenient for describing the shape of the data distribution. Figure 1.15 shows that most emails have a relatively small number of characters, while fewer emails have a very large number of characters. When data trail off to the right in this way and have a longer right tail, the shape is said to be **right skewed**.[25]

[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**.

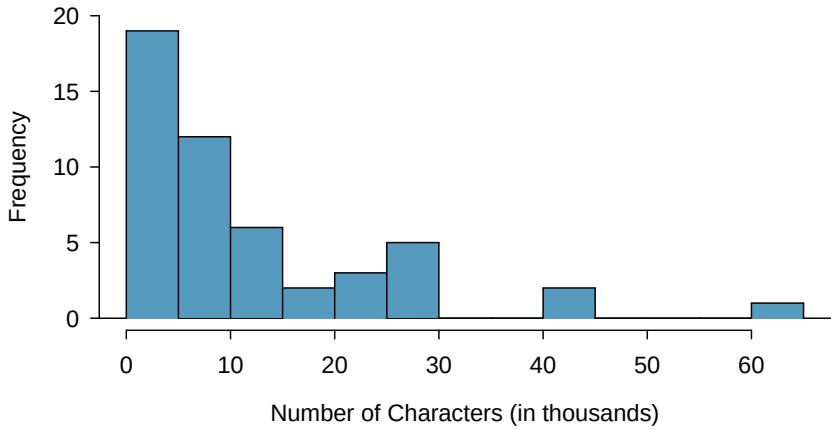


Figure 1.18: A histogram of `num_char`. This distribution is very strongly skewed to the right.

Data sets with the reverse characteristic – a long, thin tail to the left – are said to be **left skewed**. We also say that such a distribution has a long left tail. Data sets that show roughly equal trailing off in both directions are called **symmetric**.

In addition to looking at whether a distribution is skewed or symmetric, histograms can be used to identify modes. A **mode** is represented by a prominent peak in the distribution.[26] There is only one prominent peak in the histogram of `num_char`.

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 2 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.

⊙ **Guided Practice 1.11** Figure 1.15 reveals only one prominent mode in the number of characters. Is the distribution unimodal, bimodal, or multimodal?[27]

[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 data sets.

[27] Unimodal. Remember that *uni* stands for 1 (think *uni*cycles). Similarly, *bi* stands for 2 (think *bi*cycles). (We're hoping a *multicycle* will be invented to complete this analogy.)

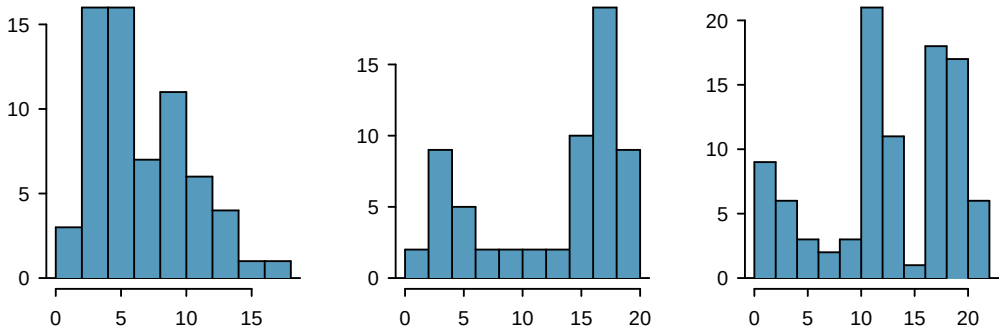


Figure 1.19: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

1.4.5 Boxplots, quantiles, outliers

A **boxplot** summarizes a dataset using five statistics while also plotting unusual observations. Figure 1.17 provides a vertical dot plot alongside a box plot of the `num_char` variable from the `email50` data set.

Figure 1.20: A vertical dot plot next to a labeled box plot for the number of characters in 50 emails. The median (6,890), splits the data into the bottom 50% and the top 50%, marked in the dot plot by horizontal dashes and open circles, respectively.

The first step in building a box plot is drawing a dark line denoting the **median**, which splits the data in half. Figure 1.17 shows 50% of the data falling below the median

(dashes) and other 50% falling above the median (open circles).

The second step in building a box plot is drawing a rectangle to represent the middle 50% of the data, which is the IQR.

Extending out from the box, the **whiskers** capture the data that fall between $1.5 \times IQR$.[28] In Figure 1.17, the upper whisker does not extend to the last three points, which is beyond $Q_3 + 1.5 \times IQR$, and so it extends only to the last point below this limit. The lower whisker stops at the lowest value, 33, since there is no additional data to reach; the lower whisker's limit is not shown in the figure because the plot does not extend down to $Q_1 - 1.5 \times IQR$. In a sense, the box is like the body of the box plot and the whiskers are like its arms trying to reach the rest of the data.

Any observation that lies beyond the whiskers is labeled with a dot. The purpose of labeling these points is to help identify any observations that appear to be unusually distant from the rest of the data. These observations are called outliers; An **outlier** is an observation that appears extreme relative to the rest of the data. In this case, it would be reasonable to classify the emails with character counts of 41,623, 42,793, and 64,401 as outliers since they are numerically distant from most of the data. Outliers can potentially provide insight into interesting properties of the data.

1.4.6 Scatterplots

A **scatterplot** provides a case-by-case view of data for two numerical variables. In Figure #, a scatterplot is used to examine the relationship between clutch volume and female body size in the `frog` dataset. In any scatterplot, each point represents a single case. Since body size was measured for 129 frogs, there are 129 points in Figure .

The `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 # shows

[28] While the choice of exactly 1.5 is arbitrary, it is the most commonly used value for box plots.

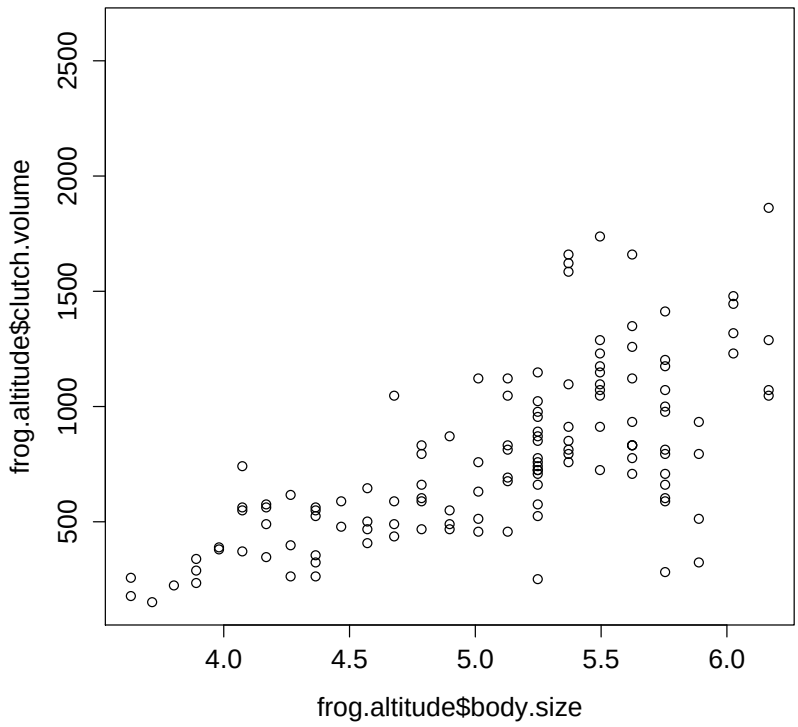


Figure 1.21: A scatterplot showing `clutch.volume` (horizontal axis) vs. `body.size` (vertical axis).

an upward trend – larger frogs tend to produce clutches with larger volume. 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 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. *plot should be fancified, and one of the heavy participants highlighted.*

Because taller people tend, naturally, to be heavier, 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. In the metric system, BMI is a person's weight in

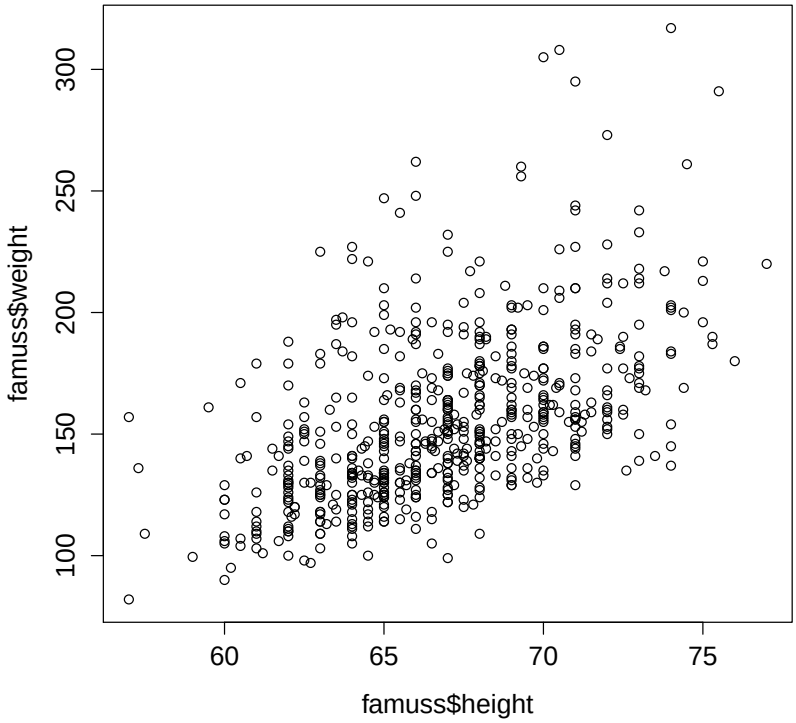


Figure 1.22: A scatterplot showing `height` (horizontal axis) vs. `weight` (vertical axis).

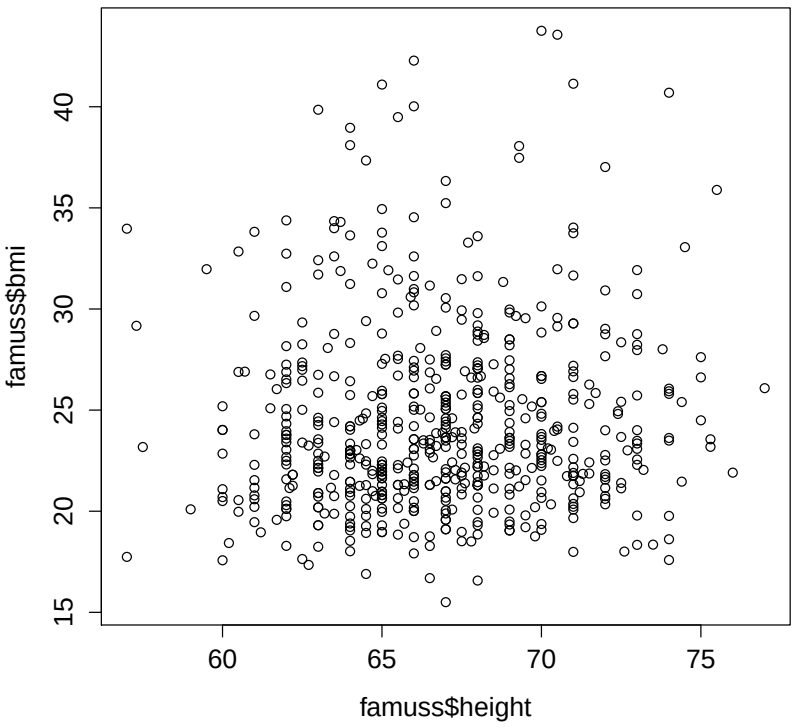


Figure 1.23: A scatterplot showing `height` (horizontal axis) vs. `bmi` (vertical axis).

kilograms (kg) divided by his or her height in meters squared. If height and weight are measured in inches and pounds, then BMI is weight in pounds divided by height in inches squared, then multiplied by 703. The `famuss` dataset 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, the US NIH, the World Health Organization and other health agencies use BMI rather that weight 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. It is generally not easy to determine definitively from a scatterplot whether two variables are independent, even in 1.20.

Caution: association does not imply causation

Labeling variables as *explanatory* and *response* does not guarantee the relationship between the two is causal, even if there is an association identified between the two variables. We use these labels only to keep track of which variable we suspect influences the other. Taller people do tend to be heavier, a variety of genetic and environmental factors influence weight as well.

Example 1.12 Consider a new data set of 54 cars with two variables: vehicle price and weight.[29] A scatterplot of vehicle price versus weight is shown in Figure **??**. What can be said about the relationship between these variables?

The relationship is evidently nonlinear, as highlighted by the dashed line. This is different from previous scatterplots we've seen, such as Figure **??** on page ?? and Figure **??**, which show relationships that are very linear.

1.4.7 Transforming data (special topic)

We should include a section on transforming data, esp since the frog data has been transformed.

[29] Subset of data from www.amstat.org/publications/jse/v1n1/datasets.lock.html

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 # is a frequency table for the `actn3.r577x` variable. Recall that `actn3.r577x` is a categorical variable that describes genotype at a particular locus 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**.

[insert frequency table for `actn3.r577x`]

none	small	big	Total
549	2827	545	3921

Table 1.24: A frequency table for the `number` variable.

Table # summarizes two variables: `race` and `actn3.r577x`. A table that summarizes data for two categorical variables in this way is called a **contingency table**. 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.

[insert race by genotype table]

Table # shows the row proportions for Table #. 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

		number			
		none	small	big	Total
spam	spam	149	168	50	367
	not spam	400	2659	495	3554
	Total	549	2827	545	3921

Table 1.25: A contingency table for spam and number.

	none	small	big	Total
spam	$149/367 = 0.406$	$168/367 = 0.458$	$50/367 = 0.136$	1.000
not spam	$400/3554 = 0.113$	$2657/3554 = 0.748$	$495/3554 = 0.139$	1.000
Total	$549/3921 = 0.140$	$2827/3921 = 0.721$	$545/3921 = 0.139$	1.000

Table 1.26: A contingency table with row proportions for the spam and number variables.

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

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

	none	small	big	Total
spam	$149/549 = 0.271$	$168/2827 = 0.059$	$50/545 = 0.092$	$367/3921 = 0.094$
not spam	$400/549 = 0.729$	$2659/2827 = 0.941$	$495/545 = 0.908$	$3684/3921 = 0.906$
Total	1.000	1.000	1.000	1.000

Table 1.27: A contingency table with column proportions for the spam and number variables.

1.5.2 Bar plots

A bar plot is a common way to display a single categorical variable. The left panel of Figure # 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 `none`), showing the proportion of observations that are in each level (i.e. in each category).

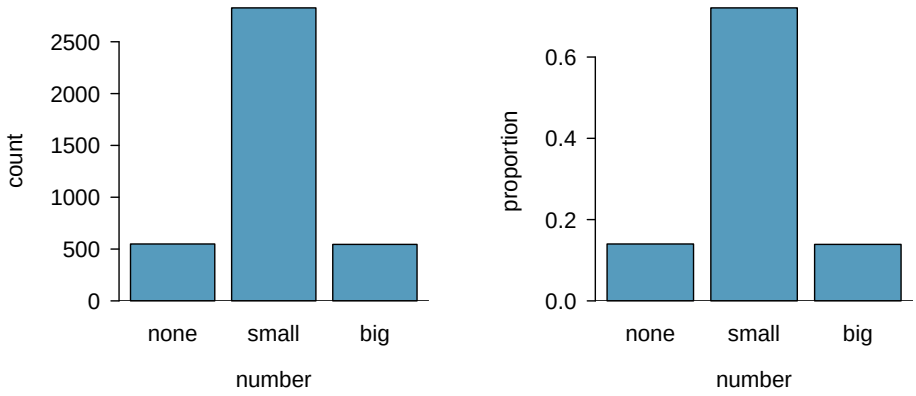


Figure 1.28: Two bar plots of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

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 representing Table # is shown in #, 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 # have been translated into a standardized segmented bar plot in Figure #, which is a helpful visualization of the races represented in each level of `actn3.r577x`.

1.5.3 Comparing numerical data across groups

Some of the more interesting investigations can be considered by examining numerical data across groups. In this section, two convenient methods are introduced: side-by-side box plots and hollow histograms.

The **side-by-side box plot** is a traditional tool for comparing across groups. Another useful plotting method uses **hollow histograms** to compare numerical data across groups.

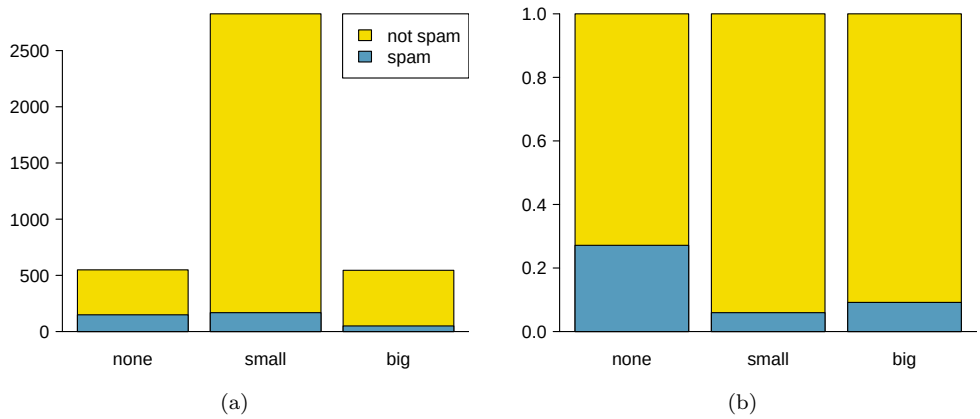


Figure 1.29: (a) Segmented bar plot for numbers found in emails, where the counts have been further broken down by `spam`. (b) Standardized version of Figure (a).

These are just the outlines of histograms of each group put on the same plot.

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. 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.

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 #. It seems that as a general rule, clutches found at higher altitudes have greater volume. However, more advanced statistical methods are required to thoroughly investigate the potential association between altitude and clutch size.

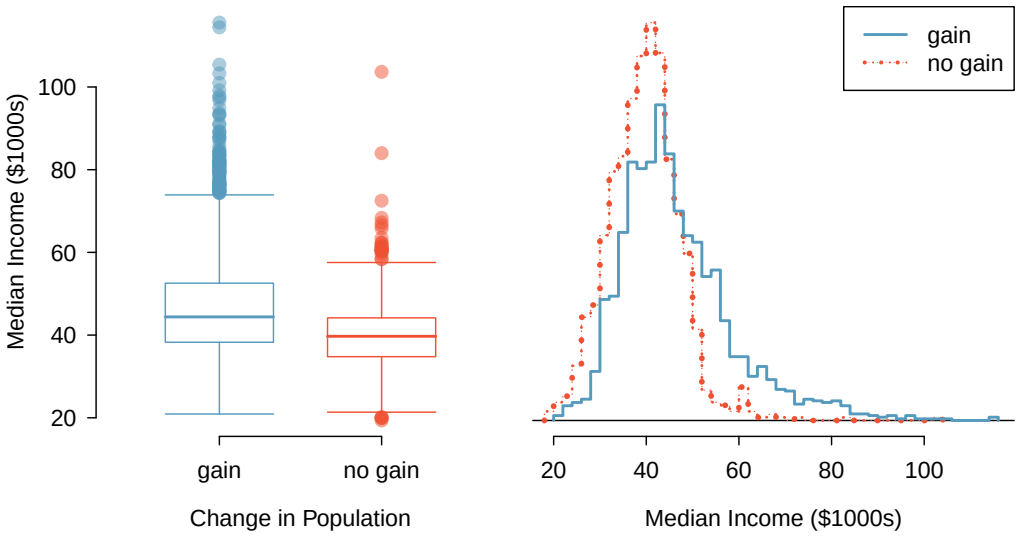


Figure 1.30: Side-by-side box plot (left panel) and hollow histograms (right panel) for `med_income`, where the counties are split by whether there was a population gain or loss from 2000 to 2010. The income data were collected between 2006 and 2010.