STAT 110
INTRODUCTION TO STATISTICAL REASONING

Spring 2024

Lecture Notes

Joshua M. Tebbs
Department of Statistics
University of South Carolina

Contents

1 Where Do Data Come From?

1.1 Introduction

Definition: Statistics is the science of data; how to interpret data, analyze data, and design studies to collect data.

- Statistics is used in all majors and areas of science, social science, and the humanities.
- "Statisticians get to play in everyone else's back yard." (John Tukey)

Example 1.1. Caffeine is commonly used to treat newborn infants for apnea of prematurity and to prevent the onset of other acute conditions. Known as "the silver bullet" in the treatment of prematurely born infants, caffeine is widely regarded within the neonatal care community to be safe and cost effective. It has also been approved by the US Food and Drug Administration for use with preterm infants due to its history of providing beneficial outcomes with no long-term adverse side effects. Cox et al. (2015) summarize a study where one of the research questions was:

- *Does treating premature infants with caffeine increase the chances of developing necrotizing enterocolitis?*

Necrotizing enterocolitis (NEC) is a serious condition characterized by infection and inflammation of the intestine. It is most commonly observed in premature infants. Left untreated, NEC can lead to serious health complications and even death.

In an 18-month period during 2008-2009, there were 615 infants admitted to the neonatal intensive care unit at Palmetto Richland Hospital in Columbia, SC.

- 35 out of 137 patients (26%) receiving caffeine developed NEC
- 10 out of 478 patients (2%) not receiving caffeine developed NEC.

What do these results say about caffeine and whether it increases the development of NEC? Does caffeine *cause* a greater percentage of NEC cases?

Definitions: Individuals are the objects observed in a study. Individuals are often people, but they don't have to be. A **variable** is a characteristic that we measure on each individual. Measurements of variables are called **data**.

- In the NEC study (Example 1.1), the individuals are the infants.
- There were many variables recorded in the NEC study. Here are some of them:
  - NEC (Yes/No)

  - Caffeine (Yes/No)
    - if "Yes," different doses were also recorded (in mg/kg/dose)
  - Birth weight (measured in grams)
  - Gestational age (measured in weeks)
  - Race (AA, Hispanic, White, Other)
  - Sex (M/F)
  - Nutrition type (Breastmilk, Fluids, Formula, TPN)
  - Maternal drug use (Yes/No)
    - if "Yes," what type (Alcohol, Cocaine, Marijuana, Tobacco)
  - Time to discharge from the NICU (measured in days)
  - Alive at discharge (Yes/No).

Definitions: A **categorical** variable places individuals into one of several groups or categories. A **quantitative** variable assumes numerical values. Measurements will be different for different individuals. This is what statisticians call **variation**.

Note: Graphs can be used to display the variation observed in one or more variables. In Figure 1.1 (next page), we use a **histogram** to display the variation in the birth weight data for the 615 infants. Birth weight is a quantitative variable. Its measurements are numerical in nature (e.g., 2402 grams).

1.2 Populations and samples

"You don't have to eat the entire pot of soup to know that it needs more salt."

Definitions: In a statistical study, the **population** is the entire group of individuals about which we want information. A **sample** is the part of the population we actually observe.

Discussion: In the NEC study (Example 1.1), the sample is the 615 infants admitted to the neonatal intensive care unit. We observed these individuals and we recorded data on them. What is the population in this example? In other words, what larger group of individuals do these 615 infants represent accurately?

Example 1.2. During December 11-15, 2023, Rasmussen Reports asked 1500 likely voters in the US to rate President Biden's job performance. Forty three percent (43%) of the respondents "approved" of his job performance.

- Sample: The 1500 likely voters contacted
- Population: Likely voters in the US.

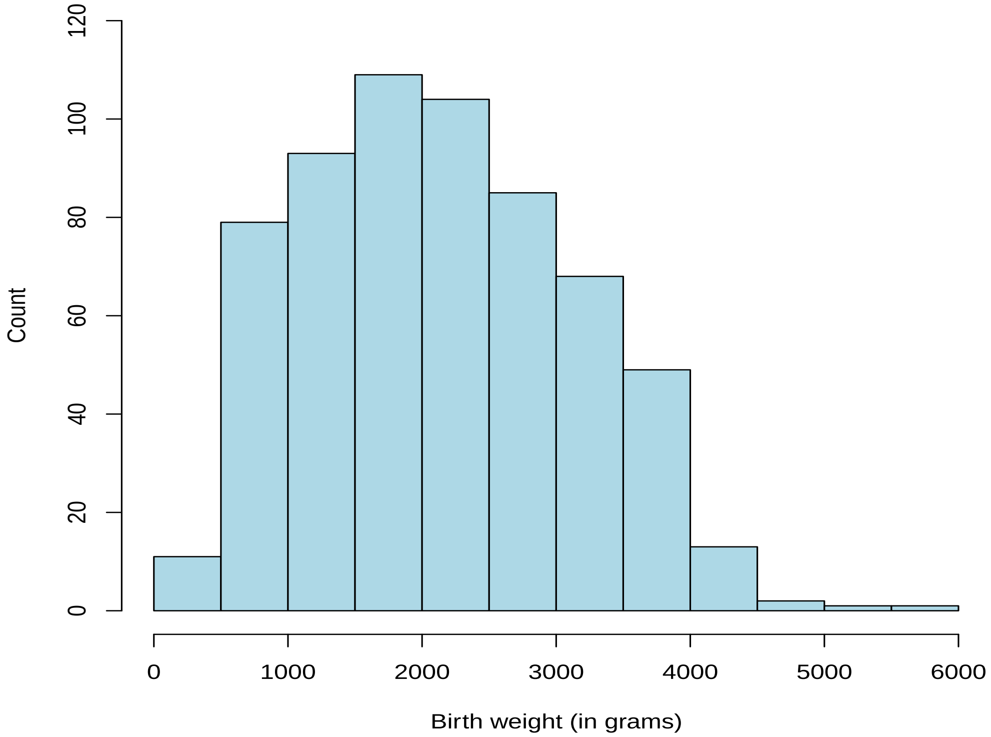


Figure 1.1: Necrotizing enterocolitis data. Histogram of birth weights for 615 infants. This figure was created using R.

Discussion: What do these results suggest about the population? The 43% approval rating is for the 1500 likely voters in the sample. Could the population approval rating be different? Rasmussen stated that the **margin of error** associated with their sample results was ±2.5%. What precisely does this mean?

Note: Example 1.2 describes the results of a **sample survey**. These are studies where individuals are contacted in person, over the phone, by mail or email, etc.

- Interpreting the results of a sample survey depends on the notions of population and sample.
- In most situations, the population is much too large to observe (e.g., all likely voters in the US, etc.). This is why we use samples.
- **Statistical inference:** What do the results from the sample suggest about the population of individuals?

Definition: A special type of survey occurs when the goal is to observe *every* individual in the population. This is called a **census**.

- The United States Constitution empowers the Congress to carry out a census for the American people. This started in 1790 and has occurred every 10 years since then.
- "The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct (Article 1, Section 2)."
- The results of the decennial census have broad implications; e.g., deciding how many representatives each state will have, apportionment of federal funds for underrepresented states/groups, etc.

1.3 Observational studies and experiments

Definition: An **observational study** passively observes individuals and measures variables of interest. There is no attempt to influence the responses.

- The NEC study (Example 1.1) and the Rasmussen survey (Example 1.2) are examples of observational studies.
- We observe individuals' responses (i.e., developed NEC/not?, approve/not?) and simply record this information.

Example 1.3. I recently reviewed a grant proposal for the Hong Kong Research Grants Council. The proposal described an observational study involving Hong Kong area high school students:

- 101 students who were "non-heavy" smartphone users; $<$ 6 hours/day
- 103 students who were "heavy" smartphone users; $\geq$ 6 hours/day.

One variable measured on each student was whether s/he experienced sleep problems (categorical). Another variable recorded was the number of steps each student took per day (quantitative).

Remark: In this example, you can imagine two populations:

- all HK high school students who are "non-heavy" smartphone users
- all HK high school students who are "heavy" smartphone users.

The investigators wanted to compare these two populations for the different variables recorded. Comparing two populations is a common goal in observational studies and experiments.

Definition: An **experiment** is a study where the investigators actively and deliberately impose some type of treatment or intervention on the individuals. This is done to see how individuals' responses are influenced by the treatment or intervention.

Example 1.4. Moore and Notz (pp 12-13) describe an experiment where welfare mothers in the US are randomly assigned to two groups:

- Group 1: No job training provided
- Group 2: Job training provided.

Unlike in an observational study, investigators actively assigned each individual (mother) to one of two treatment groups. We can then follow the two groups over time to see how they compare.

- Analyzing results from experiments usually compare the **averages** of the groups−not individual mothers.
- We could compare the groups using different variables:
  - Job placement? (Yes/No). This is a categorical variable.
  - Income (annual earnings in dollars). This is a quantitative variable.
- Statistical methods courses like STAT 201, STAT 205, or STAT 206 teach you how to perform statistical inference analyses for categorical and quantitative data. This is not the focus of this course.
- **Q:** Is it ethical to purposefully withhold job training from individual mothers?

Important: In observational studies, we passively observe the results. In experiments, we act deliberately and then observe the effect of doing so. If they are properly performed, experiments can give better information about **causality**. We will see why in Chapters 5-6. For example, does providing job training to welfare mothers *cause* an increase in annual earnings?

Example 1.5. *If a woman eats more cereal, does this increase her chances of having a male offspring?* As silly as this sounds, Matthews et al. (2008) claim they found "evidence" that this is true from an observational study.

- The results were published in a prestigious medical journal in the UK in 2008. When this "discovery" was picked up by the media, it made national headlines.
- *Doesn't Dad have his say?* From GEN 101, we know that males (XY) are heterogametic (i.e., males have non-matching sex chromosomes); females (XX) are not.
- Therefore, for the cereal claim to be plausible, we would have to either dismiss basic genetics knowledge or claim that a woman's cereal consumption helps to influence what chromosome (X or Y) the male passes on.

- Young et al. (2009) later demonstrated that this "finding" was easily explained by random chance. Investigators considered many dietary variables in the study, and they "cherry-picked" single ones which just happened to be associated with the male sex offspring (a spurious association).

Remark: The problem with many observational studies is that their "findings" cannot be **replicated**. In other words, if the same or similar study is performed again, the conclusions would be different.

- Unfortunately, silly claims like the cereal one are commonly thrust onto an uninformed public (and media) who blindly accept them as fact.
- As another example, I recently read online of a study concluding that wearing high heels causes cancer!

Discussion: The **replication crisis** is well documented. For example, in 2015, the Open Science Collaboration Project estimated that only about 1/3 of findings published in top psychology journals could be replicated.

- The replication crisis is a term that started in the social sciences (psychology, in particular), but the "hard sciences" have had their fair share of controversy too. Another recent OSCP effort showed that only 2 of 5 cancer biology experiments could be replicated.
- How should we think of replication in studies like the NEC study (Example 1.1) where we only observe a sample of 615 infants? Another research team could redo this study with a different sample of 615 premature infants and get a different result.
- We need to stop thinking that any single study or experiment is definitive proof. In fact, a provocative and highly cited article by Ioannidis (2005) makes a compelling argument that most published research findings are probably false.

References:

Cox, C., Hashem, N., Tebbs, J., Bookstaver, B., and Iskersky, V. (2015). Evaluation of caffeine and the development of necrotizing enterocolitis. *Journal of Neonatal-Perinatal Medicine* **8**, 339–347.

Ioannidis, J. (2005). Why most published research findings are false. *PLOS Science* **2**, 696–701.

Matthews, F., Johnson, P., and Neil, A. (2008). You are what your mother eats: Evidence for maternal preconception diet influencing foetal sex in humans. *Proceedings of the Royal Society B* **277**, 1661–1668.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* **349**.

Young, S., Bang, H, and Oktay, K. (2009). Cereal-induced gender selection? Most likely a multiple testing false positive. *Proceedings of the Royal Society B* **278**, 1211–1212.

2 Samples, Good and Bad

2.1 Introduction

Example 2.1. Moore and Notz (pp 3-4, 21) describe an online poll by the Michigan online news site MLive. In 2014, the site asked visitors,

Should Michigan legalize marijuana?

Of the 9684 respondents, 7906 (82%) of them said "Yes." In 2015, the Pew Research Center conducted a poll asking,

Do you think the use of marijuana should be made legal or not?

Among the 1500 participants, only 53% agreed on the legal use of marijuana. Why are these poll results so different? Here are some possible reasons:

- the sampling designs for the two polls are different: the online poll used a voluntary response sample whereas the Pew poll likely used a more sophisticated sampling design.
- the populations sampled are different: the online poll probably used mostly residents of Michigan (those who are more likely to visit MLive) whereas the Pew poll sampled US adult residents.

2.2 How to sample badly

Recall: The goal of statistical inference is to use the information in a **sample** of individuals to describe a larger **population** of individuals.

Definition: The way we select a sample from a population is called the **sampling design**. We would like our design to provide accurate results about the population.

Definition: If a sampling design systematically favors certain individuals over others, we call the design **biased**.

- A **convenience sample** collects individuals that are the easiest to contact.
- A **voluntary response sample** includes individuals who choose themselves to be included.

These designs are biased. They often underrepresent certain groups of individuals. Therefore, the sample will not be representative of the larger population.

Remark: Some students think that larger sample sizes are always better. However, large sample sizes do not make up for bad sampling designs; see Example 2.1. We have all heard of the aphorism, "Garbage in, garbage out." With larger biased samples, you simply have more garbage.

Example 2.2. You have been asked to inspect a truck shipment of oranges. The truck carries 100 boxes, each box with 100 oranges (10,000 oranges all together−this is the population). You select 5 boxes nearest to the door. In each box, you pick 5 oranges from the top. You have a sample of 25 oranges. This is a convenience sample. Why might this sample not be representative of the entire population of oranges in the shipment?

Example 2.3. In 2013, researchers in Peru aimed to assess whether there was an association between Facebook dependence and poor sleep quality. The researchers contacted students in their own classrooms before or after lectures to obtain observations on 418 undergraduate students. This is a convenience sample. This sample of students is likely not representative of all Facebook users.

Example 2.4. C-Span routinely provides viewers the opportunity to phone in and give comments about political issues. Recently, callers were asked to comment on the current state of affairs with Iran and North Korea. Do these callers' comments accurately reflect the views of the entire voting public in the US? Probably not. The calls received constitute a voluntary response sample. Individuals choose themselves for the sample, and these individuals tend to believe very strongly about issues being discussed.

Example 2.5. The 1936 presidential election between Landon (R) and Roosevelt (D) proved to shape the future of polling. *Literary Digest*, a magazine founded in 1890, had correctly predicted the outcomes of the 1916, 1920, 1924, 1928, and 1932 elections by conducting polls. Their 1936 "postal card poll" claimed to have asked one fourth of the nation's voters which candidate they intended to vote for.

- Based on their poll, *Literary Digest* predicted Landon would win the election with 57% of the popular vote and an electoral college margin of 370 to 161.
- In fact, Roosevelt won the election with 61% of the popular vote and an electoral college landslide of 523 to 8 (the largest ever). Roosevelt won 46 of 48 states, losing only Maine and Vermont.

What happened? The predictions were based on more than 2 million returned post cards. However, this was a voluntary response sample (only those who returned the post cards were included). In addition, the mailings only went to people who had 1936 auto registrations, who were listed in telephone books, and who were on *Literary Digest*'s subscription list! *Literary Digest* went bankrupt soon after this debacle.

- George Gallup, a pioneer of survey sampling techniques and inventor of the Gallup poll, correctly predicted the 1936 election using a much more representative sample of only 50,000 individuals.

- Interestingly, Gallup later badly botched the 1948 election prediction when they claimed Dewey (R) would beat Truman (D) by "5-15 percentage points." Truman won in a landslide. Gallup blamed their mistake on the fact that sampling ended three weeks before the election.
- Nate Silver, author of the FiveThirtyEight web site, gained national recognition when he correctly predicted each state's outcome in the 2012 presidential election between Romney (R) and Obama (D).
- Interestingly, Mr. Silver predicted in July 2015 that Donald Trump had only "a 5% chance" of winning the 2016 Republican nomination. In October 2016, he predicted Trump had "a 12% chance" of winning the 2016 presidential election.

2.3 Simple random samples

Recall: Convenience and voluntary response sampling designs are biased. They tend to systematically favor certain individuals over others. To avoid systematic undercoverage, use impersonal chance to select individuals.

Definition: A **simple random sample (SRS)** of size n has the same chance of being selected as any other sample of size n. As a result, each individual has the same chance of being selected.

- A simple random sampling design is **unbiased**. There is no systematic favoring of certain individuals over others.
- A simple random sampling design will give accurate results for a population over the long run. There is still no guarantee a particular SRS will be perfectly representative.
- The last remark may seem frustrating, but a SRS is a huge improvement over biased designs. It is our "best chance" to get accurate results about the population.

Example 2.6. We can use R to demonstrate how simple random samples can be taken. Suppose I want to take a SRS of $n = 5$ students from this class (the population). Suppose there are 160 students in the class, and each of you has a numerical coding assigned on my course enrollment list (the names listed below are fake).

Student	Code
Abherdine	1
Albert	2
Anderson	3
⋮	⋮
Yell	158
Zhang	159
Zynkowski	160

The R code below can be used to select 5 numerical codes between 1 and 160:

```
students = seq(1,160,1)
sample(students,5,replace=F)
```

When I ran this code, I got the following sample:

```
> sample(students,5,replace=F)
[1]  66 110  23 123  52
```

Students whose codes are 66, 110, 23, 123, and 52 would constitute the sample.

Important: Individuals in a SRS are selected at random. Therefore, if we repeated this exercise again, we would get a different sample:

```
> sample(students,5,replace=F)
[1]  98 114  20 146 86
```

Q: How many samples of size $n = 5$ are possible in this exercise?
A: Combinatoric rules from probability can be used to answer this question. It is 820,384,032! The key feature of a SRS is each of the 820,384,032 possible samples has the same chance of happening.

Remark: Selecting simple random samples from very large populations is not as straightforward as it sounds. For one, it may not be possible to get a reliable **sampling frame** (i.e., a list of all individuals in the population). For example,

- Population: USC undergraduates (Columbia campus). Size: 27,343 (as of Fall 2022)
- Population: SC (aged 18 and older). Size: 4 million (as of July 2019)
- Population: USA (aged 18 and older). Size: 255 million (as of July 2019).

In practice, those performing the study may not identify exactly what the population is, construct an exact sampling frame, and then choose a SRS from this list.

- It is more likely that "reasonable attempts" will have been made to choose individuals at random from those in the population.
- Although the sample obtained for analysis might not be a true SRS, it might not be that far off.

Note: Instead of using R's `sample` function to select simple random samples, the authors of your text suggest a **table of random digits** can be used; see Moore and Notz (Example 4, pp 27-28).

- This table is essentially a list of numbers determined at random.
- Using it will accomplish the same goal as using software like R.

Example 2.7. The College of Arts and Sciences (CAS) is the largest college at USC. Suppose Dean Samuels wants to form a committee by selecting the chairs of three ($n = 3$) departments using a simple random sample. There are 21 departments in CAS:

Department	Code	Geography	Code	Department	Code
AA Studies	01	Geography	08	Psychology	15
Anthropology	02	History	09	Religious Studies	16
Biology	03	LLC	10	Sociology	17
Chem/Biochem	04	Mathematics	11	Statistics	18
Criminal Justice	05	Philosophy	12	Theatre/Dance	19
EOE	06	Physics	13	Vis Art/Design	20
English	07	Political Sci	14	Women's/Gen Studies	21

Here are entries on line 101 in the Table of Random Digits (Moore and Notz, Table A):

19 22 39 50 34 **05** 75 62 87 **13** 96 40 91 25 31 42 54 48 28 53

Department chairs from Theatre and Dance (19), Criminal Justice (05) and Physics (13) would be selected.

Q: How many samples of size $n = 3$ are possible in this exercise?
A: There are 1,330 different samples of size $n = 3$. The key feature of a SRS is each of the 1,330 possible samples has the same chance of happening.

Had you entered instead at line 150, the random digits would have been:

07 51 **18** 89 **15** 41 26 71 68 53 84 56 97 93 67 32 33 70 33 16

and department chairs from English (07), Statistics (18), and Psychology (15) would be selected.

Summary: When selecting a sample of individuals, our goal is to choose one that is representative of the population.

- Biased designs: convenience, voluntary response. These designs suffer from systematic undercoverage of certain groups.
- Unbiased design: SRS. This design gives representative samples on average (over the long run).

Other sampling designs will be discussed in Section 4. These include stratified sampling, cluster sampling, and systematic sampling. Sampling in real life (from large populations) usually involves some form of stratification and clustering.

3 What do Samples Tell Us?

3.1 Parameters and statistics

Example 3.1. During December 3-5, 2023, Rasmussen Reports conducted a national telephone and online survey using a sample of $n = 992$ American adults. Each participant was asked,

Should Christmas be celebrated in public schools?

The survey found that 685 of the 992 adults in the sample answered "Yes" to this question (69%).

Definition: A **parameter** is a number that describes a population of individuals. Unless every individual in the population is observed, a population parameter is unknown.

Definition: A **statistic** is a number calculated from a sample of individuals. We can calculate the value of a statistic because it is based on the individuals observed in the sample.

Discussion: In Example 3.1, we can think of the population as "all American adults;" i.e., individuals aged 18 and over. There are approximately 255 million American adults (based on July 2019 `census.gov` data). Define

$$p \;=\; \text{population proportion of American adults who agree Christmas should be celebrated in public schools.}$$

Because p describes the population of all American adults (all 255 million of them), it is a parameter. We call p a **population proportion**. It is unknown.

What do we know? We do know that 685 out of the 992 American adults in the sample agreed that Christmas should be celebrated in public schools. Therefore, the **sample proportion** is

$$\widehat{p} = \frac{685}{992} \approx 0.69 \quad \text{(or 69\%)}.$$

The sample proportion $\widehat{p}$ is a statistic. It is calculated using the individuals in the sample.

Terminology: We say that the sample proportion $\widehat{p} = 0.69$ is an **estimate** of the population proportion p.

Main point: *We use sample statistics to estimate population parameters.* Want to estimate an unknown population parameter? Choose a sample from the population and use a sample statistic to estimate it. This is the idea behind statistical inference.

3.2 Bias and variability

Discussion: Rasmussen Reports found that 685 out of 992 American adults sampled are in favor of celebrating Christmas in public schools. Suppose Gallup did the same poll during the same time with the same number of individuals (992), asking the exact same question, and found that 715 out of 992 American adults were in favor of this. The sample proportion based on Gallup's sample is

$$\widehat{p} = \frac{715}{992} \approx 0.72 \quad \text{(or 72\%)}.$$

How can Gallup's sample proportion be different than Rasmussen's? That's easy. Each sample uses different people.

Important: Values of statistics like $\widehat{p}$ change from sample to sample because different samples contain different individuals. On the other hand, population parameters like p do not change. They represent the entire population.

Exercise: Let's use R to simulate many different sample proportions $\widehat{p}$. We are doing this so you can see that statistic's values do indeed change from sample to sample.

- I assumed that the population proportion is $p = 0.70$. In real life, this would be unknown, but I can pretend I know what it is for this classroom exercise.
- I used R to simulate 10,000 values of the sample proportion $\widehat{p}$ under this assumption. Each sample proportion is calculated from a **SRS** of size $n = 992$.
- I used a histogram to display the 10,000 sample proportions. This histogram is shown in Figure 3.1 (next page).

Observations: The histogram in Figure 3.1 reveals some important insights. Here are some of them:

- The sample proportion $\widehat{p}$ changes from sample to sample, but the values center at the truth about the population (i.e., at $p = 0.70$).
  - In other words, the sample proportion estimates are "correct on the average." This is what it means for an estimate to be **unbiased**.
  - We say that the sample proportion $\widehat{p}$ is an **unbiased estimate** of the population proportion p.
  - Unbiasedness (i.e., no bias) is guaranteed when we use a SRS. This is why we use them.
  - Unbiasedness is not guaranteed when we used biased sampling designs like convenience and voluntary response samples!
- The Rasmussen and Gallup sample proportion estimates (0.69 and 0.72, respectively) are close to the true population proportion $p = 0.70$, but some estimates are much further away; e.g., 0.65, 0.75, etc.

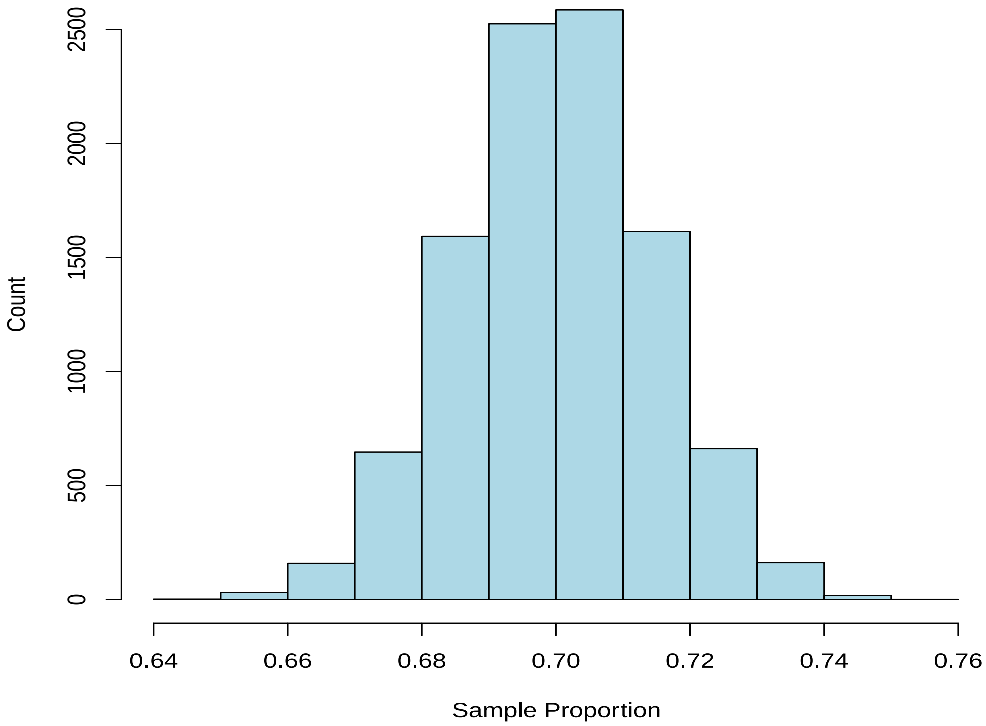


Figure 3.1: 10,000 sample proportions $\widehat{p}$. Each one is based on a SRS of $n = 992$ individuals. The population proportion is $p = 0.70$.

- The spread in the histogram gives us information on precision; i.e., how variable the sample proportion estimates are.
- Rasmussen used a sample size of $n = 992$ American adults. How can we make the sample proportion estimate $\widehat{p}$ more precise (i.e., less variable)? **Answer:** Take a larger SRS (see next simulation).

- The shape of the histogram resembles a **normal distribution**. We will study normal distributions in Chapter 13.

Another simulation: We repeat our simulation exercise under identical conditions, except we now use a sample size of

$$n = 992 \times 10 = 9920.$$

In other words, we are now pretending that each sample is <u>10 times larger</u> than the Rasmussen sample. This would emulate the situation where Rasmussen sampled 10 times as many American adults as it did originally.

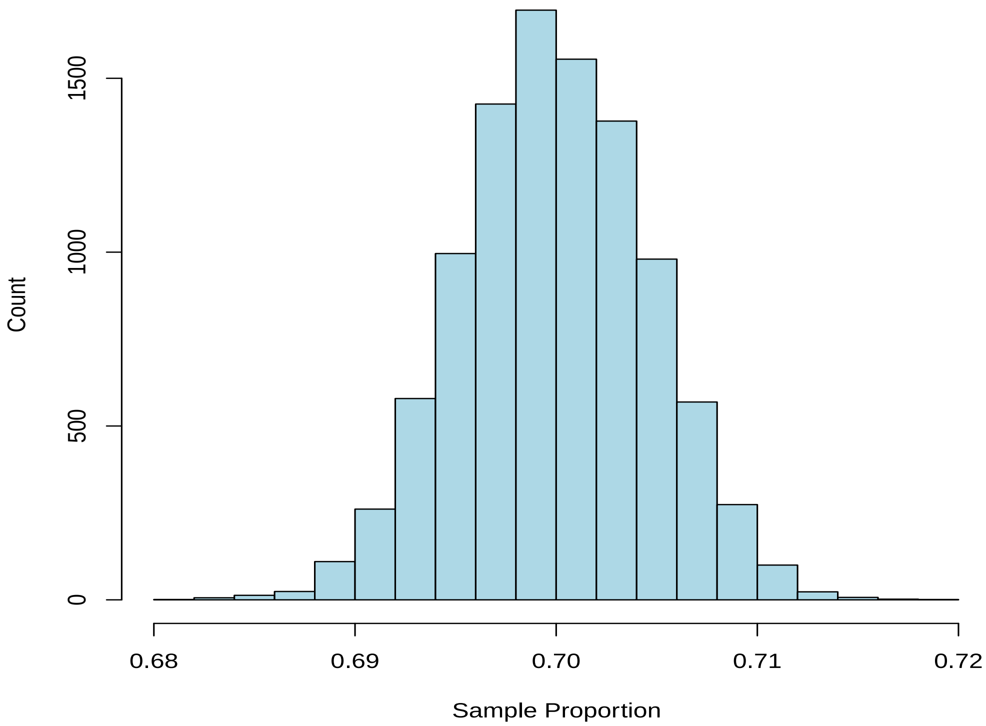
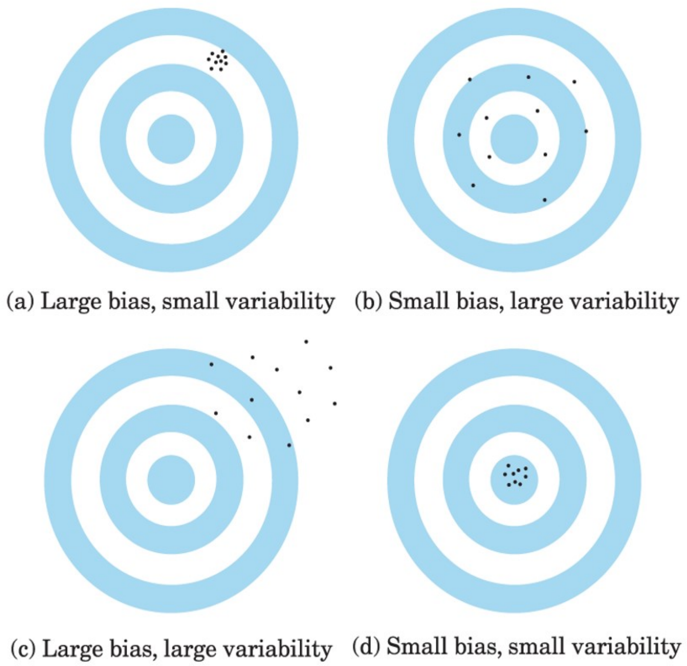


Figure 3.2: 10,000 sample proportions $\widehat{p}$. Each one is based on a SRS of $n = 9920$ individuals. The population proportion is $p = 0.70$.

Observations: The histogram in Figure 3.2 displays the results from this simulation:

- The sample proportions $\widehat{p}$ remain centered at the truth $p = 0.70$.
  - The sample proportion estimate $\widehat{p}$ remains **unbiased**. Bias does not depend on sample size (as long as a SRS is used).
- We see that the same normal distribution pattern emerges.
- The only difference from increasing the sample size (from $n = 992$ to $n = 9920$) is that the variability in the estimates has decreased (look at the scale/range on the horizontal axis).
  - In Figure 3.2 ($n = 9920$), the variability is much smaller than it was in Figure 3.1 ($n = 992$). In other words, the estimate $\widehat{p}$ is more precise.
- This suggests that we can improve the precision of the estimate $\widehat{p}$ by taking a larger SRS. Indeed, this is true.



(a) Large bias, small variability (b) Small bias, large variability

(c) Large bias, large variability (d) Small bias, small variability

Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020
W. H. Freeman and Company

Remark: We want sample estimates like $\widehat{p}$ to be **accurate** and **precise**. Although "accuracy" and "precision" sound similar in everyday English expression, they have very different meanings (at least in statistics they do).

- Accuracy deals with bias. Precision deals with variability.

Bias is a consistent repeated deviation of a sample statistic from a population parameter if we took many samples. **Variability** describes how spread out the values of a sample statistic are if we took many samples.

- The dartboard analogy above describes these two concepts beautifully. The dartboard in the lower right shows what would happen if our sample estimate had low bias and low variability. This is the best we can hope for when sampling from a population.

Main point: A good sample produces small bias (or no bias) and small variability.

- To eliminate bias, use a SRS. This design produces estimates that are **unbiased**. In other words, we neither overestimate nor underestimate the value of the population parameter on average.

- To reduce the variability in a SRS, use a larger sample. The larger the sample size, the smaller the variability (i.e., the more precise our estimates are).

3.3 Margin of error and confidence statements

Recall: In Example 3.1, we discussed the Rasmussen survey question:

Should Christmas be celebrated in public schools?

The survey found that 685 of the 992 adults in the sample answered "Yes" to this question (69%).

- Rasmussen stated "the margin of sampling error is ±3 percentage points with a 95% level of confidence." What does this statement mean?

Reality: Even if we use a SRS with a large sample size (like $n = 992$), we cannot be sure how close our estimate

$$\widehat{p} \approx 0.69$$

is to the population proportion p. We do not sample the entire population, so no one knows what p is. As a result, there will always be a degree of uncertainty with our estimate.

Q: How could we determine what p is exactly in this example?
A: Sample all 255 million American adults and ask them (i.e., perform a census). How large is this number? If you contacted one person each minute starting now, and never took any breaks, it would take you 475.6 years to complete the census! Now, you know why samples are used to **estimate** population parameters like p.

Definition: The **margin of error** is a number that describes how precise a sample estimate is.

- The smaller the margin of error, the more precise (less variable) the estimate is.
- Large margins of error indicate less precision (more variability).

Note: Rasmussen found the sample proportion of adults who believe Christmas should be celebrated in public schools to be

$$\widehat{p} = \frac{685}{992} \approx 0.69 \text{ (or 69\%)}.$$

They also stated that the margin of error was ±3 percentage points with a 95% level of confidence. This allows us to write the following **confidence statement**:

- "We are 95% confident that the proportion of American adults who believe Christmas should be celebrated in public schools is between 0.66 and 0.72 (i.e., between 66% and 72%)."

Definition: A **confidence statement** is a statement about a population parameter. It contains two parts:

- **Margin of error.** This quantifies how close the sample statistic (estimate) is to the population parameter. This is related to variability.
- **Level of confidence.** Our statement about the population is never completely certain. The level of confidence tells us how confident we are in the statement.

Discussion: The confidence statement in the Rasmussen poll example is an illustration of statistical inference. We are using a sample statistic $\widehat{p}$ and its margin of error to make a statement about the population proportion p.

- We are 95% confident the population proportion p is between 0.66 and 0.72, that is,
$$0.66 < p < 0.72.$$
This is a statement about the population of all American adults.

Q: What does the phrase "95% level of confidence" mean? This was the phrase Rasmussen used to accompany its margin of error.
A: First of all, what this means is that we are not 100% confident. This is an unfortunate reality when performing statistical inference for any population.

- Remember, the only way we can be 100% sure about the population is to perform a census (and we know what that would entail).
- Therefore, do we know for a fact that
$$0.66 < p < 0.72$$
and that this interval is correct about the population proportion p? No, we don't. In other words, it could be that the population proportion p is outside this interval.
- The phrase "95% confidence" only assures that 95% of the time our sample will produce an estimate $\widehat{p}$ whose corresponding confidence statement will be correct.
- In other words, 5% of the time, our sample will produce a confidence statement that is not correct.
- How do we know if Rasmussen's sample is one of the 95% which produced a correct statement about p? We don't.
- This is the tradeoff we must live with. We make our lives easier by sampling from a population (and not having to take a census). But the price we pay for this is there is always uncertainty about our inferences.

Q: How do we calculate the margin of error for a sample?
A: This is a hard question to answer in general. However, the following result is true.

Result: If a SRS is used, the margin of error in the sample proportion $\widehat{p}$ associated with a 95% confidence level is approximately equal to

$$\textbf{margin of error} = \frac{1}{\sqrt{n}},$$

where n is the sample size. This formula is only applicable for a SRS design with a 95% level of confidence. If either of these changed, then the formula would change.

Discussion: We now see mathematically why larger samples produce more precise statistical inference (when a SRS is used). Because the sample size n is in the denominator of this formula, increasing n will decrease the margin of error. For example,

$$\begin{aligned} n = 100 &\implies \frac{1}{\sqrt{100}} = \frac{1}{10} = 0.10 \quad \text{(or 10\%)} \\ n = 400 &\implies \frac{1}{\sqrt{400}} = \frac{1}{20} = 0.05 \quad \text{(or 5\%)} \\ n = 1600 &\implies \frac{1}{\sqrt{1600}} = \frac{1}{40} = 0.025 \quad \text{(or 2.5\%).} \end{aligned}$$

This shows that to cut the margin of error in half, the sample size must increase by four times.

Example 3.2. During September 25-October 1, 2023, Pew Research Center conducted a survey using a sample of $n = 8,842$ American adults. Each participant was asked,

Would you support or oppose the U.S. government banning TikTok?

The survey found that 38% of the adults in the sample would support the US government banning TikTok.
(a) If this was a SRS, then what is the margin of error associated with this estimate assuming a 95% level of confidence?
(b) After you calculate the margin of error, write the corresponding confidence statement.

Solutions: In part (a), we use the margin of error formula

$$\textbf{margin of error} = \frac{1}{\sqrt{8842}} \approx \frac{1}{94.03} \approx 0.01.$$

The margin of error is 0.01 or 1%. In part (b), our confidence statement is

- “We are 95% confident that the proportion of American adults who support the US government banning TikTok is between 0.37 and 0.39 (i.e., between 37% and 39%).”

Example 3.3. During October 1-23, 2023, Gallup conducted a telephone survey using a sample of $n = 1,009$ American adults. Each participant was asked,

Is there more crime in your area than there was a year ago, or less?

The survey found that 55% of the adults in the sample said there was more crime than a year ago.
(a) If this was a SRS, then what is the margin of error associated with this estimate assuming a 95% level of confidence?
(b) After you calculate the margin of error, write the corresponding confidence statement.

Solutions: In part (a), we use the margin of error formula

$$\textbf{margin of error} = \frac{1}{\sqrt{1009}} \approx \frac{1}{33.30} \approx 0.03.$$

The margin of error is 0.03 or 3%. In part (b), our confidence statement is

- "We are 95% confident that the proportion of American adults who say there is more crime than a year ago is between 0.52 and 0.58 (i.e., between 52% and 58%)."

Final thoughts: Confidence statements are a form a statistical inference. We are taking the results from a sample, and we are making a statement about the population.

Q: Can we use a larger level of confidence?
A: We can, but the margin of error formula

$$\textbf{margin of error} = \frac{1}{\sqrt{n}}$$

would no longer be correct. This formula is only approximately correct for 95% confidence statements about a population proportion p. We will show later in the course (Chapter 21) how to include different levels of confidence−including those which are larger and smaller.

Remark: The margin of error formula above is only applicable for statistical inference problems where the population is really large.

- The three examples in this chapter (Rasmussen, Pew, and Gallup) have all featured the same population:
  - American adults (aged 18 and older). Size: 255 million (as of July 2019).
- We would not want to use this formula for smaller populations, like this class (160 students) or in Example 2.7 (21 department chairs).
- Moore and Notz describe how the sampling variability of a statistic (which is what the margin of error describes numerically) is unaffected as long as the population size is at least 20 times that of the sample size. Clearly, there is nothing to worry about when we are sampling from all American adults.

4 Sample Surveys in the Real World

4.1 Introduction

Example 4.1. A Pew Research Center opinion poll conducted over the phone reports a sample of size $n = 1,507$ people. Your favorite web site posts the poll results. You read the results and are satisfied because (at the bottom of the article) you see words like "probability sample," "margin of error," and "95 percent confidence."

However, let's look more closely at what it took to secure these 1,507 responses:

	Landline	Cell	Total
Noncontacts	2,464	3,114	5,578
Not eligible	18,427	6,048	24,475
Other	104	56	160
Unknown eligibility	3,305	361	3,666
Refusals, breakoffs, and partials	4,719	4,836	9,555
Complete interview	902	605	**1,507**
Totals	29,921	15,056	44,977

It's pretty likely the web site you read did not give this information. That is, to secure 1,507 responses, Pew had to contact 44,977 people! This gives a **response rate** of

$$\frac{1507}{44977} \approx 0.034 \quad \text{(or 3.4\%)}.$$

In other words, the nonresponse rate is 96.6%!

- Reporting a large sample size like $n = 1,507$ seems impressive until you dig deeper.
- What about the 43,470 people that didn't respond?
- Do you believe the poll results are representative of the larger population now?

Reality: Getting feedback from humans in sample surveys is not easy. Many people will not participate over the phone, in person, or through mail/email. Response rates in sample surveys are usually low like in this example.

Preview: This chapter describes how sample surveys can go wrong and what can be done to mitigate the problems that arise. Not all problems can be eliminated, and even if they could, we still never get to make confidence statements (about a population) which are free of uncertainty. This chapter is illuminating, because you will learn just how bad things can be. If nothing else, you should learn to be skeptical of sample surveys you see in the news, even if it appears everything is statistically sound.

Preview: There are two types of errors that can occur in sample surveys:

1. Sampling errors
2. Nonsampling errors

4.2 Sampling errors

Terminology: Sampling errors are errors that arise through the act of selecting a sample. There are two types of sampling errors:

1. Biased sampling designs
   - convenience sample, voluntary response sample
   - these designs lead to biased samples which are not representative of the population.
2. Random sampling error.

Definition: Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.

- This is the main problem with convenience samples. By sampling those that are the "easiest to reach," you exclude many groups of individuals.
- Similarly, voluntary response samples will usually include only those individuals who feel most strongly about the issue (strongly in favor/strongly against). This leaves out most people in the middle.

We can avoid biased designs if we use simple random samples. However, **random sampling error** is unavoidable. This is the natural error that arises because a simple random sample is only a part of the population. Recall the confidence statement we wrote in Example 3.3:

- "We are 95% confident that the proportion of American adults who say there is more crime than a year ago is between 0.52 and 0.58 (i.e., between 52% and 58%)."

Important: The margin of error in a confidence statement only accounts for random sampling error. It does not account for anything else.

Q: How can we *reduce* random sampling error?
A: Use larger samples.

Q: How can we *eliminate* random sampling error?
A: Take a census.

4.3 Nonsampling errors

Conceptualization: Suppose you have selected a simple random sample from a population. You're off to a great start! Now, what can go wrong?

Terminology: Nonsampling errors are errors that arise for reasons not related to the act of sampling. Here are examples:

- Incomplete or inaccurate sampling frame
- nonresponse (this is the biggest one in sample surveys)
- poorly worded or misleading questions
- interviewer bias
- data entry or processing errors
- individuals giving false responses (either on purpose or because they don't know).

Example 4.2. Many opinion polls in the US still conduct interviews by telephone using random digit dialing. If every American adult (255 million) had exactly one phone number, then this would be an ideal way to establish a reliable sampling frame. However, which groups of individuals are <u>excluded</u>?

- Individuals without phones (about 2% of the US population?)
- Other groups, for example, military members overseas, individuals in long-term care hospitals, prisoners, etc.

In addition, certain individuals could have multiple phone lines (duplication), in which case they could be included more than once.

Recall: In Example 2.5, we talked about the *Literary Digest* example and how predicting the 1936 presidential election between Landon (R) and Roosevelt (D) was badly botched. Recall, the sampling frame consisted of

- people who had 1936 auto registrations
- people who were listed in telephone books
- people who were on *Literary Digest*'s subscription list.

Even if *Literary Digest* had taken a simple random sample from this sampling frame (they didn't), a biased sample would have likely still resulted. Unfortunately, almost 100 years later, political polling isn't that much better.

Main point: Even when simple random sampling is used, **sampling frame errors** can still lead to undercoverage. Certain groups of individuals in the population will not be available for selection. Leaving these groups out causes bias.

Terminology: Nonresponse occurs when an individual is selected for a sample but investigators are unable to obtain responses from them.

- This is usually the largest source of nonsampling error in sample surveys.
- Recall in Example 4.1, Pew experienced a nonresponse rate of almost 97%!
- If nonresponse happens for random reasons, then this may not cause bias. However, certain groups of individuals may not want to participate because they do not feel comfortable sharing their responses on certain topics. This does cause bias.

The **response rate** in a sample survey is

$$\frac{\text{number of individuals who complete the survey}}{\text{number of individuals contacted}} \times 100\%.$$

The **nonresponse rate** is simply "100% minus the response rate."

Q: What can be done to mitigate the effects of nonresponse?
A: A common tactic in sample surveys is to reach out to a nonresponding individual in another way; e.g., following up an incomplete phone call with a personal letter requesting a mail-in response. This can decrease the nonresponse rate. Other strategies include:

- substitute "similar" households of nonresponders
- use **probability weighting** to weight responses differently. For example,
  - if the response rate from a certain race group is below the national average, give them more weight.
  - if the response rate from rural areas is above the national average, give them less weight.
  - Here is an actual excerpt from a *New York Times* sample survey:

    > *"The results have been weighted to take account of household size and number of telephone lines into the residence and to adjust for variations in the sample relating to geographic region, sex, race, age, and education."*

  - Probability weighting techniques are usually complex and mathematical (especially when adjusting margins of error).

Example 4.3. In 2006, a political polling organization asked the following question:

> *"In light of the mounting casualties we have seen recently, do you approve of the way President Bush is handling the war in Iraq?"*

No one likes "mounting casualties." This question is misleading and encourages negative responses. Here are other examples:

- *"Is our government providing too much money for welfare programs?"*
    - 44 percent said "Yes." When "welfare programs" was replaced with "assistance to the poor," only 13% responded "Yes." Question wording can greatly influence the results.
- *"Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?"*
    - This question is poorly worded. 22 percent of the sample said that it was "possible." A much simpler version of this question was later asked and only 1% of the respondents said it was "possible."

The **wording of questions** in a sample survey always influences the results. Poorly worded and misleading questions can create serious nonsampling error.

Example 4.4. Singer (2005) described the results of a sample survey carried out among biology teachers in Louisiana. The following excerpt appeared in a *New York Times* editorial, published on February 4, 2005:

> *"A 1998 doctoral dissertation found that 24 percent of the biology teachers sampled in Louisiana said that creationism had a scientific foundation and that 17 percent were not sure."*

Q: Could 41% of the biology teachers in Louisiana reject or not be sure of evolution?

Remark: Another source of nonsampling error is **response error**. This happens when individuals give incorrect responses. This could happen on purpose, or it could be because the participant cannot recall specific information. The latter is common with **self-reported data**. For example, a survey may ask,

"How many cigarettes did you smoke last week?"

It is unlikely that smokers will be able to remember the exact number. In other situations, individuals may feel embarrassed about the topic being discussed or they may not feel comfortable sharing their true responses. For example, suppose someone asked you,

"Have you ever drink-spiked someone to take advantage of them sexually?"

Guilty individuals may be inclined to lie for obvious reasons.

Randomized-response technique: This is a sample survey method that can preserve individual anonymity. This encourages participants to provide truthful responses to sensitive questions. Here is an example of how it works.

Example 4.5. Frenger et al. (2016) summarize a sample survey involving professional German athletes. The question investigators wanted to ask was,

> *Have you used prohibited substances/methods with the aim to enhance your sports performance?*

This question was asked as part of a randomized-response survey. Each athlete was presented with two questions:

1. Have you used prohibited substances/methods with the aim to enhance your sports performance?
2. Are there 7 days in a week?

To decide which question an athlete would answer, s/he would be given a list of randomly generated numbers such as

```
8 4 4 3 8 1 3 3 9 3
3 3 0 4 4 9 1 7 1 6
6 7 9 6 7 1 4 1 4 9
5 0 4 4 7 8 2 5 0 2
6 2 9 0 9 3 2 3 5 1
```

and then asked to choose one number.

- If this number was a 1, 2, or 3, then s/he would answer Question 2.
- Otherwise, s/he would answer Question 1.

Individual athlete anonymity is preserved as long as the athlete never reveals which number s/he chose. Why? A "Yes" answer in this survey for any athlete could be a response to the innocuous Question 2!

Discussion: If the numbers above are truly random and equally likely, then the probability of a "Yes" response in this example is

$$0.7p + 0.3(1) = 0.7p + 0.3,$$

where

$$p = \text{population proportion of professional German athletes who have used prohibited substances/methods.}$$

Therefore, the sample proportion of "Yes" responses is not an unbiased estimate of p, but instead of $0.7p+0.3$. Simple algebra can then be used to create an unbiased estimate of p. In other words, one can estimate the population proportion of this sensitive characteristic while protecting the anonymity of each athlete in the survey.

4.4 Other (good) sampling designs

Terminology: A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance (probability) each possible sample has. Here are examples:

- Simple random sample (SRS)
- Stratifed random sample
- Cluster random sample
- Systematic random sample.

Recall in a SRS, each sample of size n has the same chance of being selected. This is the defining characteristic of a SRS.

Terminology: A **stratified random sample** arises in two steps:

1. Divide the sampling frame into groups of individuals, called **strata**. Strata could be formed by sex, race, income class, political affiliation, etc.
2. Take a SRS from each stratum and combine each SRS to form the complete sample.

The reason for stratification is simple. We can guarantee equal representation (or proportional representation) of individuals from each stratum. Most polling organizations use some type of stratification to ensure the sample selected is representative of the larger population (e.g., all American adults, all likely voters in the US, etc.).

Example 4.5. Williams (1990) summarized a large seroprevalence study in Houston, Texas to estimate HIV positivity among heterosexual males who were intravenous drug users (IVDUs) and were not receiving treatment for their addiction. Here were the results from their study:

Race	Sample size	Infected	Sample proportion
Hispanic	107	4	0.037 (3.7%)
White	214	14	0.065 (6.5%)
Black	600	59	0.098 (9.8%)
Total	921	77	

In this example, investigators split the population into three race strata: Hispanic, White, and Black. Individuals in each stratum gave demographic and risk information. Infectivity was determined by enzyme-linked immunosorbant assay (ELISA).

Discussion: Do you think each of the three samples above is a true SRS? Probably not. Given the sensitive nature of this study, there would be no way to know all members of the population.

Example 4.6. A large university has 25000 undergraduate students and 5000 graduate students. The university newspaper wants to get the opinions of students on a topic involving free speech. The Editor asks his staff to use a stratified random sample with

- 200 undergraduate students
- 50 graduate students.

There are two strata in this example: undergraduate students and graduate students.

Q: Does each sample of size 250 have the same chance of being selected? Recall this is the defining characteristic of a SRS.
A: No, it is not possible for a sample to include all undergraduate students, for example. In fact, the chance an undergraduate student would be selected is

$$\frac{200}{25000} = 0.008 \quad \text{(or 0.8\%)}.$$

The chance a graduate student would be selected is

$$\frac{50}{5000} = 0.01 \quad \text{(or 1\%)}.$$

Therefore, undergraduates and graduates have different probabilities of inclusion. This compromise is made in stratified random samples to ensure fair representation.

Results: The following responses were observed:

Stratum	Sample size	Agreed	Sample proportion
Undergraduate	200	100	0.50 (50%)
Graduate	50	20	0.40 (40%)
Total	250	120	

Discussion: Suppose we want to estimate

$$p = \text{ population proportion of \underline{all} university students who agree.}$$

How should we do this? It wouldn't be fair to simply take the average of 50% (for undergraduates) and 40% (for graduates) because the stratum sizes are different. Undergraduate input should have more weight because their stratum is bigger. A better estimate would weight the responses by the stratum sizes, that is,

$$\frac{0.5(25000) + 0.4(5000)}{30000} = \frac{12500 + 2000}{30000} = \frac{14500}{30000} \approx 0.483 \quad \text{(or 48.3\%)}.$$

This would be an unbiased estimate of p from stratified random sampling. Calculating the margin of error of this estimate would be more difficult, but it would be lower than if a SRS had been used with 250 students. Why do you think this is?

Terminology: A **cluster random sample** arises in two steps:

1. Divide the sampling frame into clusters and take a SRS of clusters.
2. Sample every individual in each cluster selected.

Example 4.7. A researcher was interested in studying the eating habits of children in the fourth grade. One of the goals was to determine if BMI was related to participation in federal programs which provide free and reduced-price meals. She oversaw a study in Augusta, GA, which has 33 public elementary schools. A cluster random sampling design was used.

- A SRS of 6 elementary schools was chosen from the 33 schools in the city.
- Each fourth-grade student in the selected schools was invited to participate in the study.
- There were a total of 329 fourth grade children who participated.

What is the population? In this example, why would a cluster random sample be easier to observe than a SRS?

Discussion: Stratified and cluster random sampling designs seem similar because in both the sampling frame is divided into smaller groups. However, the two designs are different.

- The strata in a stratified random sample consist of individuals who are homogeneous in some way (e.g., sex, race, political affiliation, etc.). On the other hand, clusters in a cluster sample will probably include individuals from all strata.
- Stratified random samples are used to ensure overall representation. However, a secondary goal is often to compare the results between/among the strata; e.g.,
  - How do individuals of different sexes compare? Different races?
  - What is President Biden's approval rating among Republicans? Democrats? Independents?
- In a cluster random sample, on the other hand, there is usually no interest in comparing the clusters themselves. Clustering is used because it makes sampling easier, especially if carrying out the survey or study involves extensive observation (as in Example 4.7).
- Cluster random samples are common in polling individuals in a large metropolitan city, for example. City blocks or neighborhoods can act as clusters. Investigators then only need to visit the clusters that were randomly selected (instead of spreading them out over the whole city).

Terminology: A **systematic random sample** arises when individuals are selected at a pre-determined increment. For example,

- every 10th customer
- every 100th web site visitor
- every 5th student
- every 6th airline passenger (for "enhanced screening"), etc.

Example 4.8. Suppose I wanted to select a systematic random sample of students from this class using the following sampling frame (which I have).

Student	Code
Abherdine	1
Albert	2
Anderson	3
⋮	⋮
Yell	158
Zhang	159
Zynkowski	160

A systematic random sample could select the 5th student on the list, the 15th, the 25th, and so on. The last student sampled would be the 155th student on the list. This would give a systematic sample of 16 students. Every tenth student is sampled.

Discussion: Elaborate sample surveys may use more than one type of probability sample. For example, suppose we wanted a representative sample of all SC adults (4 million). We could accomplish this in stages:

- Stage 1: Treat the 46 counties in SC as clusters and take a SRS of counties (clusters could also be chosen systematically).
- Stage 2: Within each selected county, take a stratified random sample of individuals (e.g., by sex, race, political affiliation, etc.).

This design would involve using multiple probability sampling methods.

References:

Frenger, M., Pitsch, W., and Emrich, E. (2016). Sport-induced substance use−An empirical study to the extent within a German sports association. *PLOS One* **11**, e0165103.

Singer, J. (2005). Afraid to discuss evolution? *Chance* **18**, 29–31.

Williams, M. (1990). HIV seroprevalence among male IVDUs in Houston, Texas. *American Journal of Public Health* **80**, 1507–1509.

5 Experiments, Good and Bad

5.1 Introduction and examples

Important: Here is an excerpt from Moore and Notz (pp 88):

> *"Observational studies passively collect data. We observe, record, or measure, but we don't interfere. Experiments actively produce data. Experimenters actively intervene by imposing some treatment in order to see what happens."*

This passage explains succinctly the difference between observational studies and experiments. Here is the specific language we use with experiments:

- The individuals under study in an experiment are called **subjects**.
- A **treatment** is a specific experimental condition applied to the subjects (e.g., drug, method of instruction, etc.).
- A **response variable** is a variable that measures an outcome or result of a study.
- An **explanatory variable** is a variable that we think explains or causes changes in the response variable.

Example 5.1. *Does aspirin reduce the rate of heart attacks?* The Physicians' Health Study Research Group (1989) summarized the results of an experiment involving 22,071 male physicians (aged 40+). One of the goals was to investigate whether taking aspirin reduces the risk of heart attack. This experiment was performed in the 1980s and the aspirin component ended in 1995 (there was a second component we will discuss later).

- 11,037 physicians were assigned to take aspirin (325 mg every other day)
- 11,034 physicians were assigned to take a placebo.

Physicians were <u>randomly assigned</u> to either the aspirin or placebo group. The experiment was double-blinded.

- Subjects $\longrightarrow$ physicians
- Treatment $\longrightarrow$ aspirin/placebo
- Response variable $\longrightarrow$ heart attack? ($1 =$ Yes; $0 =$ No)
- Explanatory variables $\longrightarrow$ aspirin/placebo (treatment), age, smoking history, diabetes status, family history, cholesterol level, blood pressure, alcohol use, exercise frequency, BMI.

Results: Here were the results from the experiment:

- 139 out of 11,037 physicians (1.26%) receiving aspirin had a heart attack
- 239 out of 11,034 physicians (2.17%) receiving placebo had a heart attack.

These results, even when adjusting for the other explanatory variables, were determined to be **statistically significant**. In plain English, this means the difference between the treatment groups (aspirin/placebo) was so large that it was likely not explained by random chance.

Discussion: There were many limitations to this study.

- Researchers mailed invitation letters to all male physicians between 40 and 84 years of age who lived in the US and who were registered with the American Medical Association (in 1981). There were 261,248 letters mailed. The response rate was

  $$\frac{22,071}{261,248} \approx 0.084 \quad \text{(or 8.4\%)}.$$

  The sample had a mix of convenience and voluntary response aspects.
- What about younger male doctors? Doctors who were not registered with the AMA?
- The study did not include female doctors.
- What about the general public? Doctors probably lead healthier lifestyles (on average) than others.

Discussion: In Example 1.1, Cox et al. (2015) summarized a study where one of the research questions was:

- *Does treating premature infants with caffeine increase the chances of developing necrotizing enterocolitis?*

In an 18-month period during 2008-2009, there were 615 infants admitted to the neonatal intensive care unit at Palmetto Richland Hospital in Columbia, SC.

- 35 out of 137 patients (26%) receiving caffeine developed NEC
- 10 out of 478 patients (2%) not receiving caffeine developed NEC.

Was this an experiment? No. Infants were never "assigned" the caffeine treatment in advance. Investigators merely observed what happened by retrospectively examining electronic health records for patients admitted to the NICU. This was an observational study.

Example 5.2. *Sex education for adolescents.* Jemmott et al. (2010) summarize the results of an experiment performed with 662 black students (6th/7th grade) recruited from 4 public schools that serve low-income communities in the northeast US. The study took place during 2001-2002. Students were randomly assigned to one of five different instruction programs:

1. Abstinence only (8 hours)
2. Safe sex instruction (8 hours); discussed condom use, STIs, etc.
3. Comprehensive (abstinence and safe sex/STI discussions; 8 hours)
4. Comprehensive (abstinence and safe sex/STI discussions; 12 hours)
5. Health promotion (8 hours); discussed general health practices (control group).

Students were followed for 24 months and self-reported sexual activity during this time.

- Subjects ⟶ students
- Treatment ⟶ educational program (A, SS, C-8, C-12, Control)
- Response variable ⟶ sexual intercourse during follow-up period? (1 = Yes; 0 = No)
- Explanatory variables ⟶ educational program (treatment), sex, age, living arrangement, previous sexual history.

Results: Students in the abstinence only treatment group were least likely to participate in sexual intercourse in the 24-month follow-up period. The results were **statistically significant**. This means the results for the abstinence only group were so different than the other groups that this difference was likely not explained by random chance.

Limitations: There were many limitations to this study.

- Self-reported data can be inaccurate (especially on sensitive topics). However, it is not possible for the investigators to monitor the children and actively record sexual activity themselves.
- The sampling design was a convenience sample. The investigators made announcements in assemblies, classrooms, and lunchrooms (i.e., locations where students are "easiest to recruit").
- The study included black students only.
- The study included only students from low-income communities in one geographic region of the US.

Example 5.3. *Does withholding feed from pigs prior to slaughter reduce the incidence of salmonella?* An experiment in North Carolina was carried out to investigate this question. Pigs who were exposed to salmonella prior to slaughter were randomly assigned to one of three treatment groups:

1. Feed withheld 0 hours prior to slaughter
2. Feed withheld 12 hours prior to slaughter
3. Feed withheld 24 hours prior to slaughter.

After slaughter, each pig's cecum (part of the large intestine) was cut open and tested for salmonella. There were two competing theories on the impact feeding would have:

1. Slaughtermen may not lacerate the entrails of lighter pigs as often. Contamination of the cecum would be minimized and therefore the percentage of salmonella cases would be smaller for lighter pigs.
2. The stress from feed withdrawal on the pigs themselves may increase the excretion of salmonella by the pigs. Therefore, the percentage of salmonella cases would be larger for lighter pigs.

- Subjects $\longrightarrow$ pigs
- Treatment $\longrightarrow$ feeding withdrawal schedule (0 hours, 12 hours, 24 hours)
- Response variable $\longrightarrow$ salmonella detected in cecum? (1 = Yes; 0 = No)
- Explanatory variables $\longrightarrow$ feeding withdrawal schedule (treatment), sex, initial weight, different marketing groups.

Results: The percentages of cases where salmonella was detected were similar among the three treatment groups (60%, 64%, and 67%, respectively). These small differences were **not statistically significant**. In other words, the differences in the percentages among the three treatment groups could have arisen simply by random chance (and thus neither of the proposed theories was supported).

Limitations: There were some limitations to this study.

- The sampling design was a convenience sample. Investigators worked with a pig development nursing site to carry out their experiment and they sampled all of the pigs at that location.
- Therefore, it is hard to know whether the results would change if the investigators included pigs from other nurseries.
- The investigators excluded underweight and overweight pigs.

5.2 Randomized comparative experiments

Terminology: A **randomized comparative experiment** is an experiment that compares two or more treatment groups. Randomization is used to assign subjects to one of the treatment groups.

Terminology: Randomization is the use of impersonal chance to assign subjects to treatment groups.

- This produces groups of subjects that should be "similar on average" before we apply the treatments.

Outline

Sample subjects from population
↓
Randomize subjects to treatments
↓
Record data on each subject
↓
Analyze data
↓
Make statement about the differences between/among treatments.

- If investigators assigned subjects to treatment groups in a biased way, then this would destroy the experiment. We could not say if a treatment difference was due to the treatments or due to the biased assignment.

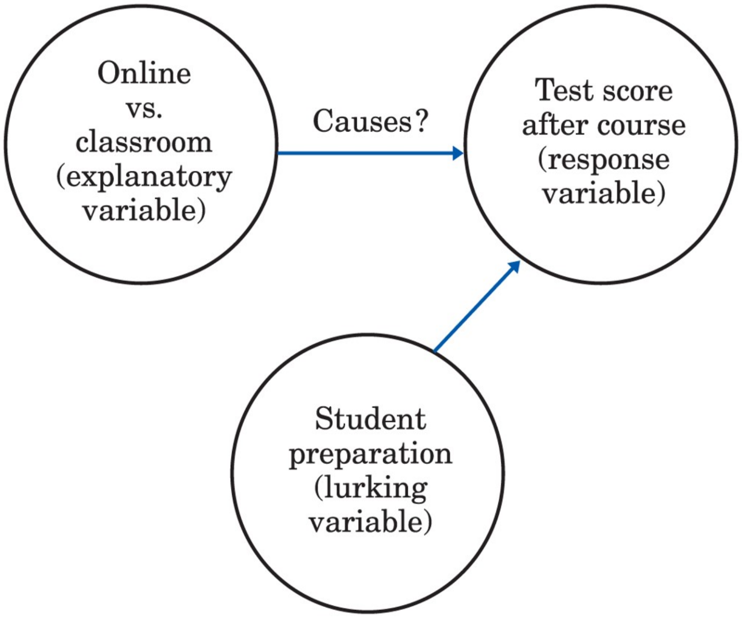
Remark: Examples 5.1-5.3 are examples of randomized comparative experiments.

- In Example 5.1, there are two treatment groups (aspirin, placebo).
- In Example 5.2, there are five treatment groups (A, SS, C-8, C-12, Control).
- In Example 5.3, there are three treatment groups (0 hrs, 12 hrs, 24 hrs)

In each experiment, subjects (i.e., physicians, students, and pigs, respectively) were randomly assigned to one of the treatment groups.

Example 5.4. Suppose we wanted to compare taking STAT 110 online and taking STAT 110 in this large lecture in-class format. In a randomized comparative experiment with 320 students, we could randomly assign 160 students to my online section. The remaining 160 students would take my in-class section.

- This experiment would be terribly difficult to implement in practice. Getting USC administration approval would be difficult, and recruiting 320 students to participate probably would be more difficult.


Online vs. classroom (explanatory variable)
Causes?
Test score after course (response variable)
Student preparation (lurking variable)

Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

- Randomizing students to the different course formats would make the two groups similar on average. This would be preferred.
- However, in real life, students sign up for the format that is best for them.
- If we did not use randomization, this would create the presence of a **lurking variable**: student preparation/background.
- Do the backgrounds of students who sign up for online and in-class sections differ systematically in some way? If we did not do a randomized comparative experiment, then any differences we saw in the treatment group outcomes (e.g., final exam score, etc.) could be attributed to this.
- In other words, the course format and student preparation/background explanatory variables are **confounded**.

Terminology: A **causal diagram** is a diagram which shows how different explanatory variables have a causal influence on the response variable.

Remark: Example 5.4 illustrates the value of **randomization**. When we randomly assign subjects to the different treatments, we create groups of subjects which are "similar on average." This mitigates the impact of lurking variables and allows us to assess causality more effectively.

> That is, does exposing subjects to different treatments **cause** a change in the response?

This is why randomized comparative experiments are considered to be the gold standard in statistics. In observational studies, it is much harder to "block out" the effects of lurking variables (more on this later).

Summary: Here is the logic behind randomized comparative experiments:

- Randomization produces groups of subjects that should be "similar on average" in all respects before we apply the treatments.
- Comparative design exposes the treatment groups to similar conditions other than the treatment they receive.
- Therefore, differences in the response variable (usually a group average or proportion) must be due to the effects of the treatments.

5.3 Practical considerations

Terminology: A **placebo** is a "dummy treatment" with no active effect on the response. In a comparative experiment, the **control group** may receive a placebo or an existing treatment (historical control). For example, to evaluate the effectiveness of a new drug or intervention, we may use a randomized comparative experiment with two groups:

- new drug/intervention versus placebo
- new drug/intervention versus an existing one.

A control group is used as a baseline for comparison.

Remark: The use of a control group helps to control the effects of lurking variables. For example, Moore and Notz (pp 91) describe a one-track experiment where patients with chronic migraine are given the drug fremanezumab. A **one-track experiment** uses only one treatment group. That is,

$$\text{All patients} \quad \longrightarrow \quad \text{fremanezumab.}$$

- If patients respond favorably to the drug (e.g., fewer headaches, less severity, etc.), we can't say if it is due to the drug or the due to a **placebo effect**. Some patients will respond favorably to any treatment or simply get better on their own.
- Therefore, if this effect is real, then we should include a second group (a control group) that receives a placebo. This will then balance out and make the two groups (fremanezumab/placebo) comparable.

Silberstein et al. (2017) summarize the results of a randomized comparative experiment with two doses of fremanezumab and a placebo. In other words, patients in the experiment were randomly assigned to one of three treatment groups:

1. Receive fremanezumab quarterly
2. Receive fremanezumab monthly
3. Receive placebo.

Patients in both the fremanezumab groups experienced a reduction in the number of headache days and migraine severity. These results were judged to be **statistically significant**. In other words, the differences between the two groups (fremanezumab quarterly versus placebo, fremanezumab monthly versus placebo) were so large that they probably were not due to random chance.

Terminology: We say an experiment is **blinded** if the subjects do not know which treatment they are receiving. An experiment is **double-blinded** if neither the investigators nor the subjects know which treatment they are receiving.

- Blinding and double-blinding are used to eliminate the effects of lurking variables.
- If the patients know which treatment they are receiving, then this might affect their willingness to participate in the experiment. Those in the placebo group may lose interest and drop out.
- Double-blinding is used to eliminate a biased assignment of subjects to the treatments. For example, a physician with a financial stake (e.g., research funding, etc.) may assign healthier patients to the new drug/intervention group with the hope they will show more favorable results.
- A randomized, double-blinded comparative experiment is "the holy grail" in medical research.

Discussion: The placebo effect can be real! Aseffi and Garry (2003) summarized the results of a randomized comparative experiment involving college students to study the effects of memory when consuming alcohol. In the experiment, 148 students were randomly assigned to two treatment groups:

1. Drinking vodka and tonic
2. Drinking tonic only.

However, the experiment did not use blinding. Students were informed what group they were assigned to.

- In reality, both groups were drinking tonic only!
- Afterwards, students were shown a sequence of slides depicting a crime. The results showed that participants who believed they were intoxicated were more "suggestible" and made "worse eyewitnesses" than those who thought they were sober. Moreover, students who thought they were drinking vodka even behaved drunk, displaying physical signs of intoxication.

Terminology: We say the differences observed in a study (an observational study or an experiment) are **statistically significant** if they are likely not due to random chance.

Illustration: Consider Example 5.4 which hypothetically compares two groups of students taking STAT 110 in different formats:

1. STAT 110 online
2. STAT 110 in-class lecture.

Suppose that, in reality, the two groups of students are similar in every way (making them comparable without lurking variables) and suppose the classes are taught in the same way except for class format.

Scenario 1: On-line section: 82% class average; In-class section: 83%.

- Is there a difference between the two groups? Yes.
- However, it is small and it could therefore have easily arisen "by random chance" (e.g., due to the effects of random assignment, inherent differences among students even if the groups are similar on average, etc.).
- Not statistically significant.

Scenario 2: On-line section: 67% class average; In-class section: 83%.

- Is there a difference between the two groups? Yes.
- The difference is quite large. It could have arisen by random chance but it is not likely. This large difference is more likely to have arisen from the difference between the class formats.
- Statistically significant.

Q: How do we determine if the results of a study are statistically significant or not?
A: This is what STAT 201, STAT 205, and STAT 206 are all about (as well as other statistics courses). Statistical methods like **confidence intervals** and **hypothesis tests** are used to determine statistical significance. We will touch on these topics only briefly at the end of this course.

Discussion: Moore and Notz (pp 97) pose the following questions:

- *Do children who are bullied suffer depression as adults?*
- *Do doctors discriminate against women suffering from heart disease?*
- *Are e-cigarettes safe?*

Unfortunately, we cannot use a randomized comparative experiment in instances such as these. We can not "assign" children to be bullied, subjects to be women, or students to smoke e-cigarettes. We can still perform comparative studies, but they will be observational like the NEC study in Example 1.1.

Terminology: An observational study is **retrospective** if it uses information on events that have taken place in the past. For example, we could

- ask adult patients if they were bullied as a child.
- ask women with heart disease if they have ever experienced discrimination on the basis of sex or gender.
- ask students if they have ever smoked e-cigarettes.

In each of these instances, we could still do a comparative study. For example, we could group adult patients retrospectively according to their bullying status, that is,

1. Experienced bullying as a child
2. Did not experience bullying as a child.

We could then record the percentage of depression cases for the two groups and compare them. However, if subjects were not randomized to the two groups, then there is no guarantee the groups will be similar to begin with. Therefore, any such comparison would be obscured by lurking variables which make the groups dissimilar.

- In other words, bullying status (bullied/not) will be **confounded** with all of the other explanatory variables which make the two groups dissimilar to begin with.
  - sex/gender; sexual orientation
  - genetic information; e.g., if parents had depression, etc.
  - experienced maltreatment as a child
  - experienced depression as a child
  - probably many, many others.
- We could determine that the difference in the percentage of depression cases between the two groups is statistically significant! *However, we have no way of knowing if this is due to the effect of bullying or if it is due to all of the other explanatory variables we could not control for.*
  - Remember, the beauty of **randomization** in comparative experiments is that it makes the treatment groups "similar on average" before the treatments are applied.
  - Therefore the effect of lurking variables is "balanced out" between/among all treatment groups.

- This is why so many observational studies (even if they are comparative in nature) cannot be replicated. The effects of the lurking variables may be too great. Therefore, if the study was performed again with a new sample of subjects, the results could change.

Reality: Assessing causality in observational studies is difficult! There are usually too many confounding variables. This is why randomized comparative experiments are preferred.

Q: If the goal is to establish a **causal relationship** between an explanatory variable (e.g., bullying status) and a response variable (e.g., depression/not), can an observational study still be useful? Can anything be done to remove or mitigate the effects of lurking variables? In other words, can observational studies be "salvaged?"
A: Yes, but the statistical methods needed to do so can be complex. **Causal inference** is an area of statistics dedicated to this question. There are two main approaches:

1. Regression adjustment
2. Matching explanatory variables.

Regression adjustment adjusts for the effects of lurking variables by including them in a regression model. **Matching** is used to make the groups "more similar" by intentionally including subjects in different groups whose explanatory variables match each other. Unfortunately, for either approach to be useful, investigators have to know beforehand all of the possible lurking variables! If any important ones are missed, then the groups being compared are still dissimilar in some systematic way. In this case, the sword of Damocles likely awaits despite the investigators' best efforts.

References:

Aseffi, S. and Garry, M. (2003). Absolut memory distortions: Alcohol placebos affect the misinformation effect. *Psychological Science* **14**, 77–80.

Jemmott et al. (2010). Efficacy of a theory-based abstinence-only intervention over 24 months. *Archives of Pediatrics and Adolescent Medicine* **164**, 152–159.

Morrow et al. (1999). Prevalence of *salmonella spp.* in the feces on farm and ceca at slaughter for a cohort of finishing pigs. *Proceedings of the 3rd International Symposium on the Epidemiology and Control of Salmonella in Pork*, 155–157.

Silberstein, S., Dodick, D., Bigal, M., Yeung, P., Goadsby, P., Blankenbiller, T., Grozinski-Wolff, M., Yang, R., Ma, Y., and Aycardi, E. (2017). Fremanezumab for the preventive treatment of chronic migraine. *New England Journal of Medicine* **377**, 2113–2122.

The Physicians' Health Study Research Group (1989). Final report on the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine* **321**, 129–135.

6 Experiments in the Real World

6.1 Introduction

Terminology: The three basic principles of experimental design are:

1. **Randomization:** the use of impersonal chance to assign subjects to treatment groups.
2. **Replication:** using enough subjects in each treatment group to reduce chance variation in the results.
3. **Control:** removing the effects of lurking variables and ensuring all subjects are treated similarly.

Here is the importance of each principle:

- Randomizing subjects to different treatment groups is our best hope to make the groups "similar on average." In other words, if there are lurking variables (e.g., sex/gender, race, etc.), then their effects will be shared equally by all treatment groups. This puts the groups on a level playing field *before* the treatments are administered.
- After the experiment is over and all the data are in, presumably we will want to make a statement about how the treatment groups compare. For example,
  - Does aspirin reduce the chance of heart attack when compared to placebo?
  - Which method of instruction is most effective at reducing the incidence of sexual intercourse during the follow-up period?
  - When should feed be withdrawn to reduce the transmission of salmonella during slaughter?

  In practice, answering these questions starts with an assessment of **statistical significance**. Are the differences we see between/among the groups so large that they would likely not be due to random chance? Using enough subjects (replication) helps us to make this type of assessment. The more subjects we use, the better our chances of finding differences between/among the groups when they truly exist.
- The third principle, control, is the most difficult to maintain especially when an experiment involves human subjects. As Moore and Notz put it, *"Treating subjects exactly alike is hard to do."*

Remark: We start with an example where things aren't all that hard. When the subjects are inanimate objects or even living things we can control easily (e.g., plants, rats, pigs, etc.), maintaining a high degree of control is easier.

Example 6.1. An engineer wants to compare the drying times of three brands of paint (Brands A, B, and C). She has 30 small wooden boards. Each board will be randomly assigned to one of the three brands of paint (10 boards per brand). The boards will then be painted using an automated system, and the drying time (in minutes) for each board will be recorded by the engineer.

- Subjects ⟶ boards
- Treatment ⟶ brands of paint
- Response variable ⟶ drying time (in minutes).

We could use R to perform the randomization. Label each board using $1, 2, ..., 30$.

```
> boards = seq(1,30,1)
> sample(boards,30,replace=F)
```

$$\underbrace{\texttt{19 29 5 26 3 10 23 13 17 9}}_{\text{assign to Brand A}} \quad \underbrace{\texttt{27 1 21 28 25 30 20 14 16 11}}_{\text{assign to Brand B}}$$

$$\underbrace{\texttt{22 2 18 7 15 6 12 8 24 4}}_{\text{assign to Brand C}}$$

The important thing is that we are using impersonal chance to assign subjects (boards) to treatments (brands of paint). We also see how replication is being used in this experiment. The engineer is using 10 boards per treatment group. This gives her more information about the brands of paint and their drying times than had she used 2 or 3 boards per group. Averaging a larger number of measurements per group reduces chance variation.

Q: What are possible lurking variables in this experiment? Are there any ways the engineer could introduce lurking variables herself?

6.2 Clinical trials

Definition: A **clinical trial** is a comparative experiment that studies the effectiveness of administering medical treatments to human subjects. A clinical trial is the clearest method of determining whether a new drug or intervention has a postulated effect. Clinical trials can generally be broken down into four phases:

- **Phase I.** Very small number of subjects. The goal is to establish the correct dosing amount and the correct dosing frequency.
- **Phase II.** Small number of subjects. The goal is to get preliminary information about the efficacy of the drug or intervention. Positive trials then move to the next phase.

- **Phase III.** Large number of patients. This is a large, carefully designed randomized comparative experiment evaluating the effects of the drug or intervention against placebo or the standard treatment. These experiments usually involve thousands of patients from all over the US (or the world).
- **Phase IV.** This is the last phase to determine if the new drug/intervention can be administered to the general public and which groups of patients would benefit the most. Side effects and safety are closely monitored.

The web sites `www.clinicaltrials.gov` and `www.centerwatch.com` summarize the results and findings from recent clinical trials. Here is an example:

Example 6.2. Scher et al. (2012) published results from a phase III trial of enzalutamide for the treatment of metastatic prostate cancer. This international, randomized, double-blind, placebo-controlled study enrolled 1,199 men who had been previously treated with chemotherapy.

- Subjects received enzalutamide 160mg daily (as four 40mg capsules) or placebo.
- The data showed enzalutamide exhibited a statistically significant benefit in overall survival compared to placebo.
  - Men treated with enzalutamide had a median survival of 18.4 months (95% confidence interval, >17.3 months) compared to 13.6 months (95% confidence interval, 11.3-15.8 months) for men treated with placebo.
- The drug was well tolerated. The most frequent adverse events were fatigue, diarrhea, and hot flush. Seizure was reported in less than 1% of patients treated with enzalutamide.

Remark: In the absence of a well-designed clinical trial, it is easy for anecdotal information about the benefit of a drug or therapy to become accepted. This can be serious and costly.

- In the 1970s, laetrile was rumored to be a “wonder drug” for cancer patients even though there was no evidence of biological activity.
- People were convinced there was a conspiracy in the medical community to keep the treatment from them.
- In 1982, a clinical trial with 175 patients (conducted at the Mayo Clinic) showed that nearly every patient experienced negative outcomes. Cancer tumors actually increased in size and some patients experienced cyanide poisoning.
- The National Institutes of Health evaluated the evidence separately and concluded that clinical trials of laetrile showed no effect against cancer.

More recently, during the pandemic, there were disputes about the effects of hydroxychloroquine and ivermectin as drug therapies to treat patients with covid-19. Reis et al. (2022) reported how over 60 clinical trials were performed to study the effects of ivermectin with no definitive conclusion reached. Some studies showed statistically significant results while others did not. The authors' study did not show a benefit from taking ivermectin in terms of reducing the percentage of hospital admissions. Here is a summary from their abstract:

> *"A total of 3515 patients were randomly assigned to receive ivermectin (679 patients), placebo (679), or another intervention (2157). Overall, 100 patients (14.7%) in the ivermectin group had a primary-outcome event, as compared with 111 (16.3%) in the placebo group (relative risk, 0.90; 95% Bayesian credible interval, 0.70 to 1.16)."*

In other words, was there a difference in the percentage of patients who were admitted to the hospital? Yes. However, this difference was not statistically significant; it could have arisen just by random chance.

Remark: Because clinical trials involve human subjects, there are important ethical issues that arise. For example, is it ethical to purposely withhold enzalutamide from patients suffering from advanced prostate cancer? Data and research ethics will be discussed in Chapter 7.

6.3 Common problems in experiments

Remark: Just as we saw with sample surveys and their potential problems (Chapter 4), there are many ways experiments can go horribly wrong. Most of these have to do with working with human subjects.

- If an experiment has any chance of providing insight on causality, then it is critical all subjects are treated alike except for the treatments.
- If subjects in different treatment groups are treated differently, this could introduce lurking variables which are confounded with the treatment. Blinding is a useful technique to avoid this when the experiment involves human subjects.
  - Blinding: the subject does not know which treatment s/he is receiving
  - Double blinding: neither the subject nor the investigator knows which treatment the subject is receiving.

Blinding can mitigate the **placebo effect**, a phenomenon which says subjects will respond to *any* treatment, even if it confers no biological or psychological effect. We already saw this in Aseffi and Garry (2003) where students displayed noticeable signs of intoxication despite drinking no alcohol (see Chapter 5). Moore and Notz (pp 113-114) give more examples:

- 42% of balding men maintained or increased their amount of hair when taking an innocuous substance designed to look like hair gel.
- 13/13 patients broke out in a rash after receiving a solution not containing poison ivy (they were told it did).

Main point: If there is a placebo effect in any experiment, then we want this effect to be the same in all treatment groups. Blinding therefore becomes critical. If subjects knew which treatment they were receiving, this might create imbalance among the groups. In this case, the placebo effect would be confounded with the effect of the treatment.

Note: Double blinding is used to remove possible bias caused by the investigators in charge of the experiment. If an investigator has a vested interest in a particular treatment, s/he may introduce bias by treating different subjects differently. For example, subjects taking the "real" treatment may get more attention or better care. This creates an unwanted lurking variable.

Remark: It might not be possible to blind subjects in an experiment. For example, in Example 5.4, students would obviously know if they are taking the online version of STAT 110 or my in-class version. Elsewhere, if an individual is assigned to a treatment group which takes a standard drug and an experimental radiation intervention in a cancer clinical trial (versus the other group which only receives the drug), then blinding may not be possible. Subjects would have to report to additional appointments where radiation is administered.

Terminology: Undercoverage can be a problem in experiments. In any experiment, subjects are first sampled from a larger population. Recall the outline for randomized comparative experiments:

Outline

Sample subjects from population
↓
Randomize subjects to treatments
↓
Record data on each subject
↓
Analyze data
↓
Make statement about the differences between/among treatments.

If the sampling design is biased (e.g., convenience, voluntary response, etc.), then we are systematically excluding certain individuals from participating. For example,

- a sociology experiment involves USC students but excludes students majoring in the hard sciences

- a clinical trial examines different nutrition methods for newborns but underrepresents mothers of certain races
  - see also Example 3 in Moore and Notz (pp 114-115) for a further discussion of minority underrepresentation in clinical trials
- an agricultural experiment compares different treatments for Ovine Johne's disease (in sheep) but includes only certain breeds.

Investigators want their conclusions to apply to an appropriate population of interest. If certain groups are not represented fairly, conclusions could be biased, or, at best, limited to only those groups studied.

Terminology: Non-adherence occurs when subjects do not follow the treatment regimen outlined by the investigator. For example, patients in a clinical trial might take their treatment at incorrect times or take an incorrect dose.

- Phase I clinical trials are designed to establish the correct dosing schedule and dosing amount. The success of the subsequent phases depends on this.

Some patients may not take the treatment at all or provide incorrect self-reported data (e.g., underreporting their pill count, etc.). This is non-adherence. Protocol inclusion criteria and pre-interviews can be useful in determining which patients are most likely to adhere to the specified treatment plan.

Terminology: Dropouts are subjects who begin the experiment but do not complete it. Especially when dealing with humans, some subjects may decide that they no longer want to participate in the study.

- Did they drop out for a reason? This may be due to undesirable side effects, lack of interest, or some other factor.
- Dropouts can cause serious bias if the reason is due to the treatment itself. If subjects drop out randomly (e.g., move away, etc.), then this may not cause large problems.

Terminology: Lack of realism occurs when we can not generalize the results of an experiment to a larger population. This can happen when an experiment is performed in very controlled settings that ultimately may not emulate real life.

- An agricultural experiment performed in a weather-controlled greenhouse setting may not adequately describe what happens on large farms.
- A psychology experiment subjecting students to irritating stimuli (e.g., to measure concentration levels, etc.); students know the experiment will be over soon.

Lack of realism is a common problem in clinical trials. The trials are usually performed on selected groups of individuals who first must meet strict inclusion criteria and then receive very careful attention for the duration of the trial. What about individuals who do not meet the inclusion criteria? If the FDA approves the new treatment, will the clinical benefit extend to "ordinary patients?"

Terminology: The **Hawthorne effect** is a change in subjects' behavior or outcomes not directly attributable to the treatment received but simply due to the "awareness of being in an experiment."

- The "Hawthorne experiments" were designed to study the effect of shop-floor lighting on worker productivity at a telephone parts factory in Chicago in the 1920s.
- However, researchers were perplexed to find that productivity improved, not just when the lighting was improved, but also when the lighting was diminished! In fact, productivity improved whenever changes were made in other variables such as working hours and rest breaks.
- The researchers concluded that the worker productivity was not being affected by the changes in working conditions, but rather by the fact that worker behavior changed due to an awareness of being watched more closely.

Some believe the Hawthorne effect is an intrinsic lurking variable in any experiment. However, this belief is not universally accepted.

6.4 Experimental designs

Terminology: In a comparative experiment which uses a **completely randomized design**, subjects are assigned to the treatment groups at random. In other words, each subject is treated as basically "the same in all regards" and then random chance decides which treatment group they will go to.

Example 6.3. An exercise science major wants to compare different training methods for powerlifters. He has recruited 60 lifters from the Columbia area. Each lifter will be assigned at random to one of the three training methods:

1. Bodybuilding training only (less weight; focuses more sustained muscle growth)
2. Powerlifting training only (high-impact heavy lifting only)
3. Bodybuilding and powerlifting training combined.

At the end of the training cycle, each lifter will perform their maximum lift on squat, bench press, and deadlift. The total weight will be recorded (combining all three lifts).

- Subjects $\longrightarrow$ lifters

- Treatment ⟶ training method
- Response variable ⟶ total weight among three maximum lifts.

If a completely randomized design is used, then each lifter is treated as being "the same in all regards" and is simply randomized to one of the three training methods.

Q: What are possible **lurking variables** here? That is, which variables might have an influence on the response variable but are not related to the training method?

- sex
- body weight
- years of experience lifting
- others?

Of course, the beauty of randomization is that the effects of all lurking variables should be "averaged out" among the three treatment groups. However, if we can identify obvious lurking variables beforehand, why not use a different experimental design which acknowledges them? Stay tuned!

Remark: In Example 6.3, the three treatment groups were created on the basis of a single variable: training method. In other completely randomized designs, the treatment groups can be created by cross-classifying two or more variables. The following example illustrates this.

Example 6.4. In the Physician's Health Study (Example 5.1, notes), we discussed how physicians were randomly assigned to take either aspirin or placebo. Although this is true, I only told you half the story. Another goal of the study was to determine if taking beta carotene reduces the incidence of cancer. Therefore, there were two variables used to form the treatment groups: aspirin use and beta carotene use. Here were the four groups:

1. Aspirin + beta carotene
2. Aspirin placebo + beta carotene
3. Aspirin + beta carotene placebo
4. Aspirin placebo + beta carotene placebo.

Each physician was assigned to one of these four treatment groups at random. That is, a completely randomized design was used. When the investigators wrote their seminal paper in 1989, the effects of beta carotene were not yet determined. However, the aspirin effect was so overwhelming that investigators could conclude statistical significance before the trial ended.

Terminology: In the language of experiments, a **block** is a group of subjects that are known (or believed) to be similar in some way. Blocks are created with the expectation that subjects in one block may respond differently to the treatment than subjects in another block.

Terminology: In a comparative experiment which uses a **randomized block design**, the subjects are first grouped into different blocks. Then, within each block, subjects are assigned to the treatments at random.

- This is called "restricted randomization." The investigator randomizes subjects to treatments within each block after they have been formed.

Importance: Blocking is a form of **control**. It is used to mitigate the effects of lurking variables.

- By separating subjects into different blocks and then randomizing subjects to treatments within each block, you are removing the effects of the blocking variable.
- Completely randomized designs do not do this. The effect of the blocking variable is merely "averaged over" by using randomization. One is then left to hope that the effect of the blocking variable is the same in each treatment group (not a guarantee).
- Therefore, blocking allows the researcher to compare the treatment groups more efficiently.

Example 6.3 (revisited). An obvious blocking variable in the powerlifting example is sex. Men generally lift heavier weight than women, so men may respond to the treatments differently than women. Therefore, if we were to "block on sex" beforehand, we would first separate the 60 lifters into two groups:

- Block 1: Female lifters
- Block 2: Male lifters.

In a randomized block design, the investigator would first randomize each female lifter to one of the three training methods (bodybuilding, powerlifting, combined). This would then be repeated for lifters in the male block. If the goal is to compare the training methods, this design removes the effects of sex as a potential lurking variable.

Example 6.5. In the adolescent sex education experiment (Example 5.2, notes), 662 6th/7th-grade students were assigned to one of five sex education instruction methods (A, SS, C-8, C-12, Control). However, a completely randomized design was not used. The investigators created four blocks by cross-classifying the students' grade and sex. Here were the blocks:

- Block 1: 6th grade/M

- Block 2: 6th grade/F
- Block 3: 7th grade/M
- Block 4: 7th grade/F.

Within each block separately, students were then randomly assigned to one of the five instruction methods (A, SS, C-8, C-12, Control). This was a randomized block design. If the goal is to compare the instruction methods (in terms of the incidence of sexual intercourse during the follow-up period), this design removes the effects of both grade and sex as potential lurking variables.

Example 6.6. Uhuri et al. (1998) considered whether using xylitol would reduce ear infection episodes in children at daycare centers. Children were recruited from 34 daycare centers in Oulu, Finland. There were a total of 857 healthy children who participated. A randomized block design was used. Here were the two blocks:

- Block 1: Children who could chew gum
- Block 2: Children who could not chew gum.

Chewing gum status is obviously closely related to age (i.e., younger children may not be able to chew gum yet). Within each block, children were randomly assigned to receive xylitol or placebo.

- In Block 1, xylitol was administered in gum form. In Block 2, xylitol was administered in syrup form.

Children were followed for a 3-month period. After the experiment ended, the authors concluded,

> *"The occurrence of [ear infection episodes] during the follow-up period was significantly lower in those who received xylitol syrup or gum, and these children required antimicrobials less often than did controls."*

There were some potential limitations in this study. There were more dropouts recorded for the xylitol treatment group than for the control group in both blocks. This might temper the authors' conclusions (especially if children dropped out for reasons due to taking xylitol). The form of the medication (gum/syrup) may also be a lurking variable. Not all children are being treated equally.

Example 6.7. An ophthalmologist would like to assess the clinical effect of a new eye drop for pink eye. She recruits 20 patients who have pink eye in both eyes. For each patient,

- the new eye drop is applied to one eye

- the standard eye drop is applied to the other eye
- which drop goes in which eye is determined at random for each patient (e.g., by flipping a fair coin, etc.).

The response variable (measured on each eye) is the time until the infection is eradicated.

Q: What is different about this experimental design?
A: Each subject receives both treatments.

Terminology: A **matched pairs design** compares two treatments. Each subject in the experiment receives both treatments.

Remark: Matched pairs experiments are common in studies where the subjects have "natural pairs" which can be observed; e.g., eyes, ears, arms, etc. They are also common in twins studies (especially involving identical twins), where one twin is exposed to one experimental condition and the other twin is exposed to the other condition.

Discussion: Matched pairs experiments are useful because they remove the effects of lurking variables which make different subjects different. Because each subject receives both treatments, the investigator is allowed to compare the treatments under perfectly homogeneous conditions.

- In our pink eye example, suppose Joe takes the new eye drop and Jim takes the standard eye drop.
- Do we get to see the response variable (time to infection eradication) for each treatment? Yes. One on Joe, and one on Jim.
- However, these measurements also incorporate every biological variable that makes Joe and Jim different people. In other words, the difference between the two responses contains
  - information on how the treatments differ
  - additional variation that exists because Joe and Jim aren't the same person.

  If the latter source of variability is large, this could hinder our ability to assess how the treatments compare.
- Matched pairs designs eliminate this extra source of variability because treatment comparisons are made on the same person.

Example 6.7. *Pre-test/post-test studies.* An accounting student wants to examine the effectiveness of a new accounting training course taught to undergraduate students. She has 100 students who will take the training course.

- Each student will take a **pre-test** to measure initial knowledge about the subject matter.

- Each student will take the training course.
- Each student will take a **post-test** (after completing the course) to measure knowledge about the subject matter.

The difference between the pre- and post-test scores serves as a measure of training effectiveness. Each student is measured under two conditions: before the training course is completed and after. Therefore, this is a matched pairs design.

- It is important to keep track of which student is which so you can "pair up" the scores; i.e., to "match" the post-test score with the correct pre-test score.
- If this is not done, then you cannot analyze the data correctly−there is no way of knowing where the "pairs" are.

Discussion: In a sense, matched pairs experiments are a special type of randomized block design. Each subject serves as its own block. Remember, blocking is a form of control, that is, to control the effects of lurking variables which may be confounded with the treatment. Matched pairs experiments not only control the effects of lurking variables, they remove them completely.

Remark: We have discussed matched pairs experiments in the context where each subject receives both treatments. However, "matching" or "finding pairs" of subjects could be done with different subjects. As Moore and Notz (pp 120) put it,

> *"Choose pairs of subjects that are as closely matched as possible. Assign one of the treatments to each subject in a pair by tossing a coin...."*

In our pink eye example, while Joe and Jim are different subjects, suppose they are the same age, the same weight, have the same dietary habits, have nearly identical health backgrounds, have the same eye characteristics, etc. Could we now think of Joe and Jim as a "pair," where Joe receives one eye drop and Jim receives the other? We haven't removed all lurking variables (Joe and Jim are still different people), but maybe most of the important ones have been removed by **matching**. This would certainly be better than using a completely randomized design, where subjects are assigned to the treatment groups randomly without any matching at all.

References:

Reis, G., Silva, E., Silva, D., et al. (2022). Effect of early treatment with ivermectin among patients with covid-19. *New England Journal of Medicine* **386**, 1721–1731.

Scher, H., Fizazi, K., Saad, F., et al. (2012). Increased survival with enzalutamide in prostate cancer after chemotherapy. *New England Journal of Medicine* **367**, 1187–1197.

Uhuri et al. (1998). A novel use of xylitol sugar in preventing acute otitis media. *Pediatrics* **102**, 879–884.

7 Data Ethics

7.1 Introduction

Discussion: We now live in a time where data are everywhere. It used to be that data were exclusively numerical values, like birth weights and IQ scores, or categorical values (e.g., 1 = had heart attack; 0 = not), and statistical science progressed in the last century by developing appropriate methods to analyze these types of data. Today, "data" can take on many new forms, for example,

- texts and social media posts,
- digital images, video, and audio,
- geospatial maps,
- notes in electronic health records,

among other forms that are constantly emerging in the "data science age." Because all fields use data in some form or another, researchers must know how to use them. In fact, convincing data are required for researchers to substantiate or to confirm their ideas. If the data refute their ideas, then they run the risk of being less productive and perhaps even marginalized by their peers.

From a researcher's perspective, often the ultimate goal of a study (e.g., survey, observational study, experiment, etc.) is to demonstrate that a **research hypothesis** is supported by the data and the resulting statistical analysis. Here are some examples:

- "Administering caffeine to premature infants increases the risk of NEC."
- "American voters' opinions on gay marriage are changing."
- "Abstinence-only sex education is preferred for at-risk youths."
- "Genes PLEKHA8, PRR25, FBXL13, VPS54, SLFN5, SNCAIP, and TGM1 are associated with autism development."
- "Excessive use of smart phones leads to insomnia and other sleep disorders."
- "HIV patients with dental insurance provided by Medicaid are more likely to have unmet dental needs than those with private insurance."
- "Common core math provides high school students with the skills they need to be successful in the workforce and in college."
- "Whites who live in the South are in favor of harsher penalties for Hispanic and black criminals."

- "Eating more cereal increases the chances of having a male child."

Obtaining empirical evidence for one's research hypothesis is usually mandatory for publishing, which serves as a measuring stick for success and accomplishment in academics and paves the way for future success. Perhaps because of this, there are some researchers who violate basic ethical principles when carrying out a research experiment or observational study. Some examples of this include

- cheating on the use of randomization during sampling or treatment assignment
- retaining only those data which are favorable to one's hypothesis
- faking the data all together.

This chapter deals with data ethics. I identify three individuals who committed blatant academic fraud by manipulating and fabricating data. Interestingly, Moore and Notz (pp 136) remark,

> *"Neither will discuss those few researchers who, in the pursuit of professional advancement, publish fake data. There is no ethical question here–faking data to advance your career is just wrong."*

Although I understand their reluctance to discuss such examples, I believe it is important (for you) to see and expose academic fraud when it is obvious.

7.2 Unethical behavior with data: Three examples

Example 7.1. *The Duke-Potti cancer research scandal.* Dr. Anil Potti's cancer research at Duke University in the mid-2000s was viewed to be revolutionary. His research was based on the theory that the best chemotherapy drug treatment could be uniquely matched to an individual tumor's DNA. Therefore, instead of viewing cancer patients as one group of individuals and developing general treatment strategies, an individual's treatment plan could be targeted specifically to that individual's cancer. This is the idea behind "personalized medicine."

- Initial data, results, and conclusions supported Potti's theory and were published in top medical journals. Of course, funding and accolades started to pour in to Potti, his mentor, and Duke University.
- Drs. Kevin Coombs and Keith Baggerly, biostatisticians at MD Anderson Cancer Center in Houston, found basic problems with the data used to make Potti's conclusions. An anonymous source pointed out a misrepresentation of Potti's academic credentials on federal grant applications.

- It was ultimately determined the data collected by Potti and his research colleagues had been manipulated. Patient data supporting Potti's groundbreaking approach were retained. Those that did not were altered to support the theory.
- Patients were assigned cancer treatments that had no effect or made things worse. Potti "resigned" from Duke, his research career destroyed.

Example 7.2. *The LaCour-gay marriage debacle.* An aspiring academician in political science, PhD student Michael LaCour (UCLA) undertook a large experimental study under the direction of Dr. Donald Green (a political scientist at Columbia). The study involved political canvassing of registered voters in California, trying to persuade voters to change their opinions of gay marriage. This study was done before gay marriage became legal in California and elsewhere.

- The hypothesis proposed by LaCour was that gay canvassers could have more impact on changing the opinions of voters than straight canvassers.
- The data LaCour "collected" demonstrated this hypothesis was supported overwhelmingly. LaCour and Green (2014) appeared in *Science*. News outlets discussed the findings extensively; LaCour was offered an academic position at Princeton.
- A few months after publication, there were questions raised about who funded the study (no sources were provided). The whistle blower was another graduate student at UCLA who had discussed the work with LaCour.
- The whistle blower also demonstrated statistically that Lacour's published results matched a theoretical model "a little too perfectly." It was suggested Lacour's data had been completely fabricated. Also, the reported response rates were "extraordinary."
- LaCour could not successfully defend his data sets nor did he provide them for others to see; he said they had been "destroyed." This prevented anyone from being able to reproduce his findings.
- In the light of these anomalies, Green asked *Science* to retract the publication. Princeton soon after rescinded its offer of employment to LaCour.

Example 7.3. *The Stewart retractions.* In 2023, Dr. Eric Stewart was fired from Florida State University for fabricating data in a number of high-impact publications. Stewart's work focused on race relations in criminal justice. He was an established researcher who received millions of dollars in federal grants.

- One of Stewart's PhD students, Justin Pickett was responsible for bringing the allegations to light. He started to notice problems with the reporting of data in Stewart's existing work.

- Pickett pointed out one of their joint articles claimed the study was based on 1,184 respondents, but that there were actually only 500. It also hand-picked data from 91 counties as opposed to including the full 326 studied. Upon discovery, Pickett and the other co-authors immediately asked for the article to be retracted.

- The response rates from sample surveys cited in Stewart's existing work were unrealistic, often in excess of 60%. Remember the Pew response rate (Chapter 4) of only 8%? In one instance, Stewart said his "friends from graduate school" helped him performed the surveys.

- Pickett (2020) performed a statistical analysis which demonstrated decimal place digits reported in Stewart's other articles violated "Benford's Law," an established probability law for how unaltered digits 0-9 should appear. The conclusion was the decimal place digits had to have been changed.

- One of the reasons cited in Dr. Stewart's termination letter was "extreme negligence in basic data management with the presence of no backup of the data." Stewart never provided coauthors with the data which could reproduce the articles' analyses.

7.3 Basic principles

Reality: Many observational studies and experiments involve human subjects. Therefore, a host of ethical issues need to be considered before a study or experiment can be performed (or even approved).

Terminology: The organization that carries out the study must have an **institutional review board** (**IRB**) that reviews the study and gives its approval. An IRB's primary mission is to protect subjects from possible harm.

- In 1974, the US Congress established the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research to develop guidelines for human subjects research (as part of the National Research Act).
  - The Act required the establishment of an IRB for all research funded by the federal government.
  - For clinical trials, these were later modified to require IRB approval for all drugs or products regulated by the Food and Drug Administration.

- IRBs must have multiple members with expertise relevant to safeguarding the rights and welfare of the subjects.
  - At least one member should be a scientist, one a non-scientist, and at least one must be unaffiliated with the institution/organization. Expertise in bioethics is also included.

  - An IRB should be made up of individuals with diverse racial, gender, and cultural backgrounds.

- IRBs approve human research studies that meet specific prerequisites.
  - The risks to the study participants are minimized.
  - The selection of study participants is equitable (when appropriate).
  - Informed consent is obtained and documented for each participant.
  - The privacy of the participants and confidentiality of the data are protected.

- Federal grant agencies like NSF and NIH require IRB approval before research grants get funded. Publications must identify the IRB who approved the study. For example, in our NEC paper, we write, "This study was approved by the Palmetto Health Institutional Review Board."

Terminology: The term **informed consent** means that subjects must agree in advance to being included in the study.

- Subjects must be told the purpose of the study and what the possible risks are.
- This itself can present ethical questions:
  - Who can give informed consent? What if the study involves young children? mentally challenged subjects? patients in a coma?
  - Are there some instances where the informed consent requirement can be "overlooked?"

Example 7.4. In Moore and Notz (pp 147-148), your authors describe a criminiology experiment involving domestic abuse calls answered by police. At the scene, police officers randomize each offender to one of two groups with different actions:

1. Arrest the offender and hold overnight in jail.
2. Warn the offender and release.

The researchers then visit the victim later to inquire about incidents of recidivism (i.e., if the offender has re-offended).

Q: Did the offenders give informed consent to be included in this study?

Definition: Confidentiality is the protection of information provided by subjects and the assurance that information about individual respondents cannot be derived from the statistics reported.

- Researchers may report averages or percentages for groups (perhaps one you are in), but they can not reveal individual information about you; e.g., your income, your answers to survey questions, etc.

Definition: Anonymity means that subjects are anonymous. Subjects' identities are not known to any researcher involved with the study.

- The randomized response survey method (Chapter 4) is an example where an individual's responses are anonymous. However, the researchers will still probably have identifying information on the subjects.
- Complete anonymity is a rarity in statistical studies. An exception might be online polls, but those are often useless.

7.4 Ethics in clinical trials

Discussion: A clinical trial involves giving treatments to human subjects. In most trials, patients are assigned to take a new treatment or some type of control treatment (e.g., placebo, standard treatment, etc.). Are clinical trials ethical?

- Although a new treatment could benefit the patient, there are also potential risks associated with it. Patients are therefore subjected to potential harm.
- If a new treatment is suspected to be better, can we justify ethically assigning some patients *not* to get it?

Here is a quote from Dr. Charles Hennekens, director of the Physician's Health Study (Example 5.1):

> *"There's a delicate balance between when to do or not do a randomized trial. On the one hand, there must be sufficient belief in the agent's potential to justify exposing half the subjects to it. On the other hand, there must be sufficient doubt about its efficacy to justify withholding it from the other half of subjects who might be assigned to placebos."*

Terminology: There should be genuine uncertainty about which treatment might be superior for each individual patient. This is known as **equipoise**.

- This means that no known superior alternative treatment is available for each patient.
- Equipoise provides the principled basis for medical research involving patients randomly assigned to different treatments.
- It is accepted that clinical trials are ethical in this setting of uncertainty.

Terminology: Because of the amount invested in clinical trials and ethical concerns, large Phase III clinical trials are monitored by an independent committee called a **Data Safety Monitoring Board** (**DSMB**). This group meets periodically throughout the trial and makes recommendations on whether the trial should be modified or stopped.

- The overall responsibility of the DSMB is to ensure the safety and well-being of patients enrolled in the trial.
- The DSMB will generally have members with diverse backgrounds including clinicians, epidemiologists, biostatisticians, data managers, and ethicists.
- DSMB members should have no conflict of interest with the studies they are monitoring; e.g., no financial holdings, no family member conflicts, etc. All discussions of the DSMB are confidential.
- The specific responsibilities of the DSMB include
  - protocol/manuscript review
  - interim reviews (study progress, quality and safety, efficacy and benefit)
  - early administrative analyses (recruitment/entry criteria, design assumptions, quality and timeliness of data collection)
  - design modifications (entry criteria, treatment dose, sample size adjustments).

Example 7.5. The Untreated Syphilis Study at Tuskegee grossly violated the equipoise principle among other ethical guidelines. This study started in 1932 in Macon County, Alabama and involved 600 black men: 399 with syphilis and 201 who did not have the disease. The purpose of the study was to see how the disease progressed in black men.

- Participants' informed consent was not collected. Researchers told the men they were being treated for "bad blood," a local term used to describe several ailments, including syphilis, anemia, and fatigue. In exchange for taking part in the study, the men received free medical exams, free meals, and burial insurance.
- By 1943, penicillin was the treatment of choice for syphilis and becoming widely available, but the participants in the study were not offered treatment.
- In 1972, a whistle-blower and an AP news story led to creating a panel to review the study. The study was ended.
- In 1974, the National Research Act was passed by the US Congress.
- In 1976, the US government agreed to a $10 million settlement, lifetime medical benefits, and burial services to all living participants. These benefits were later extended to the participants' wives and children.

References:

Lacour, M. and Green, D. (2014). When contact changes minds: An experiment on transmission of support for gay equality. *Science* **346**, 1366–1369. This article was retracted on May 28, 2015.

Pickett, J. (2020). The Stewart retractions: A quantitative and qualitative analysis. *Econ Journal Watch* **17**, 152–190.

8 Measuring

8.1 Introduction

Discussion: Statistics is the science that deals with data. As such, we need to know where data come from. To set our ideas, let's reconsider the NEC/caffeine study in Example 1.1 (notes). There were many variables recorded for each infant in the study; here are five of them:

- Gestational age (measured in weeks)
- Birth weight (measured in grams)
- Time to discharge (measured in days)
- Nutrition type (Breastmilk, Fluids, Formula, TPN)
- NEC (1 = Yes, 0 = No).

Here is what part of the NEC data set looks like (copied from my Excel file):

Infant	Gestational age	Birth weight	Time to discharge	Nutrition type	NEC
1	34	2402	25	Breastmilk	0
2	25	755	107	Fluids	1
3	33	2059	33	Formula	0
⋮	⋮	⋮	⋮	⋮	⋮
615	31	1401	72	Fluids	0

Terminology: We **measure** a variable on an individual (subject) when we assign a number or category to the variable. We use an **instrument** to make a measurement. Quantitative variables have measurements with **units** attached to them. The measurements of all variables are called **data**.

Q: What instruments were used to measure the five variables above?

- Gestational age $\longrightarrow$ ultrasound
- Birth weight $\longrightarrow$ scale
- Time to discharge $\longrightarrow$ calendar
- Nutrition type $\longrightarrow$ determined by attending obstetrician
- NEC $\longrightarrow$ x-ray (reveals if there is an irregular gas pattern in the intestine).

Discussion: For some variables, different instruments could be used. For example, take gestational age. Measuring this variable exactly involves a determination of when conception took place. Ultrasound is widely regarded as the most **accurate** and **reliable** instrument to do this. However, other instruments could be used:

- self-reported menstrual/intercourse history
- clinical observation (after birth); e.g, by measuring head circumference, birth weight, and other physical characteristics.

These instruments are generally regarded to be biased (inaccurate) and unreliable.

8.2 Measurement error and misclassification

Terminology: Measuring quantitative variables perfectly is not always possible. There are two sources of **measurement error**:

- Bias
- Random error.

We can always think of the measurement of a quantitative variable as follows:

$$\text{Measured value} = \text{True value} + \text{Bias} + \text{Random error.}$$

- We would like to know the true value of the variable for an individual. This isn't always attainable.
- A measurement process has **bias** if it systematically overstates or understates the true value.
  - Bias usually arises from the instrument used to make the measurement.
- A measurement process has **random error** if repeated measurements on the same individual give different results.
  - If random error is small, we say the measurement process is **reliable**.
  - Perfect reliability arises when the random error component is 0; i.e., repeated measurements on the same individual are exactly the same.

Example 8.1. Do NOT read this example ahead of time (i.e., before we do it in class). See next page.

Instructions: In 30 seconds, read the following passage and count the number of "F's."

> "THE NECESSITY OF TRAINING FARM HANDS FOR FIRST CLASS FARMS IN THE FATHERLY HANDLING OF FARM LIVESTOCK IS FOREMOST IN THE MINDS OF EFFECTIVE FARM OWNERS. SINCE THE FOREFATHERS OF THE FARM OWNERS TRAINED THE FARM HANDS FOR FIRST CLASS FARMS IN THE FATHERLY HANDLING OF FARM LIVESTOCK, THE FARM OWNERS FEEL THEY SHOULD CARRY ON WITH THE FORMER FAMILY TRADITION OF TRAINING FARMHANDS OF FIRST CLASS FARMS IN THE EFFECTIVE FATHERLY HANDLING OF FARM LIVE STOCK, HOWEVER FUTILE, BECAUSE OF THEIR BELIEF THAT IT FORMS THE BASIS OF EFFECTIVE FARM MANAGEMENT EFFORTS."

Number of F's on 1st reading: _____
Number of F's on 2nd reading: _____
Number of F's on 3rd reading: _____

- Did your answers systematically underestimate or overestimate the true number of F's? This is bias.

- Did your answers change from reading to reading? This is random error.

Example 8.2. Using an electronic sphygmomanometer at home while at rest, I measured my systolic blood pressure (SBP) ten times. Here were the measurements (units = mm Hg):

$$131 \quad 127 \quad 128 \quad 132 \quad 129 \quad 132 \quad 134 \quad 129 \quad 128 \quad 130$$

What are possible sources of bias in this measurement process? Is there random error? Remember the general idea:

$$\text{Measured value} = \text{True value} + \text{Bias} + \text{Random error}.$$

Consider the three hypothetical scenarios:

1. My true SBP is 130. In this case, there does not appear to be much (any?) bias. All the measurements hover around this value (5 below, 5 at or above).

2. My true SBP is 120 ("normal"). In this case, the measurement process is **positively biased**. It is systematically reporting SBP measurements that are larger than this.

3. My true SBP is 140. In this case, the measurement process is **negatively biased**. It is systematically reporting SBP measurements that are smaller than this.

Q: How do we quantify reliability?
A: We can calculate the **variance** of repeated measurements on the same individual. We can do this "by hand" or we can use R.

Remark: We will discuss the variance in more detail in Chapter 11. For now, here are the steps to calculate it by hand in Example 8.2 and elsewhere.

1. Find the average of all the measurements.
2. Take each measurement and subtract the average. Square the difference.
3. Add up the squared differences in Step 2.
4. Divide the sum in Step 3 by $n-1$ (i.e., one less than the number of measurements).

Illustration: Let's calculate the variance of the SBP measurements in Example 8.2:

Step 1: The average is

$$\frac{131+127+128+132+129+132+134+129+128+130}{10} = \frac{1300}{10} = 130$$

Step 2: Calculate each difference and square it:

$$\begin{array}{lcc}
131-130 = 1 & \Longrightarrow & 1^2 = 1 \\
127-130 = -3 & \Longrightarrow & (-3)^2 = 9 \\
128-130 = -2 & \Longrightarrow & (-2)^2 = 4 \\
132-130 = 2 & \Longrightarrow & 2^2 = 4 \\
129-130 = -1 & \Longrightarrow & (-1)^2 = 1 \\
132-130 = 2 & \Longrightarrow & 2^2 = 4 \\
134-130 = 4 & \Longrightarrow & 4^2 = 16 \\
129-130 = -1 & \Longrightarrow & (-1)^2 = 1 \\
128-130 = -2 & \Longrightarrow & (-2)^2 = 4 \\
130-130 = 0 & \Longrightarrow & 0^2 = 0
\end{array}$$

Step 3: Add up the squared differences:

$$1+9+4+4+1+4+16+1+4+0 = 44$$

Step 4: Divide the sum by one less than the number of measurements:

$$\textbf{variance} = \frac{44}{9} \approx 4.89.$$

Note: Instead of performing hand calculation (like above), you can use R to calculate this quickly:

```
> sbp = c(131,127,128,132,129,132,134,129,128,130) # enter the measurements
> mean(sbp) # calculates the average in Step 1
[1] 130
> var(sbp) # calculates the variance
[1] 4.888889
```

Interpretation: A measurement process that has small variance is said to be **reliable**.

- The smaller the variance, the more reliable the measurement process is.
- A measurement process is **perfectly reliable** if there is no random error. This happens when repeated measurements on the same individual are all the same.

To illustrate the last point, suppose my 10 SBP measurements were

130 130 130 130 130 130 130 130 130 130

In this case, the variance would be 0. There is no random error.

Discussion: Can a quantitative variable like SBP be measured perfectly? Yes, but only when there is no bias and no random error. Remember the general idea:

$$\text{Measured value} = \text{True value} + \text{Bias} + \text{Random error}.$$

- Removing **bias** can be difficult. Bias usually arises from the instrument used to make the measurement. Removing bias might mean a different instrument is needed.
- Removing **random error** (i.e., attaining perfect reliability) can also be difficult. One way to *reduce* random error is to average several measurements on the same individual. For example, in Example 8.2, the average was

  $$\frac{131+127+128+132+129+132+134+129+128+130}{10} = \frac{1300}{10} = 130.$$

  Reporting the average 130 as my SBP would be more reliable than reporting 10 separate measurements for SBP. "Averaging reduces variation" is a well-known statistical principle.

Terminology: We say that a categorical variable measurement is **misclassified** if the wrong category has been given for an individual.

Example 8.3. The State Hygienic Laboratory at University of Iowa tests thousands of residents each year for chlamydia. Here are some of the variables they record on each individual:

- Sex (M/F)
- Whether contact with an STD infected partner occurred in the last 3 months: 1 = yes; 0 = no
- Presence of symptoms (e.g., painful ejaculation, painful urination, painful menstruation, etc.): 1 = yes; 0 = no
- Chlamydia diagnosis: 1 = positive; 0 = negative.

Among the variables in this example, contact with an STD infected partner and the chlamydia diagnosis could be misclassified.

- Contact: Measurements of this variable are **self-reported**. Therefore, individuals could lie about sexual contacts if they are embarrassed. Also, the individual reporting may not be aware of the STD status of their partner(s). They could have also lost track of what happened in the three-month time frame.
- The SHL uses the Aptima Combo II Assay to perform the chlamydia test. This test is not perfect, so some diagnoses will be misclassified.
  - An individual could be positive, but the test is negative: **False negative**
  - An individual could be negative, but the test is positive: **False positive**.

8.3 Counts versus rates

Example 8.4. The following table lists the number of sports-related injuries treated in US hospital emergency rooms in 2001, along with an estimate of the number of participants (in thousands) in the sports:

Sport	# Injuries	# Participants	Rate
Basketball	646,678	26,000	24.7
Bicycle riding	600,649	54,000	11.1
Base/softball	459,542	36,000	12.7
Football	453,684	13,000	34.1
Soccer	150,449	10,000	15.0
Swimming	130,362	66,200	2.0
Volleyball	129,839	22,600	5.7
Roller skating	113,150	26,500	4.3
Weightlifting	86,398	39,200	2.2
Fishing	84,115	47,000	1.8
Horse riding	71,490	10,100	7.1
Skateboarding	56,435	8,000	7.1
Ice hockey	54,601	1,800	30.3
Golf	38,626	24,700	1.6
Tennis	29,936	16,700	1.8
Ice skating	29,047	7,900	3.7
Water skiing	26,633	9,000	3.0
Bowling	25,417	40,400	0.6

Q: How should different sports be compared in terms of ER injuries?

A: If we use the number of injuries (a count), then we would conclude that

- riding a bicycle is more dangerous than football? fishing is more dangerous than ice hockey?

Note: Here is how injury rate (Rate) was computed:

$$\text{Rate} = \frac{\text{\# Injuries}}{\text{\# Participants}}.$$

Because the number of participants is recorded in thousands, the injury rate is the number of injuries per 1000 participants. For example,

- The injury rate for bicycle riding is 11.1 injuries per 1000 participants.
- The injury rate for football is 34.1 injuries per 1000 participants.

Dividing by the number of participants (i.e., calculating a **rate**) puts these figures on a common scale, thereby facilitating a fairer comparison.

Discussion: Using counts instead of rates can be very misleading when comparing two or more groups, for example,

- "Bicycle riding results in more ER visits than football."
- "Professor Tebbs gave the fewest number of A's among all STAT 110 professors."
- "The number of annual firearm deaths in Texas (2823) is larger than in Vermont (69)."
    - These counts are based on 2014 data from the National Center for Health Statistics (part of the CDC).
    - The rate of annual firearm deaths per 100,000 inhabitants, however, is very close: 10.6/100,000 (Texas) versus 10.3/100,000 (Vermont).
    - Using a rate paints a clearer picture of how the two states compare.

Example 8.5. Then-candidate Donald Trump, during the Republican presidential primary in 2016, made the claim,

"We got the highest vote count in the history of the Republican Party."

This statement is true. The previous record was held by President George W. Bush:

- 2000: President Bush: 11.5 million votes
- 2015: President Trump: 13.3 million votes.

However, this statement is also misleading for the following reasons:

1. The population of the US in 2000 was 282 million. In 2015, it was 325 million! That is 43 million more residents in 2015, many of whom are eligible voters.
2. President Trump also holds the dubious record of having the largest number of votes *against* him during a Republican presidential primary, barely eclipsing John McCain in 2008.

Main point: When comparing two or more groups of individuals, it is safer to use a rate or a percentage. Comparing counts can be misleading because the size of the groups being compared can be very different.

8.4 Measurement validity

Remark: Some variables correspond to physical properties that are clear-cut, and the variables themselves are easy to measure:

- Birth weight (in grams)
- Whether a physician has a heart attack (1 = yes; 0 = no)
- Monthly rainfall at CAE (in inches)
- Maximum hurricane wind speed (in mph)
- Distance travelled (in miles).

Some variables are more ambiguous and can be more difficult to measure:

- intelligence
- college readiness
- human personality
- quality of a musical performance.

Definition: A measurement is said to be **valid** if it accurately represents the intended variable. How would you rate the validity of these measurements?

- IQ test score ⟶ intelligence?
- SAT score ⟶ college readiness?
- personality test ⟶ human personality?
- attendance ⟶ quality of a musical performance?

Discussion: We can clearly come up with measurements that are *not* valid. For example, measurements of height and weight are not valid measures of intelligence. The number of followers you have on Instagram or TikTok is not a valid measure of your college readiness. Whether or not you like Five Guys doesn't tell me much about your personality. A senior music major may perform a flawless solo in a graduation performance with three people in attendance.

Remark: The problem with variables like intelligence, college readiness, and personality, which are common in the social sciences, education, management, and elsewhere, is that it's not always clear what they really mean.

- Measuring physical characteristics on human subjects (e.g., height, weight, eye color, etc.) identify precisely the property being measured.
- Measuring social and psychological properties is more ambiguous. For example, as Moore and Notz put it,

  > *"If we can't agree on exactly what intelligence is, we can't agree on how to measure it."*

Definition: A measurement is said to have **predictive validity** if it can be used to predict success on tasks that are related to the property being measured.

Discussion: Take "college readiness" and think about the following two questions:

1. Is SAT score a valid measurement of college readiness?
2. Do SAT scores help to predict whether USC students will graduate within 6 years?

The first question is hard to answer and even controversial. The second question is not. Data on graduation rates overwhelmingly show students with higher SAT scores have higher graduation rates on average. Therefore, SAT score may not be entirely valid when measuring college readiness, but it does have predictive validity in determining graduation status.

Note: Assessing validity and predictive validity is done for groups of individuals as a whole; not specific individuals. There are always exceptions. We all know individuals who had lousy SAT scores who went on to graduate from college. Similarly, I also know students who had exceptional SAT scores but dropped out after the first year.

Exercise: In Example 8.3, which variable has better predictive validity in assessing chlamydia status? Remember the diagnosed status from the assay is not always correct.

1. if an individual had more than one sexual partner within the last 3 months (yes/no)
2. the number of sexual partners an individual had within the last 3 months (quantitative).

9 Do the Numbers Make Sense?

9.1 Introduction

Example 9.1. Dr. Joel Best, in the introduction of his book *Damned Lies and Statistics*, recalls a statistic a student cited in his 1994 PhD dissertation:

> *"Every year since 1950, the number of American children gunned down has doubled."*

The author, Craig Sautter, published part of his dissertation in the academic journal *Phi Delta Kappan* in 1995; the title of the article was "Standing Up to Violence." This statistic was used in his article. A number of other articles have cited Mr. Sautter's article (just do a quick Google search).

Discussion: The problem with this statement is that it's absolutely ridiculous. Suppose in 1950 there was one child "gunned down" in the US. If Sautter's claim was correct, here are the number of children that would be "gunned down" in subsequent years:

Year	#	Year	#	Year	#	Year	#	Year	#
1950	1	1959	512	1968	262,144	1977	134,217,728	1986	68,719,476,736
1951	2	1960	1,024	1969	524,288	1978	268,435,456	1987	137,438,953,472
1952	4	1961	2,048	1970	1,048,576	1979	536,870,912	1988	274,877,906,944
1953	8	1962	4,096	1971	2,097,152	1980	1,073,741,824	1989	549,755,813,888
1954	16	1963	8,192	1972	4,194,304	1981	2,147,483,648	1990	1,099,511,627,776
1955	32	1964	16,384	1973	8,388,608	1982	4,294,967,296	1991	2,199,023,255,552
1956	64	1965	32,768	1974	16,777,216	1983	8,589,934,592	1992	4,398,046,511,104
1957	128	1966	65,536	1975	33,554,432	1984	17,179,869,184	1993	8,796,093,022,208
1958	256	1967	131,072	1976	67,108,864	1985	34,359,738,368	1994	17,592,186,044,416

Therefore, when Mr. Sautter defended his dissertation in 1994, astonishingly, there were over 17 trillion children "gunned down" in the US that year. Dr. Best calls this the "worst social statistic ever."

Exercise: Let's carefully diagnose Mr. Sautter's statement:

- 1965: 32,768 children "gunned down." In 1965, the FBI identified only 9,960 criminal homicides in the entire country (including all people).
- 1975: $\approx$ 33.5 million. This is more than the number of children living in the US (about 25.4 million in 1975).
- 1978: $\approx$ 268 million. This exceeds the population of the United States (about 222 million in 1978).
- 1983: $\approx$ 8.5 billion. This exceeds the population of the world (about 4.7 billion in 1983).

- 1987: ≈ 137 billion. This exceeds the human population throughout all of history on Earth (about 110 billion).
- 1994: ≈ 17.5 trillion. This is $2^7 = 128$ times larger than the previous number!

Doubling numbers increase exponentially fast—much faster than what Mr. Sautter probably thought. But using the phrase "doubling each year" conveys that gun violence is a big problem, and this was probably his intended goal. Interestingly, Mr. Sautter (when pressed) later amended his claim to

> *"The number of American children killed each year by guns has doubled since 1950."*

This figure was taken from the Children's Defense Fund, a "non-profit" 501(c) social action organization in Washington DC (whose 2021 assets were listed at $58 million). This amended statement is actually plausible, but we learned in the last chapter that counts are not always a **valid** measurement when comparing two groups (US in 1950 versus US in 1994).

- The US population increased dramatically from 1950 to 1994 (152.3 million to 262.1 million). This is a 72.1% increase.
- A doubling of the number of gun deaths among children outpaces population growth, but not by a huge amount—certainly not by the amount Mr. Sautter wanted his audience to believe.

9.2 Percentage changes

Remark: Percentage changes are commonly reported in the media and elsewhere. For example,

- "In comparing the 2022-2023 and 2023-2024 academic years, tuition rates at private universities increased by about 5%. The costs of in-state and out-of-state tuition and fees at public universities also rose, by nearly 4% and 1.4%, respectively."
- "U.S. crude oil production fell by 7.4% in 2020, the largest annual decrease on record."
- Recognition of this fact may explain the 123% increase in ADHD prevalence among adults reported between 2007 and 2016.

To calculate a **percentage change** from one time period to the next, use this formula:

$$\text{percentage change} = \frac{\text{amount of the change}}{\text{starting value}} \times 100\%.$$

Example 9.2. In 1950, the US population was 152.3 million. In 1994, it was 262.1 million. Calculate the percentage change in the US population between 1950 and 1994.

$$\begin{aligned} \text{percentage change} &= \frac{262.1 \text{ million} - 152.3 \text{ million}}{152.3 \text{ million}} \times 100\% \\ &= \frac{109.8 \text{ million}}{152.3 \text{ million}} \times 100\% \\ &= 0.721 \times 100\% \\ &= 72.1\%. \end{aligned}$$

Between 1950 and 1994, the US population increased by 72.1%.

Example 9.3. The number of undergraduate students attending USC (Columbia campus only) in January 2023 was 25,606. In January 2024, it was 26,968. Calculate the percentage change between these two time periods.

$$\begin{aligned} \text{percentage change} &= \frac{26{,}968 \text{ students} - 25{,}606 \text{ students}}{25{,}606 \text{ students}} \times 100\% \\ &= \frac{1{,}362 \text{ students}}{25{,}606 \text{ students}} \times 100\% \\ &= 0.053 \times 100\% \\ &= 5.3\%. \end{aligned}$$

The number of undergraduate students attending USC (Columbia campus only) increased by 5.3% between January 2023 and January 2024.

Example 9.4. In 2019, the US annual oil crude production was 12.2 million barrels per day (b/d). In 2020, it was 11.3 million barrels per day. Calculate the percentage change between these two time periods.

$$\begin{aligned} \text{percentage change} &= \frac{11.3 \text{ million b/d} - 12.2 \text{ million b/d}}{12.2 \text{ million b/d}} \times 100\% \\ &= \frac{-0.9 \text{ million b/d}}{12.2 \text{ million b/d}} \times 100\% \\ &= -0.074 \times 100\% \\ &= -7.4\%. \end{aligned}$$

This corresponds to a 7.4% decrease in the US annual oil crude production between 2019 and 2020.

Remark: Percentage changes can be positive (an increase) or negative (a decrease). How large can a percentage increase be? Can a percentage increase exceed 100%?

Example 9.5. In 2021, *AmStat News* reported the following,

> *"The number of statistics and biostatistics bachelor's degrees awarded in 2020 increased 474% since 2010, capping what can justifiably be termed the decade of statistics."*

Here is where the 474% figure comes from:

- Number of STAT/BIOS degrees awarded in 2010: 861
- Number of STAT/BIOS degrees awarded in 2020: 4942.

The percentage change is

$$\begin{aligned}\frac{4942 \text{ degrees} - 861 \text{ degrees}}{861 \text{ degrees}} \times 100\% &= \frac{4081 \text{ degrees}}{861 \text{ degrees}} \times 100\% \\ &= 4.74 \times 100\% \\ &= 474\%.\end{aligned}$$

Note: A quantity can increase by any amount between two time periods.

- A 100% increase just means the quantity has doubled.
- A 200% increase means the quantity has tripled.
- A 300% increase means the quantity has quadrupled, and so on.

On the other hand, a quantity cannot decrease by more than 100%. Once a quantity loses 100% of its value, there is nothing left.

Example 9.6. An advertisement for a new product says that it

"reduces heartburn by 300%."

No it doesn't. Once heartburn has been reduced by 100%, there is no more heartburn. There may be a percentage decrease when comparing subjects who take the product and those who don't, but it does not exceed 100%.

9.3 Developing a "number sense"

Remark: We are bombarded on a daily basis by statistics and numbers, and many people believe the numbers they see without question. Developing a "number sense" is a phrase your authors use to have you ask if the numbers *do* make sense. Answer these questions before you believe what is being said:

1. What didn't they tell us?
2. Are the numbers consistent with each other?
3. Are the numbers plausible?
4. Are the numbers too good to be true?
5. Is the arithmetic right?
6. Is there a hidden agenda?

Example 9.7. Hobart and William Smith Colleges (in New York) reported that their students' SAT scores have jumped 20 points since 2006.

- In 2006, they instituted an optional-SAT-reporting policy in its admissions.
- Lower scores were less likely to be reported.
- A 20-point increase may not be surprising even if there was no change in incoming student quality.

Example 9.8. Based on July 2019 data, the Bureau of Labor Statistics (BLS) calculates the unemployment rate at 3.7%. Let's look at how this is calculated:

- The number of citizens in the civilian labor force (persons classified as employed or unemployed) is 163,351,000.
- The number of unemployed persons is 6,063,000.

Therefore, the official July 2019 unemployment rate is

$$\frac{6,063,000}{163,351,000} \approx 0.037 \quad \text{(or 3.7\%)}.$$

In July 2019, the number of American adults (aged 18 and over) was about 255,000,000. Therefore, the total number of citizens in the civilian labor force (163,351,000) excludes about 92 million adults:

- military, government employees, the retired, the disabled, persons not actively looking for work, underemployed, discouraged workers, etc.

How the BLS measures the unemployment rate depends on very precise definitions that the US government uses. These definitions are usually not provided when citing a figure like 3.7%.

Example 9.9. According to the FBI,

> *"over 26% of home burglaries take place between Memorial Day and Labor Day."*

This statistic was used in an advertisement for a home security system.

- There are 14 weeks between Memorial Day (end of May) and Labor Day (beginning of September). There are 52 weeks in a year.

$$\frac{14}{52} \approx 0.27 \quad \text{(or about 27\%)}.$$

This figure is consistent with burglaries happening at basically the same rate all year.

Example 9.10. The Census Bureau once gave a simple test of literacy in English to a random sample of 3400 people. The *New York Times* printed some of the questions under the headline,

"113% of adults in US failed this test."

The embarrassing thing is that this must have passed multiple editors, all of whom missed it or didn't know it was wrong.

Example 9.11. An article in *Organic Gardening* magazine made the claim,

> *"the US Interstate Highway System spans 3.9 million miles and is wearing out 50% faster than it can be fixed. Continuous road deterioration adds $7 billion yearly in fuel costs to motorists."*

- The distance from Charleston to Seattle is 3,000 miles. Therefore, 3.9 million is the equivalent of making 650 round trips between these two cities.
- Actual length of US Interstate Highway System = 47,856 miles.

I don't have much faith in the $7 billion annual price tag claim, but it sure sounds big. The "wearing out percentage" also sounds alarming.

Example 9.12. An article in the November 3, 2009 issue of *The Guardian* reported that,

> *"50 percent of obese people earn less than the national average income."*

The intended meaning here is that obese people are discriminated against in the workforce. However, if the population distribution of incomes is approximately symmetric (e.g., a normal distribution), then this is exactly what one would expect for all people in the workforce. In other words, this claim is actually entirely consistent with "no discrimination."

Example 9.13. Poll results for the survey question: *"Do you approve of the current bond measure?"*

Results: 52% Yes; 44% No; 15% Undecided.

These percentages add up to 111%. Oops.

Remark: Instead of Moore and Notz's 6 questions, Dr. Best's recommended questions to ask are simpler. I like these better too.

1. Who created this statistic?
2. Why was this statistic created?
3. How was this statistic created?

Discussion: Here are additional examples from Dr. Best:

1. A child advocate tells Congress that 3,000 children per year are lured with internet messages and then kidnapped.

2. Tobacco opponents attribute over 400,000 deaths per year to smoking.

3. Anti-hunger activists say that 31 million Americans regularly face hunger.

Statistics like these are usually pushed by individuals who have an agenda. For example, when I did a little background reading on the "31 million" statistic for hunger (a September 8, 2000 article in the *Pittsburgh Post-Gazette*), I found out that the following instances were classified as hunger:

- food insecurity, meaning that a household has limited or uncertain access to nutritious food

- declining food stamp use.

Therefore, the definition of what it means to be "hungry" is misleading, at best. Of course, saying a big number like "31 million" has more impact. This brings more attention to the problem, which inevitably leads to this statistic being used over and over again (an example of what Dr. Best calls a **mutant statistic**).

DLS: Dr. Best writes this compelling excerpt, which summarizes the overall theme of this chapter:

> *"Innumeracy–widespread confusion about basic mathematical ideas–means that many statistical claims about social problems don't get the critical attention they deserve. This is not simply because an innumerate public is being manipulated by advocates who cynically promote inaccurate statistics. Often, statistics about social problems originate with sincere, well-meaning people who are themselves innumerate; they may not grasp the full implications of what they are saying. Similarly, the media are not immune to innumeracy; reporters commonly repeat the figures their sources give them without bothering to think critically about them."*

And further,

> *"The result can be a social comedy. Activists want to draw attention to a problem [e.g., gun violence, homelessness, etc.]. The press asks the activists for statistics....knowing that big numbers indicate big problems....The activists produce a big estimate, and the press....simply publicizes it. The general public–most of us suffering from at least a mild case of innumeracy–tend to accept the figure without question."*

10 Graphs, Good and Bad

10.1 Graphs for categorical data

Recall: A **variable** is a characteristic that we measure on each individual.

- Categorical ⟶ places individuals into one of several groups or categories
- Quantitative ⟶ assumes numerical values which have a physical meaning.

Example 10.1. Recall the necrotizing enterocolitis (NEC) study in Example 1.1 (notes). One of the variables recorded for each infant was nutrition type. Here are the possible categories for this variable:

- Breastmilk
- Fluids (mostly sugars and salts)
- TPN (total parental nutrition; contains all essential fluids and electrolytes)
- Formula.

Nutrition type is categorical because the "values" listed above identify categories. In my Excel file which contains these data, the following codings are used:

$$1 = \text{Breastmilk; } 2 = \text{Fluids; } 3 = \text{TPN; } 4 = \text{Formula.}$$

These codings are numerical, but they do not have a physical meaning. They simply keep track of which category is which.

Definition: The **distribution** of a variable (categorical or quantitative) tells us

(a) what values the variable can have

(b) how often it has these values.

An easy way to show the distribution of data for a categorical variable is to construct a **table**. Here is the table for nutrition type in Example 10.1:

Category	Breastmilk	Fluids	TPN	Formula	Total
Count	237	60	265	43	605
Proportion	0.39	0.10	0.44	0.07	1

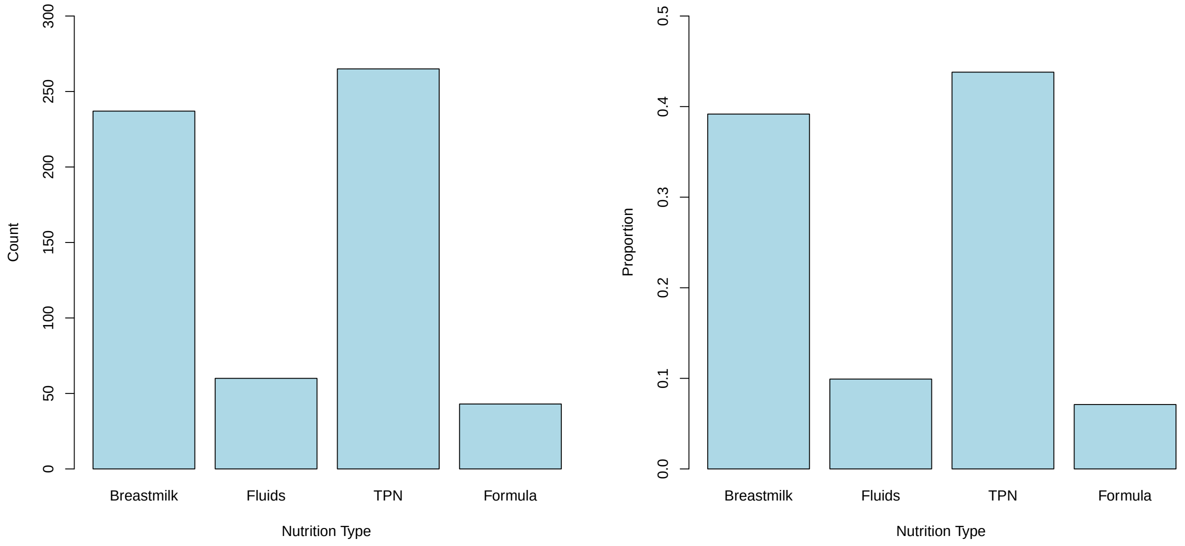


Figure 10.1: Necrotizing enterocolitis data. Bar graphs of nutrition type for 605 infants. Left: Counts. Right: Proportions.

Note: There were 615 infants in the study, but nutrition type was not recorded for 10 of the infants (i.e., these values were "missing"). The table on the last page shows the distribution for the infants whose data are not missing. Therefore, the proportions are calculated as

$$\text{Proportion} = \frac{\text{Count}}{605}$$

and are rounded to 2 digits. Note that the proportions in the categories add up to 1. This is a requirement in any distribution for a categorical variable.

Graphs: There are two graphs which display the distribution of a categorical variable: a **bar graph** and a **pie chart**. A bar graph can use either counts or proportions.

- Figure 10.1 above shows the bar graphs for nutrition type in Example 10.1. The only difference in the figures is the scale used for the vertical axis.
- Figure 10.2 (next page) shows the pie chart for nutrition type. Recall there are 360 degrees in a circle. Therefore, the angles formed in pie chart are

$$\begin{aligned}
\text{Breastmilk:} &\quad 0.39 \times 360 = 140.4 \text{ degrees} \\
\text{Fluids:} &\quad 0.10 \times 360 = 36.0 \text{ degrees} \\
\text{TPN:} &\quad 0.44 \times 360 = 158.4 \text{ degrees} \\
\text{Formula:} &\quad 0.07 \times 360 = 25.2 \text{ degrees.}
\end{aligned}$$

R does all the work, so we don't have to calculate these angles ourselves. Note the degrees above do add up to 360.

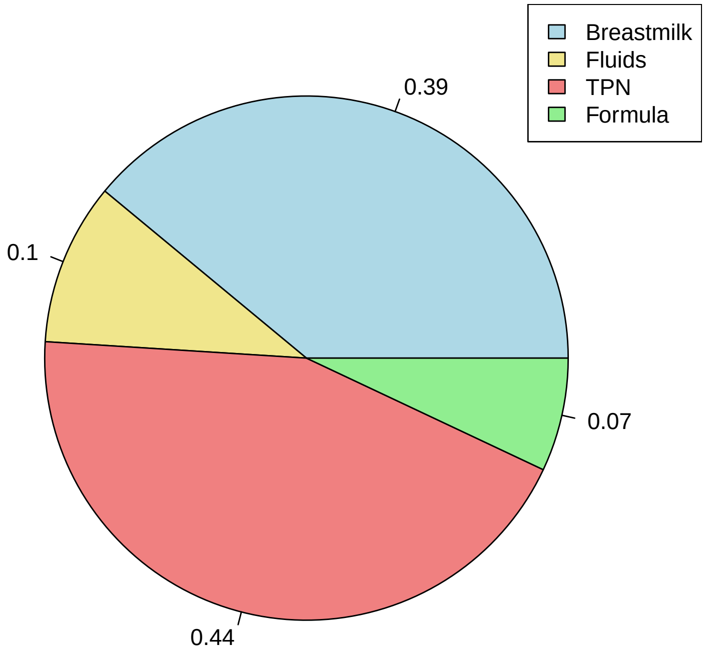


Figure 10.2: Necrotizing enterocolitis data. Pie chart of nutrition type for 605 infants.

Discussion: Both bar graphs and pie charts can be used to show the distribution of data measured on a categorical variable (like nutrition type). However, remember the data in Example 10.1 come from a **sample** of infants. Therefore, the graphs we have shown are for the sample, and these are "estimates" of the analogous graphs for the larger population. Recall this is the idea behind **statistical inference**.

Remark: Pie charts are only used when the category proportions add to 1 (i.e., the angles add to 360 degrees). However, bar graphs can be used when this is not true, as the next example shows.

Example 10.2. Here are the percentage of residents who have a bachelor's degree in (what I consider to be) the 10 southern states:

State	Percent	State	Percent
Alabama	27.4	Arkansas	25.3
Florida	33.2	Georgia	34.6
Louisiana	26.5	North Carolina	34.9
Mississippi	24.8	South Carolina	31.5
Tennessee	30.5	Virginia	41.8

Note that these percentages do not add to 100 percent. However, we can still use a bar graph to display these percentages.

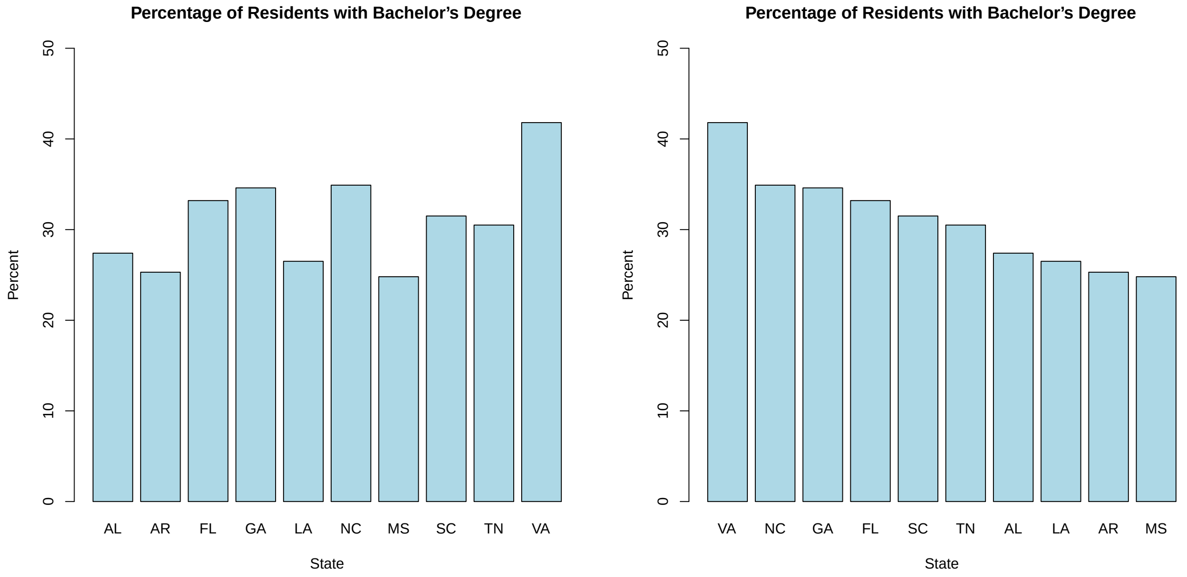
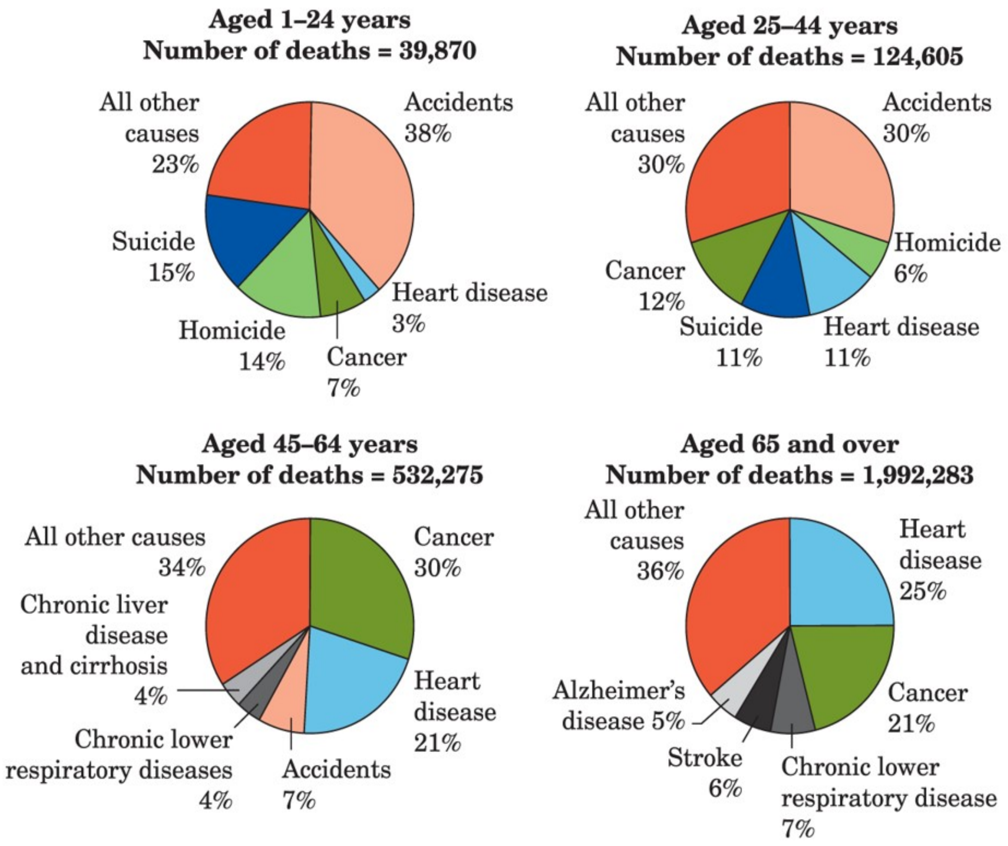


Figure 10.3: Percentage of residents with a bachelor's degree. These data were taken from the American Community Survey in 2021.

Remark: This type of bar graph is used to compare the different states; it is not used to show any type of distribution. The two bar graphs in Figure 10.3 list the same percentages; the one on the right just lists the percentages in descending order. This is easier to visualize.

Example 10.3. The National Center for Health Statistics (NCHS) keeps track of the leading causes of death in the United States. Moore and Notz (pp 214) display data on 2015 deaths taken from a NCHS annual report; I have reproduced this graph on the next page.

- This graph shows information on two categorical variables: cause of death and age group.
- Age (in years) is a quantitative variable. However, the authors have **dichotomized** it into four age categories, thereby creating a categorical variable:
  - age 1-24 years
  - age 25-44 years
  - age 45-64 years
  - age 65 and older.
- The five "leading causes" of death are shown in each age group. Leading causes in one age group will not necessarily be leading causes in another group. A sixth category is included in each age group showing "other causes."



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

10.2 Graphs for quantitative data measured over time (time series data)

Note: It is common to observe quantitative data that are measured at regular intervals over time. For example,

- Daily gas prices (in dollars)
- Monthly sales (in dollars)
- Daily high temperatures (in deg F)
- Number of home runs hit each year in MLB
- Number of tuberculosis cases per month in the United States.

Each of these variables is quantitative, and we can observe the values of each variable at each time period (e.g., each day, each month, each year, etc.). Data that are observed over time are called **time series data**.

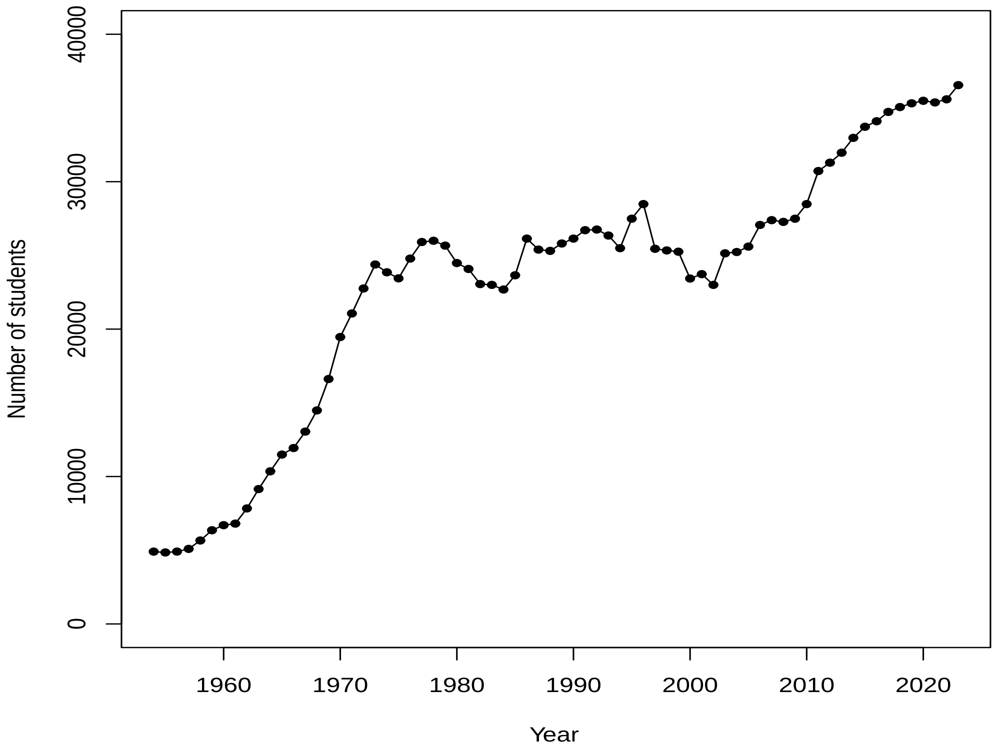


Figure 10.4: USC enrollment data. Number of students enrolled at USC (Columbia campus) in the fall semester during 1954-2023.

Example 10.4. I recorded the number of students enrolled at USC during the fall semester each year from 1954 to 2023 (Columbia campus only; undergraduate and graduate students combined). These data are available from the USC Office of Institutional Research, Assessment, and Analytics. Here is part of the data set:

Year	# Students
1954	4,906
1955	4,849
1956	4,907
⋮	⋮
2021	35,376
2022	35,587
2023	36,548

The variable is the number of students enrolled at USC (Columbia campus) in the fall semester. This is a **quantitative** variable and it is measured each year, starting in 1954 and ending in 2023. There are a total of 70 observations. This is an example of a time series data set.

Terminology: Figure 10.4 is an example of a **line graph**, which is also known as a **time series graph**. This graph shows how the values of a quantitative variable (like enrollment counts) change or evolve over time. Here are some observations:

- We see a general increase in enrollment over time (from 1954 to 2023). This is called a **trend**.
- Upon closer inspection,
  - there is an increasing trend in enrollment between 1954-1970
  - enrollment stays fairly flat between 1970-2000
  - there is another increasing trend between 2000 and 2023.

Remark: In general, look for these three characteristics when examining data over time in line graphs.

- Overall trend. A trend is a long-term increase or decrease over time.
- Seasonal variation. These are patterns which repeat themselves over time; e.g., each week, each month, each year, etc.
- Sharp deviations. These are unusual observations that deviate greatly from the overall pattern.

Example 10.5. The course STAT 520 is called "Time Series and Forecasting." Whenever I teach this course, I ask the students to do an end-of-the-semester project where they carefully analyze a time series data set of their choice. Here are four data sets taken from projects the last time I taught the course (in Fall 2013):

1. Airline mile data. Number of air miles (in 1000s) traveled by passengers in the US each month (Jan 1996-May 2005)
2. Home run data. Number of home runs hit by the Boston Red Sox each year (1909-2010)
3. Earthquake data. Number of earthquakes measuring $\geq$7.0 on the Richter Scale observed worldwide each year (1900-1998)
4. Supreme Court data. Percentage of cases granted review by the US Supreme Court each year (1926-2004).

Exercise: Line graphs for the data sets are shown in Figure 10.5 and Figure 10.6 (next two pages). Interpret what characteristics you see in each time series using the terminology defined above.

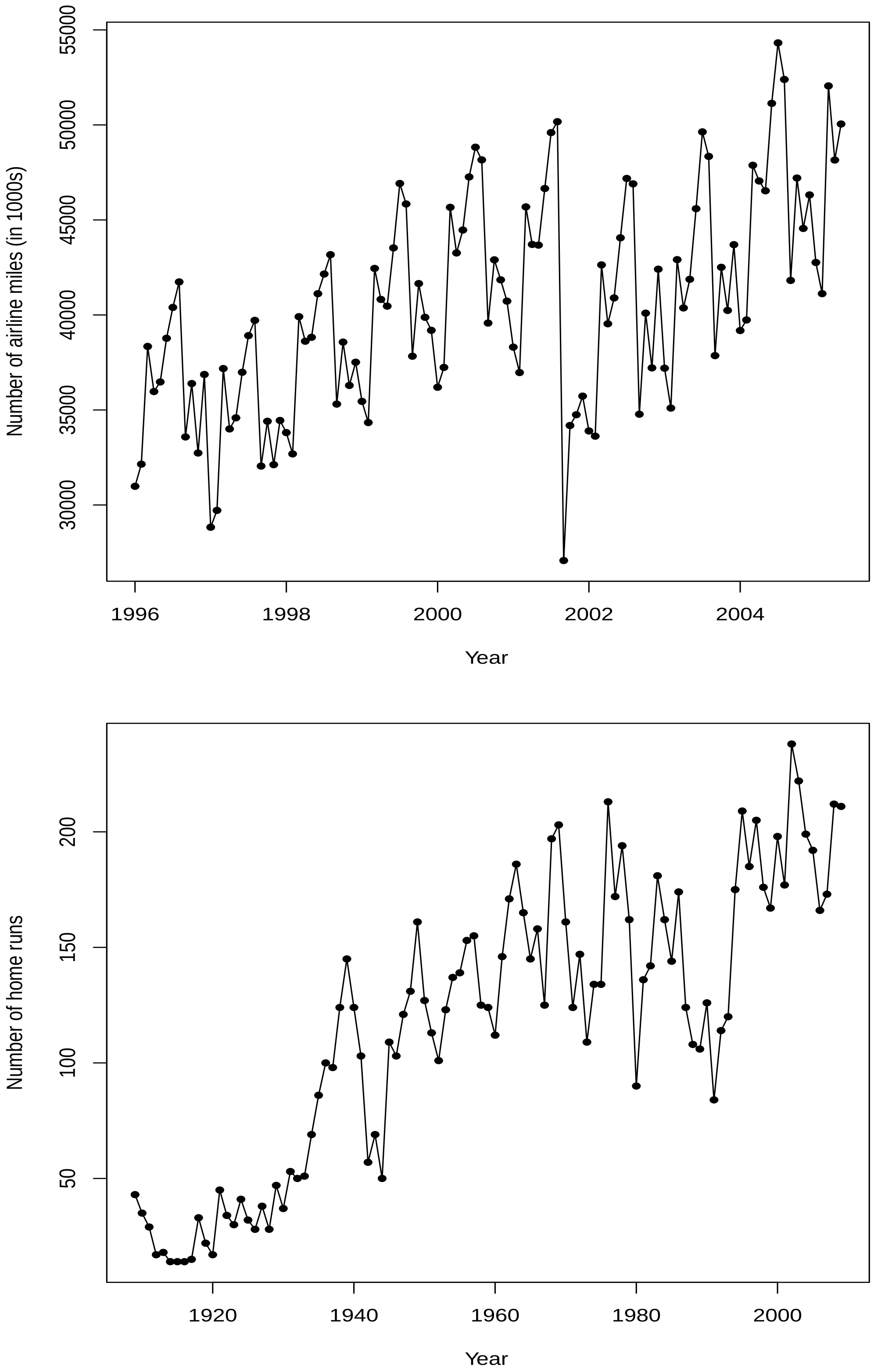


Figure 10.5: Top: Airline mile data. Bottom: Home run data.

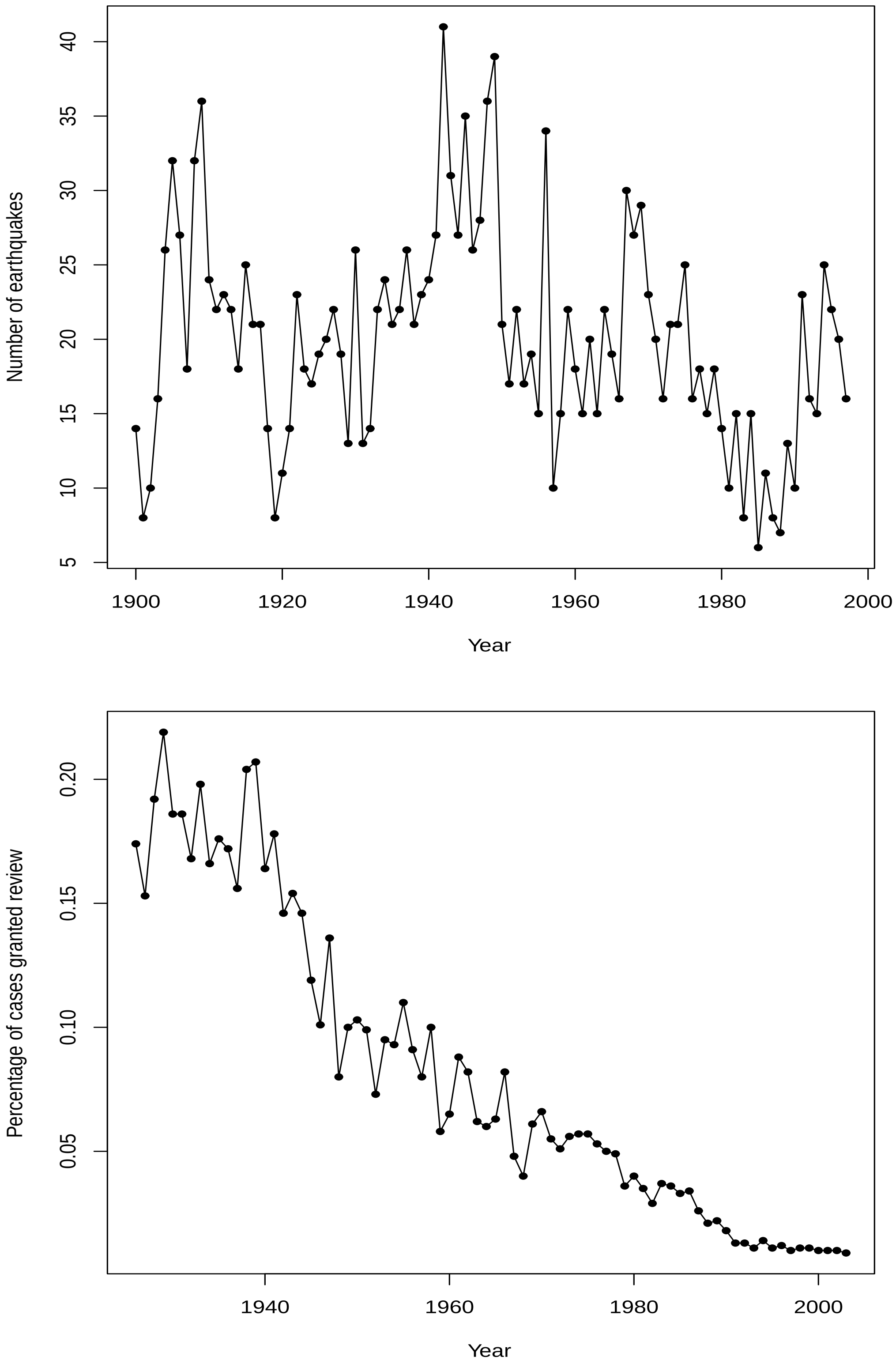


Figure 10.6: Top: Earthquake data. Bottom: Supreme Court data.

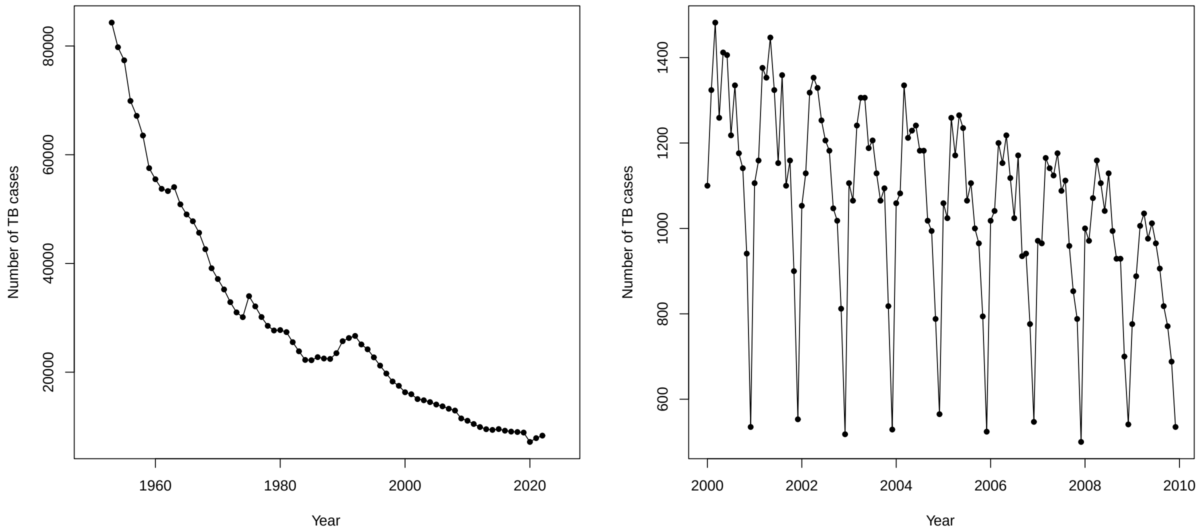


Figure 10.7: Number of TB cases. Left: Yearly data (1953-2022). Right: Monthly data (Jan 2000-Dec 2009).

Example 10.6. Tuberculosis (TB) is a bacterial infection that can spread through the lymph nodes and bloodstream to any organ in your body. It is most often found in the lungs. Most people exposed to TB never develop symptoms, because the bacteria can live in an inactive form in the body. However, if the immune system becomes compromised, TB bacteria can become active and ultimately fatal.

- Figure 10.7 (above) shows line graphs of TB case counts.
  - The graph on the left shows yearly counts between 1953-2022.
  - The graph on the right shows monthly counts between Jan 2000-Dec 2009.
- The line graph for the yearly counts shows a sharp decreasing trend over time, which is attributed to many factors, including improvements in living and social conditions and effective treatment.
- The line graph for monthly counts shows the downward trend over the 10-year span, but it also shows the seasonal pattern within each year. Monthly case counts generally repeat the same pattern each year.

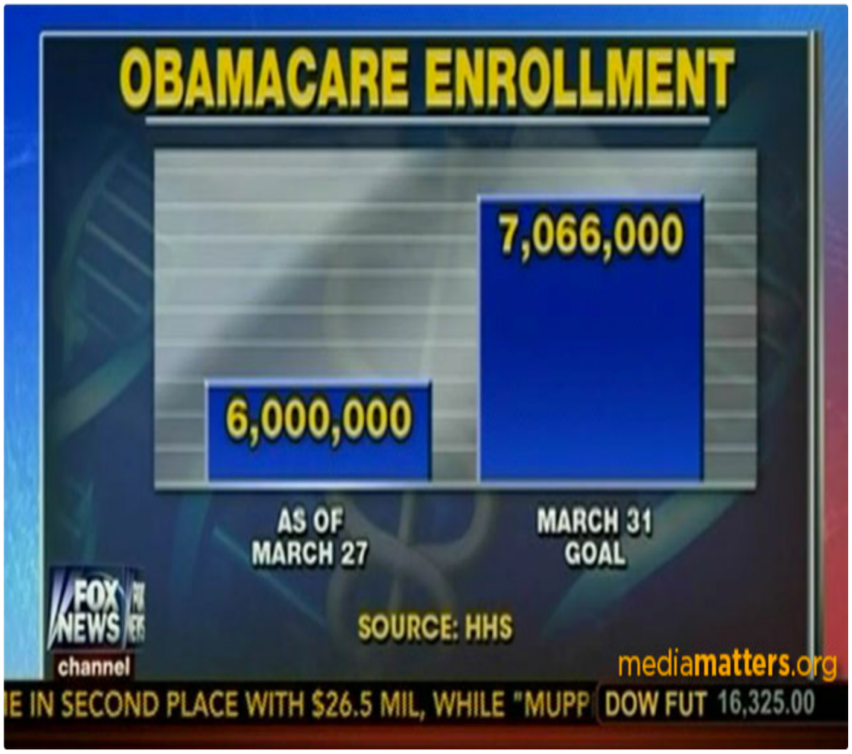
Remark: The real value of modeling time series data is to make **predictions** about what will happen in the future. For example,

- How many TB cases will be seen in the US next month? next year?
- How many students will be enrolled at USC in Fall 2025? Fall 2030?

For a given time series, statistical models can extract information about what has happened in the past and predict future values in the series that have yet to be seen.

10.3 Examples of bad and misleading graphs

Example 10.7. In March 2014, Fox News displayed this graph on one of its nightly programs:

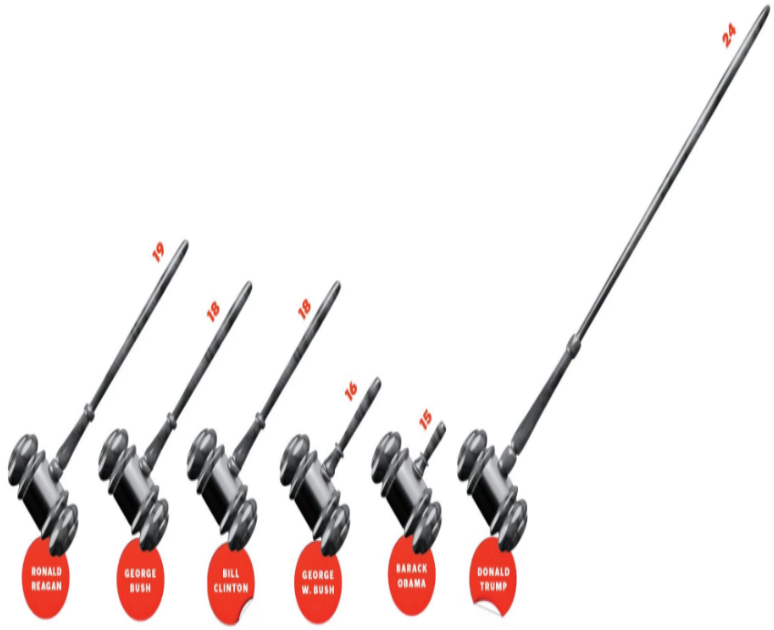


Q: Why is this graph misleading?
A: The height of the second bar is about 2.7 times larger than that of the first bar. However, if we calculate the **percentage change** in enrollment between the two time periods, we get

$$\begin{aligned}\text{percentage change} &= \frac{\text{amount of the change}}{\text{starting value}} \times 100\% \\ &= \frac{7{,}066{,}000\text{-}6{,}000{,}000}{6{,}000{,}000} \times 100\% \\ &= \frac{1{,}066{,}000}{6{,}000{,}000} \times 100\% \\ &\approx 0.178 \times 100\% \\ &= 17.8\%.\end{aligned}$$

So, the March 31 enrollment is only a 17.8% increase when compared to the enrollment on March 27. However, the difference in heights (2.7) corresponds to a 170% increase between the two time periods! This is almost 10 times larger than the actual percentage increase.

Example 10.8. In August 2018, *New York Times Magazine* displayed this graph in one of its articles about the Trump administration:



Appellate Judgeships Confirmed During First Congressional Term. Ronald Reagan, 19; George Bush, 18; Bill Clinton, 18; George W. Bush, 16; Barack Obama, 15; Donald Trump, 24. Illustration by Tracy Ma

Q: Why is this graph misleading?
A: This is one of the worst graphs I have ever seen.

- As in Example 10.7, the heights are completely distorted. For example, 24 Trump judges confirmed is a 60% increase when compared to President Obama at the same time during his presidency (15). That is,

$$\text{percentage change} = \frac{24-15}{15} \times 100\% = 60\%.$$

  However, the height of the gavels suggests this is about a 400% increase (the Trump gavel is about 5 times higher).

- And what's with the gavels to begin with? This is a classic example of **chart junk**. All gavels do in this graph is distract the reader from what the data actually say. This is likely the intended goal—to target innumerate readers.

- The text and the confirmed judge counts are intentionally kept small to hide relevant information. Innumerate readers likely focus on the height of the "Trump gavel" only.

- A much better graph of these data is shown in Figure 10.8 (next page).

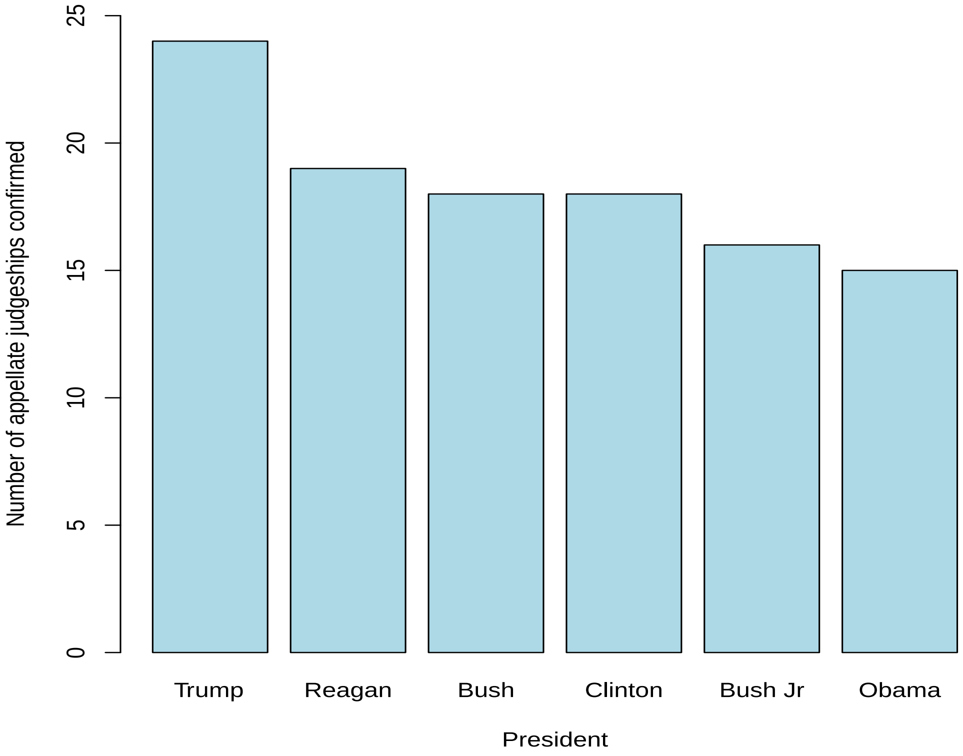


Figure 10.8: Number of appellate court judges confirmed during first term.

Example 10.9. *Is global warming real or not?* We won't answer this question in this example, but we will show two different graphs of the same data that reveal different impressions. The National Weather Service records the annual average global surface temperature (in deg F) during 1875-2023. Figure 10.9 shows two graphs of these same data. The graphs are different only in that they use different vertical axis scales.

- The top figure uses a range of 35 to 70 deg F on the vertical axis. Doing this greatly compresses the data in the figure and results in a lot of "wasted space." The effect of this is the time series looks relatively flat over time.

- The bottom figure uses a range of 48 to 56 deg F on the vertical axis. This range better corresponds to where the smallest and largest observations in the time series actually are. An increasing trend starting around 1960 is now evident.

- By changing the scales, you can greatly alter the visual impression of the data. Unfortunately, this is a common tactic the media and others use to distort and suppress information. As Moore and Motz put it,

"Because graphs speak so strongly, they can mislead the unwary."

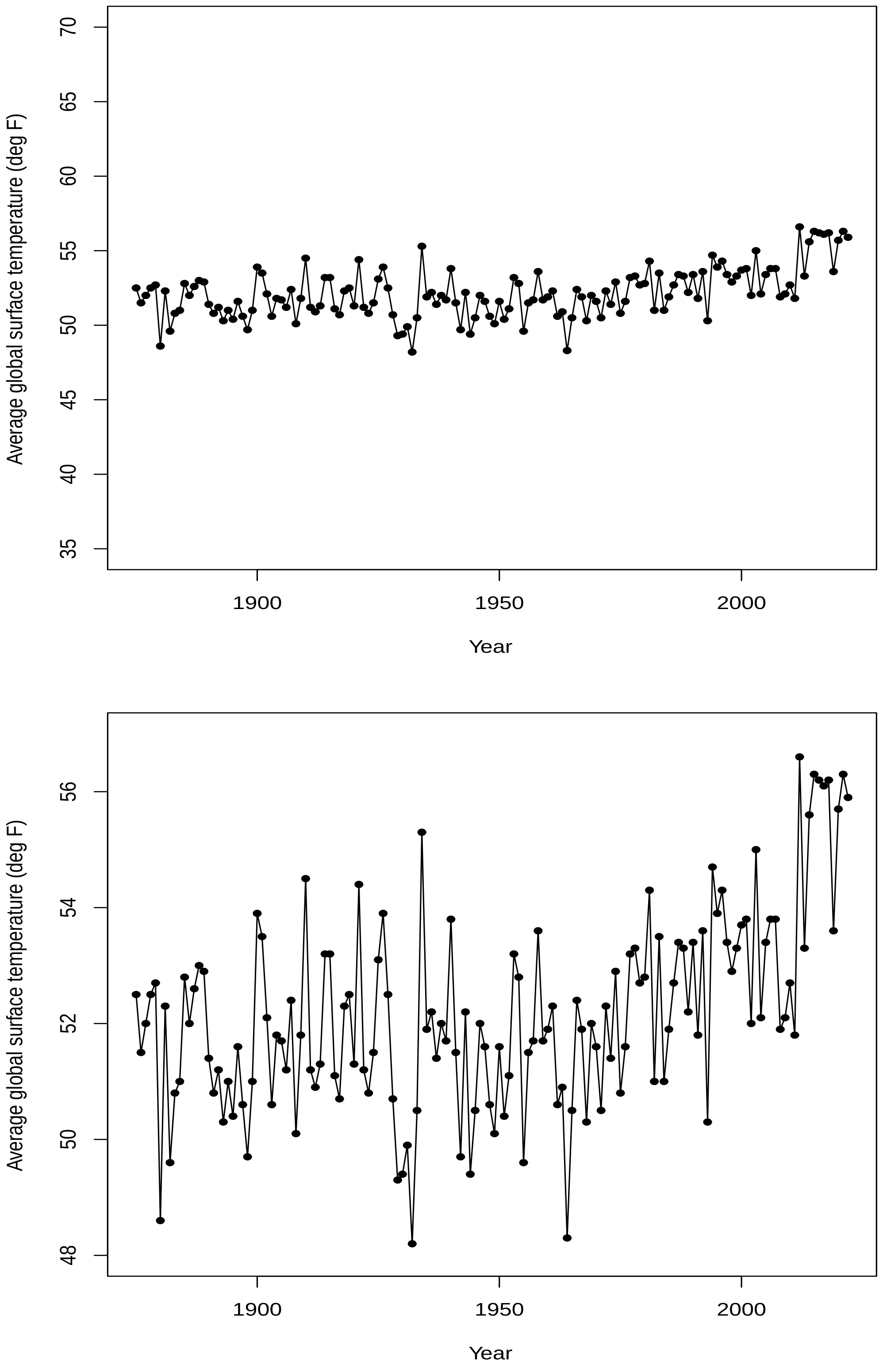


Figure 10.9: Annual average global surface temperature (in deg F) from 1875-2023. Different vertical axis scales are used.

11 Displaying Distributions with Graphs

11.1 Introduction

Recall: The **distribution** of a variable (categorical or quantitative) tells us

(a) what values the variable can have

(b) how often it has these values.

In Chapter 10, we talked about bar graphs and pie charts. These are the two most common graphs used to show the distribution of a categorical variable (e.g., nutrition type, sex/gender, race, political affiliation, etc.). Recall that categorical variables are variables whose values simply denote category membership.

Note: Different graphs are used to show the distribution of quantitative data, which is the topic of this chapter. Here are the three we will use:

- histogram
- stemplot
- boxplot (Chapter 12).

We have already seen an example of a histogram in Chapter 1, when we showed the distribution of birth weights (in grams) for 615 premature infants in the NEC-caffeine observational study. This histogram is also shown in Figure 11.1 (next page).

11.2 Histograms

Remark: Close inspection of Figure 11.1 shows that a histogram can be constructed by knowing two things: the **intervals** selected on the horizontal axis and the **counts** on the vertical axis.

- The intervals on the horizontal axis are also called **classes** or **bins**. All intervals are of the same width. In Figure 11.1, the intervals are 0-500 grams, 500-1000 grams, 1000-1500 grams, and so on. Each datum belongs to exactly one interval.
- The counts on the vertical axis count the number of observations that fall in each interval. For example,
  - there are 11 infants whose birth weight is between 0-500 grams,
  - there are 79 infants whose birth weight is between 500-1000 grams,
  - there are 92 infants whose birth weight is between 1000-1500 grams, and so on.

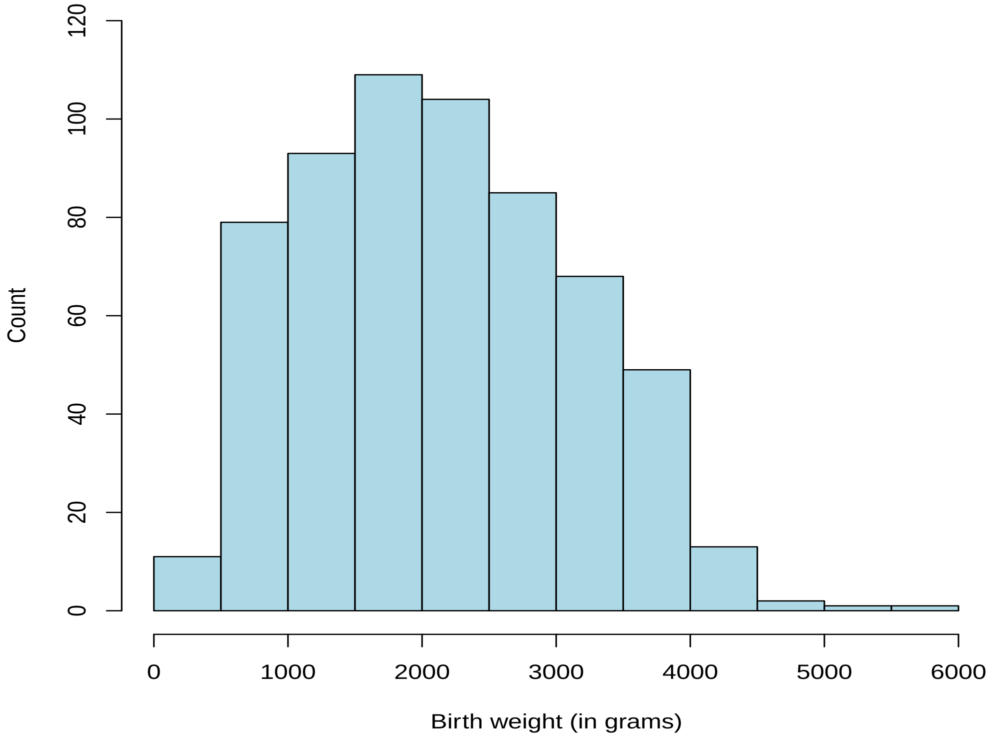


Figure 11.1: Necrotizing enterocolitis data. Histogram of birth weights for 615 infants.

These counts can be converted to proportions by division. Recall there are 615 infants shown in Figure 11.1. Therefore,

- the proportion (percentage) of infants whose birth weights are between 0 and 500 grams is
$$\frac{11}{615} \approx 0.017 \quad \text{(or 1.7\%)}.$$
- the proportion (percentage) of infants whose birth weights are between 500 and 1000 grams is
$$\frac{79}{615} \approx 0.128 \quad \text{(or 12.8\%)}.$$
- the proportion (percentage) of infants whose birth weights are between 1000 and 1500 grams is
$$\frac{92}{615} \approx 0.150 \quad \text{(or 15.0\%)}.$$

Note: If we did this calculation for all intervals in Figure 11.1, then the proportions would add up to 1 (i.e., the percentages would add up to 100%).

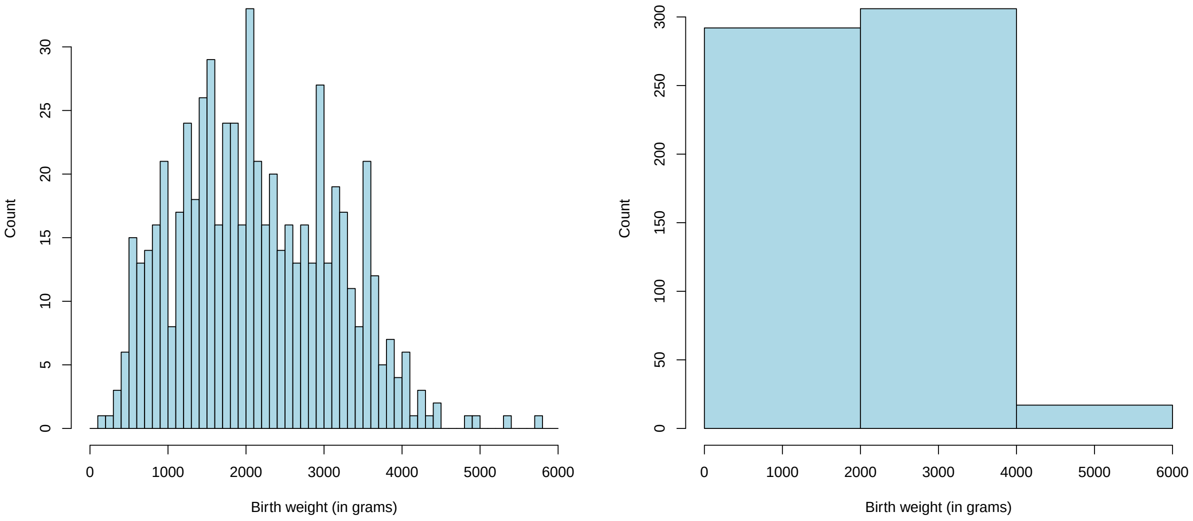


Figure 11.2: Necrotizing enterocolitis data. Histograms of birth weights for 615 infants. Left: Interval width = 100 grams; Right: Interval width = 2000 grams.

Q: When we are constructing a histogram, how should we choose the intervals and their widths?
A: There is no "right answer" to this question, but there are certainly bad ways to choose the intervals. For example, Figure 11.2 above shows what would happen if we chose the interval widths to be

- 100 grams; i.e., the intervals are 0-100 grams, 100-200 grams, 200-300 grams, and so on. This is the figure shown on the left.
- 2000 grams; i.e., the intervals are 0-2000 grams, 2000-4000 grams, and 4000-6000 grams. This is the figure shown on the right.

The figure on the left has excessive granularity while the figure on the right is highly uninformative. Neither figure does a good job at displaying the distribution of the birth weights. When you are constructing your own histograms, you want to avoid both of these extremes and select intervals that are "just right." Typically, what I do first is use the default selections for intervals that R determines, as I did in Figure 11.1. If I don't like R's selections (because the graph doesn't look right), I will tell R which intervals to try next. Iterating to the final histogram usually involves trial and error.

Important: Constructing histograms is easy. Interpreting them is more important. The first thing we should remember is **statistical inference**.

- Histograms are used to show the distribution of data recorded for a quantitative variable (like birth weight).

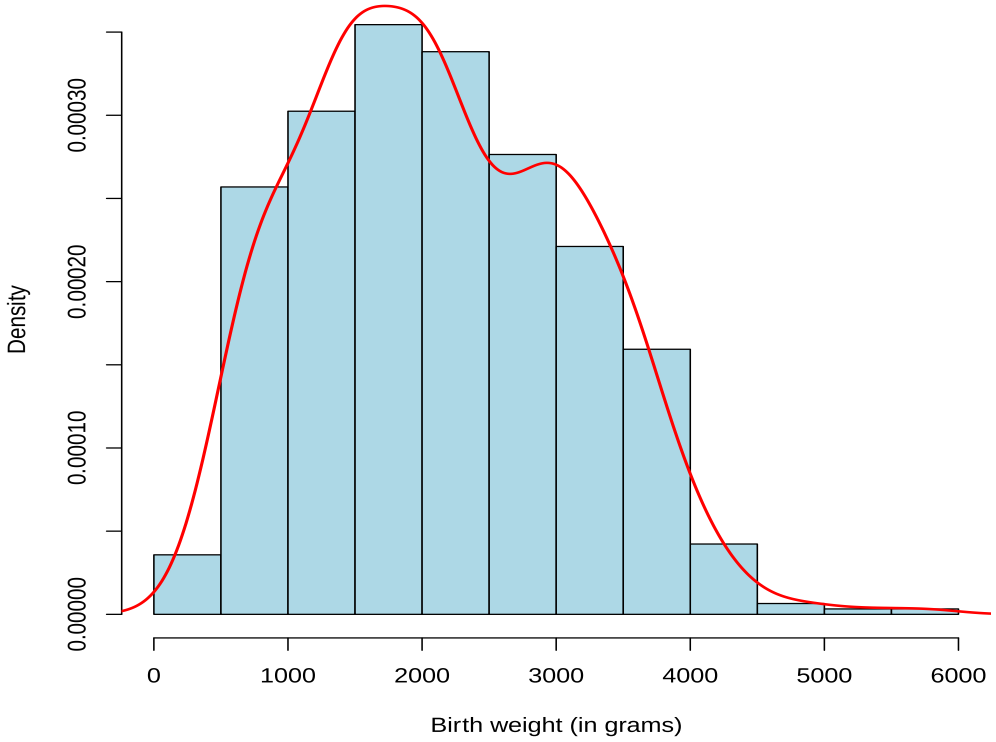


Figure 11.3: Necrotizing enterocolitis data. Histogram of birth weights for 615 infants. An estimate of the population density curve has been added.

- If the data we have are from a **sample**, and if the sample is representative of the population, then the histogram presents "an impression" of the distribution of the variable in the entire population.
- Therefore, by interpreting characteristics we see in the histogram (for the sample), we are interpreting "what might be going on" in the larger population of individuals.

Terminology: If a histogram is prepared for a sample of quantitative data, and if this sample is representative of a larger population, then the histogram is "approximating" a smooth curve that describes all individuals in this population. We call this smooth curve a **population density curve**.

- The population density curve is a mathematical model which describes the distribution of the variable (e.g., birth weights, etc.) for all individuals in the population.
- R will calculate an **estimate** of the population density curve using the data in the sample. It's only an estimate because it is determined from the sample.

- Figure 11.3 (previous page) shows an estimate of the population density curve for the birth weight data in the NEC-caffeine study.
  - The curve approximates the distribution we see in the histogram. In mathematical language, the curve is a **function**.
  - Note that the vertical axis scale has changed in Figure 11.3. This has been done automatically (by R) to ensure that the area under the population density curve estimate is equal to 1.
- Population density curves will be discussed more in Chapter 13. The **normal distribution** is the most common population density curve.

Interpretation: We will focus on the following four characteristics when we examine and describe histograms in words:

1. Center: Where does the center of the distribution fall approximately? You can think of this as the point where the histogram would balance.
2. Variability: How much variation is in the distribution? How spread out is it? What is the range of possible observations?
3. Shape: What type of shape does the distribution have? Is it symmetric or skewed (right/left)? Does it have a single peak or multiple peaks?
4. Deviations (from the overall pattern): Are there observations that fall outside the overall pattern of the distribution? These observations are called **outliers**.

Remark: In Chapter 12, we will present numerical summaries of "center" and "variability." We will also describe how to classify an observation as an outlier or not.

Discussion: Let's interpret the histogram in Figure 11.1 (or in Figure 11.3) in terms of our four characteristics:

- Center: The center of the birth weight distribution is around 2250 grams.
- Variability: All of the birth weights are between 0 and 6000 grams.
- Shape: This distribution has a single peak (around 1750 grams) and is slightly **skewed to the right**.
- Deviations: There are no obvious outliers.

Example 11.1. A researcher was interested in studying the eating habits of elementary school children. Her research team measured the body mass index (BMI) of $n = 328$ fourth-grade children sampled from public schools in Augusta, GA. An individual's BMI is calculated as follows:

$$\text{BMI} = \frac{\text{weight (kg)}}{[\text{height (m)}]^2}.$$

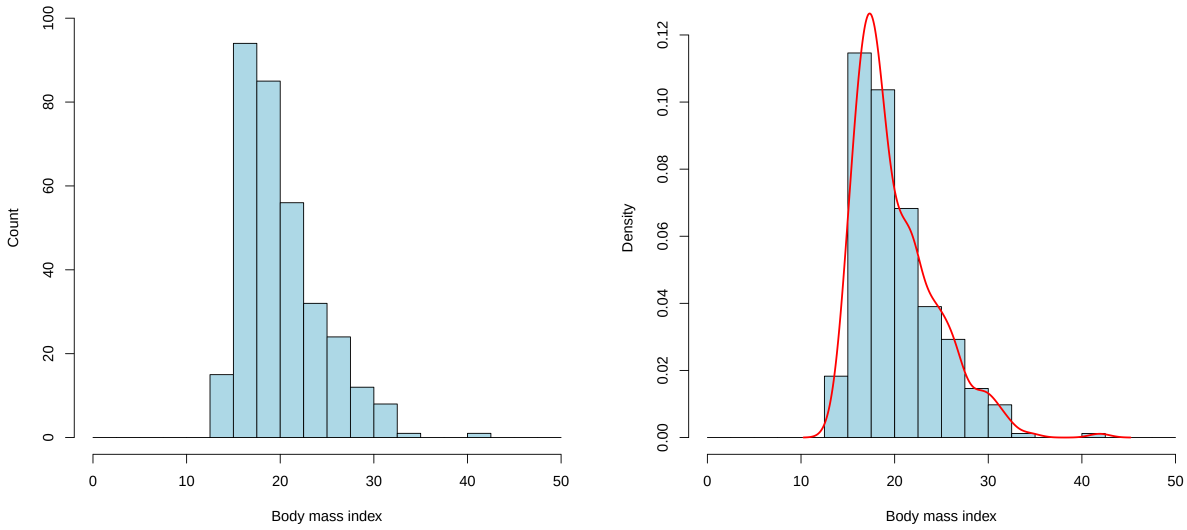


Figure 11.4: Childhood obesity study. Histogram of BMI data for 328 fourth-grade students in Augusta, GA. Right: An estimate of the population density curve has been added.

BMI is a quantitative variable. Here are the CDC guidelines for interpreting what BMI means for 10-year-old children (the approximate age of those in fourth grade):

≤ 14.5: Underweight; 14.5-19.5: Healthy; 19.5-21.5: Overweight; ≥ 21.5: Obese.

Interpretation: Here is how we would interpret the histogram in Figure 11.4 in terms of our four characteristics:

- Center: The center of the BMI distribution is around 20.
- Variability: Almost all of the BMI observations are between 12.5 and 35.
- Shape: This distribution has a single peak (around 16-17) and is **skewed to the right**.
- Deviations: The BMI observation larger than 40 is an outlier.

Note: Figure 11.4 (right) shows the same histogram but with an estimate of the population density curve added.

- The histogram corresponds to the 328 fourth-grade students in the **sample**.
- The population density curve serves as a mathematical model for all fourth-grade children in the **population**.

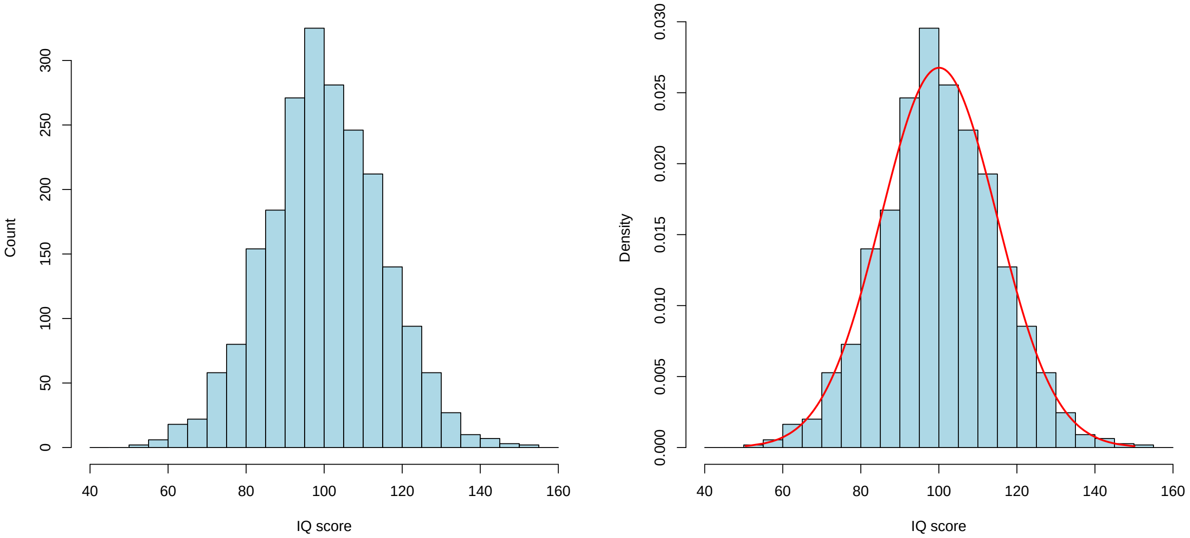


Figure 11.5: IQ data (simulated). Left: Histogram of IQ scores for a sample of $n = 2200$ Americans. Right: An estimate of the population density curve has been added.

Example 11.2. The Wechsler Adult Intelligence Scale (WAIS) is an IQ test. A recent version of this test was administered to a sample of $n = 2200$ Americans (aged 16-90). I don't have access to the data (for confidentiality reasons) so I simulated what the scores might look like given the summary information which was available. The histogram of the simulated IQ score data is shown in Figure 11.5.

Interpretation: Here is how we would interpret the histogram in Figure 11.5 in terms of our four characteristics:

- Center: The center of the IQ score distribution is around 100.
- Variability: Almost all of the IQ scores are between 50 and 150.
- Shape: This distribution has a single peak (around 100) and is **symmetric**.
- Deviations: There are no obvious outliers.

Note: Figure 11.5 (right) shows the same histogram but with an estimate of the population density curve added.

- The histogram corresponds to the 2200 Americans in the **sample**.
- The population density curve serves as a mathematical model for all Americans (aged 16-90) in the **population**.

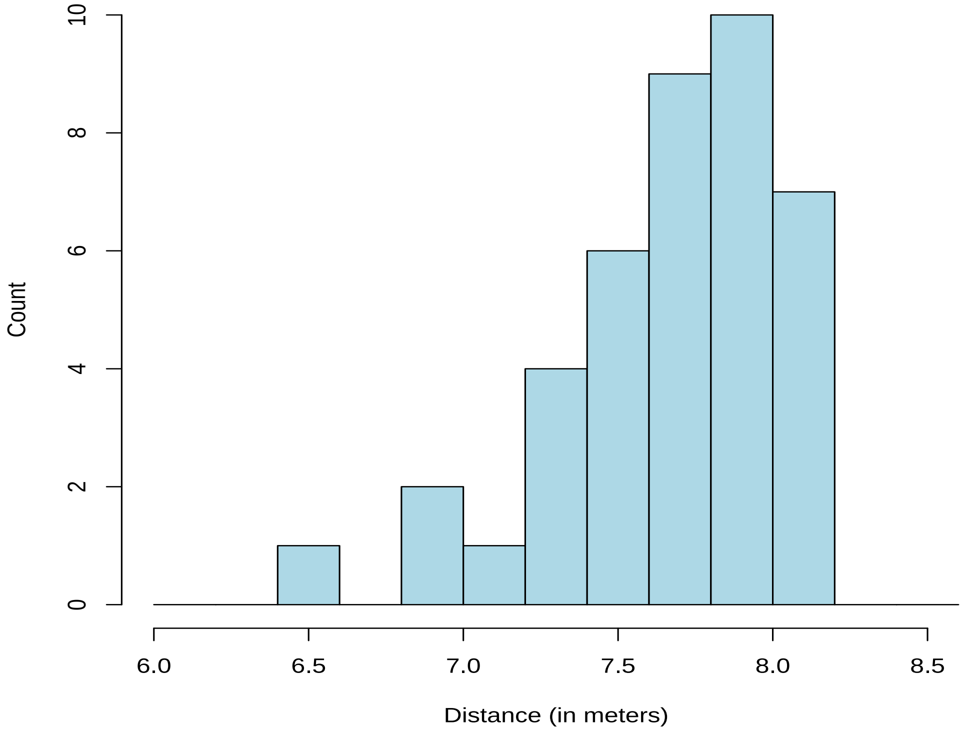


Figure 11.6: Summer Olympics data. Long-jump distances for 40 male athletes.

Example 11.3. Figure 11.6 displays the long-jump distances (in meters) for 40 male athletes participating in the 2012 Summer Olympics in London. There were 42 male athletes, but 2 were disqualified. The longest jump for each athlete is shown.

Interpretation: Here is how we would interpret the histogram in Figure 11.6 in terms of our four characteristics:

- Center: The center of the long-jump distance distribution is around 7.6 meters.
- Variability: Almost all of the long-jump distances are between 6.8 and 8.2 meters.
- Shape: This distribution has a single peak (around 7.9 meters) and is **skewed to the left**.
- Deviations: There is an outlier around 6.5 meters.

Remark: In this example, the histogram does not really correspond to a sample taken from a larger population. This histogram is constructed using all long-jump athletes who qualified for the Summer Olympics in 2012. I would argue it doesn't make sense to regard these 40 athletes as "coming from a larger population." Therefore, it is not appropriate to construct a population density curve.

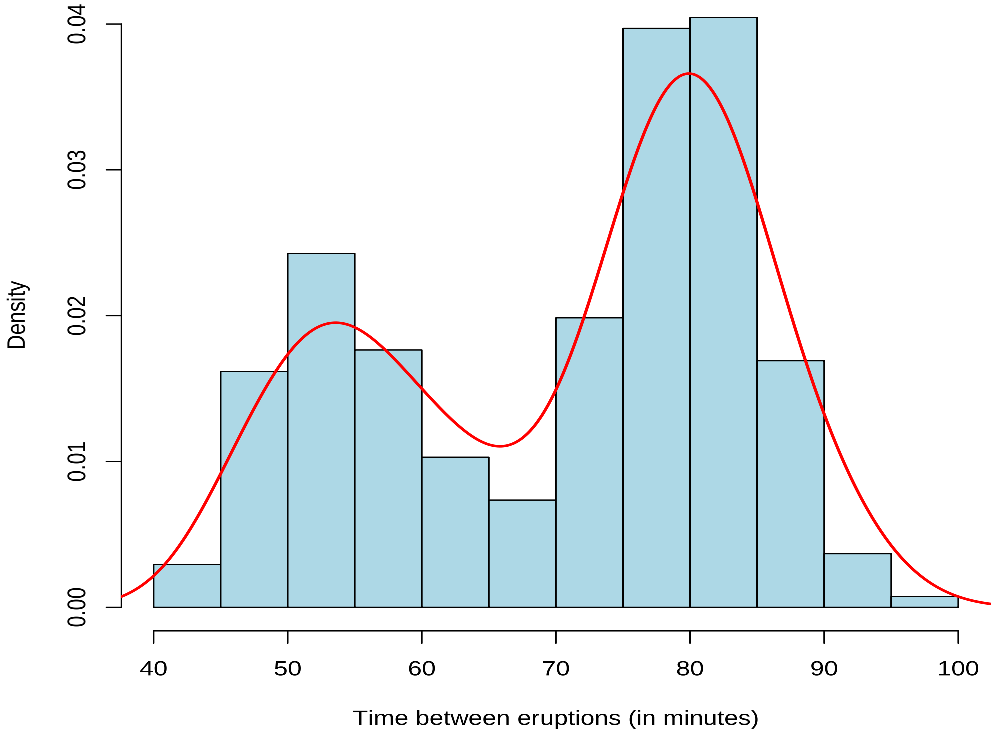


Figure 11.7: Old Faithful geyser data. The time between eruptions recorded on 272 occasions.

Example 11.4. The histogram in Figure 11.7 shows the time between eruptions (in minutes) of the Old Faithful geyser in Yellowstone National Park, USA. This time was recorded on 272 occasions.

Interpretation: Here is how we would interpret the histogram in Figure 11.7 in terms of our four characteristics:

- Shape: This distribution has a double peak—one peak around 55 minutes and another around 80 minutes. This is an example of a **bimodal distribution**.
- Center: The center of the distribution is around 70 minutes, but this assessment might be misleading because the distribution is bimodal.
- Variability: All of the times are between 40 and 100 minutes.
- Deviations: There are no obvious outliers.

Remark: What Figure 11.7 shows is that there are two different types of eruptions—one which happens more quickly and one which takes longer. There could be an excellent geological reason for this. I don't know why this would happen, but the histogram brings it to our attention.

11.3 Stemplots

Remark: For smaller data sets, a **stemplot** can be used to show the distribution of quantitative data.

- Stemplots show the distribution while retaining the numerical values in the data set. Histograms do not show the exact value of each datum (the data are grouped together in the histogram intervals).
- The idea is to separate each datum into a **stem** and a **leaf**.
  - Stems are plotted on the leftmost column.
  - Leaves are plotted on the right side in ascending order.

Example 11.5. Here are the final exam scores from an undergraduate course I taught when I was at Oklahoma State University. There were $n = 66$ students.

```
95 98 93 91 95 90 90 96 98 89 93 92 88 79 91 83 85 90 81 87 79 87
88 83 86 80 77 81 81 78 79 82 78 76 79 76 73 81 78 84 70 71 77 65
69 70 63 77 74 82 69 74 67 44 63 70 57 63 51 52 22 57 47 54 52 76
```

For these data, an obvious choice for the stem and leaf portions is

- **Stem:** 10's digit; e.g., "98" ⟶ "9"
- **Leaf:** 1's digit; e.g., "98" ⟶ "8".

```
> stem(final.exam)

  2 | 2
  3 |
  4 | 47
  5 | 122477
  6 | 3335799
  7 | 00013446667778889999
  8 | 011112233345677889
  9 | 0001123355688
```

Interpretation: Here is how we would interpret the stemplot above in terms of our four characteristics:

- Center: The center of the score distribution is somewhere between 75 and 80.
- Variability: Most of the scores are between 44 and 98.
- Shape: This distribution has a single peak (around 79) and is **skewed to the left**.
- Deviations: There is an outlier at 22.

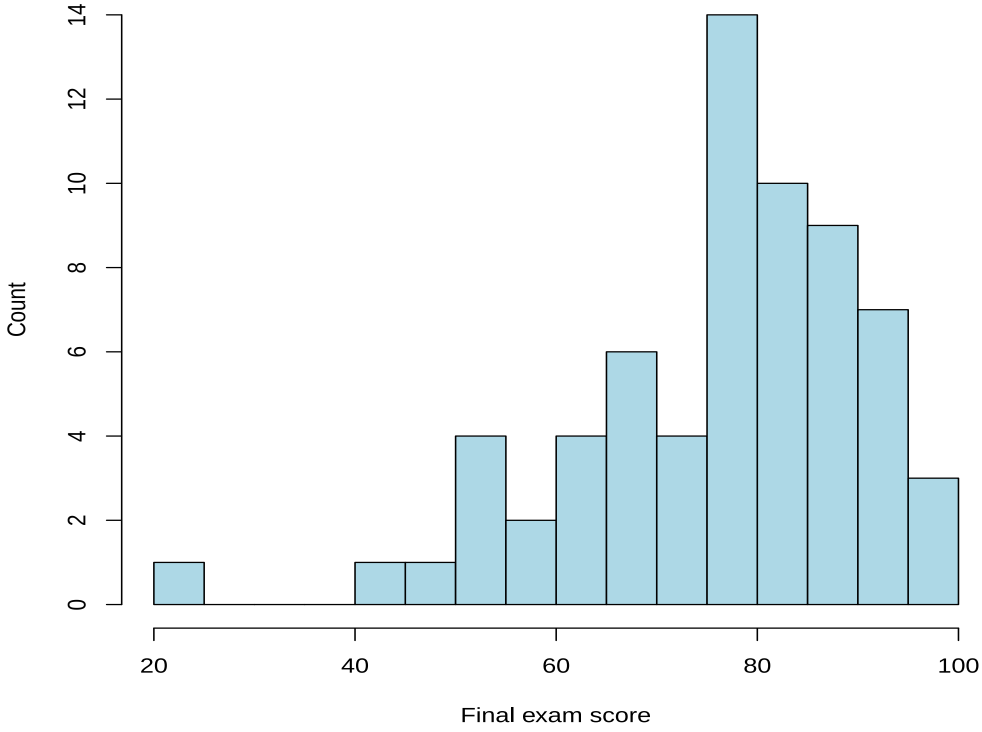


Figure 11.8: OSU data. Final exam scores for 66 undergraduate students.

Remark: It can be instructive to see the histogram and stemplot for the same data.

```
> stem(final.exam,scale=2)

  2 | 2
  2 |
  3 |
  3 |
  4 | 4
  4 | 7
  5 | 1224
  5 | 77
  6 | 333
  6 | 5799
  7 | 0001344
  7 | 6667778889999
  8 | 0111122334
  8 | 5677889
  9 | 00011233
  9 | 55688
```

12 Describing Distributions with Numbers

12.1 Introduction

Recall: In the last chapter, we learned about two graphical displays for quantitative data: histograms and stemplots. We also adopted the practice of interpreting distributions of quantitative data in terms of these four characteristics:

1. Center: Where does the center of the distribution fall approximately? You can think of this as the point where the histogram (or stemplot) would balance.
2. Variability: How much variation is in the distribution? How spread out is it? What is the range of possible observations?
3. Shape: What type of shape does the distribution have? Is it symmetric or skewed (right/left)? Does it have a single peak or multiple peaks?
4. Deviations (from the overall pattern): Are there observations that fall outside the overall pattern of the distribution? These observations are called **outliers**.

Remark: "Center" and "variability" are two important characteristics of a distribution. They are so important that we use numerical summaries to describe them. This chapter introduces these numerical summaries for quantitative data.

- We will illustrate calculations "by hand" for small data sets.
- For larger data sets, we will do our calculations in R. This is how calculations would be done in real life.

12.2 The five-number summary and boxplots

Example 12.1. Non-small cell lung cancer (NSCLC) is the most common type of lung cancer in the world, accounting for about 85% of all cases. A study in Japan examined a small group of NSCLC patients who had been treated with gefitinib and erlotinib (two cancer drugs). Here are the times until treatment failure (TTF, in months) for a sample of $n = 14$ patients:

0.8	7.5	13.4	1.4	0.5	68.9	16.1	20.4	15.6	4.2	2.4	8.2	5.3	14.0

"Treatment failure" could mean disease progression, informative dropout (e.g., experienced an adverse reaction, etc.), or death.

Ordered data: Here are the TTF data, ordered from smallest to largest. When calculating the five-number summary, you always order the data first.

$$\underbrace{0.5 \quad 0.8 \quad 1.4 \quad 2.4 \quad 4.2 \quad 5.3 \quad 7.5}_{\text{lower half}} \quad \underbrace{8.2 \quad 13.4 \quad 14.0 \quad 15.6 \quad 16.1 \quad 20.4 \quad 68.9}_{\text{upper half}}$$

Terminology: The **median** M of a data set is the middle ordered value.

- Therefore, half of the data are smaller than the median, and half of the data are larger.
- The median is one numerical summary of the "center" of a distribution.

Calculation: The median of the TTF data (above) is

$$M = \frac{7.5 + 8.2}{2} = 7.85.$$

Terminology: The **first quartile** Q_1 is the median of the lower half of a data set. The **third quartile** Q_3 is the median of the upper half.

Calculation: The quartiles of the TTF data (above) are

$$\begin{aligned} Q_1 &= 2.4 \\ Q_3 &= 15.6. \end{aligned}$$

Terminology: The **five-number summary** is a numerical summary consisting of 5 numbers:

$$\begin{aligned} \text{Minimum} &\longleftarrow \text{smallest observation} \\ Q_1 &\longleftarrow \text{first quartile} \\ M &\longleftarrow \text{median} \\ Q_3 &\longleftarrow \text{third quartile} \\ \text{Maximum} &\longleftarrow \text{largest observation.} \end{aligned}$$

Calculation: The five-number summary for the TTF data (above) is

$$\begin{aligned} \text{Minimum} &= 0.5 \\ Q_1 &= 2.4 \\ M &= 7.85 \\ Q_3 &= 15.6 \\ \text{Maximum} &= 68.9. \end{aligned}$$

Implementation in R:

```
# Enter data
TTF = c(0.8,7.5,13.4,1.4,0.5,68.9,16.1,20.4,15.6,4.2,2.4,8.2,5.3,14.0)

> sort(TTF) # order data from low to high
 [1]  0.5  0.8  1.4  2.4  4.2  5.3  7.5  8.2 13.4 14.0 15.6 16.1 20.4 68.9

> median(TTF) # median
[1] 7.85

> quantile(TTF,type=2) # 5-number summary
   0%   25%   50%   75%  100%
 0.50  2.40  7.85 15.60 68.90
```

Note: As the R output above indicates, another name for Q_1 is the 25th percentile. Another name for the median M is the 50th percentile. Another name for Q_3 is the 75th percentile. In other words,

- approximately 25% of the data are less than or equal to Q_1.
- approximately 50% of the data are less than or equal to M.
- approximately 75% of the data are less than or equal to Q_3.

Remark: The term "percentile" is commonly used when indicating a specific location on a distribution. For example, if you scored in the 75th percentile (third quartile) on an exam, then

- you did better than 75% of the students who took the exam (your score was higher than theirs)
- you did worse than 25% of the students who took the exam (they had a higher score than you).

Terminology: A **boxplot** is a graphical display for quantitative data that uses the five-number summary.

- A central box spans the quartiles (from Q_1 to Q_3).
- A solid line marks the median M.
- Lines extend from the box to the minimum and maximum observations.

The boxplot for the TTF data is shown in Figure 12.1 (next page).

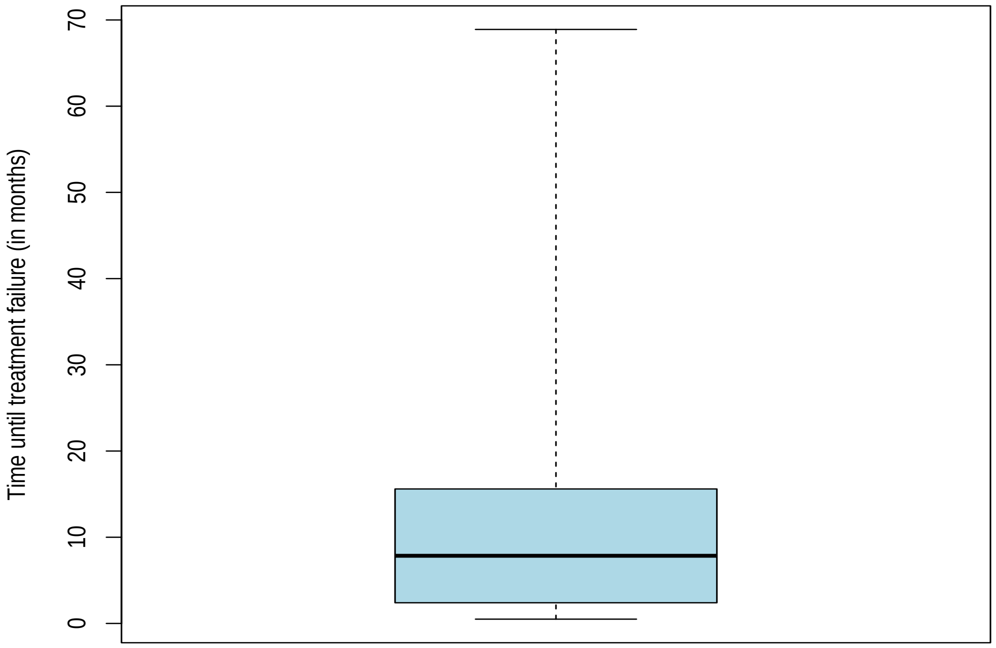


Figure 12.1: Boxplot of the TTF data in Example 12.1.

Terminology: The **interquartile range** (IQR) measures the variability (spread) in the middle 50% of a distribution of quantitative data. It is calculated as

$$\text{IQR} = Q_3 - Q_1.$$

Calculation: The IQR for the TTF data is

$$\begin{aligned} Q_1 &= 2.4 \\ Q_3 &= 15.6 \implies \text{IQR} = 15.6 - 2.4 = 13.2. \end{aligned}$$

Note: The IQR can be used to identify outliers in a data set. A common rule of thumb is to classify an observation (Obs) as an outlier if

$$\text{Obs} < Q_1 - 1.5(\text{IQR}) \quad \text{or if} \quad \text{Obs} > Q_3 + 1.5(\text{IQR}).$$

That is, if an observation is

- 1.5(IQR) **below** the first quartile ⟶ outlier on the low side
- 1.5(IQR) **above** the third quartile ⟶ outlier on the high side.

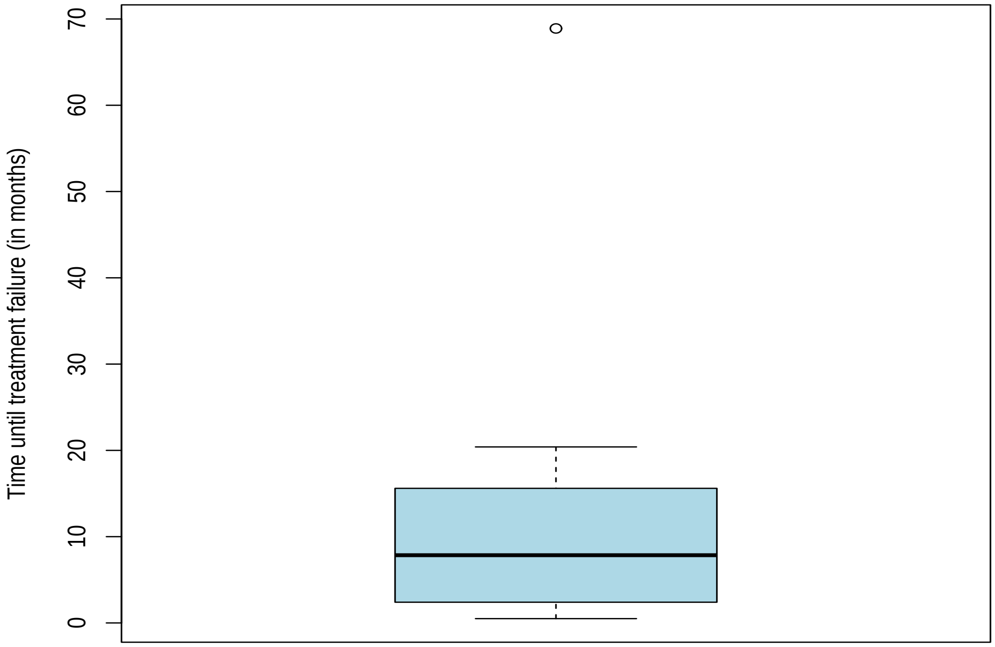


Figure 12.2: Boxplot of the TTF data in Example 12.1. The outlier "68.9" is identified.

Q: Are there any outliers in the TTF data set?
A: Let's calculate the cutoffs:

$$\begin{aligned} Q_1 - 1.5(\text{IQR}) &= 2.4 - 1.5(13.2) = -17.4 \\ Q_3 + 1.5(\text{IQR}) &= 15.6 + 1.5(13.2) = 35.4. \end{aligned}$$

Therefore,

- any TTF observation less than −17.4 months would be classified as an outlier on the low side. This cutoff doesn't make sense because TTF is positive.
- any TTF observation greater than 35.4 months would be classified as an outlier on the high side. The observation "68.9" is an outlier.

Exercise: Physicians measured the concentration of calcium (in nM) in blood samples from 15 healthy patients. Here are the data:

95 112 122 88 66 104 90 110 100 122 126 102 122 96 135

(a) Calculate the five-number summary for these data and prepare a boxplot.
(b) Are there any outliers in this data set?

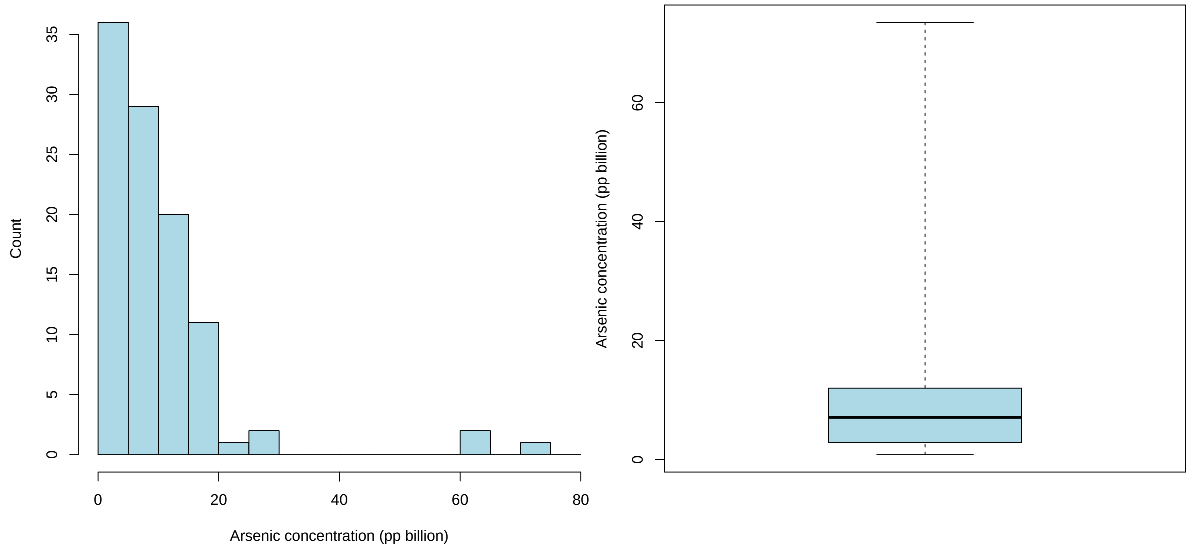


Figure 12.3: Concentration of arsenic (in ppb) in ground water for 102 wells in Texas.

Example 12.2. Arsenic (As) is a chemical element found naturally in ground water. Excessive levels may result from contamination caused by hazardous waste or industries that make or use arsenic. The histogram and boxplot in Figure 12.3 show the distribution of arsenic concentrations (in parts per billion, ppb) for a sample of $n = 102$ water wells in Texas. The data set for these observations is online.

Analysis: With a large data set like this, it is easier to do the calculations in R. The five-number summary is shown below:

```
> quantile(arsenic,type=2) # 5-number summary
  0%  25%  50%   75%  100%
 0.8  2.9  7.1  12.0  73.5
```

The median As concentration in the data set is 7.1 ppb. This is a numerical summary of the "center" of the distribution. Fifty percent of the wells in the sample have an As concentration below this value; 50% of the wells have an As concentration above.

Q: Are there any outliers in this data set?
A: Let's first calculate the interquartile range:

$$\text{IQR} = Q_3 - Q_1 = 12.0 - 2.9 = 9.1.$$

Now, let's calculate the cutoffs:

$$\begin{aligned} Q_1 - 1.5(\text{IQR}) &= 2.9 - 1.5(9.1) = -10.75 \\ Q_3 + 1.5(\text{IQR}) &= 12.0 + 1.5(9.1) = 25.65. \end{aligned}$$

Therefore,

- any As concentration less than −10.75 ppb would be classified as an outlier on the low side. Again, this cutoff doesn't make sense because As concentration is positive.

- any As concentration greater than 25.65 ppb would be classified as an outlier on the high side. There are multiple observations larger than this:

  ```
  > arsenic[arsenic>25.65]
  [1] 63.0 73.5 28.0 28.1 62.1
  ```

  These five As concentrations would be classified as outliers using our 1.5(IQR) rule of thumb.

Example 12.3. The Survey of Study Habits and Attitudes (SSHA) is a psychological test designed to measure the motivation, study habits, and attitudes toward learning of college students. A university gives the SSHA to random samples of female and male first-year students:

Female	154	109	137	115	152	140	154	178	101	
	103	126	126	137	165	165	129	200	148	
Male	108	140	114	91	180	115	126	92	169	146
	109	132	75	88	113	151	70	115	187	104

The female sample included 18 students. The male sample included 20 students. The data set for these observations is online. Which group had a larger median score?

Analysis: This is an example where we have two groups that we would like to compare (a common goal in observational studies and randomized comparative experiments). It is easy to calculate the median score for each group:

```
> median(female)
[1] 138.5
> median(male)
[1] 114.5
```

It is more interesting to compare both groups in terms of their overall distributions. This can be done visually by plotting the boxplots for each sex side-by-side; see Figure 12.4.

- Which sample has more variability (spread) in its scores?

- What is the shape of each distribution?

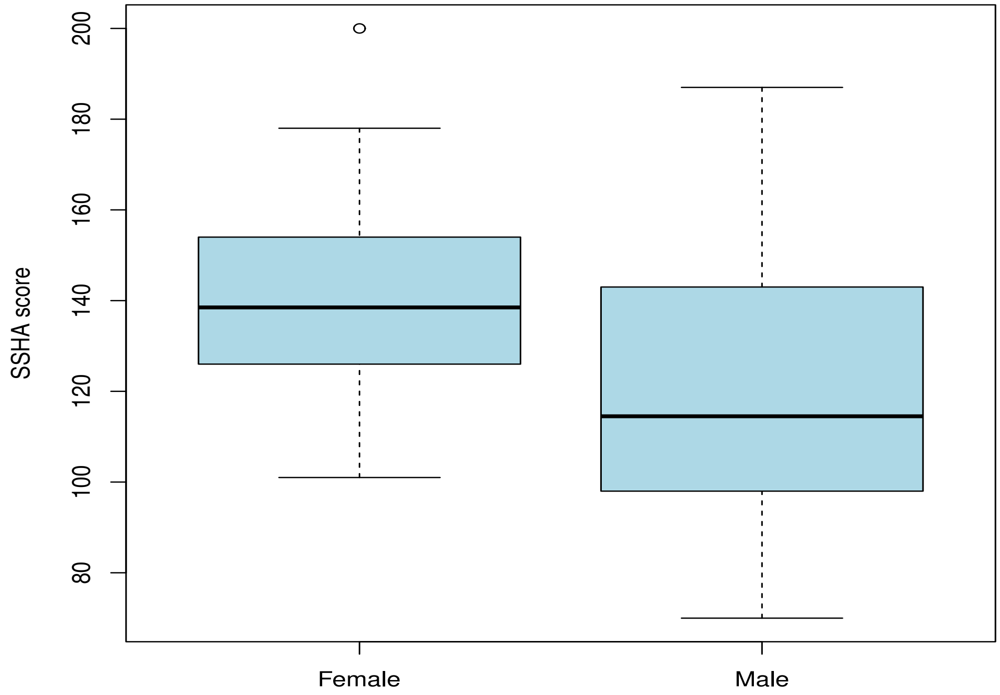


Figure 12.4: SSHA test scores for 18 female and 20 male students. The female score at 200 has been declared an outlier by the 1.5(IQR) rule.

Discussion: What do these two samples say about the populations of all female and male first-year students at this university?

- This is a question dealing with **statistical inference**.
- That is, we are trying to use the information in the samples to make a statement about the larger populations.

Remark: We can also use side-by-side boxplots for more than two groups. For example, Figure 12.5 shows the distributions of birth weights in the NEC study (in Example 1.1) for each nutrition type.

- We have already looked at the histogram of all the birth weights in Figure 11.1. Recall there were 615 infants in the study, and the histogram shows the birth weights for all of them.
- The new insight gleaned from Figure 12.5 is that we can examine the distribution of birth weight separately for each nutrition group.
- In essence, we are looking at how the distribution of one variable (birth weight) changes depending on the value another variable (nutrition type).

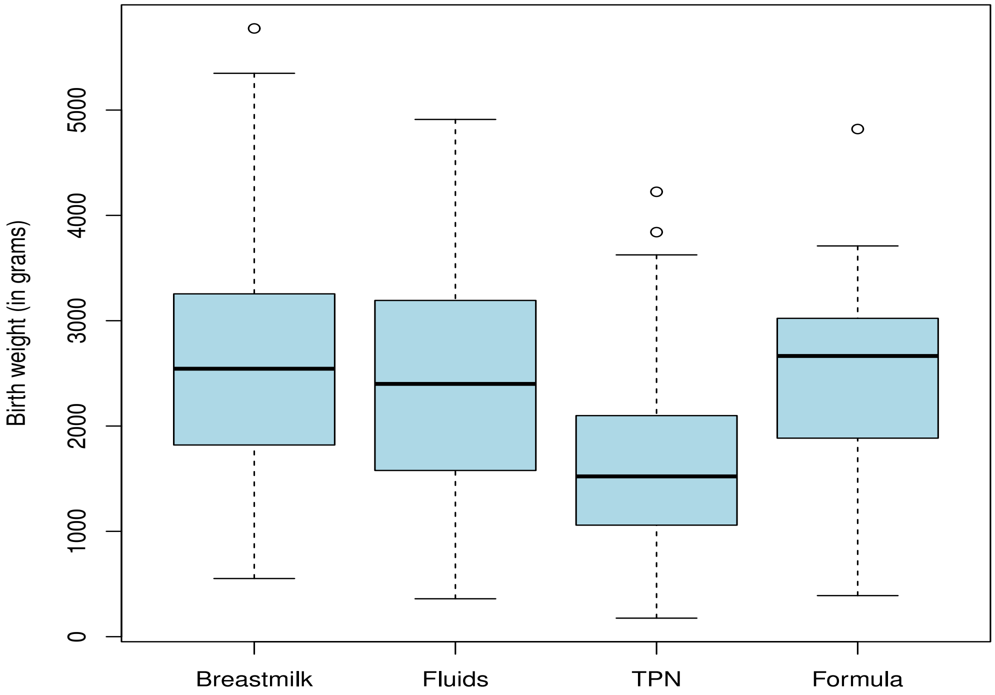


Figure 12.5: NEC data. Boxplots of birth weight (in grams) by nutrition type.

12.3 Mean and standard deviation

Definition: With a set of observations $x_1, x_2, ..., x_n$, the **mean** $\overline{x}$ (pronounced "x-bar") is calculated as

$$\overline{x} \;=\; \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}.$$

In other words, the mean is the **average** of the n observations $x_1, x_2, ..., x_n$.

- The symbol

  $$\sum$$

  is the capital Greek letter "sigma." It simply means "add."

- Physically, the mean $\overline{x}$ is the precise balancing point of a distribution of observations (e.g., it is the point where a histogram would balance).

- Like the median M, the mean $\overline{x}$ is a numerical summary of the "center" of a distribution.

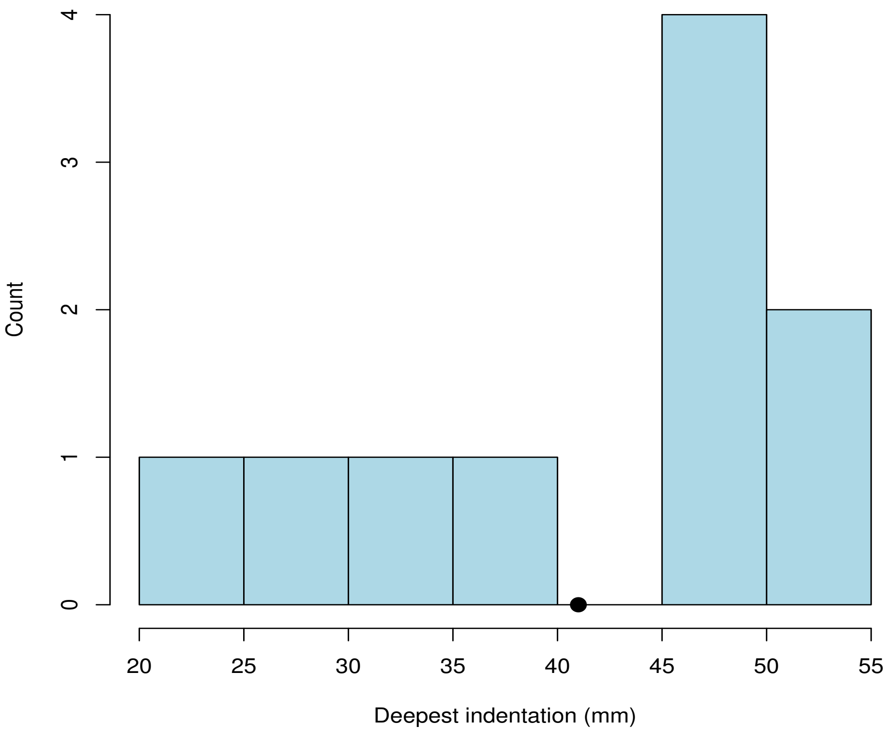


Figure 12.6: Clay model data. Depth of the deepest indentation (in mm) for a sample of $n = 10$ clay models. The mean $\overline{x} = 41.0$ is identified with a solid circle.

Example 12.4. The US Army recently commissioned a study to assess how deeply a bullet penetrates ceramic body armor. A cylindrical clay model was layered under an armor vest. Projectiles were then fired, causing indentations in the clay. The deepest indentation in the clay model was measured as an indication of survivability. Here are the deepest indentations (measured in millimeters, mm) for a sample of $n = 10$ clay models:

22.6 25.9 34.9 35.6 45.4 46.5 47.7 49.4 50.7 51.3

The sum of the observations is

$$\sum x = 22.6 + 25.9 + 34.9 + 35.6 + 45.4 + 46.5 + 47.7 + 49.4 + 50.7 + 51.3 = 410.0.$$

Therefore,

$$\overline{x} = \frac{\sum x}{10} = \frac{410.0}{10} = 41.0.$$

The mean indentation is 41.0 mm.

Implementation in R:

```
# Enter data
indentation = c(22.6,25.9,34.9,35.6,45.4,46.5,47.7,49.4,50.7,51.3)

> mean(indentation)
[1] 41
```

Definition: With a set of observations $x_1, x_2, ..., x_n$, the **standard deviation** s is

$$s = \sqrt{\frac{\sum(x-\overline{x})^2}{n-1}}.$$

- The standard deviation can be interpreted as "an average distance from the mean."
  - In other words, each observation is a certain distance from the mean. The standard deviation measures the average distance for all observations.
- Therefore, the standard deviation is a numerical summary of the variability in a distribution.
- The standard deviation is the square root of the variance. Recall we used the variance of a set of observations (on the same measurement method) to assess instrument reliability in Chapter 8.
- The table below shows how the standard deviation is calculated for the clay model data in Example 12.4; see also Figure 12.7 (next page).

Individual	Observation	Squared distance from the mean
1	22.6	$(22.6-41.0)^2 = (-18.4)^2 = 338.56$
2	25.9	$(25.9-41.0)^2 = (-15.1)^2 = 228.01$
3	34.9	$(34.9-41.0)^2 = (-6.1)^2 = 37.21$
4	35.6	$(35.6-41.0)^2 = (-5.4)^2 = 29.16$
5	45.4	$(45.4-41.0)^2 = (4.4)^2 = 19.36$
6	46.5	$(46.5-41.0)^2 = (5.5)^2 = 30.25$
7	47.7	$(47.7-41.0)^2 = (6.7)^2 = 44.89$
8	49.4	$(49.4-41.0)^2 = (8.4)^2 = 70.56$
9	50.7	$(50.7-41.0)^2 = (9.7)^2 = 94.09$
10	51.3	$(51.3-41.0)^2 = (10.3)^2 = 106.09$
		$\sum(x-\overline{x})^2 = 998.18$

Therefore,

$$s = \sqrt{\frac{\sum(x-\overline{x})^2}{n-1}} = \sqrt{\frac{998.18}{9}} \approx \sqrt{110.91} \approx 10.5.$$

The standard deviation is 10.5 mm.

Implementation in R:

```
# Enter data
indentation = c(22.6,25.9,34.9,35.6,45.4,46.5,47.7,49.4,50.7,51.3)

> sd(indentation)
[1] 10.53133
```

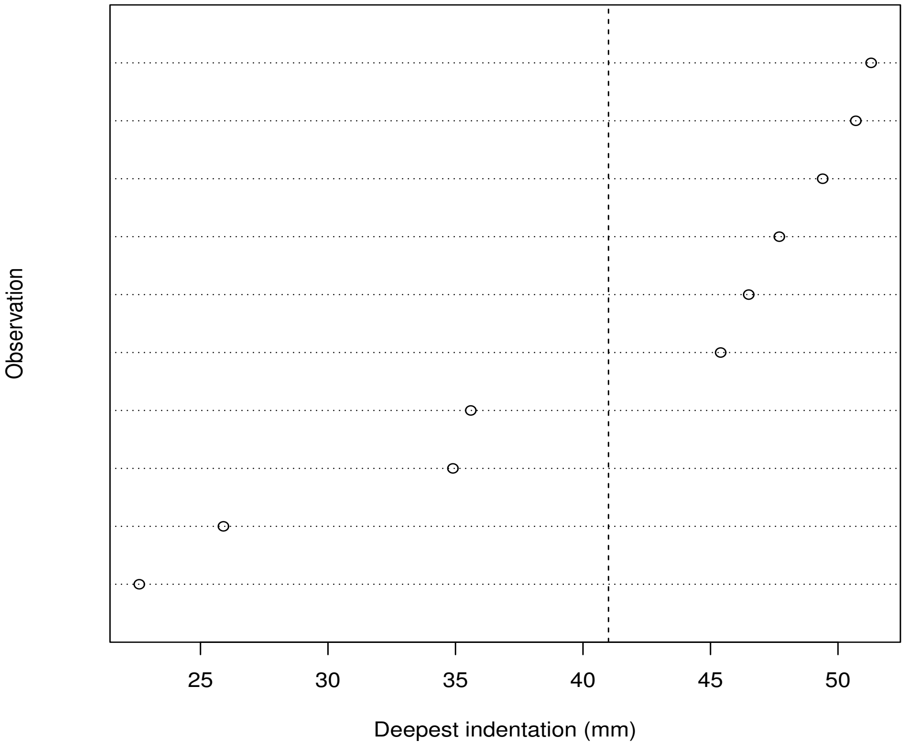


Figure 12.7: Clay model data. Each observation is shown. A dotted vertical line is shown at $\overline{x} = 41.0$.

Interpretation: Here are some things to keep in mind when interpreting the standard deviation.

- The standard deviation is a numerical summary of the **variability** in a distribution. It is the most commonly used number to do this.
  - The larger the standard deviation, the more variability in the distribution.
  - The smaller the standard deviation, the less variability in the distribution.
- **Q:** What is the smallest possible value the standard deviation can be? Look at the picture above. When is there "no variability" among the 10 observations?

- **A:** The smallest the standard deviation can be is 0. This happens when all of the observations are the same.

```
# Enter data
indentation = c(41,41,41,41,41,41,41,41,41,41)

> sd(indentation)
[1] 0
```

- The **variance** is defined as

$$s^2 = \frac{\sum(x-\overline{x})^2}{n-1}.$$

  The standard deviation is the square root of the variance.

  - The variance also measures the variability in a distribution. However, it is measured in **squared units** (not the original units of the data).
  - This is less meaningful if, for example, the data are measured in dollars, months, points, etc.

- The standard deviation is measured in the **original units** of the data.
  - For example, the clay armor indentation data in Example 12.4 are measured in millimeters (mm). The standard deviation is also measured in millimeters.

- Other statistics books introduce the variance first; then the standard deviation. Moore and Notz did use the variance to assess instrument reliability in Chapter 8, but they do not use it in this chapter.

Example 12.5. Researchers performed an experiment to compare the living environments of mice after receiving radiation. There were 181 mice in the experiment, all of which received a radiation dose of 300 r at the age of 5-6 weeks. After radiation, mice were randomized to

- **Group 1:** Conventional laboratory environment (99 mice)
- **Group 2:** Germ-free environment (82 mice).

The response variable measured on each mouse was the time until death (in days). Figure 12.8 shows side-by-side boxplots of the distributions of death times for both environments. The data set for these observations is online.

(a) Which environment has a larger mean time until death?
(b) Which environment has a larger standard deviation?

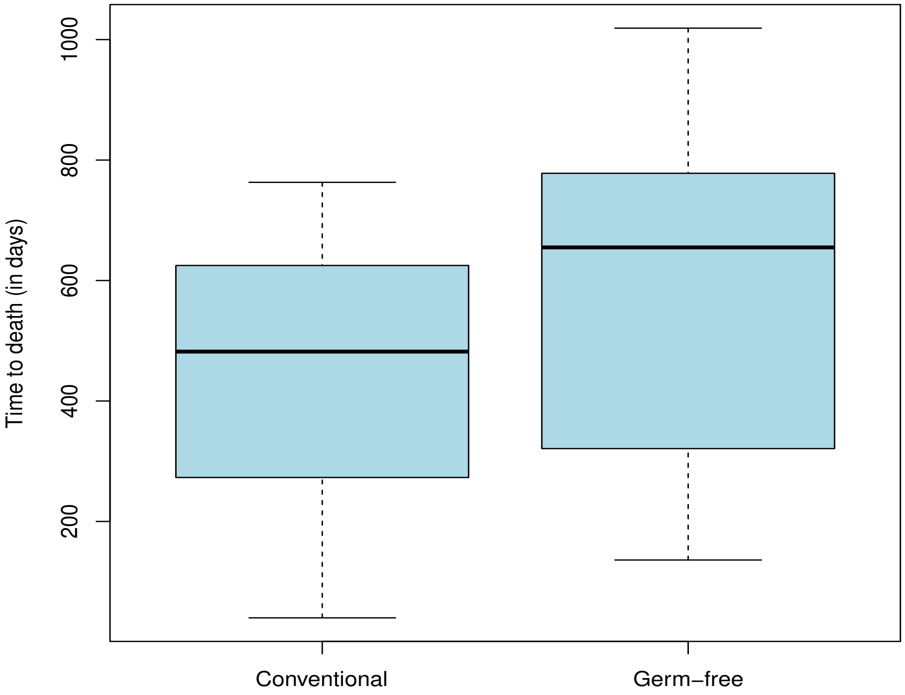


Figure 12.8: Mice data. Time to death (in days) for different environments.

We perform all the calculations in R. Here are the means for both environments:

```
> mean(conventional)
[1] 456.596
> mean(germ.free)
[1] 582.561
```

Therefore, the germ-free environment has a larger mean time to death. This makes sense; the mice live longer on average in a sterile environment. Here are the standard deviations for both environments:

```
> sd(conventional)
[1] 195.9897
> sd(germ.free)
[1] 254.4978
```

Therefore, the times until death are more variable for the germ-free environment than for the conventional environment.

12.4 Choosing numerical descriptions

Summary: We have talked about two numerical descriptions of distributions of quantitative data: (a) the five-number summary and (b) the mean and standard deviation.

Q: Which numerical description should we use?
A: It depends on what the shape of the distribution is. It also depends on if there are any outliers in the data set.

Discussion: Let's revisit the TTF data in Example 12.1. I have ordered the data from low to high:

0.5 0.8 1.4 2.4 4.2 5.3 7.5 8.2 13.4 14.0 15.6 16.1 20.4 68.9

Here are the median and mean for the data:

```
> median(TTF)
[1] 7.85
> mean(TTF)
[1] 12.76
```

Q: Why is the mean larger?
A: It is largely due to the outlier 68.9.

Let's remove the outlier and calculate the median and mean again:

0.5 0.8 1.4 2.4 4.2 5.3 7.5 8.2 13.4 14.0 15.6 16.1 20.4

Here are the median and mean now:

```
> median(TTF)
[1] 7.5
> mean(TTF)
[1] 8.45
```

Remark: Outliers can have a big impact on the value of the mean.

- Outliers on the high side will increase the mean.
- Outliers on the low side will decrease the mean.

On the other hand, outliers will usually have a small impact on where the median is (if it has an impact at all). This makes sense because the median ignores the extreme observations on each side. *We say that the mean's value is sensitive to outliers.*

Illustration: Suppose I record the annual starting salary (in \$1000s) for 5 USC graduates:

44 47 52 58 61

The median is 52. If the starting salaries were

44 47 52 58 350

the median would still be 52. *We say that the median's value is resistant to outliers.*

Observation: The value of the standard deviation s is also sensitive to outliers. This makes sense because one uses the mean $\overline{x}$ to calculate the standard deviation. If the mean is sensitive to outliers, the standard deviation's value will be too. Here are the values of the standard deviation for the two data sets above:

```
> salary = c(44,47,52,58,61)
> round(sd(salary),1)
[1] 7.2

> salary = c(44,47,52,58,350)
> round(sd(salary),1)
[1] 134.2
```

The outlier "350" has increased the standard deviation substantially.

Q: What's the main point?
A: When we provide numerical descriptions of "center" and "variability," we should choose the description that accounts for the shape of the distribution and is resistant to outliers. This means

- when the shape of the distribution is approximately symmetric with no outliers, use the mean as a description of the "center" and use the standard deviation as a description of the "variability."
- when the shape of the distribution is heavily skewed and/or has extreme outliers, use descriptions in the five-number summary:
  - use the median as a description of the "center"
  - use the IQR as a description of the "variability" (recall IQR describes the variation in the middle 50% of the distribution).
- Always graph your data first to see what the shape of the distribution is and to determine if any outliers are present. This will inform you which numerical description to use.

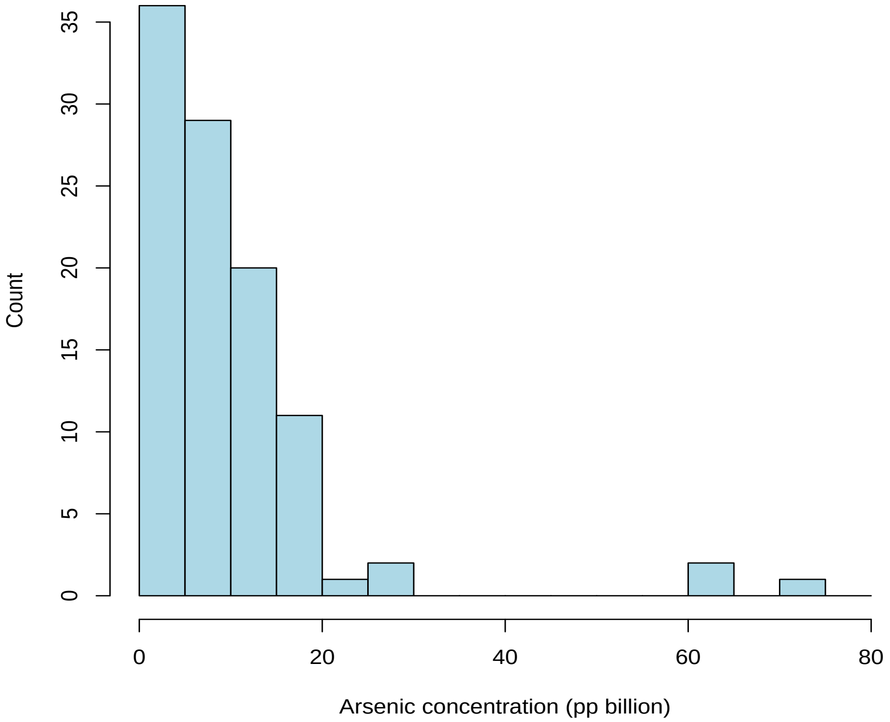


Figure 12.9: Concentration of arsenic (in ppb) in ground water for 102 wells in Texas.

Example 12.2 (continued). Figure 12.9 shows the distribution of arsenic (As) concentrations (in ppb) for a sample of $n = 102$ water wells in Texas. Which numerical description should we use to measure the "center?" Here are the median and mean:

```
> median(arsenic)
[1] 7.1
> mean(arsenic)
[1] 9.7
```

This distribution is heavily skewed right and has extreme outliers. The median is a better numerical description of the "center" of the distribution. What about "variability?" The standard deviation is

```
> sd(arsenic)
[1] 11.5
```

However, we know the standard deviation will be sensitive to the outliers on the high side (i.e., they will inflate its value). It is better to report the IQR as a description of the variation of the middle 50% of the distribution:

$$\text{IQR} = Q_3 - Q_1 = 12.0 - 2.9 = 9.1.$$

The IQR's value is not influenced by the outliers (i.e., it is resistant to them).

13 Normal Distributions

13.1 Introduction

Example 13.1. A biologist is studying green sea turtles inhabiting the Grand Cayman Islands. One variable of interest is the length of the turtle's curved shell. He catches a sample of $n = 76$ turtles and measures the length of each shell (in cm). A histogram of these observations is shown below. The data set for these observations is online.

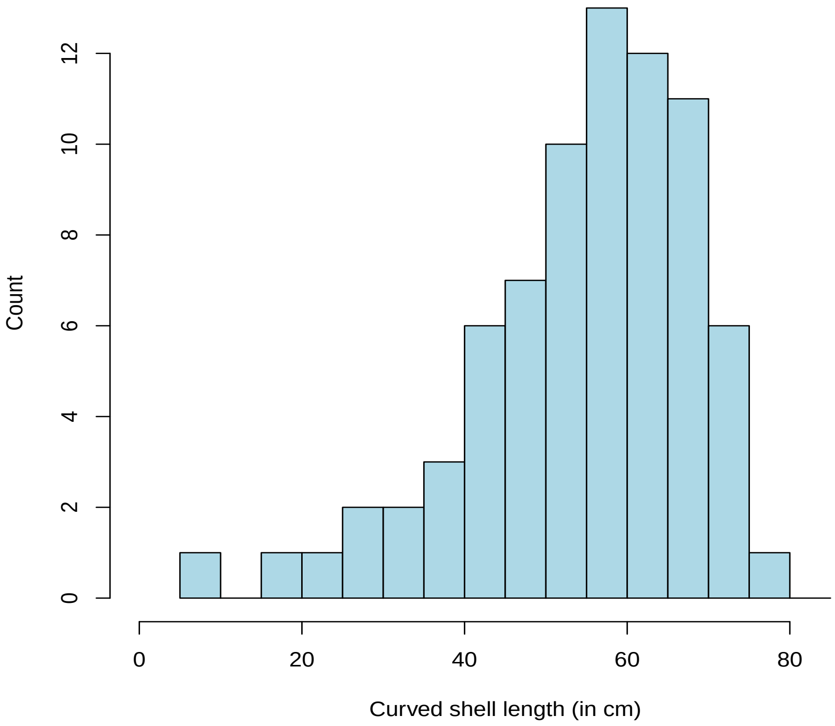


Figure 13.1: Turtle data. Shell lengths for a sample of $n = 76$ green sea turtles.

Review: We are accustomed to describing four characteristics of a quantitative distribution (center, variability, shape, and deviations). From the last chapter, we now have numerical summaries that can assist us with this description.

1. We have two numerical summaries of the center of the distribution:

```
> mean(shell)
[1] 54.54
> median(shell)
[1] 56.79
```

The median is a better measure of the center here because the distribution is skewed to the left and there are possible outliers.

2. We have two numerical summaries of the variability of the distribution:

```
> quantile(shell,type=2)
   0%    25%    50%    75%   100%
 9.34  47.61  56.79  64.85  77.34
> sd(shell)
[1] 13.68
```

- The middle 50% of shell length distribution is between 47.61 and 64.85 cm. The IQR is
$$\text{IQR} = Q_3 - Q_1 = 64.85 - 47.61 = 17.24 \text{ cm.}$$
- The standard deviation gives the average distance of each observation to the mean. The average distance here is 13.68 cm.

3. Overall shape: This distribution has a single peak slightly below 60 cm and is skewed to the left.

4. Let's calculate the cutoffs for outlier detection:

$$\begin{aligned} Q_1 - 1.5(\text{IQR}) &= 47.61 - 1.5(17.24) = 21.75 \\ Q_3 + 1.5(\text{IQR}) &= 64.85 + 1.5(17.24) = 90.71. \end{aligned}$$

Therefore,

- any shell length less than 21.75 cm would be classified as an outlier on the low side. There are two turtles in the sample satisfying this:

```
> shell[shell<21.75]
[1] 17.65  9.34
```

- any shell length greater than 90.71 cm would be classified as an outlier on the high side. There are no outliers on the high side.

Important: The histogram in Figure 13.1 shows the distribution of the shell lengths for the 76 turtles in the **sample**.

- Recall: A population density curve is a smooth curve that approximates the histogram; see Figure 13.2 (next page).
- What is a reasonable description of the **population** in this example?
- If the sample of turtles is representative of this population, then the population density curve serves as a mathematical model for the distribution of shell lengths for all turtles in the population.

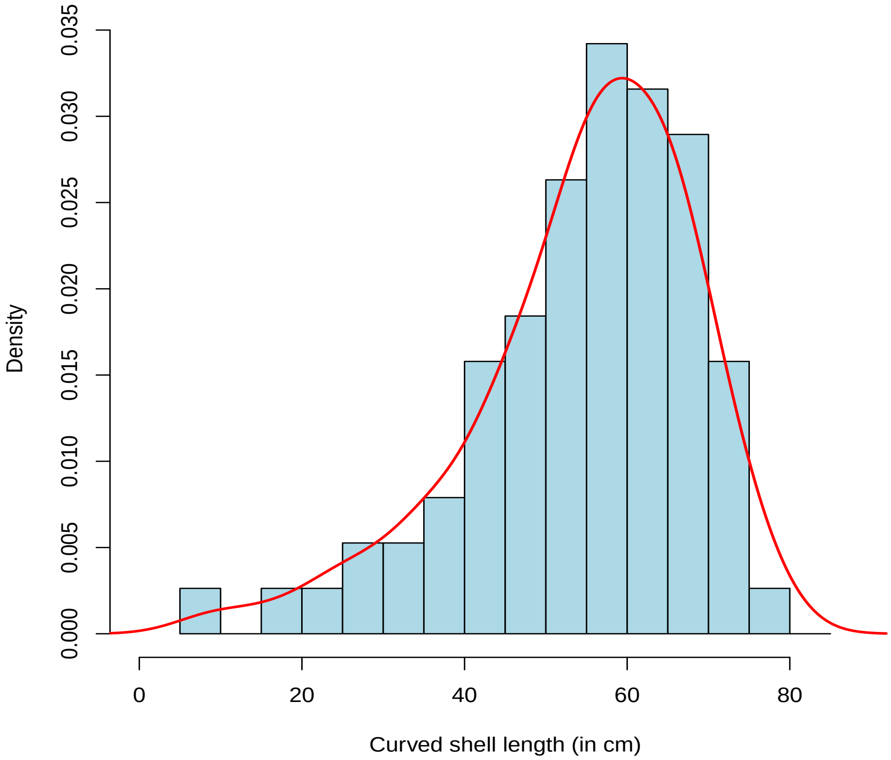


Figure 13.2: Turtle data. Shell lengths for a sample of $n = 76$ green sea turtles. An estimate of the population density curve has been added.

13.2 Population density curves

Terminology: Population density curves are curves that describe the distribution of a quantitative variable for all individuals in a population. Any population density curve has these three properties:

1. the curve is non-negative (i.e., it must reside above the horizontal axis).
2. the total area under the curve is 1 (or 100%).
3. the **area** under the curve over a given range represents the **proportion** of individuals in the population that fall in that range.
   - You can also interpret this area as the **probability** a single individual in the population falls in that range.
   - We will discuss probability in Chapters 17-18.

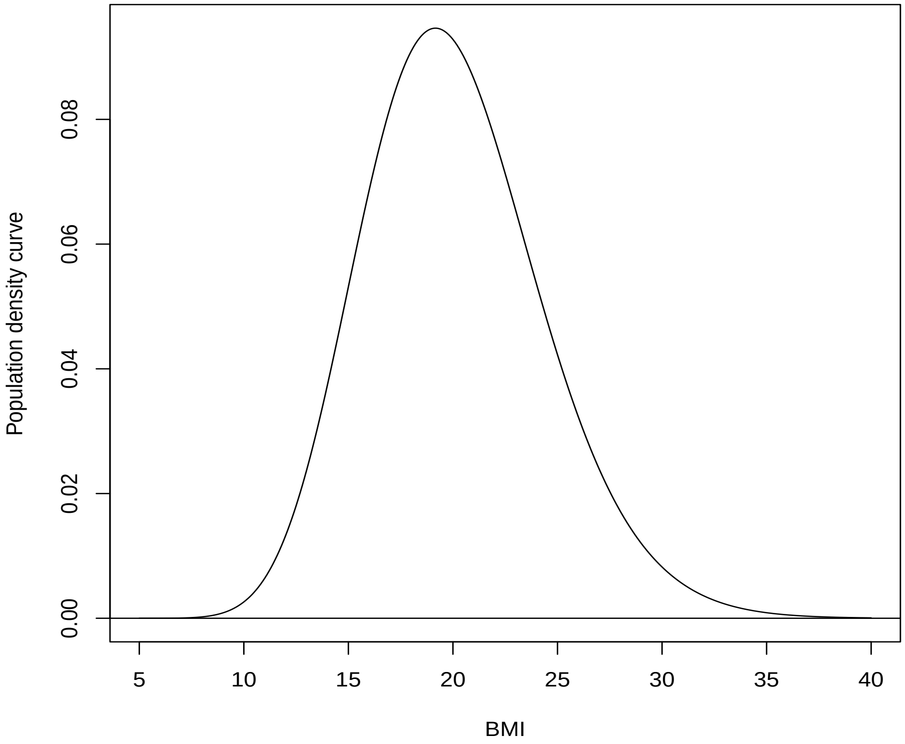


Figure 13.3: Population density curve for the BMI of fourth-grade children in Augusta, GA.

Example 13.2. Figure 13.3 shows the population density curve for the BMI of fourth-grade children in Augusta, GA. The CDC guidelines for interpreting BMI for 10-year-old children (the approximate age of those in fourth grade) are given below:

$\leq$ 14.5: Underweight; 14.5-19.5: Healthy; 19.5-21.5: Overweight; $\geq$ 21.5: Obese.

Q: What proportion of the population falls in each category?
A: We determine this by finding the **area** under the curve. See Figure 13.4 (next page).

- the proportion of fourth-grade children in Augusta who are underweight is 0.085.
- the proportion of fourth-grade children in Augusta who are at a healthy weight is 0.389.
- the proportion of fourth-grade children in Augusta who are overweight is 0.179.
- the proportion of fourth-grade children in Augusta who are obese is 0.347.

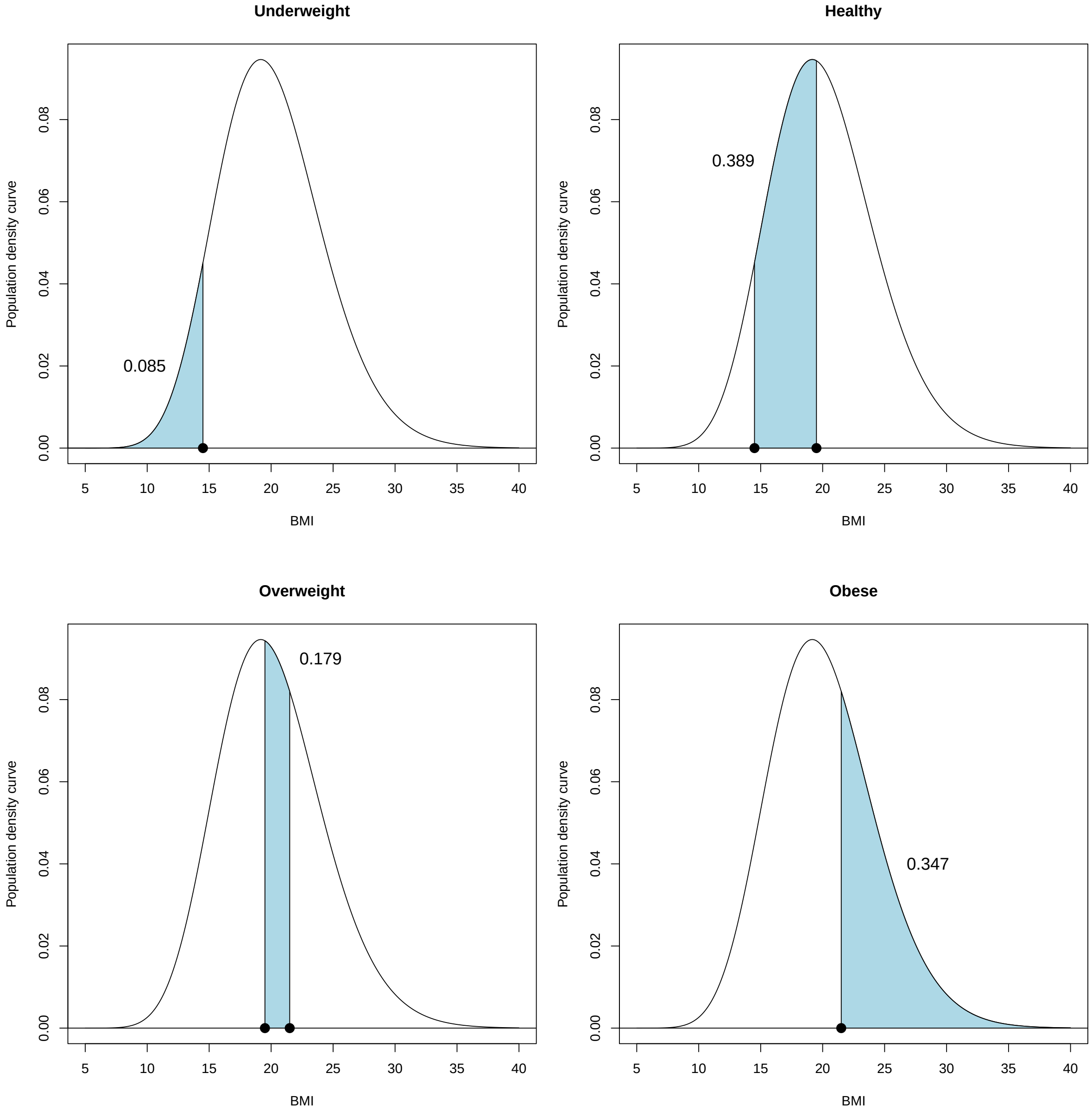


Figure 13.4: Population density curve for the BMI of fourth-grade children in Augusta, GA. The proportions of underweight, healthy, overweight, and obese students are shown. Note that the proportions (areas) add to 1. R was used to determine these areas.

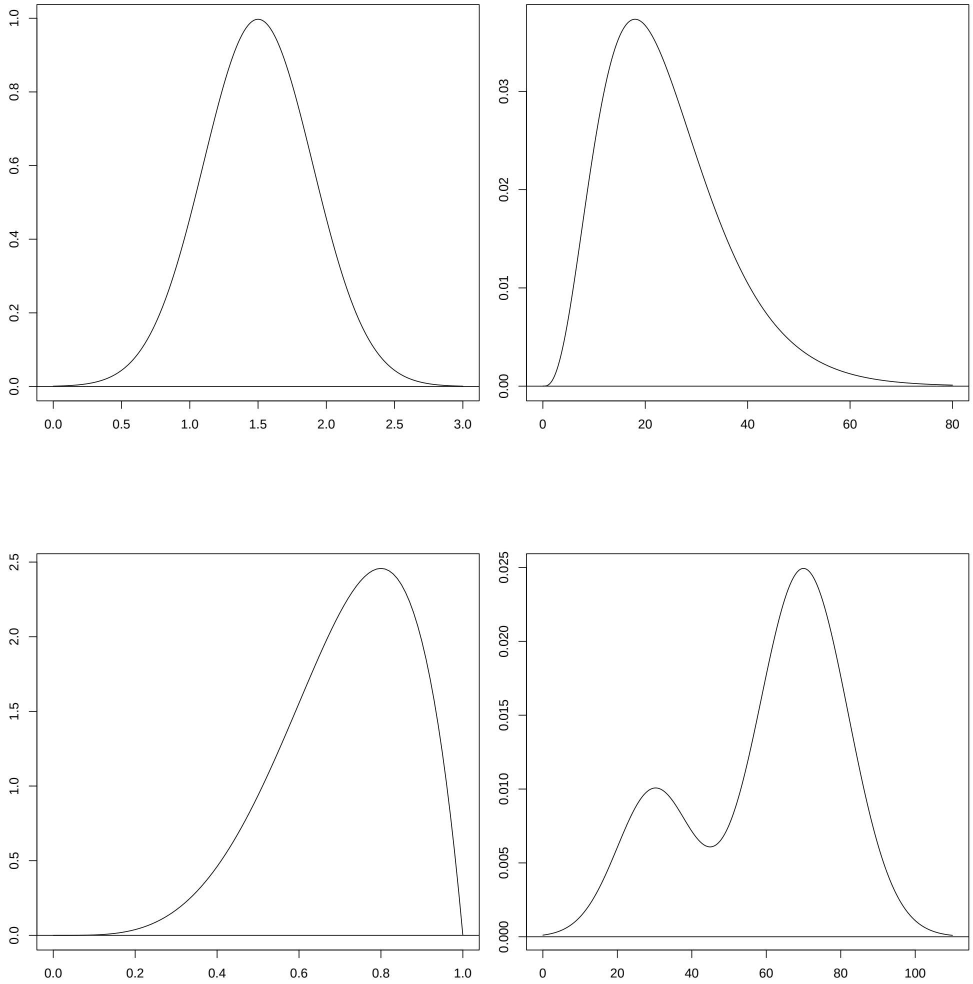
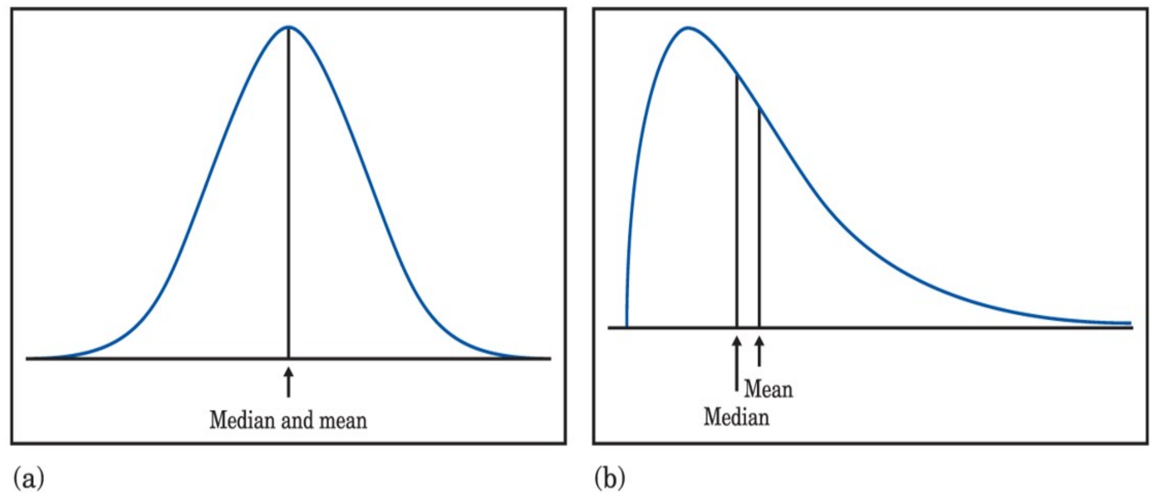


Figure 13.5: Examples of density curves for quantitative variables. The horizontal axis gives the range of possible values for the variable.

Remark: Just like histograms, population density curves come in all shapes. Figure 13.5 shows population density curves that are symmetric, skewed, and bimodal.

- The **mean** for a population density curve is the point on the horizontal axis where the curve would balance; think of this as "the center of gravity."
- The **median** for a population density curve is the point on the horizontal axis which divides the area under the curve in half; i.e., 50% of the area is above the median; 50% is below.
- The **mode** for a population density curve is the point on the horizontal axis where the curve is at its highest value.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Q: For population density curves, how do the mean and median compare?
A: It depends on the shape of the distribution. If the population density curve is

- **symmetric** ⟶ mean and median are equal
- **skewed right** ⟶ mean is greater than the median
  - outliers on the high side increase the value of the mean
- **skewed left** ⟶ mean is less than the median
  - outliers on the low side decrease the value of the mean.

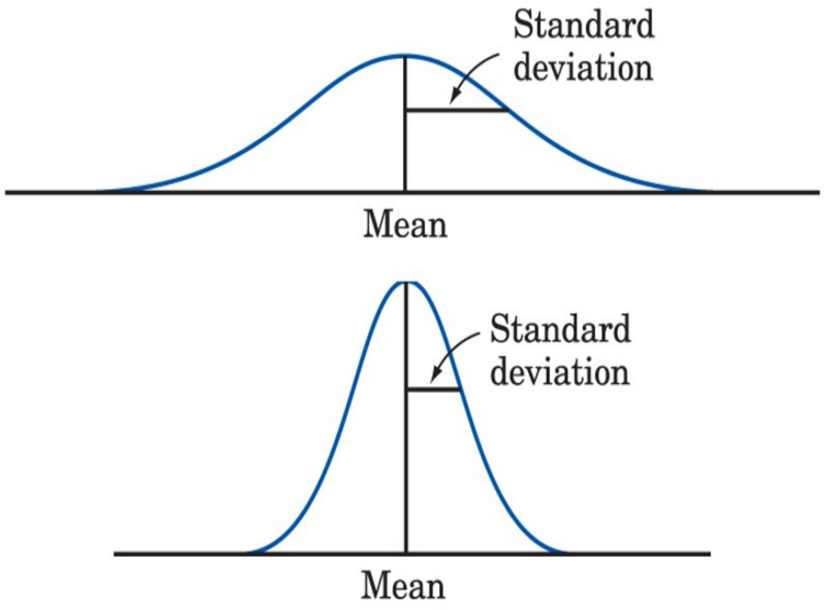
Recall: The value of the mean is sensitive to unusually high or low observations (the median is not). Therefore, the mean will move in the direction of these observations.

Notation: The notions of "center" and "variation" remain important with population density curves. We use the following notation:

- The **mean** for a population density curve is denoted by μ.
- The **standard deviation** for a population density curve is denoted by σ.

The table below summarizes the differences in notation for "sample" and "population:"

	Sample	Population
	Histogram	Density curve
Mean	$\overline{x}$	μ
Standard deviation	s	σ



Moore/Notz, *Statistics: Concepts and Controversies*, 10e,
© 2020 W. H. Freeman and Company

Q: Why do we use different notation for histograms and population density curves?
A: Histograms show the distribution for a sample of individuals. Population density curves describe all individuals in the population.

- The sample mean $\overline{x}$ and sample standard deviation s are calculated from the sample of individuals. They are **statistics**.
- The population mean μ and population standard deviation σ describe the entire population. They are **parameters**.

13.3 Normal distributions

Remark: The **normal distribution** is the most common population density curve. A normal distribution is described by two numbers:

- the population mean μ
- the population standard deviation σ.

The mean measures where the center is. The standard deviation measures how "spread out" the distribution is. Normal distributions are symmetric and have a single peak in the middle. They have a "bell-shaped" appearance.

Mathematics: There is a function $f(x)$ that describes the normal distribution mathematically. It is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\ e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

The good news is that we will never use this formula. However, it is insightful to note that the normal density curve is a **function** and has a specific mathematical form. If nothing else, we see the function depends on the mean μ and the standard deviation σ.

Recall: For any population density curve,

- the **area** under the curve over a given range represents the **proportion** of individuals in the population that fall in that range.

Q: How do we find areas under the normal density curve?
A: This is what we investigate the rest of this chapter.

13.3.1 Standard scores

Terminology: The **standard score** for any observation x is calculated as follows:

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{x-\mu}{\sigma}.$$

Here are some facts about standard scores:

- A standard score z indicates how many standard deviations an observation x falls above or below the mean μ.
  - If a standard score $z < 0$, then x is below the mean μ.
  - If a standard score $z > 0$, then x is above the mean μ.
- Standard scores are **unitless** (i.e., they have no units attached to them). This makes standard scores useful if we want to compare observations from different populations, as shown in the next example.

Example 13.3. SAT mathematics scores follow a normal distribution with mean 500 and standard deviation 100. ACT mathematics scores follow a normal distribution with mean 18 and standard deviation 6.

- Steve scored a 650 on the SAT mathematics exam.
- David scored a 21 on the ACT mathematics exam.

Q: Who did better?

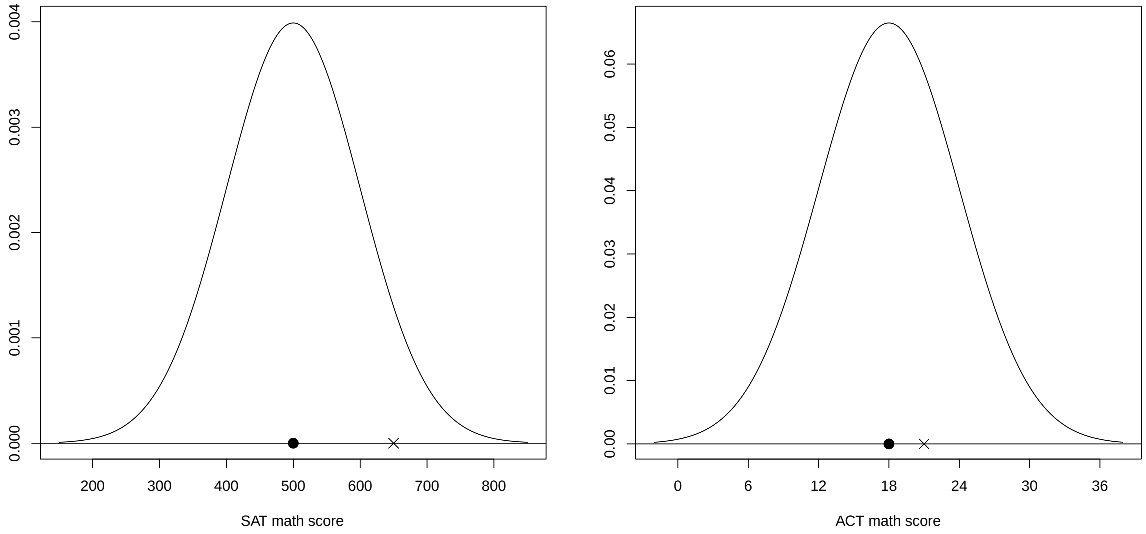


Figure 13.6: SAT and ACT mathematics score distributions. Left: SAT. Right: ACT. In each distribution, the population mean is identified with a solid circle. The symbol "×" identifies the student's exam score.

A: We can answer this by calculating each student's standard score:

- Steve's standard score is

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{650 - 500}{100} = 1.5.$$

  Interpretation: Steve's SAT score is 1.5 standard deviations above the mean.

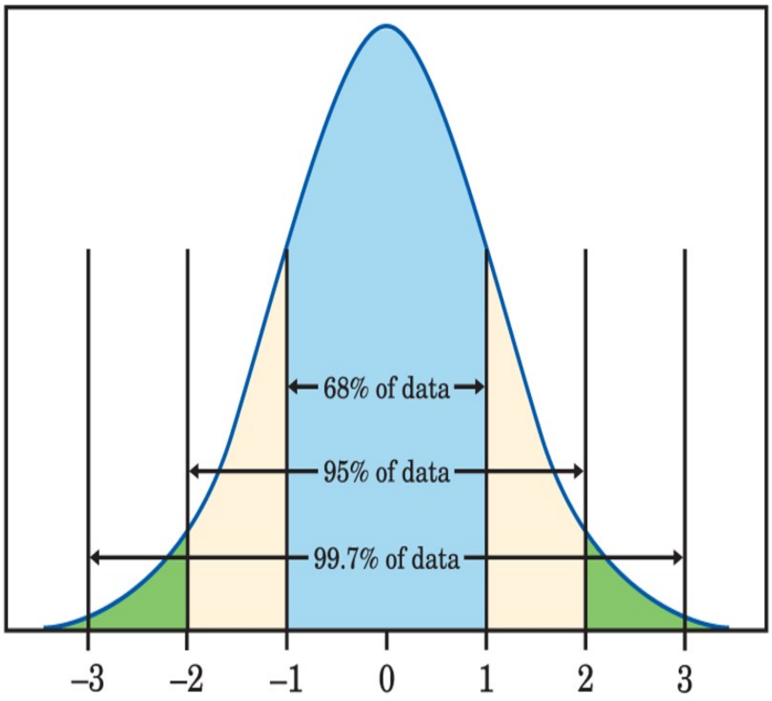
- David's standard score is

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{21 - 18}{6} = 0.5.$$

  Interpretation: David's ACT score is 0.5 standard deviations above the mean.

- Both standard scores are positive, so Steve and David both did better than the mean. However, Steve did better than David because his standard score is larger. See Figure 13.6 above.

Exercise: Alex got a 400 on the SAT mathematics exam. What is his standard score? Ben got a 18 on the ACT mathematics exam. What is his standard score? Write an interpretation of each standard score. Who did better?



Moore/Notz, *Statistics: Concepts and Controversies*, 10e,
© 2020 W. H. Freeman and Company

Figure 13.7: Normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

13.3.2 68-95-99.7 Rule

68-95-99.7 Rule: For any normal distribution, approximately

- 68% of the observations will be within **one** standard deviation of the mean
- 95% of the observations will be within **two** standard deviations of the mean
- 99.7% (or almost all) of the observations will be within **three** standard deviations of the mean.

Remark: In any normal distribution, it would be very unusual for an observation to be more than 3 standard deviations away from the mean (i.e., either 3 standard deviations below the mean or 3 standard deviations above).

- Translation: Standard scores $z < -3$ or $z > 3$ are highly unlikely.
- From the 68-95-99.7 Rule, only about 0.3% of the observations will be more than 3 standard deviations from the mean. As a decimal, this is 0.003 (or 3 out of 1000).

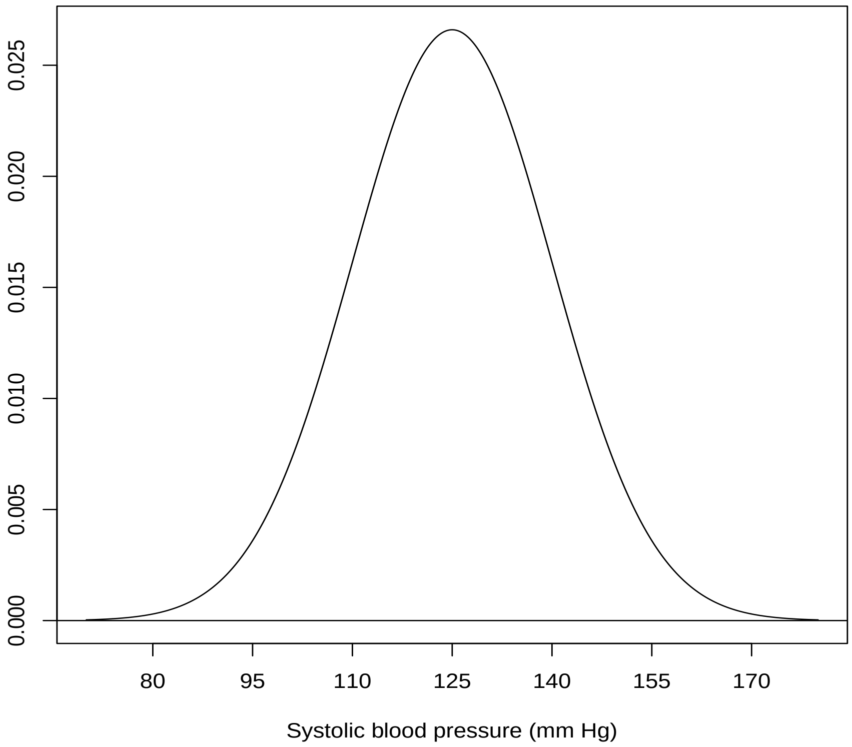


Figure 13.8: Normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 15$.

Example 13.4. The World Health Organization uses a normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 15$ to describe the systolic blood pressure (SBP, measured in mm Hg) of American males (aged 18 and over). This population density curve is shown above.

(a) Form intervals 1, 2, and 3 standard deviations from the mean. Interpret.
Solution: Our calculations are:

$$\begin{array}{rlcl}
\text{One standard deviation:} & \mu - \sigma = 125 - 15 = 110 & \text{and} & \mu + \sigma = 125 + 15 = 140 \\
\text{Two standard deviations:} & \mu - 2\sigma = 125 - 30 = 95 & \text{and} & \mu + 2\sigma = 125 + 30 = 155 \\
\text{Three standard deviations:} & \mu - 3\sigma = 125 - 45 = 80 & \text{and} & \mu + 3\sigma = 125 + 45 = 170
\end{array}$$

These intervals are interpreted as follows:

- About 68% of American males have SBP between 110 and 140 mm Hg.
- About 95% of American males have SBP between 95 and 155 mm Hg.
- About 99.7% of American males have SBP between 80 and 170 mm Hg.

In-class exercises: Use the 68-95-99.7 Rule. Draw a picture in each part.

(b) What percentage of American males have a SBP of 95 mm Hg or lower?

(c) What percentage of American males have a SBP of 140 mm Hg or higher? This is regarded as “high blood pressure.”

(d) What percentage of American males have a SBP between 95 and 140 mm Hg?

13.3.3 Finding any area under the normal density curve

Remark: In Example 13.4, we used the 68-95-99.7 Rule to determine what areas were needed in parts (b), (c), and (d); see the previous page. We could do this because

- in part (b), 95 mm Hg is exactly two standard deviations below the mean; i.e.,

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{95 - 125}{15} = -2.$$

- in part (c), 140 mm Hg is exactly one standard deviation above the mean; i.e.,

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{140 - 125}{15} = 1.$$

Important: When the standard scores are ±1, ±2, or ±3, we can use the 68-95-99.7 Rule to determine areas under the normal density curve.

Q: What percentage of American males have SBP below 100 mm Hg?
A: The standard score of 100 mm Hg is

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{100 - 125}{15} \approx -1.7.$$

Therefore, we cannot use the 68-95-99.7 Rule to answer this question! We need additional assistance−either from R or from Table B in Moore and Notz.

Remark: As the example above shows, we might want to do a normal distribution calculation with standard scores that are not ±1, ±2, or ±3. How do we do this?

Calculating normal areas using R:

- To find the **area to the left** of an observation x, do the following:
  1. Calculate the standard score z.
  2. Use the R command `pnorm(z)`.
- To find the **area to the right**, use `1-pnorm(z)` instead.

Note: For the standard score $z = -1.7$ (above), we have

```
> pnorm(-1.7)
[1] 0.0446
```

Interpretation: About 4.46% of American males have SBP below 100 mm Hg. See Figure 13.9 (next page).

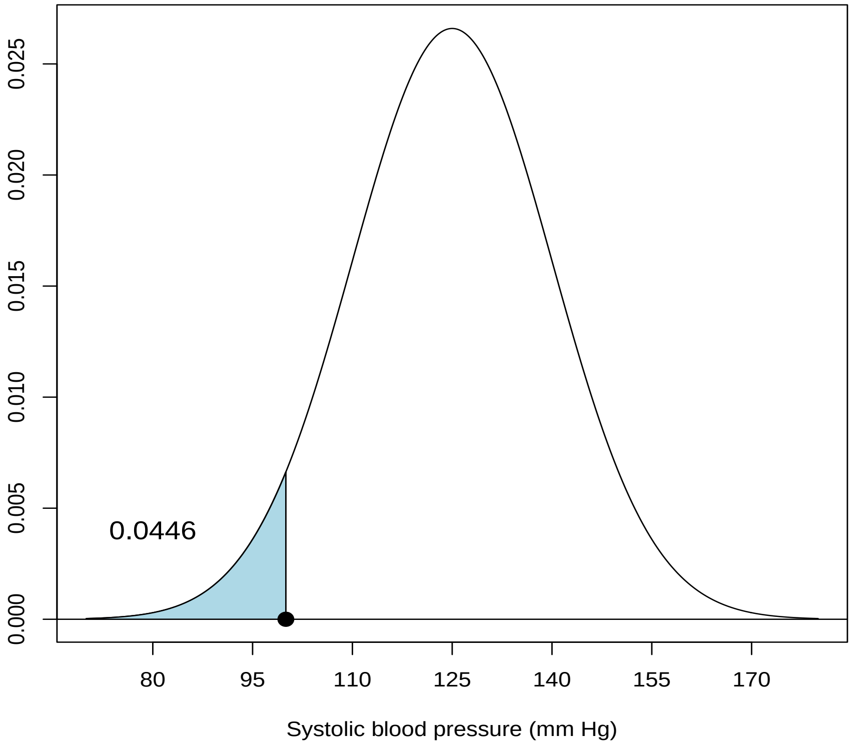


Figure 13.9: Normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 15$. The area to the left of 100 mm Hg is shown shaded.

Calculating normal areas using Moore and Notz's Table B (see next page):

- To find the **area to the left** of an observation x, do the following:
  1. Calculate the standard score z.
  2. Find the standard score z on Table B and read off the percentage.
  3. Table B gives percentages. To convert the percentage to a decimal, divide by 100.
- To find the **area to the right**, take the percentage listed and subtract it from 100%.

Q: What percentage of American males have SBP above 130 mm Hg? Use Table B.
A: The standard score of 130 mm Hg is

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{130 - 125}{15} \approx 0.3.$$

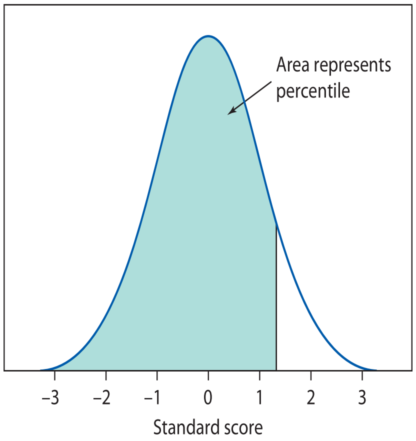


Table B Percentiles of the Normal distributions

Standard score	→ Percentile	Standard score	→ Percentile	Standard score	→ Percentile
−3.4	0.03	−1.1	13.57	1.2	88.49
−3.3	0.05	−1.0	15.87	1.3	90.32
−3.2	0.07	−0.9	18.41	1.4	91.92
−3.1	0.10	−0.8	21.19	1.5	93.32
−3.0	0.13	−0.7	24.20	1.6	94.52
−2.9	0.19	−0.6	27.42	1.7	95.54
−2.8	0.26	−0.5	30.85	1.8	96.41
−2.7	0.35	−0.4	34.46	1.9	97.13
−2.6	0.47	−0.3	38.21	2.0	97.73
−2.5	0.62	−0.2	42.07	2.1	98.21
−2.4	0.82	−0.1	46.02	2.2	98.61
−2.3	1.07	0.0	50.00	2.3	98.93
−2.2	1.39	0.1	53.98	2.4	99.18
−2.1	1.79	0.2	57.93	2.5	99.38
−2.0	2.27	0.3	61.79	2.6	99.53
−1.9	2.87	0.4	65.54	2.7	99.65
−1.8	3.59	0.5	69.15	2.8	99.74
−1.7	4.46	0.6	72.58	2.9	99.81
−1.6	5.48	0.7	75.80	3.0	99.87
−1.5	6.68	0.8	78.81	3.1	99.90
−1.4	8.08	0.9	81.59	3.2	99.93
−1.3	9.68	1.0	84.13	3.3	99.95
−1.2	11.51	1.1	86.43	3.4	99.97

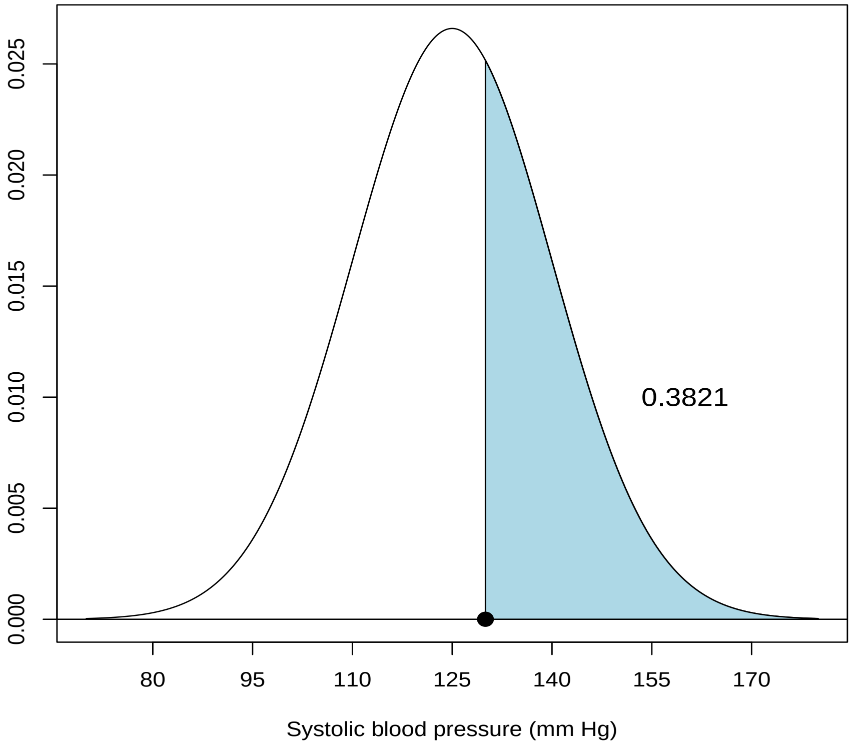


Figure 13.10: Normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 15$. The area to the right of 130 mm Hg is shown shaded.

Reading off Table B, the corresponding percentage is 61.79%.

- This is the percentage of American males who have SBP below 130 mm Hg.
- Therefore, the percentage of American males who have SBP above 130 mm Hg is

$$100\% - 61.79\% = 38.21\%.$$

In R (which reports decimals; not percentages),

```
> 1-pnorm(0.3)
[1] 0.3821
```

Q: What percentage of American males have SBP between 100 and 130 mm Hg?
A: Recall that

- the percentage of American males with SBP below 100 mm Hg is 4.46%.
- the percentage of American males with SBP above 130 mm Hg is 38.21%.

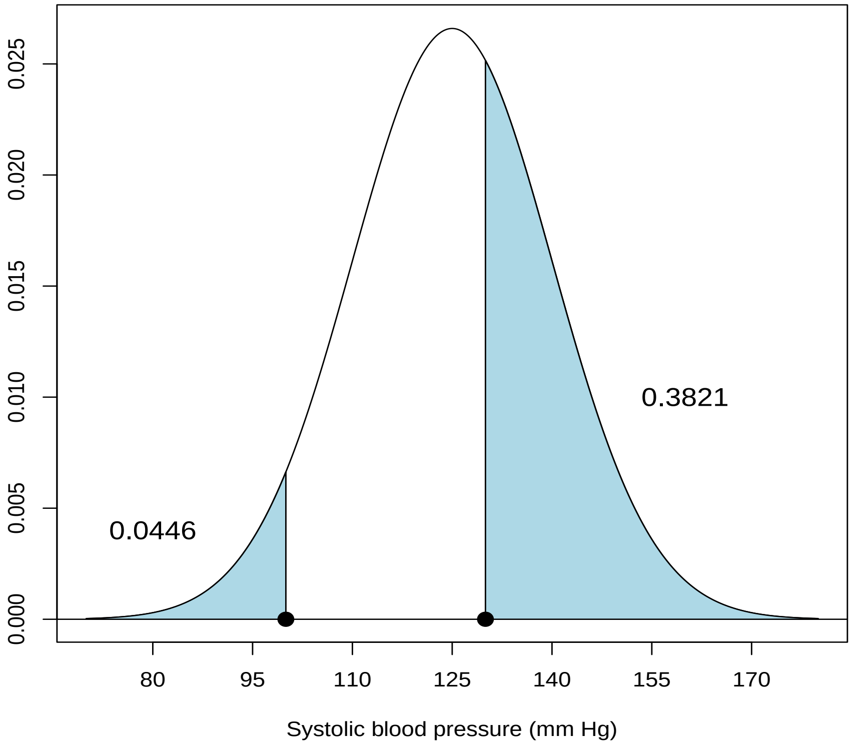


Figure 13.11: Normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 15$.

Therefore, the percentage of American males with SBP between 100 and 130 mm Hg is

$$100\% - 4.46\% - 38.21\% = 57.33\%.$$

See Figure 13.11 above.

Recall: The area under any population density curve is 1 (or 100%).

Example 13.5. For a population of runners, the time it takes to complete the New York Marathon (in minutes) is normally distributed with mean $\mu = 250$ and standard deviation $\sigma = 40$. Use this distribution to answer the questions below:

(a) What percentage of runners will finish the marathon in less than 180 minutes?
(b) What percentage of runners will take at least 310 minutes to complete the marathon (i.e., take longer than 310 minutes)?
(c) What percentage of runners will finish the marathon between 180 and 310 minutes?

The population density curve is shown in Figure 13.12 (next page). Use Table B.

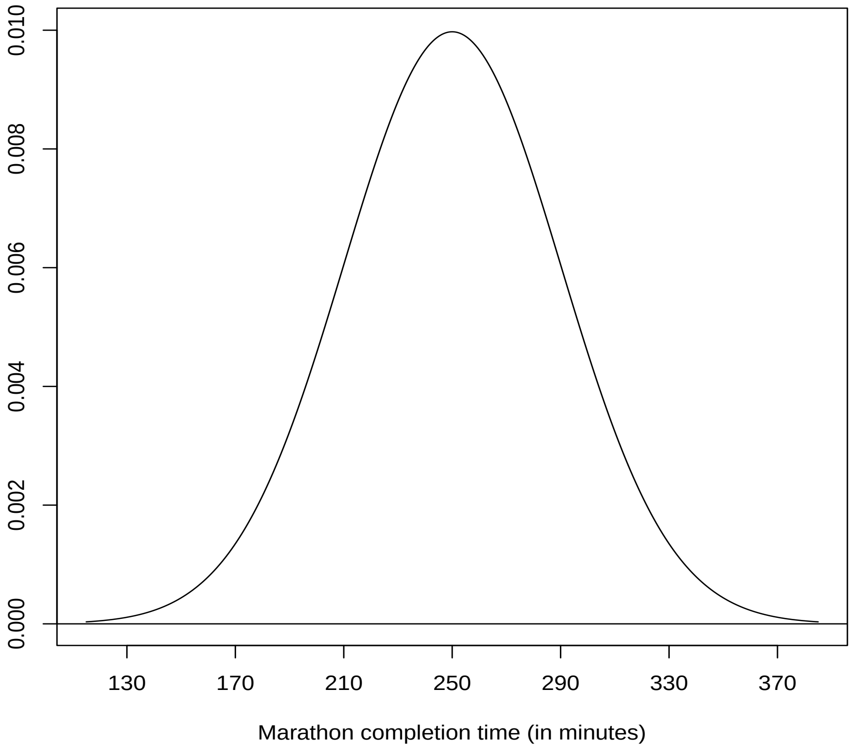


Figure 13.12: Normal distribution with mean $\mu = 250$ and standard deviation $\sigma = 40$.

Solutions:

14 Describing Relationships: Scatterplots and Correlation

14.1 Introduction

Remark: In the last three chapters, we have examined various aspects of quantitative distributions:

- graphical displays: histograms, stem plots, boxplots
- numerical summaries: mean/median, standard deviation, five-number summary
- population density curves, normal distributions.

A common theme in each example was that we were looking at one quantitative variable; e.g., birth weights, arsenic concentrations, turtle shell lengths, systolic blood pressure, marathon times, etc. Our goal was to examine the distribution−both graphically and with numerical summaries−for one quantitative variable.

Preview: In most experiments and observational studies, investigators are interested in more than one variable. For example,

- BMI and the number of calories consumed at school-provided lunches
- SAT score and first-year college GPA
- amount of state-level funding (dollars spent per student) and high school graduation rates
- daily Google stock returns and daily Walmart stock returns
- MLB team winning percentage and total team salary (in dollars).

In each of these examples, investigators might want to know if there is a relationship between the two variables. For example, are certain values of one variable usually associated with certain values of the other? If so, how strong is the relationship? Investigators might also want to build statistical models which formally describe the mathematical relationship between the variables; this forms the basis for a statistical method called **regression** (Chapter 15).

14.2 Scatterplots

Example 14.1. Manatees are large, slow-moving creatures found along the coast of Florida. Many manatees are injured or killed by boats. The table on the next page shows the number of boats registered in Florida (in thousands) and the number manatees killed each year between 1977 and 2018.

Table 14.1: Manatee data. Number of Florida boat registrations (thousands) and manatees killed by boats during 1977-2018.

Year	Boats	Manatees	Year	Boats	Manatees	Year	Boats	Manatees
1977	447	13	1992	679	38	2007	1027	73
1978	460	21	1993	678	35	2008	1010	90
1979	481	24	1994	696	49	2009	982	97
1980	498	16	1995	713	42	2010	942	83
1981	513	24	1996	732	60	2011	922	88
1982	512	20	1997	755	54	2012	902	82
1983	526	15	1998	809	66	2013	897	73
1984	559	34	1999	830	82	2014	900	69
1985	585	33	2000	880	78	2015	916	86
1986	614	33	2001	944	81	2016	931	106
1987	645	39	2002	962	95	2017	944	111
1988	675	43	2003	978	73	2018	951	124
1989	711	50	2004	983	69			
1990	719	47	2005	1010	79			
1991	681	53	2006	1024	92			

Analysis: There are two quantitative variables in this example: the number of boat registrations (in thousands) and the number of manatees killed by boats. Figure 14.1 (next page) shows distributions of each variable separately.

- Histograms show distributions while ignoring the time aspect.
- Line graphs show how the values of each variable change over time.

Limitation: These displays consider only one variable at a time. They do not show how the two variables are potentially related.

Definition: A **scatterplot** is a graphical display that shows the relationship between two quantitative variables measured on the same individuals.

- The values of one variable appear on the horizontal axis; the values of the other variable appear on the vertical axis.
- Scatterplots give a visual impression of how the two variables behave together.

Manatee data: Figure 14.2 shows the scatterplot for the manatee data in Table 14.1. This graph shows a **positive linear relationship** between the two variables. As the values of one variable increase, so do the other's values and the relationship is linear (i.e., a straight line) in form.

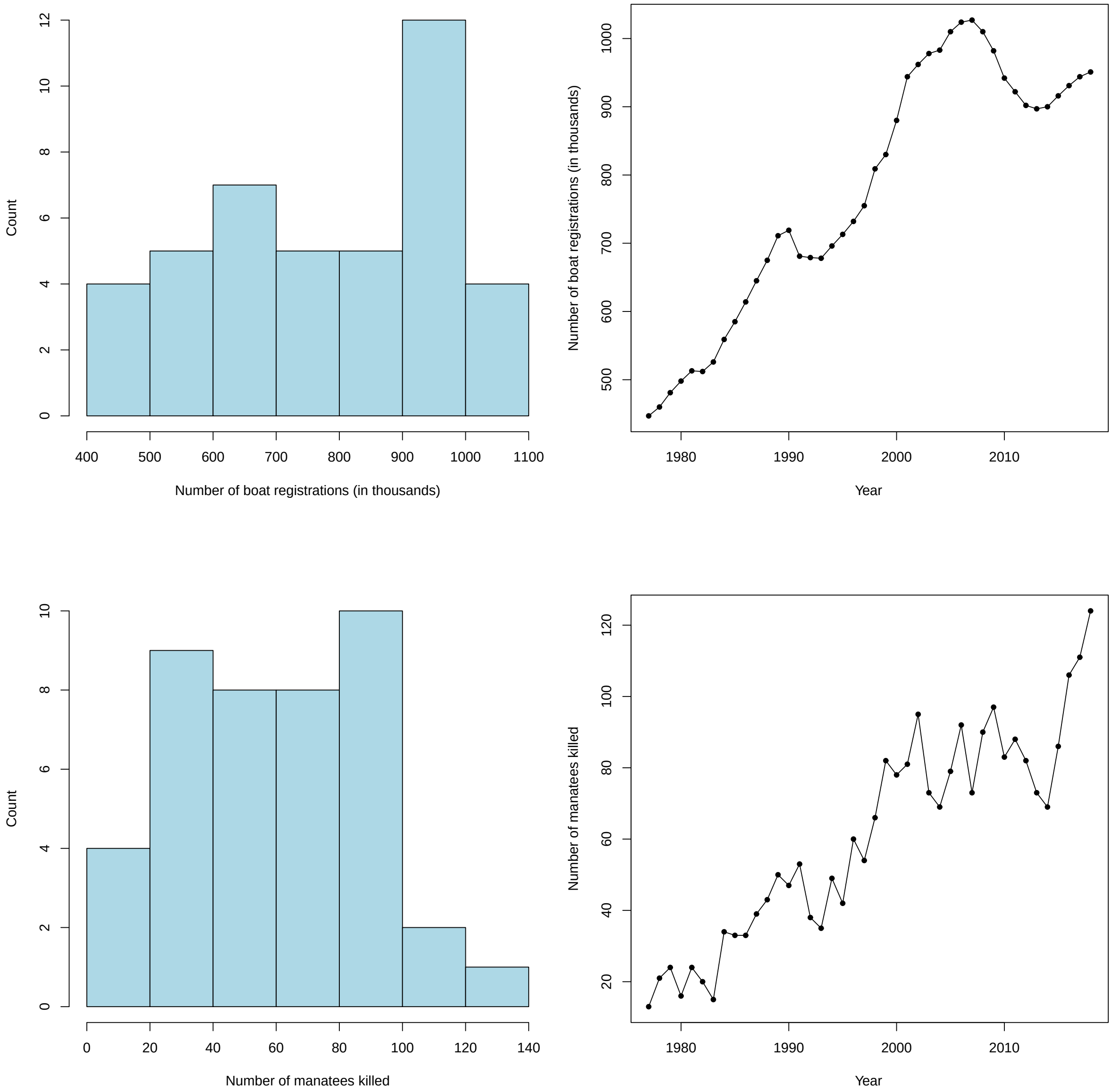


Figure 14.1: Manatee data. Top: Number of boat registrations in Florida (in thousands). Bottom: Number of manatees killed by boats.

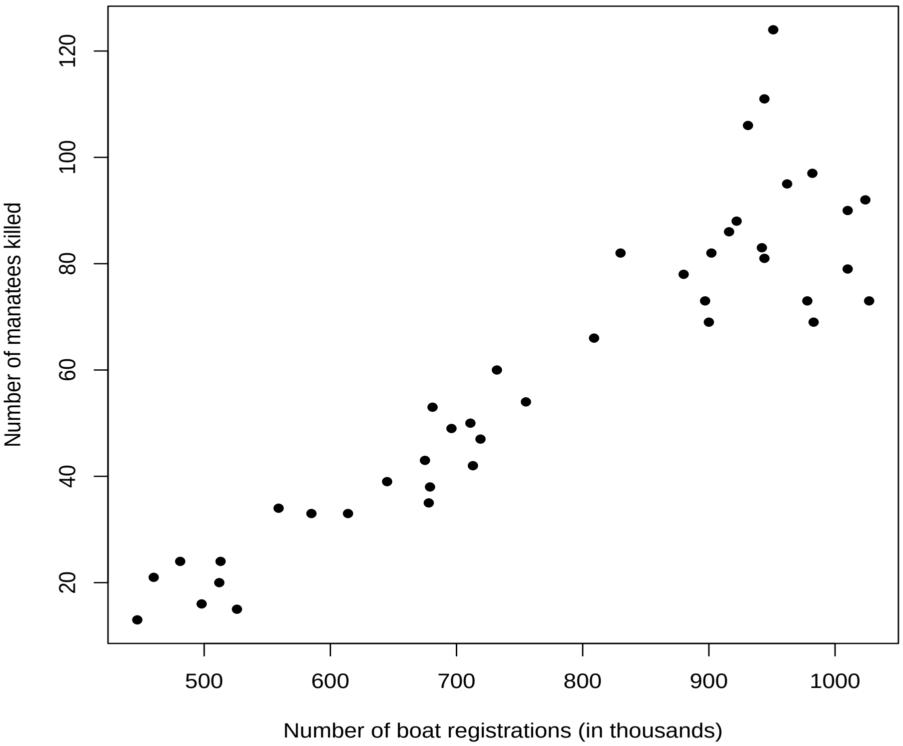


Figure 14.2: Manatee data. Scatterplot of the number of boat registrations in Florida (in thousands) and the number of manatees killed by boats.

Example 14.2. Elementary school performance in Florida is based on a standardized exam called the Florida Comprehensive Assessment Test (FCAT). The authors of a study published in *Journal of Educational and Behavioural Statistics* examined the relationship between

$$\begin{aligned} x &= \text{percentage of students below the poverty level} \\ y &= \text{average FCAT reading score} \end{aligned}$$

for a sample of $n = 22$ Florida elementary schools. A scatterplot of these observations is shown in Figure 14.3 (next page). The data set for these observations is online.

Analysis: There are two quantitative variables in this example: the percentage of students below the poverty level and the average FCAT reading score.

- Figure 14.3 shows a **negative linear relationship** between the two variables. As the values of one variable increase, the other variable's values tend to decrease and the relationship is linear (i.e., a straight line) in form.
- We use x and y to denote the variables in this example. This is standard notation and will be used later.

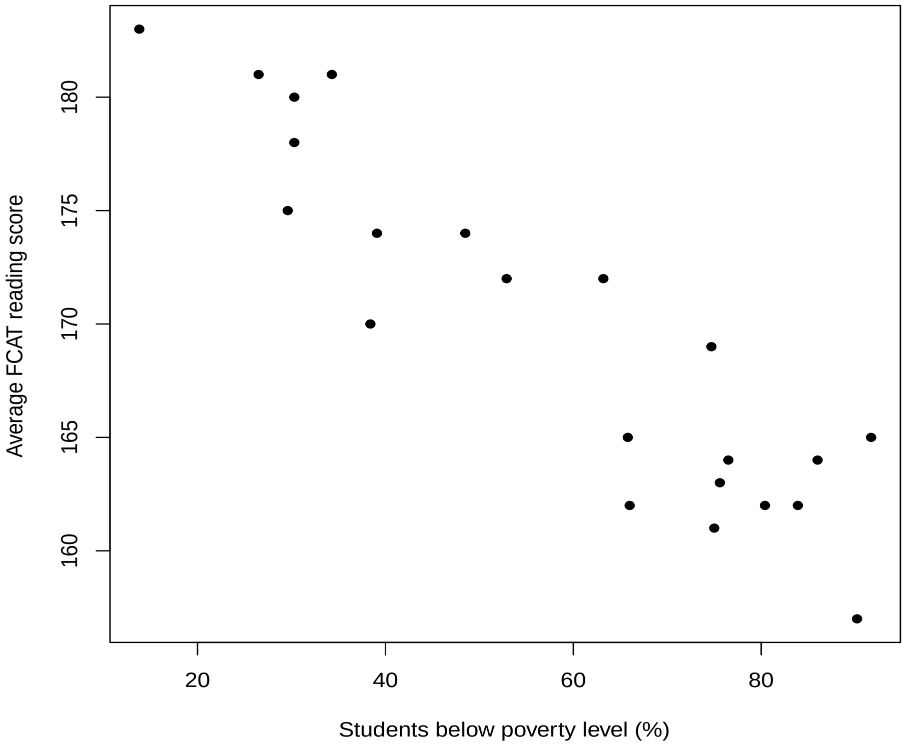


Figure 14.3: FCAT data. Scatterplot of the percentage of students below the poverty level (x) and average FCAT reading score (y) for a sample of $n = 22$ elementary schools.

Remark: Constructing scatterplots is easy. Interpreting them is more important. The first thing we should remember is **statistical inference**.

- Scatterplots are used to show the relationship between observations for two quantitative variables.
- If the observations we have are from a representative sample (as in Example 14.2), then the scatterplot presents an impression of the underlying relationship between the two variables in the population.
- Therefore, by interpreting characteristics we see in the scatterplots, we are interpreting what may be "going on" in the larger population of individuals.
- **Q:** Does the pattern we see in Figure 14.3 generalize to the population of all elementary schools in Florida?
  - This is a question about statistical inference.

Interpretation: We will focus on the following four characteristics when we examine and describe scatterplots:

1. Form: Are there straight-line (linear) patterns or curved patterns? Do observations tend to fall into clusters?
2. Direction: Are the variables positively related or negatively related?
3. Strength: Is the relationship strong, moderate, or weak? See Figure 14.4 (next page).
    - The scatterplot may no display no discernible relationship between the two variables; e.g., a "random scatter of points."
4. Deviations from the overall pattern (e.g., outliers, etc.).

Definitions: Two quantitative variables are **positively related** when an increase in one variable tends to accompany an increase in the other. They are **negatively related** when an increase in one variable tends to accompany a decrease in the other.

- Example 14.1: The number of boat registrations in Florida (x) and the number of manatees killed by boats (y) are positively related.
- Example 14.2: The percentage of students below the poverty level (x) and the average FCAT reading score (y) are negatively related.
- In both examples, the variables show a linear (straight-line) relationship, and the straight-line relationship is strong. There are no outliers in either example.

Notation: As noted earlier, it is common to use the letters x and y to denote the variables in a scatterplot.

- We use y to denote the **response variable** in an experiment or observational study.
- We use x to denote an **explanatory variable**.
- In most contexts, the goal is to determine how the response variable depends on the explanatory variable; e.g.,
    - How does the number of manatee deaths (y) depend on the number of boat registrations (x)?
    - How does the average FCAT reading score (y) depend on the percentage of students below the poverty level (x)?
- In some examples, it might be difficult to say which variable is which (e.g., height versus weight or weight versus height?). This is fine. We can still use a scatterplot to show how the variables are related.

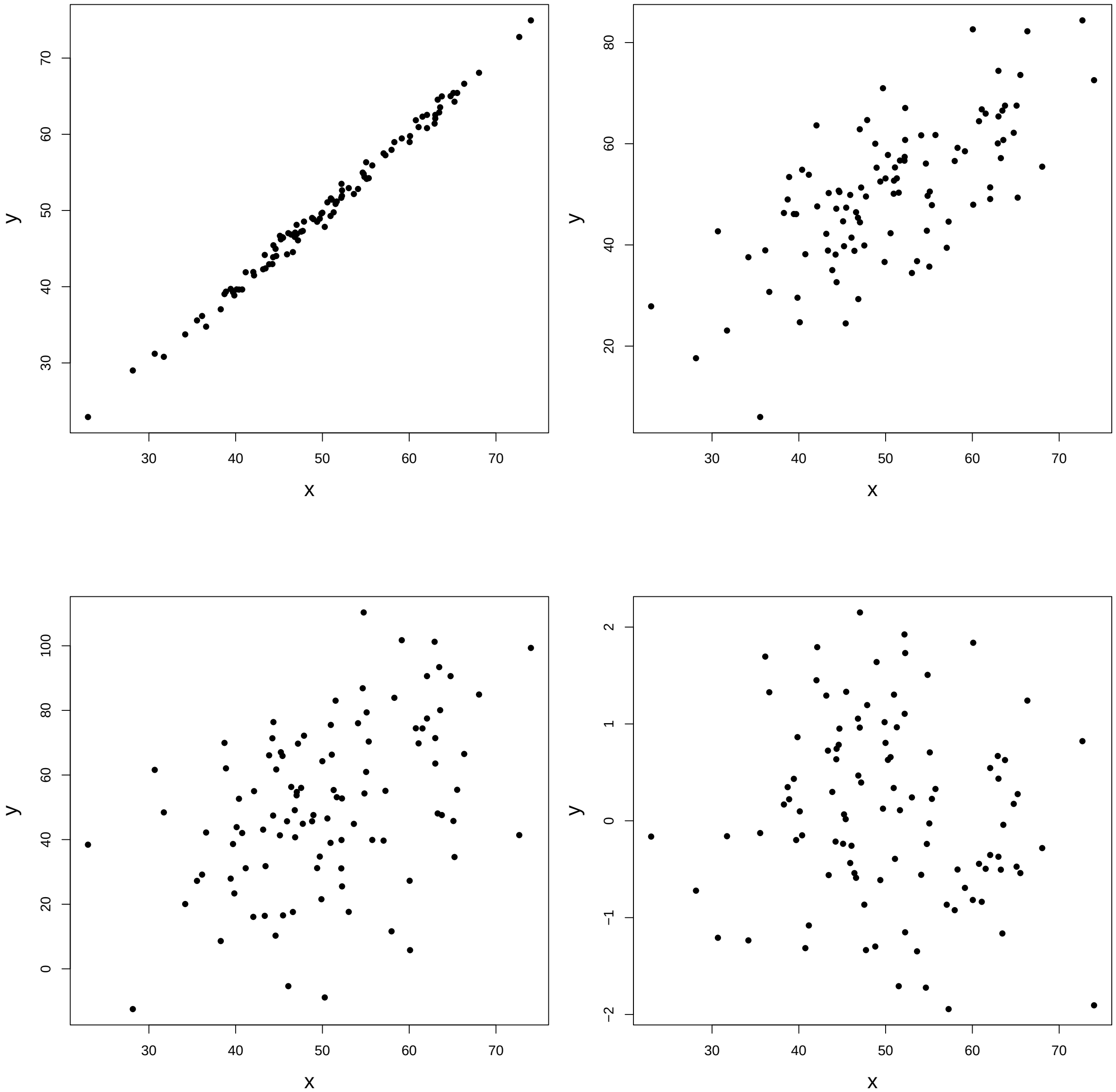


Figure 14.4: Scatterplots with positive linear relationships. Upper left: Very strong. Upper right: Moderate-to-strong. Lower left: Weak-to-moderate. Lower right: No linear relationship (random scatter).

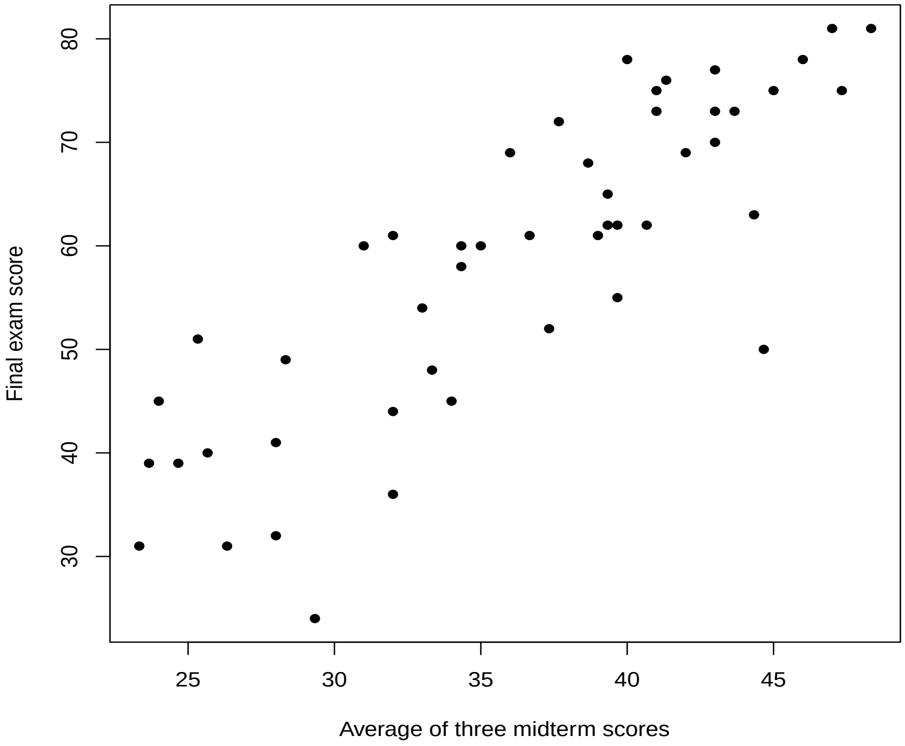


Figure 14.5: STAT 110 data. Scatterplot of the final exam score (y) and the average score of the three midterms (x).

Example 14.3. The data in Figure 14.5 show midterm scores and final exam scores for students in one of my previous STAT 110 classes (I'm not going to tell you which one). Specifically,

$$\begin{aligned} x &= \text{average midterm score (out of 3 midterms)} \\ y &= \text{final exam score.} \end{aligned}$$

Here is how we would interpret the scatterplot in Figure 14.5 in terms of our four characteristics:

- Form: There is a linear (straight-line) relationship between the average midterm score and the final exam score.
- Direction: The relationship is positive.
- Strength: The linear relationship is strong.
- Deviations: There is possibly one outlier (i.e., the student whose average midterm score was around 45).

Remark: Treating x as an explanatory variable and y as a response variable makes sense in Example 14.3.

- How might a student's final exam score (y) depend on the student's average midterm score (x)?
- Imagine you are a student whose average midterm score is 40. What would you **predict** your final exam score to be?

It wouldn't make (as much) sense to ask these questions with the roles of x and y reversed, especially if you think about using midterm scores to predict a final exam score.

14.3 Correlation

Goal: Scatterplots give a visual impression of how two quantitative variables are related. We now wish to summarize relationships numerically.

Definition: The **correlation** is a number that describes the strength and direction of the straight-line (linear) relationship between two quantitative variables.

- The correlation is denoted by r.

Formula: For a data set with n individuals, the correlation is computed by

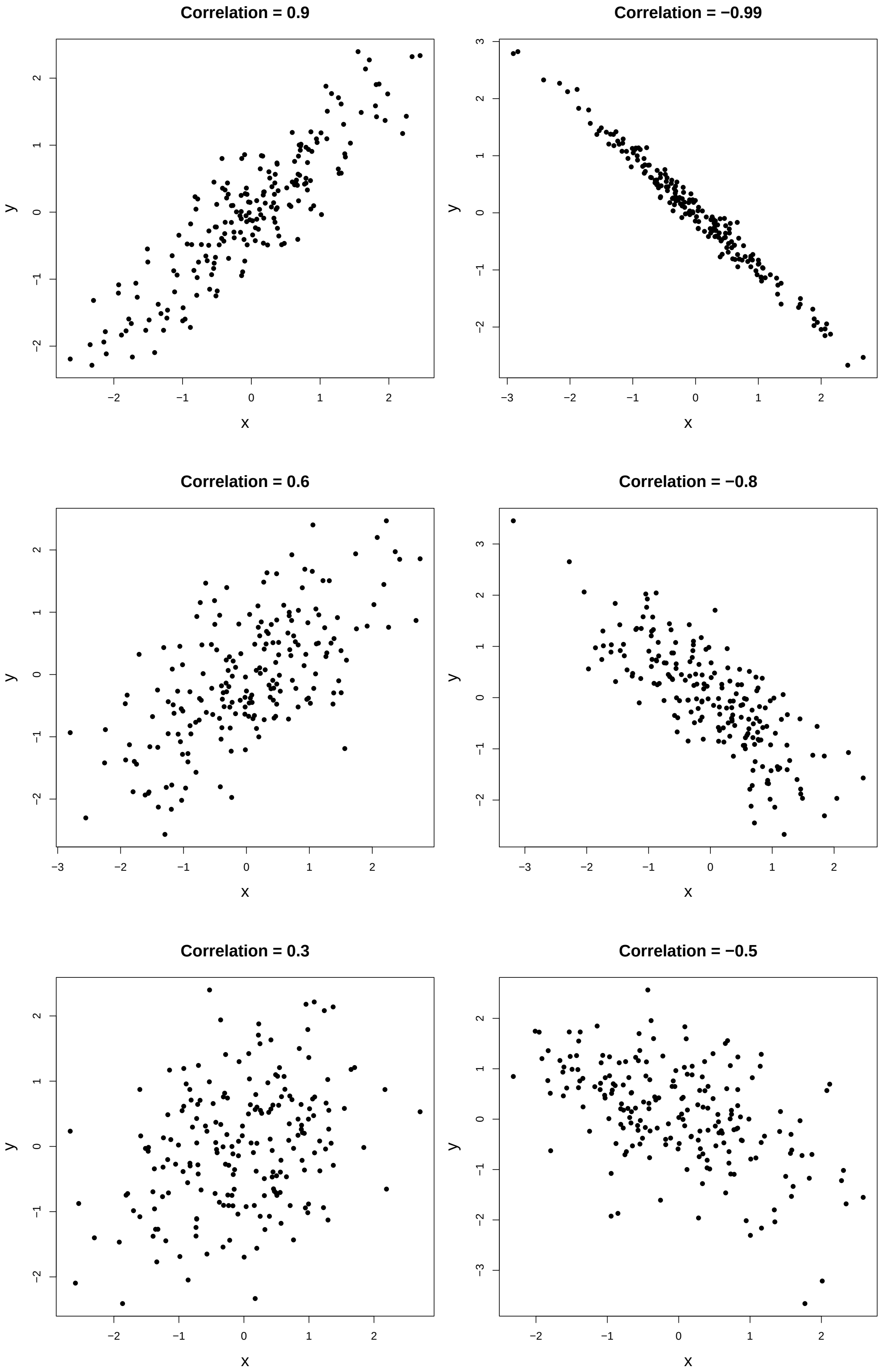
$$r = \frac{1}{n-1}\sum\left(\frac{x-\overline{x}}{s_x}\right)\left(\frac{y-\overline{y}}{s_y}\right),$$

where $\overline{x}$ and $\overline{y}$ are the sample means and s_x and s_y are the sample standard deviations.

- Example 4 in Moore and Notz (pp 326) gives an example where r is calculated "by hand."
  - We will use software (R) to calculate r.
  - It is more important to understand how to interpret the value of r (instead of doing hand calculation).
  - Many researchers interpret r incorrectly and confuse "correlation" with "causation."
- I do think it is interesting to note the terms

$$\frac{x-\overline{x}}{s_x} \quad \text{and} \quad \frac{y-\overline{y}}{s_y}$$

in the formula are the (sample) **standard scores** of x and y, respectively.



Correlation = 0.9

Correlation = −0.99

Correlation = 0.6

Correlation = −0.8

Correlation = 0.3

Correlation = −0.5

Correlation fact #1:

- If two variables (x and y) have a positive linear relationship, then $r > 0$.
- If two variables have a negative linear relationship, then $r < 0$.
- Note: $r = 0$ means that x and y have no linear relationship. We say that "x and y are uncorrelated."

Examples are shown on the previous page. Left column: $r > 0$. Right column: $r < 0$.

Correlation fact #2: The correlation r is always between -1 and 1; i.e.,

$$-1 \leq r \leq 1.$$

What happens at the extreme cases?

- If $r = 1$, then all of the data fall on a straight line with positive slope.
- If $r = -1$, then all of the data fall on a straight line with negative slope.
- In either case, the relationship between the two variables x and y is perfectly linear.
- Perfect relationships are an extreme rarity with real life data; e.g., see the scatterplots in Examples 14.1 and 14.2. There are always sources of variability that make a relationship not perfect, for example,
  - natural sampling variability
  - other (lurking) variables that influence how x and y are related.

Illustration: Let's calculate the correlation in Example 14.1 (manatee data) and Example 14.2 (FCAT data). Scatterplots are shown on the next page.

```
> cor(boats,manatees) # manatee data (Example 14.1)
[1] 0.92
> cor(poverty,FCAT.reading) # FCAT data (Example 14.2)
[1] -0.92
```

In both examples, the relationship between the two variables is linear and this relationship is strong.

- Example 14.1: The number of boat registrations in Florida and the number of manatees killed by boats are positively related ($r = 0.92$).
- Example 14.2: The percentage of students below the poverty level and the average FCAT reading score are negatively related ($r = -0.92$).

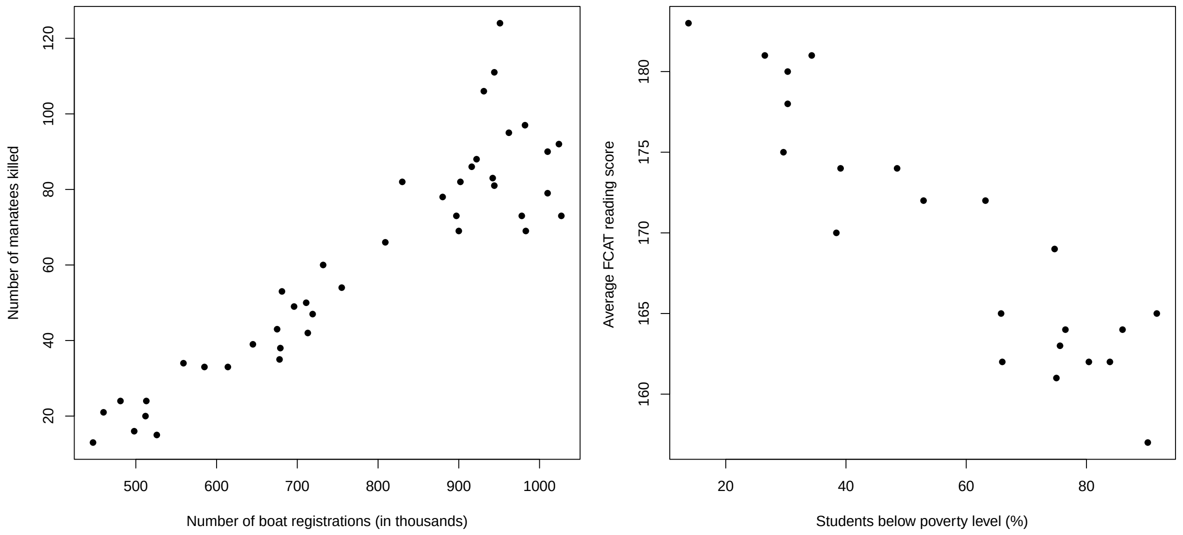


Figure 14.6: Scatterplots. Left: Manatee data (Example 14.1). Right: FCAT data (Example 14.2).

Exercise: Examine the scatterplot in Figure 14.5 for the STAT 110 exam data in Example 14.3. What do you think the correlation is? Is it positive or negative? Is it closer to 1 (perfect linear relationship) or closer to 0 (no linear relationship)?

Correlation fact #3: The correlation r is **unitless**; i.e., there are no units attached to it.

- The reason this is true is because the standard scores

$$\frac{x-\overline{x}}{s_x} \quad \text{and} \quad \frac{y-\overline{y}}{s_y}$$

  are also unitless (Chapter 13).

- This means that you could change the units of your data and it would not change the value of r. For example,
  - changing height measurements in inches to measurements in centimeters
  - changing temperature measurements in deg F to measurements in deg C
  - changing percentages below the poverty level (Example 14.2) to proportions below the poverty level.

Statements like this are wrong: "We found a strong positive correlation between the starting income levels of twins separated at birth ($r = 0.82$ dollars)."

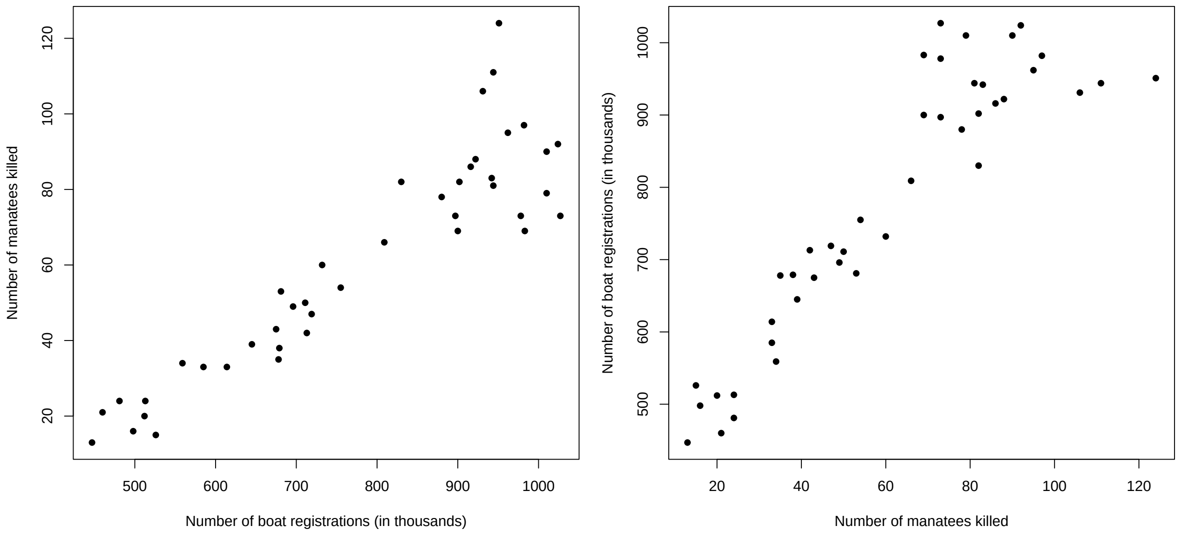


Figure 14.7: Left: Scatterplot of the number of boat registrations (in thousands) and the number of manatees killed by boats. Right: The axes have been switched.

Correlation fact #4: When calculating the correlation r, it makes no difference what you call x and what you call y. The correlation will be the same.

- In other words, the correlation r ignores the distinction between the explanatory variable x and the response variable y.

This also makes sense if you look at the formula:

$$r = \frac{1}{n-1} \sum \left(\frac{x - \overline{x}}{s_x} \right) \left(\frac{y - \overline{y}}{s_y} \right).$$

If we switch the order of the standard scores, it doesn't matter because we are multiplying them together.

```
> cor(boats,manatees) # Left
[1] 0.92
> cor(manatees,boats) # Right
[1] 0.92
```

Correlation fact #5: The correlation r measures the strength and the direction of linear (straight-line) relationships.

- The correlation does not describe a curved relationship, no matter how strong that relationship is.

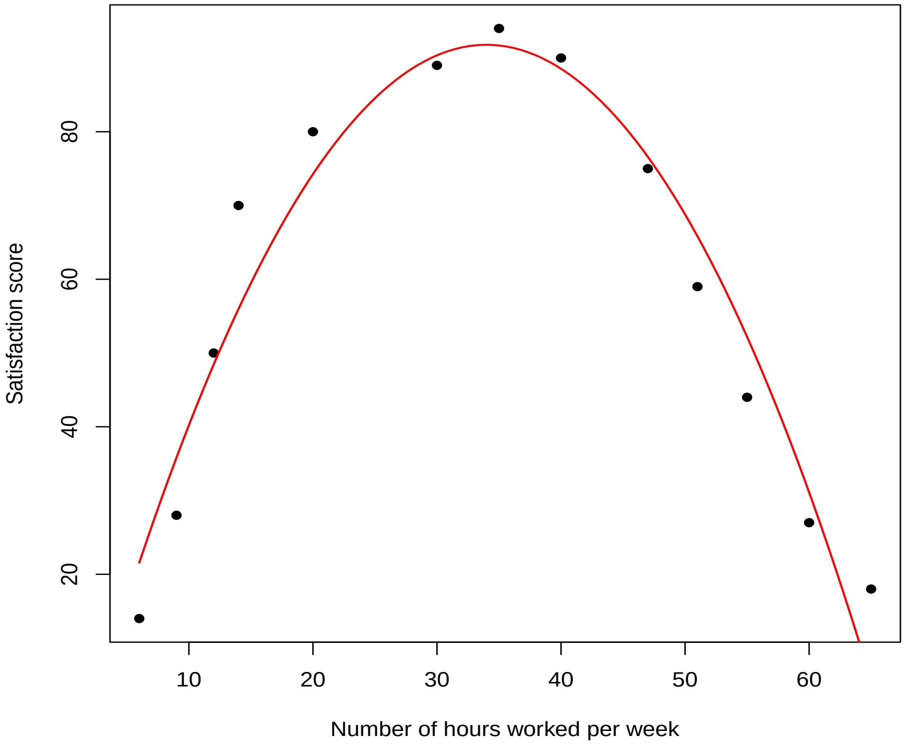


Figure 14.8: Job satisfaction data. Scatterplot of the number of hours worked per week (x) and employee satisfaction scores (y). A quadratic curve has been added.

Example 14.4. An organizational manager wants to understand the relationship between the number of hours worked (per week) and his employees' reported levels of "satisfaction and happiness." The manager observes a sample of $n = 13$ employees. The data set for these observations is online.

- The scatterplot in Figure 14.8 above shows these two variables are strongly related! It's just the relationship is not a straight line.
- Let's calculate r:

  ```
  > cor(hours,satisfaction)
  [1] -0.05
  ```

  This illustrates the limitation of correlation in describing relationships. The correlation does not describe curved relationships. It only describes the strength and direction of linear ones.

Statements like this are wrong: "There is a weak correlation between the number of hours worked and satisfaction scores ($r = -0.05$). This suggests the two variables are not related."

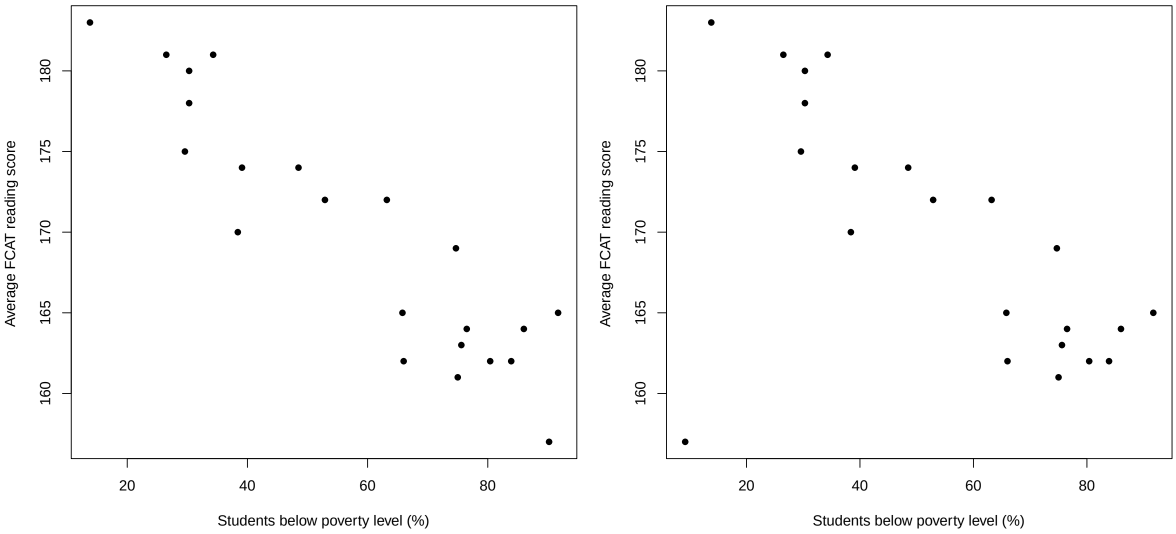


Figure 14.9: FCAT data. Left: Scatterplot of the percentage of students below the poverty level (x) and average FCAT reading score (y) for a sample of $n = 22$ elementary schools. Right: Scatterplot with outlier.

Correlation fact #6: The value of the correlation r is sensitive to outliers. In other words, observations that don't follow the overall pattern in a scatterplot can greatly distort the value of r.

Recall: In Example 14.2, the correlation between the percentage of students below the poverty level and the average FCAT reading score was $r = -0.92$.

```
> cor(poverty,FCAT.reading) # FCAT data (Example 14.2)
[1] -0.92
```

- Suppose the poverty percentage "90.2" in the original scatterplot (left, lower right observation) was mistakenly recorded as "9.2."
- The scatterplot with this error (now, an obvious outlier) is shown above on the right.

  ```
  > cor(poverty,FCAT.reading) # with outlier
  [1] -0.63
  ```

  The correlation has changed from $r = -0.92$ to $r = -0.63$ as a result of this error.
- This illustrates how the value of r can be highly sensitive to outliers.

Remark: Always plot your data first! Be cautious about misinterpreting the value of r when outliers are present.

15 Describing Relationships: Regression, Prediction, and Causation

15.1 Introduction

Preview: A common practice in the social sciences, public health, education, engineering, and other areas, is describing the relationship between two quantitative variables. For example,

- breaking and entering crime rates and motor vehicle theft rates in South Carolina
- percentage of residents with covid vaccination and hospital admissions
- SAT math and SAT reading scores for college-bound seniors
- machine filtration rate and moisture percentage of sewage sludge specimens.

In the last chapter, we learned the correlation r is a number that describes the straight-line relationship between two variables. In this chapter, we describe straight-line relationships using **regression equations**.

Example 15.1. Isolated systolic hypertension, which is an elevation in systolic but not diastolic blood pressure, is the most prevalent type of hypertension (especially among the elderly). An observational study investigated the relationship between

$$\begin{aligned} x &= \text{age (in years)} \\ y &= \text{systolic blood pressure (SBP, measured in mm Hg)} \end{aligned}$$

in adult males with one or more risk factors for this type of hypertension. There were $n = 30$ subjects in the study. The data are shown below.

Subject	Age	SBP	Subject	Age	SBP	Subject	Age	SBP
1	39	144	11	64	162	21	36	136
2	47	220	12	56	150	22	50	142
3	45	138	13	59	140	23	39	120
4	47	145	14	34	110	24	21	120
5	65	162	15	42	128	25	44	160
6	46	142	16	48	130	26	53	158
7	67	170	17	45	135	27	63	144
8	42	124	18	17	114	28	29	130
9	67	158	19	20	116	29	25	125
10	56	154	20	19	124	30	69	175

A scatterplot of these data is shown in Figure 15.1 (next page).

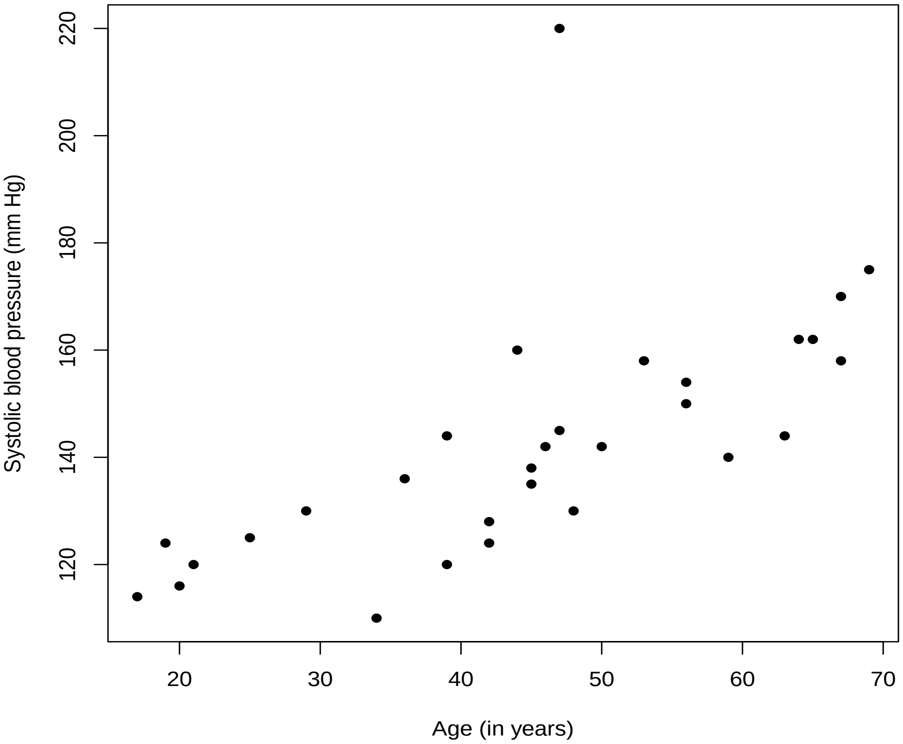


Figure 15.1: Hypertension data. Scatterplot of age (x, measured in years) and systolic blood pressure (y, measured in mm Hg) for $n = 30$ adult males.

Discussion: Here is how we would interpret the scatterplot in Figure 15.1 in terms of our four characteristics:

- Form: There is a linear (straight-line) relationship between age and SBP.
- Direction: The relationship is positive.
- Strength: The linear relationship is moderately strong ($r = 0.66$).
- Deviations: There is an obvious outlier: the 47-year old male whose SBP is 220 mm Hg.

```
> cor(age,SBP)
[1] 0.66
```

Definition: A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes.

- We often use a regression line to **predict** the value of y for a given value of x.

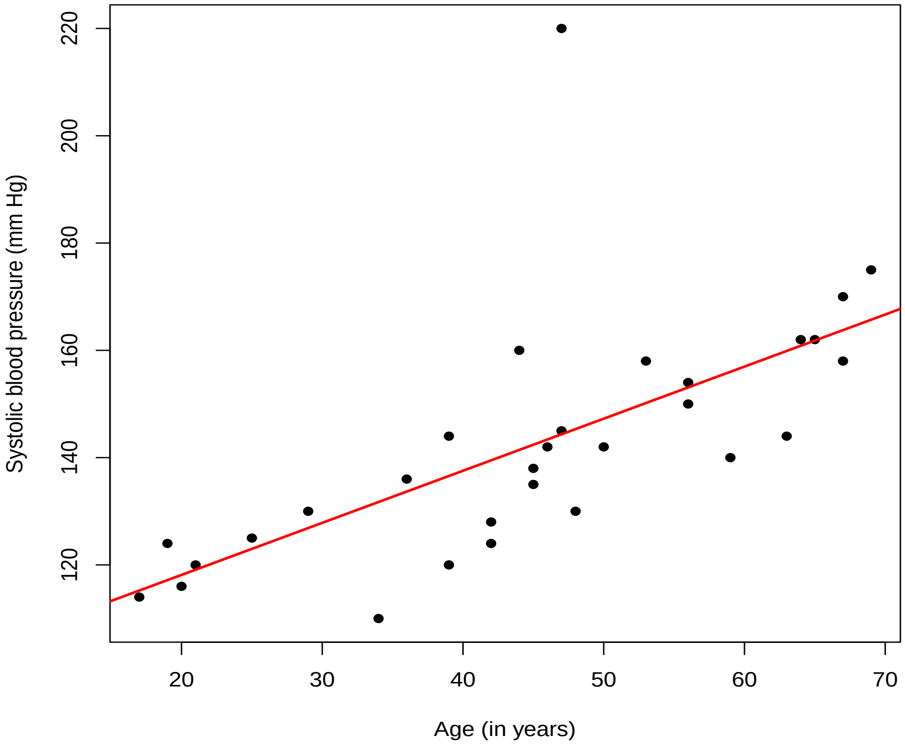


Figure 15.2: Hypertension data. Scatterplot of age and systolic blood pressure for $n = 30$ adult males. The least-squares regression line is added.

- For example, what SBP would we predict for a male who is 50 years old?
- The **least-squares regression line** has been added to the scatterplot above.
  - Where does this line come from; i.e., what is meant by "least squares?"
  - How do we use this line for prediction?

15.2 Regression equations and prediction

Algebra review: Suppose y is a response variable (on the vertical axis) and x is an explanatory variable (on the horizontal axis). A **straight line** relating y to x has an equation of the form

$$y = a + bx,$$

where b is the slope of the line and a is the intercept.

- The **slope** b is the amount by which the response y changes when x increases by one unit.

- In other words, for every one unit increase in x,
  - the response y increases by b if $b > 0$. This is a positive slope.
  - the response y decreases by b if $b < 0$. This is a negative slope.
- The **intercept** a is the value of y when $x = 0$. This is where the line intersects with the vertical axis.

Remark: With real data, like those in Example 15.1, we will never be able to find one line that fits the data perfectly; i.e., a line that goes through all the points. Why? Think about all of the sources of variability that are present in Example 15.1:

- biological variables that affect SBP (other than age)
- environmental variables that affect SBP
- genetic variables that affect SBP
- natural sampling variability! This is a sample of 30 males. Different samples would give different males and hence different observations.

All of these sources of variability make the relationship between age and SBP not perfect.

Q: So, how do we find the equation of the regression line?
A: We find the equation by using the **method of least squares**.

Definition: The **least-squares regression line** is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

- With the data from Example 15.1, Figure 15.3 shows vertical distances for subjects whose ages are between 20 and 35.
  - Imagine calculating these distances and squaring them for all $n = 30$ subjects in the data set.
  - Adding the squared distances gives the sum of squared vertical distances.
  - The least-squares regression line is the one line that makes this quantity as small as possible. This is what we mean by "best fit."
- We will use R to find the equation of the least-squares regression line (the "best-fit line").
  - Essentially, this boils down to R calculating the slope b and the intercept a for us.
  - The equation of the least-squares regression line is then

    $$y = a + bx.$$

    This is called a **regression equation**.

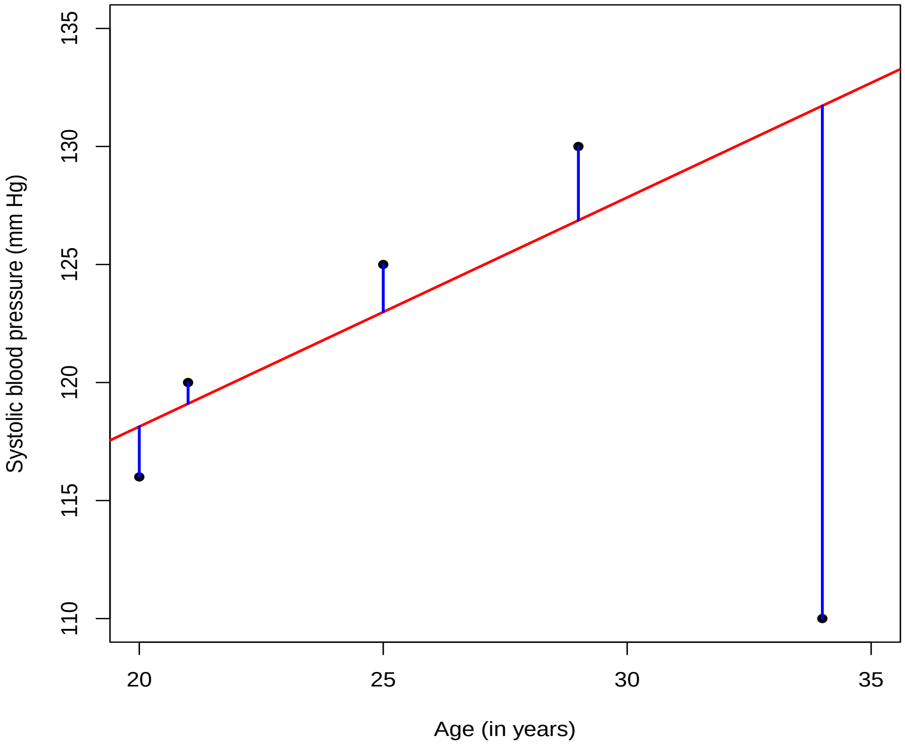


Figure 15.3: Hypertension data. Five subjects aged 20-35. The least-squares regression line is shown. Vertical distances from the data points to the line are shown.

Implementation in R:

```
> # Calculate intercept (a) and slope (b) of least-squares regression line
> fit = lm(SBP~age)
> fit

Coefficients:
(Intercept)          age
       98.7         0.97
```

Analysis: This output gives:

$$\begin{aligned} a &= 98.7 \\ b &= 0.97. \end{aligned}$$

The regression equation is

$$y = 98.7 + 0.97x, \quad \text{or, in other words,} \quad \text{SBP} = 98.7 + (0.97 \times \text{age}).$$

This is the line shown in Figure 15.2.

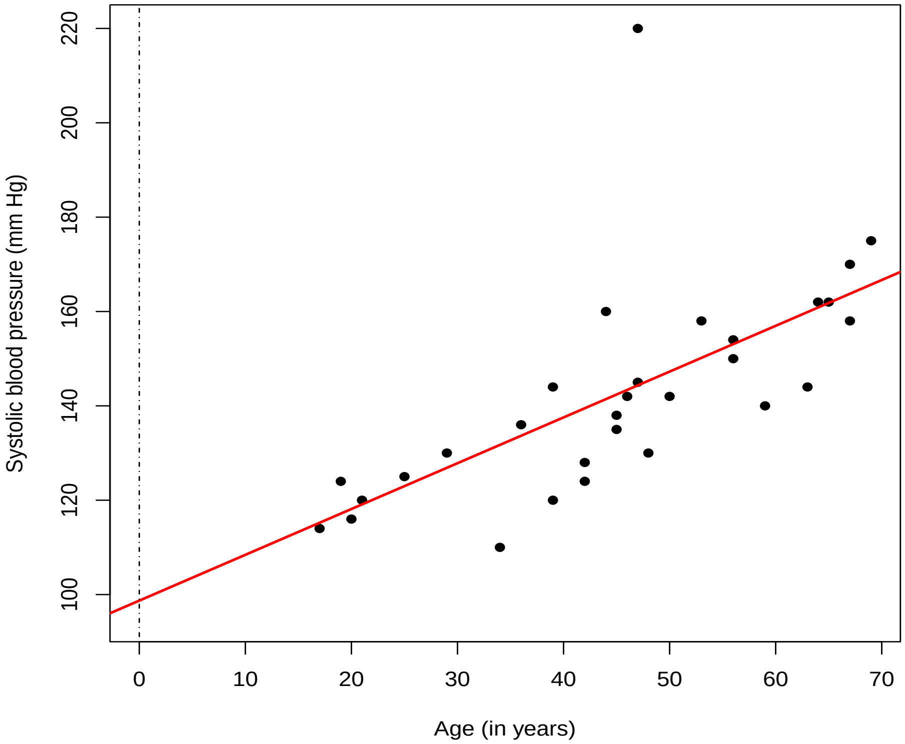


Figure 15.4: Hypertension data. Scatterplot of age and systolic blood pressure with the least-squares regression line. The horizontal axis has been extended to age = 0 years. A dotted vertical line at $x = 0$ is shown.

Interpretation:

- The slope $b = 0.97$ is interpreted as follows:

> "For a one-year increase in age, we would expect SBP to **increase** by 0.97 mm Hg."

- The intercept $a = 98.7$ is interpreted as follows:

> "For a person whose age is $x = 0$, we would expect SBP to be 98.7 mm Hg."

Remark: In most regression problems (including this one), the **slope** b is the most important to interpret. It describes how much we can expect the response variable to increase (or decrease) when the explanatory variable increases by one unit. On the other hand, the **intercept** a (i.e., the value of y when $x = 0$) is usually less meaningful.

- The notion of $x = 0$ refers to a male subject (in this example) whose age is "0;" a male subject who has just been born.

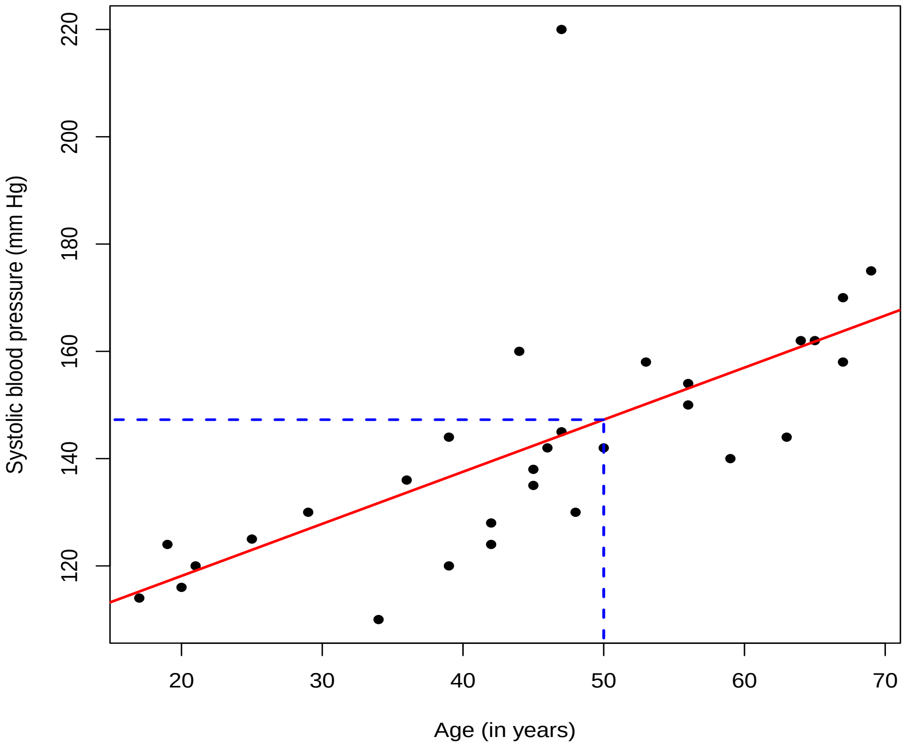


Figure 15.5: Hypertension data. Predicting SBP when age = 50 years.

- Even if this notion is accepted, it is hard to put too much faith in the assertion that a newborn's SBP is 98.7 mm Hg based on the data from this study.
- Why? The value $x = 0$ is well outside the range of the ages of the other subjects in the study (ages 17-69). Therefore, predicting a newborn's SBP to 98.7 mm Hg based on this study is an example of **extrapolation**.

Prediction: Use the regression equation to predict the SBP for a male subject who is 50 years old.
Solution. First, note that making this prediction makes sense. The age $x = 50$ falls right in the middle of the ages included in the study; see Figure 15.5 (above). There is no extrapolation here.

To make the prediction, simply plug in "50" for "age" in the regression equation:

$$\begin{aligned}\text{SBP} &= 98.7 + (0.97 \times \text{age}) \\ &= 98.7 + (0.97 \times 50) \;=\; 147.2 \text{ mm Hg}.\end{aligned}$$

The regression equation (model) would predict a 50-year old male to have SBP equal to 147.2 mm Hg.

Remarks:

- Predictions are best when the regression line fits the data well.
  - The better the fit, the better the predictions.
  - If a straight line is not a good description of the data, then the predictions may be bad.
- **Extrapolation** occurs when we make a prediction using a value of x outside the range of the x values seen in the study.
  - For example, using $x = 10$ years or $x = 90$ years in Example 15.1. These would both be extrapolations.
  - For an extrapolated prediction to be accurate, we would have to have faith the straight-line relationship holds for x values outside the range where we have observed data.
  - This may be difficult or impossible to assess if there is no empirical evidence (data) to support it.

15.3 Correlation and regression

Remark: Correlation and regression are statistical techniques used with data for two quantitative variables.

- The correlation r measures the strength and direction of a straight-line relationship with a single number; e.g., $r = 0.92$, $r = -0.66$, etc.
- A regression line describes the relationship with a mathematical equation.

Both correlation and regression can be affected by **outliers**. We already know this is true with correlation from the last chapter. The impact of outliers on the least-squares regression equation depends on where the outliers are.

Illustration: Figure 15.6 (next page) shows the hypertension data in Example 15.1.

- Original data: $r = 0.66$. Regression equation: SBP $= 98.7 + (0.97 \times \text{age})$
- Outlier removed: $r = 0.84$. Regression equation: SBP $= 97.1 + (0.95 \times \text{age})$.

Therefore, the correlation increased significantly when we removed the outlier. However, the regression equation did not change very much (see Figure 15.6).

- Although the outlier is clearly an outlier, it doesn't influence the regression equation that much. We would say this observation is **not influential**.

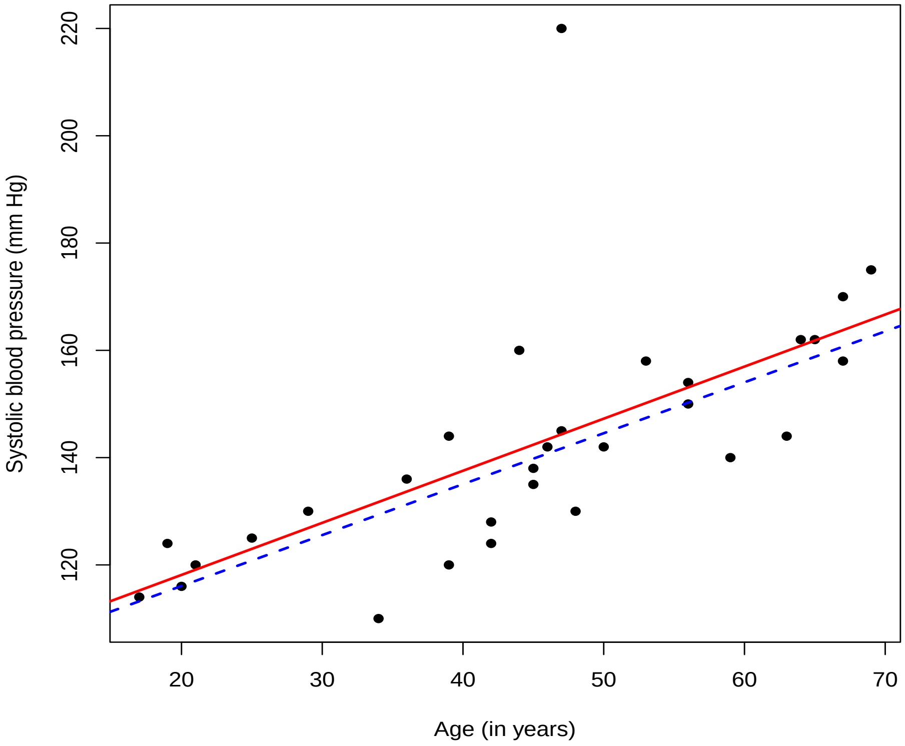


Figure 15.6: Hypertension data. Least-squares regression line with (red, solid) and without (blue, dashed) the original outlier.

Illustration: Figure 15.7 (next page) shows the hypertension data in Example 15.1, but now the outlier has been changed from $(47, 220)$ to $(17, 220)$. That is, this subject's age was recorded incorrectly as 17.

- Original data: $r = 0.66$. Regression equation: $\text{SBP} = 98.7 + (0.97 \times \text{age})$
- Outlier changed: $r = 0.40$. Regression equation: $\text{SBP} = 117.6 + (0.57 \times \text{age})$.

Therefore, the correlation decreased significantly when the outlier switched positions. In addition, the least-squares regression line changed drastically too (see Figure 15.7).

- This outlier changes the regression equation by a large amount. We would say this observation is **influential**.
- SBP predictions for younger and older subjects would be impacted heavily. Interestingly, predictions for our 50-year old male would be about the same.

Main point: Outliers have the potential to greatly distort the equation of the least-squares regression line. Always plot your data first!

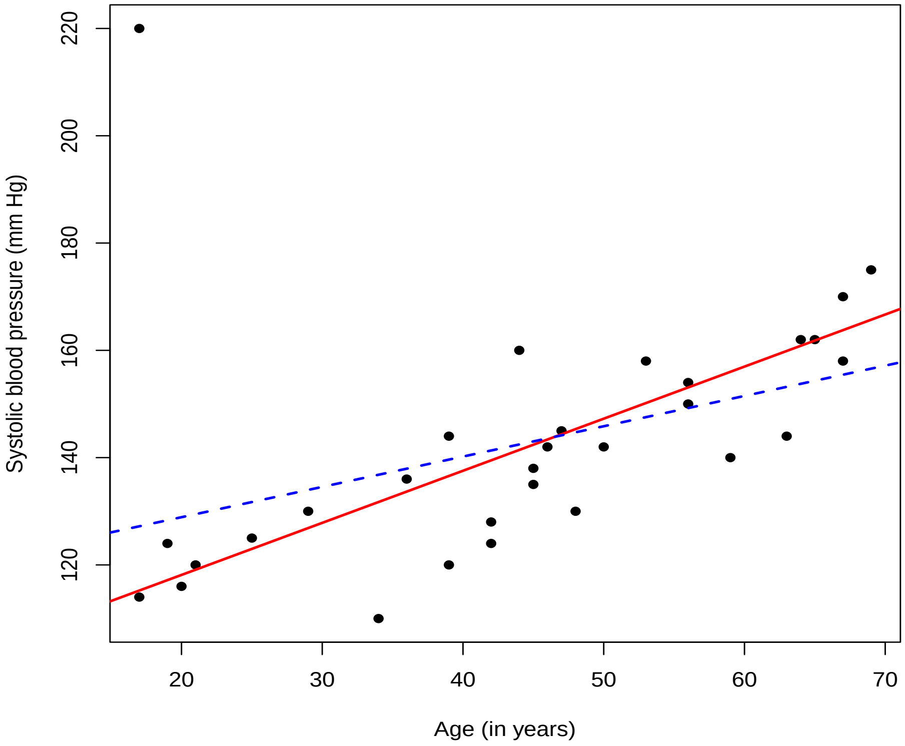


Figure 15.7: Hypertension data. Least-squares regression line with the outlier changed.

Summary: The usefulness of the least-squares regression line for prediction depends on the strength of the straight-line relationship between the two variables.

- The stronger the straight-line relationship, the more precise the predictions will be.
- The correlation r measures the strength of this relationship.

Definition: In a regression analysis, one way to assess how well a straight line fits the data is to compute the **square of the correlation**: r^2.

- It has a specific interpretation: r^2 gives the proportion of the variation in the response variable y explained by the straight-line relationship with the explanatory variable x.
  - The larger r^2 is, the more variation explained by the regression line (this leads to more precise predictions).
  - Caution: The interpretation above assumes the relationship between x and y is linear (i.e., a straight-line relationship).
  - Therefore, if you calculate r^2 when the true relationship between x and y is not linear (e.g., curved, etc.), then r^2 will not make sense.

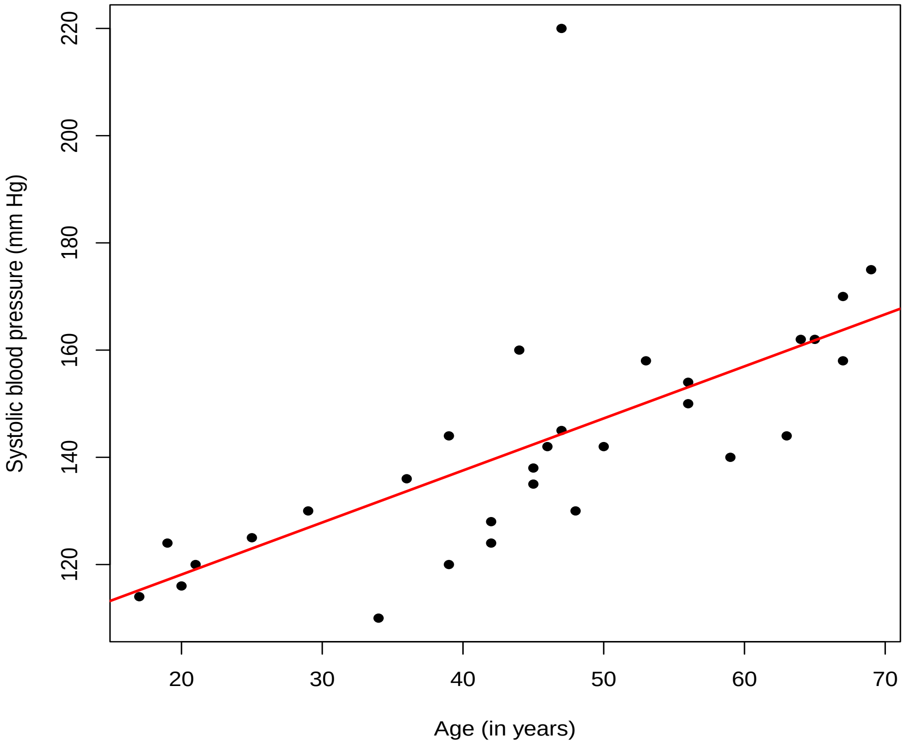


Figure 15.8: Hypertension data. Scatterplot of age and systolic blood pressure for $n = 30$ adult males. The least-squares regression line is added.

Illustration: In Example 15.1, we calculated the correlation to be $r = 0.66$. The square of the correlation is

$$r^2 = (0.66)^2 \approx 0.44.$$

Interpretation: Approximately 44% of the variability in the systolic blood pressure data is explained by the straight-line relationship with age.

- This means approximately 56% of the variability in the systolic blood pressure data is explained by other sources:
  - biological variables that affect SBP (other than age)
  - environmental variables that affect SBP
  - genetic variables that affect SBP
  - other variables that affect SBP.

Q: When would $r^2 = 1$?
A: This would happen only when all the data fell on a straight line with positive slope ($r = 1$) or with negative slope ($r = -1$).

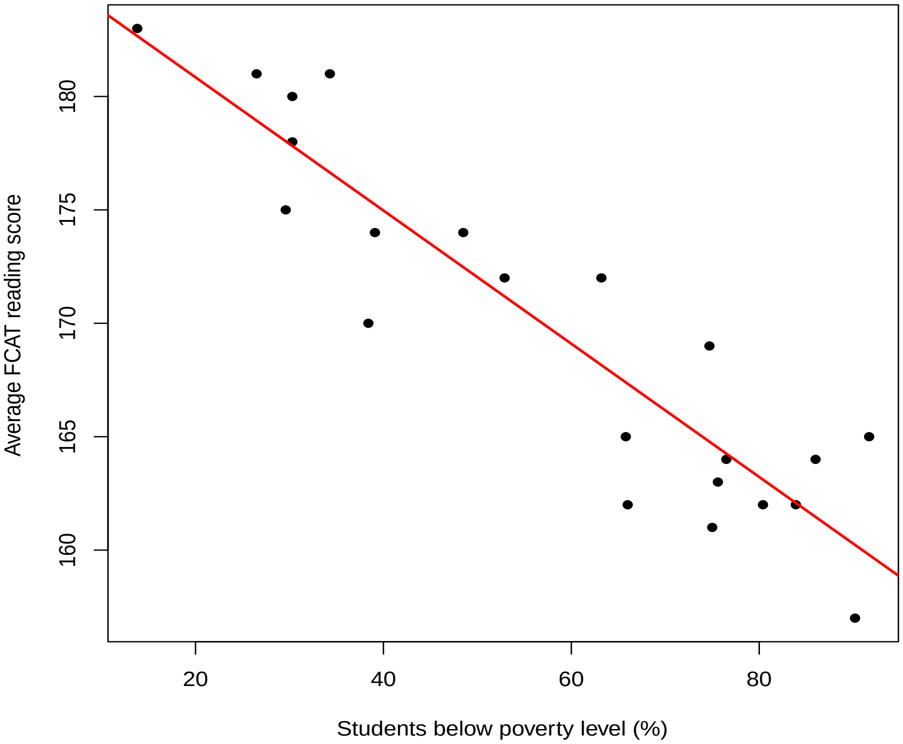


Figure 15.9: FCAT data. Scatterplot of the percentage of students below the poverty level (x) and average FCAT reading score (y) for a sample of $n = 22$ elementary schools. The least-squares regression line is added.

Example 15.2 (continuation of Example 14.2). The authors of a study published in *Journal of Educational and Behavioural Statistics* examined the relationship between

$$\begin{aligned} x &= \text{percentage of students below the poverty level} \\ y &= \text{average FCAT reading score} \end{aligned}$$

for a sample of $n = 22$ Florida elementary schools. A scatterplot of these observations is shown in Figure 15.9 (above). The least-squares regression line is superimposed.

Implementation in R:

```
> # Calculate intercept (a) and slope (b) of least-squares regression line
> fit = lm(FCAT.reading˜poverty)
> fit

Coefficients:
(Intercept)      poverty
      186.7        -0.29
```

```
> cor(poverty,FCAT.reading) # correlation
[1] -0.92
```

Analysis: This output gives:

$$\begin{aligned} a &= 186.7 \\ b &= -0.29. \end{aligned}$$

The regression equation is

$$y = 186.7 - 0.29x,$$

or, in other words,

$$\text{Average FCAT reading score} = 186.7 - (0.29 \times \%\text{ students below poverty level}).$$

Questions:
(a) What would you predict the average FCAT reading score to be for a school with 40% of its students below the poverty level?
(b) Calculate r^2 and interpret what it means.

Solutions. (a) To make this prediction, simply plug $x = 40$ in the regression equation:

$$\begin{aligned} y &= 186.7 - 0.29x \\ &= 186.7 - (0.29 \times 40) \ = \ 175.1. \end{aligned}$$

The regression equation (model) would predict this school's average FCAT reading score to be 175.1.

(b) The square of the correlation is

$$r^2 = (-0.92)^2 \approx 0.85.$$

Interpretation: Approximately 85% of the variability in the average FCAT reading score data is explained by the straight-line relationship with the percentage of students below the poverty level.

- This means approximately 15% of the variability in the average FCAT reading score data is explained by other sources:
  - quality of instruction
  - school size
  - school resources (e.g., after-school programs, etc.)
  - other variables that affect average FCAT reading scores.

Remark: We now revisit an example where the value of r^2 would have little or no meaning.

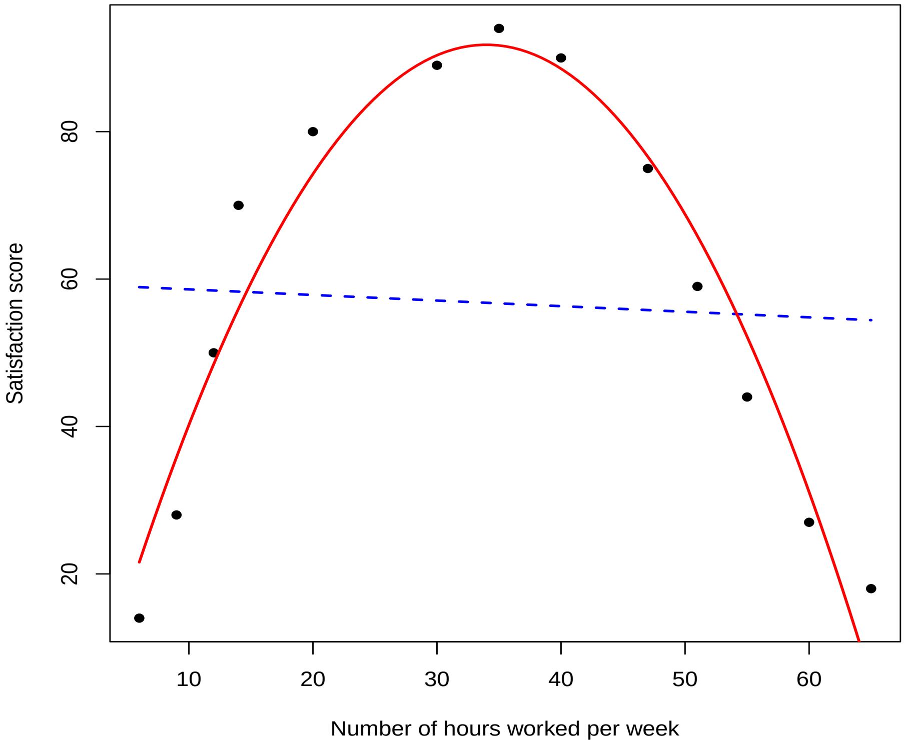


Figure 15.10: Job satisfaction data. Scatterplot of the number of hours worked per week (x) and employee satisfaction scores (y). The least-squares regression line and quadratic curve have been added.

Example 15.3 (continuation of Example 14.4). An organizational manager wants to understand the relationship between the number of hours worked (per week) and his employees' reported levels of "satisfaction and happiness." The manager observes a sample of $n = 13$ employees and records both variables for each employee.

```
> cor(hours,satisfaction)
[1] -0.05
```

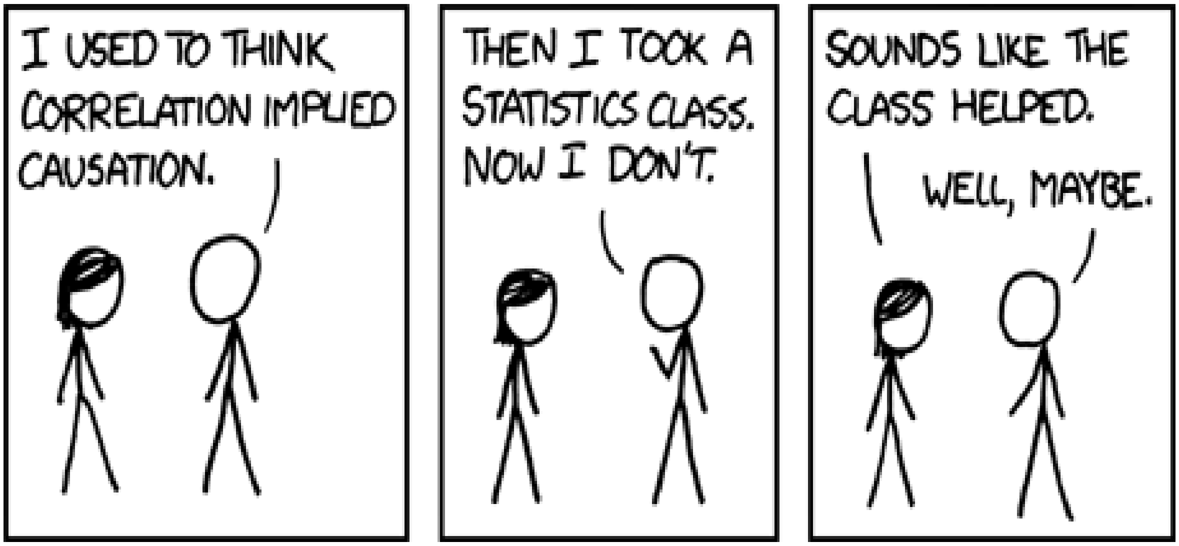
Figure 15.10 above shows a scatterplot of the observations with

- the least-squares regression line superimposed (dashed line)
- the least-squares quadratic regression curve superimposed (solid curve).

The correlation is $r = -0.05$, so the square of the correlation is

$$r^2 = (-0.05)^2 = 0.0025.$$

Interpretation: Approximately 0.25% of the variability in the satisfaction score data is explained by the straight-line relationship with the number of hours worked per week.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Discussion: Although the previous interpretation is technically correct, it would only make sense if the true relationship between the number of hours worked per week (x) and satisfaction scores (y) was **linear** (i.e., a straight line). This is another reminder how the correlation r describes straight-line relationships only. The same is true for the square of the correlation r^2.

15.4 Correlation $\neq$ causation

Important: Just because two variables are correlated−even if this correlation is very strong−this does not necessarily mean there is a **causal relationship** between the variables. In other words, we cannot conclude

- "Aging in men *causes* SBP to increase." (Example 15.1)
- "An increase in the percentage of students below the poverty level *causes* average FCAT reading scores to decline." (Example 15.2).

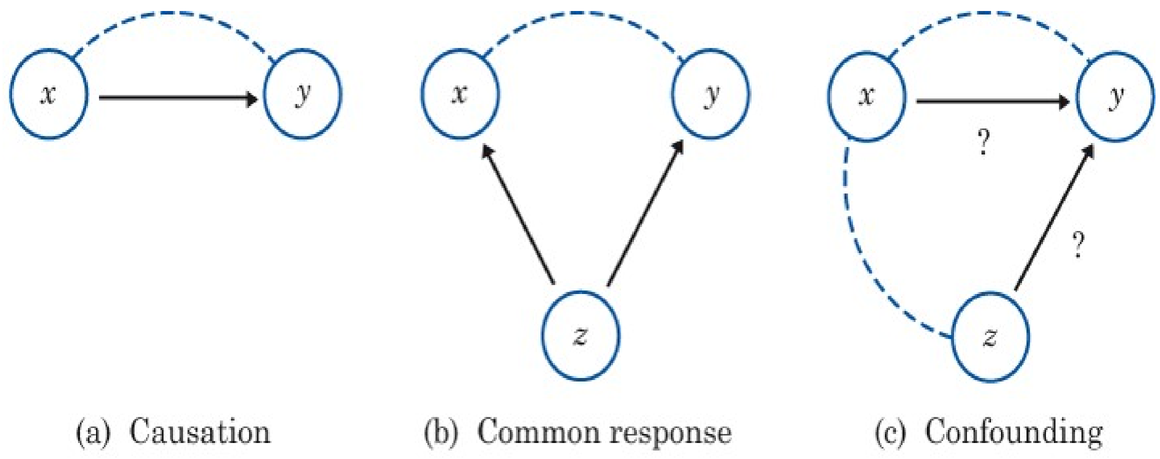
This important fact has spawned the well-known statistical aphorism:

"Correlation does not imply causation."

The correlation r only documents a special type of relationship or association between two variables: the degree to which the two variables are linearly related. It doesn't necessarily have anything to do with causality.

Q: Why?
A: Correlations could be spurious or coincidental.

- A Chicago newspaper reported "there is a strong correlation between the number of fire trucks at the scene of a fire and the amount of damage that the fire does."



(a) Causation (b) Common response (c) Confounding

- Are these two variables correlated?

$$\begin{aligned} x &= \text{number of fire trucks} \\ y &= \text{damage the fire does (measured in dollars).} \end{aligned}$$

  Probably, but there is no causal link here. An increase in the number of fire trucks at the scene of a fire does not *cause* more damage.

- This is a **spurious correlation**. The reason these variables are correlated is because of the presence of a third variable: the severity of the fire.
- Severe fires lead to both more trucks and more damage (common response).

- Some correlations are **coincidental**. For example, a recent study found a strong negative correlation between

$$\begin{aligned} x &= \text{number of lemons imported from Mexico} \\ y &= \text{highway fatality rate in the United States.} \end{aligned}$$

  Does increasing the number of lemons imported from Mexico *cause* the highway fatality rate in the US to decline? Probably not.

Important: The best way to determine if a causal relationship between two variables x and y exists is to perform a randomized comparative experiment.

- It is easy to find spurious and coincidental correlations in observational studies.
- However, these associations are usually not **repeatable**; i.e., if the study was performed again, there is no guarantee the same correlation would be found.
- Properly designed experiments control for the effects of lurking variables by blocking.
- Establishing causal relationships is much more difficult than simply documenting an association using the correlation.

17 Thinking about Chance

17.1 What is probability?

Remark: We are bombarded everyday with numbers describing the chance of something happening in the future (perhaps to us). For example,

- "The chance of winning the Powerball lottery is 1 in 292,000,000."
- "Subjects not wearing condoms are 3 times more likely to contract an STD when compared to subjects who wear condoms regularly."
- "The chance of dying in a plane crash is about 1 in 11,240 over one's entire lifetime. For a car crash, it's about 1 in 93" (**Source:** Insurance Information Institute).
- "The ESPN Analytics Match Predictor gives USC a 51.4% chance of beating Oregon State" (March 30, 2024).
- "There is a 70% chance of rain tomorrow."
- "The probability Donald Trump will win the Republican nomination is 5%" (Nate Silver, July 2015).

In some cases, numbers like these come from models based on mathematics or statistics (which may be trustworthy but not always). In other cases, numbers may arise from personal feeling or emotion. These can be highly inaccurate. We want to describe probability in a way that makes sense mathematically so we can have a better understanding of it means. This will also allow us to understand what probability does *not* mean.

Definition: Probability is the mathematics of chance. When we talk about chance, we are usually referring to a phenomenon that is **random**; i.e., the outcome of the phenomenon cannot be predicted with certainty.

- Flipping a coin. There are 2 possible outcomes (H, T).
- Rolling a die. There are 6 possible outcomes (1, 2, 3, 4, 5, 6).
- 2024 Presidential election. There are 3 possible outcomes (Biden, Trump, Other).
- 2024 NCAA Basketball Final Four. There are 4 possible outcomes:
  - Men: Connecticut, Alabama, Purdue, NC State
  - Women: South Carolina, NC State, Iowa, Connecticut.
- Disease diagnosis. There are 2 possible outcomes (positive, negative).

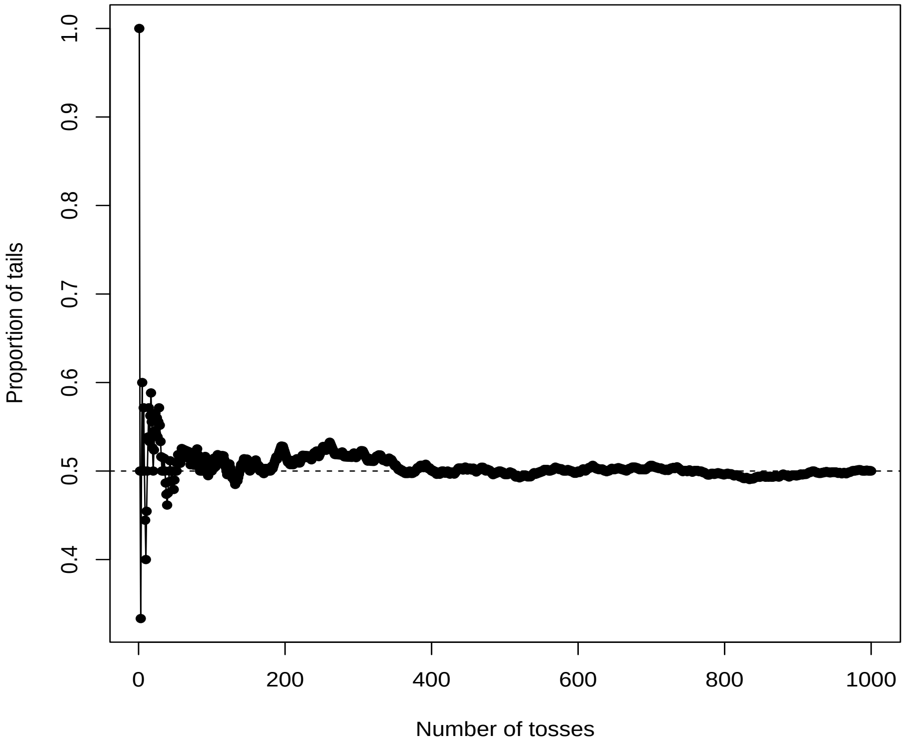


Figure 17.1: Line graph of the proportion of tails flipped in a series of 1000 flips. A horizontal line at 0.5 has been added.

Example 17.1. I will flip a fair coin. There are 2 possible outcomes (H, T) and both outcomes have the same chance of occurring. What is the probability of flipping a T?
Solution. Our simple intuition suggests the answer is 0.5. There are only two outcomes, each of them is equally likely, and we know one of them must occur.

Simulation: Another way to think about probability is that it's the proportion of times a "T" would occur if we flipped the coin many independent times.

- By "independent," I mean that the outcome on any one flip is not influenced by the others.

Figure 17.1 shows what could happen if we flipped the coin 1000 times and used a line graph to plot the proportion of "T" outcomes over the course of the 1000 flips. Not surprisingly, the proportion of "T" outcomes approaches 0.5 and stays close to it.

Moral: Chance behavior is unpredictable in the **short run**, but it has a regular and predictable pattern in the **long run**. The probability of an outcome refers to what happens in the long run.

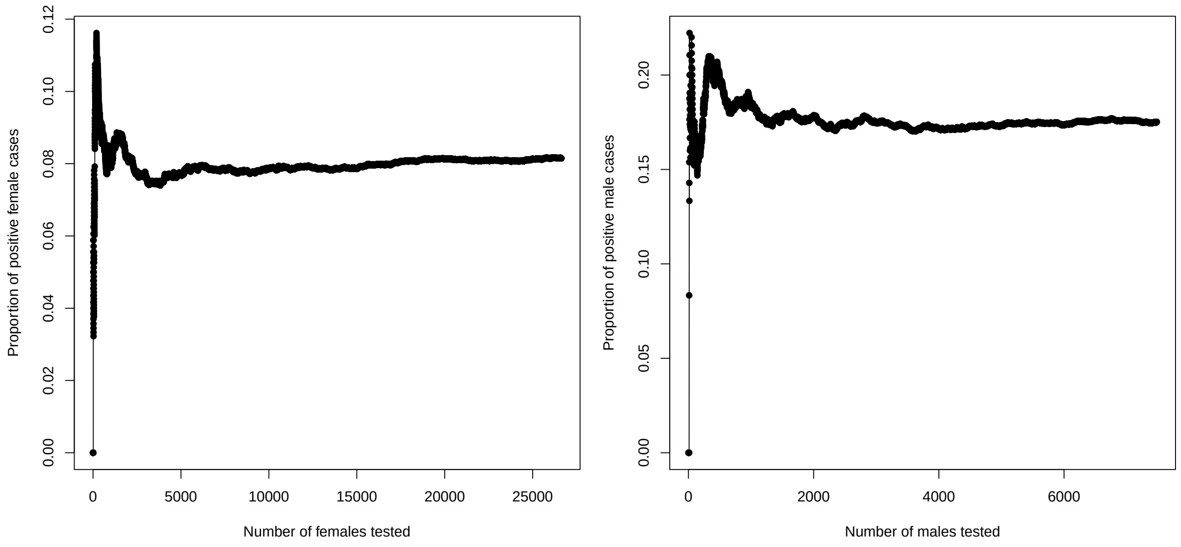


Figure 17.2: State Hygienic Laboratory data. Line graphs of the proportion of positive chlamydia cases for females (left) and males (right).

Example 17.2. The State Hygienic Laboratory (SHL) at University of Iowa tests "high-risk" Iowa residents for chlamydia. What is the probability a female tests positive? a male tests positive?

Simulation: I used 2013 data collected from the SHL to simulate both probabilities. These data include

- 26,630 female chlamydia test outcomes
- 7,481 male chlamydia test outcomes.

Figure 17.2 (above) shows the proportion of positive cases over the course of the 2013 calendar year for both sexes. What would we estimate the probability of a positive chlamydia outcome to be for females? for males?

Summary: Probability describes the proportion of times an outcome would occur in a very long series of independent repetitions.

- If the probability of an outcome is 0.25 (or 25%), then we would expect the outcome to occur about 25% of the time over the long run.
- This does not mean the outcome will occur exactly 1 time for every 4 repetitions. This is a "short run" statement that considers only 4 repetitions.

Fact: A probability is a number between 0 and 1. It can also be converted to a percentage (e.g., 0.25 ⟶ 25%).

- The larger the probability for a particular outcome, the more likely it is to occur.
- The smaller the probability for a particular outcome, the less likely it is to occur.
- Outcomes with probability 1 will always occur. Outcomes with probability 0 will never occur.
  - Can you think of practical examples of these extremes? It's harder than you might think!

Example 17.3. Vioxx was introduced by Merck in 1999 as an alternative to naproxen for treating pain due to osteoarthritis. It was marketed as an anti-inflammatory drug that would be "as effective" as naproxen but would reduce the incidence of gastrointestinal complications and pain (which it did). The problem is that it also increased the incidence of heart attacks. In Merck's VIGOR study,

- there were 20 heart attacks out of 4,047 patients taking Vioxx.
- there were 4 heart attacks out of 4,029 patients taking naproxen.

If there was really no difference between the heart attack rates for the two drugs, the probability of an outcome this extreme in favor of Vioxx is about 0.0008, or about 8 out of 10,000.

Q: Could something like this have happened strictly by chance?
A: Yes, but it is very unlikely because the probability 0.0008 is so small. Merck paid a big price for it too.

17.2 Probability myths

Review: Probability provides a language to describe the long-term regularity of random behavior. It does not describe what happens in the "short term."

Myth 1: *The myth of short-term regularity.* Probabilities stabilize in the long run; see Figures 17.1 and 17.2. Some people will confuse this interpretation with what happens in the short run.

Example 17.4. Consider the following examples:

1. At a roulette wheel, the following colors have been observed in the last 10 plays:

```
R G B B B B B B B B
```

Q: You walk up to the table with this information. What should you bet on the next play?
A: If the plays are independent, then the previous 10 plays don't matter (this is "short-run" thinking). The probability of the outcomes remains the same for each play:

Red: 18/38 Black: 18/38 Green: 2/38.

Remark: Some people will argue that because the previous 8 plays have been a "B," some mysterious "law of probability" must swing the next play back to a different outcome (R or G). This is an incorrect conjecture about short-term behavior (i.e., the last 8 plays).

2. I toss a fair coin 10 times. Which outcome looks the most random?

   `HTHTHTHTHT` `HHHHHHHHHH` `HTTHTHTTHH`

   Remark: Some people will argue the first two outcomes are not random; the first alternates perfectly and the second one is all heads. In fact, all three outcomes have the same chance of occurring:

$$\left(\frac{1}{2}\right)^{10} = \frac{1}{1024}.$$

3. What does it mean for an outcome to be **random**? Pick a "random" number (integer) between 1 and 100.

   Remark: If you ask people this question, chances are they will not give a "random" number. It will be a number that, to them, *seems* random. True randomness means that each number between 1 and 100 has the same chance of occurring. Figure 17.4 simulates this exercise for 100,000 repetitions.

4. Growing up in Iowa, I always watched Chicago Cubs games which were announced by the late Harry Caray. When the Cubs were behind late in the game, Harry would try to rally the viewers at home by making statements like,

   > *"At bat is Sandburg, who is hitting .250 for the season. He's 0 for 3 today, so he is due for hit."*

   Remark: A batting average can be interpreted as a probability of getting a hit. This refers to long-term behavior over the course of the entire season; not what Sandburg will do on the next at bat.

5. Empirical data suggests boys and girls are born at roughly the same proportions (about 50/50). A couple's first three children are boys.

   Q: What is the probability the next child will be a girl?
   A: It is still 0.5. It doesn't matter what happened on the first three births.

   Remark: Having children is like flipping coins or rolling dice. Coins and dice have no memory. A coin doesn't know the first 3 flips were tails, and it doesn't know to "try to even things out" on the next toss.

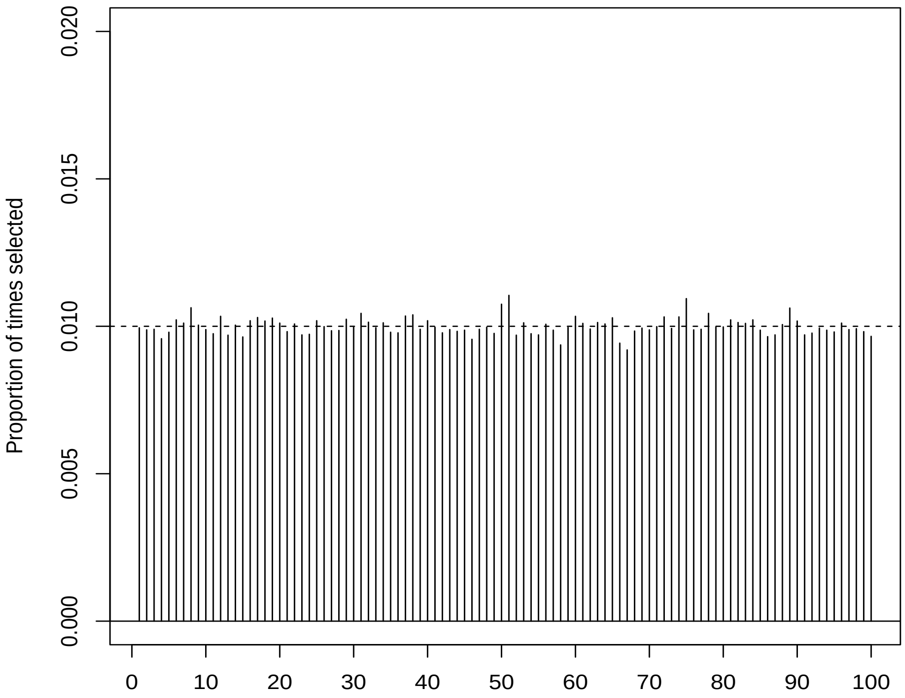


Figure 17.3: Distribution of the proportion of 100,000 randomly selected numbers among 1, 2, ..., 100. This represents what could happen if the process of selecting a "random number between 1 and 100" was truly random.

Myth 2: *The myth of surprising coincidences.* Certain outcomes may be unlikely, but this does not mean they cannot occur. Even outcomes with small probability do occur from time to time. When "surprising" events occur (especially if they are bad), it is common to seek out a *cause* for why they happened. Most of the time, surprising outcomes are explained by random chance and nothing more.

Example 17.5. Consider the following examples:

1. *Winning the lottery twice.* Winning any state or national lottery is unlikely. The probability of winning one twice is incredibly small. However, it can and does happen.
   - In 1986, Evelyn Adams won the NJ state lottery for a second time ($1.5 and $3.9 million payoffs).
   - Robert Humphries (PA) won his second lottery two years later ($6.8 million total).

2. *The Birthday Problem.* Chances are, there are two people sitting in this room with the same birthday (e.g., April 15). Does this sound surprising? It might, but it isn't surprising at all. For n different people (no multiple births), the probability there will be at least one shared birthday is given in the table below:

n	Probability	n	Probability
2	0.0027	28	0.6544
4	0.0163	30	0.7063
6	0.0404	32	0.7533
8	0.0743	34	0.7953
10	0.1169	36	0.8321
12	0.1670	38	0.8640
14	0.2231	40	0.8912
16	0.2836	50	0.9703
18	0.3469	60	0.9941
20	0.4114	70	0.9991
22	0.4756	80	>0.9999
24	0.5383	90	>0.9999
26	0.5982	100	>0.9999

Interesting: You only need $n = 23$ people to have the probability of a shared birthday be at least 0.5 (50%). Most people think you need many more people than this.

3. *The Ohio lottery.* On November 18, 2006, Ohio State beat Michigan 42-39 in college football. This game was of particular interest because OSU was ranked #1 and Michigan was ranked #2 at the time. The very next day, the Ohio State lottery number chosen was 4239−identical to the score of the game. A CNN announcer, when reporting the "anomalous" lottery selection said,

    "I'm not a mathematician, but what is the chance of that?"

    The probability of the outcome "4239" is 1/10000−the same as any other outcome. The CNN announcer made a big deal out of something that isn't. There is also some "data snooping" going on here. The number 4239 was only on the CNN announcer's mind because of the football game the day before. What if the score of the game had been something else?

4. *Streaks.* One of the most stunning winning streaks in MLB took place in 2002 when the Oakland A's won 20 games in a row, breaking the all-time American League record. What makes the streak more interesting is that the team consisted of "value players" only; i.e., players that had been chosen based on statistics not commonly used in MLB at the time (and few/no superstars). The story behind this team was the subject of the motion picture blockbuster *Moneyball.*

    Q: What is the probability of a MLB team winning 20 games in a row?
    A: This would be difficult to calculate without making restrictive assumptions. It's

very small. Based on where the Oakland A's were that year in terms of overall winning percentage when the streak started, an approximation based on independent game outcomes is about 1 in 1 million.

Interesting: The 2017 Cleveland Indians won 22 games in a row, setting a new American League record.

5. *Cancer clusters.* In 1984, residents of Randolph, MA, counted 67 cancer cases in their 250 residences. This cluster of cancer cases seemed unusual, especially because there was a nearby chemical plant (fearing that the water supply had been contaminated). However, given that cancer was the cause of about 23% of the deaths in the US at the time, 67 out of 250 cases (about 27%) is not all that unusual. Random chance emerged as the best explanation.

17.3 The law of averages

Remark: Suppose we observe a **simple random sample** (SRS) from a population of individuals. Informally, the law of averages states that

- the sample proportion $\widehat{p}$ will get close to population proportion p when the sample size gets larger
- the sample mean $\overline{x}$ will get close to population mean μ when the sample size gets larger.

Recall: In Chapter 3, we learned when we have a simple random sample (SRS) of individuals, the sample proportion $\widehat{p}$ will be an unbiased estimate of the population proportion p.

- "Unbiased estimate" means the sample proportion $\widehat{p}$ will estimate p correctly on average over many hypothetical random samples.
- Therefore, the law of averages essentially is a statement about what happens to the variability of $\widehat{p}$ as the sample size n gets larger. Let's examine this by recalling the margin of error.

  Result: If a SRS is used, the margin of error in the sample proportion $\widehat{p}$ associated with a 95% confidence level is approximately equal to

  $$\textbf{margin of error} = \frac{1}{\sqrt{n}},$$

  where n is the sample size. This formula is only applicable for a SRS design with a 95% level of confidence.

Q: What happens to the margin of error when the sample size increases? Let's see:

$$\begin{aligned}
n = 100 &\implies \frac{1}{\sqrt{100}} = 0.10\\
n = 1000 &\implies \frac{1}{\sqrt{1000}} \approx 0.03\\
n = 10000 &\implies \frac{1}{\sqrt{10000}} = 0.01\\
n = 100000 &\implies \frac{1}{\sqrt{100000}} \approx 0.003\\
n = 1000000 &\implies \frac{1}{\sqrt{1000000}} = 0.001.
\end{aligned}$$

Revelation: If the sample proportion $\widehat{p}$ is unbiased, and its variability (margin of error) tends towards to 0 as the sample size gets larger and larger, then it must get close to the population proportion p. This is what the law of averages says will happen.

Remark: A similar phenomenon occurs for the sample mean $\overline{x}$ calculated from an SRS. It will get close to the population mean μ when the sample size becomes larger and larger.

Example 17.6. In Example 13.4, we described how the World Health Organization uses a normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 15$ to describe the systolic blood pressure (SBP, measured in mm Hg) of American males (aged 18 and over).

- This population density curve is shown on the next page (top, left).
- To show how the law of averages works, I used R to simulate SBP observations from a simple random sample of 10,000 American males from this population distribution.
- I kept track and updated the sample mean $\overline{x}$ as each American male was sampled.
- I then prepared a line graph of the sample means $\overline{x}$ as they were updated. This graph is on the next page (top, right).

Q: What does the sample mean SBP $\overline{x}$ get close to as the sample size gets larger and larger?
A: The population mean $\mu = 125$. This is what the law of averages says should happen.

Important: The law of averages has important implications for **statistical inference**. It guarantees that when we observe a simple random sample (SRS) from a population,

- the (sample) statistic $\widehat{p}$ will be a better estimate of the (population) parameter p when the sample size n gets larger
- the (sample) statistic $\overline{x}$ will be a better estimate of the (population) parameter μ when the sample size n gets larger.

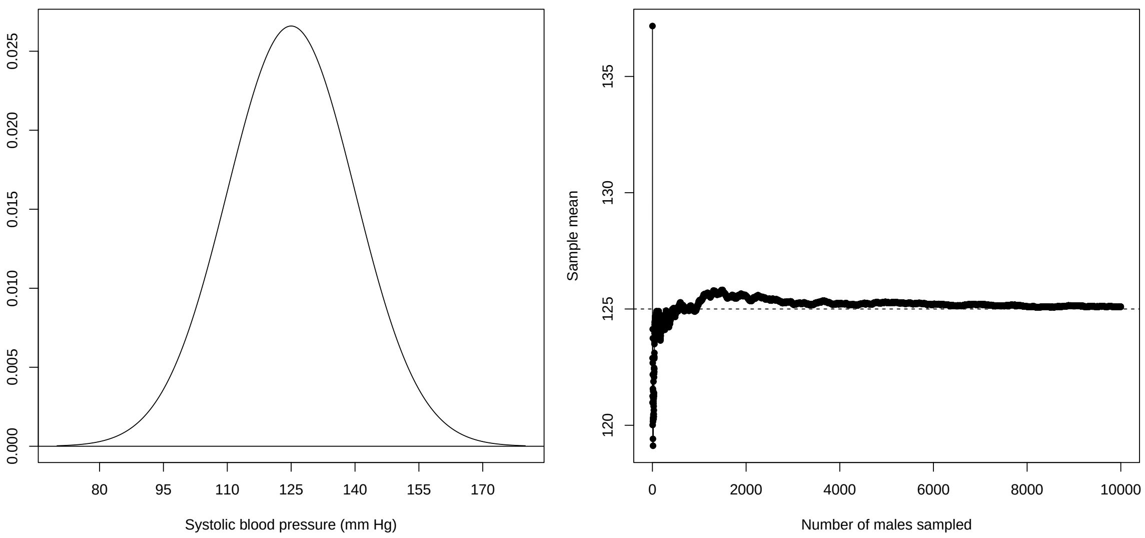


Figure 17.4: SBP for American males. Left: Normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 15$. Right: Line graph of sample means $\overline{x}$ calculated as each American male is sampled from this population.

17.4 Personal probabilities

Example 17.7. On April 1, 2024, the NCAA basketball final four teams were determined for both men and women:

- Men: Connecticut, Alabama, Purdue, NC State
- Women: South Carolina, NC State, Iowa, Connecticut.

Based on my knowledge of college basketball, I assigned probabilities to the teams I thought would win the championship. For the men,

Team	Connecticut	Alabama	Purdue	NC State
Probability	0.70	0.08	0.20	0.02

For the women,

Team	South Carolina	NC State	Iowa	Connecticut
Probability	0.40	0.10	0.40	0.10

Discussion: The only satisfaction one should get from these tables is that the probabilities add up to 1 for both men and women separately. Other than that, the numbers in the tables are based on my personal feeling and judgement. They are not based on our definition of probability as a "long-run proportion." Your assignment of probabilities might be different than mine.

- We cannot assign probabilities in this example by performing the NCAA Final Four a large number of times and keeping track of the proportion of times a certain team would win.
- The numbers in the tables are examples of **personal probabilities**. They are based on one's personal feeling about the chance that something will happen.
- Personal probabilities and probabilities determined as "long-run proportions" don't always align. The former can be biased by personal feeling or judgement. The latter is more mathematically satisfying, but not always practical. The Final Four is played only one time−not a large number of times.

Q: Is a personal probability even a bona fide probability?
A: It might be to you if it's yours, but probably not otherwise. The real danger of a personal probability is that we use them to make decisions as we go about our daily life. If these personal probabilities are based on biased information, then we could end up making bad decisions.

Remark: Personal probabilities aren't all bad, and they can be updated as our personal feeling and judgement change. For example, suppose you woke up this morning with a sore throat.

- You find out from your social media accounts that "strep throat is going around." You assign a personal probability of 0.50 to the outcome you have strep throat.
- You then look in the mirror to see what your throat looks like. Your tonsils appear to be red and swollen. You update your personal probability to 0.70.
- Later that day, your doctor tells you, "it looks like strep throat but I can't be sure." You update your personal probability to 0.80.
- An in-house strep test at your doctor's office says you don't have strep. You update your personal probability to 0.10.
- A culture is performed on a swab specimen from your tonsils and is negative for strep bacteria. You update your personal probability to 0.01. Based on this, you conclude you likely don't have strep throat.

Remark: As we continue to get more information, our personal feelings change so personal probabilities change along with it to incorporate that information. This is the central theme behind **Bayesian statistics**. This is a framework for how we update probabilities in the presence of new information (data). An important part of this framework is **Bayes' Rule**, which is taught in statistics courses like STAT 201.

17.5 Probabliity and risk

Example 17.8. Probabilities are often used to characterize **risk**. One can think of "risk" as being exposed to danger, harm, or even death. The Insurance Information Institute reports the lifetime odds of death from different types of accidental causes:

Accident	Odds
Drug poisoning/overdose	1 in 43
Motor vehicle	**1 in 93**
Firearm assault	1 in 208
Exposure to smoke/fire	1 in 1287
Falls from steps/stairs	1 in 1577
Drowning	1 in 5186
Fall from ladder/scaffolding	1 in 7830
Firearm accidental discharge	1 in 9522
Airplane/space transport	**1 in 11240**
Weather-related storm	1 in 20098
Flood	1 in 45908
Dog bite/attack	1 in 53843
Earthquake	1 in 99120

Remark: Before we get into this table, first note the "odds" of something happening is not the same as the probability it happens, although many people think these two terms are synonymous. If the probability of an outcome is p, then the odds of the outcome is

$$\text{odds} = \frac{p}{1-p}.$$

Therefore, if the odds of drug poisoning/overdose is 1 in 43, then

$$\text{odds} = \frac{1}{43} = \frac{1/44}{43/44} \implies p = \frac{1}{44} \approx 0.023.$$

In other words, an odds of "1 in 43" means 1 death for every 43 non-deaths. More commonly, this would be written as "43 to 1."

Q: Why are some people scared of flying but not of driving?
A: The chance of dying in an automobile accident is about 120 times greater!

Remark: You are much more likely to die from heart disease (ranks 1st) or cancer (ranks 2nd). Why don't we worry more about these? The risk of these causes of death is much greater than those shown above. Shouldn't our worries and fears align better with the levels of risk involved? As Moore and Notz put it (pp 418),

> *"Few of us would leave a baby sleeping alone in a house while we drove off on a 10-minute errand, even though car-crash risks are much greater than home risks."*

18 Probability Models

18.1 Probability models for populations

Terminology: A **probability model** for a population of individuals is a description including two things:

- a collection of possible outcomes for individuals in the population
- a probability for each outcome.

Example 18.1. The following table describes a probability model for weight categories of all American males (aged 18 and over):

Outcome	Underweight	Healthy	Overweight	Obese
Probability	0.01	0.32	0.43	0.24

This model lists four **outcomes**: underweight, healthy, overweight, and obese.

Conceptualization: Imagine observing one American male randomly selected from this population. The probabilities for each outcome are 0.01, 0.32, 0.43, and 0.24, respectively. Note these probabilities add up to 1. This is a requirement for any probability model.

Probability rules: For a probability model to be valid, the following must be true:

1. Each probability must be between 0 and 1.
2. All probabilities taken together must add to 1.
3. If two events have no outcomes in common, the probability either event will occur is the sum of the individual probabilities.

Q: What is the probability an American male randomly selected from the population is either overweight or obese?
A: We add the probabilities

$$P(\text{Overweight}) + P(\text{Obese}) = 0.43 + 0.24 = 0.67.$$

We use the letter "P" as short for "probability." The notation

$$P(\text{Overweight})$$

is read, "the probability of overweight." The notation

$$P(\text{Obese})$$

is read, "the probability of obese."

Example 18.2. The Centers for Disease Control and Prevention uses the following probability model to describe the acquisition of hepatitis C (HCV) among all American adults:

Outcome	Probability
IVDU	0.60
Unprotected sex	0.25
Transfusion-related	0.05
Occupational	0.03
Other/Unknown	??

This model lists five **outcomes**: IVDU, unprotected sex, transfusion-related, occupational, and other/unknown.

Q: What is the correct probability for the outcome "Other/Unknown?"
A: The probabilities taken together must add to 1. Therefore,

$$0.60 + 0.25 + 0.05 + 0.03 + \ ?? \ = 1 \implies P(\text{Other/Unknown}) = 0.07.$$

Conceptualization: Imagine observing one HCV-infected American randomly selected from the population of all HCV-infected Americans. The probabilities for each outcome are 0.60, 0.25, 0.05, 0.03, and 0.07, respectively.

Example 18.3. In the game of craps, two fair dice are rolled initially to start the game. There are

$$6 \times 6 = 36$$

different outcomes from rolling the two dice (see next page). Here is a probability model for the **sum** of the two faces.

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

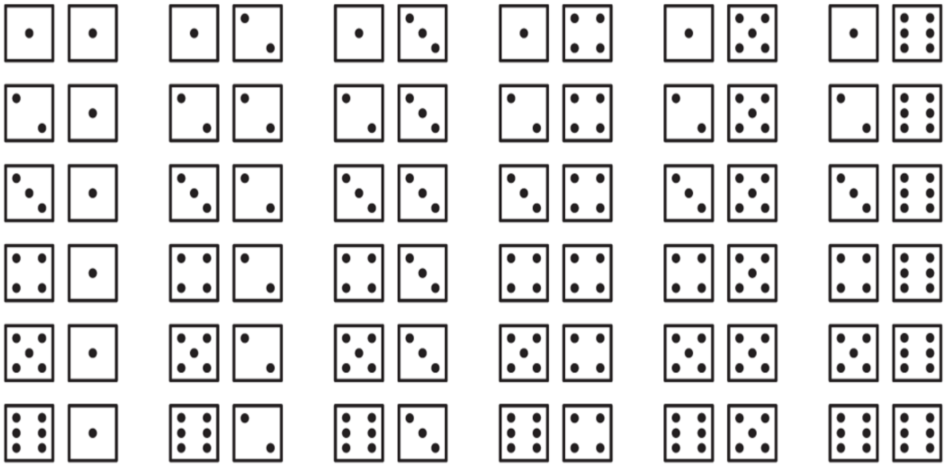
Q: What is the probability of rolling a "7" or an "11?"
A: We add the probabilities

$$P(\text{rolling a 7 or 11}) = P(7) + P(11) = \frac{6}{36} + \frac{2}{36} = \frac{8}{36} \approx 0.22.$$

Q: What is the probability of **not** rolling a "7" or an "11?"
A:

$$P(\text{not rolling a 7 or 11}) \approx 1 - 0.22 = 0.78.$$

Remark: The last calculation is an example of the **complement rule** in probability: "The probability an event doesn't occur is 1 minus the probability it does occur."



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Recall: Population density curves are curves that describe the distribution of a quantitative variable for all individuals in a population. Any population density curve has these three properties:

1. the curve is non-negative (i.e., it must reside above the horizontal axis).
2. the total area under the curve is 1 (or 100%).
3. the **area** under the curve over a given range represents the **proportion** of individuals in the population that fall in that range.
    - You can also interpret this area as the **probability** a single individual in the population falls in that range.

Revelation: Population density curves are probability models for individuals in a population!

Example 18.4. The distribution of hours of sleep per school night among high-school seniors is normally distributed with mean $\mu = 6.6$ hours and standard deviation $\sigma = 1.3$ hours. This is a probability model for all high-school seniors (the population).

Q: What is the probability a randomly selected high-school senior from this population sleeps more than 8 hours?
A: This is a "normal calculation" just like we did in Chapter 13.

We first calculate the **standard score**:

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{8 - 6.6}{1.3} \approx 1.1.$$

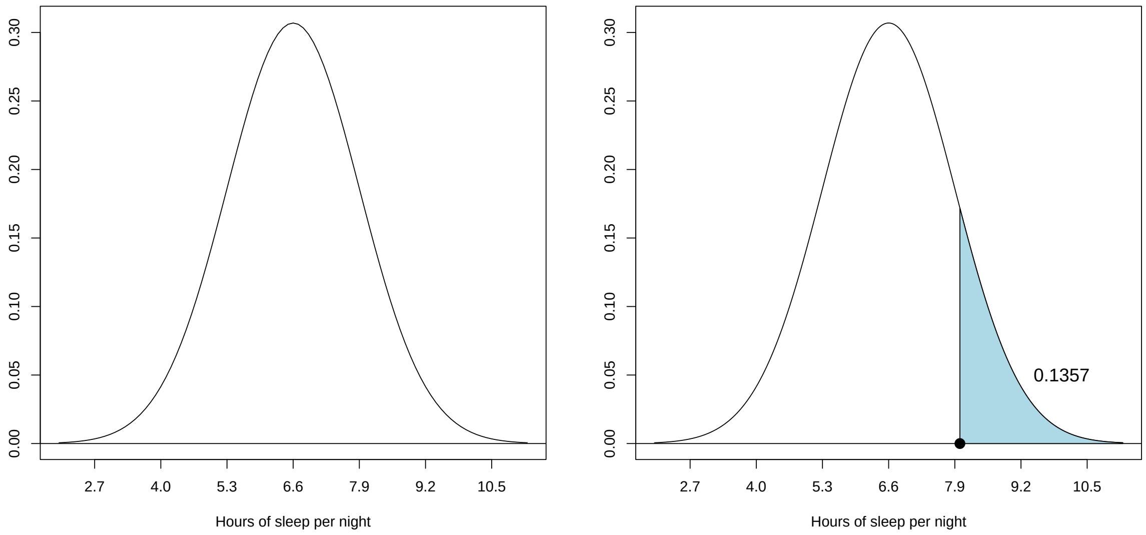


Figure 18.1: Left: Normal population density curve with mean $\mu = 6.6$ hours and standard deviation $\sigma = 1.3$ hours. Right: The area to the right of 8 hours is shaded (this is the probability).

Moore and Notz's Table B gives the percentage of the population **to the left** of the standard score $z = 1.1$:

$$\text{Percentage} = 86.43\% \longrightarrow \text{probability of getting } \underline{\text{less than}} \text{ 8 hours of sleep.}$$

The percentage of the population to **to the right** of $z = 1.1$ is

$$\begin{aligned}\text{Percentage} &= 100\% - 86.43\% \\ &= 13.57\% \longrightarrow \text{probability of getting } \underline{\text{more than}} \text{ 8 hours of sleep.}\end{aligned}$$

As a probability (between 0 and 1), this is 0.1357; see Figure 18.1 above.

Implementation in R:

```
> 1-pnorm(1.1)
[1] 0.1357
```

Note: Normal population density curves are probability models for populations when the quantitative variable of interest (e.g., hours of sleep per night) follows a normal distribution.

Exercise: In Example 18.4, find the probability a randomly selected high-school senior gets less than 4 hours of sleep. You can use the 68-95-99.7 rule to determine this.

18.2 Probability models for sampling

18.2.1 Sampling distribution of the sample proportion $\widehat{p}$

Example 18.5. During September 28-29, 2016, Rasmussen Reports conducted a national telephone and online survey using a sample of $n = 1000$ American adults. Each participant was asked:

> *Do you trust the media to accurately fact-check the presidential candidates' comments?*

The survey found 290 of the 1000 adults in the sample answered "Yes" to this question. In other words, the **sample proportion** is

$$\widehat{p} = \frac{290}{1000} = 0.29 \quad \text{(or 29\%)}.$$

Discussion: Suppose the **population proportion** is $p = 0.30$. In other words, among all 255 million American adults (aged 18 and over), suppose that 30% of this population trusts the media to accurately fact-check presidential candidates.

- Of course, we don't know if this is correct, but we are just exploring.
- Under the assumption that $p = 0.30$, we will use simulation to see what sample proportions $\widehat{p}$ we *could get* with 1000 Americans sampled.
- We did similar simulations in Chapter 3 assuming a simple random sample (SRS). We will make the same assumption here.

Results: I used R to simulate 10,000 values of the sample proportion $\widehat{p}$ under the assumption that $p = 0.30$. Here are the first 100 sample proportions I got:

```
> round(sample.prop,2) # round to 2 dp
  [1] 0.30 0.28 0.31 0.31 0.28 0.28 0.31 0.30 0.29 0.29 0.31 0.28 0.29 0.30
 [15] 0.29 0.28 0.32 0.29 0.29 0.29 0.31 0.32 0.29 0.30 0.30 0.29 0.30 0.28
 [29] 0.30 0.31 0.29 0.30 0.28 0.31 0.29 0.29 0.31 0.30 0.32 0.31 0.28 0.30
 [43] 0.30 0.29 0.27 0.28 0.31 0.30 0.30 0.29 0.28 0.30 0.30 0.34 0.29 0.30
 [57] 0.32 0.29 0.29 0.29 0.29 0.28 0.30 0.28 0.30 0.31 0.31 0.29 0.33 0.31
 [71] 0.32 0.29 0.30 0.31 0.30 0.27 0.30 0.32 0.31 0.29 0.31 0.29 0.29 0.30
 [85] 0.30 0.30 0.29 0.29 0.31 0.31 0.30 0.32 0.29 0.27 0.30 0.28 0.32 0.32
 [99] 0.26 0.27
```

I used a histogram to show all 10,000 sample proportions $\widehat{p}$ that were simulated; this is shown in Figure 18.2 (next page).

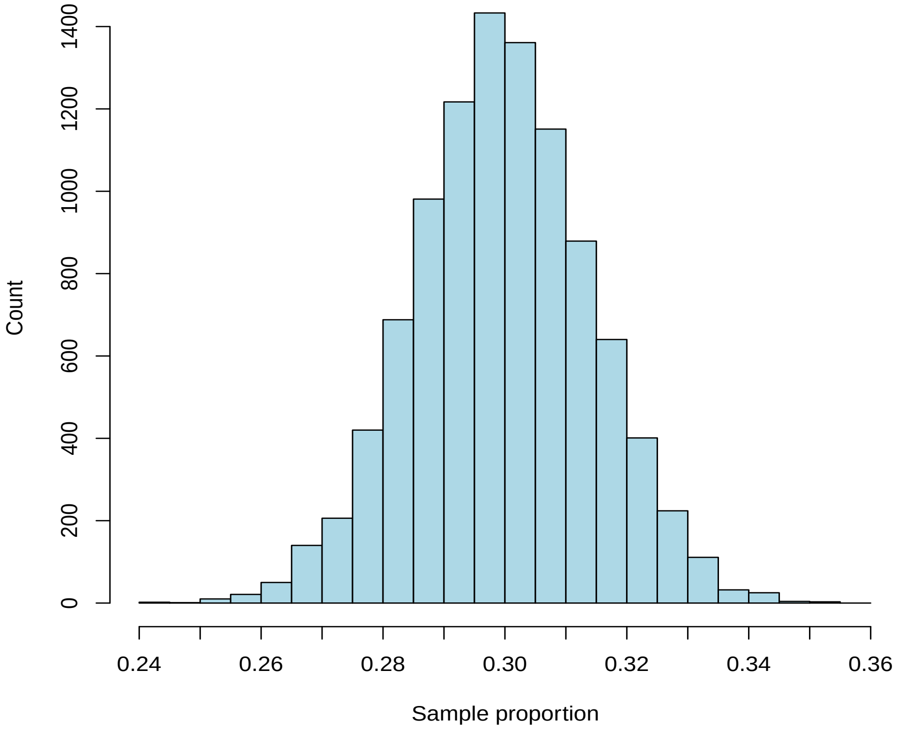


Figure 18.2: Sampling distribution of the sample proportion $\widehat{p}$ when $n = 1000$ and $p = 0.30$.

Main point: The value of the sample proportion $\widehat{p}$ will vary from sample to sample. It therefore makes sense to think about the following question:

*"What **probability model** describes how the sample proportion $\widehat{p}$ behaves?"*

Result: Take a SRS of size n from a large population of individuals, where p is the population proportion. For large samples,

- the sampling distribution of $\widehat{p}$ is described by a normal density curve
- the **mean** of the sampling distribution is p
- the **standard deviation** of the sampling distribution is

$$\sqrt{\frac{p(1-p)}{n}}.$$

Illustration: The normal density curve with mean $p = 0.30$ and standard deviation

$$\sqrt{\frac{0.30(1-0.30)}{1000}} \approx 0.014.$$

is shown in Figure 18.3 (next page).

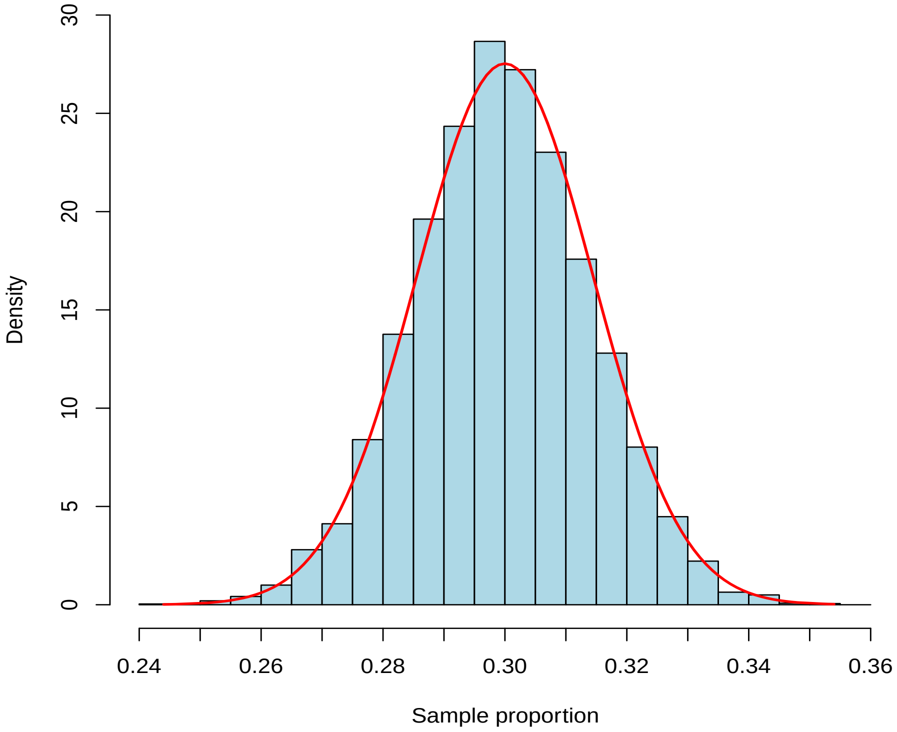


Figure 18.3: Sampling distribution of the sample proportion $\widehat{p}$ when $n = 1000$ and $p = 0.30$. A normal density curve has been superimposed.

Exercise: Use the 68-95-99.7 rule to form intervals 1, 2, and 3 standard deviations from the mean. Interpret.

Solution. The normal density curve above has mean 0.30 and standard deviation 0.014. Therefore,

- 68% of the sample proportions $\widehat{p}$ should be within $(0.286, 0.314)$.
  - Recall that Rasmussen's sample proportion $\widehat{p} = 0.29$ falls in this interval.
  - Was this poll result inconsistent with $p = 0.30$?
- 95% of the sample proportions $\widehat{p}$ should be within $(0.272, 0.328)$.
- 99.7% of the sample proportions $\widehat{p}$ should be within $(0.258, 0.342)$.

Q: Which sample proportion $\widehat{p}$ below would be inconsistent with the hypothesis that the population proportion is $p = 0.30$?

(a) $\widehat{p} = 0.30$
(b) $\widehat{p} = 0.28$
(c) $\widehat{p} = 0.32$
(d) $\widehat{p} = 0.24$

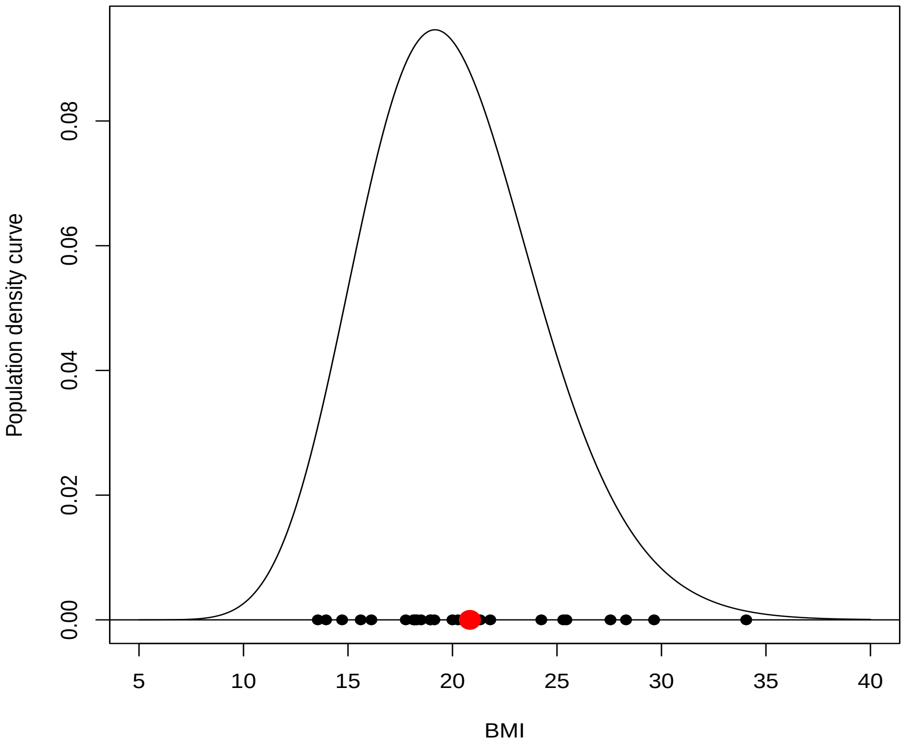


Figure 18.4: Population density curve for the BMI of fourth-grade children in Augusta, GA. A simple random sample of $n = 25$ BMI observations is shown using dark circles. The sample mean $\overline{x}$ is shown using a larger red circle.

18.2.2 Sampling distribution of the sample mean $\overline{x}$

Example 18.6. Figure 18.4 shows the population density curve for the BMI of fourth-grade children in Augusta, GA. This density curve is an example of a gamma distribution and has

$$\begin{aligned} \mu &= 20.1 \longleftarrow \text{mean} \\ \sigma &= 4.3 \longleftarrow \text{standard deviation.} \end{aligned}$$

These are population-level parameters that describe the mean and standard deviation of all fourth-grade children in the population.

Sample: A simple random sample of $n = 25$ fourth grade children is observed and the BMI of each child is determined (see Figure 18.4). Here are the summaries:

```
> mean(bmi) # sample mean
[1] 20.8
> sd(bmi) # sample standard deviation
[1] 5.1
```

These are the statistics $\overline{x}$ and s, respectively, calculated from the sample of 25 children.

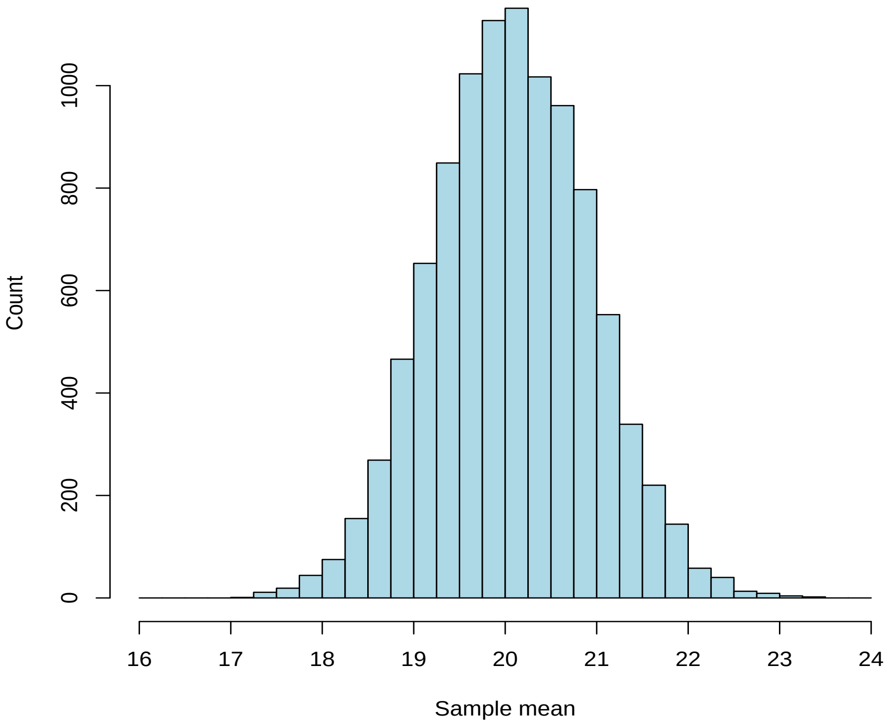


Figure 18.5: Sampling distribution of the sample mean $\overline{x}$ in Example 18.6 when $n = 25$.

Discussion: Let's do another simulation. This time, we will simulate 10,000 samples of $n = 25$ children from the population with mean $\mu = 20.1$ and $\sigma = 4.3$.

- We assume the gamma distribution in Figure 18.4 is an accurate probability model for the population of all fourth-grade children in Augusta, GA.
- With each sample, we will calculate the sample mean $\overline{x}$.

Results: I used R to simulate 10,000 values of the sample mean $\overline{x}$. Here are the first 100 sample means I got:

```
> round(sample.mean[1:100],1) # round to 1 dp
  [1] 20.9 21.3 18.4 19.0 18.8 20.6 21.0 21.1 20.0 19.1 19.3 20.9 19.7 19.2
 [15] 20.8 19.3 19.6 18.8 20.0 18.7 19.2 21.4 19.6 19.0 19.3 19.4 19.7 19.0
 [29] 20.6 22.0 20.0 19.0 20.2 19.5 20.3 19.0 19.2 20.9 20.3 19.1 20.5 19.9
 [43] 19.7 20.7 19.8 19.9 18.9 19.0 20.5 19.6 19.5 18.6 20.6 20.4 19.3 20.6
 [57] 20.4 18.8 20.6 19.8 20.1 21.9 20.1 20.6 20.6 21.5 20.5 20.7 22.3 20.0
 [71] 20.0 20.3 19.6 19.0 21.0 21.0 18.6 18.9 20.6 19.0 18.8 19.2 19.9 19.6
 [85] 20.5 19.6 20.6 19.2 19.3 20.3 19.2 19.5 20.1 20.2 20.9 20.1 19.2 20.5
 [99] 20.3 20.0
```

I used a histogram to show all 10,000 sample means $\overline{x}$ that were simulated; this is shown in Figure 18.5 (above).

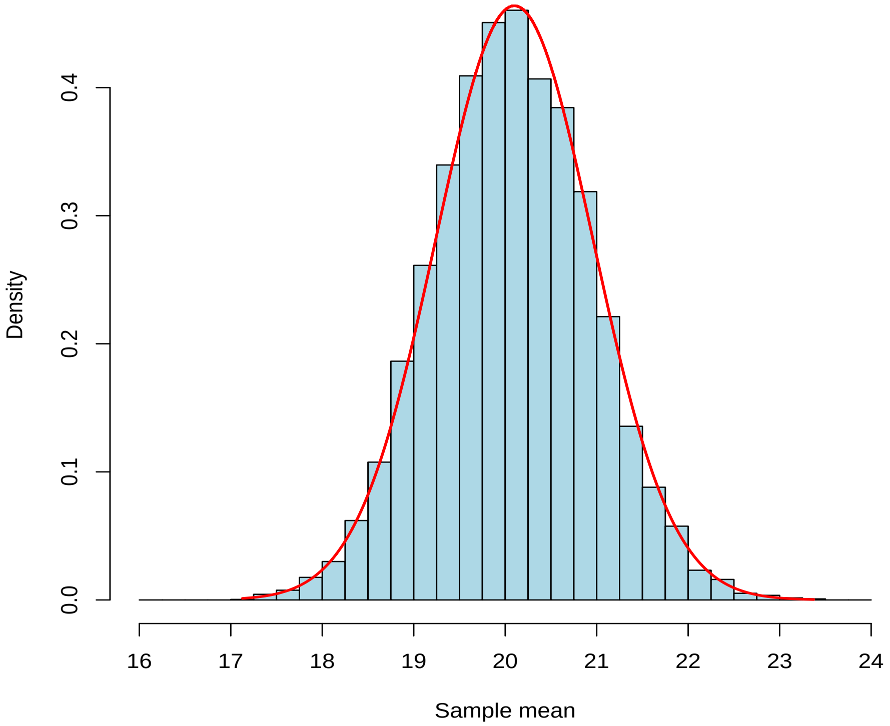


Figure 18.6: Sampling distribution of the sample mean $\overline{x}$ in Example 18.6 when $n = 25$. A normal density curve has been superimposed.

Main point: The value of the sample mean $\overline{x}$ will vary from sample to sample. It therefore makes sense to think about the following question:

*"What **probability model** describes how the sample mean $\overline{x}$ behaves?"*

Result: Take a SRS of size n from a large population of individuals with mean μ and standard deviation σ. For large samples,

- the sampling distribution of $\overline{x}$ is described by a normal density curve
- the **mean** of the sampling distribution is μ
- the **standard deviation** of the sampling distribution is

$$\frac{\sigma}{\sqrt{n}}.$$

Illustration: The normal density curve with mean $\mu = 20.1$ and standard deviation

$$\frac{4.3}{\sqrt{25}} = 0.86$$

is shown in Figure 18.6 (above).

21 What is a Confidence Interval?

21.1 Introduction

Preview: Confidence intervals are a form a **statistical inference**. Recall that statistical inference is the process of using sample information (statistics) to estimate characteristics of a population (parameters). The root word of "inference" is "infer." We would like "to infer" something about a population.

Remark: Mathematicians and engineers do this all the time with functions and equations. What makes statistics different is that we acknowledge real life variability, and, as such, there will always be uncertainty with our conclusions. The following excerpt, lifted liberally from Moore and Notz (pp 485), summarizes this idea succinctly:

> *"Drawing conclusions in mathematics is a matter of starting from a hypothesis and using logical argument to prove without doubt that a conclusion follows. Statistics isn't like that. Statistical conclusions are uncertain because the sample isn't the entire population. So statistical inference has to not only state conclusions but also say how uncertain they are."*

Informally, a confidence interval is an interval of numbers that we "think" should include a population parameter. We use the term "think" because, as the excerpt above reminds us, we can't be 100% sure. However, we have the ability to construct the interval in a way that *should* include the parameter with a high degree of certainty (probability). We start by revisiting our oft-discussed survey question from Chapter 3.

Example 3.1 (continued). During December 3-5, 2023, Rasmussen Reports conducted a national telephone and online survey using a sample of $n = 992$ American adults. Each participant was asked,

Should Christmas be celebrated in public schools?

The survey found 685 of the 992 adults in the sample answered "Yes" to this question.

- The **sample proportion** is

$$\widehat{p} = \frac{685}{992} \approx 0.69 \quad \text{(or 69\%)}.$$

  The sample proportion $\widehat{p}$ is a statistic. It is calculated from the 992 individuals in the sample.

- The **population proportion** is

$$p = \text{proportion of } \underline{\text{all American adults}} \text{ who agree Christmas should be celebrated in public schools.}$$

The population proportion p is a parameter. It describes the population of all American adults (about 255 million of us). Because we don't get to see every individual in the population, the population proportion p is unknown.

- We use the sample proportion $\widehat{p}$ to **estimate** the population proportion p.

Recall: If a simple random sample (SRS) is used, the margin of error in the sample proportion $\widehat{p}$ associated with a 95% confidence level is approximately equal to

$$\textbf{margin of error} = \frac{1}{\sqrt{n}},$$

where n is the sample size. This formula is only applicable for a SRS design with a 95% level of confidence. If either of these changed, then the formula would change (we will discuss this in this chapter).

Recall: We used the margin of error to write **confidence statements** for the population proportion p. If Rasmussen used a SRS, then the margin of error associated with $\widehat{p}$ is

$$\textbf{margin of error} = \frac{1}{\sqrt{992}} \approx \frac{1}{31.50} \approx 0.03 \text{ (or 3\%)}.$$

Recall the margin of error describes only the random sampling error associated with $\widehat{p}$. This is the natural error that arises because we are sampling from a large population (i.e., different SRSs will produce different samples and hence different values of $\widehat{p}$). The margin of error describes this source of error. The margin of error does not describe non-sampling errors.

Recall: We use the sample proportion $\widehat{p}$ and the margin of error to write the following confidence statement for p:

- "We are 95% confident that the proportion of all American adults who believe Christmas should be celebrated in public schools is between 0.66 and 0.72 (i.e., between 66% and 72%)."

Terminology: A **95% confidence interval** is an interval that is guaranteed to capture a population parameter in 95% of all samples.

Revelation: When we wrote confidence statements for a population proportion p back in Chapter 3, we were actually calculating 95% confidence intervals for p; in Example 3.1, this interval is

$$(0.66, 0.72) \iff 0.66 < p < 0.72.$$

However, note that in Chapter 3,

- calculations were restricted to a 95% level of confidence. What if we want to use another level of confidence?

  - e.g., 90% (less confidence), 99% (more confidence), etc.
  - What is the “price” we must pay for being more confident?

- the quick formula for margin of error; i.e.,

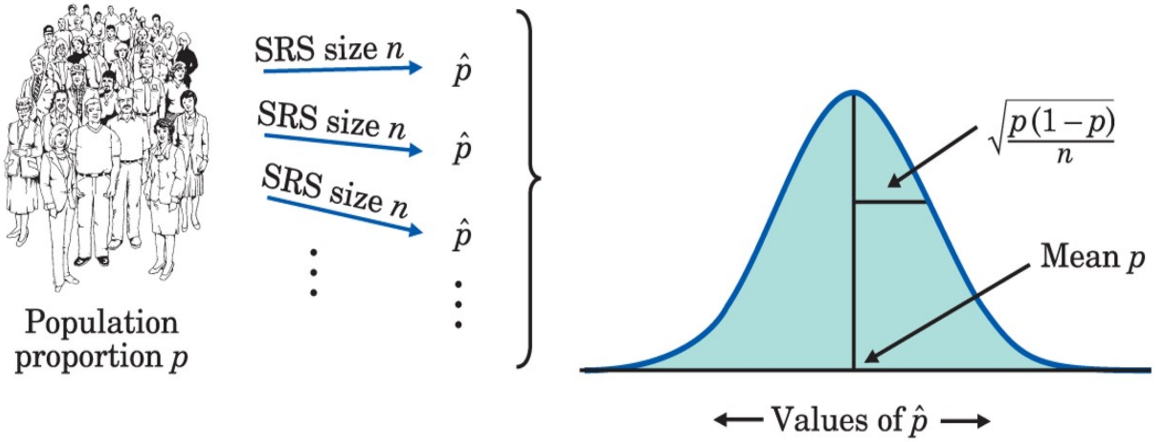
$$\textbf{margin of error} = \frac{1}{\sqrt{n}},$$

  is a conservative approximation for the true margin of error.

Preview: We now aspire to make our confidence interval calculations for p more precise (Section 21.2). We will also discuss how to write a confidence interval for a **population mean** μ (Section 21.3).

21.2 Confidence intervals for a population proportion p

Remark: We begin by recalling an important result from Chapter 18, which is illustrated in the picture below:



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Result: Take a SRS of size n from a large population of individuals, where p is the population proportion. For large samples,

- the sampling distribution of $\widehat{p}$ is described by a normal density curve
- the **mean** of the sampling distribution is p
- the **standard deviation** of the sampling distribution is

$$\sqrt{\frac{p(1-p)}{n}}.$$

Remark: The sampling distribution result on the preceding page tells us the probability model for the sample proportion $\widehat{p}$. In other words, it describes the distribution of all possible sample proportions $\widehat{p}$ we can expect to see when observing simple random samples from a population with (population) proportion p.

- From the 68-95-99.7 rule, we know about 95% of all sample proportions $\widehat{p}$ will be within 2 standard deviations of the mean. In other words, the interval

  $$p \pm 2\sqrt{\frac{p(1-p)}{n}}$$

  will contain about 95% of all possible sample proportions $\widehat{p}$.

- The preceding bullet is an application of the 68-95-99.7 rule. It's important too, because, mathematically, the statement above is the same as the following: "about 95% of the time, the interval

  $$\widehat{p} \pm 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

  will contain the population proportion p."

- This discovery leads to the following result.

Result: Choose a SRS of size n from a large population with proportion p. A **95% confidence interval** for the population proportion p is

$$\widehat{p} \pm 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Example 21.1. The Women's Interagency HIV Study is a large ongoing prospective study, funded by the National Institutes of Health, that follows women in the United States living with HIV. In a recent survey, researchers observed a sample of 1,288 women from this cohort.

- Among the 1,288 women in the sample, 399 of them self-reported having been abused as a child.

- The sample proportion of childhood abuse victims is therefore

  $$\widehat{p} = \frac{399}{1288} \approx 0.31.$$

Q: What is the population here?
A: A reasonable answer is "all HIV positive women living in the United States." Based on 2019 CDC estimates, there are about 258,000 such women.

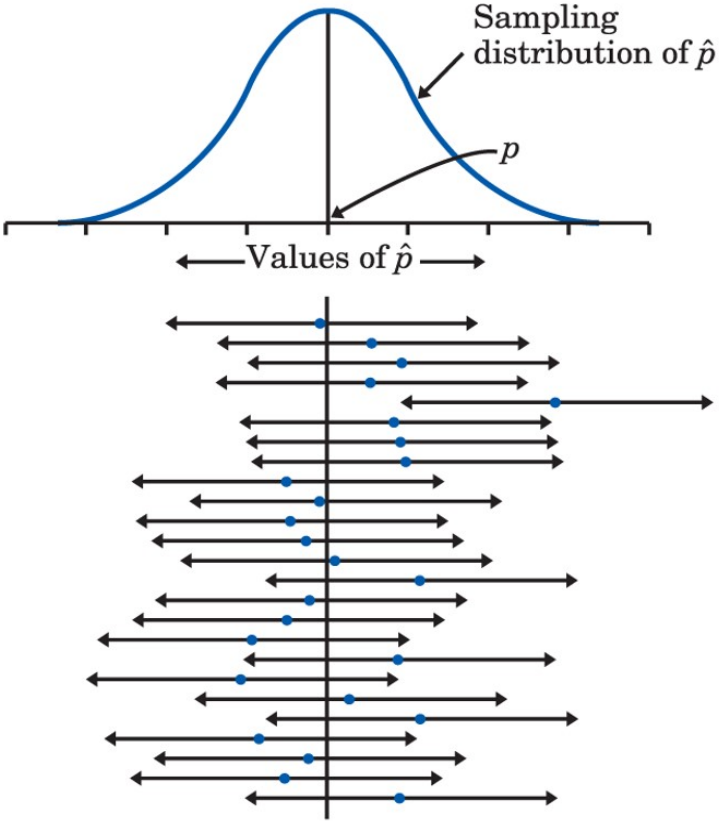
Assuming this sample is a SRS, let's calculate a 95% confidence interval for p, the proportion of women in this population who experienced abuse as a child:

$$\begin{aligned}\widehat{p} \pm 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} &\implies 0.31 \pm 2\sqrt{\frac{0.31(1-0.31)}{1288}} \\ &\implies 0.31 \pm 0.03 \\ &\implies (0.28, 0.34).\end{aligned}$$

Interpretation: We are 95% confident the proportion of women in this population who experienced abuse as a child is between 0.28 and 0.34 (i.e., between 28% and 34%).

Q: What exactly is meant by the phrase "95% confident?"
A: This means 95% of the samples will produce a confidence interval that contains the population proportion p.

- This means that, 5% of the time, our 95% confidence interval will <u>not</u> contain the population proportion!



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Q: A 95% confidence interval for a population proportion p is

$$\widehat{p} \pm 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

When we wrote 95% confidence statements in Chapter 3, we used the formula

$$\widehat{p} \pm \frac{1}{\sqrt{n}}.$$

How do these formulas compare?
A: The first formula is more precise. The margin of error in the first formula

$$2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

is based on the sampling distribution of $\widehat{p}$. The second formula from Chapter 3 is OK to use as an approximation, but it is ultimately conservative (i.e., the interval it produces is a little too wide). Both formulas have the same form:

$$\text{estimate} \pm \text{margin of error}.$$

Q: What if we want to use other levels of confidence (i.e., other than 95%)?

Terminology: A **level** C **confidence interval** is an interval that will contain a population parameter in C% of all samples.

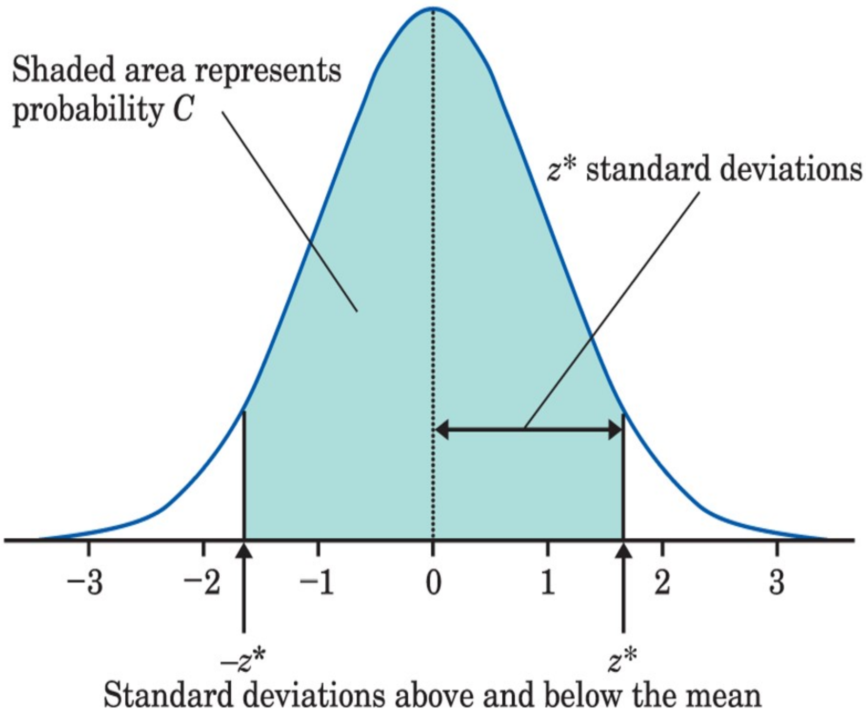
- C is a percentage between 0 and 100. We have already discussed 95% confidence intervals; i.e., $C = 95\%$.
  - $C = 80\%$, $C = 90\%$, $C = 95\%$, and $C = 99\%$ are commonly used; $C = 95\%$ is by far the most common.
  - Larger $C \Longrightarrow$ more confidence.

Result: Choose a SRS of size n from a large population with proportion p. A **level** C **confidence interval** for the population proportion p is

$$\widehat{p} \pm z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},$$

where z^* is a **critical value** which depends on the confidence level C.

- **Q:** What is z^* if $C = 95\%$? **A:** $z^* \approx 2$. We get this from the 68-95-99.7 rule.
- The usefulness of this result is now we can calculate confidence intervals at any level of confidence we want.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Remark: The critical value z^* is a percentile from the normal distribution with mean 0 and standard deviation 1 (Table B in Moore and Notz). Here are values of z^* for commonly used levels of confidence:

C	80%	90%	95%	99%
z^*	1.28	1.64	1.96	2.58

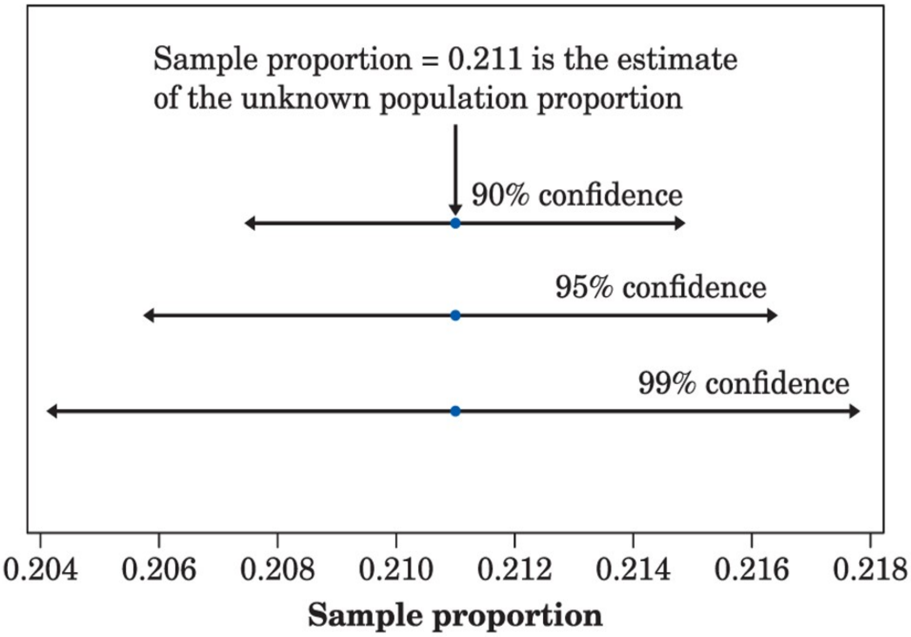
Note that the value of $z^* = 1.96$ for 95% confidence is very close to 2.

- The value $z^* = 1.96$ is an exact value that gives 95% confidence; the value "2" is based on the 68-95-99.7 rule, which is just an approximation.

Example 21.2. The National Survey of Student Engagement (NSSE) asked a sample of 23,915 college seniors (in the United States and Canada) what their immediate plans were after graduation.

- Of the 23,915 seniors who answered this question, 5038 indicated they planned to go to graduate school or professional school (e.g., law school, medical school, etc.).
- The sample proportion is

$$\widehat{p} = \frac{5038}{23915} \approx 0.211.$$



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020
W. H. Freeman and Company

Q: What is the population here?
A: A reasonable answer is "all college seniors in the United States and Canada." Based on estimates from *Statista*, there are about 5-6 million students in this population.

Assuming this sample is a SRS, let's calculate a **95% confidence interval** for p, the proportion of all college seniors in the United States and Canada who plan to go to graduate school or professional school:

$$\begin{aligned} \widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} &\implies 0.211 \pm 1.96\sqrt{\frac{0.211(1-0.211)}{23915}} \\ &\implies 0.211 \pm 0.005 \\ &\implies (0.206, 0.216). \end{aligned}$$

Interpretation: We are 95% confident the proportion of seniors in this population who plan to go to graduate school or professional school is between 0.206 and 0.216 (i.e., between 20.6% and 21.6%).

Exercise: Let's calculate 90% and 99% confidence intervals for p in this problem and compare all three intervals (see figure above). A **90% confidence interval** for p is

$$\begin{aligned} \widehat{p} \pm 1.64\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} &\implies 0.211 \pm 1.64\sqrt{\frac{0.211(1-0.211)}{23915}} \\ &\implies 0.211 \pm 0.004 \\ &\implies (0.207, 0.215). \end{aligned}$$

A **99% confidence interval** for p is

$$\widehat{p} \pm 2.58\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \implies 0.211 \pm 2.58\sqrt{\frac{0.211(1-0.211)}{23915}}$$
$$\implies 0.211 \pm 0.007$$
$$\implies (0.204, 0.218).$$

Observation: The 90% confidence interval is shorter than the 95% confidence interval. The 99% confidence interval is longer. Both of these observations make intuitive sense. The more confidence we require, the longer the intervals must be to guarantee this.

21.3 Confidence interval for a population mean μ

Remark: We now switch gears and discuss how to estimate a **population mean** μ with a level C confidence interval.

- We are still doing statistical inference. The only difference is now we are interested in a different population parameter.
- A population mean μ describes the center of a population density curve; i.e., the curve that describes the distribution of a **quantitative variable** in the population.
- Therefore, we are now interested in means of quantitative variables; e.g., the population mean
  - birth weight (in kg) for premature infants
  - BMI for fourth-graders in Augusta, GA
  - IQ score for American adults
  - arsenic concentration (in ppb) of ground water wells in Texas
  - turtle shell length (in cm) in the Grand Cayman Islands
  - systolic blood pressure (in mm Hg) among American males.

Recall: Take a SRS of size n from a large population of individuals with mean μ and standard deviation σ. For large samples,

- the sampling distribution of $\overline{x}$ is described by a normal density curve
- the **mean** of the sampling distribution is μ
- the **standard deviation** of the sampling distribution is

$$\frac{\sigma}{\sqrt{n}}.$$

In a more advanced course, this result would be known as the **Central Limit Theorem**. Informally, it says that "averages are normally distributed in large samples." This leads to the following result.

Result: Choose a SRS of size n from a large population with mean μ. A **level** C **confidence interval** for the population mean μ is

$$\overline{x} \pm z^* \left(\frac{s}{\sqrt{n}} \right).$$

- The confidence interval formula has the same form:

  estimate ± margin of error.

- $\overline{x}$ is the sample mean and s is the sample standard deviation. These are calculated from the sample.

- The critical value z^* is the same as before; i.e.,

C	80%	90%	95%	99%
z^*	1.28	1.64	1.96	2.58

Example 21.3. In 1998, as an advertising campaign, the Nabisco Company announced a "1000 Chips Challenge," claiming that every 18-ounce bag of their Chips Ahoy! cookies contained at least 1000 chocolate chips. Dedicated statistics students at the Air Force Academy randomly selected 16 bags of cookies and counted the number of chocolate chips. These data are listed below:

Bag	Chips	Bag	Chips	Bag	Chips	Bag	Chips
1	1087	5	1191	9	1244	13	1325
2	1121	6	1200	10	1258	14	1345
3	1132	7	1214	11	1270	15	1356
4	1135	8	1219	12	1295	16	1419

Q: What is the population here?
A: A reasonable answer is "all 18-ounce bags of Chips Ahoy! cookies produced by Nabisco." There are likely millions of bags produced every day. We will assume the Air Force students had a representative SRS from this population.

Illustration: Let's calculate a **95% confidence interval** for μ, the population mean number of chocolate chips in an 18-ounce bag of Chips Ahoy! cookies.

- Note that this interval will not allow us to assess Nabisco's claim of "every 18-ounce bag of their Chips Ahoy! cookies containing at least 1000 chocolate chips."

- The claim Nabisco made was for every bag in the population! The confidence interval we will calculate is for the mean of the population.

Implementation in R:

```
# enter data
chips = c(1087,1121,1132,1135,1191,1200,1214,1219,1244,1258,1270,1295,1325,
    1345,1356,1419)
> mean(chips) # sample mean
[1] 1238.2
> sd(chips) # sample standard deviation
[1] 94.3
```

For this sample of 16 bags, the sample mean and sample standard deviation are

$$\begin{aligned} \overline{x} &= 1238.2 \text{ chips} \\ s &= 94.3 \text{ chips.} \end{aligned}$$

A 95% confidence interval for the population mean number of chocolate chips μ is

$$\begin{aligned} \overline{x} \pm z^* \left(\frac{s}{\sqrt{n}}\right) &\Longrightarrow 1238.2 \pm 1.96 \left(\frac{94.3}{\sqrt{16}}\right) \\ &\Longrightarrow 1238.2 \pm 46.2 \\ &\Longrightarrow (1192.0, 1284.4). \end{aligned}$$

Interpretation: We are 95% confident that the population mean number of chocolate chips in an 18-ounce bag of Chips Ahoy! cookies is between 1192.0 and 1284.4 chips.

Remarks: We finish with some remarks and words of caution.

- The confidence interval formula for μ we have presented is only applicable when the sample is a SRS. Other sampling designs would use different formulas.
- Outliers can distort the confidence interval for μ. This makes sense because both $\overline{x}$ and s are sensitive to them.
- When the sample size n is small (like in Example 21.3), the confidence interval formula might be inappropriate.
  - Reason: The sampling distribution of $\overline{x}$ may not resemble a normal distribution very well. The Central Limit Theorem is a statement about large samples; not small ones.
  - The authors of your textbook recommend using the interval only when the sample size $n \geq 15$. However, this is only a guideline.
  - If the underlying population density curve is skewed, this may not be a good guideline (you might need a larger sample).
- This is a good reminder to plot your data with a suitable graph (e.g., histogram, boxplot, stemplot). This will allow you to see if there is severe skewness or outliers present.