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?

The topics scientists investigate are as diverse as the questions they ask. However, many of these investigations can be addressed with a small number of data collection techniques, analytic tools, and fundamental concepts in statistical inference. This chapter provides a glimpse into these and other themes we will encounter throughout the rest of the book. We introduce the basic principles of each branch and learn some tools along the way. We will encounter applications from other fields, some of which are not typically associated with science but nonetheless can benefit from statistical study.

1.1 Case study: using stents to prevent strokes

Section 1.1 introduces a classic challenge in statistics: evaluating the efficacy of a medical treatment. Terms in this section, and indeed much of this chapter, will all be revisited later in the text. The plan for now is simply to get a sense of the role statistics can play in practice.

In this section we will consider an experiment that studies effectiveness of stents in treating patients at risk of stroke.[1] Stents are devices put inside blood vessels that assist

[1] Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. New England Journal of Medicine 365:993-1003. http://www.nejm.org/doi/full/10.1056/NEJMoa1105335. NY Times article reporting on the study: http://www.nytimes.com/2011/09/08/health/research/08stent.html.

in patient recovery after cardiac events and reduce the risk of an additional heart attack or death. Many doctors have hoped that there would be similar benefits for patients at risk of stroke. We start by writing the principle question the researchers hope to answer:

> Does the use of stents reduce the risk of stroke?

The researchers who asked this question collected data on 451 at-risk patients. Each volunteer patient was randomly assigned to one of two groups:

> **Treatment group**. Patients in the treatment group received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle modification.
>
> **Control group**. Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

Researchers randomly assigned 224 patients to the treatment group and 227 to the control group. In this study, the control group provides a reference point against which we can measure the medical impact of stents in the treatment group.

Researchers studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment. The results of 5 patients are summarized in Table 1.1. Patient outcomes are recorded as "stroke" or "no event", representing whether or not the patient had a stroke at the end of a time period.

Patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
⋮	⋮	⋮	
450	control	no event	no event
451	control	no event	no event

Table 1.1: Results for five patients from the stent study.

Considering data from each patient individually would be a long, cumbersome path towards answering the original research question. Instead, performing a statistical data analysis allows us to consider all of the data at once. Table 1.2 summarizes the raw data in a more helpful way. In this table, we can quickly see what happened over the entire study. For instance, to identify the number of patients in the treatment group who had a stroke within 30 days, we look on the left-side of the table at the intersection of the treatment and stroke: 33.

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Table 1.2: Descriptive statistics for the stent study.

⊙ **Exercise 1.1** Of the 224 patients in the treatment group, 45 had a stroke by the end of the first year. Using these two numbers, compute the proportion of patients in the treatment group who had a stroke by the end of their first year. (Please note: answers to all in-text exercises are provided using footnotes.)[2]

We can compute summary statistics from the table. A **summary statistic** is a single number summarizing a large amount of data.[3] For instance, the primary results of the study after 1 year could be described by two summary statistics: the proportion of people who had a stroke in the treatment and control groups.

Proportion who had a stroke in the treatment (stent) group: $45/224 = 0.20 = 20\%$.

Proportion who had a stroke in the control group: $28/227 = 0.12 = 12\%$.

These two summary statistics are useful in looking for differences in the groups, and we are in for a surprise: an additional 8% of patients in the treatment group had a stroke! This is important for two reasons. First, it is contrary to what doctors expected, which was that stents would *reduce* the rate of strokes. Second, it leads to a statistical question: do the data show a "real" difference between the groups?

This second question is subtle. Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won't observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process. It is 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 sample size), the less believable it is that the difference is due to chance. So what we are really asking is the following: is the difference so large that we should reject the notion that it was due to chance?

While we don't yet have our statistical tools to fully address this question on our own, we can comprehend the conclusions of the published analysis: there was compelling evidence of harm by stents in this study of stroke patients.

Be careful: do not generalize the results of this study to all patients and all stents. This study looked at patients with very specific characteristics who volunteered to be a part of this study and who may not be representative of all stroke patients. In addition, there are many types of stents and this study only considered the self-expanding Wingspan stent (Boston Scientific). However, this study does leave us with an important lesson: we should keep our eyes open for surprises.

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

Table 1.3 displays rows 1, 2, 3, and 50 of a data set concerning 50 emails received during early 2012. These observations will be referred to as the `email50` data set, and they are a random sample from a larger data set that we will see in Section 1.7.

[2] The proportion of the 224 patients who had a stroke within 365 days: $45/224 = 0.20$.

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

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

Table 1.3: Four rows from the email50 data matrix.

variable	description
spam	Specifies whether the message was spam
num_char	The number of characters in the email
line_breaks	The number of line breaks in the email (not including text wrapping)
format	Indicates if the email contained special formatting, such as bolding, tables, or links, which would indicate the message is in HTML format
number	Indicates whether the email contained no number, a small number (under 1 million), or a large number

Table 1.4: Variables and their descriptions for the email50 data set.

Each row in the table represents a single email or **case**.[4] The columns represent characteristics, called **variables**, for each of the emails. For example, the first row represents email 1, which is a not spam, contains 21,705 characters, 551 line breaks, is written in HTML format, and contains only small numbers.

In practice, it is especially important to ask clarifying questions to ensure important aspects of the data are understood. For instance, it is always important to be sure we know what each variable means and the units of measurement. Descriptions of all five email variables are given in Table 1.4.

The data in Table 1.3 represent a **data matrix**, which is a common way to organize data. Each row of a data matrix corresponds to a unique case, and each column corresponds to a variable. A data matrix for the stroke study introduced in Section 1.1 is shown in Table 1.1 on page 2, where the cases were patients and there were three variables recorded for each patient.

Data matrices are a convenient way to record and store data. If another individual or case is added to the data set, an additional row can be easily added. Similarly, another column can be added for a new variable.

⊙ **Exercise 1.2** We consider a publicly available data set that summarizes information about the 3,143 counties in the United States, and we call this the county data set. This data set includes information about each county: its name, the state where it resides, its population in 2000 and 2010, per capita federal spending, poverty rate, and five additional characteristics. How might these data be organized in a data matrix? Reminder: look in the footnotes for answers to in-text exercises.[5]

Seven rows of the county data set are shown in Table 1.5, and the variables are summarized in Table 1.6. These data were collected from the US Census website.[6]

[4] A case is also sometimes called a **unit of observation** or an **observational unit**.

[5] Each county may be viewed as a case, and there are eleven pieces of information recorded for each case. A table with 3,143 rows and 11 columns could hold these data, where each row represents a county and each column represents a particular piece of information.

[6] http://quickfacts.census.gov/qfd/index.html

	name	state	pop2000	pop2010	fed_spend	poverty	homeownership	multiunit	income	med_income	smoking_ban
1	Autauga	AL	43671	54571	6.068	10.6	77.5	7.2	24568	53255	none
2	Baldwin	AL	140415	182265	6.140	12.2	76.7	22.6	26469	50147	none
3	Barbour	AL	29038	27457	8.752	25.0	68.0	11.1	15875	33219	none
4	Bibb	AL	20826	22915	7.122	12.6	82.9	6.6	19918	41770	none
5	Blount	AL	51024	57322	5.131	13.4	82.0	3.7	21070	45549	none
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Washakie	WY	8289	8533	8.714	5.6	70.9	10.0	28557	48379	none
3143	Weston	WY	6644	7208	6.695	7.9	77.9	6.5	28463	53853	none

Table 1.5: Seven rows from the `county` data set.

variable	description
`name`	County name
`state`	State where the county resides (also including the District of Columbia)
`pop2000`	Population in 2000
`pop2010`	Population in 2010
`fed_spend`	Federal spending per capita
`poverty`	Percent of the population in poverty
`homeownership`	Percent of the population that lives in their own home or lives with the owner (e.g. children living with parents who own the home)
`multiunit`	Percent of living units that are in multi-unit structures (e.g. apartments)
`income`	Income per capita
`med_income`	Median household income for the county, where a household's income equals the total income of its occupants who are 15 years or older
`smoking_ban`	Type of county-wide smoking ban in place at the end of 2011, which takes one of three values: `none`, `partial`, or `comprehensive`, where a `comprehensive` ban means smoking was not permitted in restaurants, bars, or workplaces, and `partial` means smoking was banned in at least one of those three locations

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

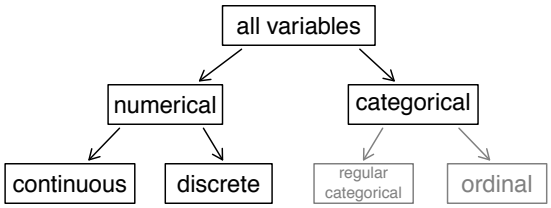


Figure 1.7: Breakdown of variables into their respective types.

1.2.2 Types of variables

Examine the `fed_spend`, `pop2010`, `state`, and `smoking_ban` variables in the `county` data set. Each of these variables is inherently different from the other three yet many of them share certain characteristics.

First consider `fed_spend`, which is said to be a **numerical** variable since it can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those 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 meaning.

The `pop2010` variable is also numerical, although it seems to be a little different than `fed_spend`. This variable of the population count can only take whole non-negative numbers (`0`, `1`, `2`, ...). For this reason, the population variable is said to be **discrete** since it can only take numerical values with jumps. On the other hand, the federal spending variable is said to be **continuous**.

The variable `state` can take up to 51 values after accounting for Washington, DC: `AL`, ..., and `WY`. Because the responses themselves are categories, `state` is called a **categorical** variable,[7] and the possible values are called the variable's **levels**.

Finally, consider the `smoking_ban` variable, which describes the type of county-wide smoking ban and takes values `none`, `partial`, or `comprehensive` in each county. This variable seems to be a hybrid: it is a categorical variable but the levels have a natural ordering. A variable with these properties is called an **ordinal** variable. To simplify analyses, any ordinal variables in this book will be treated as categorical variables.

● **Example 1.3** Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

The number of siblings and student height represent numerical variables. Because the number of siblings is a count, it is discrete. Height varies continuously, so it is a continuous numerical variable. The last variable classifies students into two categories – those who have and those who have not taken a statistics course – which makes this variable categorical.

⊙ **Exercise 1.4** Consider the variables `group` and `outcome` (at 30 days) from the stent study in Section 1.1. Are these numerical or categorical variables?[8]

[7] Sometimes also called a **nominal** variable.

[8] There are only two possible values for each variable, and in both cases they describe categories. Thus, each are categorical variables.

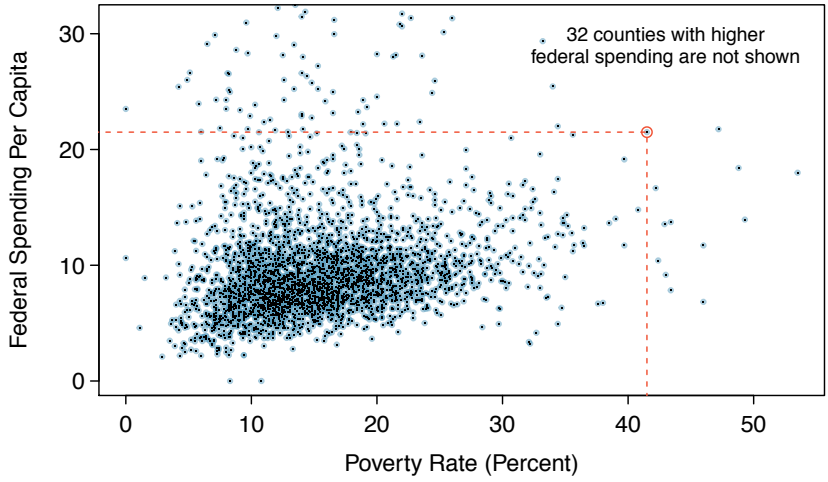


Figure 1.8: A scatterplot showing `fed_spend` against `poverty`. Owsley County of Kentucky, with a poverty rate of 41.5% and federal spending of $21.50 per capita, is highlighted.

1.2.3 Relationships between variables

Many analyses are motivated by a researcher looking for a relationship between two or more variables. A social scientist may like to answer some of the following questions:

(1) Is federal spending, on average, higher or lower in counties with high rates of poverty?

(2) If homeownership is lower than the national average in one county, will the percent of multi-unit structures in that county likely be above or below the national average?

(3) Which counties have a higher average income: those that enact one or more smoking bans or those that do not?

To answer these questions, data must be collected, such as the `county` data set shown in Table 1.5. Examining summary statistics could provide insights for each of the three questions about counties. Additionally, graphs can be used to visually summarize data and are useful for answering such questions as well.

Scatterplots are one type of graph used to study the relationship between two numerical variables. Figure 1.8 compares the variables `fed_spend` and `poverty`. Each point on the plot represents a single county. For instance, the highlighted dot corresponds to County 1088 in the `county` data set: Owsley County, Kentucky, which had a poverty rate of 41.5% and federal spending of $21.50 per capita. The scatterplot suggests a relationship between the two variables: counties with a high poverty rate also tend to have slightly more federal spending. We might brainstorm as to why this relationship exists and investigate each idea to determine which is the most reasonable explanation.

⊙ **Exercise 1.5** Examine the variables in the `email50` data set, which are described in Table 1.4 on page 4. Create two questions about the relationships between these variables that are of interest to you.[9]

[9]Two sample questions: (1) Intuition suggests that if there are many line breaks in an email then there would tend to also be many characters: does this hold true? (2) Is there a connection between whether an email format is plain text (versus HTML) and whether it is a spam message?

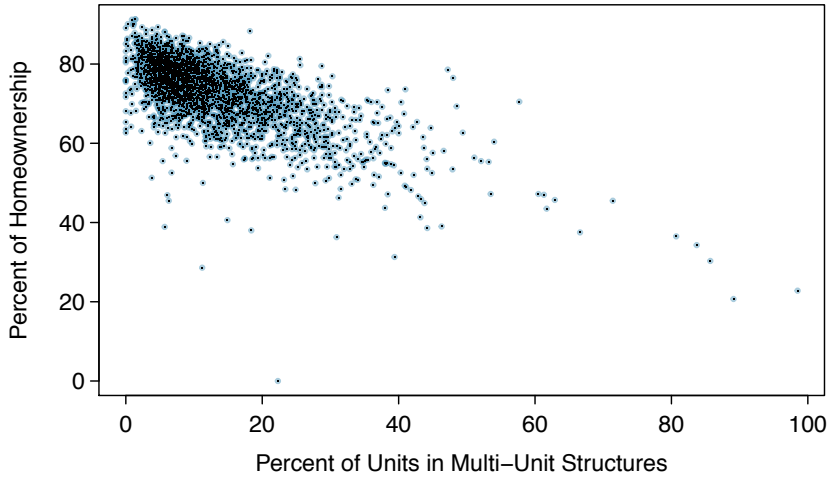


Figure 1.9: A scatterplot of homeownership versus the percent of units that are in multi-unit structures for all 3,143 counties. Interested readers may find an image of this plot with an additional third variable, county population, presented at www.openintro.org/stat/down/MHP.png.

The `fed_spend` and `poverty` variables are said to be associated because the plot shows a discernible pattern. When two variables show some connection with one another, they are called **associated** variables. Associated variables can also be called **dependent** variables and vice-versa.

Example 1.6 This example examines the relationship between homeownership and the percent of units in multi-unit structures (e.g. apartments, condos), which is visualized using a scatterplot in Figure 1.9. Are these variables associated?

It appears that the larger the fraction of units in multi-unit structures, the lower the homeownership rate. Since there is some relationship between the variables, they are associated.

Because there is a downward trend in Figure 1.9 – counties with more units in multi-unit structures are associated with lower homeownership – these variables are said to be **negatively associated**. A **positive association** is shown in the relationship between the `poverty` and `fed_spend` variables represented in Figure 1.8, where counties with higher poverty rates tend to receive more federal spending per capita.

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.

Associated or independent, not both
A pair of variables are either related in some way (associated) or not (independent). No pair of variables is both associated and independent.

1.3 Overview of data collection principles

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases should be studied and what variables are important. It is also important to consider *how* data are collected so that they are reliable and help achieve the research goals.

1.3.1 Populations and samples

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last 5 years, what is the average time to degree for Duke undergraduate students?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each research question 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. Often times, it is too expensive to collect data for every case in a population. Instead, a sample is taken. 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 to provide an estimate of the population average and answer the research question.

⊙ **Exercise 1.7** For the second and third questions above, identify the target population and what represents an individual case.[10]

1.3.2 Anecdotal evidence

Consider the following possible responses to the three research questions:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Each of the conclusions are based on some data. However, there are two problems. First, the data only represent one or two cases. Second, and more importantly, it is unclear whether these cases are actually representative of the population. Data collected in this haphazard fashion are called **anecdotal evidence**.

[10] (2) Notice that the first question is only relevant to students who complete their degree; the average cannot be computed using a student who never finished her degree. Thus, only Duke undergraduate students who have graduated in the last five years represent cases in the population under consideration. Each such student would represent an individual case. (3) A person with severe heart disease represents a case. The population includes all people with severe heart disease.

Figure 1.10: In February 2010, some media pundits cited one large snow storm as valid evidence against global warming. As comedian Jon Stewart pointed out, "It's one storm, in one region, of one country."

February 10th, 2010.

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.

Anecdotal evidence typically is composed of unusual cases that we recall based on their striking characteristics. For instance, we are more likely to remember the two people we met who took 7 years to graduate than the six others who graduated in four years. Instead of looking at the most unusual cases, we should examine a sample of many cases that represent the population.

1.3.3 Sampling from a population

We might try to estimate the time to graduation for Duke undergraduates in the last 5 years by collecting a sample of students. All graduates in the last 5 years represent the *population*, and graduates who are selected for review are collectively called the *sample*. In general, we always seek to *randomly* select a sample from a population. The most basic type of random selection is equivalent to how raffles are conducted. For example, in selecting graduates, we could write each graduate's name on a raffle ticket and draw 100 tickets. The selected names would represent a random sample of 100 graduates.

Why pick a sample randomly? Why not just pick a sample by hand? Consider the following scenario.

Example 1.8 Suppose we ask a student who happens to be majoring in nutrition to select several graduates for the study. What kind of students do you think she might collect? Do you think her sample would be representative of all graduates?

Perhaps she would pick a disproportionate number of graduates from health-related fields. Or perhaps her selection would be well-representative of the population. When selecting samples by hand, we run the risk of picking a *biased* sample, even if that bias is unintentional or difficult to discern.

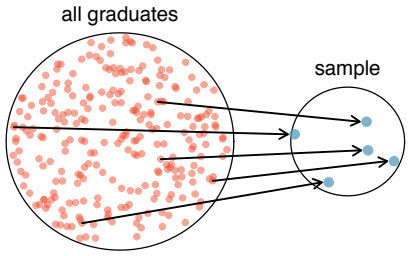


Figure 1.11: In this graphic, five graduates are randomly selected from the population to be included in the sample.

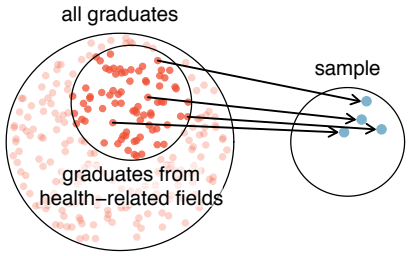


Figure 1.12: Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionally often.

If someone was permitted to pick and choose exactly which graduates were included in the sample, it is entirely possible that the sample could be skewed to that person's interests, which may be entirely unintentional. This introduces **bias** into a sample. Sampling randomly helps resolve this problem. The most basic random sample is called a **simple random sample**, and it is the equivalent of using a raffle to select cases. This means that each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample.

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.

Another common downfall is a **convenience sample**, where individuals who are easily accessible are more likely to be included in the sample. For instance, if a political survey is done by stopping people walking in the Bronx, this will not represent all of New York City. It is often difficult to discern what sub-population a convenience sample represents.

⊙ **Exercise 1.9** 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?[11]

[11] 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 should data on the subject become available.

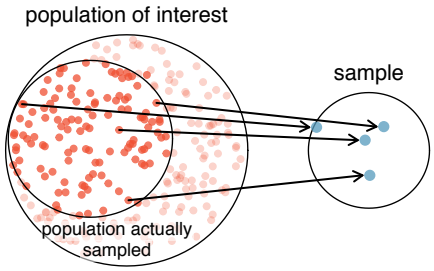


Figure 1.13: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

1.3.4 Explanatory and response variables

Consider the following question from page 7 for the `county` data set:

(1) Is federal spending, on average, higher or lower in counties with high rates of poverty?

If we suspect poverty might affect spending in a county, then poverty is the **explanatory** variable and federal spending is the **response** variable in the relationship.[12] If there are many variables, it may be possible to consider a number of them as explanatory variables.

TIP: Explanatory and response variables
To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other and plan an appropriate analysis.

explanatory variable → might affect → response variable

Caution: association does not imply causation
Labeling variables as *explanatory* and *response* does not guarantee the relationship between the two is actually 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 affects the other.

In some cases, there is no explanatory or response variable. Consider the following question from page 7:

(2) If homeownership is lower than the national average in one county, will the percent of multi-unit structures in that county likely be above or below the national average?

It is difficult to decide which of these variables should be considered the explanatory and response variable, i.e. the direction is ambiguous, so no explanatory or response labels are suggested here.

[12] Sometimes the explanatory variable is called the **independent** variable and the response variable is called the **dependent** variable. However, this becomes confusing since a *pair* of variables might be independent or dependent, so we avoid this language.

1.3.5 Introducing observational studies and experiments

There are two primary types of data collection: observational studies and experiments.

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. In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.

When researchers want to investigate the possibility of a causal connection, they conduct an **experiment**. Usually there will be both an explanatory and a response variable. For instance, we may suspect administering a drug will reduce mortality in heart attack patients over the following year. To check if there really is a causal connection between the explanatory variable and the response, researchers will collect a sample of individuals and split them into groups. The individuals in each group are *assigned* a treatment. When individuals are randomly assigned to a group, the experiment is called a **randomized experiment**. For example, each heart attack patient in the drug trial could be randomly assigned, perhaps by flipping a coin, into one of two groups: the first group receives a **placebo** (fake treatment) and the second group receives the drug. See the case study in Section 1.1 for another example of an experiment, though that study did not employ a placebo.

TIP: association $\neq$ causation
In general, association does not imply causation, and causation can only be inferred from a randomized experiment.

1.4 Observational studies and sampling strategies

1.4.1 Observational studies

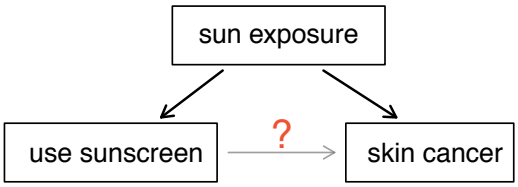
Generally, data in observational studies are collected only by monitoring what occurs, while experiments require the primary explanatory variable in a study be assigned for each subject by the researchers.

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.

⊙ **Exercise 1.10** 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?[13]

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.

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



Sun exposure is what is called a **confounding variable**,[14] 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.

In the same way, the `county` data set is an observational study with confounding variables, and its data cannot easily be used to make causal conclusions.

⊙ **Exercise 1.11** Figure 1.9 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 1.9.[15]

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.[16] 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. retails sales) while the federal government retrospectively collected others during the 2010 census (e.g. county population counts).

1.4.2 Three sampling methods (special topic)

Almost all statistical methods are based on the notion of implied randomness. If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable. Here we consider three random sampling techniques: simple, stratified, and cluster sampling. Figure 1.14 provides a graphical representation of these techniques.

Simple random sampling is probably the most intuitive form of random sampling. Consider the salaries of Major League Baseball (MLB) players, where each player is a member of one of the league's 30 teams. To take a simple random sample of 120 baseball players and their salaries from the 2010 season, we could write the names of that season's 828 players onto slips of paper, drop the slips into a bucket, shake the bucket around until

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

[15] Answers 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.

[16] http://www.channing.harvard.edu/nhs/

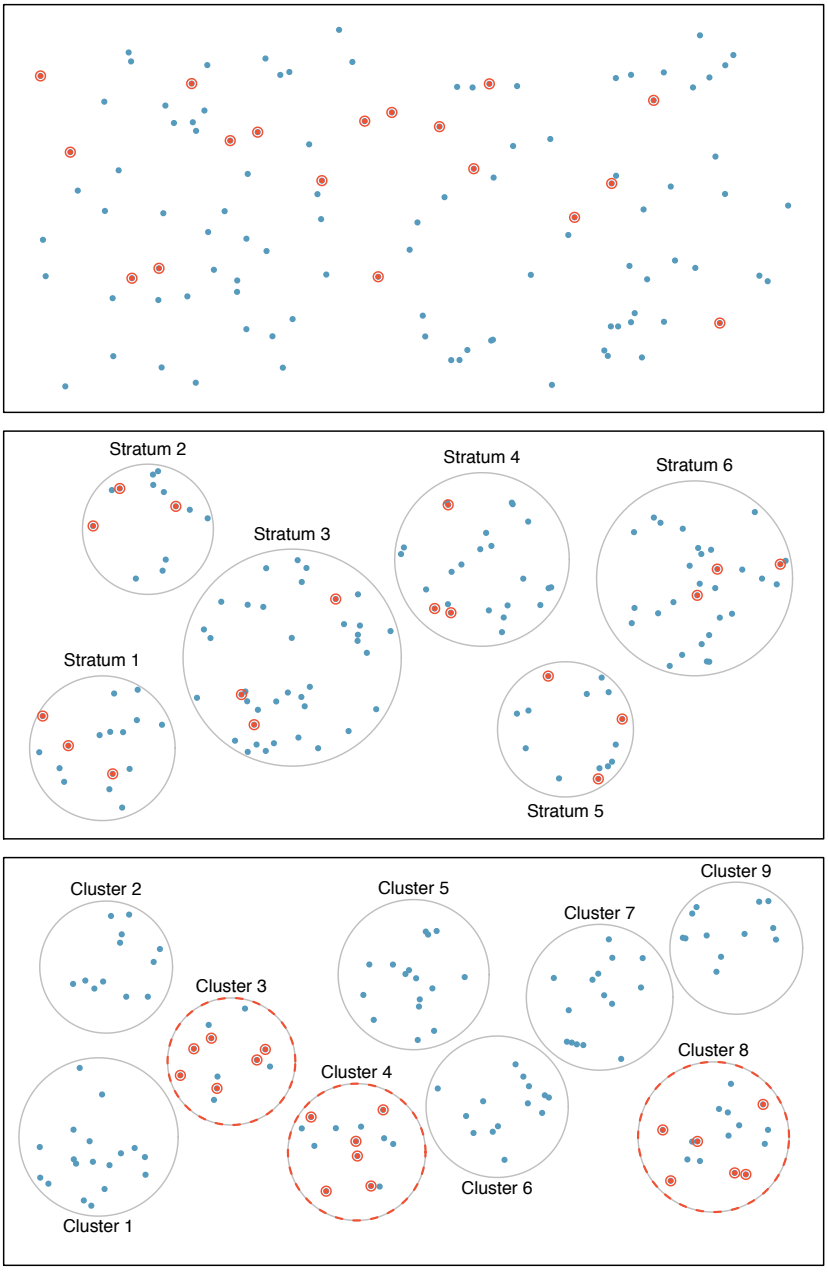


Figure 1.14: Examples of simple random, stratified, and cluster sampling. In the top panel, simple random sampling was used to randomly select the 18 cases. In the middle panel, stratified sampling was used: cases were grouped into strata, and then simple random sampling was employed within each stratum. In the bottom panel, cluster sampling was used, where data were binned into nine clusters, three of the clusters were randomly selected, and six cases were randomly sampled in each of these clusters.

we are sure the names are all mixed up, then draw out slips until we have the sample of 120 players. In general, a sample is referred to as "simple random" if each case in the population has an equal chance of being included in the final sample *and* knowing that a case is included in a sample does not provide useful information about which other cases are included.

Stratified sampling is a divide-and-conquer sampling strategy. The population is divided into groups called **strata**. The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each stratum. In the baseball salary example, the teams could represent the strata; some teams have a lot more money (we're looking at you, Yankees). Then we might randomly sample 4 players from each team for a total of 120 players.

Stratified sampling is especially useful when the cases in each stratum are very similar with respect to the outcome of interest. The downside is that analyzing data from a stratified sample is a more complex task than analyzing data from a simple random sample. The analysis methods introduced in this book would need to be extended to analyze data collected using stratified sampling.

Example 1.12 Why would it be good for cases within each stratum to be very similar?

We might get a more stable estimate for the subpopulation in a stratum if the cases are very similar. These improved estimates for each subpopulation will help us build a reliable estimate for the full population.

A **cluster sample** is much like a two-stage simple random sample. We break up the population into many groups, called **clusters**. Then we sample a fixed number of clusters and collect a simple random sample within each cluster. This technique is similar to stratified sampling in its process, except that there is no requirement in cluster sampling to sample from every cluster. Stratified sampling requires observations be sampled from every stratum.

Sometimes cluster sampling can be a more economical random sampling technique than the alternatives. Also, unlike stratified sampling, cluster sampling is most helpful when there is a lot of case-to-case variability within a cluster but the clusters themselves don't look very different from one another. For example, if neighborhoods represented clusters, then this sampling method works best when the neighborhoods are very diverse. A downside of cluster sampling is that more advanced analysis techniques are typically required, though the methods in this book can be extended to handle such data.

Example 1.13 Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What sampling method should be employed?

A simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive. Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals. However, cluster sampling seems like a very good idea. First, we might randomly select half the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample and would still give us reliable information.

1.5 Experiments

Studies where the researchers assign treatments to cases are called **experiments**. When this assignment includes randomization, e.g. using a coin flip to decide which treatment a patient receives, it is called a **randomized experiment**. Randomized experiments are fundamentally important when trying to show a causal connection between two variables.

1.5.1 Principles of experimental design

Randomized experiments are generally built on four principles.

Controlling. Researchers assign treatments to cases, and they do their best to **control** any other differences in the groups. For example, when patients take a drug in pill form, some patients take the pill with only a sip of water while others may have it with an entire glass of water. To control for water consumption, a doctor may ask all patients to drink a 12 ounce glass of water with the pill.

Randomization. Researchers randomize patients into treatment groups to account for variables that cannot be controlled. For example, some patients may be more susceptible to a disease than others due to their dietary habits. Randomizing patients into the treatment or control group helps even out such differences, and it also prevents accidental bias from entering the study.

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. Additionally, a group of scientists may replicate an entire study to verify an earlier finding.

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 1.15. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

It is important to incorporate 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 this book may be extended to analyze data collected using blocking.

1.5.2 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.[17] In particular, researchers wanted to know if the drug reduced deaths in patients.

[17] Anturane Reinfarction Trial Research Group. 1980. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. New England Journal of Medicine 302(5):250-256.

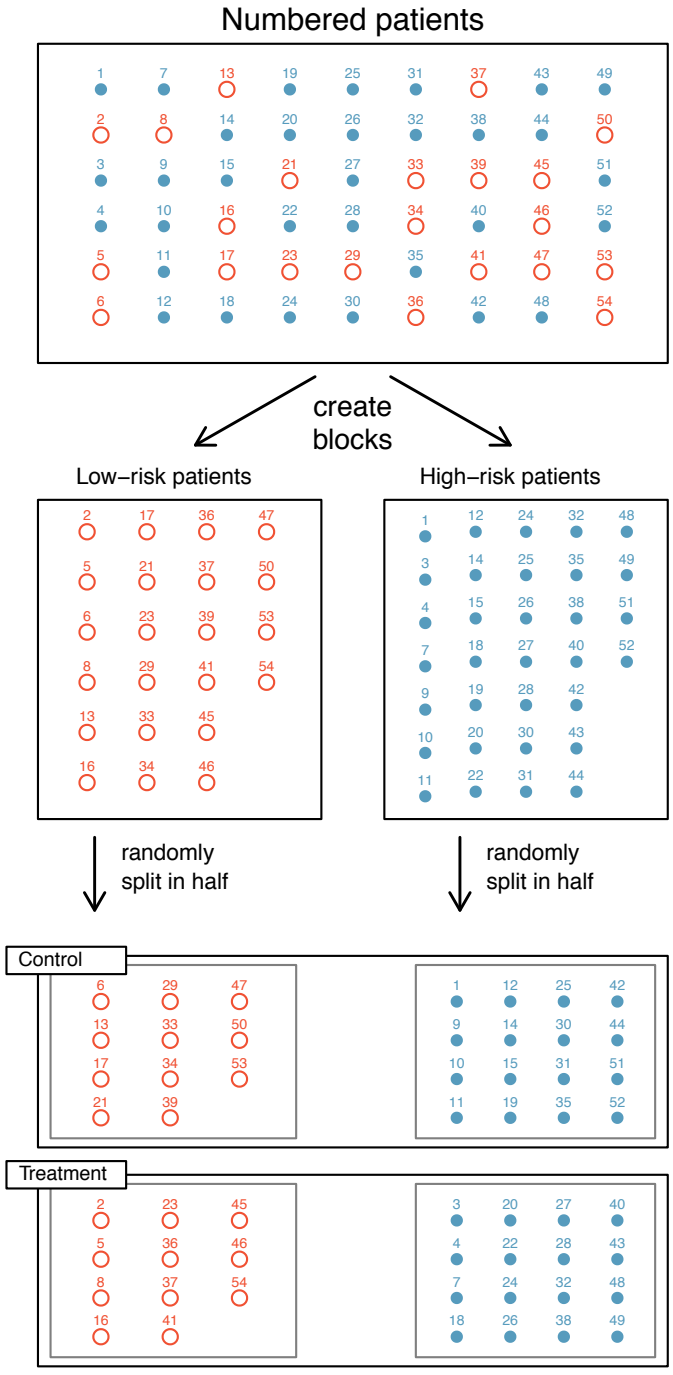


Figure 1.15: Blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly divided into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both the low-risk and high-risk categories.

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[18] 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 fancy new drug that you anticipate will help you. On the other hand, a person in the other group doesn't receive the drug and sits idly, hoping her participation doesn't increase her risk of death. These perspectives suggest there are actually two effects: the one of interest is the effectiveness of the drug, and the second is an emotional effect that is difficult to quantify.

Researchers aren't usually interested in the emotional effect, which might bias the study. To circumvent this problem, researchers 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 **blind**. But there is one problem: if a patient doesn't receive a treatment, she will know she is in the control group. The solution to this problem is to give fake treatments to patients in the control group. A fake treatment is called a **placebo**, and an effective placebo is the key to making a study truly blind. A classic example of a placebo is a sugar pill that is made to look like the actual treatment pill. Often times, a placebo results in a slight but real improvement in patients. This effect has been dubbed the **placebo effect**.

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. 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 treatment.[19]

⊙ **Exercise 1.14** Look back to the study in Section 1.1 where researchers were testing whether stents were effective at reducing strokes in at-risk patients. Is this an experiment? Was the study blinded? Was it double-blinded?[20]

1.6 Examining numerical data

In this section we will be introduced to techniques for exploring and summarizing numerical variables. The `email50` and `county` data sets from Section 1.2 provide rich opportunities for examples. Recall that outcomes of numerical variables are numbers on which it is reasonable to perform basic arithmetic operations. For example, the `pop2010` variable, which represents the populations of counties in 2010, is numerical since we can sensibly discuss the difference or ratio of the populations in two counties. On the other hand, area codes and zip codes are not numerical, but rather they are categorical variables.

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

[19] 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.

[20] 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.

1.6.1 Scatterplots for paired data

A **scatterplot** provides a case-by-case view of data for two numerical variables. In Figure 1.8 on page 7, a scatterplot was used to examine how federal spending and poverty were related in the `county` data set. Another scatterplot is shown in Figure 1.16, comparing the number of line breaks (`line_breaks`) and number of characters (`num_char`) in emails for the `email50` data set. In any scatterplot, each point represents a single case. Since there are 50 cases in `email50`, there are 50 points in Figure 1.16.

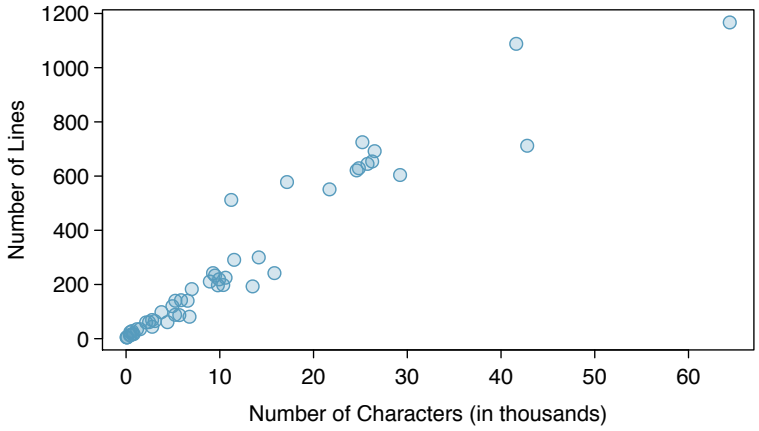


Figure 1.16: A scatterplot of `line_breaks` versus `num_char` for the `email50` data.

To put the number of characters in perspective, this paragraph has 363 characters. Looking at Figure 1.16, it seems that some emails are incredibly verbose! Upon further investigation, we would actually find that most of the long emails use the HTML format, which means most of the characters in those emails are used to format the email rather than provide text.

⊙ **Exercise 1.15** What do scatterplots reveal about the data, and how might they be useful?[21]

● **Example 1.16** Consider a new data set of 54 cars with two variables: vehicle price and weight.[22] A scatterplot of vehicle price versus weight is shown in Figure 1.17. 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 1.8 on page 7 and Figure 1.16, which show relationships that are very linear.

⊙ **Exercise 1.17** Describe two variables that would have a horseshoe shaped association in a scatterplot.[23]

[21] Answers may vary. Scatterplots are helpful in quickly spotting associations relating variables, whether those associations come in the form of simple trends or whether those relationships are more complex.

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

[23] Consider the case where your vertical axis represents something "good" and your horizontal axis represents something that is only good in moderation. Health and water consumption fit this description since water becomes toxic when consumed in excessive quantities.

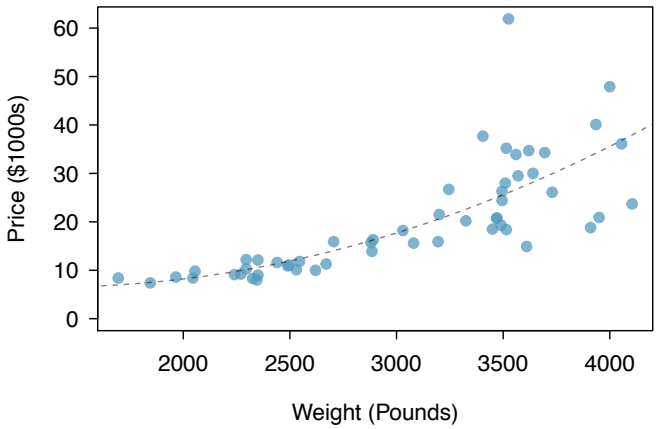


Figure 1.17: A scatterplot of `price` versus `weight` for 54 cars.

1.6.2 Dot plots and the mean

Sometimes two variables is one too many: only one variable may be of interest. In these cases, a dot plot provides the most basic of displays. A **dot plot** is a one-variable scatterplot; an example using the number of characters from 50 emails is shown in Figure 1.18. A stacked version of this dot plot is shown in Figure 1.19.

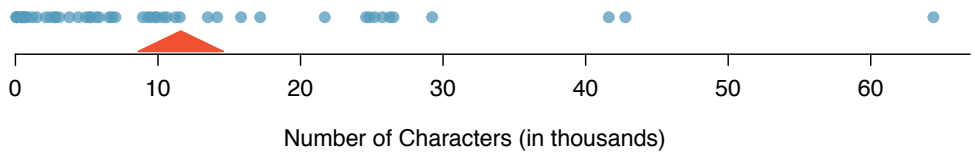


Figure 1.18: A dot plot of `num_char` for the `email50` data set.

The **mean**, sometimes called the average, is a common way to measure the center of a **distribution** of data. To find the mean number of characters in the 50 emails, we add up all the character counts and divide by the number of emails. For computational convenience, the number of characters is listed in the thousands and rounded to the first decimal.

$$\bar{x} = \frac{21.7 + 7.0 + \cdots + 15.8}{50} = 11.6 \tag{1.18}$$

The sample mean is often labeled $\bar{x}$. The letter x is being used as a generic placeholder for the variable of interest, `num_char`, and the bar says it is the average number of characters in the 50 emails was 11,600. It is useful to think of the mean as the balancing point of the distribution. The sample mean is shown as a triangle in Figures 1.18 and 1.19.

$\bar{x}$
sample mean

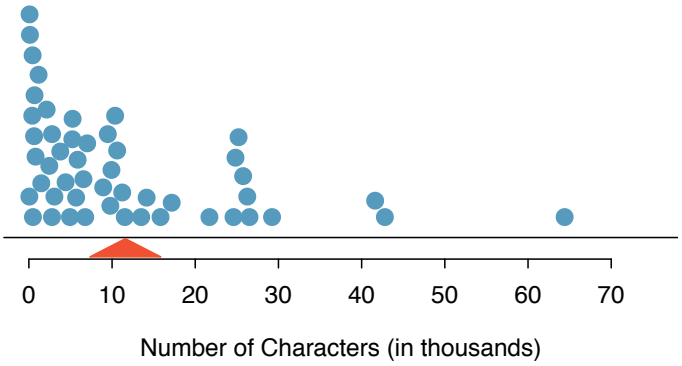


Figure 1.19: A stacked dot plot of `num_char` for the `email50` data set.

Mean

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

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1.19}$$

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

n
sample size

⊙ **Exercise 1.20** Examine Equations (1.18) and (1.19) above. What does x_1 correspond to? And x_2? Can you infer a general meaning to what x_i might represent?[24]

⊙ **Exercise 1.21** What was n in this sample of emails?[25]

The `email50` data set represents a sample from a larger population of emails that were received in January and March. We could compute a mean for this population in the same way as the sample mean, however, the population mean has a special label: μ. The symbol μ is the Greek letter *mu* and represents the average of all observations in the population. Sometimes a subscript, such as $_x$, is used to represent which variable the population mean refers to, e.g. μ_x.

μ
population mean

● **Example 1.22** The average number of characters across all emails can be estimated using the sample data. Based on the sample of 50 emails, what would be a reasonable estimate of μ_x, the mean number of characters in all emails in the `email` data set? (Recall that `email50` is a sample from `email`.)

The sample mean, 11,600, may provide a reasonable estimate of μ_x. While this number will not be perfect, it provides a *point estimate* of the population mean. In Chapter 4 and beyond, we will develop tools to characterize the accuracy of point estimates, and we will find that point estimates based on larger samples tend to be more accurate than those based on smaller samples.

[24] x_1 corresponds to the number of characters in the first email in the sample (21.7, in thousands), x_2 to the number of characters in the second email (7.0, in thousands), and x_i corresponds to the number of characters in the i^{th} email in the data set.

[25] The sample size was $n = 50$.

Example 1.23 We might like to compute the average income per person in the US. To do so, we might first think to take the mean of the per capita incomes across the 3,143 counties in the `county` data set. What would be a better approach?

The `county` data set is special in that each county actually represents many individual people. If we were to simply average across the `income` variable, we would be treating counties with 5,000 and 5,000,000 residents equally in the calculations. Instead, we should compute the total income for each county, add up all the counties' totals, and then divide by the number of people in all the counties. If we completed these steps with the `county` data, we would find that the per capita income for the US is $27,348.43. Had we computed the *simple* mean of per capita income across counties, the result would have been just $22,504.70!

Example 1.23 used what is called a **weighted mean**, which will not be a key topic in this textbook. However, we have provided an online supplement on weighted means for interested readers:

http://www.openintro.org/stat/down/supp/wtdmean.pdf

1.6.3 Histograms and shape

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 1.20. These binned counts are plotted as bars in Figure 1.21 into what is called a **histogram**, which resembles the stacked dot plot shown in Figure 1.19.

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 1.20: 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 1.21 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**.[26]

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**.

[26] 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**.

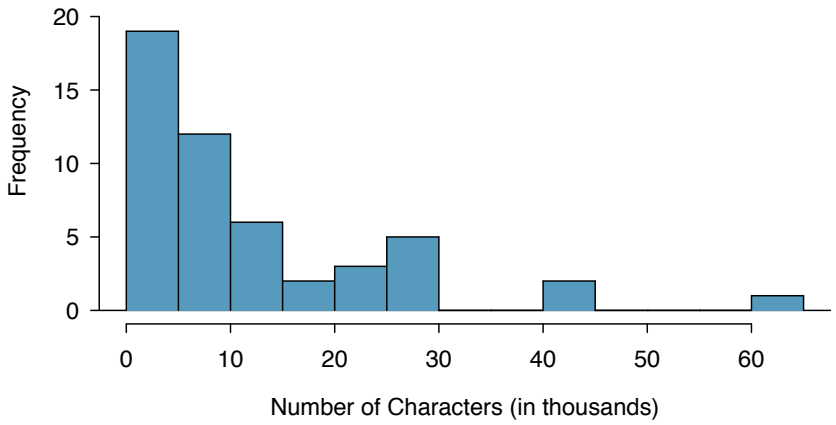


Figure 1.21: A histogram of `num_char`. This distribution is very strongly skewed to the right.

Long tails to identify skew
When data trail off in one direction, the distribution has a **long tail**. If a distribution has a long left tail, it is left skewed. If a distribution has a long right tail, it is right skewed.

⊙ **Exercise 1.24** Take a look at the dot plots in Figures 1.18 and 1.19. Can you see the skew in the data? Is it easier to see the skew in this histogram or the dot plots?[27]

⊙ **Exercise 1.25** Besides the mean (since it was labeled), what can you see in the dot plots that you cannot see in the histogram?[28]

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.[29] There is only one prominent peak in the histogram of `num_char`.

Figure 1.22 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.

⊙ **Exercise 1.26** Figure 1.21 reveals only one prominent mode in the number of characters. Is the distribution unimodal, bimodal, or multimodal?[30]

[27] The skew is visible in all three plots, though the flat dot plot is the least useful. The stacked dot plot and histogram are helpful visualizations for identifying skew.

[28] Character counts for individual emails.

[29] 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.

[30] 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.)

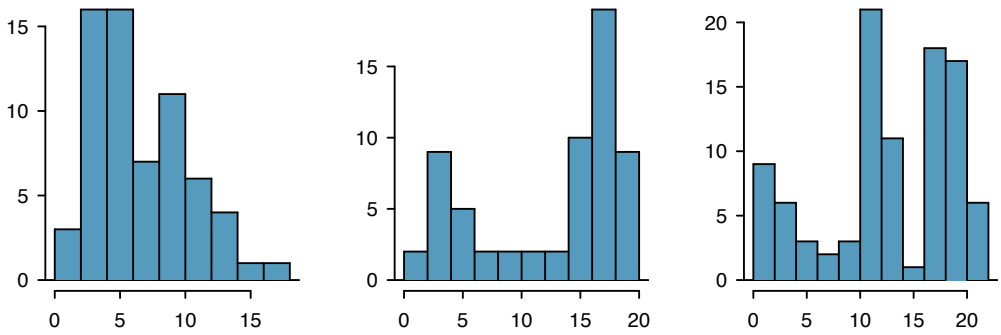


Figure 1.22: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

⊙ **Exercise 1.27** Height measurements of young students and adult teachers at a K-3 elementary school were taken. How many modes would you anticipate in this height data set?[31]

TIP: Looking for modes
Looking for modes isn't about finding a clear and correct answer about the number of modes in a distribution, which is why *prominent* is not rigorously defined in this book. The important part of this examination is to better understand your data and how it might be structured.

1.6.4 Variance and standard deviation

The mean was introduced as a method to describe the center of a data set, but the variability in the data is also important. Here, we introduce two measures of variability: the variance and the standard deviation. Both of these are very useful in data analysis, even though their formulas are a bit tedious to calculate by hand. The standard deviation is the easier of the two to understand, and it roughly describes how far away the typical observation is from the mean.

We call the distance of an observation from its mean its **deviation**. Below are the deviations for the 1^{st}, 2^{nd}, 3^{rd}, and 50^{th} observations in the `num_char` variable. For computational convenience, the number of characters is listed in the thousands and rounded to the first decimal.

$$\begin{aligned} x_1 - \bar{x} &= 21.7 - 11.6 = 10.1 \\ x_2 - \bar{x} &= 7.0 - 11.6 = -4.6 \\ x_2 - \bar{x} &= 0.6 - 11.6 = -11.0 \\ &\vdots \\ x_{50} - \bar{x} &= 15.8 - 11.6 = 4.2 \end{aligned}$$

[31] There might be two height groups visible in the data set: one of the students and one of the adults. That is, the data are probably bimodal.

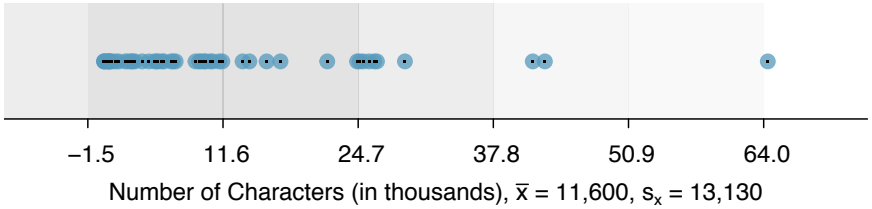


Figure 1.23: In the `num_char` data, 41 of the 50 emails (82%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% are within 2 standard deviations, though this rule of thumb is less accurate for skewed data, as shown in this example.

s^2
sample variance

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

$$\begin{aligned} s^2 &= \frac{10.1^2 + (-4.6)^2 + (-11.0)^2 + \cdots + 4.2^2}{50 - 1} \\ &= \frac{102.01 + 21.16 + 121.00 + \cdots + 17.64}{49} \\ &= 172.44 \end{aligned}$$

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. Notice that squaring the deviations does two things. First, it makes large values much larger, seen by comparing 10.1^2, $(-4.6)^2$, $(-11.0)^2$, and 4.2^2. Second, it gets rid of any negative signs.

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

s
sample standard deviation

$$s = \sqrt{172.44} = 13.13$$

The standard deviation of the number of characters in an email is about 13.13 thousand. A subscript of $_x$ may be added to the variance and standard deviation, i.e. s_x^2 and s_x, as a reminder that these are the variance and standard deviation of the observations represented by x_1, x_2, ..., x_n. The $_x$ subscript is usually omitted when it is clear which data the variance or standard deviation is referencing.

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.

σ^2
population variance

σ
population standard deviation

Formulas and methods used to compute the variance and standard deviation for a population are similar to those used for a sample.[32] However, like the mean, the population values have special symbols: σ^2 for the variance and σ for the standard deviation. The symbol σ is the Greek letter *sigma*.

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

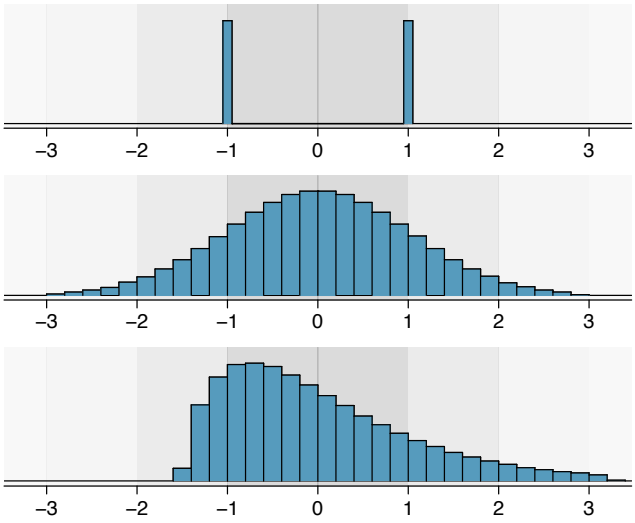


Figure 1.24: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

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 1.23 and 1.24, these percentages are not strict rules.

⊙ **Exercise 1.28** On page 23, 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 1.24 as an example, explain why such a description is important.[33]

● **Example 1.29** Describe the distribution of the `num_char` variable using the histogram in Figure 1.21 on page 24. 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 4 we will use the variance and standard deviation to assess how close the sample mean is to the population mean.

[33] Figure 1.24 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.

1.6.5 Box plots, quartiles, and the median

A **box plot** summarizes a data set using five statistics while also plotting unusual observations. Figure 1.25 provides a vertical dot plot alongside a box plot of the `num_char` variable from the `email50` data set.

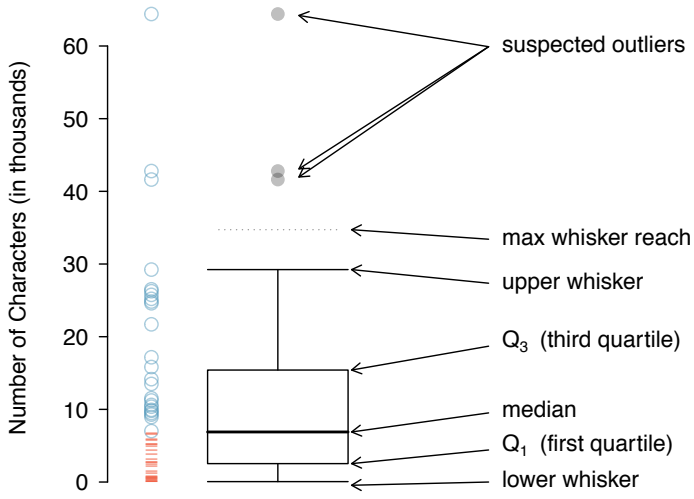


Figure 1.25: A vertical dot plot next to a labeled box plot for the number of characters in 50 emails. The median (6,890), splits the data into the bottom 50% and the top 50%, marked in the dot plot by horizontal dashes and open circles, respectively.

The first step in building a box plot is drawing a dark line denoting the **median**, which splits the data in half. Figure 1.25 shows 50% of the data falling below the median (dashes) and other 50% falling above the median (open circles). There are 50 character counts in the data set (an even number) so the data are perfectly split into two groups of 25. We take the median in this case to be the average of the two observations closest to the 50^{th} percentile: $(6{,}768 + 7{,}012)/2 = 6{,}890$. When there are an odd number of observations, there will be exactly one observation that splits the data into two halves, and in this case that observation is the median (no average needed).

Median: the number in the middle
If the data are ordered from smallest to largest, the **median** is the observation right in the middle. If there are an even number of observations, there will be two values in the middle, and the median is taken as their average.

The second step in building a box plot is drawing a rectangle to represent the middle 50% of the data. The total length of the box, shown vertically in Figure 1.25, is called the **interquartile range** (IQR, for short). It, like the standard deviation, is a measure of variability in data. The more variable the data, the larger the standard deviation and IQR. The two boundaries of the box are called 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), and these are often labeled Q_1 and Q_3, respectively.

Interquartile range (IQR)
The IQR is the length of the box in a box plot. It is computed as

$$IQR = Q_3 - Q_1$$

where Q_1 and Q_3 are the 25^{th} and 75^{th} percentiles.

⊙ **Exercise 1.30** What percent of the data fall between Q_1 and the median? What percent is between the median and Q_3?[34]

Extending out from the box, the **whiskers** attempt to capture the data outside of the box, however, their reach is never allowed to be more than $1.5 \times IQR$.[35] They capture everything within this reach. In Figure 1.25, 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 – instead of just extending the whiskers to the minimum and maximum observed values – is to help identify any observations that appear to be unusually distant from the rest of the data. Unusually distant observations are called **outliers**. 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 are extreme
An **outlier** is an observation that appears extreme relative to the rest of the data.

TIP: Why it is important to look for outliers
Examination of data for possible outliers serves many useful purposes, including

1. Identifying strong skew in the distribution.
2. Identifying data collection or entry errors. For instance, we re-examined the email purported to have 64,401 characters to ensure this value was accurate.
3. Providing insight into interesting properties of the data.

⊙ **Exercise 1.31** The observation 64,401, a suspected outlier, was found to be an accurate observation. What would such an observation suggest about the nature of character counts in emails?[36]

⊙ **Exercise 1.32** Using Figure 1.25, estimate the following values for `num_char` in the `email50` data set: (a) Q_1, (b) Q_3, and (c) IQR.[37]

[34] Since Q_1 and Q_3 capture the middle 50% of the data and the median splits the data in the middle, 25% of the data fall between Q_1 and the median, and another 25% falls between the median and Q_3.

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

[36] That occasionally there may be very long emails.

[37] These visual estimates will vary a little from one person to the next: Q_1 = 3,000, Q_3 = 15,000, IQR = $Q_3 - Q_1$ = 12,000. (The true values: Q_1 = 2,536, Q_3 = 15,411, IQR = 12,875.)

1.6.6 Robust statistics

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 1.26, and sample statistics are computed under each scenario in Table 1.27.

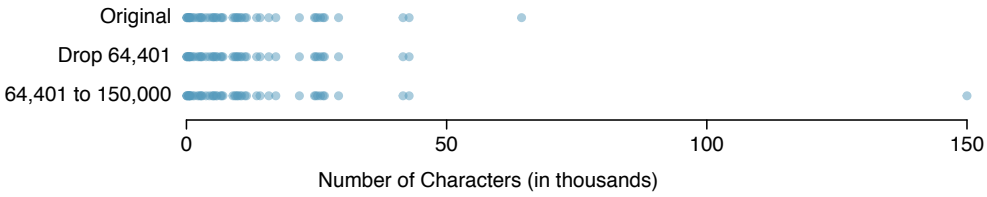


Figure 1.26: Dot plots of the original character count data and two modified data sets.

	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original `num_char` 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 1.27: A comparison of how the median, IQR, mean ($\bar{x}$), and standard deviation (s) change when extreme observations are present.

⊙ **Exercise 1.33** (a) Which is more affected by extreme observations, the mean or median? Table 1.27 may be helpful. (b) Is the standard deviation or IQR more affected by extreme observations?[38]

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.

● **Example 1.34** The median and IQR do not change much under the three scenarios in Table 1.27. Why might this be the case?

The median and IQR are only sensitive to numbers near Q_1, the median, and Q_3. Since values in these regions are relatively stable – there aren't large jumps between observations – the median and IQR estimates are also quite stable.

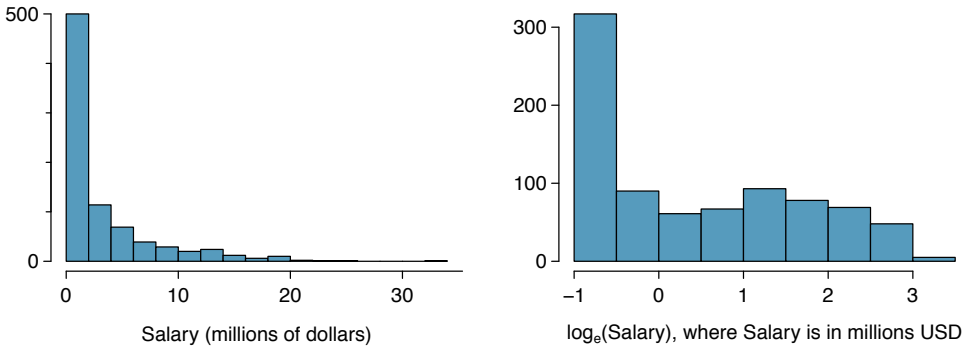
⊙ **Exercise 1.35** The distribution of vehicle prices tends to be right skewed, with a few luxury and sports cars lingering out into the right tail. If you were searching for a new car and cared about price, should you be more interested in the mean or median price of vehicles sold, assuming you are in the market for a regular car?[39]

[38] (a) Mean is affected more. (b) Standard deviation is affected more. Complete explanations are provided in the material following Exercise 1.33.

[39] Buyers of a "regular car" should be concerned about the median price. High-end car sales can drastically inflate the mean price while the median will be more robust to the influence of those sales.

1.6.7 Transforming data (special topic)

When data are very strongly skewed, we sometimes transform them so they are easier to model. Consider the histogram of salaries for Major League Baseball players' salaries from 2010, which is shown in Figure 1.28(a).



(a) (b)

Figure 1.28: (a) Histogram of MLB player salaries for 2010, in millions of dollars. (b) Histogram of the log-transformed MLB player salaries for 2010.

Example 1.36 The histogram of MLB player salaries is useful in that we can see the data are extremely skewed and centered (as gauged by the median) at about \$1 million. What isn't useful about this plot?

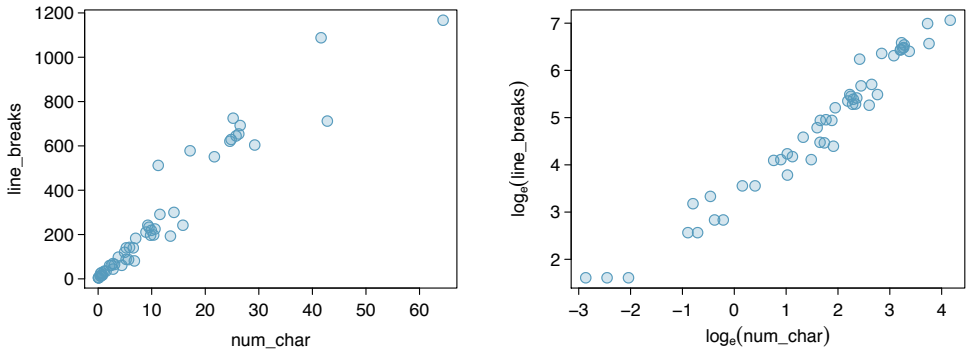
Most of the data are collected into one bin in the histogram and the data are so strongly skewed that many details in the data are obscured.

There are some standard transformations that are often applied when much of the data cluster near zero (relative to the larger values in the data set) and all observations are positive. A **transformation** is a rescaling of the data using a function. For instance, a plot of the natural logarithm[40] of player salaries results in a new histogram in Figure 1.28(b). Transformed data are sometimes easier to work with when applying statistical models because the transformed data are much less skewed and outliers are usually less extreme.

Transformations can also be applied to one or both variables in a scatterplot. A scatterplot of the `line_breaks` and `num_char` variables is shown in Figure 1.29(a), which was earlier shown in Figure 1.16. We can see a positive association between the variables and that many observations are clustered near zero. In Chapter 7, we might want to use a straight line to model the data. However, we'll find that the data in their current state cannot be modeled very well. Figure 1.29(b) shows a scatterplot where both the `line_breaks` and `num_char` variables have been transformed using a log (base e) transformation. While there is a positive association in each plot, the transformed data show a steadier trend, which is easier to model than the untransformed data.

Transformations other than the logarithm can be useful, too. For instance, the square root ($\sqrt{\text{original observation}}$) and inverse ($\frac{1}{\text{original observation}}$) are used by statisticians. Common goals in transforming data are to see the data structure differently, reduce skew, assist in modeling, or straighten a nonlinear relationship in a scatterplot.

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



(a) (b)

Figure 1.29: (a) Scatterplot of `line_breaks` against `num_char` for 50 emails. (b) A scatterplot of the same data but where each variable has been log-transformed.

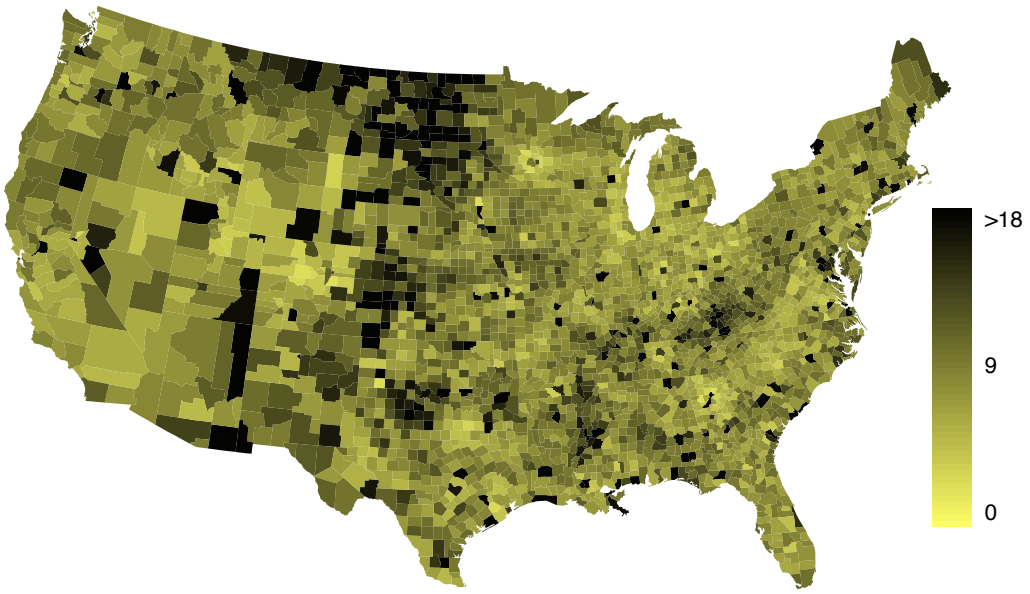
1.6.8 Mapping data (special topic)

The `county` data set offers many numerical variables that we could plot using dot plots, scatterplots, or box plots, but these miss the true nature of the data. Rather, when we encounter geographic data, we should map it using an **intensity map**, where colors are used to show higher and lower values of a variable. Figures 1.30 and 1.31 shows intensity maps for federal spending per capita (`fed_spend`), poverty rate in percent (`poverty`), homeownership rate in percent (`homeownership`), and median household income (`med_income`). The color key indicates which colors correspond to which values. Note that the intensity maps are not generally very helpful for getting precise values in any given county, but they are very helpful for seeing geographic trends and generating interesting research questions.

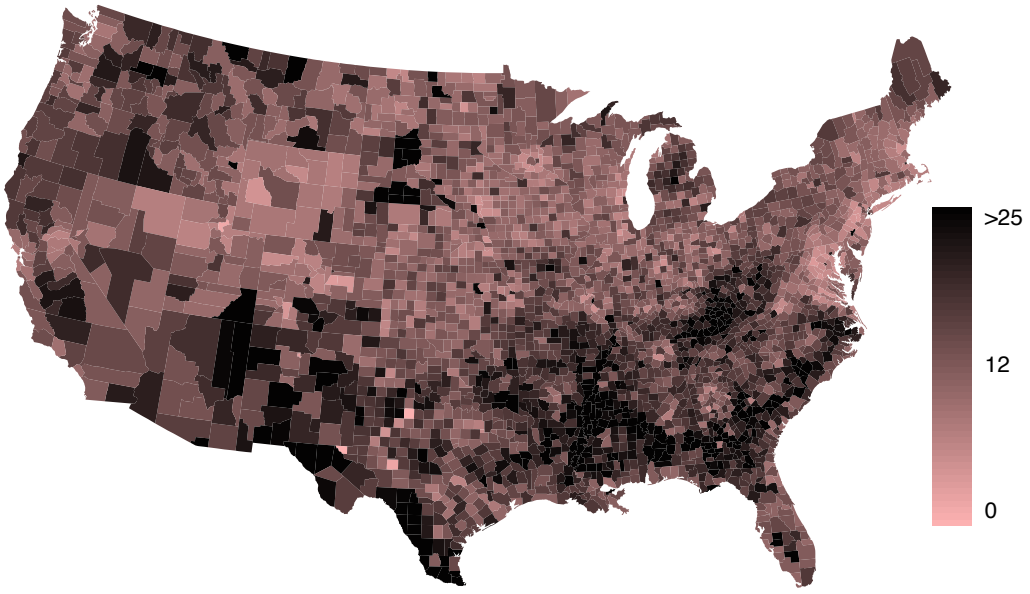
Example 1.37 What interesting features are evident in the `fed_spend` and `poverty` intensity maps?

The federal spending intensity map shows substantial spending in the Dakotas and along the central-to-western part of the Canadian border, which may be related to the oil boom in this region. There are several other patches of federal spending, such as a vertical strip in eastern Utah and Arizona and the area where Colorado, Nebraska, and Kansas meet. There are also seemingly random counties with very high federal spending relative to their neighbors. If we did not cap the federal spending range at $18 per capita, we would actually find that some counties have extremely high federal spending while there is almost no federal spending in the neighboring counties. These high-spending counties might contain military bases, companies with large government contracts, or other government facilities with many employees.

Poverty rates are evidently higher in a few locations. Notably, the deep south shows higher poverty rates, as does the southwest border of Texas. The vertical strip of eastern Utah and Arizona, noted above for its higher federal spending, also appears to have higher rates of poverty (though generally little correspondence is seen between the two variables). High poverty rates are evident in the Mississippi flood plains a little north of New Orleans and also in a large section of Kentucky and West Virginia.

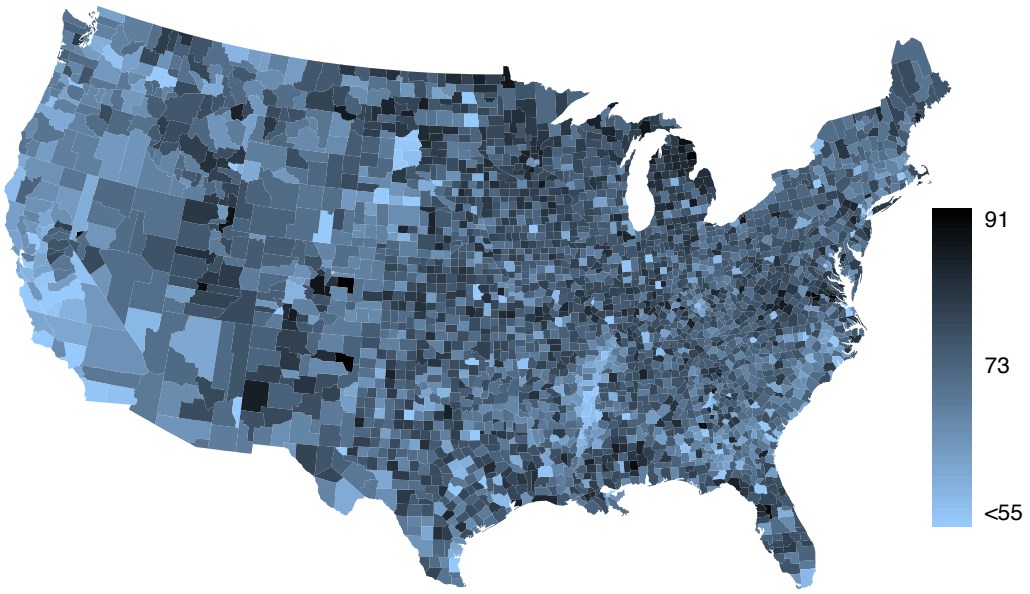


(a)

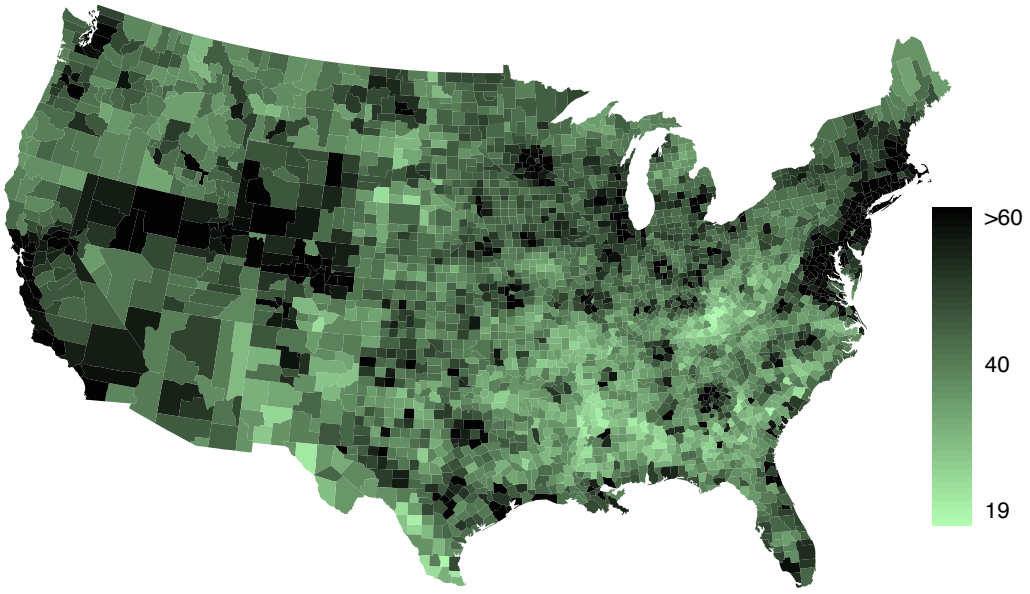


(b)

Figure 1.30: (a) Map of federal spending (dollars per capita). (b) Intensity map of poverty rate (percent).



(a)



(b)

Figure 1.31: (a) Intensity map of homeownership rate (percent). (b) Intensity map of median household income ($1000s).

⊙ **Exercise 1.38** What interesting features are evident in the `med_income` intensity map?[41]

1.7 Considering categorical data

Like numerical data, categorical data can also be organized and analyzed. In this section, we will introduce tables and other basic tools for categorical data that are used throughout this book. The `email50` data set represents a sample from a larger email data set called `email`. This larger data set contains information on 3,921 emails. In this section we will examine whether the presence of numbers, small or large, in an email provides any useful value in classifying email as spam or not spam.

1.7.1 Contingency tables and bar plots

Table 1.32 summarizes two variables: `spam` and `number`. Recall that `number` is a categorical variable that describes whether an email contains no numbers, only small numbers (values under 1 million), or at least one big number (a value of 1 million or more). 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 value 149 corresponds to the number of emails in the data set that are spam *and* had no number listed in the email. Row and column totals are also included. The **row totals** provide the total counts across each row (e.g. $149 + 168 + 50 = 367$), and **column totals** are total counts down each column.

A table for a single variable is called a **frequency table**. Table 1.33 is a frequency table for the `number` variable. If we replaced the counts with percentages or proportions, the table would be called a **relative frequency table**.

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

Table 1.32: A contingency table for `spam` and `number`.

none	small	big	Total
549	2827	545	3921

Table 1.33: A frequency table for the `number` variable.

A bar plot is a common way to display a single categorical variable. The left panel of Figure 1.34 shows a **bar plot** for the `number` variable. In the right panel, the counts are converted into proportions (e.g. $549/3921 = 0.140$ for `none`), showing the proportion of observations that are in each level (i.e. in each category).

[41] Note: answers will vary. There is a very strong correspondence between high earning and metropolitan areas. You might look for large cities you are familiar with and try to spot them on the map as dark spots.

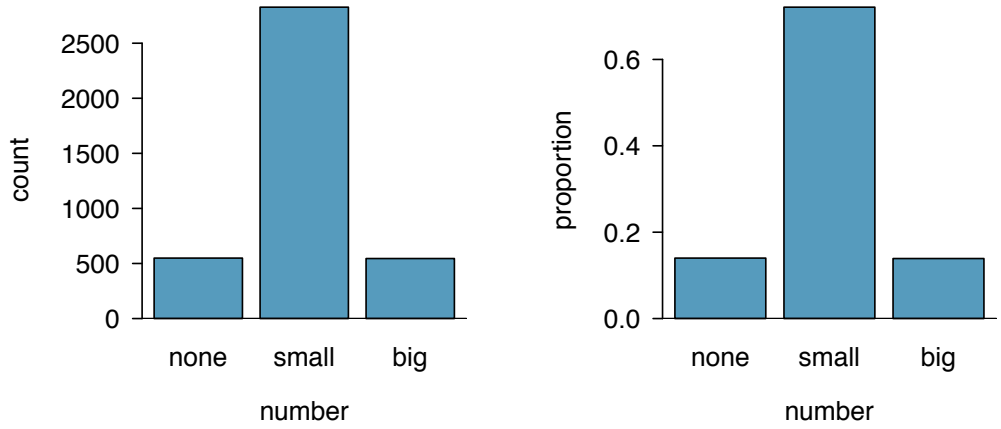


Figure 1.34: Two bar plots of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

1.7.2 Row and column proportions

Table 1.35 shows the row proportions for Table 1.32. The **row proportions** are computed as the counts divided by their row totals. The value 149 at the intersection of `spam` and `none` is replaced by $149/367 = 0.416$, i.e. 149 divided by its row total, 367. So what does 0.416 represent? It corresponds to the proportion of spam emails in the sample that do not have any numbers.

	none	small	big	Total
spam	$149/367 = 0.416$	$168/367 = 0.458$	$50/367 = 0.136$	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 1.35: A contingency table with row proportions for the `spam` and `number` variables.

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 1.36 shows such a table, and here the value 0.271 indicates that 27.1% of emails with no numbers were spam. This rate of spam is much higher compared to emails with only small numbers (5.9%) or big numbers (9.2%). Because these spam rates vary between the three levels of `number` (`none`, `small`, `big`), this provides evidence that the `spam` and `number` variables are associated.

	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 1.36: A contingency table with column proportions for the `spam` and `number` variables.

We could also have checked for an association between `spam` and `number` in Table 1.35 using row proportions. When comparing these row proportions, we would look down columns to see if the fraction of emails with no numbers, small numbers, and big numbers varied from `spam` to `not spam`.

⊙ **Exercise 1.39** What does 0.458 represent in Table 1.35? What does 0.059 represent in Table 1.36?[42]

⊙ **Exercise 1.40** What does 0.139 represent in Table 1.35? What does 0.908 represent in the Table 1.36?[43]

● **Example 1.41** Data scientists use statistics to filter spam from incoming email messages. By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy. One of those characteristics is whether the email contains no numbers, small numbers, or big numbers. Another characteristic is whether or not an email has any HTML content. A contingency table for the `spam` and `format` variables from the `email` data set are shown in Table 1.37. Recall that an HTML email is an email with the capacity for special formatting, e.g. bold text. In Table 1.37, which would be more helpful to someone hoping to classify email as spam or regular email: row or column proportions?

Such a person would be interested in how the proportion of spam changes within each email format. This corresponds to column proportions: the proportion of spam in plain text emails and the proportion of emails in HTML emails.

If we generate the column proportions, we can see that a higher fraction of plain text emails are spam ($209/1195 = 17.5\%$) than compared to HTML emails ($158/2726 = 5.8\%$). This information on its own is insufficient to classify an email as spam or not spam, as over 80% of plain text emails are not spam. Yet, when we carefully combine this information with many other characteristics, such as `number` and other variables, we stand a reasonable chance of being able to classify some email as spam or not spam. This is a topic we will return to in Chapter 8.

	text	HTML	Total
spam	209	158	367
not spam	986	2568	3554
Total	1195	2726	3921

Table 1.37: A contingency table for `spam` and `format`.

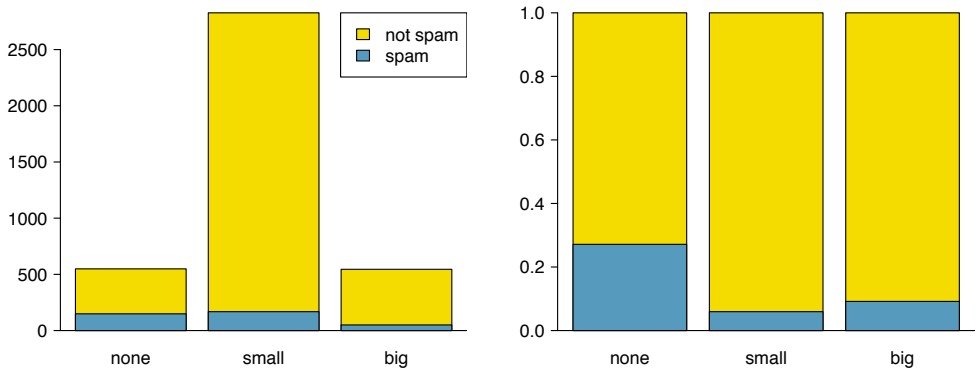
Example 1.41 points out that row and column proportions are not equivalent. Before settling on one form for a table, it is important to consider each to ensure that the most useful table is constructed.

⊙ **Exercise 1.42** Look back to Tables 1.35 and 1.36. Which would be more useful to someone hoping to identify spam emails using the `number` variable?[44]

[42] 0.458 represents the proportion of spam emails that had a small number. 0.058 represents the fraction of emails with small numbers that are spam.

[43] 0.139 represents the fraction of non-spam email that had a big number. 0.908 represents the fraction of emails with big numbers that are non-spam emails.

[44] The column proportions in Table 1.36 will probably be most useful, which makes it easier to see that emails with small numbers spam about 5.9% of the time (relatively rare). We would also see that about 27.1% of emails with no numbers are spam, and 9.2% of emails with big numbers are spam.



(a) (b)

Figure 1.38: (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).

1.7.3 Segmented bar and mosaic plots

Contingency tables using row or column proportions are especially useful for examining how two categorical variables are related. Segmented bar and mosaic plots provide a way to visualize the information in these tables.

A **segmented bar plot** is a graphical display of contingency table information. For example, a segmented bar plot representing Table 1.36 is shown in Figure 1.38(a), where we have first created a bar plot using the `number` variable and then divided each group by the levels of `spam`. The column proportions of Table 1.36 have been translated into a standardized segmented bar plot in Figure 1.38(b), which is a helpful visualization of the fraction of spam emails in each level of `number`.

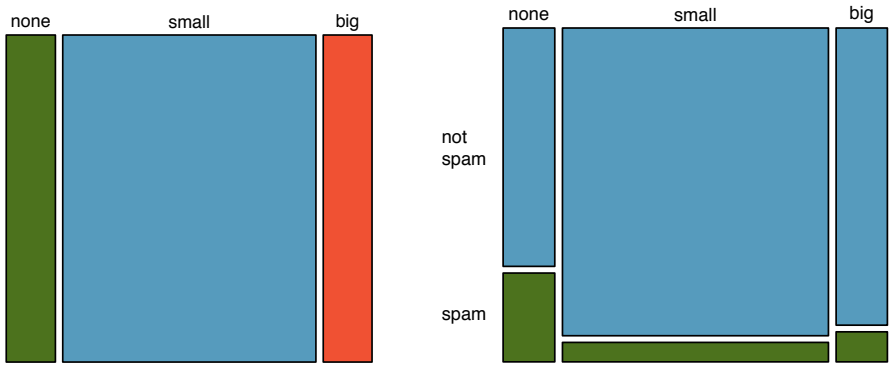
Example 1.43 Examine both of the segmented bar plots. Which is more useful?

Figure 1.38(a) contains more information, but Figure 1.38(b) presents the information more clearly. This second plot makes it clear that emails with no number have a relatively high rate of spam email – about 27%! On the other hand, less than 10% of email with small or big numbers are spam.

Since the proportion of spam changes across the groups in Figure 1.38(b), we can conclude the variables are dependent, which is something we were also able to discern using table proportions. Because both the `none` and `big` groups have relatively few observations compared to the `small` group, the association is more difficult to see in Figure 1.38(a).

In some other cases, a segmented bar plot that is not standardized will be more useful in communicating important information. Before settling on a particular segmented bar plot, create standardized and non-standardized forms and decide which is more effective at communicating features of the data.

A **mosaic plot** is a graphical display of contingency table information that is similar to a bar plot for one variable or a segmented bar plot when using two variables. Figure 1.39(a) shows a mosaic plot for the `number` variable. Each column represents a level of `number`, and the column widths correspond to the proportion of emails of each number type. For instance, there are fewer emails with no numbers than emails with only small numbers, so



(a) (b)

Figure 1.39: The one-variable mosaic plot for `number` and the two-variable mosaic plot for both `number` and `spam`.

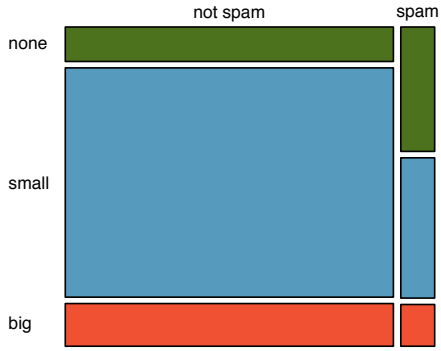


Figure 1.40: Mosaic plot where emails are grouped by the `number` variable after they've been divided into `spam` and `not spam`.

the no number email column is slimmer. In general, mosaic plots use box *areas* to represent the number of observations that box represents.

This one-variable mosaic plot is further divided into pieces in Figure 1.39(b) using the `spam` variable. Each column is split proportionally according to the fraction of emails that were spam in each number category. For example, the second column, representing emails with only small numbers, was divided into emails that were spam (lower) and not spam (upper). As another example, the bottom of the third column represents spam emails that had big numbers, and the upper part of the third column represents regular emails that had big numbers. We can again use this plot to see that the `spam` and `number` variables are associated since some columns are divided in different vertical locations than others, which was the same technique used for checking an association in the standardized version of the segmented bar plot.

In a similar way, a mosaic plot representing row proportions of Table 1.32 could be constructed, as shown in Figure 1.40. However, because it is more insightful for this application to consider the fraction of spam in each category of the `number` variable, we prefer Figure 1.39(b).

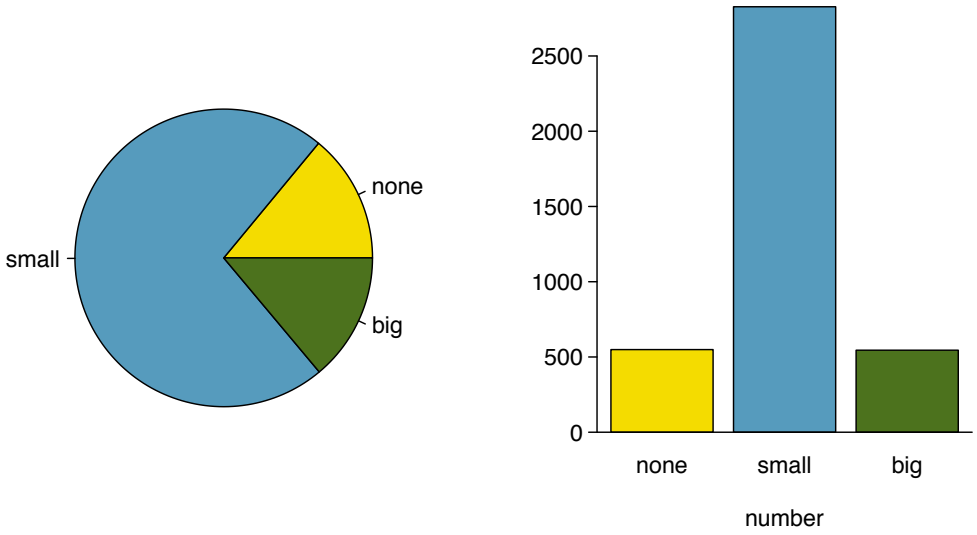


Figure 1.41: A pie chart and bar plot of `number` for the `email` data set.

1.7.4 The only pie chart you will see in this book

While pie charts are well known, they are not typically as useful as other charts in a data analysis. A **pie chart** is shown in Figure 1.41 alongside a bar plot. It is generally more difficult to compare group sizes in a pie chart than in a bar plot, especially when categories have nearly identical counts or proportions. In the case of the `none` and `big` categories, the difference is so slight you may be unable to distinguish any difference in group sizes for either plot!

1.7.5 Comparing numerical data across groups

Some of the more interesting investigations can be considered by examining numerical data across groups. The methods required here aren't really new. All that is required is to make a numerical plot for each group. Here two convenient methods are introduced: side-by-side box plots and hollow histograms.

We will take a look again at the `county` data set and compare the median household income for counties that gained population from 2000 to 2010 versus counties that had no gain. While we might like to make a causal connection here, remember that these are observational data and so such an interpretation would be unjustified.

There were 2,041 counties where the population increased from 2000 to 2010, and there were 1,099 counties with no gain (all but one were a loss). A random sample of 100 counties from the first group and 50 from the second group are shown in Table 1.42 to give a better sense of some of the raw data.

The **side-by-side box plot** is a traditional tool for comparing across groups. An example is shown in the left panel of Figure 1.43, where there are two box plots, one for each group, placed into one plotting window and drawn on the same scale.

Another useful plotting method uses **hollow histograms** to compare numerical data across groups. These are just the outlines of histograms of each group put on the same plot, as shown in the right panel of Figure 1.43.

population gain						no gain		
41.2	33.1	30.4	37.3	79.1	34.5	40.3	33.5	34.8
22.9	39.9	31.4	45.1	50.6	59.4	29.5	31.8	41.3
47.9	36.4	42.2	43.2	31.8	36.9	28	39.1	42.8
50.1	27.3	37.5	53.5	26.1	57.2	38.1	39.5	22.3
57.4	42.6	40.6	48.8	28.1	29.4	43.3	37.5	47.1
43.8	26	33.8	35.7	38.5	42.3	43.7	36.7	36
41.3	40.5	68.3	31	46.7	30.5	35.8	38.7	39.8
68.3	48.3	38.7	62	37.6	32.2	46	42.3	48.2
42.6	53.6	50.7	35.1	30.6	56.8	38.6	31.9	31.1
66.4	41.4	34.3	38.9	37.3	41.7	37.6	29.3	30.1
51.9	83.3	46.3	48.4	40.8	42.6	57.5	32.6	31.1
44.5	34	48.7	45.2	34.7	32.2	46.2	26.5	40.1
39.4	38.6	40	57.3	45.2	33.1	38.4	46.7	25.9
43.8	71.7	45.1	32.2	63.3	54.7	36.4	41.5	45.7
71.3	36.3	36.4	41	37	66.7	39.7	37	37.7
50.2	45.8	45.7	60.2	53.1		21.4	29.3	50.1
35.8	40.4	51.5	66.4	36.1		43.6	39.8	

Table 1.42: In this table, median household income (in $1000s) from a random sample of 100 counties that held steady or gained population over 2000-2010 are shown on the left. Median incomes from a random sample of 50 counties that had no population gain are shown on the right.

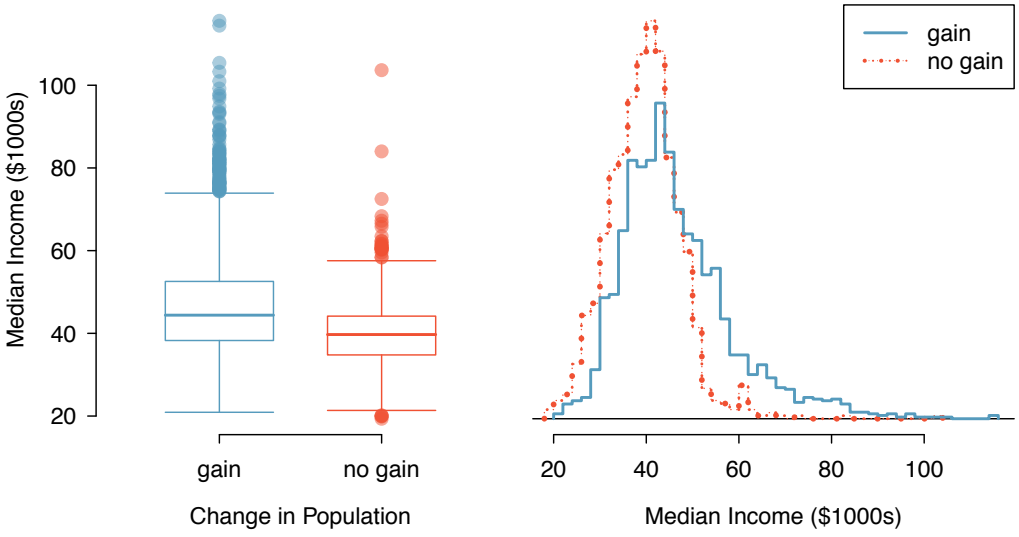


Figure 1.43: Side-by-side box plot (left panel) and hollow histograms (right panel) for `med_income`, where the counties are split by whether there was a population gain or loss from 2000 to 2010. The income data were collected between 2006 and 2010.

⊙ **Exercise 1.44** Use the plots in Figure 1.43 to compare the incomes for counties across the two groups. What do you notice about the approximate center of each group? What do you notice about the variability between groups? Is the shape relatively consistent between groups? How many *prominent* modes are there for each group?[45]

⊙ **Exercise 1.45** What components of each plot in Figure 1.43 do you find most useful?[46]

1.8 Case study: gender discrimination (special topic)

● **Example 1.46** Suppose your professor splits the students in class into two groups: students on the left and students on the right. If $\hat{p}_L$ and $\hat{p}_R$ represent the proportion of students who own an Apple product on the left and right, respectively, would you be surprised if $\hat{p}_L$ did not exactly equal $\hat{p}_R$?

While the proportions would probably be close to each other, it would be unusual for them to be exactly the same. We would probably observe a small difference due to chance.

⊙ **Exercise 1.47** If we don't think the side of the room a person sits on in class is related to whether the person owns an Apple product, what assumption are we making about the relationship between these two variables?[47]

1.8.1 Variability within data

We consider a study investigating gender discrimination in the 1970s, which is set in the context of personnel decisions within a bank.[48] The research question we hope to answer is, "Are females unfairly discriminated against in promotion decisions made by male managers?"

The participants in this study are 48 male bank supervisors attending a management institute at the University of North Carolina in 1972. They were asked to assume the role of the personnel director of a bank and were given a personnel file to judge whether the person should be promoted to a branch manager position. The files given to the participants were identical, except that half of them indicated the candidate was male and the other half indicated the candidate was female. These files were randomly assigned to the subjects.

[45] Answers may vary a little. The counties with population gains tend to have higher income (median of about $45,000) versus counties without a gain (median of about $40,000). The variability is also slightly larger for the population gain group. This is evident in the IQR, which is about 50% bigger in the *gain* group. Both distributions show slight to moderate right skew and are unimodal. There is a secondary small bump at about $60,000 for the *no gain* group, visible in the hollow histogram plot, that seems out of place. (Looking into the data set, we would find that 8 of these 15 counties are in Alaska and Texas.) The box plots indicate there are many observations far above the median in each group, though we should anticipate that many observations will fall beyond the whiskers when using such a large data set.

[46] Answers will vary. The side-by-side box plots are especially useful for comparing centers and spreads, while the hollow histograms are more useful for seeing distribution shape, skew, and groups of anomalies.

[47] We would be assuming that these two variables are independent.

[48] Rosen B and Jerdee T. 1974. Influence of sex role stereotypes on personnel decisions. Journal of Applied Psychology 59(1):9-14.

Exercise 1.48 Is this an observational study or an experiment? What implications does the study type have on what can be inferred from the results?[49]

For each supervisor we record the gender associated with the assigned file and the promotion decision. Using the results of the study summarized in Table 1.44, we would like to evaluate if females are unfairly discriminated against in promotion decisions. In this study, a smaller proportion of females are promoted than males (0.583 versus 0.875), but it is unclear whether the difference provides *convincing evidence* that females are unfairly discriminated against.

		decision		
		promoted	not promoted	Total
gender	male	21	3	24
	female	14	10	24
	Total	35	13	48

Table 1.44: Summary results for the gender discrimination study.

Example 1.49 Statisticians are sometimes called upon to evaluate the strength of evidence. When looking at the rates of promotion for males and females in this study, what comes to mind as we try to determine whether the data show convincing evidence of a real difference?

The observed promotion rates (58.3% for females versus 87.5% for males) suggest there might be discrimination against women in promotion decisions. However, we cannot be sure if the observed difference represents discrimination or is just from random chance. Generally there is a little bit of fluctuation in sample data, and we wouldn't expect the sample proportions to be *exactly* equal, even if the truth was that the promotion decisions were independent of gender.

Example 1.49 is a reminder that the observed outcomes in the sample may not perfectly reflect the true relationships between variables in the underlying population. Table 1.44 shows there were 7 fewer promotions in the female group than in the male group, a difference in promotion rates of 29.2% $\left(\frac{21}{24} - \frac{14}{24} = 0.292\right)$. This difference is large, but the sample size for the study is small, making it unclear if this observed difference represents discrimination or whether it is simply due to chance. We label these two competing claims, H_0 and H_A:

H_0: **Independence model.** The variables `gender` and `decision` are independent. They have no relationship, and the observed difference between the proportion of males and females who were promoted, 29.2%, was due to chance.

H_A: **Alternative model.** The variables `gender` and `decision` are *not* independent. The difference in promotion rates of 29.2% was not due to chance, and equally qualified females are less likely to be promoted than males.

What would it mean if the independence model, which says the variables `gender` and `decision` are unrelated, is true? It would mean each banker was going to decide whether to promote the candidate without regard to the gender indicated on the file. That is,

[49] The study is an experiment, as subjects were randomly assigned a male file or a female file. Since this is an experiment, the results can be used to evaluate a causal relationship between gender of a candidate and the promotion decision.

the difference in the promotion percentages was due to the way the files were randomly divided to the bankers, and the randomization just happened to give rise to a relatively large difference of 29.2%.

Consider the alternative model: bankers were influenced by which gender was listed on the personnel file. If this was true, and especially if this influence was substantial, we would expect to see some difference in the promotion rates of male and female candidates. If this gender bias was against females, we would expect a smaller fraction of promotion decisions for female personnel files relative to the male files.

We choose between these two competing claims by assessing if the data conflict so much with H_0 that the independence model cannot be deemed reasonable. If this is the case, and the data support H_A, then we will reject the notion of independence and conclude there was discrimination.

1.8.2 Simulating the study

Table 1.44 shows that 35 bank supervisors recommended promotion and 13 did not. Now, suppose the banker's decisions were independent of gender. Then, if we conducted the experiment again with a different random arrangement of files, differences in promotion rates would be based only on random fluctuation. We can actually perform this **randomization**, which simulates what would have happened if the bankers decisions had been independent of gender but we had distributed the files differently.

In this **simulation**, we thoroughly shuffle 48 personnel files, 24 labeled `male_sim` and 24 labeled `female_sim`, and deal these files into two stacks. We will deal 35 files into the first stack, which will represent the 35 supervisors who recommended promotion. The second stack will have 13 files, and it will represent the 13 supervisors who recommended against promotion. Then, as we did with the original data, we tabulate the results and determine the fraction of `male_sim` and `female_sim` who were promoted. The randomization of files in this simulation is independent of the promotion decisions, which means any difference in the two fractions is entirely due to chance. Table 1.45 show the results of such a simulation.

		`decision`		
		promoted	not promoted	Total
`gender_sim`	`male_sim`	18	6	24
	`female_sim`	17	7	24
	Total	35	13	48

Table 1.45: Simulation results, where any difference in promotion rates between `male_sim` and `female_sim` is purely due to chance.

⊙ **Exercise 1.50** What is the difference in promotion rates between the two simulated groups in Table 1.45? How does this compare to the observed 29.2% in the actual groups?[50]

[50] $18/24 - 17/24 = 0.042$ or about 4.2% in favor of the men. This difference due to chance is much smaller than the difference observed in the actual groups.

1.8.3 Checking for independence

We computed one possible difference under the independence model in Exercise 1.50, which represents one difference due to chance. While in this first simulation, we physically dealt out files, it is more efficient to perform this simulation using a computer. Repeating the simulation on a computer, we get another difference due to chance: -0.042. And another: 0.208. And so on until we repeat the simulation enough times that we have a good idea of what represents the *distribution of differences from chance alone*. Figure 1.46 shows a plot of the differences found from 100 simulations, where each dot represents a simulated difference between the proportions of male and female files that were recommended for promotion.

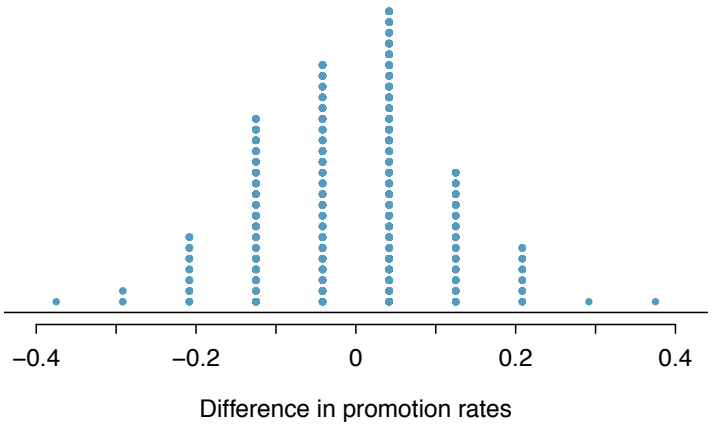


Figure 1.46: A stacked dot plot of differences from 100 simulations produced under the independence model, H_0, where `gender_sim` and `decision` are independent. Two of the 100 simulations had a difference of at least 29.2%, the difference observed in the study.

Note that the distribution of these simulated differences is centered around 0. We simulated these differences assuming that the independence model was true, and under this condition, we expect the difference to be zero with some random fluctation. We would generally be surprised to see a difference of *exactly* 0: sometimes, just by chance, the difference is higher than 0, and other times it is lower than zero.

Example 1.51 How often would you observe a difference of at least 29.2% (0.292) according to Figure 1.46? Often, sometimes, rarely, or never?

It appears that a difference of at least 29.2% due to chance alone would only happen about 2% of the time according to Figure 1.46. Such a low probability indicates a rare event.

The difference of 29.2% being a rare event suggests two possible interpretations of the results of the study:

H_0 **Independence model.** Gender has no effect on promotion decision, and we observed a difference that would only happen rarely.

H_A **Alternative model.** Gender has an effect on promotion decision, and what we observed was actually due to equally qualified women being discriminated against in promotion decisions, which explains the large difference of 29.2%.

Based on the simulations, we have two options. (1) We conclude that the study results do not provide strong evidence against the independence model. That is, we do not have sufficiently strong evidence to conclude there was gender discrimination. (2) We conclude the evidence is sufficiently strong to reject H_0 and assert that there was gender discrimination. When we conduct formal studies, usually we reject the notion that we just happened to observe a rare event.[51] So in this case, we reject the independence model in favor of the alternative. That is, we are concluding the data provide strong evidence of gender discrimination against women by the supervisors.

One field of statistics, statistical inference, is built on evaluating whether such differences are due to chance. In statistical inference, statisticians evaluate which model is most reasonable given the data. Errors do occur, just like rare events, and we might choose the wrong model. While we do not always choose correctly, statistical inference gives us tools to control and evaluate how often these errors occur. In Chapter 4, we give a formal introduction to the problem of model selection. We spend the next two chapters building a foundation of probability and theory necessary to make that discussion rigorous.

[51] This reasoning does not generally extend to anecdotal observations. Each of us observes incredibly rare events every day, events we could not possibly hope to predict. However, in the non-rigorous setting of anecdotal evidence, almost anything may appear to be a rare event, so the idea of looking for rare events in day-to-day activities is treacherous. For example, we might look at the lottery: there was only a 1 in 176 million chance that the Mega Millions numbers for the largest jackpot in history (March 30, 2012) would be (2, 4, 23, 38, 46) with a Mega ball of (23), but nonetheless those numbers came up! However, no matter what numbers had turned up, they would have had the same incredibly rare odds. That is, *any set of numbers we could have observed would ultimately be incredibly rare.* This type of situation is typical of our daily lives: each possible event in itself seems incredibly rare, but if we consider every alternative, those outcomes are also incredibly rare. We should be cautious not to misinterpret such anecdotal evidence.

1.9 Exercises

1.9.1 Case study

1.1 Migraine and accupuncture. A migraine is a particularly painful type of headache, which patients sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a randomized controlled study where 89 females diagnosed with migraine headaches were randomly assigned to one of two groups: treatment or control. 43 patients in the treatment group received acupuncture that is specifically designed to treat migraines. 46 patients in the control group received placebo acupuncture (needle insertion at nonacupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. Results are summarized in the contingency table below.[52]

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

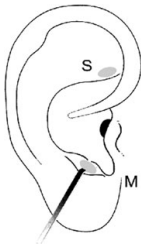


Figure from the original paper displaying the appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks.

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

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

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

1.2 Sinusitis and antibiotics, Part I. Researchers studying the effect of antibiotic treatment for acute sinusitis compared to symptomatic treatments randomly assigned 166 adults diagnosed with acute sinusitis to one of two groups: treatment or control. Study participants received either a 10-day course of amoxicillin (an antibiotic) or a placebo similar in appearance and taste. The placebo consisted of symptomatic treatments such as acetaminophen, nasal decongestants, etc. At the end of the 10-day period patients were asked if they experienced significant improvement in symptoms. The distribution of responses are summarized below.[53]

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

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

(b) At first glance, which treatment appears to be more effective for sinusitis?

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

[52] G. Allais et al. "Ear acupuncture in the treatment of migraine attacks: a randomized trial on the efficacy of appropriate versus inappropriate acupoints". In: *Neurological Sciences* 32.1 (2011), pp. 173–175.

[53] J.M. Garbutt et al. "Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial". In: *JAMA: The Journal of the American Medical Association* 307.7 (2012), pp. 685–692.

1.9.2 Data basics

1.3 Identify study components, Part I. Identify (i) the cases, (ii) the variables and their types, and (iii) the main research question in the studies described below.

(a) Researchers collected data to examine the relationship between pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.[54]

(b) The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.[55]

1.4 Identify study components, Part II. Identify (i) the cases, (ii) the variables and their types, and (iii) the main research question of the studies described below.

(a) While obesity is measured based on body fat percentage (more than 35% body fat for women and more than 25% for men), precisely measuring body fat percentage is difficult. Body mass index (BMI), calculated as the ratio $weight/height^2$, is often used as an alternative indicator for obesity. A common criticism of BMI is that it assumes the same relative body fat percentage regardless of age, sex, or ethnicity. In order to determine how useful BMI is for predicting body fat percentage across age, sex and ethnic groups, researchers studied 202 black and 504 white adults who resided in or near New York City, were ages 20-94 years old, had BMIs of 18-35 kg/m^2, and who volunteered to be a part of the study. Participants reported their age, sex, and ethnicity and were measured for weight and height. Body fat percentage was measured by submerging the participants in water.[56]

(b) In a study of the relationship between socio-economic class and unethical behavior, 129 University of California undergraduates at Berkeley were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that they were for children in a nearby laboratory, but that they could take some if they wanted. Participants completed unrelated tasks and then reported the number of candies they had taken. It was found that those in the upper-class rank condition took more candy than did those in the lower-rank condition.[57]

[54] B. Ritz et al. "Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993". In: *Epidemiology* 11.5 (2000), pp. 502–511.

[55] J. McGowan. "Health Education: Does the Buteyko Institute Method make a difference?" In: *Thorax* 58 (2003).

[56] Gallagher et al. "How useful is body mass index for comparison of body fatness across age, sex, and ethnic groups?" In: *American Journal of Epidemiology* 143.3 (1996), pp. 228–239.

[57] P.K. Piff et al. "Higher social class predicts increased unethical behavior". In: *Proceedings of the National Academy of Sciences* (2012).

1.5 Fisher's irises. Sir Ronald Aylmer Fisher was an English statistician, evolutionary biologist, and geneticist who worked on a data set that contained sepal length and width, and petal length and width from three species of iris flowers (*setosa*, *versicolor* and *virginica*). There were 50 flowers from each species in the data set.[58]

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

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

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

1.6 Smoking habits of UK residents. A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that "£" stands for British Pounds Sterling, "cig" stands for cigarettes, and "N/A" refers to a missing component of the data.[59]

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

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

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

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

1.9.3 Overview of data collection principles

1.7 Generalizability and causality, Part I. Identify the population of interest and the sample in the the studies described in Exercise 1.3. Also comment on whether or not the results of the study can be generalized to the population and if the findings of the study can be used to establish causal relationships.

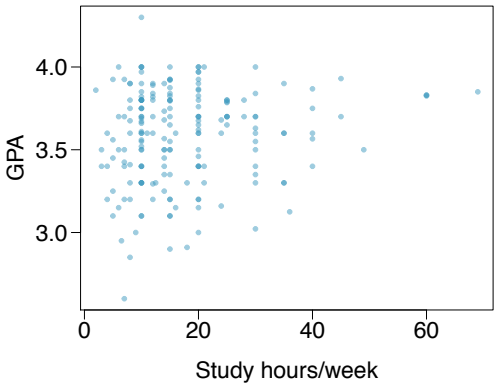
1.8 Generalizability and causality, Part II. Identify the population of interest and the sample in the the studies described in Exercise 1.4. Also comment on whether or not the results of the study can be generalized to the population and if the findings of the study can be used to establish causal relationships.

[58] Photo by rtclauss on Flickr, Iris.; R.A Fisher. "The Use of Multiple Measurements in Taxonomic Problems". In: *Annals of Eugenics* 7 (1936), pp. 179–188.

[59] Stats4Schools, Smoking.

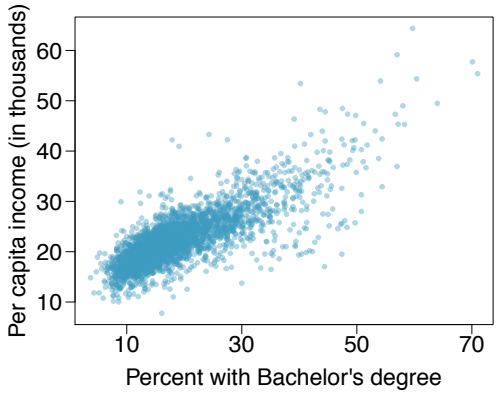
1.9 GPA and study time. A survey was conducted on 218 undergraduates from Duke University who took an introductory statistics course in Spring 2012. Among many other questions, this survey asked them about their GPA and the number of hours they spent studying per week. The scatterplot below displays the relationship between these two variables.

(a) What is the explanatory variable and what is the response variable?
(b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.
(c) Is this an experiment or an observational study?
(d) Can we conclude that studying longer hours leads to higher GPAs?



1.10 Income and education. The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,143 counties in the US in 2010.

(a) What are the explanatory and response variables?
(b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.
(c) Can we conclude that having a bachelor's degree increases one's income?



1.9.4 Observational studies and sampling strategies

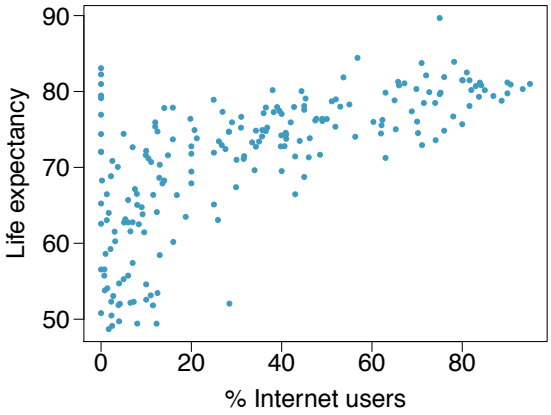
1.11 Propose a sampling strategy. A large college class has 160 students. All 160 students attend the lectures together, but the students are divided into 4 groups, each of 40 students, for lab sections administered by different teaching assistants. The professor wants to conduct a survey about how satisfied the students are with the course, and he believes that the lab section a student is in might affect the student's overall satisfaction with the course.

(a) What type of study is this?

(b) Suggest a sampling strategy for carrying out this study.

1.12 Internet use and life expectancy. The scatterplot below shows the relationship between estimated life expectancy at birth as of 2012[60] and percentage of internet users in 2010[61] in 208 countries.

(a) Describe the relationship between life expectancy and percentage of internet users.
(b) What type of study is this?
(c) State a possible confounding variable that might explain this relationship and describe its potential effect.



1.13 Random digit dialing. The Gallup Poll uses a procedure called random digit dialing, which creates phone numbers based on a list of all area codes in America in conjunction with the associated number of residential households in each area code. Give a possible reason the Gallup Poll chooses to use random digit dialing instead of picking phone numbers from the phone book.

1.14 Sampling strategies. A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey. Three research strategies for collecting data are described below. In each, name the sampling method proposed and any bias you might expect.

(a) He randomly samples 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day.
(b) He gives out the survey only to his friends, and makes sure each one of them fills out the survey.
(c) He posts a link to an online survey on his Facebook wall and asks his friends to fill out the survey.

1.15 Family size. Suppose we want to estimate family size, where family is defined as one or more parents living with children. If we select students at random at an elementary school and ask them what their family size is, will our average be biased? If so, will it overestimate or underestimate the true value?

[60] CIA Factbook, Country Comparison: Life Expectancy at Birth, 2012.
[61] ITU World Telecommunication/ICT Indicators database, World Telecommunication/ICT Indicators Database, 2012.

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

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

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

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

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

(a) An article called *Risks: Smokers Found More Prone to Dementia* states the following:[62]

> "Researchers analyzed the data of 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50 to 60 years old. Twenty-three years later, about one-quarter of the group, or 5,367, had dementia, including 1,136 with Alzheimers disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37 percent more likely than nonsmokers to develop dementia, and the risks went up sharply with increased smoking; 44 percent for one to two packs a day; and twice the risk for more than two packs."

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

(b) Another article called *The School Bully Is Sleepy* states the following:[63]

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

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

1.18 Shyness on Facebook. Given the anonymity afforded to individuals in online interactions, researchers hypothesized that shy individuals would have more favorable attitudes toward Facebook and that shyness would be positively correlated with time spent on Facebook. They also hypothesized that shy individuals would have fewer Facebook "Friends" just like they have fewer friends than non-shy individuals have in the offline world. Data were collected on 103 undergraduate students at a university in southwestern Ontario via online questionnaires. The study states "Participants were recruited through the university's psychology participation pool. After indicating an interest in the study, participants were sent an e-mail containing the study's URL as well as the necessary login credentials." Are the results of this study generalizable to the population of all Facebook users?[64]

[62] R.C. Rabin. "Risks: Smokers Found More Prone to Dementia". In: *New York Times* (2010).

[63] T. Parker-Pope. "The School Bully Is Sleepy". In: *New York Times* (2011).

[64] E.S. Orr et al. "The influence of shyness on the use of Facebook in an undergraduate sample". In: *CyberPsychology & Behavior* 12.3 (2009), pp. 337–340.

1.9.5 Experiments

1.19 Vitamin supplements. In order to assess the effectiveness of taking large doses of vitamin C in reducing the duration of the common cold, researchers recruited 400 healthy volunteers from staff and students at a university. A quarter of the patients were assigned a placebo, and the rest were evenly divided between 1g Vitamin C, 3g Vitamin C, or 3g Vitamin C plus additives to be taken at onset of a cold for the following two days. All tablets had identical appearance and packaging. The nurses who handed the prescribed pills to the patients knew which patient received which treatment, but the researchers assessing the patients when they were sick did not. No significant differences were observed in any measure of cold duration or severity between the four medication groups, and the placebo group had the shortest duration of symptoms.[65]

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

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

(c) Were the patients blinded to their treatment?

(d) Was this study double-blind?

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

1.20 Soda preference. You would like to conduct an experiment in class to see if your classmates prefer the taste of regular Coke or Diet Coke. Briefly outline a design for this study.

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

(a) What type of study is this?

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

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

(d) Does this study make use of blinding?

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

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

[65] C. Audera et al. "Mega-dose vitamin C in treatment of the common cold: a randomised controlled trial". In: *Medical Journal of Australia* 175.7 (2001), pp. 359–362.

1.22 Chia seeds and weight loss. Chia Pets – those terra-cotta figurines that sprout fuzzy green hair – made the chia plant a household name. But chia has gained an entirely new reputation as a diet supplement. In one 2009 study, a team of researchers recruited 38 men and divided them evenly into two groups: treatment or control. They also recruited 38 women, and they randomly placed half of these participants into the treatment group and the other half into the control group. One group was given 25 grams of chia seeds twice a day, and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no significant difference between the groups in appetite or weight loss.[66]

(a) What type of study is this?

(b) What are the experimental and control treatments in this study?

(c) Has blocking been used in this study? If so, what is the blocking variable?

(d) Has blinding been used in this study?

(e) Comment on whether or not we can make a causal statement, and indicate whether or not we can generalize the conclusion to the population at large.

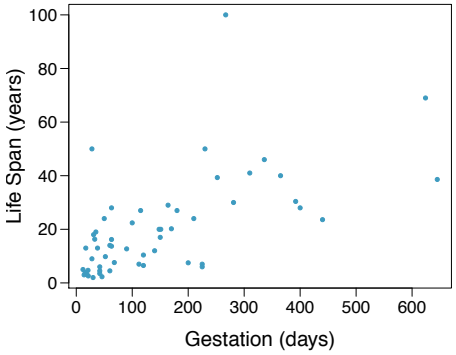
1.9.6 Examining numerical data

1.23 Mammal life spans. Data were collected on life spans (in years) and gestation lengths (in days) for 62 mammals. A scatterplot of life span versus length of gestation is shown below.[67]

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

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

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



1.24 Office productivity. Office productivity is relatively low when the employees feel no stress about their work or job security. However, high levels of stress can also lead to reduced employee productivity. Sketch a plot to represent the relationship between stress and productivity.

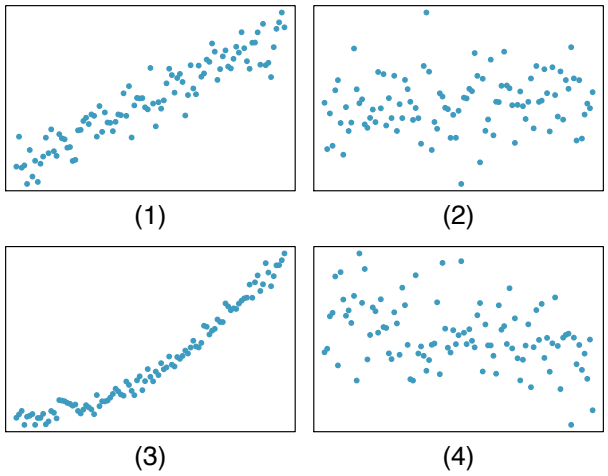
[66] D.C. Nieman et al. "Chia seed does not promote weight loss or alter disease risk factors in overweight adults". In: *Nutrition Research* 29.6 (2009), pp. 414–418.

[67] T. Allison and D.V. Cicchetti. "Sleep in mammals: ecological and constitutional correlates". In: *Arch. Hydrobiol* 75 (1975), p. 442.

1.25 Associations. Indicate which of the plots show a

(a) positive association
(b) negative association
(c) no association

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



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

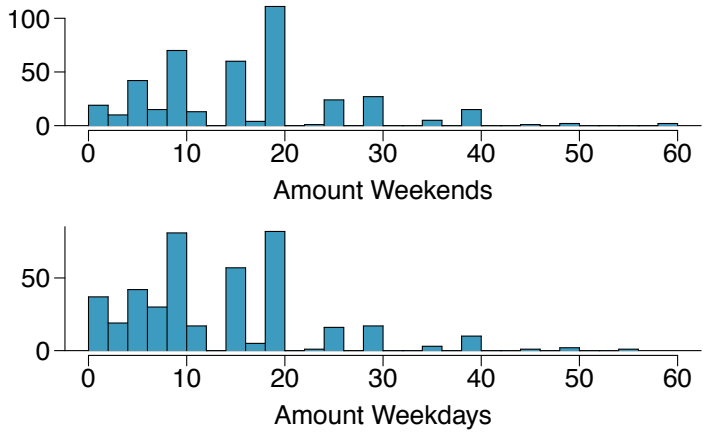
(a) A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night.
(b) American households spent an average of about $52 in 2007 on Halloween merchandise such as costumes, decorations and candy. To see if this number had changed, researchers conducted a new survey in 2008 before industry numbers were reported. The survey included 1,500 households and found that average Halloween spending was $58 per household.
(c) The average GPA of students in 2001 at a private university was 3.37. A survey on a sample of 203 students from this university yielded an average GPA of 3.59 in Spring semester of 2012.

1.27 Make-up exam. In a class of 25 students, 24 of them took an exam in class and 1 student took a make-up exam the following day. The professor graded the first batch of 24 exams and found an average score of 74 points with a standard deviation of 8.9 points. The student who took the make-up the following day scored 64 points on the exam.

(a) Does the new student's score increase or decrease the average score?
(b) What is the new average?
(c) Does the new student's score increase or decrease the standard deviation of the scores?

1.28 Days off at a mining plant. Workers at a particular mining site receive an average of 35 days paid vacation, which is lower than the national average. The manager of this plant is under pressure from a local union to increase the amount of paid time off. However, he does not want to give more days off to the workers because that would be costly. Instead he decides he should fire 10 employees in such a way as to raise the average number of days off that are reported by his employees. In order to achieve this goal, should he fire employees who have the most number of days off, least number of days off, or those who have about the average number of days off?

1.29 Smoking habits of UK residents, Part I. Exercise 1.6 introduces a data set on the smoking habits of UK residents. Below are histograms displaying the distributions of the number of cigarettes smoked on weekdays and weekends, excluding non-smokers. Describe the two distributions and compare them.



1.30 Stats scores. Below are the final scores of 20 introductory statistics students.

79, 83, 57, 82, 94, 83, 72, 74, 73, 71,
66, 89, 78, 81, 78, 81, 88, 69, 77, 79

Draw a histogram of these data and describe the distribution.

1.31 Smoking habits of UK residents, Part II. A random sample of 5 smokers from the data set discussed in Exercises 1.6 and 1.29 is provided below.

gender	age	maritalStatus	grossIncome	smoke	amtWeekends	amtWeekdays
Female	51	Married	£2,600 to £5,200	Yes	20 cig/day	20 cig/day
Male	24	Single	£10,400 to £15,600	Yes	20 cig/day	15 cig/day
Female	33	Married	£10,400 to £15,600	Yes	20 cig/day	10 cig/day
Female	17	Single	£5,200 to £10,400	Yes	20 cig/day	15 cig/day
Female	76	Widowed	£5,200 to £10,400	Yes	20 cig/day	20 cig/day

(a) Find the mean amount of cigarettes smoked on weekdays and weekends by these 5 respondents.

(b) Find the standard deviation of the amount of cigarettes smoked on weekdays and on weekends by these 5 respondents. Is the variability higher on weekends or on weekdays?

1.32 Factory defective rate. A factory quality control manager decides to investigate the percentage of defective items produced each day. Within a given work week (Monday through Friday) the percentage of defective items produced was 2%, 1.4%, 4%, 3%, 2.2%.

(a) Calculate the mean for these data.

(b) Calculate the standard deviation for these data, showing each step in detail.

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

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

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

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

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

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

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

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

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

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

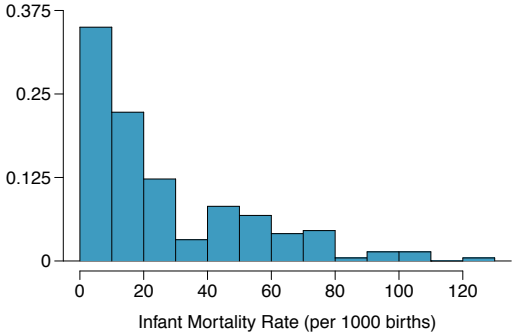
1.35 Box plot. Create a box plot for the data given in Exercise 1.30. The five number summary provided below may be useful.

Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

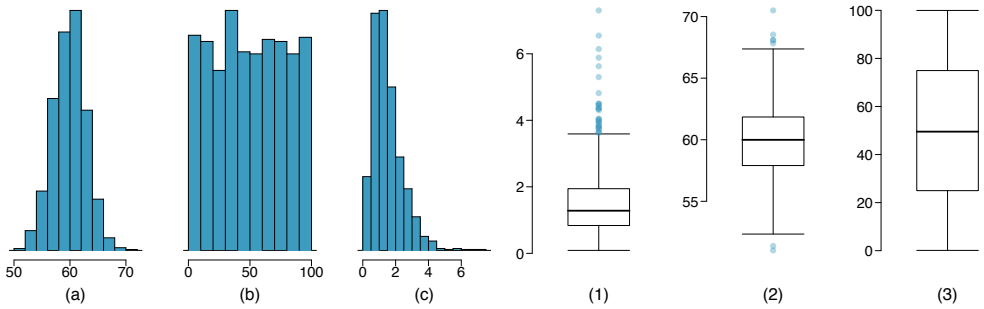
1.36 Infant mortality. The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates in 2012 for 222 countries.[68]

(a) Estimate Q1, the median, and Q3 from the histogram.

(b) Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning.

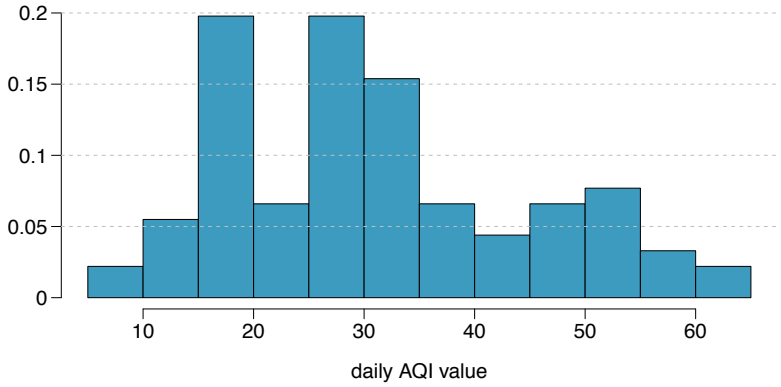


1.37 Matching histograms and box plots. Describe the distribution in the histograms below and match them to the box plots.



[68] CIA Factbook, Country Comparison: Infant Mortality Rate, 2012.

1.38 Air quality. Daily air quality is measured by the air quality index (AQI) reported by the Environmental Protection Agency. This index reports the pollution level and what associated health effects might be a concern. The index is calculated for five major air pollutants regulated by the Clean Air Act. and takes values from 0 to 300, where a higher value indicates lower air quality. AQI was reported for a sample of 91 days in 2011 in Durham, NC. The relative frequency histogram below shows the distribution of the AQI values on these days.[69]

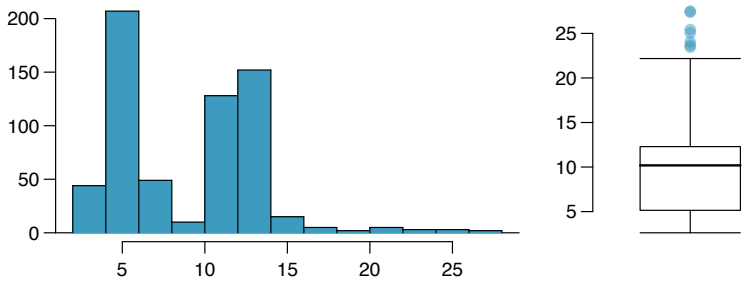


(a) Estimate the median AQI value of this sample.

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

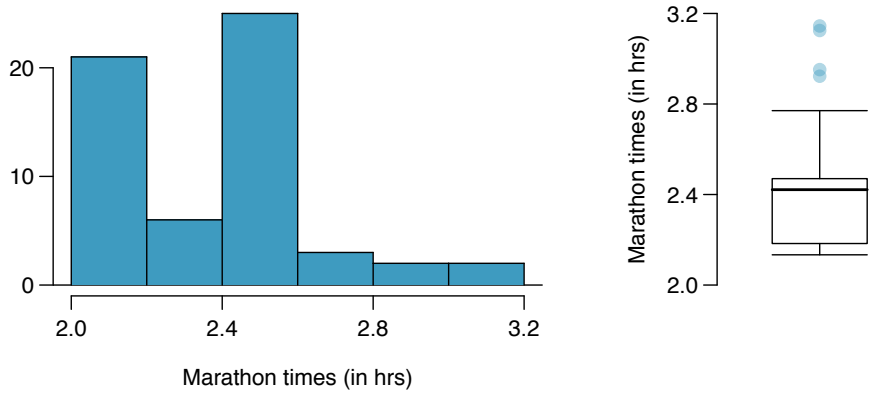
(c) Estimate Q1, Q3, and IQR for the distribution.

1.39 Histograms and box plots. Compare the two plots below. What characteristics of the distribution are apparent in the histogram and not in the box plot? What characteristics are apparent in the box plot but not in the histogram?



[69] US Environmental Protection Agency, AirData, 2011.

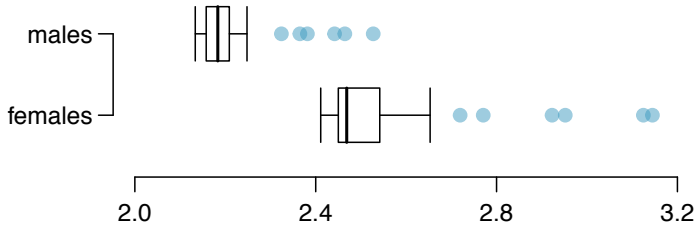
1.40 Marathon winners. The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1980 and 1999.



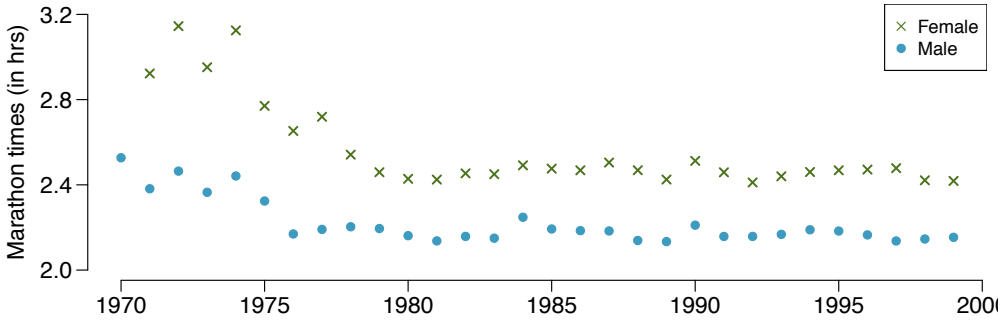
(a) What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

(b) What may be the reason for the bimodal distribution? Explain.

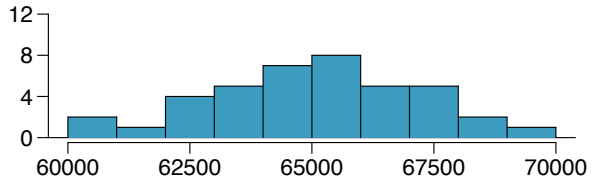
(c) Compare the distribution of marathon times for men and women based on the box plot shown below.



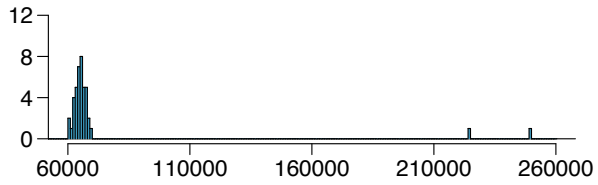
(d) The time series plot shown below is another way to look at these data. Describe what is visible in this plot but not in the others.



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



(1)



(2)

	(1)	(2)
n	40	42
Min.	60,680	60,680
1st Qu.	63,620	63,710
Median	65,240	65,350
Mean	65,090	73,300
3rd Qu.	66,160	66,540
Max.	69,890	250,000
SD	2,122	3,7321

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

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

1.42 Distributions and appropriate statistics. For each of the following, describe whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR.

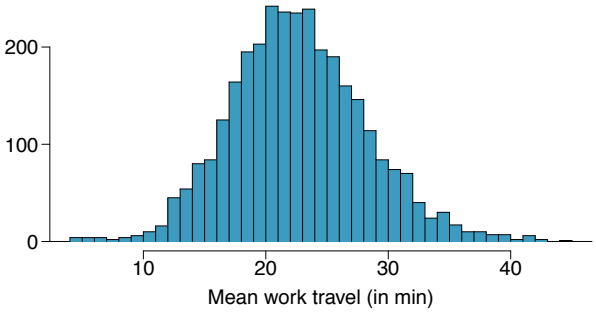
(a) Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.

(b) Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.

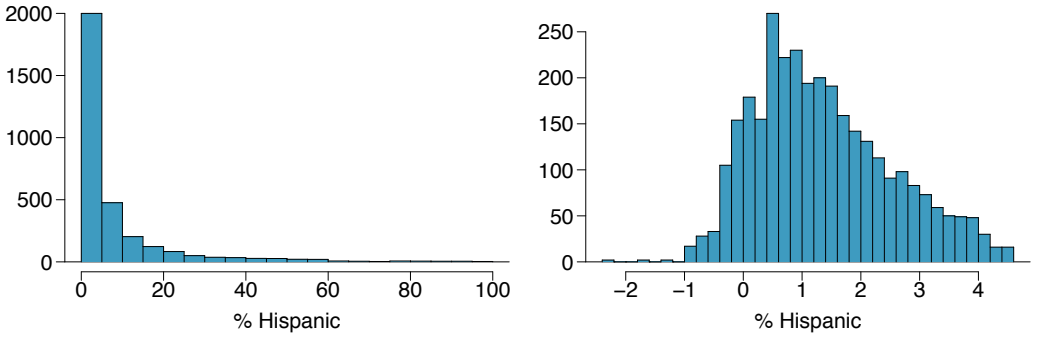
(c) Number of alcoholic drinks consumed by college students in a given week.

(d) Annual salaries of the employees at a Fortune 500 company.

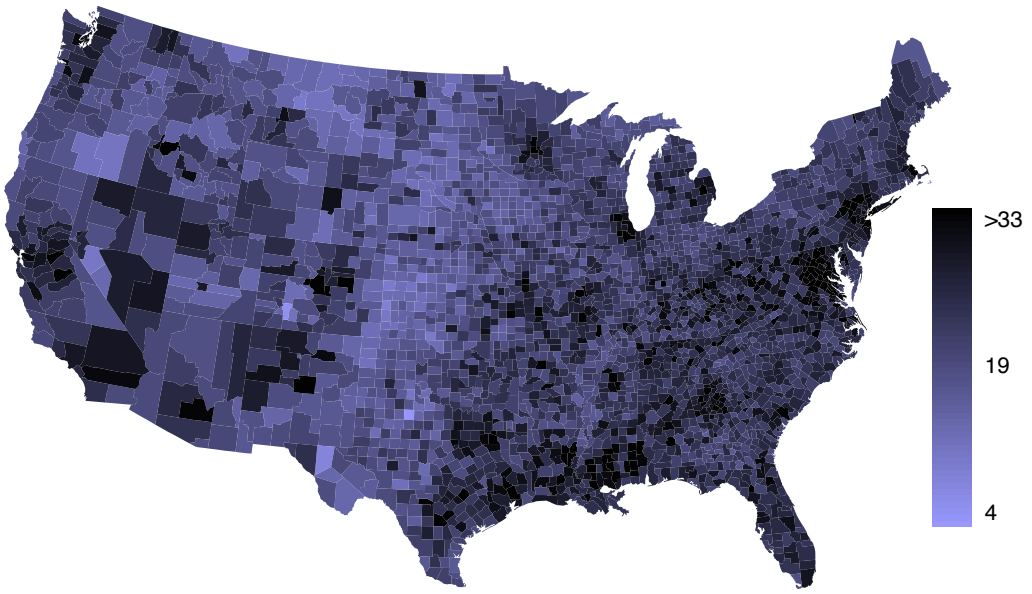
1.43 Commuting times, Part I. The histogram to the right shows the distribution of mean commuting times in 3,143 US counties in 2010. Describe the distribution and comment on whether or not a log transformation may be advisable for these data.



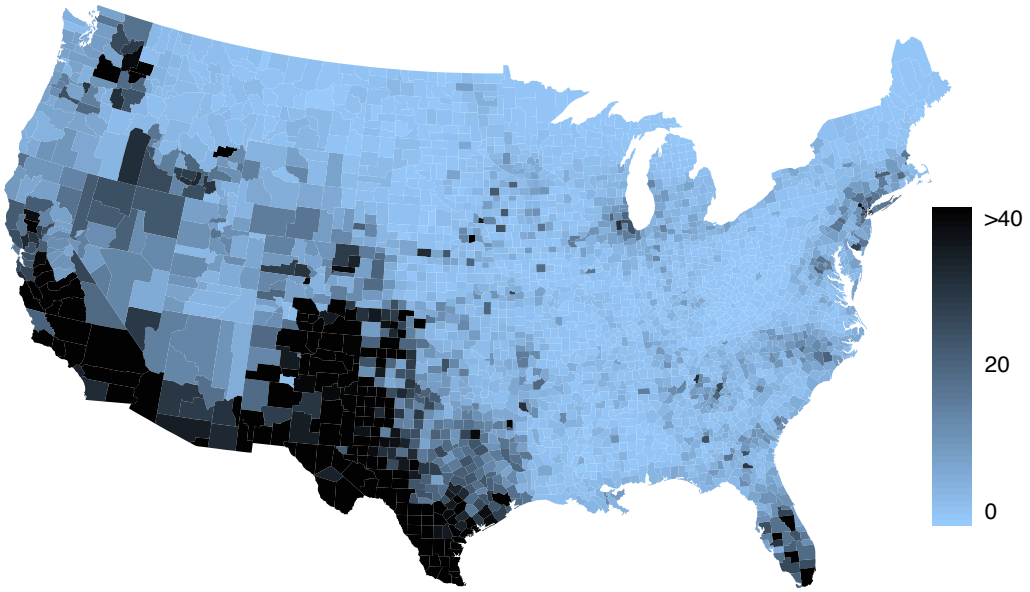
1.44 Hispanic population, Part I. The histogram below shows the distribution of the percentage of the population that is Hispanic in 3,143 counties in the US in 2010. Also shown is a histogram of logs of these values. Describe the distribution and comment on why we might want to use log-transformed values in analyzing or modeling these data.



1.45 Commuting times, Part II. Exercise 1.43 displays histograms of mean commuting times in 3,143 US counties in 2010. Describe the spatial distribution of commuting times using the map below.



1.46 Hispanic population, Part II. Exercise 1.44 displays histograms of the distribution of the percentage of the population that is Hispanic in 3,143 counties in the US in 2010.



(a) What features of this distribution are apparent in the map but not in the histogram?

(b) What features are apparent in the histogram but not the map?

(c) Is one visualization more appropriate or helpful than the other? Explain your reasoning.

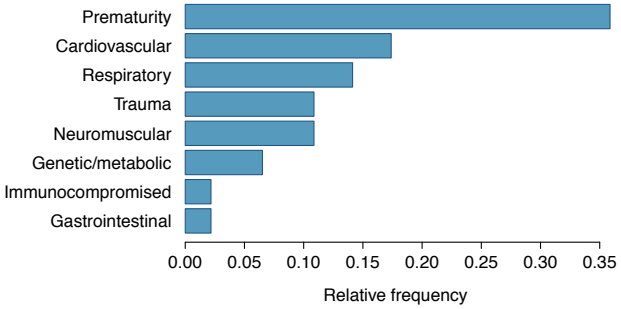
1.9.7 Considering categorical data

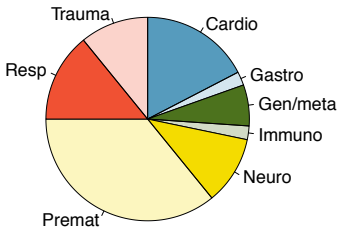
1.47 Antibiotic use in children. The bar plot and the pie chart below show the distribution of pre-existing medical conditions of children involved in a study on the optimal duration of antibiotic use in treatment of tracheitis, which is an upper respiratory infection.

(a) What features are apparent in the bar plot but not in the pie chart?

(b) What features are apparent in the pie chart but not in the bar plot?

(c) Which graph would you prefer to use for displaying these categorical data?





1.48 Views on immigration. 910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.[70]

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

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

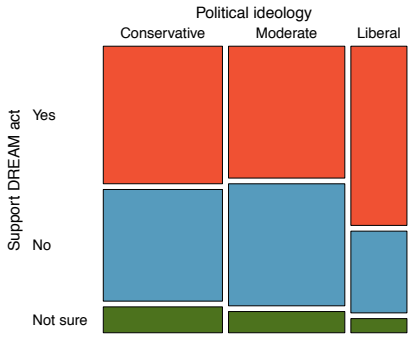
(b) What percent of these Tampa, FL voters are in favor of the citizenship option?

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

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

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

1.49 Views on the DREAM Act. The same survey from Exercise 1.48 also asked respondents if they support the DREAM Act, a proposed law which would provide a path to citizenship for people brought illegally to the US as children. Based on the mosaic plot shown on the right, are views on the DREAM Act and political ideology independent?



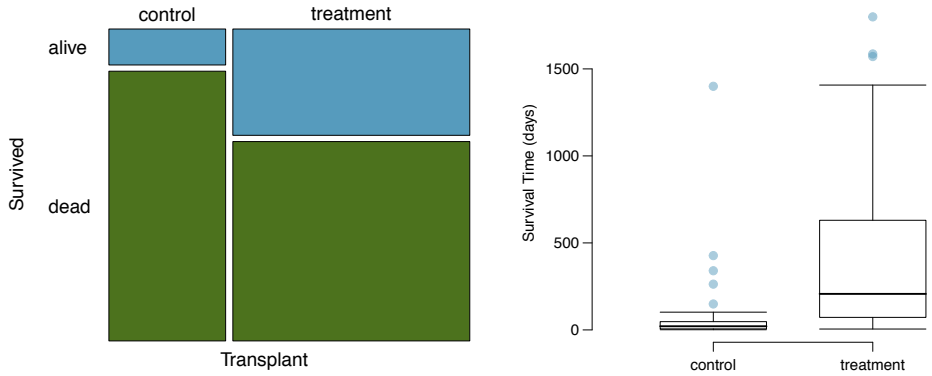
1.50 Heart transplants, Part I. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable `transplant` indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Another variable called `survived` was used to indicate whether or not the patient was alive at the end of the study. Figures may be found on the next page.[71]

(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

(b) What do the box plots suggest about the efficacy (effectiveness) of transplants?

[70] SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.

[71] B. Turnbull et al. "Survivorship of Heart Transplant Data". In: *Journal of the American Statistical Association* 69 (1974), pp. 74–80.



1.9.8 Case study: gender discrimination

1.51 Side effects of Avandia, Part I. Rosiglitazone is the active ingredient in the controversial type 2 diabetes medicine Avandia and has been linked to an increased risk of serious cardiovascular problems such as stroke, heart failure, and death. A common alternative treatment is pioglitazone, the active ingredient in a diabetes medicine called Actos. In a nationwide retrospective observational study of 227,571 Medicare beneficiaries aged 65 years or older, it was found that 2,593 of the 67,593 patients using rosiglitazone and 5,386 of the 159,978 using pioglitazone had serious cardiovascular problems. These data are summarized in the contingency table below.[72]

		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

Determine if each of the following statements is true or false. If false, explain why. *Be careful:* The reasoning may be wrong even if the statement's conclusion is correct. In such cases, the statement should be considered false.

(a) Since more patients on pioglitazone had cardiovascular problems (5,386 vs. 2,593), we can conclude that the rate of cardiovascular problems for those on a pioglitazone treatment is higher.

(b) The data suggest that diabetic patients who are taking rosiglitazone are more likely to have cardiovascular problems since the rate of incidence was (2,593 / 67,593 = 0.038) 3.8% for patients on this treatment, while it was only (5,386 / 159,978 = 0.034) 3.4% for patients on pioglitazone.

(c) The fact that the rate of incidence is higher for the rosiglitazone group proves that rosiglitazone causes serious cardiovascular problems.

(d) Based on the information provided so far, we cannot tell if the difference between the rates of incidences is due to a relationship between the two variables or due to chance.

[72] D.J. Graham et al. "Risk of acute myocardial infarction, stroke, heart failure, and death in elderly Medicare patients treated with rosiglitazone or pioglitazone". In: *JAMA* 304.4 (2010), p. 411. ISSN: 0098-7484.

1.52 Heart transplants, Part II. Exercise 1.50 introduces the Stanford Heart Transplant Study. Of the 34 patients in the control group, 4 were alive at the end of the study. Of the 69 patients in the treatment group, 24 were alive. The contingency table below summarizes these results.

		Group		
		Control	Treatment	Total
Outcome	Alive	4	24	28
	Dead	30	45	75
	Total	34	69	103

(a) What proportion of patients in the treatment group and what proportion of patients in the control group died?

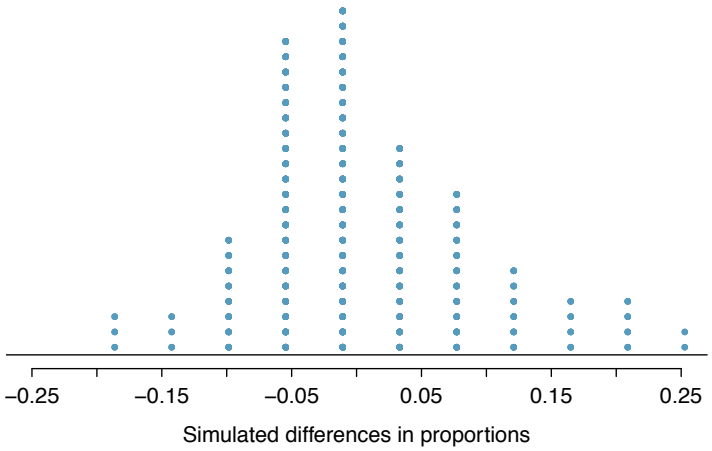
(b) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

i. What are the claims being tested?

ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

> We write *alive* on _____ cards representing patients who were alive at the end of the study, and *dead* on _____ cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size _____ representing treatment, and another group of size _____ representing control. We calculate the difference between the proportion of *dead* cards in the treatment and control groups (treatment - control) and record this value. We repeat this many times to build a distribution centered at _____. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are _____. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis (independence model) should be rejected in favor of the alternative.

iii. What do the simulation results shown below suggest about the effectiveness of the transplant program?



1.53 Side effects of Avandia, Part II. Exercise 1.51 introduces a study that compares the rates of serious cardiovascular problems for diabetic patients on rosiglitazone and pioglitazone treatments. The table below summarizes the results of the study.

		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

(a) What proportion of all patients had cardiovascular problems?

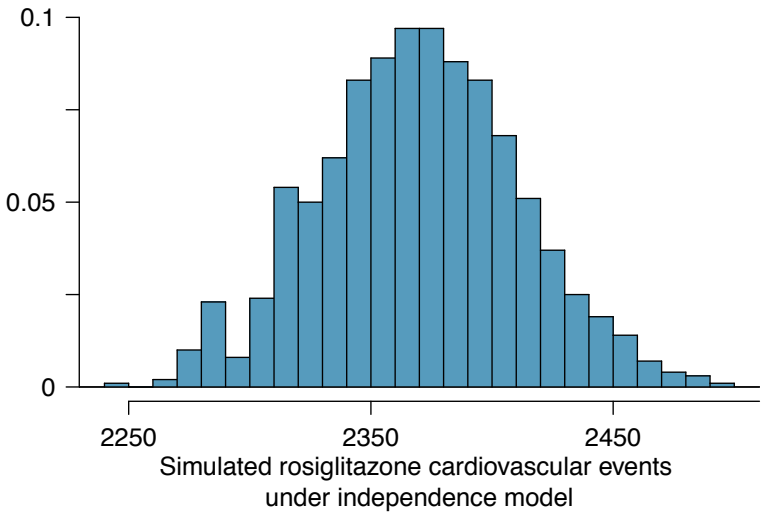
(b) If the type of treatment and having cardiovascular problems were independent, about how many patients in the rosiglitazone group would we expect to have had cardiovascular problems?

(c) We can investigate the relationship between outcome and treatment in this study using a randomization technique. While in reality we would carry out the simulations required for randomization using statistical software, suppose we actually simulate using index cards. In order to simulate from the independence model, which states that the outcomes were independent of the treatment, we write whether or not each patient had a cardiovascular problem on cards, shuffled all the cards together, then deal them into two groups of size 67,593 and 159,978. We repeat this simulation 1,000 times and each time record the number of people in the rosiglitazone group who had cardiovascular problems. Below is a relative frequency histogram of these counts.

i. What are the claims being tested?

ii. Compared to the number calculated in part (b), which would provide more support for the alternative hypothesis, *more* or *fewer* patients with cardiovascular problems in the rosiglitazone group?

iii. What do the simulation results suggest about the relationship between taking rosiglitazone and having cardiovascular problems in diabetic patients?



1.54 Sinusitis and antibiotics, Part II. Researchers studying the effect of antibiotic treatment compared to symptomatic treatment for acute sinusitis randomly assigned 166 adults diagnosed with sinusitis into two groups (as discussed in Exercise 1.2). Participants in the antibiotic group received a 10-day course of an antibiotic, and the rest received symptomatic treatments as a placebo. These pills had the same taste and packaging as the antibiotic. At the end of the 10-day period patients were asked if they experienced improvement in symptoms since the beginning of the study. The distribution of responses is summarized below.[73]

		Self reported improvement in symptoms		
		Yes	No	Total
Treatment	Antibiotic	66	19	85
	Placebo	65	16	81
	Total	131	35	166

(a) What type of a study is this?

(b) Does this study make use of blinding?

(c) At first glance, does antibiotic or placebo appear to be more effective for the treatment of sinusitis? Explain your reasoning using appropriate statistics.

(d) There are two competing claims that this study is used to compare: the independence model and the alternative model. Write out these competing claims in easy-to-understand language and in the context of the application. *Hint:* The researchers are studying the effectiveness of antibiotic treatment.

(e) Based on your finding in (c), does the evidence favor the alternative model? If not, then explain why. If so, what would you do to check if whether this is strong evidence?

[73] J.M. Garbutt et al. "Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial". In: *JAMA: The Journal of the American Medical Association* 307.7 (2012), pp. 685–692.