

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]

Copyright © 2015. Third Edition.

This textbook is available under a Creative Commons license. Visit openintro.org for a free PDF, to download the textbook's source files, or for more information about the license.

Contents

1	Introduction to data	8
1.1	Case study	9
1.2	Data basics	11
1.3	Examining numerical data	19
1.4	Considering categorical data	31

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.

¹Full 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.

²Diez 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 small number of 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

¹Du 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
	\vdots	\vdots	\vdots
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.

²Formally, 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 than 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 mm³. 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 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 reporting telephone area codes as numerical since their average, sum, and difference have no clear

⁴Thompson 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
⁵Pescatello L, et al. Highlights from the functional single nucleotide polymorphisms associated with human muscle size and strength or FAMuSS study, *BioMed Research International* 2013.
⁶Clarkson P, et al., *Journal of Applied Physiology* 99: 154163, 2005.

variable	description
<code>sex</code>	Sex of the participant
<code>age</code>	Age in years
<code>race</code>	Recorded as African AM (African American), Caucasian, Hispanic and Other.
<code>height</code>	Height in inches
<code>weight</code>	Weight in lbs
<code>actn3.r577x</code>	Genotype at the location r577x in the gene actn3. The four genotypes observed were CC, CT and TT
<code>ndrm.ch</code>	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.

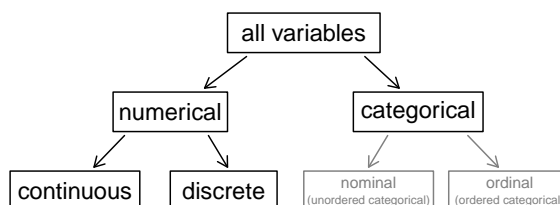


Figure 1.6: Breakdown of variables into their respective types.

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 `actn3.3577x` 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.

⁷sometimes 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 the following question: 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 to muscle function. However, the ACTN3 gene is highly conserved, and may potentially

⁸All 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.

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*

1.2.4 Introducing experiments and observational studies

Experiments and observational studies are the two primary types of study designs.

revisions to here 25jul 16:57

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 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 receives a **placebo** (fake treatment) and the second group receives the 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

⁹Two 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.

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.

1.2.5 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, in the stent study, patients in both groups received medical management, which included medications, management of stroke risk factors, and counseling on lifestyle modification. Effectively controlling variables is essential for making meaningful comparisons between treatment and control groups.

Randomization. Researchers randomize patients into treatment groups to account for variables that cannot be controlled. For example, some patients may have been more susceptible to stroke than others due to dietary habits or genetic risk factors. Randomizing patients into the treatment or control group helps even out such differences. In situations where researchers suspect that variables other than the treatment 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**. 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, we **replicate** by collecting a sufficiently large sample.

1.2.6 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 can be treacherous and is not recommended. Thus, observational studies are generally only sufficient to show associations.

⊙ **Guided Practice 1.4** Figure ?? shows a negative association between the homeownership rate and the percentage of multi-unit structures in a county. However, it is unreasonable to conclude that there is a causal relationship between the two variables. Suggest one or more other variables that might explain the relationship visible in Figure ??.¹⁰

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 1989.¹¹ 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, such as **county**, may contain both prospectively- and retrospectively-collected variables. Local governments prospectively collect some variables as events unfolded (e.g. retail sales) while the federal government retrospectively collected others during the 2010 census (e.g. county population counts).

[briefly discuss simple random sampling]

¹⁰Answers will vary. Population density may be important. If a county is very dense, then this may require a larger fraction of residents to live in multi-unit structures. Additionally, the high density may contribute to increases in property value, making homeownership infeasible for many residents.

¹¹www.channing.harvard.edu/nhs

1.3 Examining numerical data

some revisions from JV in this section

This section introduces techniques for exploring and summarizing numerical variables. The `frog_altitude` dataset from Section 1.2 provides rich opportunities for examples.

1.3.1 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.

$$\bar{x} = \frac{177.8 + 257.0 + \cdots + 933.3}{431} = 882.5mm^3 \quad (1.5)$$

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.

\bar{x}
sample
mean

Mean

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

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad (1.6)$$

where x_1, x_2, \dots, 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 216th observation in the `clutch.volume` variable: 831.8 mm^2 .

1.3.2 Measures of spread: standard deviation and interquartile range

The standard deviation roughly describes how far away the typical observation is from the mean. The distance of an observation from its mean is its **deviation**. Below are the deviations for the 1st, 2nd, 3rd, and 431th 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 :

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

We divide by $n - 1$, rather than dividing by n , when computing the variance; you need not worry about this mathematical nuance for the material in this textbook.

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

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

The standard deviation of clutch volume for the egg clutches observed is about 380 mm^3 .

s
sample
standard
deviation

insert tip box for SD formula

Variability can also be described by the **interquartile range** (IQR). To calculate the IQR, find the **first quartile** (the 25th percentile, i.e. 25% of the data fall below this value) and the **third quartile** (the 75th 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 $1096.0 - 609.6 = 486.4(mm)^3$.

1.3.3 Robust statistics

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.

adapt this for the largest value in `clutch.size`

How are the sample statistics of the `num_char` data set affected by the observation, 64,401? What would have happened if this email wasn't observed? What would happen to these summary statistics if the observation at 64,401 had been even larger, say 150,000? These scenarios are plotted alongside the original data in Figure 1.7, and sample statistics are computed under each scenario in Table 1.8.

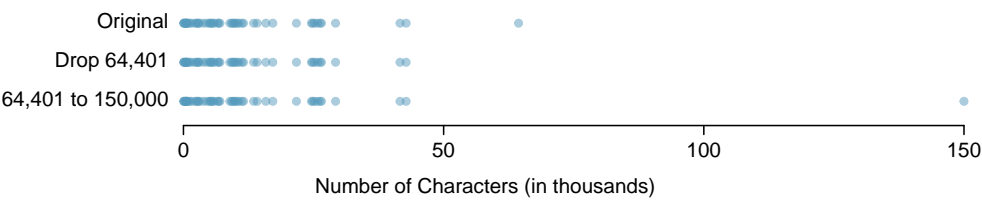


Figure 1.7: Dot plots of the original character count data and two modified data sets.

scenario	robust		not robust	
	median	IQR	\bar{x}	s
original <code>num_char</code> data	6,890	12,875	11,600	13,130
drop 66,924 observation	6,768	11,702	10,521	10,798
move 66,924 to 150,000	6,890	12,875	13,310	22,434

Table 1.8: A comparison of how the median, IQR, mean (\bar{x}), and standard deviation (s) change when extreme observations are present.

1.3.4 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.

An example using the number of characters from 50 emails is shown in Figure 1.9. A stacked version of this dot plot is shown in Figure 1.10.

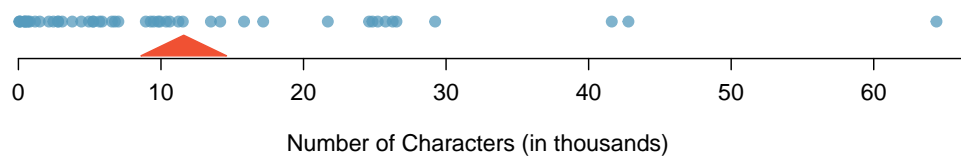


Figure 1.9: A dot plot of `num_char` for the `email50` data set.

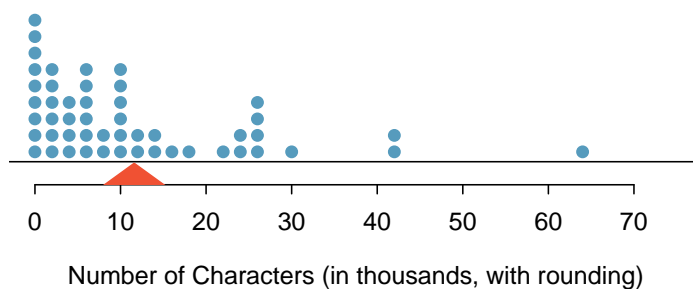


Figure 1.10: A stacked dot plot of `num_char` for the `email50` data set. The values have been rounded to the nearest 2,000 in this plot.

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 bin. This tabulation is shown in Table 1.11. These binned counts are plotted as bars in Figure 1.12 into what is called a **histogram**, which resembles the stacked dot plot shown in Figure 1.10.

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 1.11: The counts for the binned `num_char` data.

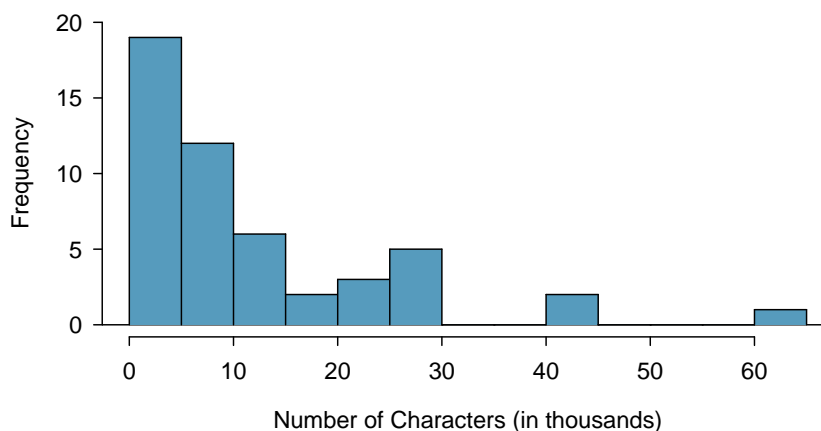


Figure 1.12: A histogram of `num_char`. This distribution is very strongly skewed to the right.

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 1.12 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**.¹²

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

¹²Other 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**.

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.¹³ There is only one prominent peak in the histogram of `num_char`.

Figure 1.13 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.

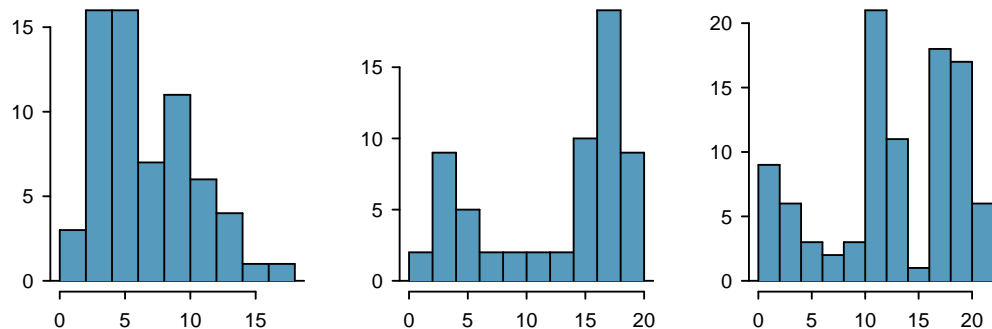


Figure 1.13: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

🕒 **Guided Practice 1.7** Figure 1.12 reveals only one prominent mode in the number of characters. Is the distribution unimodal, bimodal, or multimodal?¹⁴

1.3.5 Boxplots, quantiles, outliers

A **boxplot** summarizes a dataset using five statistics while also plotting unusual observations. Figure 1.14 provides a vertical dot plot alongside a box plot of the `num_char` variable from the `email150` data set.

¹³Another 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.

¹⁴Unimodal. Remember that *uni* stands for 1 (think *unicycles*). Similarly, *bi* stands for 2 (think *bicycles*). (We're hoping a *multicycle* will be invented to complete this analogy.)

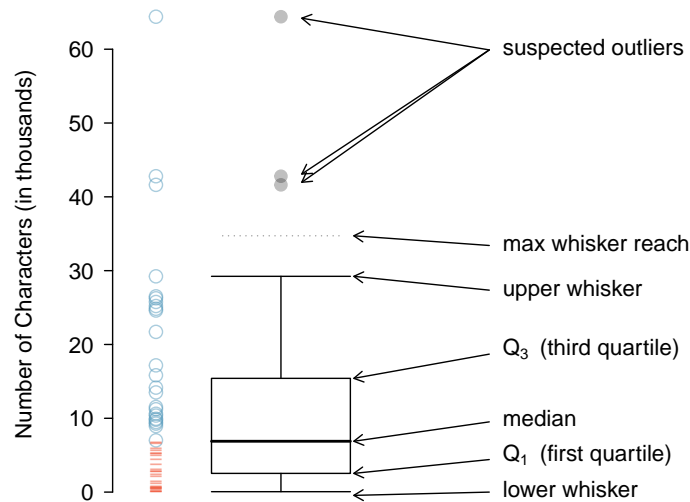


Figure 1.14: A vertical dot plot next to a labeled box plot for the number of characters in 50 emails. The median (6,890), splits the data into the bottom 50% and the top 50%, marked in the dot plot by horizontal dashes and open circles, respectively.

The first step in building a box plot is drawing a dark line denoting the **median**, which splits the data in half. Figure 1.14 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 $1.5 \times IQR$.¹⁵ In Figure 1.14, 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

¹⁵While the choice of exactly 1.5 is arbitrary, it is the most commonly used value for box plots.

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.

1.3.6 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 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 1.16 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 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

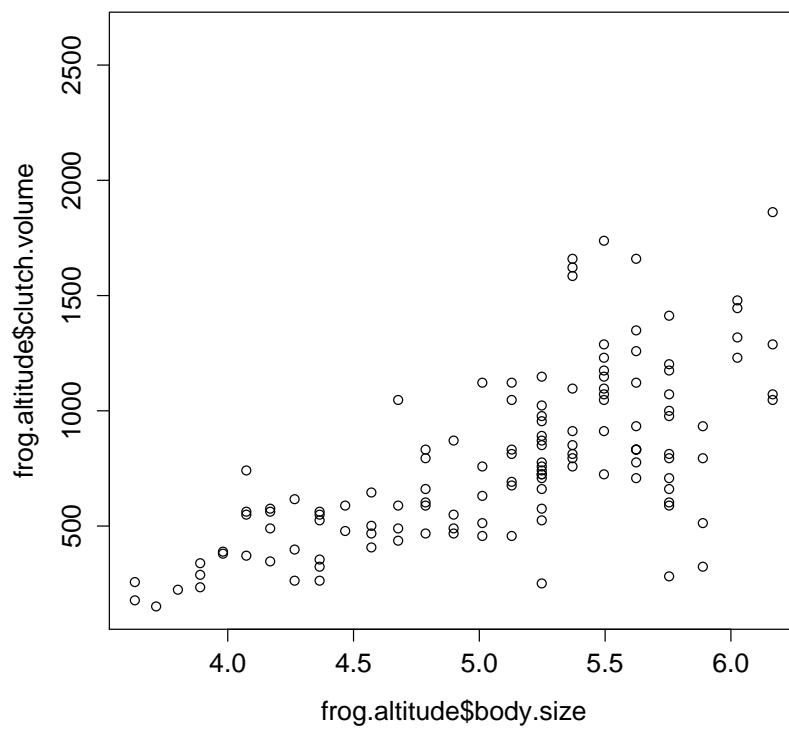


Figure 1.15: A scatterplot showing `clutch.volume` (horizontal axis) vs. `body.size` (vertical axis).

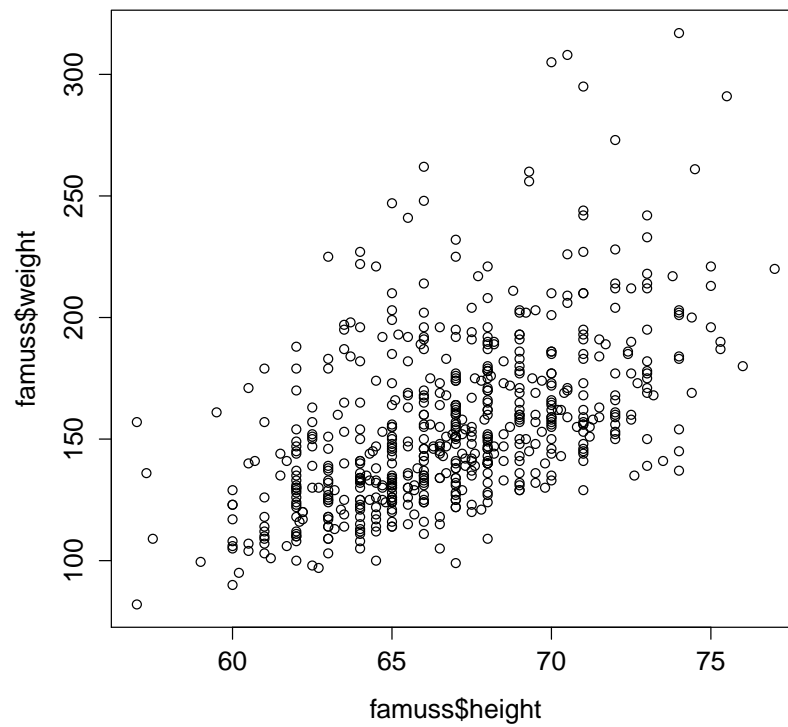


Figure 1.16: A scatterplot showing `height` (horizontal axis) vs. `weight` (vertical axis).

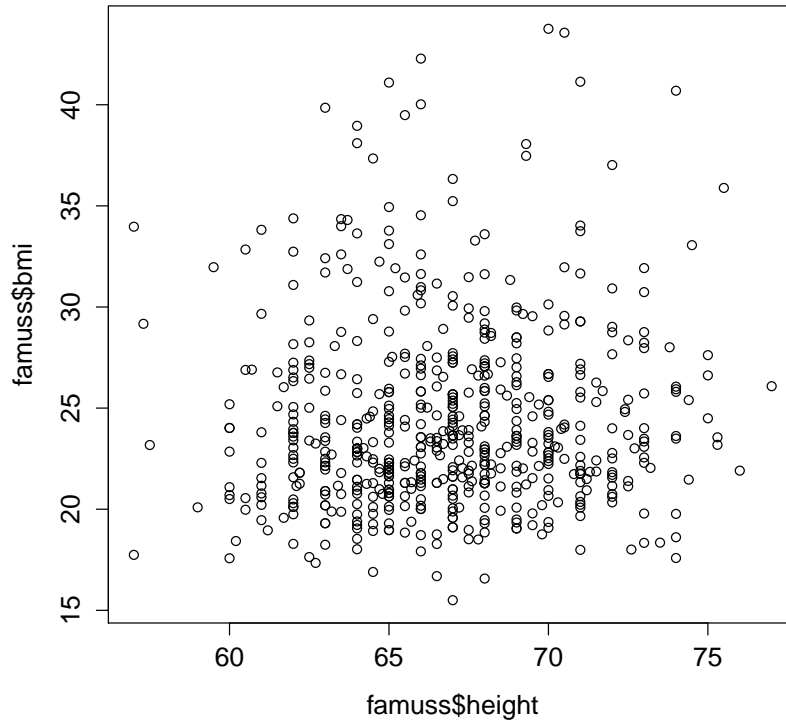


Figure 1.17: A scatterplot showing `height` (horizontal axis) vs. `bmi` (vertical axis).

participant, and figure 1.17 shows the relationship between `height` and `bmi`. The strong upward trend in Figure 1.16 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 than 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 1.17.

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 1.8** Consider a new data set of 54 cars with two variables: vehicle price and weight.¹⁶ 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.

1.3.7 Transforming data (special topic)

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

¹⁶Subset of data from www.amstat.org/publications/jse/v1n1/datasets.lock.html

1.4 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.

1.4.1 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 1.18: 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			Total
		none	small	big	
spam	spam	149	168	50	367
	not spam	400	2659	495	3554
	Total	549	2827	545	3921

Table 1.19: 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 1.20: 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 1.21: A contingency table with column proportions for the **spam** and **number** variables.

1.4.2 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).

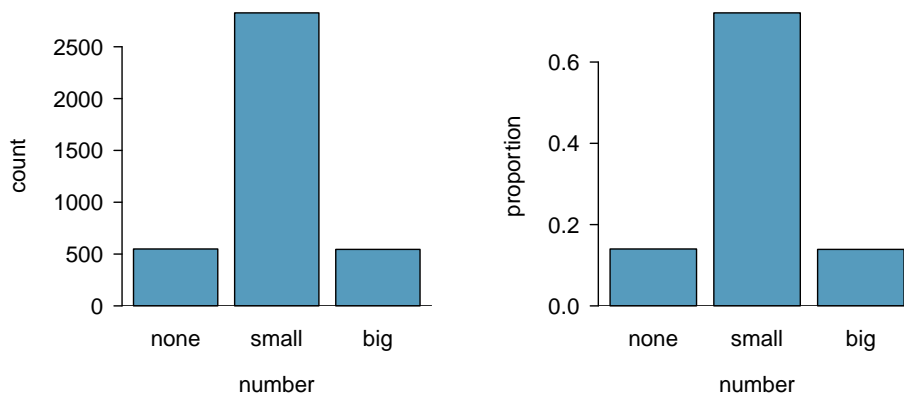


Figure 1.22: Two bar plots of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

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`.

1.4.3 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.

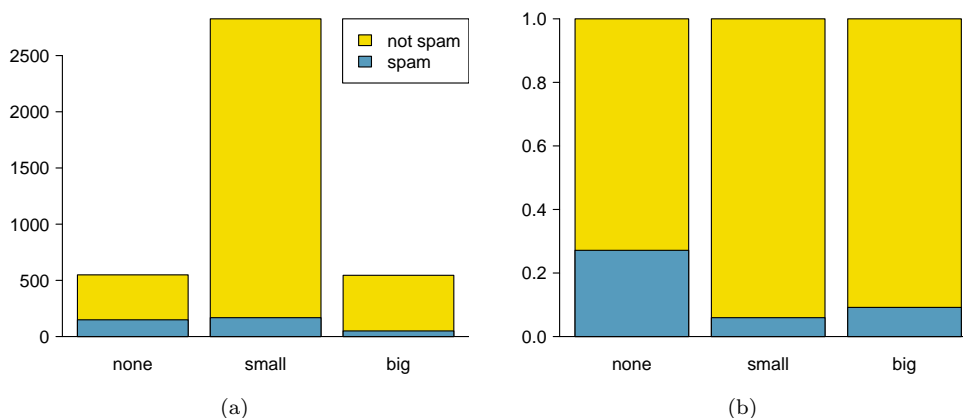


Figure 1.23: (a) Segmented bar plot for numbers found in emails, where the counts have been further broken down by **spam**. (b) Standardized version of Figure (a).

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 **CT**. 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.

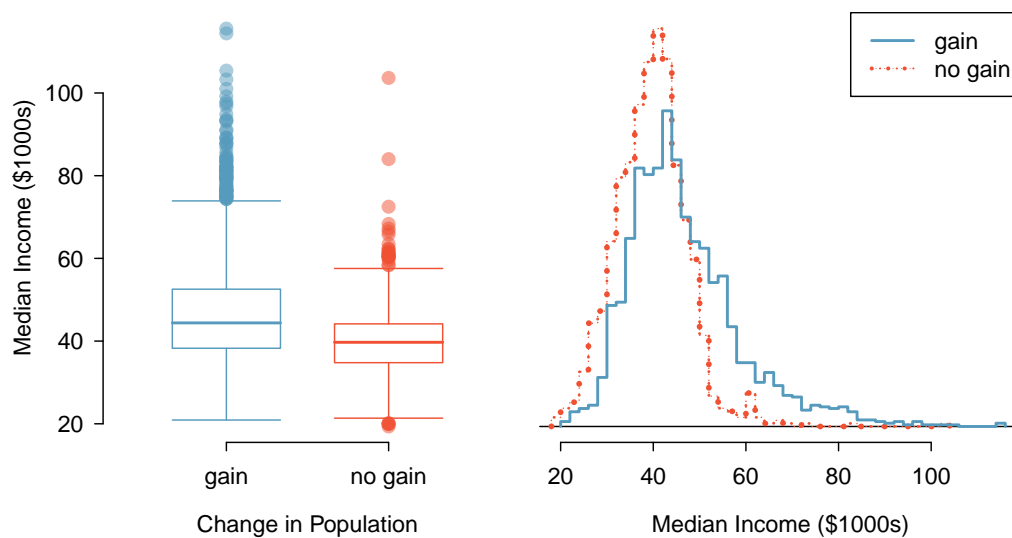


Figure 1.24: Side-by-side box plot (left panel) and hollow histograms (right panel) for `med_income`, where the counties are split by whether there was a population gain or loss from 2000 to 2010. The income data were collected between 2006 and 2010.