STAT 110
INTRODUCTION TO STATISTICAL REASONING

Spring 2017

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 disciplines!
- "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.

Research Question: Does treating premature infants with caffeine increase the chances of developing necrotizing enterocolitis?

Necrotizing enterocolitis (NEC) is a serious disease 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 (about 26 percent) receiving caffeine developed NEC
- 10 out of 478 patients (about 2 percent) not receiving caffeine developed NEC.

Reference: Cox et al. (2015). Evaluation of caffeine and the development of necrotizing enterocolitis. *Journal of Neonatal-Perinatal Medicine* **8**, 339-347.

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. These measurements are called **data**.

- In the NEC study, the individuals are the infants (patients).
- 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)
  - Gender (M/F)
  - Nutrition type (Breastmilk, Fluids, Formula, TPN)
  - Maternal drug use (Yes/No)
    - if "Yes," what type (Alcohol, Cocaine, Marijuana, Tobacco)
  - Time to discharge (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. For example, in Figure 1.1, we use a **histogram** to display the variation in the birth weight data for the 615 infants. Note that birth weight is a **quantitative** variable.

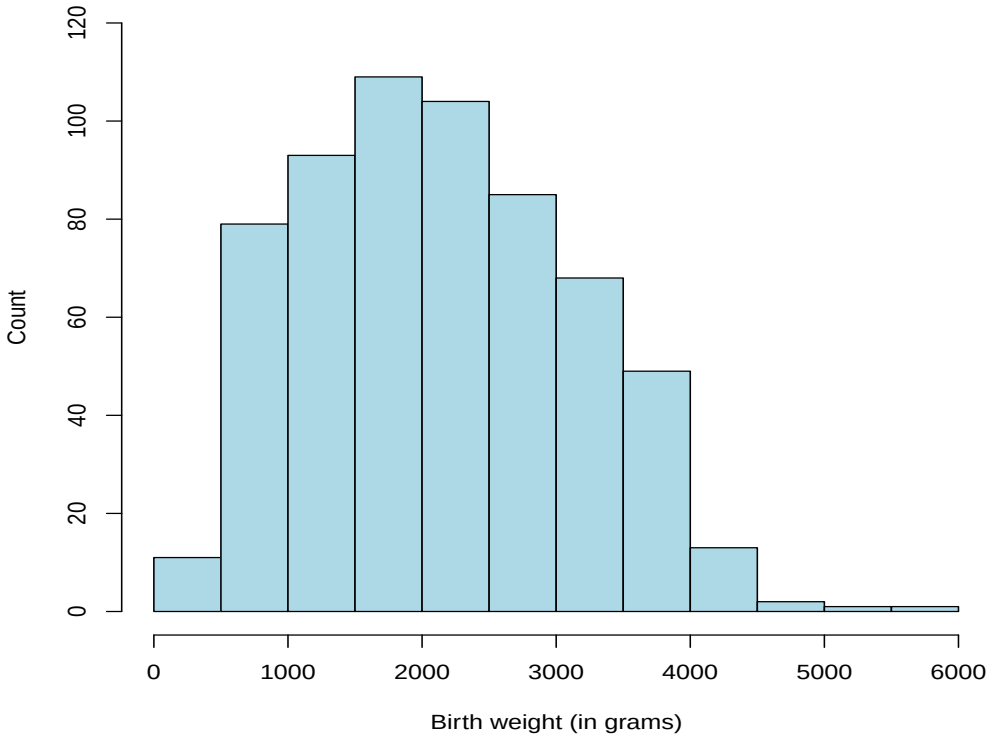


Figure 1.1: Necrotizing enterocolitis study. Histogram of birth weights (measured in grams) for 615 infants. This figure was created using R.

1.2 Populations and samples

Definition: 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, the sample is the 615 infants admitted to the neonatal intensive care unit. We observed these individuals and we recorded variables 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 July 24-26, 2016, Rasmussen Reports asked 1500 likely voters in the US to rate President Obama's job performance. Forty eight percent (48%) of the respondents "approved" of his job performance.

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

Discussion: What do these results suggest about the population? The 48% approval rating is for the sample. Could the population approval rating be different? Rasmussen stated that the **margin of error** associated with their sample results was $\pm 2.5\%$. What precisely does this mean?

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

- Additional examples of surveys: see Moore and Notz (pp 9-10).
- Interpreting the results of a survey depend on the notions of population and sample.
- In most situations, the population is 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 in "such manner as they shall by Law direct" (Article I, Section 2). This started in 1790 and has occurred every 10 years since then.

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.

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.

- Welfare example; see Moore and Notz (pp 13).
- Properly performed, experiments give better information about **cause and effect**. Observational studies generally give little or no information about this.

Main point: In observational studies, we passively observe the results. In experiments, we act deliberately (and then observe the effect of doing so).

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

- students who were "non-heavy" smartphone users (101 students); < 3 hours/day
- students who were "heavy" smartphone users (103 students); ≥ 3 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: (1) HK high school students who are "non-heavy" smartphone users and (2) HK high school students who are "heavy" smartphone users. The investigators wanted to compare these two populations for the different variables recorded. Comparing populations is a common goal in observational studies (and in experiments too).

Example 1.4. *If a woman eats more cereal, does this increase her chances of having a male child?* As outlandish as this sounds, researchers in the UK 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 the father have his say? From GEN 101, we know that males are heterogametic; females 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.
- It was later demonstrated that this "phenomenon" was easily explained by random chance.

Reference: Matthews et al. (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.

Reference: Young et al. (2009). Cereal-induced gender selection? Most likely a multiple testing false positive. *Proceedings of the Royal Society B* **278**, 1211-1212.

Remark: The problem with observational studies is that their "findings" are often not **repeatable**. In other words, if the same study is performed again, would the same conclusion be made? Unfortunately, silly claims 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!

2 Samples, Good and Bad

2.1 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. If a 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 and rarely provide accurate information about a population. These designs often underrepresent certain groups of individuals.

Example 2.1. Convenience sample. You have been asked to inspect a truck shipment of oranges. You select 5 crates nearest to the door. In each crate, you pick 5 oranges from the top. You have a sample of 25 oranges. However, this sample is probably not representative of the entire shipment.

Example 2.2. Convenience sample. You decide to survey USC undergraduates on whether they think USC professors are overpaid. You ask the first 50 students you see outside of the Strom Thurmond gym on one Saturday afternoon. What groups of USC students are underrepresented in this sample?

Example 2.3. Voluntary response sample. 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?

Example 2.4. 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 that Landon would win the election with 57.1% of the popular vote and an electoral college margin of 370 to 161.
- Roosevelt won the election with 60.8% 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 sample was a voluntary response sample (only those who returned the post cards were included). In addition, the **sampling frame** was biased. The mailings 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.

Note: George Gallup, pioneer of survey sampling techniques and inventor of the Gallup poll, correctly predicted the 1936 election using a **simple random 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.

Note: 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, he also predicted in July 2015 that Donald Trump had only "a 5 percent chance" of winning the 2016 Republican nomination!

2.2 Simple random samples

Recall: Convenience and voluntary response sampling designs are biased. They tend to systematically favor certain individuals over others. To avoid systematic underrepresentation, 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.
- On average, a simple random sampling design will give accurate results for a population.

Example 2.5. 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 200 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
⋮	⋮
Zhang	199
Zynkowski	200

The R code below can be used to select 5 numerical codes between 1 and 200:

```
students = seq(1,200,1)
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.

Note: 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  20 146 186
```

Note: Instead of using R's `sample` function to select simple random samples, the authors of your text suggest using a **table of random digits**; see Moore and Notz (pp 27-30). This is essentially a list of numbers determined at random, and using it will accomplish the same goal as using R.

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: 24,876 (as of August 2015)
- Population: SC (aged 18 and older). Size: 3.8 million (as of July 2015)
- Population: USA (aged 18 and older). Size: 248 million (as of July 2015).

In practice, those performing the study may not identify exactly what the population is, construct a 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 mathematically, it might not be that far off.

2.3 Final comments

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

- Unbiased design: SRS
- Biased designs: convenience, voluntary response

Note: Other unbiased sampling designs are available; for example,

- **Stratified sampling:** individuals are first separated into strata (different groups; e.g., gender, race, income class, etc.). Then a SRS is taken from each stratum. This design ensures that individuals from different strata are represented.
- **Cluster sampling:** select a sample of clusters (e.g., city blocks, high schools, etc.). Then sample all individuals in each cluster selected. This type of sample is easier to construct than a SRS.
- **Systematic sampling:** Individuals are selected in a systematic way but in one that does not intentionally bias the results; e.g.,
  - Asking every USC student with ID number ending in "1" to fill out a survey
  - Selecting every 1000th visitor to a popular web site.

The analysis of samples from these designs is more difficult than that from a SRS.

Warning: When reading newspaper/online articles, the following phrases usually indicate that the results are biased, most likely due to poor sampling designs:

- "This is not a scientific poll."
- "These results may not be representative of individuals in the general public."
- "....based on a list of those individuals who responded."

3 What do Samples Tell Us?

3.1 Parameters and statistics

Example 3.1. During July 24-25, 2016, Rasmussen Reports conducted a national telephone and online survey using a SRS of $n = 1000$ American adults. Each participant was asked:

Should police officers be required to wear body cameras while on duty?

The survey found that 700 of the 1000 adults in the sample answered "Yes" to this question. Rasmussen stated "the margin of sampling error is $\pm$ 3 percentage points with a 95% level of confidence." What does this statement mean?

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 248 million American adults (based on July 2015 `census.gov` data). Define

p = population proportion of American adults who agree police officers should wear body cameras while on duty.

Because p describes the population of all American adults (all 248 million of them), it is a parameter. We call p a **population proportion**. It is unknown.

What do we know? We do know that 700 out of the 1000 American adults in the sample agreed that police should wear body cameras while on duty. Therefore, the **sample**

proportion is

$$\widehat{p} = \frac{700}{1000} = 0.70 \ \ (\text{or } 70\%).$$

The sample proportion $\widehat{p}$ is a statistic, because we calculated it using the individuals in the sample.

Terminology: We say that the sample proportion $\widehat{p} = 0.70$ 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 SRS from the population and use a sample statistic to estimate it. This is the idea behind statistical inference.

3.2 Accuracy and precision

Remark: We want our sample estimates 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.

Discussion: Rasmussen Reports found that 700 out of 1000 American adults sampled are in favor of police officers wearing body cameras while on duty. Suppose Gallup did the same poll during the same time, asking the exact same question, and found that 720 out of 1000 American adults were in favor of this. The sample proportion based on Gallup's sample is

$$\widehat{p} = \frac{720}{1000} = 0.72 \ \ (\text{or } 72\%).$$

How can Gallup's sample proportion be different than Rasmussen's? That's easy. Each sample uses a different 1000 American adults.

Important: This discussion highlights an important fact: Values of statistics like $\widehat{p}$ change from sample to sample because different samples contain different individuals.

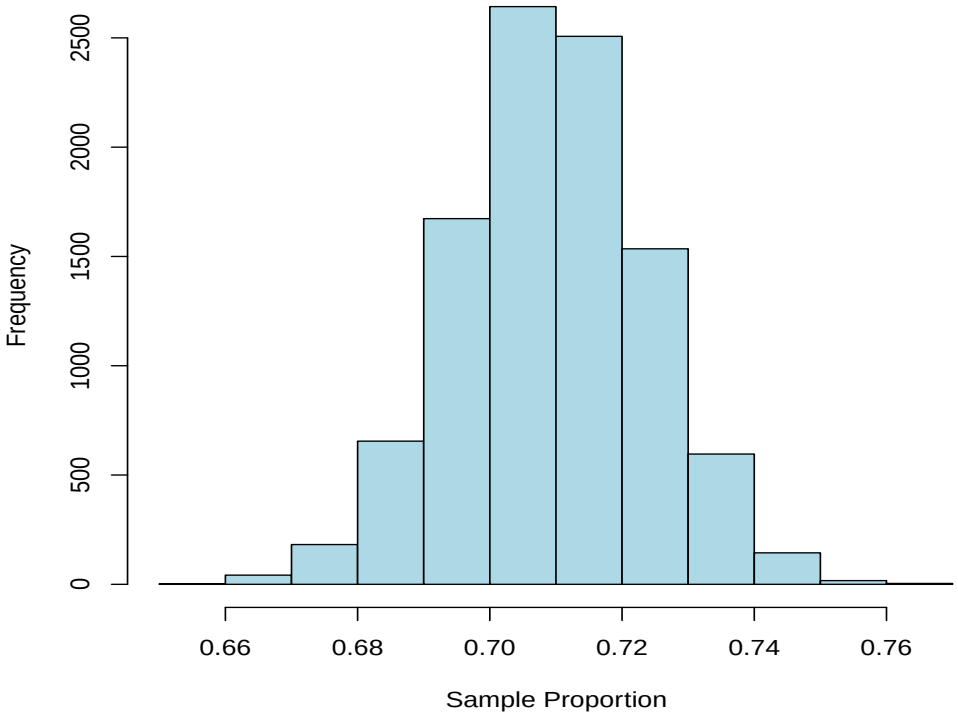


Figure 3.1: 10,000 simulated sample proportions $\widehat{p}$. Each one is based on a sample size of $n = 1000$ individuals. The population proportion is assumed to be $p = 0.71$.

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 $p = 0.71$ (halfway in between the Rasmussen and Gallup estimates).
- I used R to simulate 10,000 values of the sample proportion $\widehat{p}$ under this assumption. Each sample proportion is based on a sample of size $n = 1000$.
- I used a **histogram** to display the 10,000 sample proportions. This histogram is above in Figure 3.1.

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 $p = 0.71$).
  - In other words, the sample proportion estimates are "correct on the average." This is what it means for an estimate to be **unbiased**.
  - **Terminology:** 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.
- The Rasmussen and Gallup sample proportion estimates (0.70 and 0.72, respectively) are close to the true population proportion $p = 0.71$, but some estimates are much further away; e.g., 0.66, 0.76, etc.
  - 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 $n = 1000$ American adults. How can we make the sample proportion estimate $\widehat{p}$ more precise (i.e., less variable)? **Answer:** Take a larger sample (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 different sample sizes:

- $n = 100$ individuals (a smaller sample size)
- $n = 10,000$ individuals (a larger sample size).

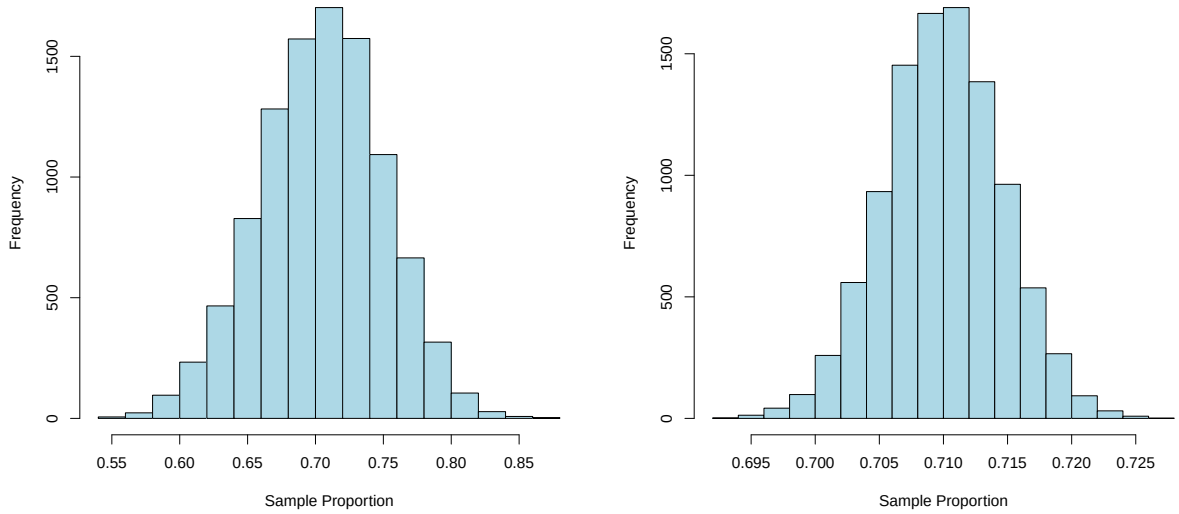


Figure 3.2: 10,000 simulated sample proportions $\widehat{p}$. The population proportion is assumed to be $p = 0.71$. Left: Sample size is $n = 100$. Right: Sample size is $n = 10,000$.

Observations: The histograms in Figure 3.2 display the results from this simulation:

- In both cases, the sample proportions $\widehat{p}$ remain centered at the truth $p = 0.71$.
  - 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 changing the sample sizes is that now the variability in the estimates has changed (look at the scale/range on the horizontal axis):
  - For the smaller sample size $n = 100$ (left), the variability is larger. In other words, the estimate $\widehat{p}$ is less precise.
  - For the larger sample size $n = 10,000$ (right), the variability is smaller. In other words, the estimate $\widehat{p}$ is more precise.

Summary: When attempting to estimate a population parameter (like p), there are two important concepts to consider:

- **Bias:** a consistent, repeated deviation of a sample statistic from a population parameter when we take repeated samples.
- **Variability:** this describes how "spread out" the values of a sample statistic are when we take repeated samples.

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

- To reduce bias, use a SRS. The simple random sampling 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 n, 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 police officers be required to wear body cameras while on duty?

The survey found that 700 of the 1000 sampled American adults answered "Yes" to this question. Rasmussen stated "the margin of sampling error is $\pm$ 3 percentage points with a 95% level of confidence." We now describe what this means.

Remark: Even if we use a SRS with a large sample size (like $n = 1000$), we cannot be sure how close our estimate $\widehat{p}$ is to the population proportion p. We do not sample the entire population, so only the Oracle knows what p is. As a result, there will always be a degree of uncertainty with our estimate.

Definition: The **margin of error** is a numerical value that quantifies the uncertainty in an estimate.

- The larger the margin of error, the more uncertainty in the estimate.
- The smaller the margin of error, the less uncertainty in the estimate.

Note: Rasmussen found the sample proportion of adults who agree police officers should wear body cameras while on duty to be

$$\widehat{p} = \frac{700}{1000} = 0.70 \quad (\text{or } 70\%).$$

They also stated that the margin of error was $\pm$ 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 agree police officers should wear body cameras while on duty is between 0.67 and 0.73 (i.e., between 67% and 73%)."

Calculation: In a SRS, the margin of error in the sample proportion $\widehat{p}$ associated with a 95% confidence level is approximately equal to $1/\sqrt{n}$; that is,

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

where n is the sample size. In the Rasmussen example,

$$\frac{1}{\sqrt{1000}} \approx \frac{1}{31.62} \approx 0.0316 \quad (\text{that is, about } 3\%).$$

This is where Rasmussen's 3% figure comes from!

Exercise: Using the formula above, calculate the margin of error when (a) $n = 100$ and (b) $n = 10,000$. Using the Rasmussen estimate of $\widehat{p} = 0.70$, write a confidence statement for each sample size.

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

- **Margin of error.** This quantifies how close the sample statistic is to the population parameter.
- **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.

Important: The margin of error formula

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

applies only for a 95% confidence level. This formula would change slightly if we wanted more or less confidence (see Chapter 21).

Remark: What does "95% confidence" really mean?

- It first reminds us that when we make a confidence statement, we are not 100% certain in its truthfulness.
  - We do not sample the entire population! Therefore, we can never be 100% confident in our statement.
- Suppose we took many simple random samples from the same population (all of the same size) and wrote a confidence statement with each one. The phrase "95% confidence" means that 95 percent of our confidence statements will be correct.
- Therefore, 5% of the time, our confidence statement will not be correct. That is, the true value of the population parameter will fall outside the range we specify; e.g., 0.67 to 0.73.
- Unfortunately, we never get to know whether our confidence statement is truly correct. This is through no fault of our own; rather, it is the reality in which we must live.

4 Sample Surveys in the Real World

4.1 Introduction

Example 4.1. A Pew Research Center opinion poll conducted over the phone talks to 1000 people chosen at random and announces its results. Your favorite web site posts the poll results. You read these results and are satisfied because you see words like, "random sample," "margin of error," and "95 percent confidence."

However, let's look more closely at what it took to secure the 1000 responses.

No answer	938
Answered but refused	678
Not eligible	221
Incomplete interview	42
Complete interview	1000
Total called	2879

- 2879 individuals were needed to get 1000 responses!
- The **response rate** is
$$\frac{1000}{2879} \approx 0.35 \quad (\text{or } 35\%)$$
- The **nonresponse rate** is 65%!
- Do you believe the 1000 individuals are representative of a larger population? What about the 1879 individuals that did not contribute?

Reality: Getting feedback from human subjects is not easy. Many people will not participate in surveys over the phone, in person, or through email.

Terminology: There are two types of **sampling errors** that can occur in surveys:

1. Random sampling errors
2. Nonsampling errors

4.2 Random sampling errors

Definition: Random sampling error is the natural error that arises when sampling from a population.

- Even if we use a SRS, random sampling errors are still present. It is the deviation between a sample statistic and a population parameter caused by chance.
- The margin of error in a confidence statement includes only random sampling error. It does not include other sources of error.
- Taking a larger SRS reduces the margin of error but it will not eliminate random sampling error completely.
- How do you completely remove random sampling error? Take a census.

Remark: Any degree to which the sample is "unlike" the population contributes to random sampling error. This is why convenience and voluntary response sampling designs produce poor results. Samples from these designs are rarely representative of a larger population. Individuals in these samples are not selected by chance.

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.

- Undercoverage also results when using an incomplete **sampling frame**, such as in the *Literary Digest* example (see Chapter 2). If the sampling frame excludes certain groups of individuals, then these individuals can never be sampled.

Example 4.2. Many opinion polls still conduct interviews by telephone using random digit dialing. Which groups of individuals would be excluded? Here are a few:

- Individuals without phones (about 2% of the US population?)
- Individuals on no-call lists
- Other groups, for example, military members overseas, individuals in hospitals, prisons, etc.
- Dialing restricted to contiguous 48 US states? AK, HI.

4.3 Nonsampling errors

Definition: Nonsampling errors are errors that arise from sources not related to the act of sampling. Here are some examples:

- Data entry errors
- Nonresponse
- Poorly worded/misleading questions
- Interviewer bias (e.g., coercing participants to respond a certain way, etc.)
- Response error (e.g., participant falsifies responses to questions, etc.).

Definition: **Nonresponse** is the failure to obtain data from an individual selected for a sample. This is biggest source of nonsampling error in sample surveys. See the Pew Research Center in Example 4.1−the nonresponse rate was 65 percent!

Discussion: Here are some examples of poorly worded or misleading questions:

1. *"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.
2. *"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 percent responded "Yes." Question wording can greatly influence the results!
3. *"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 percent of the respondents said it was "possible."

Interviewer bias can also sway responses towards a single desired direction. **Response error** occurs when individuals do not answer truthfully. An individual may not be truthful if the survey deals with sensitive topics.

Discussion: 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 really reject evolution?

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

Discussion: Questioning individuals about sexual practices, criminal activity, or other sensitive topics, is difficult. The **randomized-response technique** is a survey method that can be used when individual anonymity needs to be preserved. This will encourage participants to provide truthful responses.

We illustrate how a randomized-response survey could be performed. The topic is **drink-spiking**. In a classical survey setting, suppose someone asked you *"Have you ever drink-spiked someone to take advantage of them sexually?"* Guilty individuals would probably lie if asked this question, because they don't want to admit having done this.

To preserve individual anonymity (i.e., to encourage truthful responses), a randomized-response survey poses 2 questions; for example,

1. Have you ever drink-spiked someone to take advantage of them sexually?
2. Is the last digit of your SSN even?

Participants are then asked to flip a fair coin. If it is "heads," they answer Question 1. If it is "tails," they answer Question 2. **The result of the flip is not revealed to the interviewer.** Individual anonymity is preserved because

- a "Yes" response doesn't imply the individual is guilty of drink-spiking. S/he could be answering "Yes" to Question 2.
- a "No" response doesn't imply the individual is innocent of drink-spiking. S/he could be answering "No" to Question 2 (and still be guilty of drink-spiking).

The fraction of "Yes" responses can be used to estimate p, the population proportion of drink spikers. Provided the sample is a SRS, one would solve the following equation for p to find the estimate:

$$\text{Fraction of "Yes" responses} = 0.5p + 0.5(0.5).$$

Reference: Campbell, C. and Joiner, B. (1973). How to get the answer without being sure you've asked the question. *American Statistician* **27**, 229-231.

4.4 Other sampling designs

Note: In practice, most sample surveys use more complex sampling designs than a SRS. In part, this is done to mitigate the impact of nonsampling errors; i.e., it is done to ensure an accurate representation of the population.

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 gender, race, income class, political affiliation, etc.

2. Take a SRS from each stratum and combine each SRS to form the complete sample.

Example 4.3. A large seroprevalence study in Houston, Texas used a stratified sample to estimate HIV positivity among heterosexual males who were intravenous drug users (IVDUs) and were not receiving treatment for their addiction. **Note:** This is an older study (results published in 1990) but at the time it was important because very little was known about this population of individuals.

Race	Number of individuals	Number infected	Sample proportion
Hispanic	107	4	0.037
White	214	14	0.065
Black	600	59	0.098

In this example, this population of IVDUs was stratified by race (Hispanic, White, Black).

Exercise: Treating each stratum's sample as a SRS, calculate the margin of error associated with each sample proportion (assuming a 95% level of confidence).

Terminology: A **cluster 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 4.4. 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 elementary schools. A cluster sampling strategy was used.

- A SRS of 6 elementary schools was chosen from the 33 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.

Discussion: Stratified and cluster 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 sample consist of individuals who are homogeneous in some way (e.g., gender, race, income class, political affiliation, etc.). On the other hand, clusters in a cluster sample will include individuals from many different strata.
- Stratified samples are used to ensure overall representation (i.e., to mitigate the impact of nonsampling errors). However, a secondary goal is often to compare the results between/among the strata; e.g.,
  - How do individuals of different genders compare? Different races?

  In a cluster sample, on the other hand, there is usually no interest in comparing the clusters themselves. Clustering is used because it makes the sampling process easier. Every individual in each cluster is sampled.

Summary: We have talked about different sampling designs. These designs can be grouped into two categories:

- **Probability samples:** a sample that is chosen by chance (SRS, stratified, cluster, systematic)
- **Nonprobability samples:** a sample that is not chosen by chance (convenience, voluntary response).

5 Experiments, Good and Bad

5.1 Introduction

Note: Here is a quote from Moore and Notz (pp 91):

> "Observational studies are just **passive** data collection. We observe, record, or measure, but we don't interfere. Experiments are **active** data production. Experimenters actively intervene by imposing some treatment in order to see what happens."

Terminology: This is the language we use with experiments:

- The individuals in an experiment are called **subjects** (or **experimental units**).
- A **treatment** is a specific experimental condition applied to the subjects.
- 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 was a large experiment involving 22,071 male physicians (aged 40+). One of the goals was to investigate whether taking aspirin reduces the risk of heart attacks. This experiment was performed in the 1980s and the aspirin component ended in 1987. A complete description is located at `http://phs.bwh.harvard.edu/phs1.htm`.

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

Assignment to aspirin/placebo was **randomized**. The experiment was **double-blinded**.

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

Results: There were 139 heart attacks among those assigned to aspirin. There were 239 heart attacks among those assigned to placebo. This result was determined to be **statistically significant**, meaning that the difference between the treatment groups (aspirin/placebo) was so large that it is likely not explained by random chance.

Reference: Final report on the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine* **321**, 129-135. Published in 1989.

Example 5.2. *Sex education for adolescents.* An experiment was performed with 662 AA students (6th/7th grade) recruited from 4 public schools that serve low-income AA communities in the northeast US. The study took place during 2001-2002. Students were randomized 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 practices during this time.

- Subjects = students

- Treatment = educational programs (A, SS, C-8, C-12, Control)
- Response variable = sexual intercourse during follow-up period? (1 = Yes; 0 = No)
- Explanatory variables = educational program, gender, 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**.

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

Example 5.3. *Does withholding feed from pigs prior to slaughter reduce the impact of salmonella?* An experiment in North Carolina was carried out to investigate this issue. 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 = pigs
- Treatment = feeding withdrawal schedule (0 hours, 12 hours, 24 hours)
- Response variable = salmonella detected in cecum? (1 = Yes; 0 = No)
- Explanatory variables = feeding withdrawal schedule, gender, initial weight, different marketing groups.

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

Reference: Morrow et al. (1999). Prevalence of *salmonella spp.* in the feces on farm and ceca at slaughter for a cohort of finishing pigs. ISECSP (conference paper) 155-157.

5.2 Randomized comparative experiments

Note: Here is the outline of a **randomized comparative experiment**:

General Outline of Experimentation

Sample subjects from population

↓

Randomize subjects to treatments

↓

Record data on each subject

↓

Analyze data

↓

Make statement about the differences among treatments.

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

- This produces groups of subjects that should be similar before we apply the treatments.
- If subjects were assigned to treatment groups in a biased way, then this could destroy the experiment. We could not say if a treatment difference was due to the treatments or the biased assignment.

Terminology: A **comparative experiment** is an experiment that compares two or more treatment groups. The experiments in Examples 5.1-5.3 are comparative experiments.

- In Example 5.1 (PHS), there are two treatment groups (aspirin/placebo).

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

- new drug versus placebo
- new drug versus standard drug.

A control group is used as a basis for comparison. The use of a control group also helps to control the effects of lurking variables.

Definitions: A **lurking variable** is a variable that has an important effect on the relationship between the response variable and treatment but is not one of the explanatory variables studied. Two variables are **confounded** when their effects on a response variable can not be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables.

Example 5.4. Suppose we wanted to compare taking STAT 110 online and taking STAT 110 in this large lecture in-class format.

Design 1: Assign all females to the online section and all males to the in-class section.

- In this extreme case, the sections would be **completely confounded** with gender.
- If we found a difference in student performance, we would have no way of knowing whether this was due to the sections or the effect of gender.

Design 2: Let each student choose which section s/he wants to take (this is no longer an experiment; it is an observational study).

- If students who choose the online section tend to have better quantitative backgrounds or are academically more mature, then the sections would be **partially confounded** with these variables.
- In this example, quantitative background and academic maturity are potential **lurking variables**. They are potentially important in describing the response variable, but this "design" does not account for them.

Design 3: Randomly assign each student to one of the two sections.

- This would create two groups of students who are similar on average.
- Students are exposed to similar conditions; e.g., same material covered, same exam formats, same instructor (if possible), etc.
- Treating the groups similarly ensures that any lurking variables operate in the same way for both sections (so that their effects are "averaged out").
- Any differences we see in the response variable (e.g., final course percentage, etc.) is due to the difference of the sections.

Definition: An observed effect so large that it would rarely occur by chance is called **statistically significant**. For example, in Example 5.4, consider the following two scenarios:

- Scenario 1: On-line section: 84% class average; In-class section: 83%. The difference here is small and could be due to chance.
- Scenario 2: On-line section: 84% class average; In-class section: 66%. The difference here is large and may not be due to chance. The difference may arise because the online method of instruction is truly better.

Discussion: *Can giving a placebo actually create a lurking variable?* In some experiments involving humans, there could be an effect associated with a placebo treatment. A subject might think the fact s/he "is taking something" may itself have an effect on the subject's response. This is called the **placebo effect**.

Example 5.5. A study published in *Psychological Science* found that subjects who believed they were consuming alcohol still experienced weakened judgement and an impaired memory. For the experiment, 148 students were randomized to two groups:

- Group 1: Drinking vodka and tonic
- Group 2: Drinking tonic only.

Students were informed what group they were assigned to (i.e., the students were not **blinded**). 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, Group 1 students even behaved drunk, displaying physical signs of intoxication!

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

5.3 Experiments versus observational studies

Remark: Properly designed experiments are the "gold standard" for examining the relationship between a response variable and treatment. They are our best chance at establishing a **cause and effect** relationship.

- Randomization produces groups of subjects that should be similar on average.
- Comparative design exposes groups of subjects to similar conditions. The only difference between the groups is what treatment they receive.
- Therefore, differences we observe in the response variable can be attributed to the treatment itself.

Remark: Now you see why observational studies are less informative than designed experiments.

- Because randomization is not used in observational studies, there is no guarantee that the groups we want to compare are similar. We merely observe different groups without interfering.
- The effects of lurking variables are not controlled, so any differences we see in the response variable between or among different groups could be confounded with these variables. Because of this, establishing a cause and effect relationship in observational studies is difficult or impossible.
- There is a technique called **retrospective matching** that can be used in observational studies to deal with lurking variables. It basically involves putting individuals in certain groups formed "after the fact" and then making comparisons for these groups. See Moore and Notz (pp 102-104).

Remark: Observational studies are not useless. In some situations, it may not be ethical to perform a designed experiment and so an observational study is the best we can do. For example, can we design an experiment to prove smoking **causes** cancer?

6 Experiments in the Real World

6.1 Introduction

Recall: A successful experiment adopts the following principles:

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

These are the **three principles of experimental design**.

Remark: One aspect of a successful experiment is that all subjects are treated alike except for the treatments. If subjects in different treatment groups are treated differently, this could bias the experiment. Blinding is a useful technique to avoid this type of bias.

- Blinding: the subject does not know which treatment s/he is receiving
- Double blinding: neither the subject nor the investigator knows which treatment s/he is receiving.

Blinding can help to mitigate the impact of a **placebo effect**. If the subject does not know which treatment is being administered, this may prevent them from thinking, for example, "I'm only getting the placebo." At the same time, some subjects will respond no matter what they are taking−recall the vodka/tonic experiment in Example 5.5. Here are more examples of this:

- 42 percent 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. 2/13 patients developed a rash after receiving a solution containing poison ivy!

Sometimes it is not possible to blind subjects in an experiment. For example, if an individual is assigned to a treatment group which takes a standard drug and an experimental radiation regimen in a cancer clinical trial (versus the other group which only receives the drug), then obviously blinding is not possible.

Remark: Double blinding is used to remove bias caused by the investigator him/herself. If the investigator has a vested interest in a particular treatment, s/he may introduce bias by treating different subjects differently; e.g., treat placebo subjects differently than those who take the "real" treatment.

6.2 Clinical trials

Definition: A **clinical trial** is an 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.

- The Physicians' Health Study in Example 5.1 is an example.

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 post-marketing phase to determine if the new drug/intervention can be administered to the general public. Side effects are closely monitored.

Example 6.1. The web site `www.centerwatch.com` summarizes the results and findings from recent clinical trials. Here is an example:

August 27, 2012. Medivation and Astellas Pharma published results from a phase III trial of enzalutamide for the treatment of metastatic prostate cancer. This international, randomized, double-blind, placebo-controlled study, AFFIRM, enrolled 1,199 men who had been previously treated with chemotherapy.

- Subjects received enzalutamide 160mg daily (as four 40mg capsules), or placebo.
- 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 representing a 37% reduction in the risk of death.
- The drug was well tolerated. The most frequent adverse events were fatigue, diarrhea, and hot flush. Seizure was reported in less than 1% of enzalutamide-treated patients.

Remark: In the absence of well-designed clinical trials, 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.

Remark: Because clinical trials involve human subjects, there are important ethical issues that arise. This topic will be discussed in Chapter 7.

6.3 Common problems

Remark: Just as in sample surveys, **undercoverage** can be a problem in designed experiments. In any experiment, subjects are first sampled from a larger population. Therefore, if the sampling design is biased, then we are excluding certain individuals from participating. For example,

- a psychology experiment involving USC students but excluding students majoring in the sciences
- a medical experiment examining different nutrition methods for newborns but excluding mothers in certain racial groups
- an agricultural experiment comparing different treatments for Ovine Johne's disease (in sheep) but including only certain breeds.

We want our conclusions to apply to an appropriate population of interest. If certain groups are not represented, our conclusions will 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.

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 cause bias because we don't get to see how these subjects respond to the treatment.

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 situations.

- 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.

Remark: When experimental subjects are humans, the **Hawthorne effect** is a change in subjects' behavior or outcomes not directly attributable to the treatment received but simply to the "awareness of being in an experiment."

- The benefits of the treatment found in a controlled clinical trial may vanish when the treatment is introduced into a larger population; see Example 7 in Moore and Notz (pp 119-120).

6.4 Experimental designs

6.4.1 Completely randomized designs

Terminology: In a **completely randomized design**, all of the subjects (individuals) are allocated at random among all of the treatments.

Frame of reference: Put all subjects' names/identifiers into a single hat. Then draw randomly from the hat and assign to the treatment groups.

Example 6.2. An engineer is interested in comparing 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).

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

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

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

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

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

This is a completely randomized design because all subjects (boards) are assigned at random among all treatment groups (brands of paint).

Remark: In Example 6.2, the treatment groups arise from one dependent variable: brand of paint. However, a completely randomized design can also be performed when there is more than one dependent variable.

Example 6.3. 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 rate of cancer. Therefore, in the PHS, there were actually two dependent variables used to form the treatment groups.

Dependent variable 1: Aspirin (aspirin/placebo)
Dependent variable 2: Beta carotene (beta carotene/placebo).

Here are the four treatment groups that were used in the PHS:

- Treatment group 1: Aspirin/beta carotene
- Treatment group 2: Aspirin-placebo/beta carotene
- Treatment group 3: Aspirin/beta carotene-placebo
- Treatment group 4: Aspirin-placebo/beta carotene-placebo.

The PHS was a completely randomized design because all physicians were assigned to one of these four treatment groups at random.

6.4.2 Randomized block designs

Terminology: In the language of experimental design, a **block** is a group of subjects that are known to be similar in some way. Therefore, subjects in the same block could respond to treatments differently than subjects in another block.

Discussion: As a generic example, suppose that we are comparing two weight-loss treatments for human subjects. A completely randomized design would assign subjects to the two treatments completely at random. However, in this setting, we could also acknowledge that there are potential blocking variables:

- Gender: could individuals of different genders respond differently to the weight-loss treatment?

- Age: could individuals of different ages respond differently?
- Initial weight (e.g., overweight, obese, severely obese): could individuals of different body compositions respond differently?

Terminology: In a **randomized block design**, subjects are assigned to treatments at random, but separately within each block.

Discussion: Why use blocking? Blocking is a form of **control**.

- By separating subjects into different blocks and then randomizing within block, you eliminate the blocking variable as a potential lurking variable.
- In other words, you are controlling for the effects of this blocking variable.
- This enables the experimenter to compare the treatments more effectively (the effects of outside lurking variables have been controlled for by blocking).

Example 6.4. In the adolescent sex-education study (Example 5.2, notes), there were 662 6th/7th-grade students who were assigned to one of five sex-education treatment groups (A, SS, C-8, C-12, Control). However, the students were not assigned completely at random. The researchers used two blocking factors, grade and gender, which created the four blocks:

- Block 1: 6th grade/M
- Block 2: 6th grade/F
- Block 3: 7th grade/M
- Block 4: 7th grade/F.

Within each block, students were then randomly assigned to one of the five education groups (A, SS, C-8, C-12, Control). This was a randomized block design.

Example 6.5. Researchers in Finland wanted to evaluate the use of xylitol to potentially 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.

- Children were first separated into two groups (blocks): whether or not the subject could chew gum. This is clearly related to the age of the child−very young children may not be able to chew gum yet.
- Within each block, children were randomly assigned to receive xylitol or control.
  - For the gum-chewing block, xylitol was administered in gum form.
  - For the non-gum-chewing block, xylitol was administered in syrup form.

Children were then followed for a 3-month period. 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."

Potential problems: There were more dropouts recorded for the xylitol treatment group than for the control group (in both blocks). The form of the medication (gum/syrup) may also be a lurking variable!

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

6.4.3 Matched pairs design

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

- Prototypical example: Coke/Pepsi taste test

- A matched pairs design is a special randomized block design:
  - Each subject serves as its own block
  - The order of the treatment given (treatment 1/treatment 2) is randomized if possible.

Example 6.6. *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.

Discussion: Each student is measured twice: The pre-test score and post-test score.

- 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.

Final comment: The design of an experiment is important. If an experiment is designed poorly, then this may inhibit the researcher from learning from the experiment. Unfortunately, some researchers reach out to the statistician for advice "after the fact." By then, it is usually too late to salvage the situation.

7 Data Ethics

7.1 Introduction

Discussion: From an investigator's perspective, often the ultimate goal of an experiment or observational study is to show that a **research hypothesis** is supported by the data and statistical analysis. For example,

- "Our new method of treatment is superior to the standard treatment."
- "American voters' opinions on gay marriage can be changed."
- "Abstinence-only sex education is preferred for at-risk youths."
- "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."
- "Eating more cereal increases the chances of having a male child."

In most applications, the investigator will posit a research hypothesis like one of these above. An experiment or observational study is then designed to investigate whether or not the research hypothesis is supported.

Remark: If data from an experiment or observational study contradict an investigator's research hypothesis, s/he could end up looking bad in the process. This can sometimes lead to unethical behavior on the investigator's part; e.g.,

- 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.

Remark: The authors of your text do not discuss these types of issues, conceding that "there is no ethical question here" and that such practices "are just wrong." However, we should also be aware that unethical behavior in research happens.

7.2 Unethical behavior in research: Two examples

Example 7.1. *The Duke-Potti scandal.* Dr. Anil Potti's cancer research at Duke University in the mid-2000s was viewed as 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 generically as one group of individuals and developing general treatment strategies, treatments could be targeted to each individual cancer patient (the idea behind "personalized medicine").

- Initial data, results, and conclusions supported Potti's theories and were published in top medical journals (of course, the funding and accolades started to pour in).
- Drs. Kevin Coombs and Keith Baggerly, biostatisticians at MD Anderson Cancer Center, found problems with the data and the resulting conclusions.
- It was later found that the data collected by Potti and his research colleagues had been falsified.
  - Patient DNA data supporting Potti's ground-breaking approach to cancer treatment were retained. Those that did not were altered to support the theory.
- End result: Patients were ultimately enrolled in cancer treatment therapies that had no effect or made things worse.
- Potti "resigned" from Duke in disgrace, his research career ultimately destroyed.
- **60 Minutes clip:** `https://www.youtube.com/watch?v=W5sZTNPMQRM`

Example 7.2. *The Michael LaCour-gay marriage debacle.* An aspiring academician in political science, PhD student Michael LaCour (UCLA) undertook a large experimental study under the direction of 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.

- The hypothesis proposed by LaCour was that gay canvassers could have more impact on changing the opinions of voters than that of straight canvassers.
- The data LaCour "collected" demonstrated this hypothesis was supported overwhelmingly. LaCour and Green published their results in *Science*, perhaps the most prestigious journal in all fields.
- News outlets discussed the findings extensively; LaCour was offered an academic position at Princeton.
- A few months after publication (in 2014), there were questions raised about who funded the study. 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 that the Lacour's data had been completely fabricated.
- LaCour could not successfully defend his data sets nor could he reproduce them for others to see; they had been "destroyed."
- In the light of these anomalies, Green (LaCour's co-author) asked *Science* to retract the publication. Princeton soon after rescinded its offer of employment to LaCour.

Reference: Lacour, M. and Green, D. (2014). When contact changes minds: An experiment on transmission of support for gay equality. *Science* **346**, 1366-1369.

Note: This article was retracted on May 28, 2015.

7.3 Basic data ethics

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 overall 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 the 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.
  - The 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.

- 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 groups of human subjects where the informed consent requirement can be "overlooked?"

Example 7.3. In Moore and Notz (pp 147-148), your authors describe a social science 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.

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

Definitions: Data confidentiality is the protection of information provided by subjects and the assurance that information about individual respondents cannot be derived from the statistics reported. **Data anonymity** is the protection of the individual's data from all persons, even those that are involved with the study.

7.4 More on 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?

Quote from Dr. Charles Hennekens (director of the PHS trial), **boldface added**:

> "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.
- It is accepted that clinical trials are ethical in the setting of uncertainty.
- The **Tuskegee syphilis study** grossly violated the equipoise principle; see Example 4 in Moore and Notz (pp 143).
  - This study ended in 1972; shortly after this date, the National Research Act was passed by the US Congress.

8 Measuring

8.1 Introduction

Importance: Statistics is the science of 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 three of them:

- Gestational age (measured in weeks)
- Birth weight (measured in grams)
- Time to discharge (measured in days).

Terminology: We **measure** a variable on an individual (subject) when we assign a number to the variable. We use an **instrument** to make a measurement. For variables that are quantitative, measurements have **units** attached to them. The resulting measurements are called **data**.

Here is what part of the NEC data set looks like (copied from my Excel file; I altered the measurements for confidentiality reasons):

Infant	Gestational age	Birth weight	Time to discharge
1	34	2402	25
2	25	755	107
3	34	2059	33
⋮	⋮	⋮	⋮
615	31	1401	72

Discussion: Some variables have intuitive meanings and are easy to measure. In engineering, science, agricultural, and medical applications, for example, variables are usually well defined and there is little ambiguity in what is being measured.

- Engineering: dimensions of a part (length/width/height), temperature of a solution, strength of a concrete mixture, time until an engine fails, etc.
- Science: cell count, PH of a chemical toxin, distance traveled, hurricane wind speed, age of geological event, etc.
- Agriculture: crop yield, rainfall amount, soil composition, livestock physical characteristics, etc.
- Medicine: systolic blood pressure, body temperature, blood glucose level, birth weight, amount of virus present, time to death, etc.

On the other hand, in the "soft sciences," education, and humanities, variables can be quite a bit more vague, and, as a result, harder to measure.

- Psychology/sociology: intelligence, motivation, sexuality, aspects of mental health, happiness, severity of depression, etc.
- Other soft sciences: unemployment rate, parolee recidivism rates, support for political candidate or social position, etc.
- Education: level of preparation for college, teaching ability, effectiveness of instruction, student participation, etc.
- Humanities: quality of musical performance, historical relevance, artistic impression, quality of translation, etc.

Remark: Statisticians don't like ambiguity. We like variables to be well defined, straightforward to measure, and meaningful to the investigation at hand.

- For example, suppose we want to compare student teachers from USC and Clemson. *Does one program do a better job at preparing its students to teach K-12?*
- How do we measure "preparation?" Smart people may disagree on how this should be done. If we can't agree on how something should be measured, how valid is any measurement you make?

8.2 Counts versus rates

Example 8.1. 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? 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 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 much larger than in Vermont (69)."
  - These counts are based on 2014 data from the National Center for Health Statistics (part of CDC).
  - The rate of annual firearm deaths per 100,000 inhabitants is very close: 10.6/100,000 (Texas) versus 10.3/100,000 (Vermont). This paints a clearer picture of how the states compare.
  - Safest state? Hawaii (2.6/100,000). Worst state? Louisiana (18.9/100,000).
- Example 4 (Measuring highway safety); see Moore and Notz (pp 163-164).

Main point: When comparing two or more groups of individuals, it is usually safer to use a rate or a percentage. Comparing counts can be misleading.

8.3 Measurement validity

Definition: A variable is a **valid** measure of a property if it is relevant as a representation of that property. For example, we have just learned that rates are usually more valid than counts.

Definition: A variable has **predictive validity** if it can be used to predict success on tasks that are related to the property being measured. Statisticians like variables with high degrees of predictive ability (this makes our jobs easier).

Discussion: How would you rate the predictive validity of each variable in the following situations?

- Mammogram image $\longrightarrow$ breast cancer? e.g., how well does a mammogram image predict the presence of breast cancer?
- SAT score $\longrightarrow$ success in college? e.g., how well does a student's SAT score predict his/her success in college?
- Number of sexual partners (with unprotected sex) $\longrightarrow$ STD infection?
- Number of fights with partner $\longrightarrow$ severity of depression?
- Unemployment rate $\longrightarrow$ strength of economy?
- Body mass index $\longrightarrow$ academic achievement in grade school?

What about these?

- Number of gun deaths in each state $\longrightarrow$ risk of being killed by a gun
- High temperatures in 2016 $\longrightarrow$ global warming?
- Insurance type $\longrightarrow$ quality of dental care
- Hair color $\longrightarrow$ admitted to law school?

8.4 Measurement error

Remark: Just because a variable has predictive ability does not mean that we can measure that variable perfectly. For example, each of these variables (left-hand side) has predictive ability to assess the corresponding property (right-hand side):

- Your weight $\longrightarrow$ risk of diabetes?
- Systolic blood pressure $\longrightarrow$ risk of heart attack?
- Air quality in Columbia, SC $\longrightarrow$ incidence of asthma?

What might be some difficulties with obtaining a perfect measurement of these variables?

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

- Bias
- Random error.

Example 8.2. Let's take measuring my weight as an example. When I step on my scale at home, I get a measurement of my weight. This morning (August 29, 2016 at 5.30am), it was 228.8 lbs. This is my measured weight.

Q: Is this my true weight? Probably not. My scale sits on a soft tile floor which may overestimate my true weight. Suppose that my scale systematically over-measures my weight by 2 lbs; i.e.,

$$\text{Measured weight} = \text{True weight} + 2 \text{ lbs.}$$

This source of measurement error is the **bias**. My scale systematically over-measures my true weight each time.

So, you might say, "Well, your weight is 226.8 lbs then." Not so fast! When I stepped on the scale just seconds later, I weighed 228.6. Seconds later, I weighed 228.8 again. Seconds later 228.9, and so on.

In a two-minute span, I weighed myself 10 times. Here were the weights I recorded:

228.8 228.6 228.8 228.9 228.7 228.7 228.6 228.8 228.9 228.7

The source of error we see in these 10 measurements is not related to the scale being biased (2 lbs over-shooting each time). The variation we see in these measurements is because of **random error**.

Therefore, we can think of my measured weight as being composed of three parts:

$$\text{Measured weight} = \text{True weight} + 2 \text{ lbs} + \text{Random error}.$$

The true weight is what I am *trying* to measure. The bias part is the systematic departure from the true weight (due to the scale). The random error part arises because repeated measurements give different results.

Q: What would the 10 measurements above look like if there was **no random error**?
A: They would all be the same (assuming the 2-lb bias was the same for each measurement).

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

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

- 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 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 8.3. Do NOT read this example ahead of time (i.e., before we do it in class).

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 above systematically underestimate or overestimate the true number of F's? This is bias.
- Did your answers above change from reading to reading? This is random error.

Summary: A quantitative variable can be measured perfectly, but only when there is no bias and no random error.

- Removing **bias** can be difficult. Bias usually arises from the instrument used to make the measurement (e.g., bathroom scale, blood-pressure meter, etc.).
- Removing **random error** (i.e., attaining perfect reliability) can also be difficult. The average of several measurements on the same individual is more reliable than a single measurement. "Averaging reduces variation" is a well-known statistical principle. See Example 9 in Moore and Notz (pp 170-171).

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 that a PhD student cited in his 1994 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. Other articles have cited this statistic since.

The problem with this statement, of course, is that it is not true. Not only is it not true, it's absolutely ridiculous. Suppose that in 1950, there was **one child** "gunned down" in the United States. For Sautter's claim to be correct, here are the number of children that must have been "gunned down" since 1950:

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

Dr. Best calls the statement above the "worst social statistic ever."

Discussion: Let's have some fun in mocking 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 during 1975).
- **1978:** $\approx$ 268 million. This exceeds the population of the United States (about 222 million during 1978).
- **1983:** $\approx$ 8.5 billion. This exceeds the population of the world (about 4.7 billion during 1983).
- **1987:** $\approx$ 137 billion. This exceeds the human population throughout all of human history (about 110 billion).
- **1994:** $\approx$ 17.5 trillion. I don't even know how big this number is.
- **2013:** $\approx$ 9 quintillion. This exceeds the number of grains of sand on earth (about 7.5 quintillion, estimated).

The author (when pressed) later amended his claim. His amended claim was that

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

The author said this figure was taken from the Children's Defense Fund (a "not-for-profit" organization operating out of Washington DC). This statement is actually plausible, but we learned in the last chapter that **counts** may not be **valid** when describing a phenomenon such as this.

- The population of the United States also nearly doubled during this time (152.27 million versus 262.13 million). This is a 72 percent increase in the population!
- Therefore, the fact that "the number.....has doubled" is not so surprising.

9.2 Advice from Moore and Notz

Reality: People often perceive statistics as a "fishy science." Some believe you can use statistics to prove anything.

Advice: Moore and Notz advise you to ask 6 questions when you read anything citing statistics. 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? (e.g., # American children "gunned down.")
4. Are the numbers too good to be true?
5. Is the arithmetic right?
6. Is there a hidden agenda? (In the media, the answer to this is usually, "Yes.")

Example 9.2. What didn't they tell us? 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 (hence, the sharp increase).
- A 20-point increase may not be surprising (even if there was no change in incoming student quality).
- Current US NWR Ranking: #61 among liberal arts schools.

Example 9.3. What didn't they tell us? Based on July 2016 data, the Bureau of Labor Statistics calculates the unemployment rate at 4.9%. Let's look at how this is calculated:

- The number of citizens in the civilian labor force (persons classified as employed or unemployed) is 157,833,000.
- The number of unemployed persons is 7,770,000.

Therefore, the official July 2016 unemployment rate is

$$\frac{7,770,000}{157,833,000} \approx 0.0492 \quad (\text{or } 4.9\%).$$

The total US population is 318,900,000 citizens. There are about 250,000,000 adults (aged 18+). Therefore, the total number of citizens in the civilian labor force (157,833,000) excludes many adults:

- military, government employees, the retired, the disabled, persons not actively looking for work, underemployed, discouraged workers, etc.
- **Main point:** 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 4.9%.

Example 9.4. Are the numbers plausible? According to FBI statistics, "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 the same rate all year.

Example 9.5. Is the arithmetic right? 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."

Example 9.6. Is the arithmetic right? Another innumerate *New York Times* writer on a medical testing company, written March 31, 2016:

> "The report also contained what appeared to be comparisons between results from Theranoss proprietary technology and the same samples run on conventional equipment. It notes that the results "should have been within 20 percent of one another." But for one test the results differed by 21 to 130 percent based on nine random samples. For another, the difference ranged from 21 to 39 percent, and for a third it ranged from 22 to 146 percent."

I'm reminded of the following quote: *"Give plenty of statistics. It does not matter that they should be accurate, or even intelligible, as long as there is enough of them."*

Discussion: Some innumerate writers do not understand percents. Percentage changes are even more confusing for them. To calculate a **percentage change** from one time period to the next, use this formula:

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

Example 9.7. My 401K balance on January 1, 2016 was \$7,400,000. (**Note:** I have inflated this number quite a bit). My April 1, 2016 401K balance was \$7,700,000. The amount of the change from Q1 to Q2 was

$$\text{newer value} - \text{older value} = \$7,700,000 - \$7,400,000 = \$300,000.$$

The percentage change is

$$\text{percentage change} = \frac{\$300,000}{\$7,400,000} \approx 0.041 \quad \text{(or about 4.1\%)}$$

My 401K has **increased** by 4.1% from Q1 to Q2.

Example 9.8. The taxable value of my house was \$220,000 in 2005. The taxable value of my house in 2016 was \$181,000. The amount of the change from 2005 to 2016 was

$$\text{newer value} - \text{older value} = \$181,000 - \$220,000 = -\$39,000.$$

The percentage change is

$$\text{percentage change} = \frac{-\$39,000}{\$220,000} \approx -0.177 \;\; (\text{or about } -17.7\%)$$

My house has **decreased** in taxable value by 17.7% from 2005 to 2016.

Remark: A quantity can increase by any amount. A **100% increase** just means that the quantity has doubled, a 200% increase means that the quantity has tripled, and so on. On the other hand, a quantity can not decrease by more than 100%. Once a quantity loses 100% of its value, then there is nothing left.

Example 9.9. Are the numbers consistent with each other? A researcher was performing an experiment on 6 sets of 20 mice each. He reported in the *Journal of Experimental Medicine* the percentage of successes (e.g., cured/not cured) with each set. The percentages he reported were 53, 58, 63, 46, 48, 67. Are these numbers consistent with the experiment? How do you get 53 percent successes with 20 animals?

Example 9.10. Are the numbers plausible? *Organic Gardening* magazine once said that "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 only 3,000 miles. Therefore, 3.9 million is the equivalent of making 650 round trips between these two cities.
- True length of US Interstate Highway System = 47,856 miles.

Example 9.11. Are the numbers plausible? 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."

- If the distribution of incomes is approximately symmetric (e.g., a normal distribution), then this is exactly what one would expect!

Example 9.12. Is the arithmetic right? Poll results for the survey question: "Do you approve of the current bond measure?"

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

Example 9.13. Are the numbers too good to be true? Could 41% of the biology teachers in Louisiana really reject evolution?

9.3 Advice from Dr. Best

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

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

Examples (from DLS):

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.

Discussion: Statistics like these are usually pushed by individuals who have an agenda. 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**).

Conclusion: 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."

As Dr. Best writes 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−tends to accept the figure without question."

My philosophy: "Never believe what you hear in the news (without fully investigating it yourself)."

10 Graphs, Good and Bad

10.1 Categorical variables

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

- **Categorical** $\longrightarrow$ places individuals into one of several groups or categories
- **Quantitative** $\longrightarrow$ 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
- TPN (total parental nutrition; contains all essential fluids and electrolytes and is infused through an IV line)
- 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 tells us (a) what values the variable takes and (b) how often it takes these values.

- A **table** can be used to show the distribution of a **categorical variable**.

Here is the table for nutrition type in the NEC study:

Category	**Breastmilk**	**Fluids**	**TPN**	**Formula**	Totals
Count	237	60	265	43	605
Proportion	0.39	0.10	0.44	0.07	1.00

Remark: 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 above shows the distribution of nutrition type for the infants whose data are not missing. Therefore, the proportions are calculated as

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

Note that the proportions in the categories add up to 1.

Note: We can show the distribution of a categorical variable using graphs: a **bar graph** or a **pie chart**. A bar graph can use either **counts** or **proportions**:

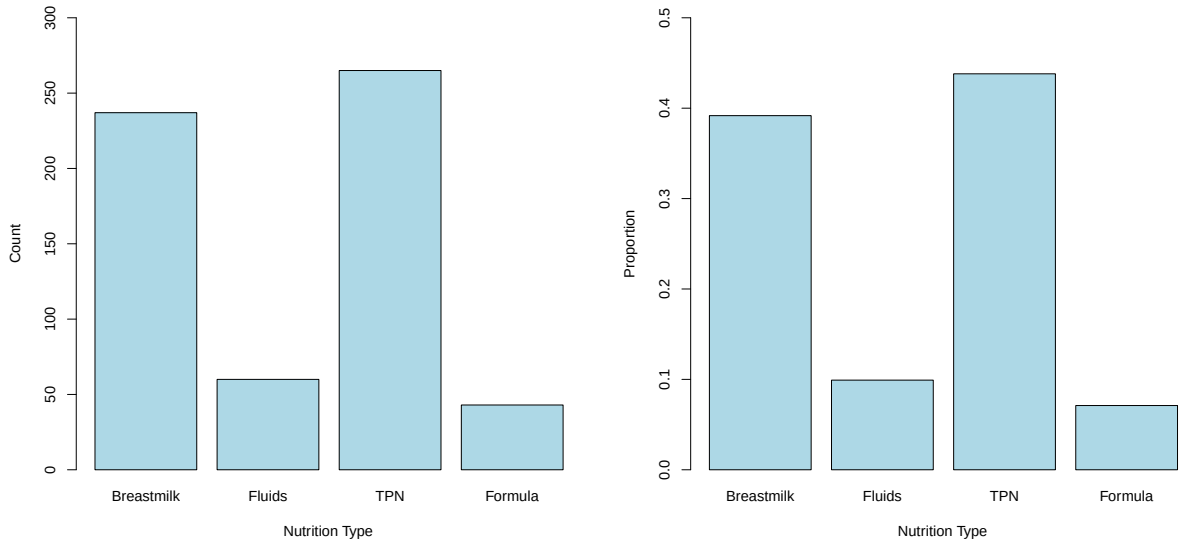


Figure 10.1: Necrotizing enterocolitis data. Bar graphs of nutrition type for 605 infants. Left: Counts. Right: Proportions. Both figures were created using R.

Note: The only difference in the figures above is the **scale** used for the vertical axis.

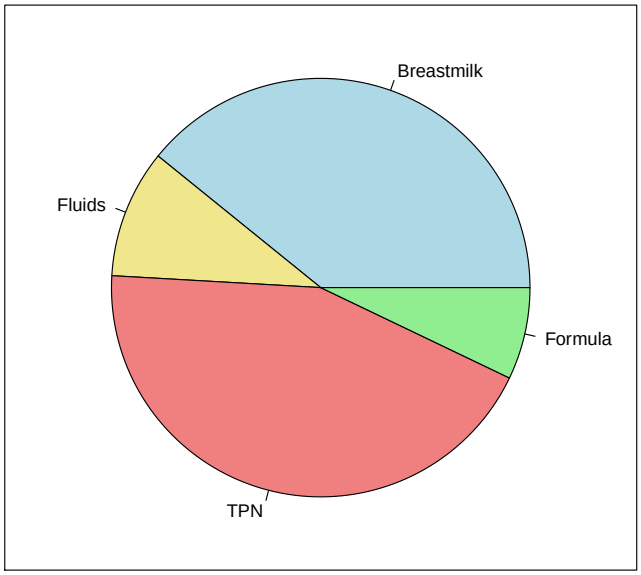


Figure 10.2: Necrotizing enterocolitis data. Pie chart of nutrition type for 605 infants. This figure was created using R.

Note: The distribution of a categorical variable can also be shown using a **pie chart**. Recall there are 360 degrees in a circle. For the nutrition type data, the angles formed in pie chart (in Figure 10.2 above) are

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

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

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	22.0	Arkansas	18.9
Florida	25.3	Georgia	27.5
Louisiana	21.4	North Carolina	26.5
Mississippi	19.6	South Carolina	24.3
Tennessee	23.0	Virginia	34.0

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

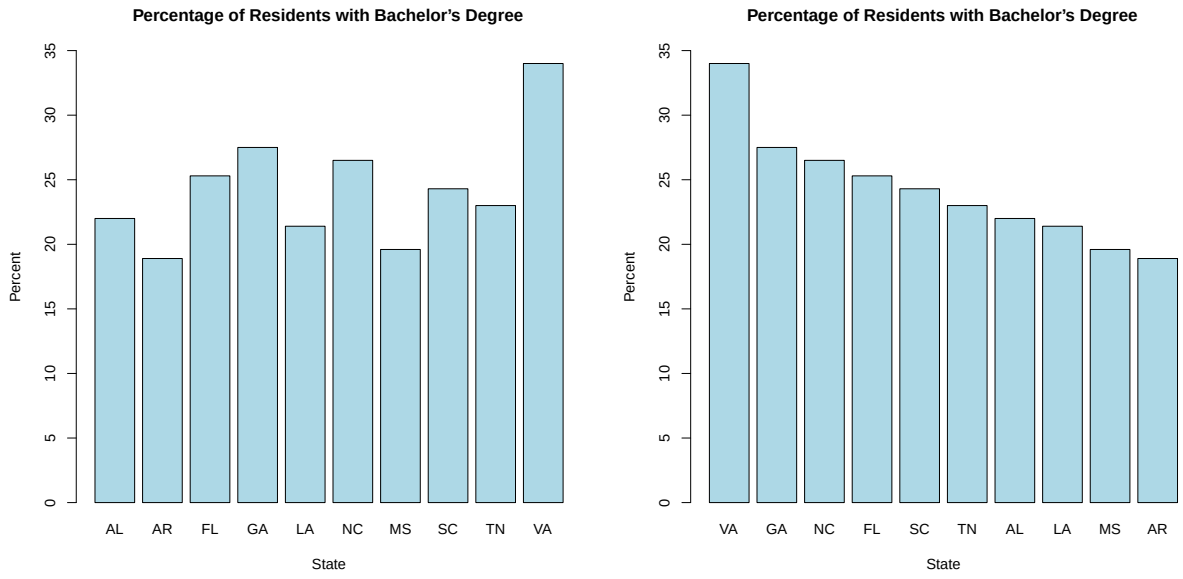


Figure 10.3: Percentage of residents with bachelor's degree for 10 southern states. These data were taken from the American Community Survey in 2011.

Remark: This type of bar graph is used for visual comparison purposes; 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.

10.2 Line graphs (for time series data)

Remark: It is very common to observe quantitative data **over time**. For example,

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

Each of the variables in these examples is quantitative, and we can observe the variable over time. Data that arise over time are called **time series data**. A **line graph** is used to plot time series data over time.

Example 10.3. I recorded the number of students enrolled at USC during the fall semester each year from 1954 to 2015 (Columbia campus only; undergraduate and graduate students combined). These data are available from the 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
⋮	⋮
2013	31,964
2014	32,972
2015	33,724

The variable is the number of fall students enrolled at USC Columbia campus. This is a **quantitative** variable.

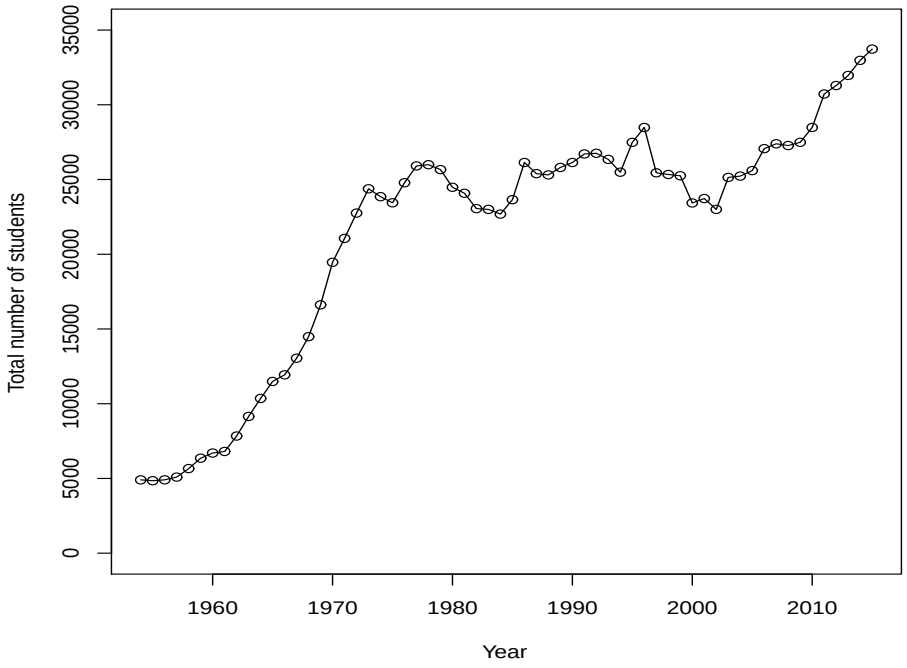


Figure 10.4: USC enrollment data. Total number of students registered for classes on the Columbia campus each fall during 1954-2015. This is an example of a **line graph**.

- There is an **upward trend** in the number of students during 1954-2015.
- Why are graphs like this useful? We can look for trends and other patterns in the data. This information can be useful for making **predictions**.

Remark: Examining data over time in line graphs should include identifying the following characteristics:

- Trends $\longrightarrow$ a long-term upward or downward movement over time
- Sharp deviations $\longrightarrow$ unusual observations that deviate from the general trend
- Seasonal variation $\longrightarrow$ patterns that repeat themselves over time; e.g., each quarter, each year, etc.

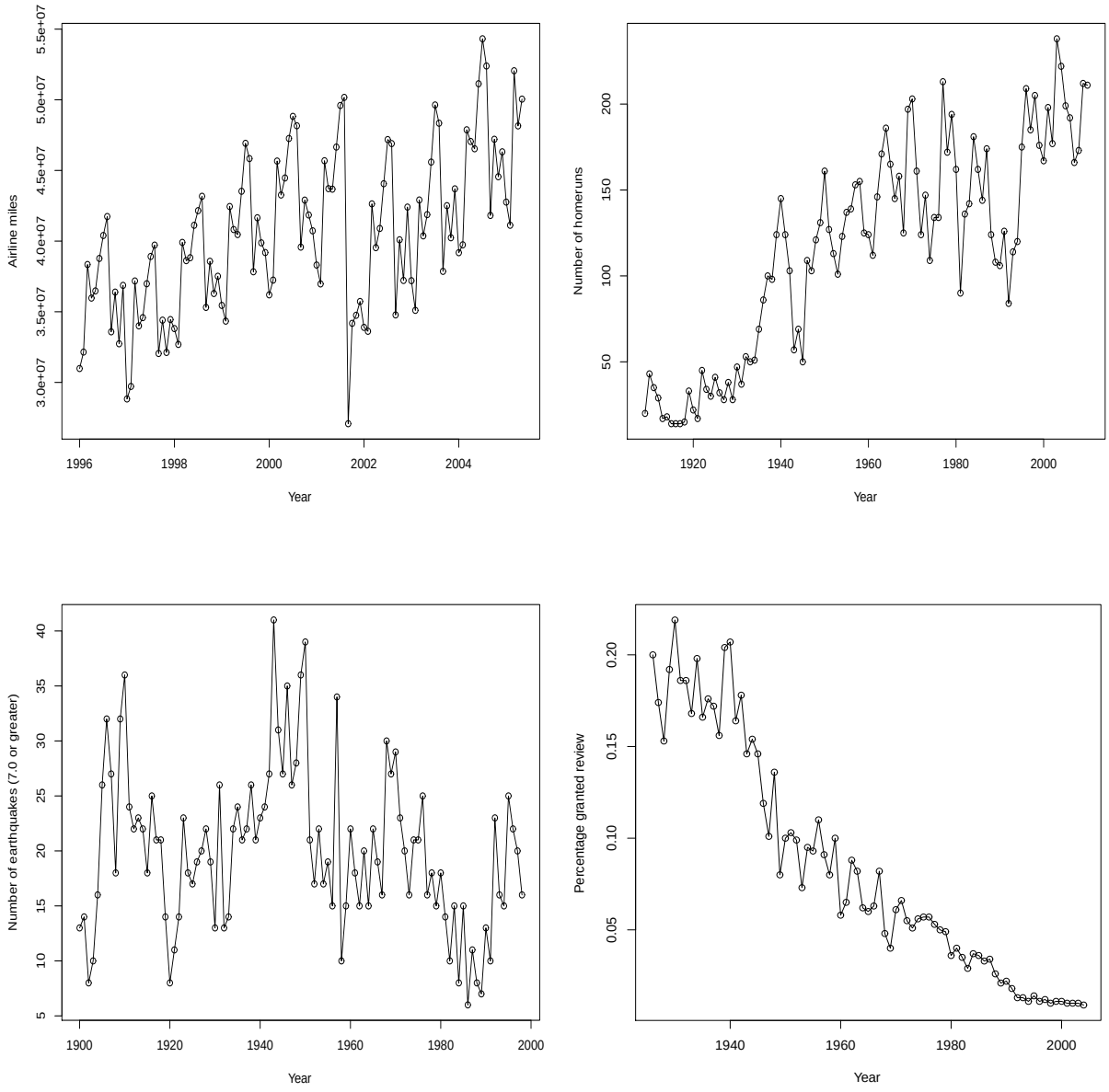


Figure 10.5: **Upper left:** Number of airmiles (in 1000s) traveled by passengers in the US (Jan 1996 to May 2005). **Upper right:** Number of home runs hit by the Boston Red Sox (1909-2010). **Lower left:** Number of earthquakes measuring ≥ 7.0 on the Richter Scale observed worldwide (1900-1998). **Lower right:** Percentage of cases granted review by the US Supreme Court (1926-2004).

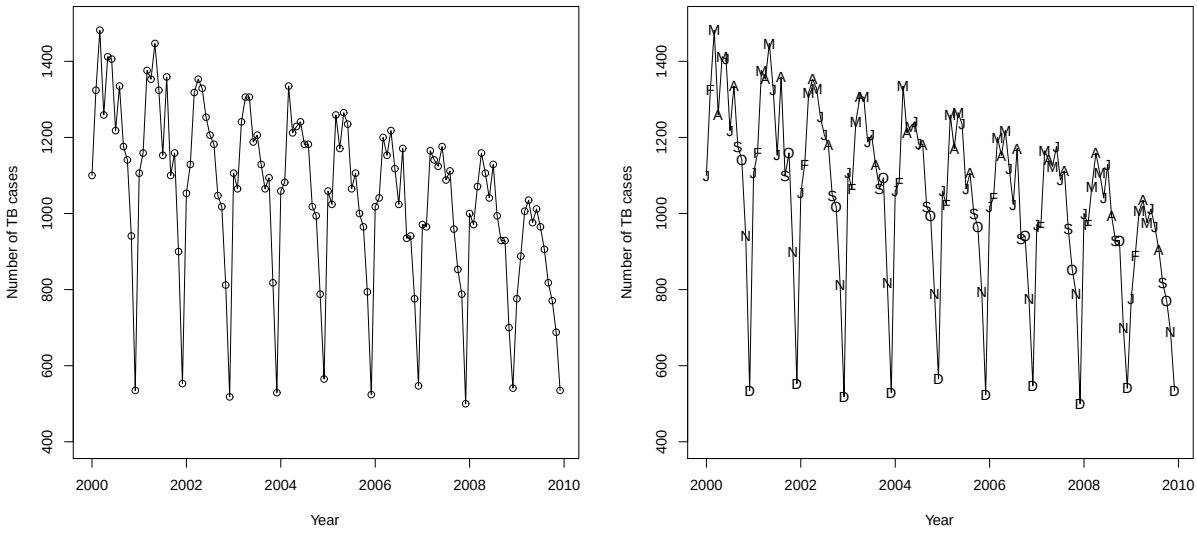


Figure 10.6: Tuberculosis data. Left: Number of TB cases per month in the United States: Jan 2000 through Dec 2009. Right: Monthly plotting symbols added.

Example 10.4. 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. But if the immune system weakens, TB bacteria can become active and ultimately fatal if left untreated.

I recorded the number of TB cases per month reported in the US from the CDC (from Jan 2000 through Dec 2009). There are two noticeable patterns in the data:

- There is **downward trend** over time (yet the minimum counts in December appear to be stable).
- There is **seasonal variation** in the data; i.e., the pattern within each year repeats itself every 12 months.

Note: If you wanted to **predict** the future number of cases ("future" meaning starting in Jan 2010), then your predictions should account for these two patterns.

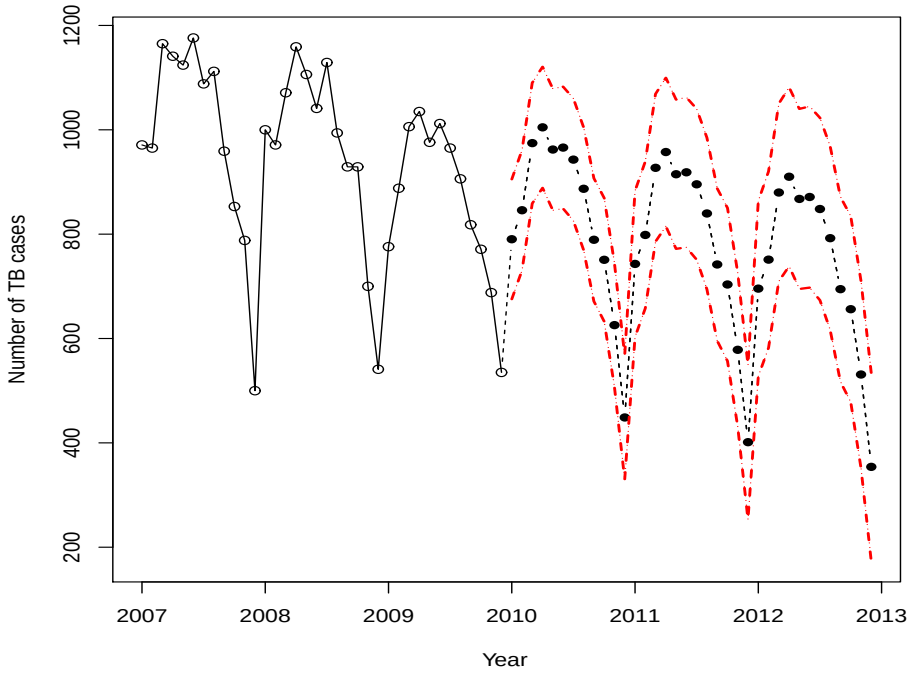


Figure 10.7: Tuberculosis data. Prediction limits for Jan 2010 through Dec 2012.

Predictions: I used a **statistical model** to make predictions for the number of TB cases that would be reported for the next 36 months (Jan 2010 through Dec 2012). This model accounts for both patterns noted above.

- The predictions are shown in Figure 10.7, starting in Jan 2010 (solid dot symbols).
- The predictions are calculated from a statistical model that I used to analyze the data (STAT 520).
- The red bands are called **prediction bands**. We can be 95 percent confident that a future value will be within its corresponding prediction interval limits.
  - For example, a 95 percent **prediction interval** for the TB count in January 2012 is (527, 864); i.e., there will be between 527 and 864 reported cases.
  - This is a prediction we would have made back in December 2009.

10.3 Examples of bad/misleading graphs

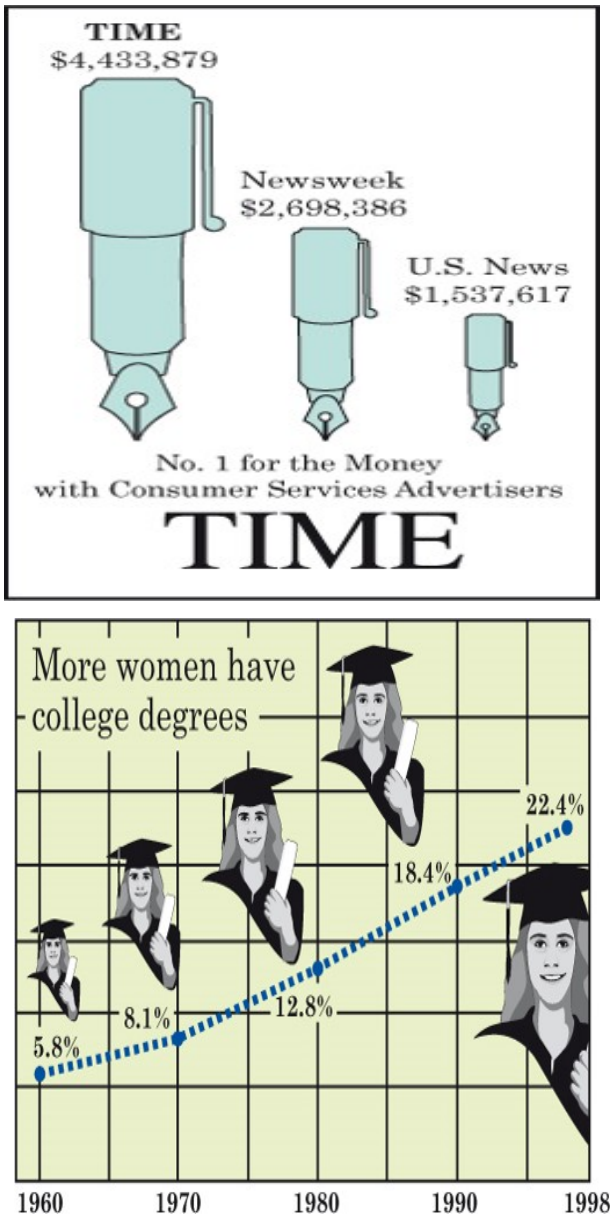


Figure 10.8: Examples of chart junk.

Remark: Apparently, presenting statistical information is so boring that graphs need to be "fancied up" to attract readers. These graphs look nice but serve little purpose other than to mislead and confuse the reader. Graphical displays that display **chart junk** hinder delivering the main message. **Note:** *USA Today* is the worst.

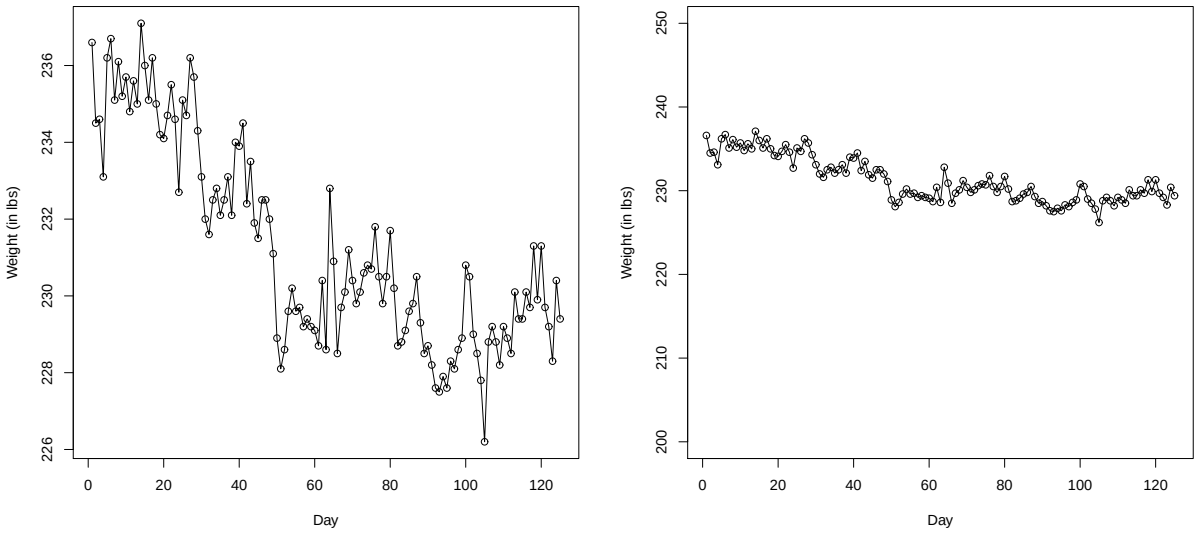


Figure 10.9: Weight data. My weight (in lbs) recorded daily from May 11, 2016 through September 13, 2016 (125 consecutive days). The same data are plotted in each line graph.

Example 10.5. As part of a plan to lose weight, I weighed myself each morning from May 11 through September 13, 2016. A line graph of my daily weight measurements is shown in Figure 10.9. There are 125 daily measurements.

- Both figures in Figure 10.9 plot the same data; the only difference is in the vertical axis scale.
  - The figure on the left is what R chose by default (226-236 lbs).
  - For the figure on the right, I purposefully specified the vertical axis range to be 200-250 lbs. Doing this greatly compresses the data in the figure.
- **Moral:** By changing the scales, you can greatly alter the visual impression of the data.
- Moore and Notz: "Because graphs speak so strongly, they can mislead the unwary." It is very easy to mislead readers with misleading graphs.
- See Example 6 in Moore and Notz (pp 222-223).

11 Displaying Distributions with Graphs

11.1 Histograms

Recall: The **distribution** of a variable tells us (a) what values the variable takes and (b) how often it takes these values. We use graphs to display distributions:

- Categorical variables: bar graph, pie chart
- Quantitative variables: histogram, stemplot, and boxplot (Chapter 12)
  - We can also use a line graph for a quantitative variable if we want to emphasize its time series aspect (see Chapter 10).

Categorical variables take on just a few values (e.g., Breastmilk, Fluids, TPN, Formula). However, **quantitative** variables are numerical and can take on many values. Therefore, different graphical displays are needed. The most common graph used to show the distribution of a quantitative variable is a **histogram**; see, e.g., Figure 11.1.

Remark: Here are Moore and Notz's directions on how to construct a histogram; see MN (pp 240-241):

1. Divide the observations into intervals on the real number line (classes); keep the intervals of equal width.
2. Count the number of observations in each interval; make a table of these counts.
3. Record intervals on the horizontal axis and plot the counts on the vertical axis.

Remark: Here are Tebbs's directions on how to construct a histogram:

1. Ignore Moore and Notz's directions.
2. Use R!

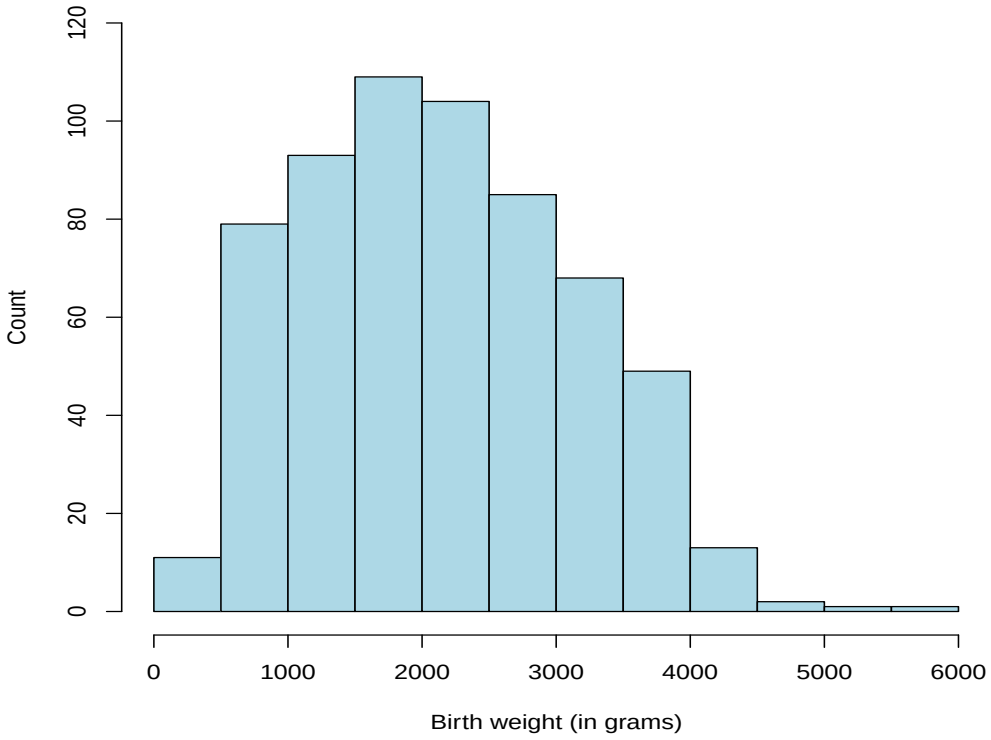


Figure 11.1: Necrotizing enterocolitis study. Histogram of birth weights for 615 infants. This figure was created using R's **default settings** for interval width.

Example 11.1. Recall the necrotizing enterocolitis study in Example 1.1 (notes). There were 615 infants in the study. The birth weight (measured in grams) was recorded for each infant. A histogram of these observations is shown in Figure 11.1.

- For these data, R's default selection for the **interval width** is 500 grams. This is the width of each interval shown in Figure 11.1.
- Default settings in R (or other software) are usually adequate. However, these can be changed. In Figure 11.2 (next page),
  - I used interval widths of 100 grams (left). This might be too narrow.
  - I used interval widths of 2000 grams (right). This is clearly too wide.

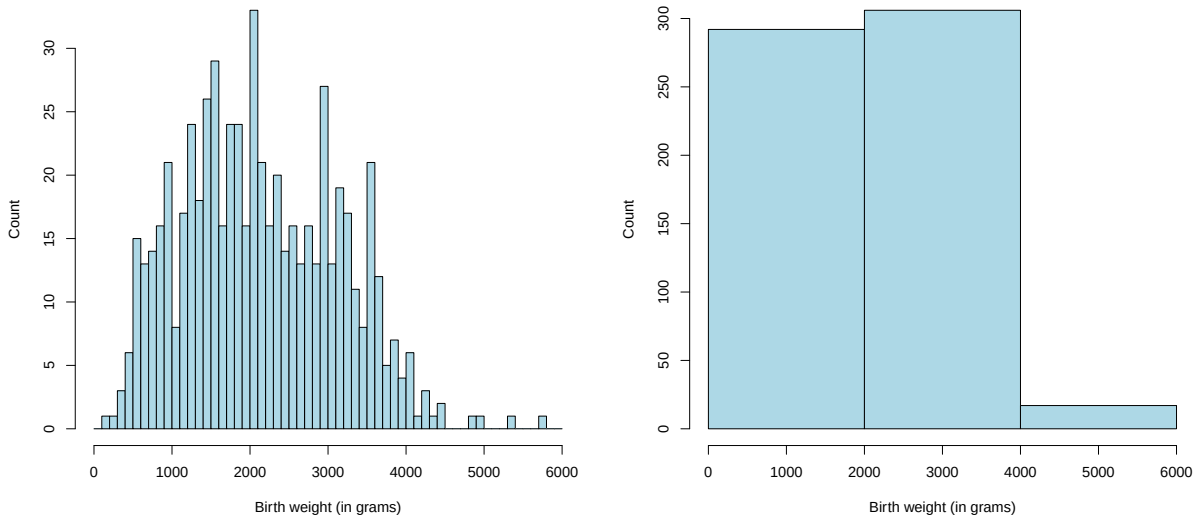


Figure 11.2: Necrotizing enterocolitis study. Histograms of birth weights for 615 infants. Left: Interval width = 100 grams; Right: Interval width = 2000 grams.

My advice: When making a histogram in R, try the default settings first. If you do not like this, use trial and error to get the appearance that best displays the distribution.

- There is no "right" way to pick the interval widths, but there are certainly bad ways to do it.

11.2 Interpreting histograms

Remark: 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 observations recorded for a quantitative variable (like birth weight).
- If the observations we have are from a **sample**, then the histogram presents an impression of the underlying distribution for the variable in the **population**.

- Therefore, by interpreting characteristics we see in the histogram, we are interpreting what may be "going on" in the larger population of individuals.

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

1. Overall pattern of the distribution
    - Center: Where does the center of the distribution fall approximately?
    - Spread: How much variation is in the distribution? How spread out is it? What is the range of possible values?
    - Shape: What type of shape does the distribution have?
2. Deviations from the overall pattern (e.g., outliers, etc.)
    - **Definition:** An **outlier** is an individual observation that falls outside the overall pattern of the distribution.

Preview: In Chapter 12, we will introduce numerical values (statistics) that quantify the notions of "center" and "spread."

Example 11.2. In an observational study examining aspects related to childhood obesity, Baxter and colleagues (2012) measured the body mass index (BMI) of $n = 328$ fourth-grade children sampled from a large public school district in Augusta, GA. An individual's BMI is calculated as follows:

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

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.

A histogram of the BMI data from the study is shown in Figure 11.3.

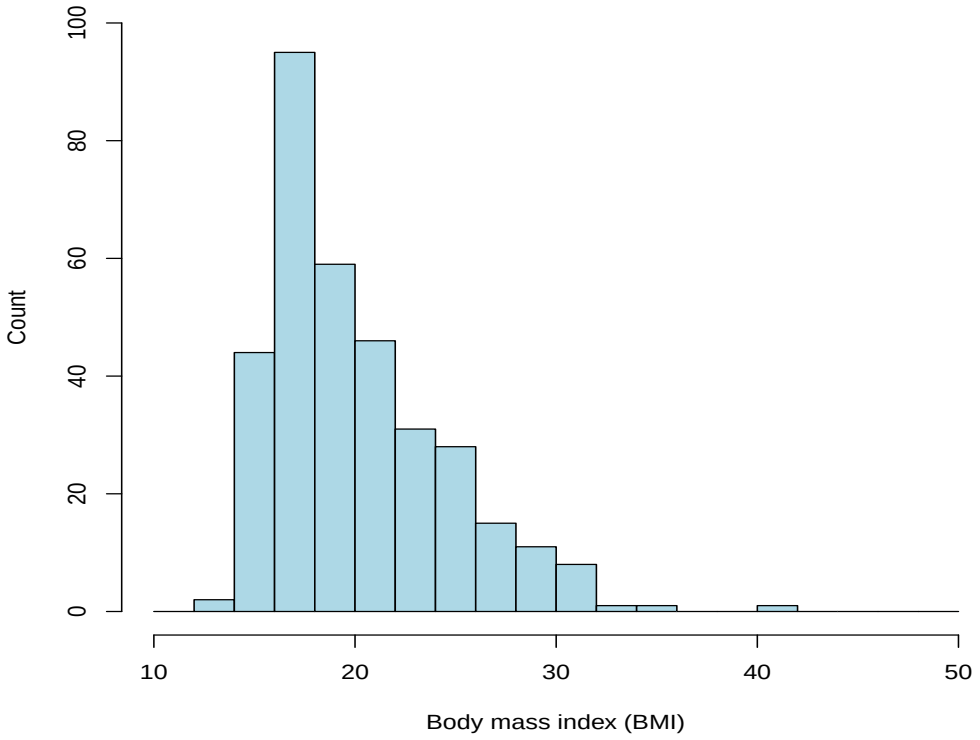


Figure 11.3: Childhood obesity study. Histogram of BMI data for $n = 328$ fourth grade students. **Note:** I made alterations to R's default settings in constructing this histogram.

Interpretation:

- The center of the distribution is around 20 (which recall is the cutoff between "Healthy" and "Overweight").
- Most of the BMI measurements are between 14 and 32 (a rough range).
  - There are a couple of observations less than 14. There are also a few observations greater than 32. Clearly, the observation larger than 40 is an outlier.
- Shape? This distribution has a single peak (around 18) and is **skewed to the right** side.

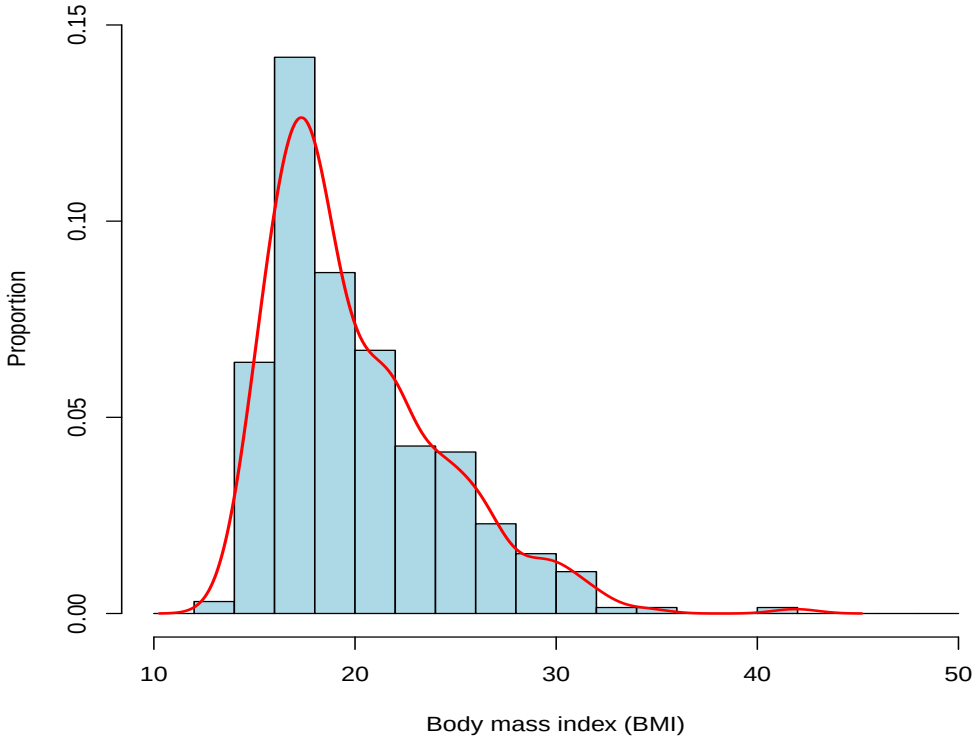


Figure 11.4: Childhood obesity study. Histogram of BMI data for $n = 328$ fourth grade students. An estimate of the population density curve has been added.

Discussion: The histogram shows the distribution of the **sample** values (i.e., the BMI observations for the 328 children in the sample).

- If the sample is representative of a larger population (e.g., all fourth-grade children in Augusta), then the histogram may be "approximating" a smooth curve that describes this population.
- We call this smooth curve the **population density curve**. This curve describes how the variable is distributed in the population (see Chapter 13).
- R can calculate an **estimate** of this population density curve using the sample of measurements (see Figure 11.4 above).

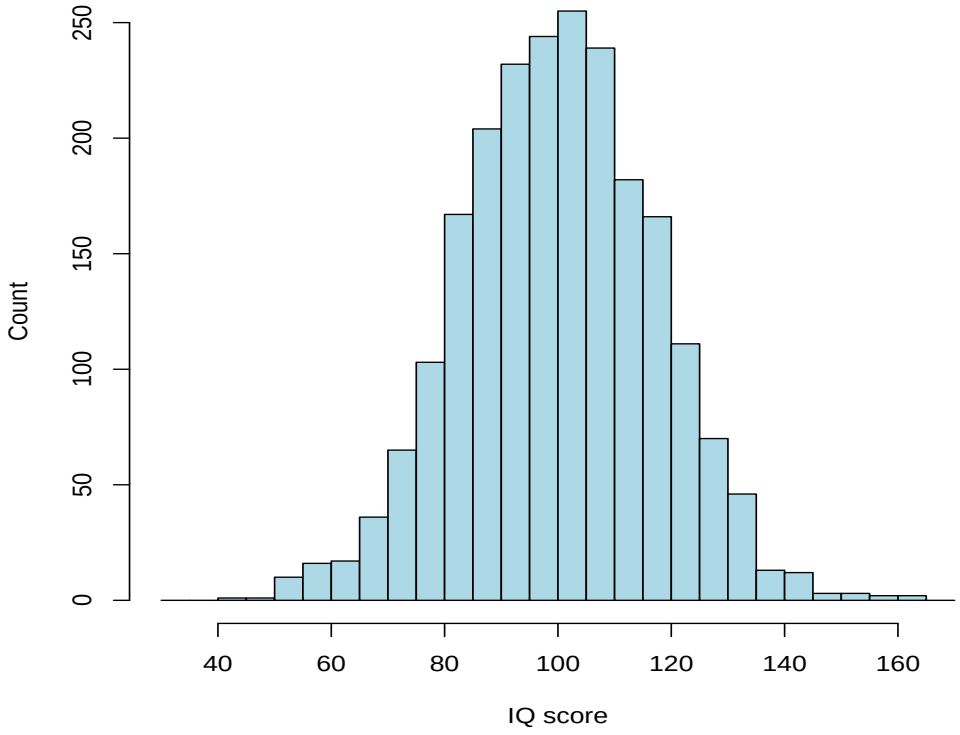


Figure 11.5: Wechsler Adult Intelligence Scale (WAIS-IV) IQ data. Histogram of IQ scores for a sample of $n = 2200$ Americans.

Example 11.3. The Wechsler Adult Intelligence Scale (WAIS) is an IQ test. A recent version of this test was administered to a sample of $n = 2200$ individuals in the United States (aged 16-90). The histogram of the scores is shown in Figure 11.5.

Interpretation:

- The center of the distribution is around 100.
- Most of the IQ scores are between 60 and 140, but there are a few outside this range. There are no striking outliers.
- Shape? This distribution has a single peak (around 100) and is approximately **symmetric**.

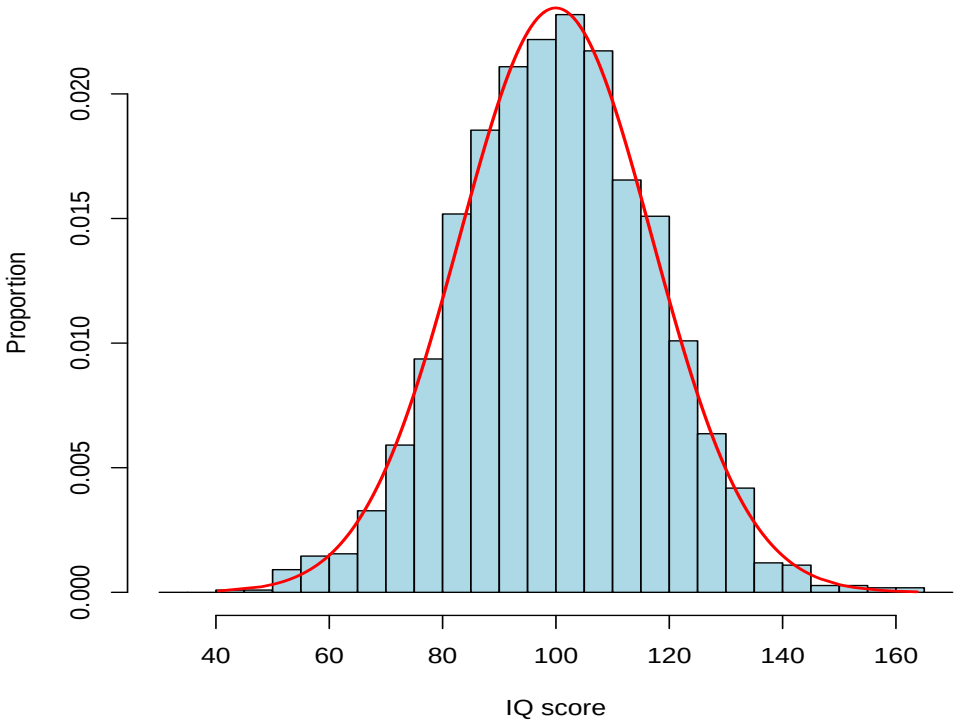


Figure 11.6: Wechsler Adult Intelligence Scale (WAIS-IV) IQ data. Histogram of IQ scores for a sample of $n = 2200$ Americans. An estimate of the population density curve has been added.

Discussion: A smooth curve has been superimposed over the IQ sample data in Figure 11.6. This is an **estimate** of the population density curve.

- In this example, the population density curve describes the distribution of IQ scores for the entire population of individuals (i.e., all Americans).
- I asked R to estimate the smooth curve assuming that the population distribution was **normal** (i.e., a **normal distribution**).
- The normal distribution is the most common population density curve. We will study normal distributions in Chapter 13.

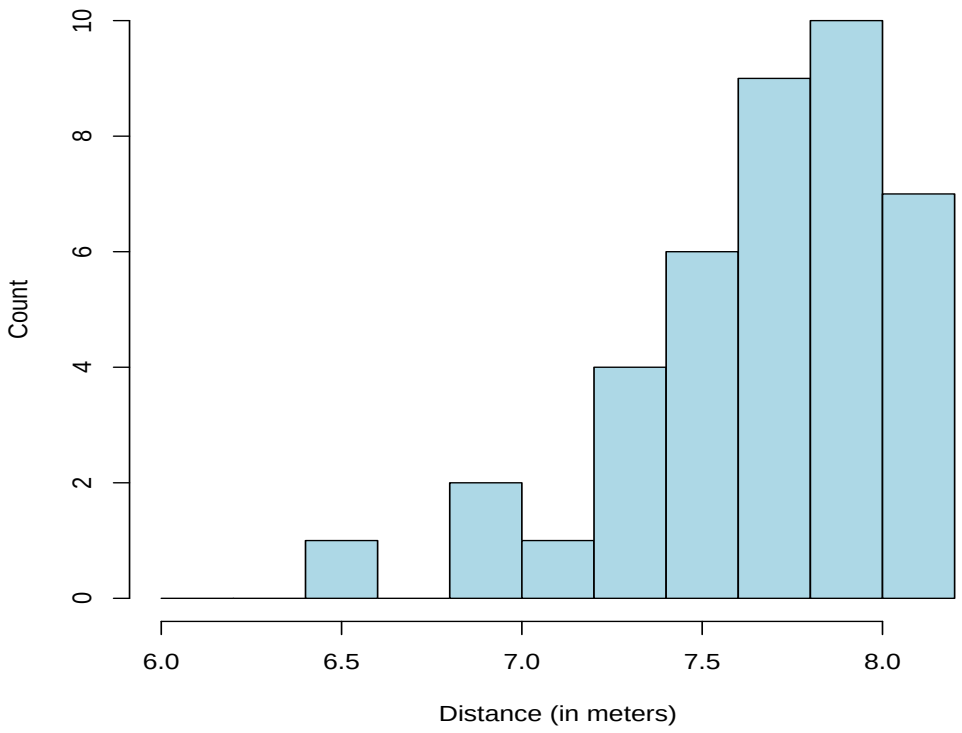


Figure 11.7: 2012 Summer Olympics data. Long-jump distances (in meters) for 40 male athletes.

Example 11.4. Figure 11.7 displays the long-jump distances 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:

- The center of the distribution is around 7.7-7.8 meters.
- Most of the distances are between 7.0 and 8.2 meters, but there are a few below 7.0. The observation around 6.5 meters may be an outlier.
- Shape? This distribution has a single peak (around 7.8 meters) and is **skewed to the left** side.

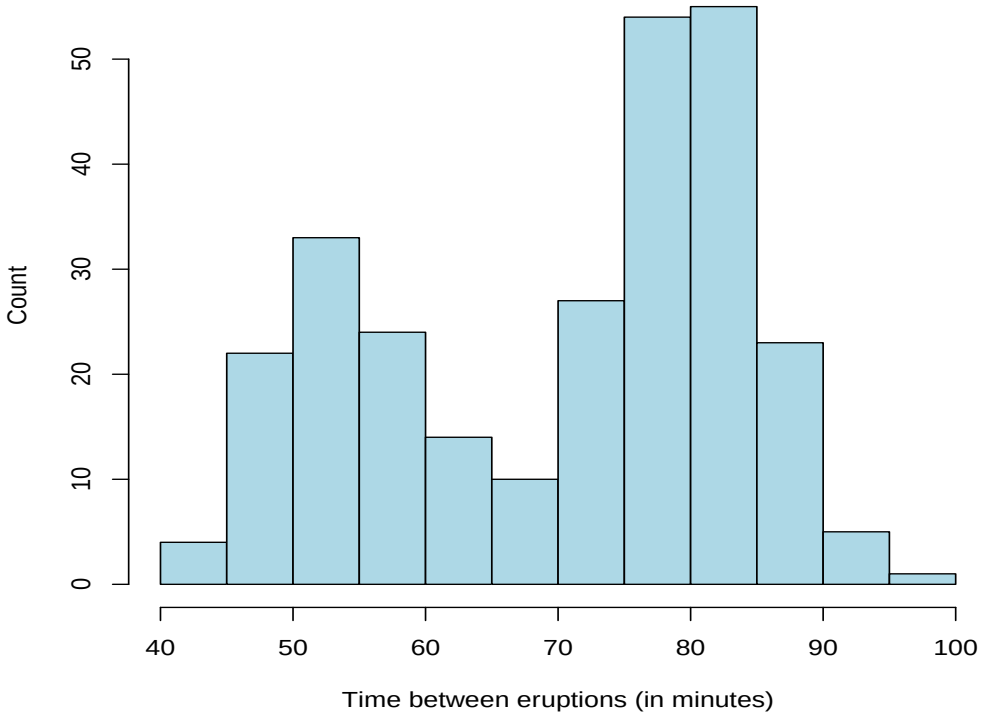


Figure 11.8: Old Faithful geyser data. The time between consecutive eruptions recorded on 272 occasions.

Example 11.5. The histogram in Figure 11.8 shows the times between eruptions of the Old Faithful geyser in Yellowstone National Park, USA, recorded on 272 occasions. These data were collected in the 1980s.

Interpretation:

- Shape? This distribution has a double peak−one peak around 50 minutes and another around 80 minutes. This is an example of a **bimodal distribution**.
- The center of the distribution is around 70 minutes, but this figure may be misleading because the distribution is bimodal.
- All of the times are between 40 minutes and 100 minutes. There are no outliers.

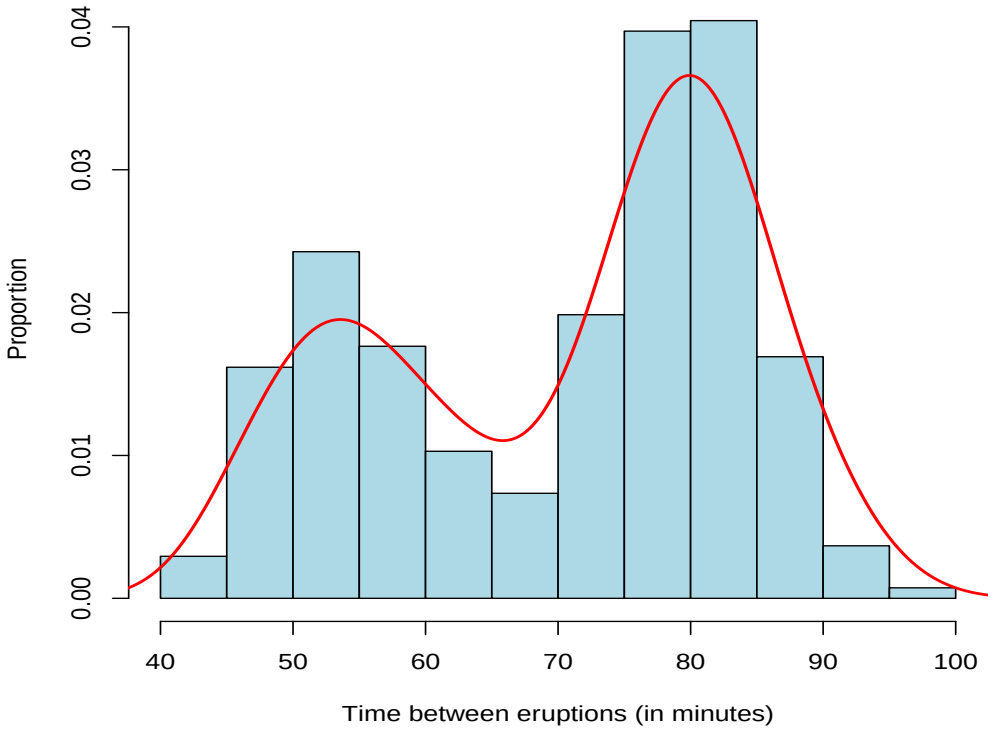


Figure 11.9: Old Faithful geyser data. The time between consecutive eruptions recorded on 272 occasions. An estimate of the population density curve has been added.

11.3 Stemplots

Note: For smaller data sets, say 20-100 observations, a **stemplot** can be useful when showing the distribution of a quantitative variable.

- Stemplots display the distribution while retaining the numerical values in the data set. The idea is to separate each data value into a **stem** and a **leaf**.
  - Stems are plotted on the leftmost column.
  - Leaves are plotted on the right side in ascending order.
- Ignore Moore and Notz's directions on pp 249. Use R!

Example 11.6. Here are the final exam scores from an undergraduate course I taught when I was at Oklahoma State University. There were 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" $\longrightarrow$ "9"
- **Leaf:** 1's digit; e.g., "98" $\longrightarrow$ "8"

Here is the **stemplot** for these data (constructed using R):

```
> stem(final.exam)

  2 | 2
  3 |
  4 | 47
  5 | 122477
  6 | 3335799
  7 | 00013446667778889999
  8 | 01111223345677889
  9 | 0001123355688
```

Interpretation:

- The center of the distribution is somewhere between 70 and 80.
- Most of the scores are between 44 and 98. There is an obvious outlier at 22.
- Shape? This distribution is **skewed to the left** side.

Remark: It can be instructive to see both the histogram and the stemplot for the same data. We do this on the next page.

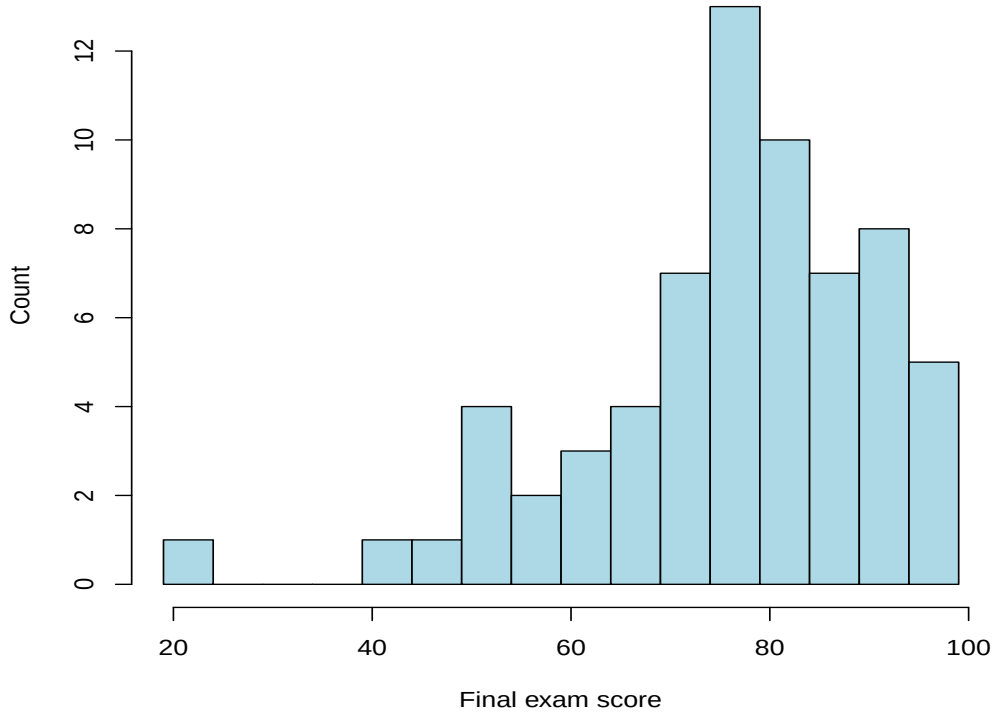


Figure 11.10: OSU final exam data. Final exam scores for 66 students.

```
> 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 Chapter 11, we used histograms and stemplots as graphs to display the distribution of a **quantitative** variable.

- Two important characteristics of a histogram (or stemplot) are its "center" and its "spread."
- In this chapter, we introduce numbers (statistics) that describe the center and spread of a distribution. We also introduce a new graph to display quantitative distributions−the **boxplot**.
- This chapter is only relevant for quantitative variables. We will avoid excessive hand calculations, relying on R whenever possible.

12.2 Median, quartiles, 5-number summary, and boxplots

Example 12.1. Non-small cell lung cancer (NSCLC) is the most common type of lung cancer in humans (roughly 85% of all cases). A study in Japan examined a small group of NSCLC patients who had been treated with both gefitinib and erlotinib (two cancer drugs). Here are the times until treatment failure (TTF, in months) for $n = 14$ 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, withdrawal from treatment due to adverse reaction, or death.

Source: Takeda et al. (2012). Clinical impact of switching to a second EGFR-TKI after a severe AE related to a first EGFR-TKI in EGFR-mutated NSCLC. *Japanese Journal of Clinical Oncology* **42**, 528-533.

Here are the TTF observations, ordered from low to high:

$$\underbrace{0.5 \quad 0.8 \quad 1.4 \quad 2.4 \quad 4.2 \quad 5.3 \quad 7.5}_{\text{lower half}} \quad \underbrace{8.2 \quad 13.4 \quad 14.0 \quad 15.6 \quad 16.1 \quad 20.4 \quad 68.9}_{\text{upper half}}$$

Definition: The **median** M is the midpoint of a distribution. Half of the observations are smaller; half are larger. The median measures the "center" of a distribution.

- For the TTF observations in Example 12.1, the median is calculated as

$$M = \frac{7.5 + 8.2}{2} = 7.85,$$

the average of the middle two observations. The median TTF is 7.85 months.

Definitions: The **first quartile** Q_1 is the median of the lower half of the observations. The **third quartile** Q_3 is the median of the upper half.

- For the TTF observations in Example 12.1, the quartiles are

$$\begin{aligned} Q_1 &= 2.4 \\ Q_3 &= 15.6. \end{aligned}$$

Definitions: The **minimum** (Min) is the smallest observation in a data set. The **maximum** (Max) is the largest.

- For the TTF observations in Example 12.1, the minimum and maximum are

$$\begin{aligned} \text{Min} &= 0.5 \\ \text{Max} &= 68.9. \end{aligned}$$

Definition: The **5-number summary** of a data set consists of these 5 values:

$$\text{Min}, Q_1, M, Q_3, \text{Max}.$$

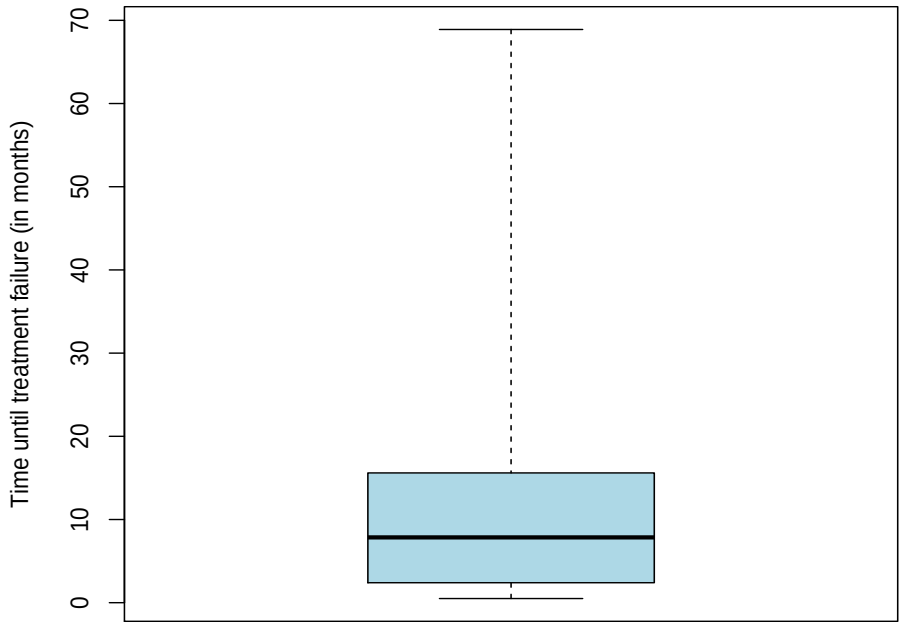


Figure 12.1: NSCLC data. Boxplot of the times until treatment failure (in months) for $n = 14$ patients. This figure was created using R.

- For the TTF observations in Example 12.1, the 5-number summary is

$$\text{Min} = 0.5 \quad Q_1 = 2.4 \quad M = 7.85 \quad Q_3 = 15.6 \quad \text{Max} = 68.9.$$

Definition: A **boxplot** is a graphical display that uses the 5-number summary.

- A central box spans the quartiles Q_1 and Q_3.
- A solid line marks the median M.
- Lines extend from the box to the minimum and maximum values.

A boxplot for the TTF data in Example 12.1 is shown in Figure 12.1.

Note: The median and quartiles are easy to calculate in R:

```
> sort(TTF) # sort 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
```

As the `quantile` output in R suggests, we can think of

- Q_1 as the "25th percentile" of the distribution (1/4 of the data fall below this)
- M as the "50th percentile" of the distribution (1/2 of the data fall below this)
- Q_3 as the "75th percentile" of the distribution (3/4 of the data fall below this).

Definition: The **interquartile range**, IQR, of a distribution is the difference between the quartiles; i.e.,

$$\text{IQR} = Q_3 - Q_1.$$

- For the TTF observations in Example 12.1, the interquartile range

$$\text{IQR} = 15.6 - 2.4 = 13.2.$$

  This is the width of the central "box" in the boxplot shown in Figure 12.1.

Interpretation: The IQR measures the "spread" in the **middle 50%** of a distribution; for example,

- approximately 50% of the patient TTF observations are within the 13.2 month range identified by the IQR (i.e., 2.4 months to 15.6 months).

Outliers: The IQR is commonly cited when measuring the "spread" in a data distribution. It is also used to classify observations as **outliers**. 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 $\longrightarrow$ outlier on the low side
- 1.5(IQR) **above** the third quartile $\longrightarrow$ outlier on the high side.

Example 12.2. Arsenic is a chemical element (As) 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 12.2 show the distribution of arsenic concentrations (in parts per billion, ppb) for a random sample of $n = 102$ water wells in Texas.

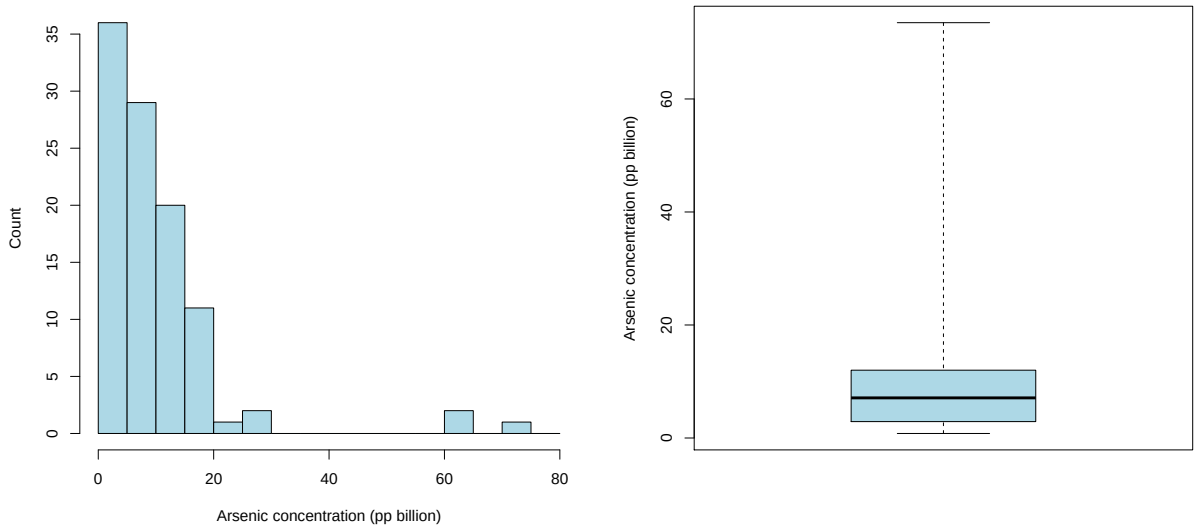


Figure 12.2: Arsenic data. Concentration of arsenic (in ppb) in ground water for a random sample of $n = 102$ wells in Texas.

Calculations: With $n = 102$ observations in the sample, it is best to use R to calculate the 5-number summary; here is the output:

```
> quantile(arsenic,type=2) # 5-number summary
  0%  25%  50%  75% 100%
 0.8  2.9  7.1 12.0 73.5
```

The interquartile range of the distribution is

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &= 12.0 - 2.9 = 9.1. \end{aligned}$$

- **Interpretation:** Approximately 50% of the arsenic concentrations are within the 9.1 ppb range identified by the IQR (i.e., 2.9 ppb to 12.0 ppb).

Outliers:

- Any observation falling **below**

$$Q_1 - 1.5(\text{IQR}) = 2.9 - 1.5(9.1) = -10.75$$

would be considered an outlier. This does not make sense because arsenic concentrations (ppb) must be greater than or equal to 0.

- Any observation falling **above**

$$Q_3 + 1.5(\text{IQR}) = 12.0 + 1.5(9.1) = 25.65$$

would be considered an outlier. There are **5** wells in the data set that have concentrations larger than 25.65 ppb:

```
> sort(arsenic[arsenic>25.65])
[1] 28.0 28.1 62.1 63.0 73.5
```

- The boxplot that identifies these outliers is shown in Figure 12.3.

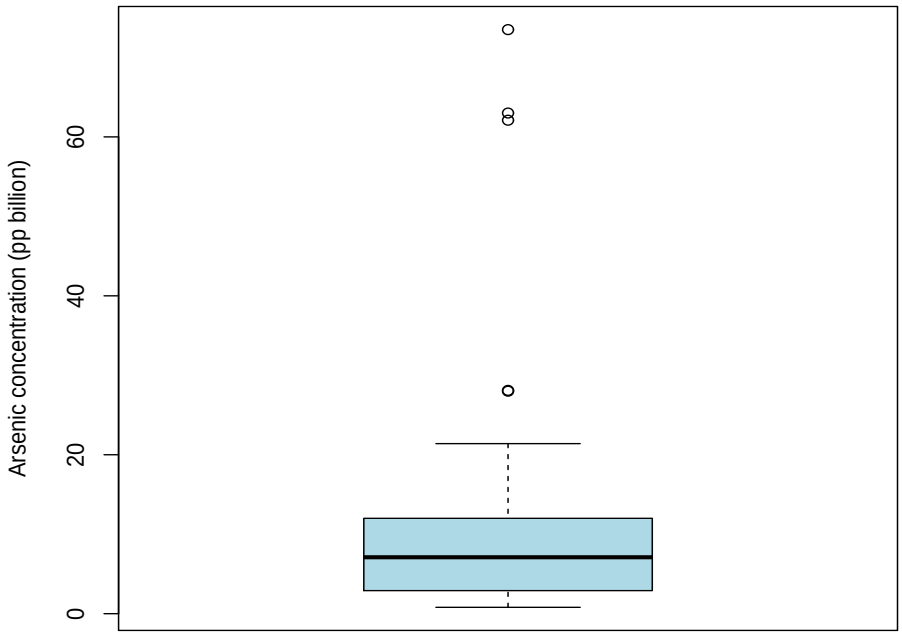


Figure 12.3: Arsenic data. Concentration of arsenic (in ppb) in ground water for a random sample of $n = 102$ wells in Texas. Five outliers on the high side have been identified (28.0, 28.1, 62.1, 63.0, and 73.5).

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 private college 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. Side-by-side boxplots of the data are in Figure 12.4.

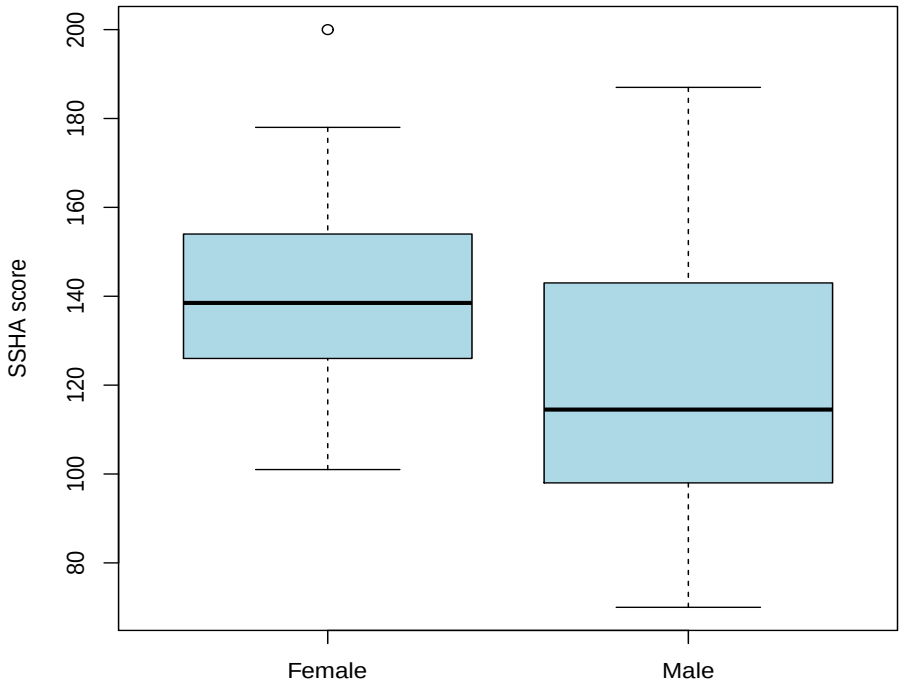


Figure 12.4: SSHA exam score data. Exam scores for 18 female and 20 male students. The female score at 200 has been declared an outlier by the 1.5(IQR) rule.

Discussion: Graphing boxplots side by side (as in Figure 12.4) allows us to **compare** two data distributions. For example,

- Which sample (female or male) has a larger median score?
- Which sample has more variability (spread) in its scores?
- What is the shape of each distribution?

Statistical inference: What do these two samples say about the larger populations of female and male students? Are you prepared to conclude that

- females score better on the SSHA?
- females have better motivation, study habits, and attitudes?

12.3 Mean and standard deviation

Definition: With a sample of observations $x_1, x_2, ..., x_n$, the **mean** $\overline{x}$ (pronounced "x-bar") is given by

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}.$$

In other words, the mean is the **average** of the n values $x_1, x_2, ..., x_n$.

- The symbol

$$\sum$$

  is the capital Greek letter "sigma." It simply means "add."

- The mean $\overline{x}$ is the balancing point of a distribution.

- Like the median M, the mean $\overline{x}$ is a measure of the **center** of a distribution.

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 (of someone wearing the armor). 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.

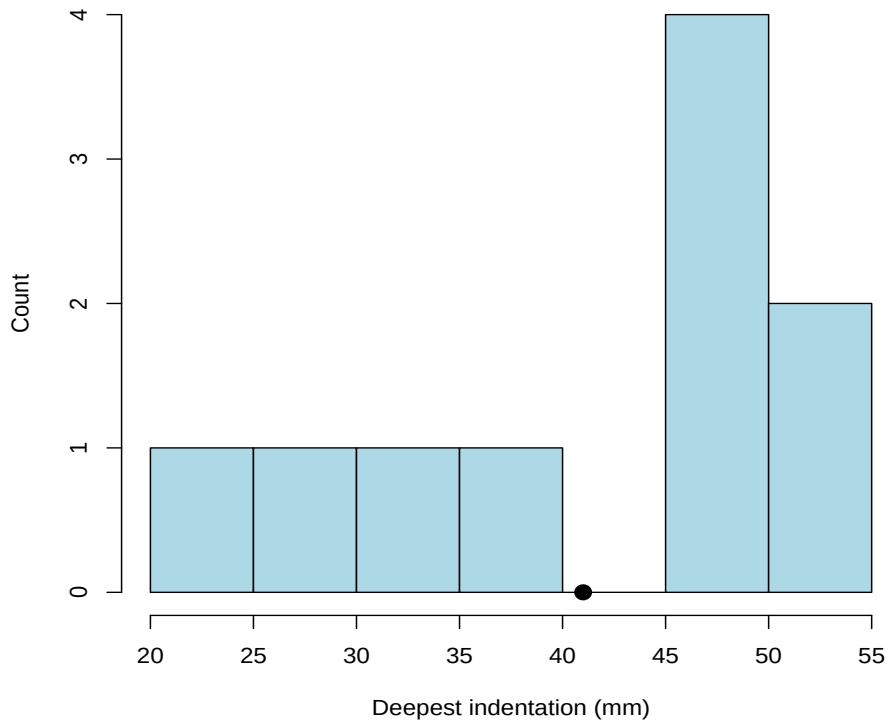


Figure 12.5: Clay model data. Depth of the deepest indentation (in mm) for a sample of $n = 10$ clay models. The mean $\overline{x} = 41.0$ is identified with a solid circle.

R: We can also calculate the mean in R:

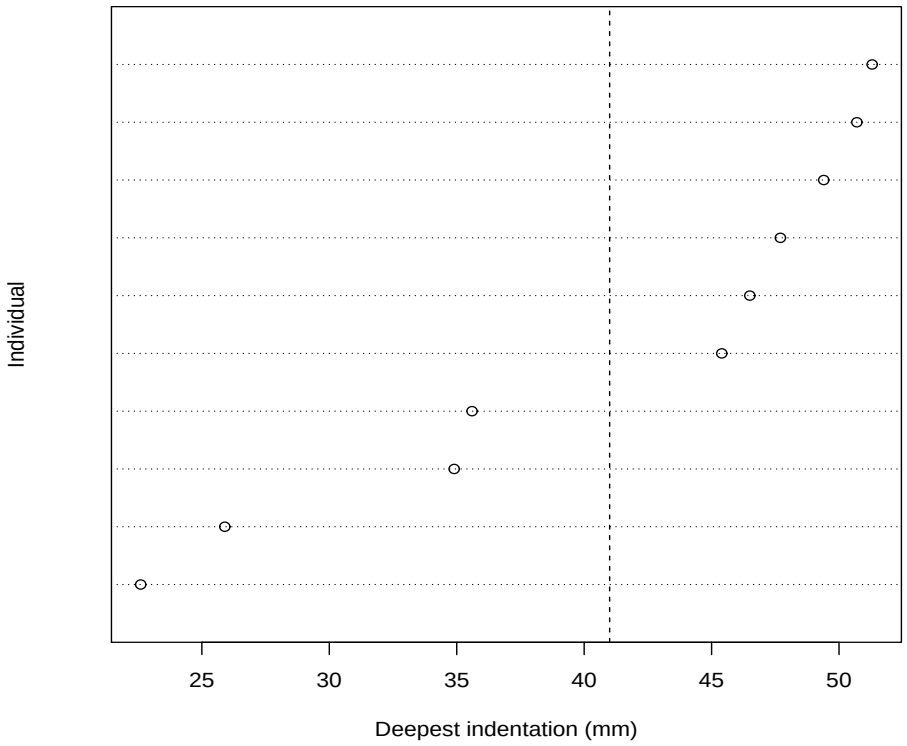
```
indentation = c(22.6,25.9,34.9,35.6,45.4,46.5,47.7,49.4,50.7,51.3)
> mean(indentation)
[1] 41
```

Definition: With a sample of observations $x_1, x_2, ..., x_n$, the **standard deviation** s is given by

$$s = \sqrt{\frac{\sum(x - \overline{x})^2}{n-1}}.$$

The standard deviation can be interpreted as "an average distance from the mean."

Clay model data: Here is a figure and table that explains how the standard deviation is calculated with the clay model indentation data:



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,

$$s = \sqrt{\frac{\sum(x-\overline{x})^2}{n-1}} = \sqrt{\frac{998.18}{9}} = \sqrt{110.91} = 10.53.$$

The standard deviation is 10.53 mm.

R: It is easier to calculate the standard deviation in R:

```
> sd(indentation) # standard deviation
[1] 10.53133
```

Remark: Here are some guidelines on how to interpret the standard deviation:

- The standard deviation is measured in the **original units** of the data.
- The larger the standard deviation, the more variation (spread) in the distribution.
- The smallest the standard deviation can be is $s = 0$. This occurs when all the observations are the same. For example,

  ```
  > data = c(5,5,5,5,5)
  > sd(data)
  [1] 0
  ```

  In the data set $(5, 5, 5, 5, 5)$, there is "no spread."

- The **variance** is defined by

  $$s^2 = \frac{\sum(x-\overline{x})^2}{n-1}.$$

  The standard deviation is the (positive) square root of the variance.

  - The variance also measures the spread of a distribution. However, it is measured in **squared units** (not original units). This is less meaningful if, for example, the data are measured in dollars, months, points, etc.

  Other statistics books introduce the variance first; then the standard deviation. Moore and Notz don't even bother with the variance.

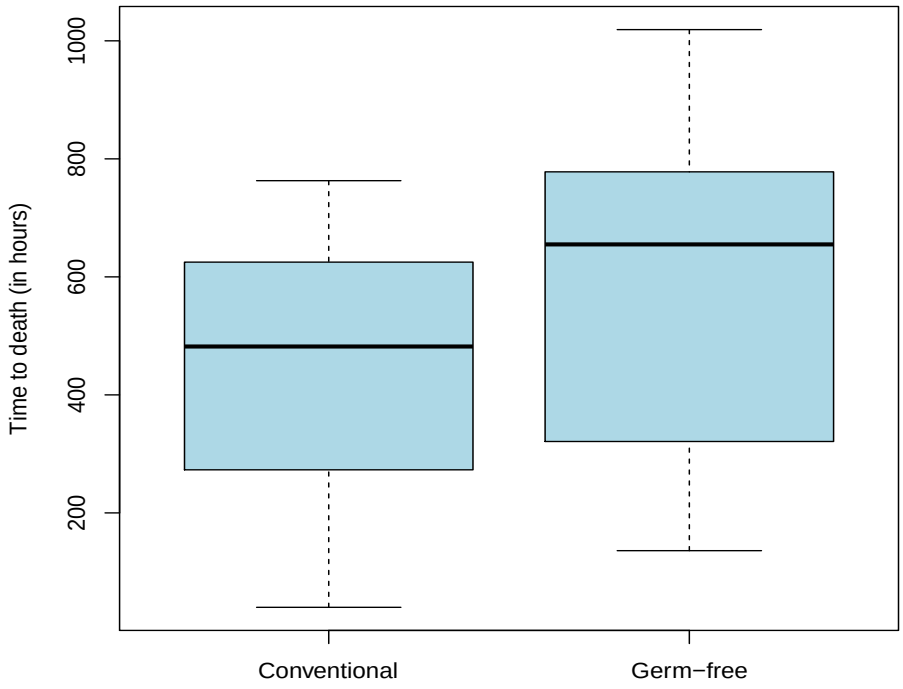


Figure 12.6: Mice data. Time to death (in hours) for two different environments.

Example 12.5. Carolan and Tebbs (2005) summarize the results of an experiment designed 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).

Q: Which group has a larger **mean** time until death?

```
> mean(conventional)
[1] 456.596
> mean(germ.free)
[1] 582.561
```

Q: Which group has a larger **standard deviation**?

```
> sd(conventional)
[1] 195.9897
> sd(germ.free)
[1] 254.4978
```

12.4 Choosing numerical descriptions

Discussion: Let's revisit the time to treatment failure 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] 12.76429
```

Q: Why is the mean larger?

A: It is largely due to the outlier observation 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

```
> median(TTF.no.outlier)
[1] 7.5
> mean(TTF.no.outlier)
[1] 8.446154
```

Important: We have discussed two statistics that describe the **center** of a distribution:

- Median = midpoint of the distribution (50% below and 50% above)
- Mean = average of the observations.

As the TTF data illustrate, the value of the mean can be heavily influenced by **outliers**:

- Unusually high observations will increase the mean.
- Unusually low observations will decrease the mean.
- In either case, the mean may not be an accurate representation of the center of a distribution.
- Outliers (on either side) do not affect the median greatly.

Important: The effect of outliers on the **standard deviation** is similar:

```
> sd(TTF)
[1] 17.40546
> sd(TTF.no.outlier)
[1] 6.737534
```

Removing the outlier has affected the standard deviation dramatically! It is now much smaller.

Advice: Choosing numerical descriptions that describe "center" and "spread" depends on the **shape** of the distribution.

- Symmetric distributions: Use $\overline{x}$ and s to describe center and spread, respectively. In symmetric distributions, there are generally no outliers.
- Skewed distributions: Use M and IQR to describe center and spread, respectively. Skewed distributions can contain numerous outliers. The values of M and IQR are less affected by outliers (and therefore are better representations).

13 Normal distributions

13.1 Density curves

Recall: We have examined different graphs that show the distribution of observations for a **quantitative** variable:

- Histogram, stemplot, boxplot.

We know to investigate the following physical characteristics:

- Center $\longrightarrow$ can summarize numerically using median or mean
- Spread $\longrightarrow$ this refers to the **variation** we see in the distribution; can summarize numerically using standard deviation, IQR, or the 5-number summary
- Shape $\longrightarrow$ Is the distribution symmetric? Skewed to the right or left? Bimodal?
- Outliers $\longrightarrow$ these are unusual observations that do not follow the regular pattern of the distribution; can use 1.5(IQR) rule to check.

Important: Observations we have (and display in a histogram, for example) are often a **sample** taken from a larger **population** of individuals. Therefore, the following question emerges as being relevant:

> What does the distribution I see in the sample say about the distribution in the larger population?

This question deals with **statistical inference**.

Example 13.1. A biologist is studying green sea turtles inhabiting the Grand Cayman Islands. One variable of interest to the biologist is the length of the turtle's curved shell. She catches $n = 76$ turtles and measures the length of each turtle's shell (in centimeters, cm). A histogram of these observations is shown in Figure 13.1.

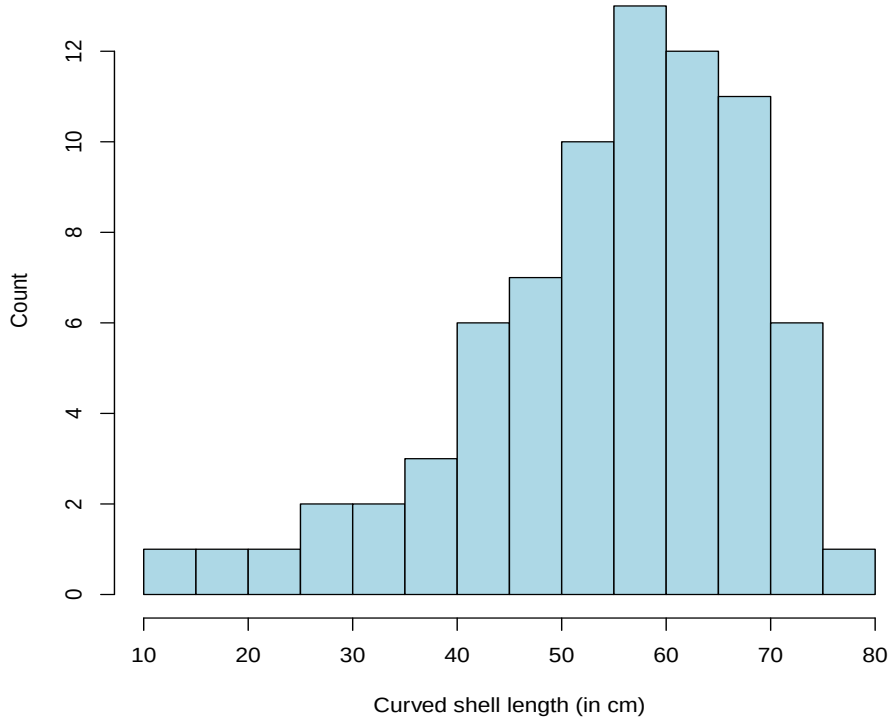


Figure 13.1: Turtle data. Shell lengths for a sample of $n = 76$ green sea turtles.

Here is the **5-number summary** for the sample:

```
> quantile(shell.length,type=2)
   0%   25%   50%   75%  100%
12.34 47.61 56.78 64.85 77.34
```

- Center $\longrightarrow$ The median shell length is 56.78 cm.
- Spread $\longrightarrow$ The IQR is $64.85 - 47.61 = 17.24$ cm. About 50% of the turtles have shell lengths between 47.61 cm and 64.85 cm.
- Shape $\longrightarrow$ The distribution is skewed to the left.
- Outliers $\longrightarrow$ The 1.5(IQR) rule declares the smallest two turtles (12.34 cm and 17.65 cm) to be outliers on the low side. Do you think this is reasonable?

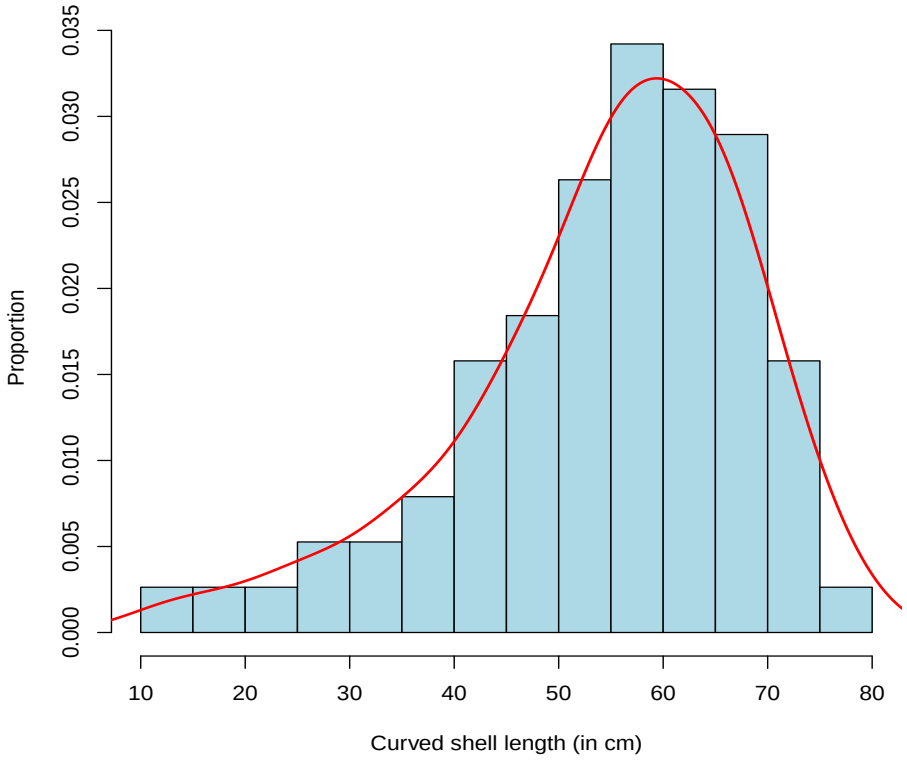


Figure 13.2: Turtle data. Shell lengths for a sample of $n = 76$ green sea turtles. An estimate of the population density curve has been added.

Remark: Sometimes the overall pattern of a histogram is so "regular" that we can describe it by a smooth curve. We call this smooth curve a **density curve**.

Important: If the sample is representative of the population, then the density curve is an approximate description of the population.

- In this light, we should call it the **population density curve** because it describes the population.
- For example, the population density curve in Figure 13.2 describes the distribution of shell lengths for all green sea turtles in the population.
- The histogram is only for the $n = 76$ turtles caught in the sample.

To reiterate this very important last point, Moore and Notz write,

> "A histogram is a plot of data obtained from a sample. The density curve is intended to reflect the idealized shape of the population distribution."

Just like histograms, **population density curves** come in all shapes:

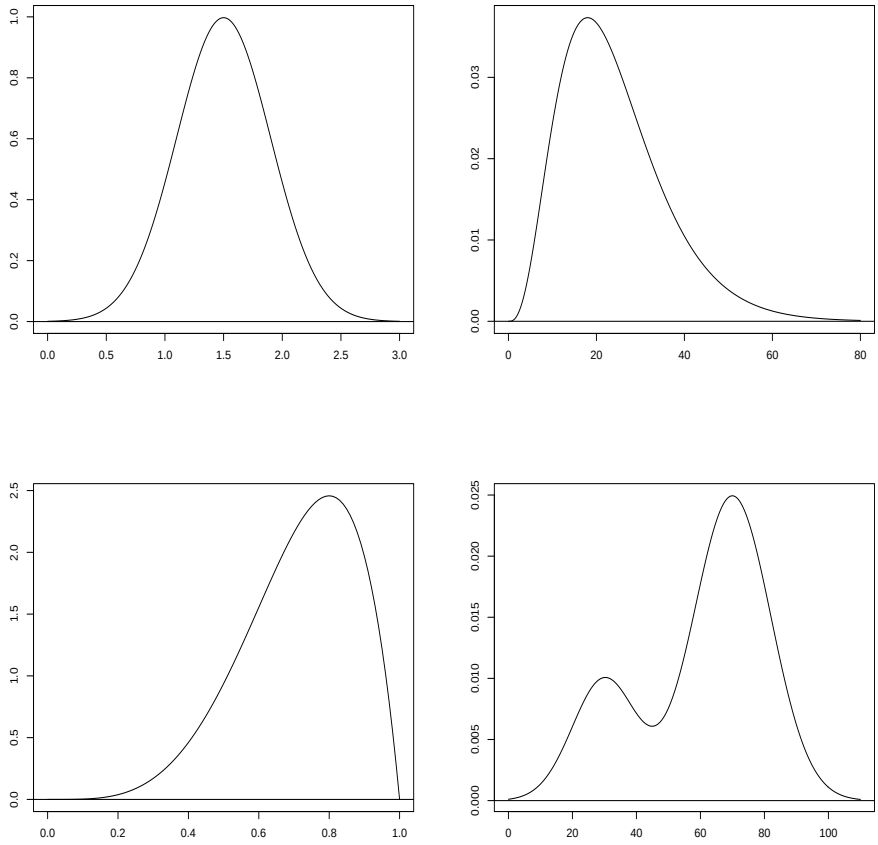


Figure 13.3: Examples of density curves.

Properties: Any population density curve has these properties:

1. the curve is non-negative.
2. the total area under the curve is 1.
3. the **area** under the curve over a given range represents the **proportion** of individuals in the population that fall in that range.

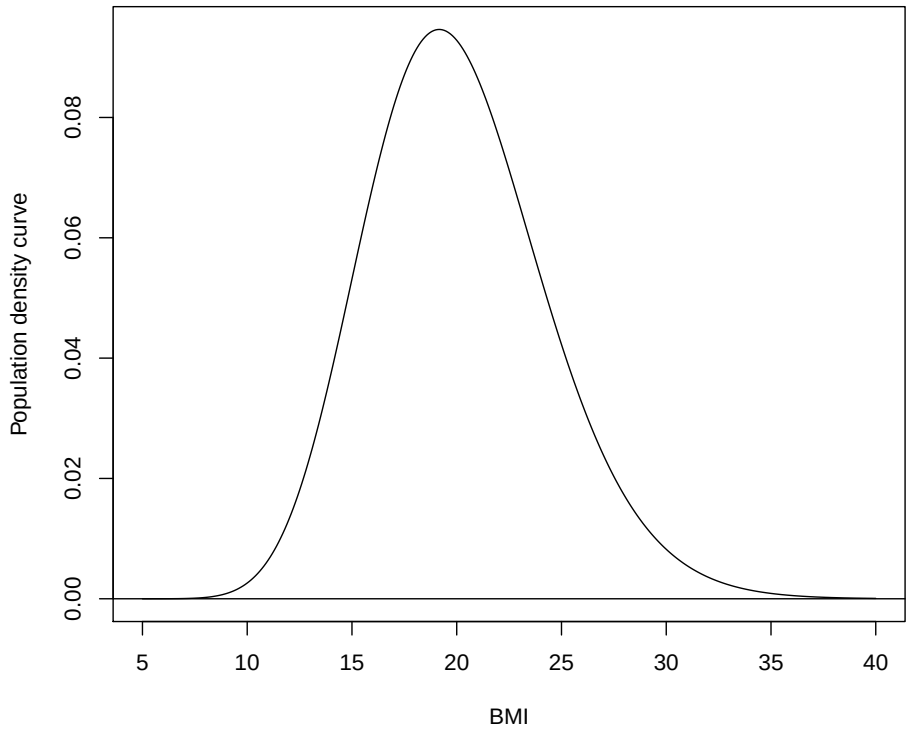


Figure 13.4: Population density curve for the BMI of fourth-grade children in Augusta, GA.

Example 13.2. Figure 13.4 shows the population density curve for the BMI of all 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:

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

Q: What proportion of the population falls in each category?

A: We can determine this by finding the **area** under the curve in the given ranges;

- i.e., ≤ 14.5, between 14.5 and 19.5, between 19.5 and 21.5, and ≥ 21.5.
- These proportions (areas) are shown in Figure 13.5.
- I used R to find these areas.

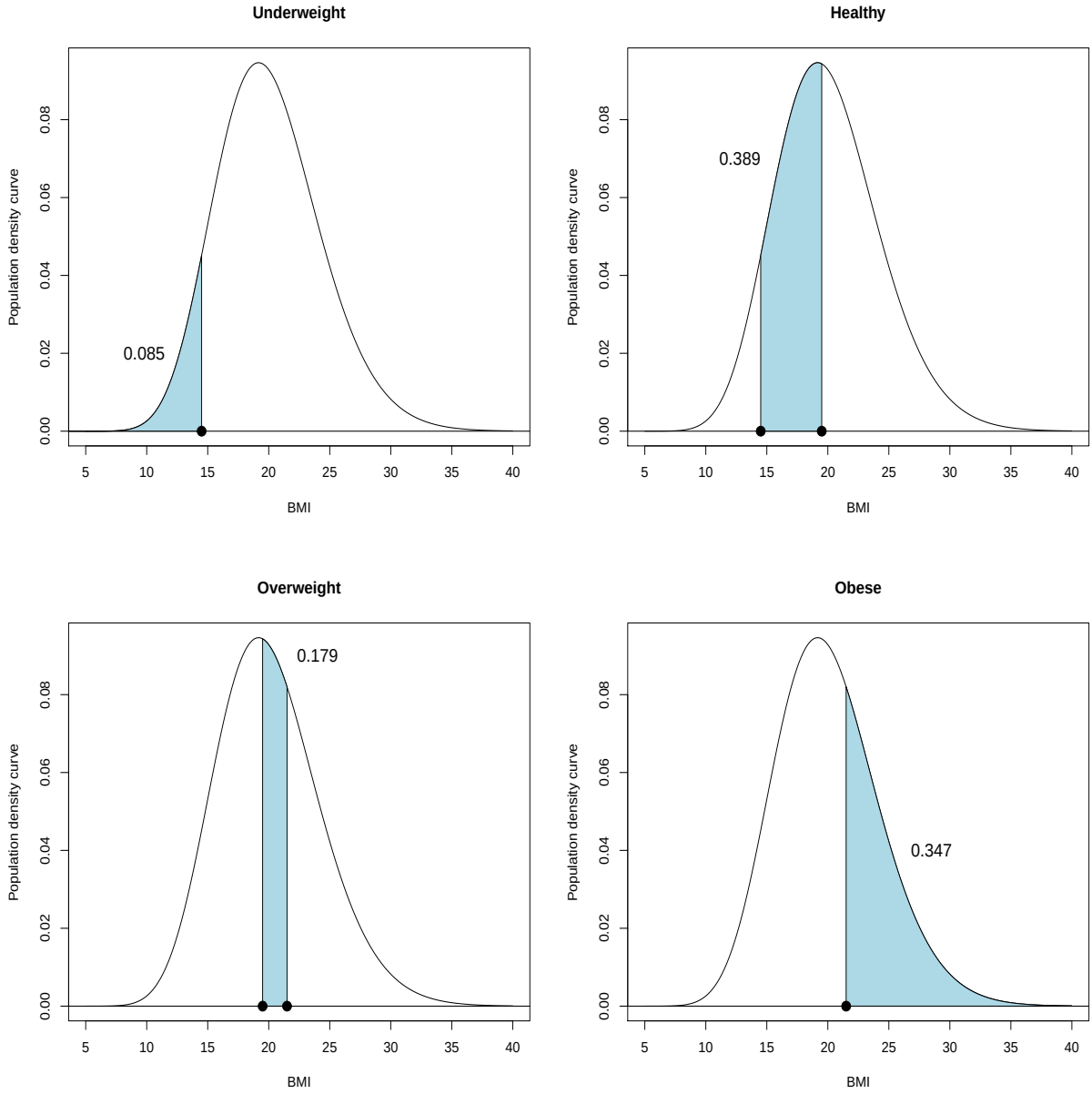


Figure 13.5: Population density curve for the BMI of fourth-grade children in Augusta, GA. Upper left: Proportion of the population who are **underweight**. Upper right: **Healthy**. Lower left: **Overweight**. Lower right: **Obese**. In each subfigure, the proportion equals the area of the shaded region. Note that the areas add to 1.

13.2 Center and spread for a population density curve

Note: The notions of "center" and "spread" 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

Important: The mean and standard deviation of the sample ($\overline{x}$ and s) are **statistics**. The mean and standard deviation of the population (μ and σ) are **parameters**.

- The mean and standard deviation (μ and σ) are important population parameters for the **normal distributions**, which we will start discussing shortly.

Recall: The median and mean are both measures of "center." Recall the difference:

- Median = midpoint of the distribution; the point that divides the area under the population density curve in half (**50% below and 50% above**)
- Mean = average value; balance point of the distribution.

Q: For population density curves, how do the median and mean compare?
A: It depends on the **shape** of the distribution. If the population density curve is

- **symmetric** $\longrightarrow$ median and mean are equal

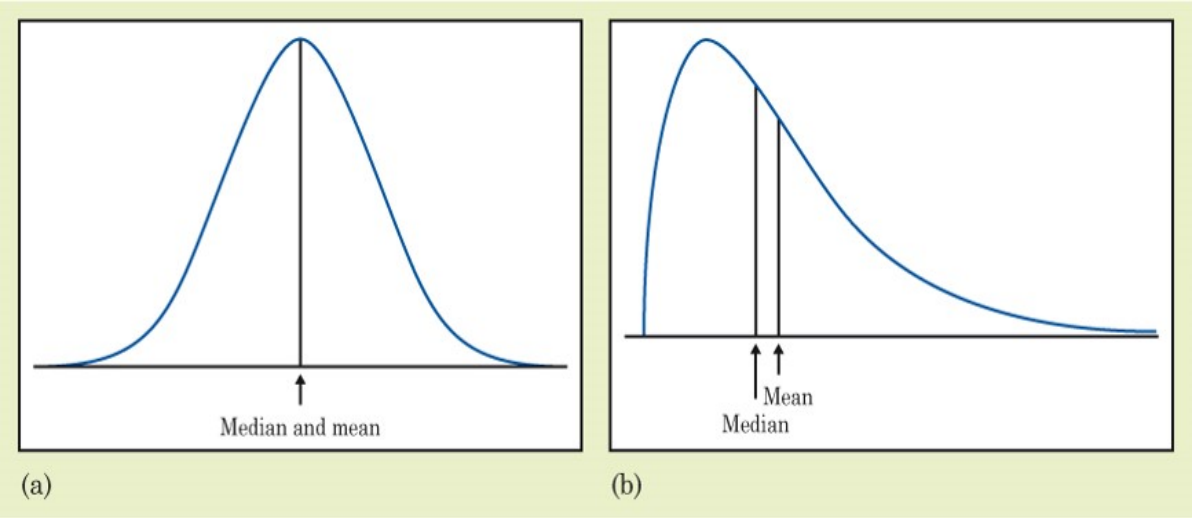


Figure 13.6: Two population density curves. Left: Symmetric; median = mean. Right: Skewed right; median < mean.

- **skewed right** $\longrightarrow$ median is less than the mean (see Figure 13.6)
- **skewed left** $\longrightarrow$ median is greater than the mean.

Recall: The mean is influenced by unusually high or low observations (the median is not). Therefore, the mean will move in the direction of these observations.

13.3 Normal distributions

Note: The most famous (and important) population density curve is the **normal distribution**. A normal distribution is characterized by 2 pieces of information:

- the mean μ
- the standard deviation σ.

The mean measures where the **center** is. The standard deviation measures how "spread out" the distribution is. All normal distributions are symmetric and unimodal.

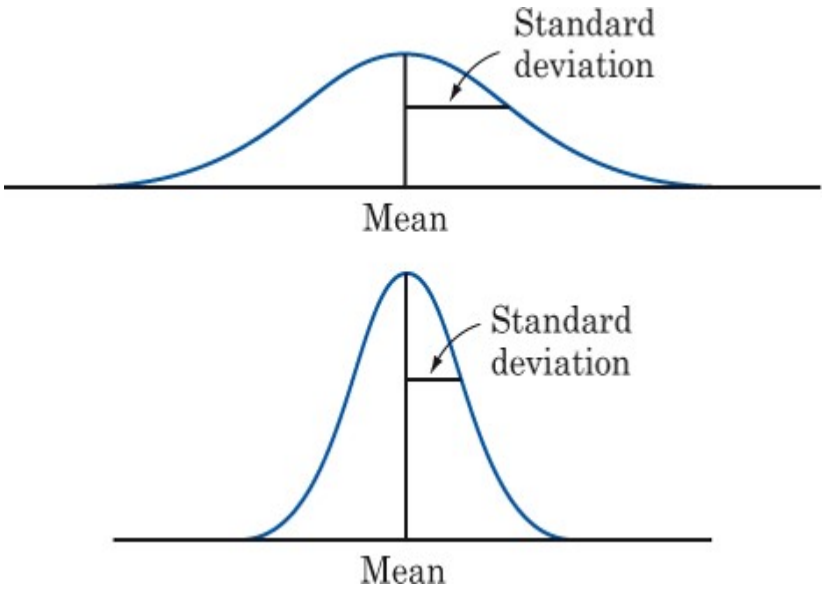


Figure 13.7: Two normal distributions. Same means, different standard deviations.

13.3.1 68-95-99.7 rule

Important: 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 deviation of the mean
- 99.7% (or almost all) of the observations will be within **three** standard deviations of the mean.

This is known as the **68-95-99.7 rule**. An illustrative display is given in Figure 15.6.

Note: In any normal distribution, it would be **very unusual** for an observation to be farther than 3 standard deviations away from the mean (i.e., either 3 standard deviations below the mean or 3 standard deviations above).

- By the **68-95-99.7 rule**, we know that only about 0.3% of the observations will do so (that is, 3 out of 1000).

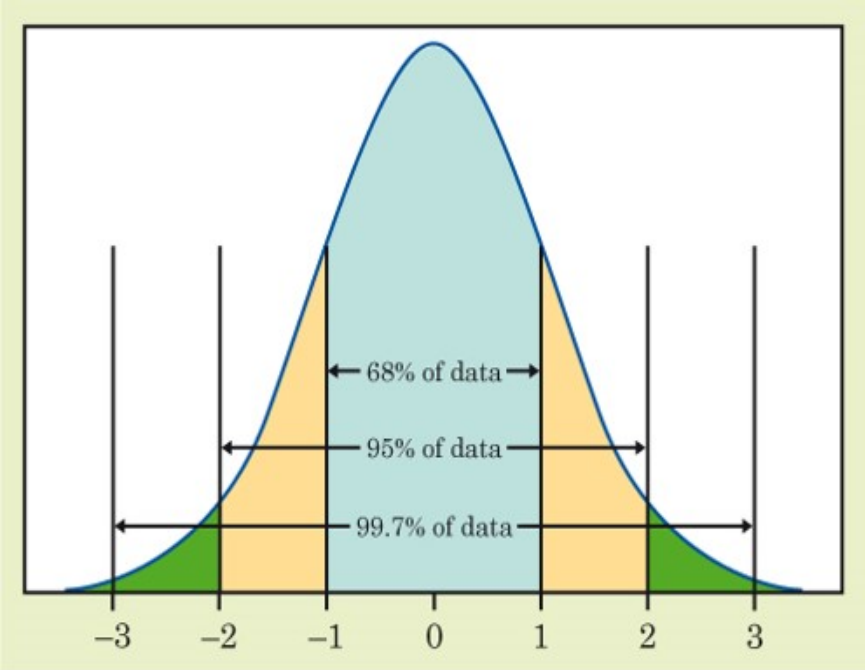


Figure 13.8: **68-95-99.7 rule:** Illustration using a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

Example 13.3. The World Health Organization uses a normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 15$ to describe the systolic blood pressure (SBP) of American males (aged 18 and over). SBP is measured in millimeters of mercury (mmHg). This population density curve is shown in Figure 13.9.

Questions:

(a) Form intervals 1, 2, and 3 standard deviations from the mean. Interpret.

(b) What percentage of the American male population has a SBP of 140 mmHg or higher? (**Note:** This is regarded as "high" blood pressure).

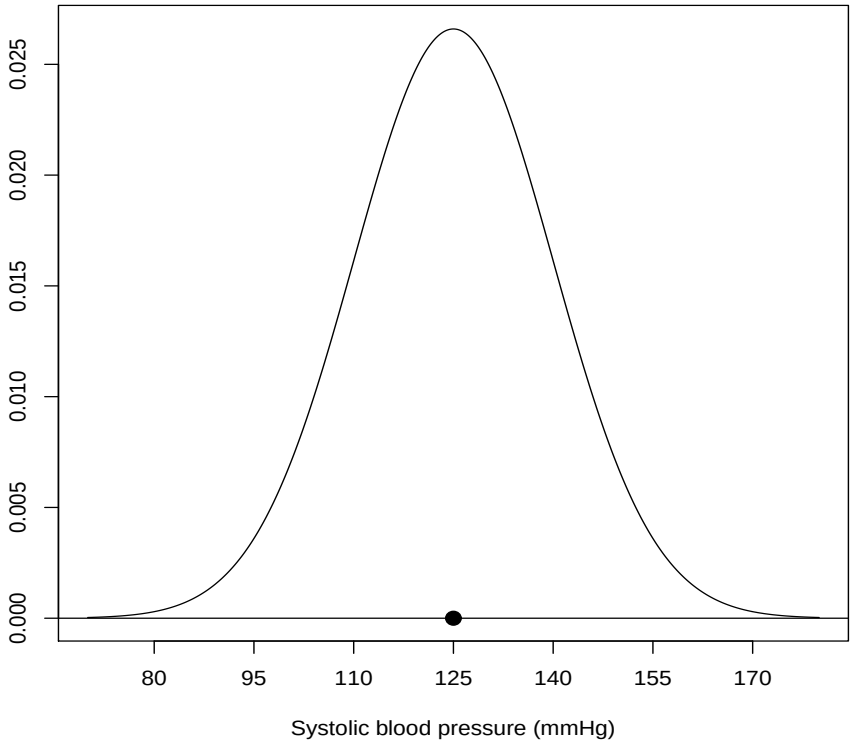


Figure 13.9: Population density curve for the systolic blood pressure of American males. A solid circle identifies the population mean $\mu = 125$ mmHg. The population standard deviation is $\sigma = 15$ mmHg.

(c) What percentage of the American male population has a SBP of 95 mmHg or lower?

(d) What percentage of the American male population has a SBP between 95 mmHg and 140 mmHg?

13.3.2 Standard scores

Definition: The **standard score** of an 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 13.4. 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.

- Jeannie scored a 650 on the SAT mathematics exam.
- Gerald scored a 21 on the ACT mathematics exam.

Who did better? We can answer this by calculating each student's standard score.

Jeannie's standard score is

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{650 - 500}{100} = 1.5.$$

Gerald's standard score is

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{21 - 18}{6} = 0.5.$$

Jeannie did better. Her standard score is larger. See Figure 13.10.

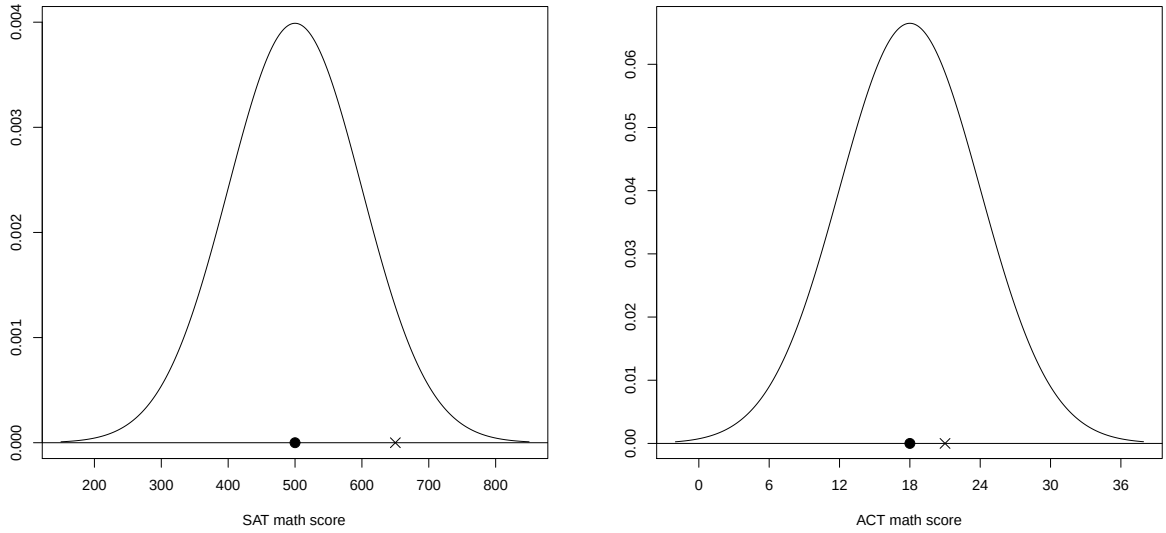


Figure 13.10: SAT and ACT mathematics score distributions. Left: SAT. Right: ACT. In each distribution, the population mean is identified with a solid circle. The symbol "×" is used to identify student's exam scores.

Question: Jeannie scored better than what percentage of the population (who took the SAT)? What about Gerald for the ACT?

- To answer these questions, we need to find the **area** under each curve (to the left of each "×" mark in Figure 13.10).

13.3.3 Calculating areas under the normal curve (using R)

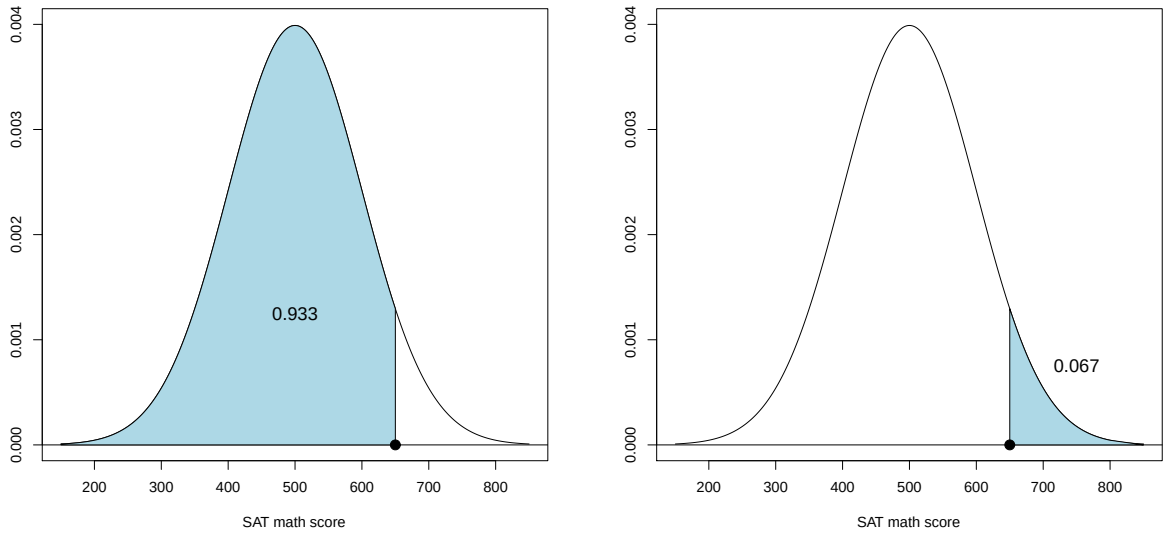
Important: We can use R to find the area under any normal curve.

- 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.

Example 13.4 (continued). SAT mathematics scores follow a normal distribution with mean 500 and standard deviation 100.

- Jeannie scored a 650 on the SAT. Her standard score is $z = 1.5$.

```
> pnorm(1.5) # area to the left
[1] 0.9331928
> 1-pnorm(1.5) # area to the right
[1] 0.0668072
```



Interpretation: Jeannie scored better than about 93.3% of the population on the SAT mathematics exam. About 6.7% of the population did better than her.

Exercise: Repeat this calculation for Gerald (who made a 21 on the ACT).

Example 13.5. For a population of drivers (e.g., all drivers in SC, etc.), suppose the reaction time to brake during in-traffic driving follows a normal distribution with mean $\mu = 1.5$ seconds and standard deviation $\sigma = 0.4$ seconds.

(a) What percentage of the population has a reaction time longer than 2 seconds?

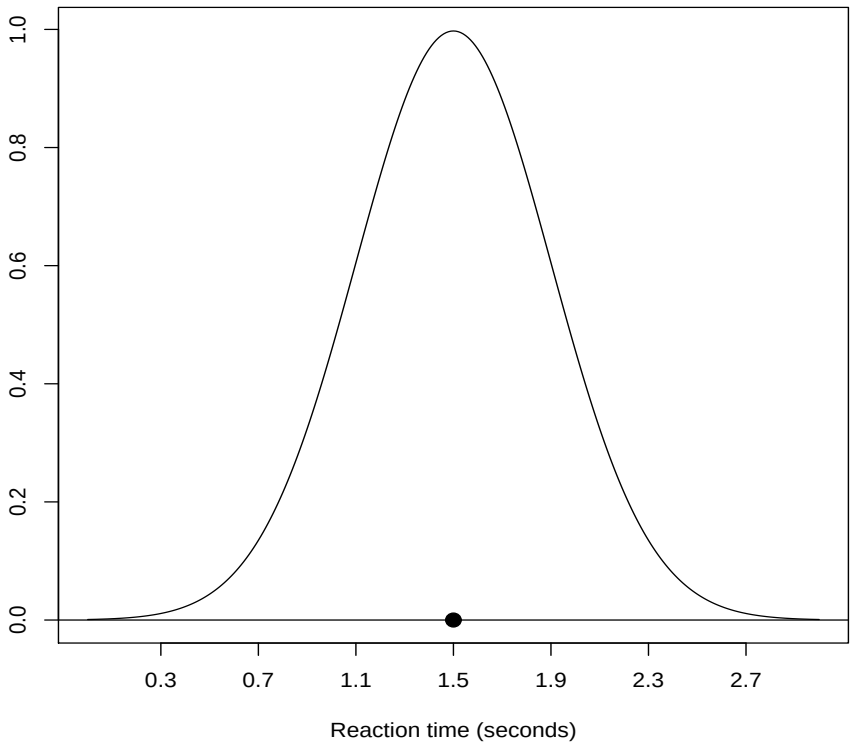


Figure 13.11: Population density curve for reaction time. A solid circle identifies the population mean $\mu = 1.5$ seconds. The population standard deviation is $\sigma = 0.4$ seconds.

(b) What percentage of the population has a reaction time **between** 1 and 2 seconds?

(c) The slowest 2.5% of the population has a reaction time of at least how long? *Hint*: You can use the 68-95-99.7 rule here.

14 Describing Relationships: Scatterplots and Correlation

14.1 Introduction

Remark: By the phrase "describing relationships" in statistics, we mean that we are interested in how two (or more) variables relate to each other.

- Most observational studies and experiments record observations on multiple variables (not just one). It therefore makes sense to think about how variables might be related.

Example 14.1. As part of a waste removal project, a new compression machine for processing sewage sludge was studied. Engineers were interested in the two variables:

$$x = \text{machine filtration rate (measured in kg/m/hr)}$$
$$y = \text{moisture of compressed pellets (measured as a \%)}.$$

Engineers collected observations from a sample of $n = 20$ individual sewage specimens. These observations are recorded below:

Specimen	x	y	Specimen	x	y
1	125.3	77.9	11	159.5	79.9
2	98.2	76.8	12	145.8	79.0
3	201.4	81.5	13	75.1	76.7
4	147.3	79.8	14	151.4	78.2
5	145.9	78.2	15	144.2	79.5
6	124.7	78.3	16	125.0	78.1
7	112.2	77.5	17	198.8	81.5
8	120.2	77.0	18	132.5	77.0
9	161.2	80.1	19	159.6	79.0
10	178.9	80.2	20	110.7	78.6

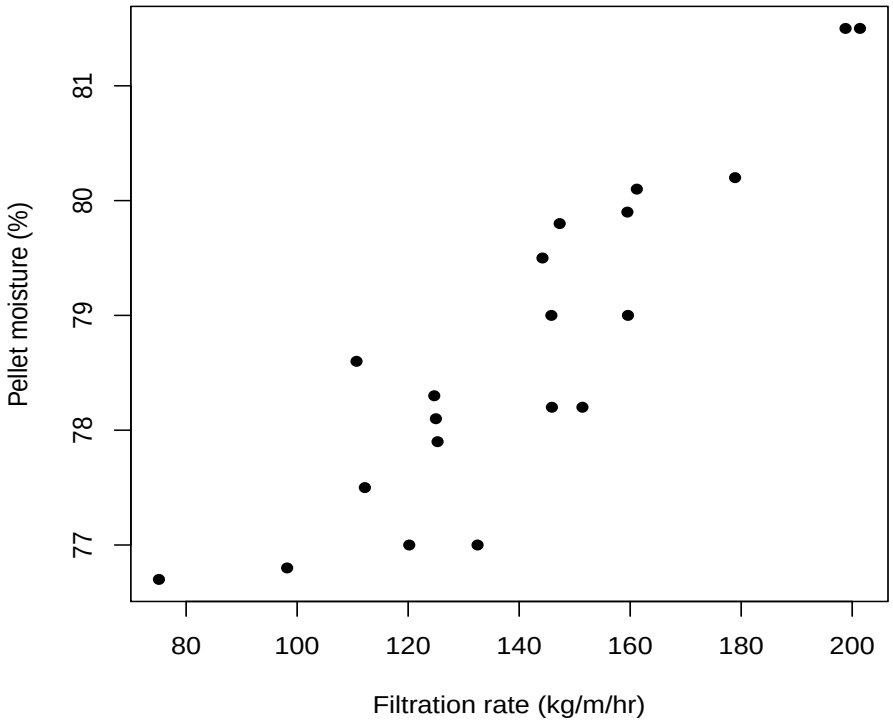


Figure 14.1: Scatterplot of filtration rate (x, measured in kg/m/hr) and pellet moisture (y, measured as a %) for a sample of $n = 20$ sewage specimens.

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.
- Figure 14.1 shows the scatterplot between filtration rate (x) and pellet moisture (y) in Example 14.1. This graph shows a **positive linear relationship** between the two variables.

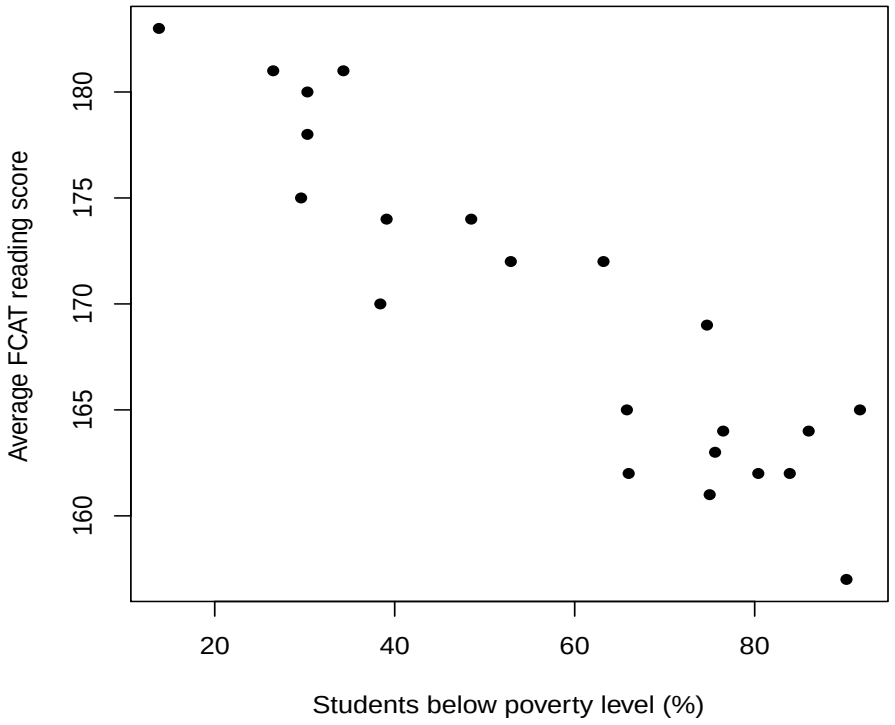


Figure 14.2: Scatterplot of the percentage of students below the poverty level (x) and average FCAT reading score (y) for a simple random sample of $n = 22$ Florida elementary schools.

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 simple random sample of $n = 22$ Florida elementary schools. A scatterplot of these observations is shown in Figure 14.2 (the data are available online). This graph shows a **negative linear relationship** between the two variables.

14.2 Interpreting scatterplots

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**, then the scatterplot presents an impression of the underlying relationship for 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.

Interpretation: We will focus on the following characteristics when we examine and describe scatterplots:

1. Overall pattern:
   - Form: Are there straight-line (linear) patterns or curved patterns? Do observations tend to fall into clusters?
   - Direction: Are the variables positively related or negatively related?
   - Strength: Is the relationship strong, moderate, or mild? Perhaps there is no relationship (e.g., a random scatter of points)?
2. 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: Filtration rate (x) and pellet moisture (y) are positively related.
- Example 14.2: The percentage of students below the poverty level (x) and average FCAT reading score (y) are negatively related.

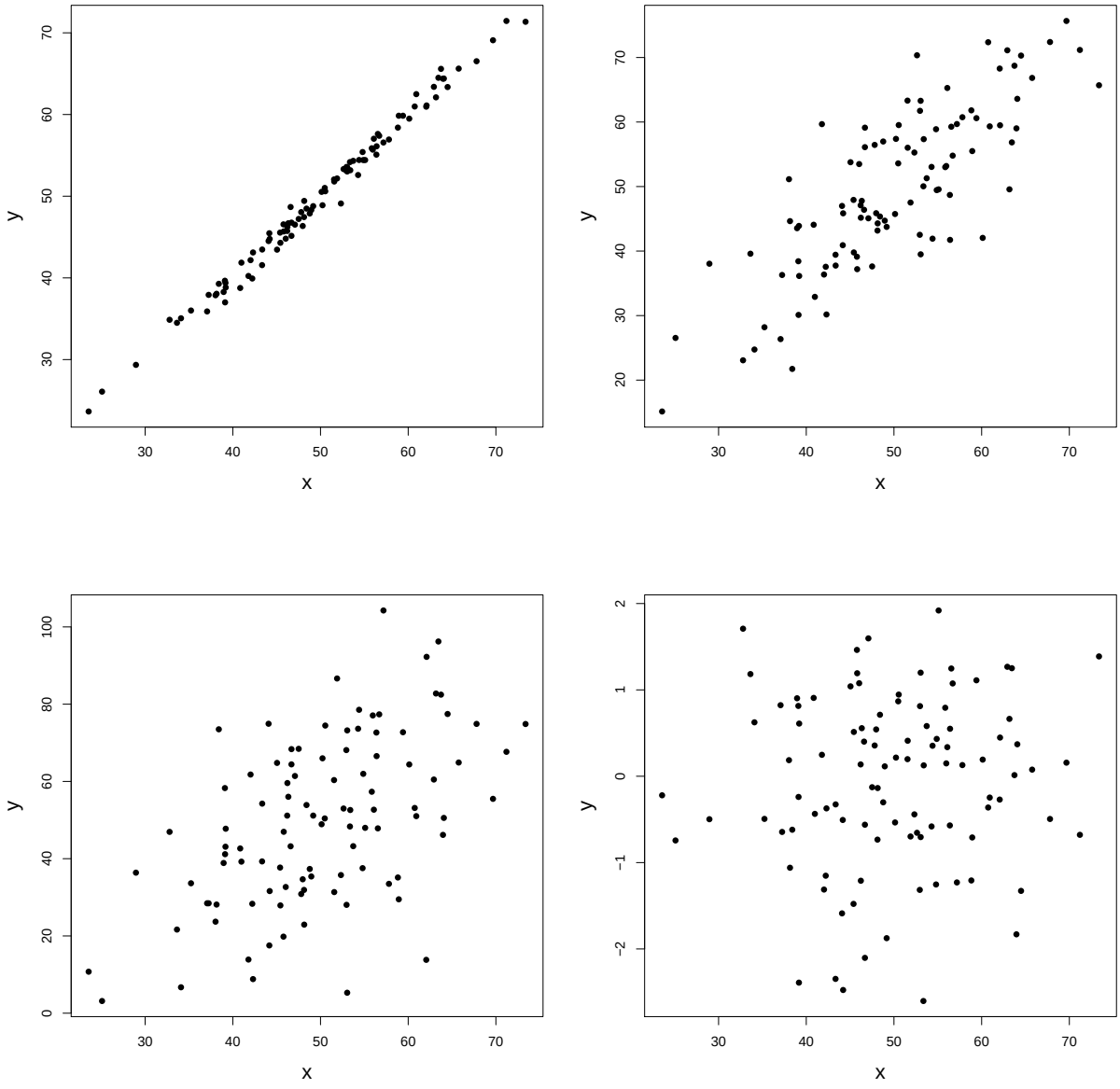


Figure 14.3: Scatterplot examples. **Upper left:** Very strong, positive linear relationship. **Upper right:** Moderate-to-strong, positive linear relationship. **Lower left:** Mild, positive linear relationship. **Lower right:** No linear relationship (random scatter).

14.3 Correlation

Goal: We would like to study the relationship between two quantitative variables, x and y. Scatterplots give us a visual display of the relationship. We now wish to summarize this relationship **numerically**.

Definition: The **correlation** is a numerical summary that describes the strength and direction of the straight-line (linear) relationship between two quantitative variables.

- The correlation is denoted by r.

Formula: With a sample of n individuals, the correlation is computed by the following formula:

$$r = \frac{1}{n-1} \sum \left(\frac{x - \overline{x}}{s_x} \right) \left(\frac{y - \overline{y}}{s_y} \right),$$

where $\overline{x}$ and $\overline{y}$ are the sample means and s_x and s_y are the sample standard deviations. Note that the terms

$$\frac{x - \overline{x}}{s_x} \quad \text{and} \quad \frac{y - \overline{y}}{s_y}$$

are the **sample standardized values** of x and y, respectively.

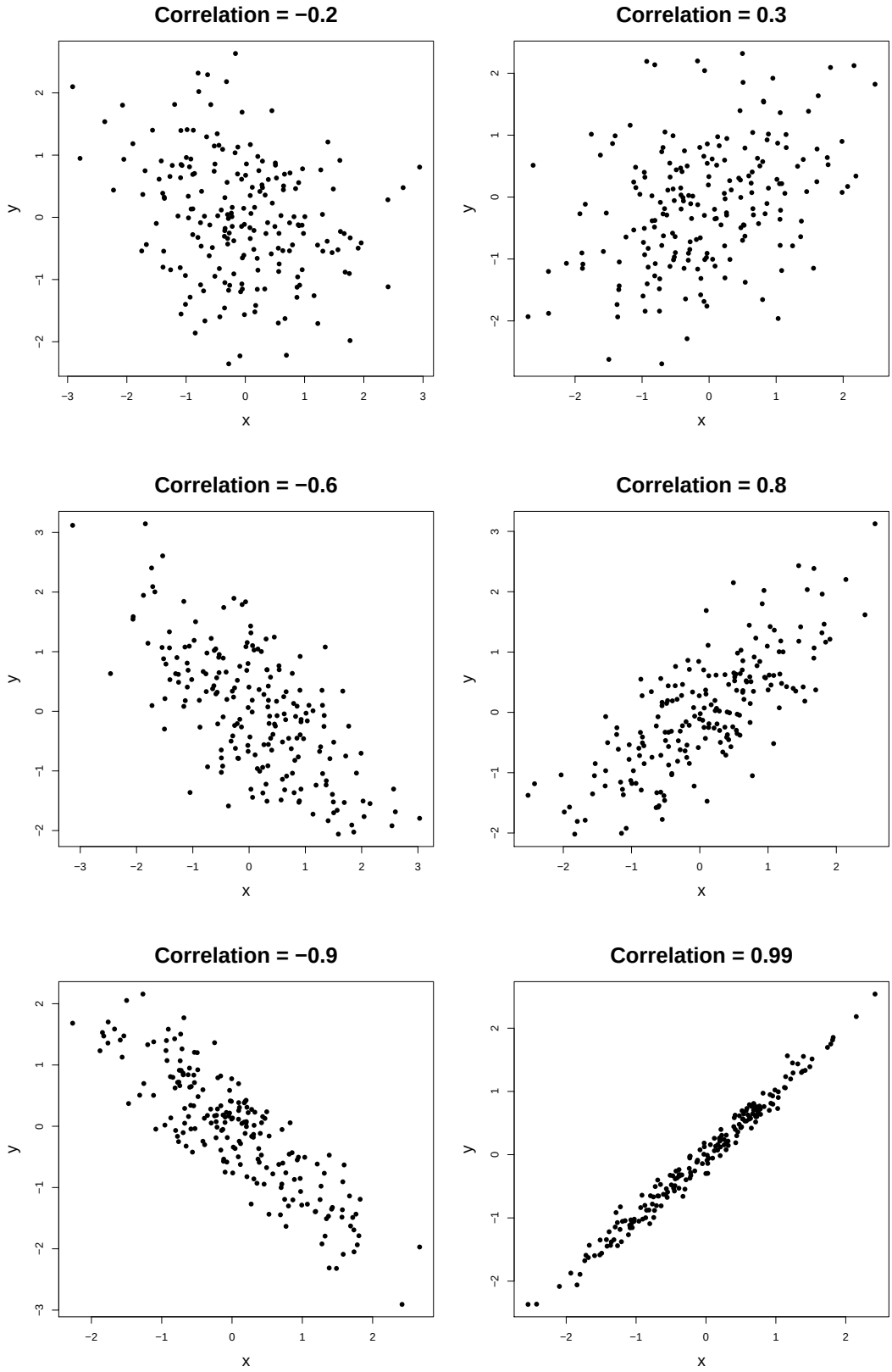
Important: We will use R to compute the correlation with real data. It will be our job to understand what its value means.

Discussion: Let's discuss the important facts about the correlation r.

Fact 1: If x and y have a positive linear relationship, then $r > 0$. If x and y have a negative linear relationship, then $r < 0$.

- **Note:** $r = 0$ means that x and y have no linear relationship.

See examples on the next page. Left column: $r < 0$. Right column: $r > 0$.


Correlation = −0.2
Correlation = 0.3
Correlation = −0.6
Correlation = 0.8
Correlation = −0.9
Correlation = 0.99
x
y

Fact 2: The correlation r is always between -1 and 1; i.e.,

$$-1 \leq r \leq 1.$$

What happens at the endpoints?

- 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**.
  - **Remark:** Perfect relationships are a rarity with real life data; e.g., see the scatterplots in Examples 14.1 and 14.2. There are almost always other sources of variation that make a relationship not perfect (e.g., lurking variables, etc.).

Illustration: Let's use R to calculate the correlation r for the data in Examples 14.1 and 14.2.

- **Example 14.1:**

```
> cor(filtration.rate,moisture)
[1] 0.8943937
```

The correlation between filtration rate (x) and pellet moisture (y) is $r \approx 0.89$. This represents a very strong, positive linear relationship.

- **Example 14.2:**

```
> cor(poverty,FCAT.reading)
[1] -0.9169302
```

The correlation between the percentage of students below the poverty level (x) and the average FCAT reading score (y) is $r \approx -0.92$. This represents a very strong, negative linear relationship.

Fact 3: The correlation r is **unitless**; i.e., there are no units attached to it (e.g., dollars, cm, etc.).

- This also means that you could change the units of your data (e.g., inches to cm, percentages to proportions, etc.), and this would not change the value of r.

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 which variable is the explanatory variable x and which one is the response variable y.
- To illustrate this, consider the data in Example 14.2 and calculate the correlation in both ways:

```
> cor(poverty,FCAT.reading) # correlation of x and y
[1] -0.9169302
> cor(FCAT.reading,poverty) # correlation of y and x
[1] -0.9169302
```

We get the same answer.

Fact 5: The correlation r only measures the strength and the direction of **straight-line (linear) relationships**.

- The correlation does not describe a curved relationship, no matter how strong that relationship is.

Example 14.3. A study was conducted to determine if there was a relationship between

$$x = \text{arterial oxygen tension (measured in mmHg)}$$
$$y = \text{cerebral blood flow}$$

in humans. Sixteen adults participated in the study and the variables x and y were measured on each individual. A scatterplot of the data is shown in Figure 14.4.

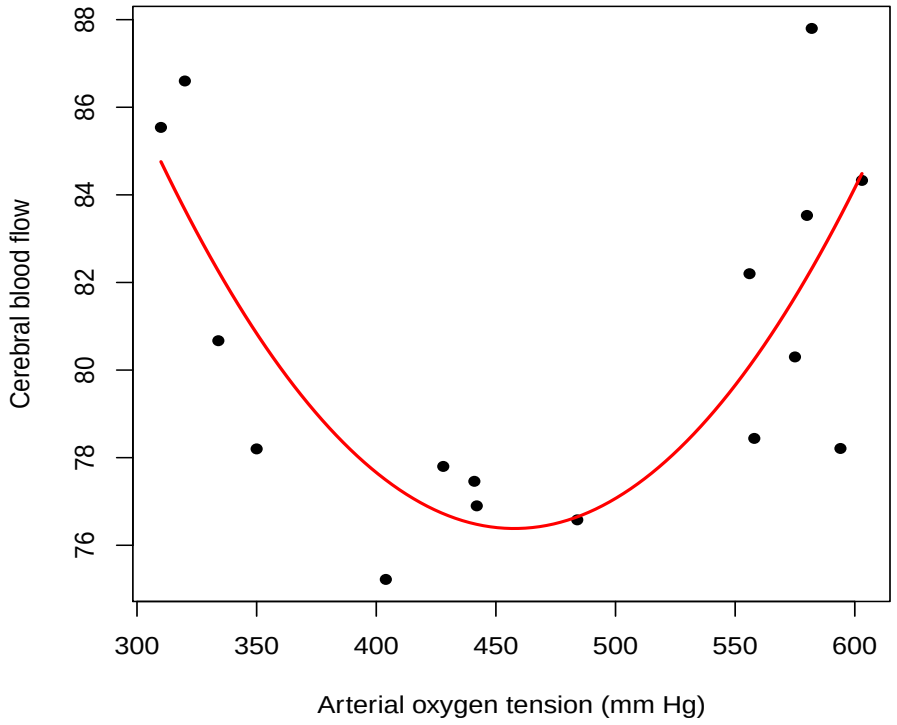


Figure 14.4: Scatterplot of arterial oxygen tension (x, measured in mmHg) and cerebral blood flow (y) for a sample of $n = 16$ adults. A quadratic curve has been added to emphasize the relationship.

Discussion: Clearly, there is a relationship between arterial oxygen tension (x) and cerebral blood flow (y). However, the relationship is not a straight-line relationship. It is better described as a **curved** (i.e., quadratic) relationship.

- Because the relationship in Figure 14.4 is not a linear one, the correlation r here is useless. It does not describe curved relationships.

```
> cor(oxygen.tension,blood.flow)
[1] 0.0703052
```

The correlation $r \approx 0.07$, which is very close to 0. However, the (curved) relationship between x and y here is pretty strong.

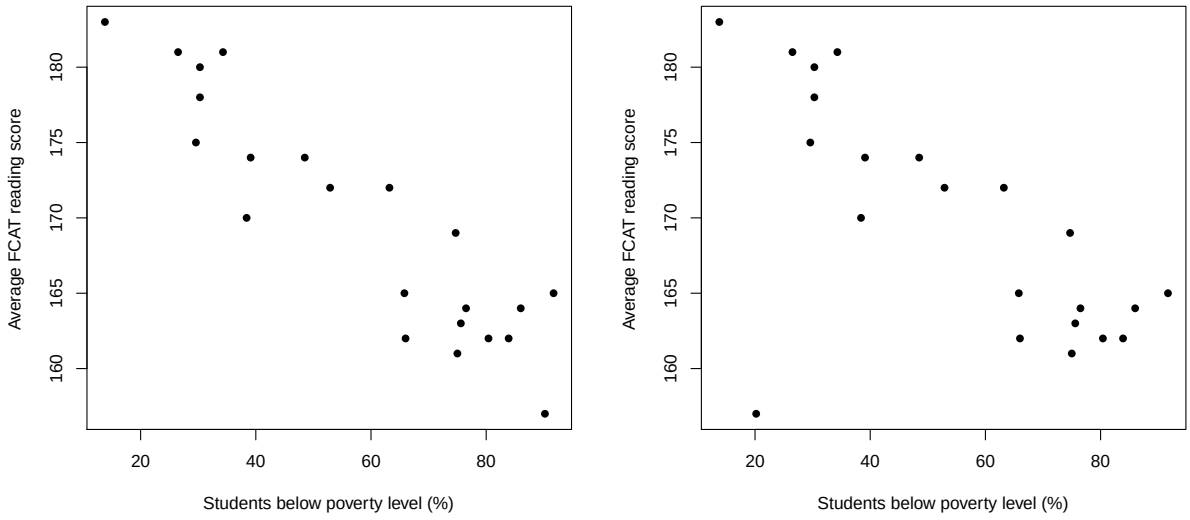


Figure 14.5: Left: Scatterplot of the percentage of students below the poverty level (x) and average FCAT reading score (y) for a simple random sample of $n = 22$ Florida elementary schools. Right: One observation changed to now be an obvious outlier.

Fact 6: The value of the correlation r can be highly affected by **outliers**.

- Just one or two outliers might completely change one's impression of the strength of the relationship.
- In Example 14.2, suppose we changed just one observation in Figure 14.5 (it's easy to see which one it is). Note what this does to the value of r:

```
> cor(poverty,FCAT.reading) # original data
[1] -0.9169302
> cor(poverty.outlier,FCAT.reading) # data with outlier created
[1] -0.6849228
```

  The value of the correlation has changed from $r \approx -0.92$ to $r \approx -0.68$ after altering only one observation!

Moral: Always plot your data first!

15 Describing Relationships: Regression, Prediction, and Causation

15.1 Introduction

Remark: A problem that arises in medicine, engineering, the social sciences, and other areas, is describing the relationship between two quantitative variables. In the last chapter, we learned the correlation r is a numerical summary of straight-line relationships. In this chapter, we would like to quantify straight-line relationships using a **regression model**.

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 in the elderly). An observational study investigated the relationship between

$$x = \text{age (in years)}$$
$$y = \text{systolic blood pressure (SBP, measured in mmHg)}$$

in adult males. There were $n = 30$ individuals in the study. The data are shown below:

Ind	x	y	Ind	x	y	Ind	x	y
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.

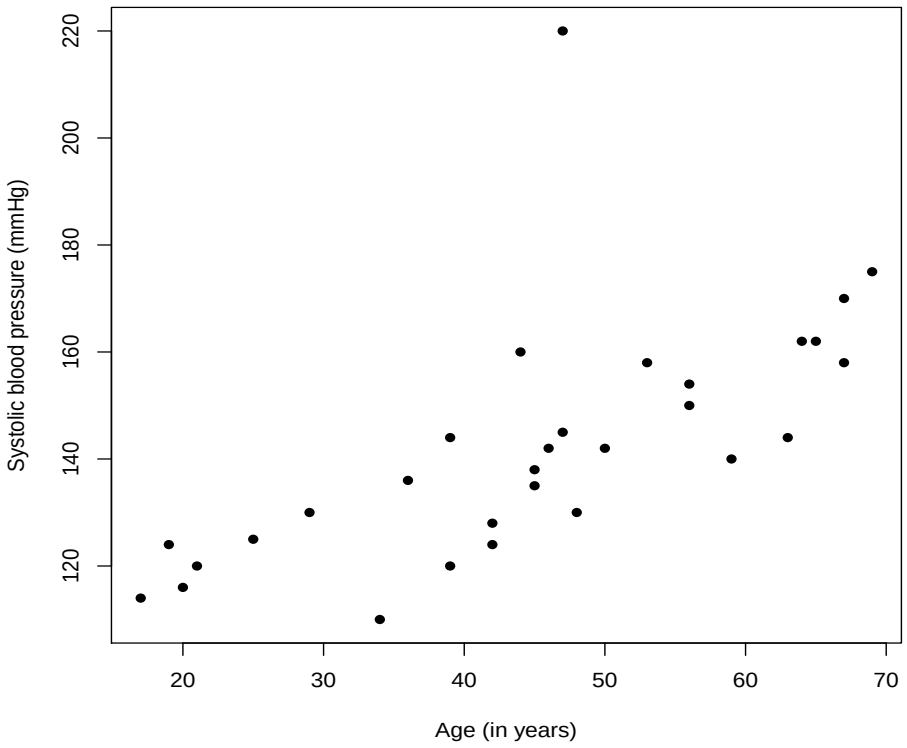


Figure 15.1: Scatterplot of age (x, measured in years) and SBP (y, measured in mmHg) for a sample of $n = 30$ adult males.

Discussion: When interpreting scatterplots, we know to investigate the following physical characteristics:

- Form $\longrightarrow$ there is a clear straight-line (linear) relationship between age and SBP.
- Direction $\longrightarrow$ age and SBP are positively related.
- Strength $\longrightarrow$ the straight-line relationship is moderately strong; the correlation is $r \approx 0.66$.
- Outliers $\longrightarrow$ there is an obvious outlier (i.e., the 47 year-old whose SBP is 220).

```
> cor(age,SBP)
[1] 0.6575673
```

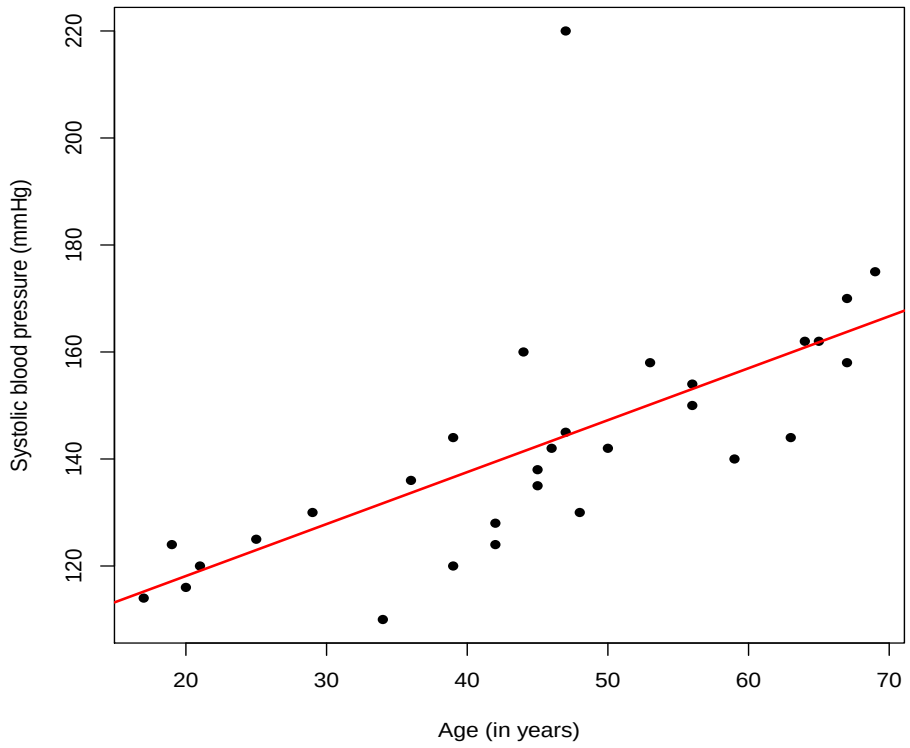


Figure 15.2: Scatterplot of age (x, measured in years) and SBP (y, measured in mmHg) for a sample of $n = 30$ adult males. The least-squares regression line has been added.

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.
- For example, in Example 15.1, what would we predict the SBP to be for a male who is 43 years old?
- The **least-squares regression line** has been added to the scatterplot in Figure 15.2.
  - Where does this line come from?
  - How can we use this line for prediction?

15.2 Regression equations and prediction

Recall: Suppose that y is the response variable (on the vertical axis) and that x is the 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 y-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 (or decreases) by b units.
- The **y-intercept** a is the value of the response y when $x = 0$.

Q: With real data (like those in Example 15.1), we will never be able to find a line that "fits perfectly;" i.e., a line that goes through all the points. So, how do we find the equation of the regression line?
A: We find the equation by using **least squares**.

Definition: The **least-squares regression line** is the line that minimizes the sum of squared vertical distances from the data points to the line.

- We call this the line that "best fits" the data.
- By "best," we mean informally that the average squared vertical distance (from the points to the line) is as small as possible.
- There is only one line that meets this criterion, and we will use R to find it.
- This line can be used for **prediction**, as we will see shortly.

R: We use R to find the equation of the least-squares regression line in Example 15.1:

```
> fit = lm(SBP~age)
> fit
Coefficients:
(Intercept)          age
    98.7147       0.9709
```

Interpretation: This output gives the following values:

$$a \approx 98.7$$
$$b \approx 0.97.$$

The **equation** of the least-squares regression line is

$$y = 98.7 + 0.97x,$$

or, in other words,

$$\text{SBP} = 98.7 + 0.97 \text{ age}.$$

This line is shown in Figure 15.2.

Interpretation:

- The slope $b \approx 0.97$ is interpreted as follows:

  > "For a one-year increase in age, we would expect the SBP to **increase** by 0.97 mmHg."

- The y-intercept $a = 98.7$ is interpreted as follows:

  > "For a person whose age is $x = 0$, we would expect the SBP to be 98.7 mmHg."

Remark: In most regression problems (including this one), the **slope** is the most important value to interpret. Its value describes how much we can expect the response to increase (or decrease) when the explanatory variable increases by one unit.

On the other hand, the **y-intercept** a (i.e., the value of y when $x = 0$) is usually less meaningful.

- The notion of $x = 0$ refers to a person (in this example) whose age is "0;" an individual who has just been born.
- 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 mmHg based on the data from this study.
- Why? The value $x = 0$ is well outside the range of the ages of the other individuals in the study (ages 17-69). Therefore, saying that a newborn's SBP is 98.7 mmHg based on this study is an **extrapolation**.

Prediction: I am 43 years old. Based on the results from this study, what would you predict my SBP to be?
Answer: First, note that making this prediction makes sense (i.e., the age "43" falls "right in the middle" of the ages used in the study; see Figure 15.2). This is now easy to calculate; simply insert the age "43" into the regression equation:

$$\text{SBP} = 98.7 + 0.97(43) \approx 140.4.$$

We would predict my SBP to be 140.4 mmHg.

Remarks:

- Predictions are best when the regression line "fits" the data well.
  - If a straight line is not adequate for the data, then predictions may be biased.
- **Extrapolation** occurs when we make a prediction for y using a value of x outside the range of the x values used in the study/experiment.
  - In order for an extrapolated prediction to be accurate, the straight-line relationship must hold for x values outside the range where we have observed data. Unfortunately, this may be difficult or impossible to assess.

15.3 Correlation and regression

Note: 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., $r = 0.95$, $r = -0.20$, etc.
- A regression line is a mathematical equation that describes the relationship.

Both correlation and regression can be affected by **outliers**, as shown in Figure 15.3.

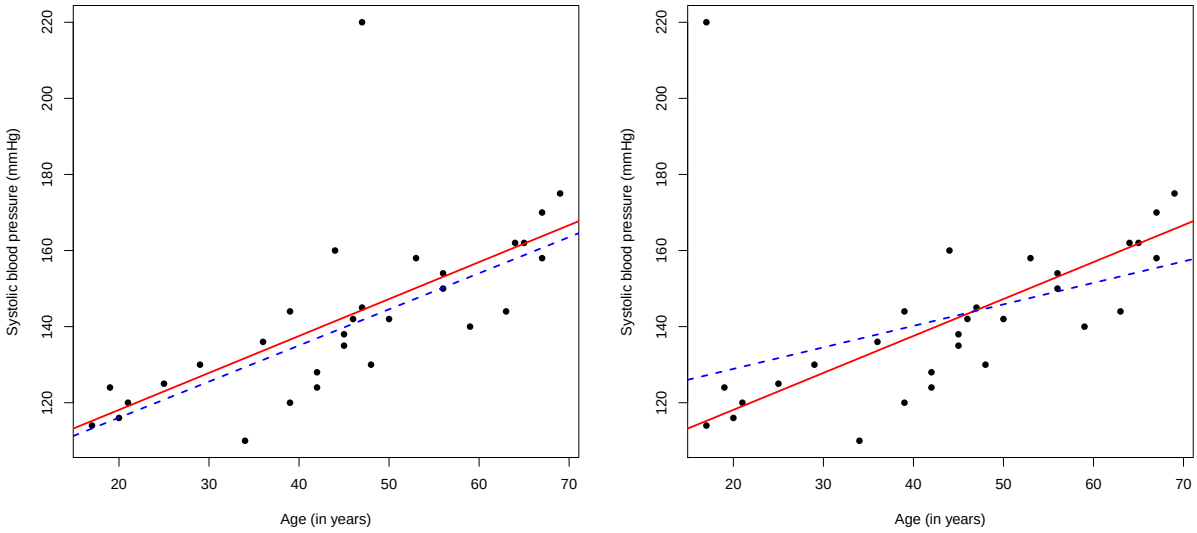


Figure 15.3: Scatterplot of age (x, measured in years) and SBP (y, measured in mmHg) for a sample of $n = 30$ adult males. **Left:** Least-squares regression line with (red, solid) and without (blue, dashed) the original outlier. **Right:** Least-squares regression line with the outlier changed.

Original outlier: Age = 47; SBP = 220. Observation removed (left).
Original outlier: Age = 47; SBP = 220. Age changed to 17 (right).

Main point: Outliers have the potential to greatly alter the equation of the least-squares regression line! Always plot your data first!!

Discussion: The usefulness of the regression line for prediction depends on the strength of the straight-line relationship between the two variables.

- The stronger the straight-line relationship, the better the predictions.
- The correlation r measures the strength of this relationship!

Definition: In a regression analysis, one way to measure how well a straight line fits the data is to compute the **square of the correlation** r^2.

- The value of r^2 gives the proportion of the total variation in the y data explained by the straight-line relationship with the explanatory variable x.

Example 15.1 (continued). For the observational study in Example 15.1, investigators measured

$$\begin{aligned} x &= \text{age (in years)} \\ y &= \text{systolic blood pressure (SBP, measured in mmHg)} \end{aligned}$$

for a sample of 30 adult males. The correlation between the two variables in the sample is $r \approx 0.66$.

```
> cor(age,SBP)
[1] 0.6575673
```

The square of the correlation is

$$r^2 \approx (0.66)^2 \approx 0.44.$$

Interpretation: Approximately 44% of the variability in the SBP data is explained by the straight-line relationship with age (x).

- This means that approximately 56% of the variability in the SBP data is explained by other variables (e.g., family history/genetics, exercise frequency, diet, smoking habits, overall health, etc.).

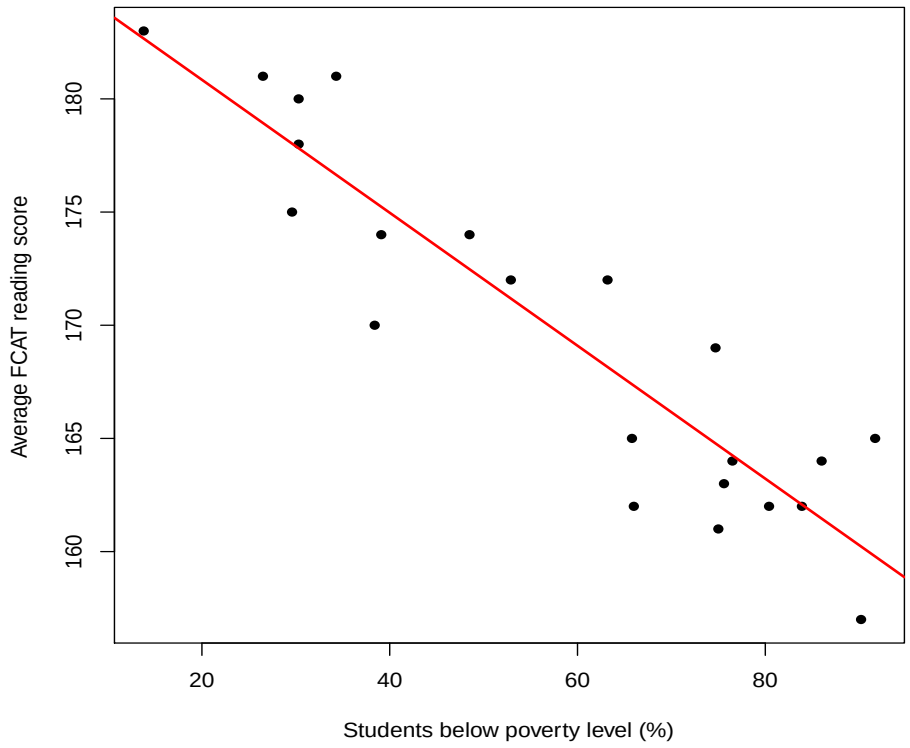


Figure 15.4: Scatterplot of the percentage of students below the poverty level (x) and average FCAT reading score (y) for a simple random sample of $n = 22$ Florida elementary schools. The least-squares regression line has been added.

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

$$x = \text{percentage of students below the poverty level}$$
$$y = \text{average FCAT reading score.}$$

for a simple random sample of $n = 22$ Florida elementary schools. The correlation is $r \approx -0.92$.

```
> cor(poverty,FCAT.reading)
[1] -0.9169302
```

The square of the correlation is

$$r^2 \approx (-0.92)^2 \approx 0.85.$$

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 (x).

- This means that approximately 15% of the variability in the average FCAT reading score data is explained by other variables (e.g., teacher quality, location of school, after-school program availability, etc.).

15.4 Assessing causation

Important: Just because two variables are correlated does not necessarily mean that there is a **causal relationship** between them. In other words, we can **not** say that

- "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 necessarily imply causation.

Discussion: Is there a causal link between the following variables?

- Smoking $\Longrightarrow$ cancer; i.e., does smoking *cause* cancer?
- Availability of guns $\Longrightarrow$ homicide rate; i.e., do more guns *cause* an increase in the homicide rate?
- SAT score $\Longrightarrow$ GPA in college; i.e., does an increase in SAT score *cause* GPAs to increase in college?

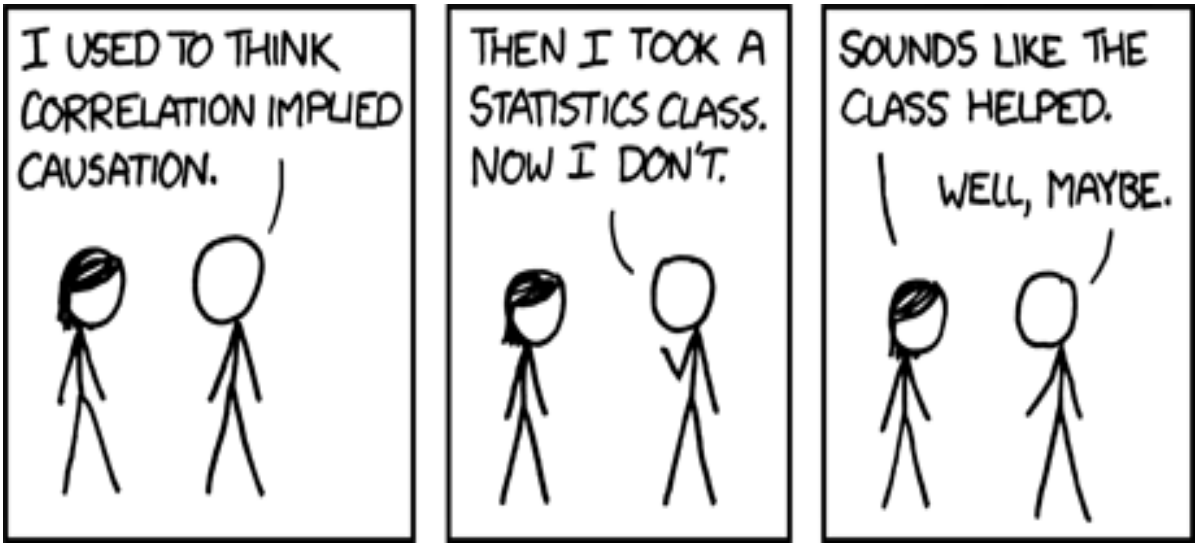


Figure 15.5: My favorite comic ever.

Discussion: It is easy to assess whether two variables are correlated (i.e., an "association" between variables). Assessing causality is far more difficult.

- Two variables may be correlated but clearly do not have a causal relationship.
- For example, a Chicago newspaper reported that "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."
  - Are these two variables correlated?

    $$x = \text{number of fire trucks}$$
    $$y = \text{damage the fire does (measured in \$)}.$$

    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 two variables are correlated is because of the presence of a third (lurking) variable, namely, the severity of the fire. Severe fires lead to both more trucks and more damage.

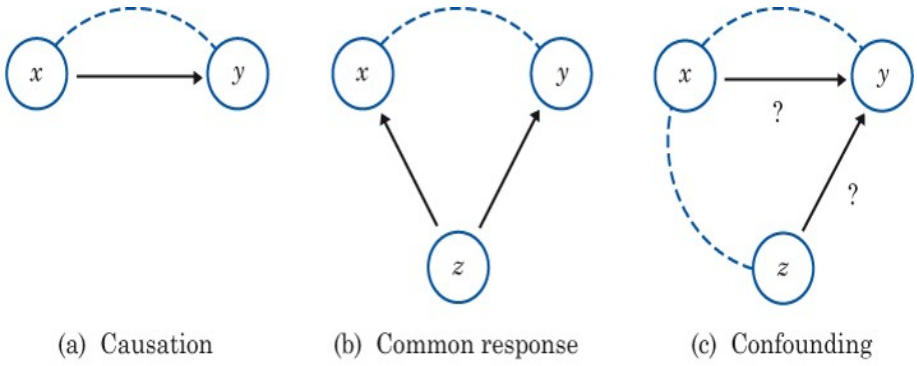


Figure 15.6: Some reasons why two variables x and y may be correlated.

- Some correlations are **coincidental**. For example, a recent study found a strong negative correlation between

$$x = \text{number of lemons imported from Mexico}$$
$$y = \text{highway fatality rate in the United States.}$$

  Does increasing the number of lemons imported from Mexico *cause* the highway fatality rate in the US to decline? Probably not.

Discussion: It is easy to find examples like the fire truck example and the lemon example. Don't be fooled when someone attempts to convince you that there is a causal relationship between two variables. Chances are good that there isn't.

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 may not be **repeatable**; i.e., if the observational study was performed again, there is no guarantee the same correlation would be found; e.g.,
  - eating more cereal $\Longrightarrow$ more male offspring.
- Designed experiments control for the effects of **lurking variables**. The presence of lurking variables usually leads to spurious correlations.

17 Thinking about Chance

17.1 Introduction

Remark: We are inundated everyday with numbers quantifying 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 10,000 over one's entire lifetime." For a car crash, it's about 1 in 600 (**Source:** Insurance Information Institute).
- "Duke is a 2:1 favorite to win the ACC men's basketball regular season championship this year."
- "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 are usually trustworthy). In other cases, numbers may arise from personal feeling or emotion (which may not be trustworthy).

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 specific 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).
- 2016 Presidential election. There are 3 possible outcomes (Trump, Clinton, Other).

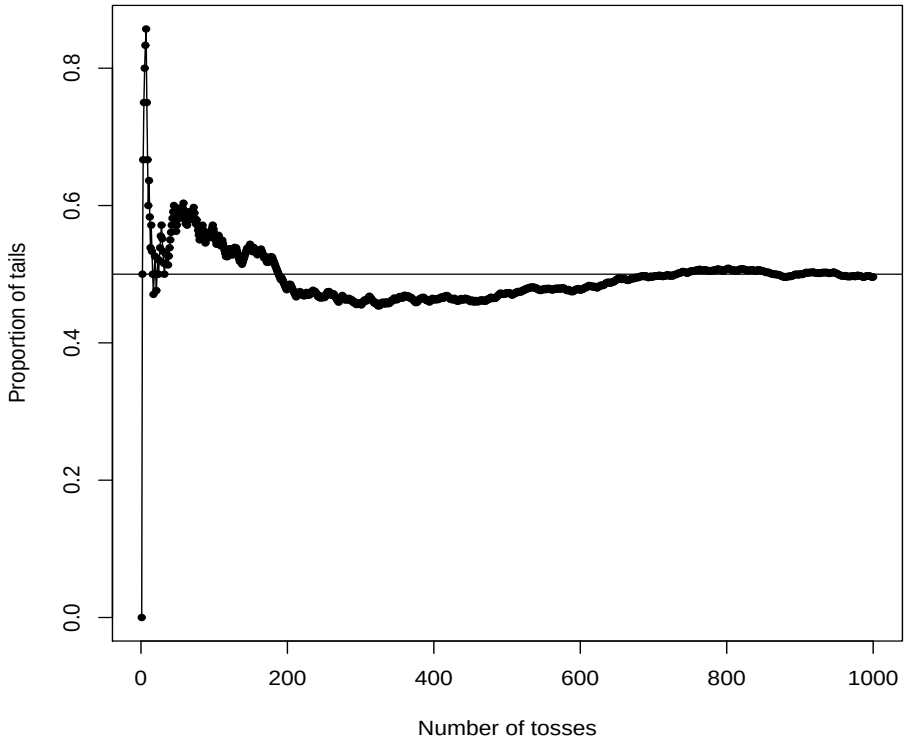


Figure 17.1: Line graph of the proportion of tails flipped in a series of 1000 flips. A horizontal line at 0.5 has been added.

Important: 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.

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?

R: I asked R to **simulate** 1000 flips. The 1000 flip results are shown in Figure 17.1. On the vertical axis is the proportion of tails flipped, charted over the number of tosses. What happens to the proportion of tails towards the end of the graph? What is this proportion getting "close" to? This is precisely what we mean by "in the long run."

Discovery: One way to think about **probability** is that it describes the proportion of times an outcome would occur in a very long series of repetitions. A probability is a number between 0 and 1 (although it can be converted to a percentage too).

- The higher the probability for a particular outcome (i.e., the closer it is to 1), the more likely it is to occur.
- The lower the probability for a particular outcome (i.e., the closer it is to 0), the less likely it is to occur.
- Outcomes with probability 1 will always occur. Outcomes with probability 0 will never occur.

17.2 Probability myths

17.2.1 Short-term regularity

Discussion: Probability provides a language to describe the long-term regularity of random behavior.

Example 17.2. Consider the following situations where one places too much faith in the **myth of short-term regularity**.

1. At a roulette wheel, the following colors have been observed in the last 15 plays:

   R G B R B R B R B B B B B B B

   Q: We walk up to the table with this information. What should we bet on the next play?
   A: If the plays are **independent**, then it doesn't matter. The probability of the following outcomes remains the same for each play:

   Red: 18/38 Black: 18/38 Green: 2/38.

2. Growing up in Iowa, I always watched Chicago Cubs games (which were announced by the late Harry Caray). Because the Cubs were usually behind in the 9th inning, Harry would try to rally the crowd by making statements like the following:

   > "At bat is Sandburg, who is hitting .250 for the season. He's 0 for 3 today, so he is due for hit."

   Problem: A batting average can be interpreted as a probability (specifically, of getting a hit). This refers to long-term behavior; not what Sandburg will do on the next at bat.

3. Empirical data suggests that 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 that the next child will be a girl?
   A: It is still 50/50. 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 that the first 3 flips were tails, and it doesn't know to "try to even things out" on the next toss.

17.2.2 Surprising coincidences

Remark: Certain outcomes may be unlikely, but this does not mean that they cannot occur. Even outcomes with small probability do occur from time to time.

Example 17.3. Consider each of these "surprise events:"

1. *Winning the lottery twice.* In 1986, Evelyn Marie Adams won the NJ state lottery for a second time (1.5 and 3.9 million dollar payoffs). Robert Humphries (PA) won his second lottery two years later (6.8 million total).

2. In Merck's VIGOR study, comparing Vioxx to naproxen (a nonsteroidal anti-inflammatory drug),

- there were 20 heart attacks out of 4,047 patients taking Vioxx.
- there were 4 heart attacks out of 4,029 patients taking naproxen (a control drug).

If there was really no difference between the heart attack rates for the two drugs, the probability of a 20-4 split in heart attacks is approximately 0.00014, or about 1.4 in 10,000.

Q: Could something like this have happened strictly by chance?

A: Yes, but it is very unlikely. Merck paid a big price for it too.

3. Chances are, there are two people sitting in this room with the same birthday. Does this sound surprising? For n different people (no twins, no triplets, etc.), 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 to be at least 0.5 (50 percent). Most people think you need many more people than this!

4. On November 18, 2006, Ohio State beat Michigan 42-39 in a college football match-up. This game was of particular interest because OSU was ranked #1 and UM 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). That is, "4239" is no more unlikely than any other 4-digit lottery number. **Moral:** Don't make a big deal out of something that isn't.

5. *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 (previously, it was 19 games). 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. A good approximation is 9.54×10^7, or about 1 in 1 million.

Remark: When "surprising" events occur (especially if they are bad), it is not uncommon for some to seek out a cause for why they have happened. Most of the time, these surprising events are likely explained by chance.

Example 17.4. 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 is the cause of about 23% of the deaths in the US, 67 out of 250 cases (about 27%) is not all that unusual.

17.3 The law of averages

Remark: There is a very strong connection between probability and our discussion of sample and population proportions. Recall the main example we discussed in Chapter 3.

Example 3.1. During July 24-25, 2016, Rasmussen Reports conducted a national telephone and online survey using a SRS of $n = 1000$ American adults. Each participant was asked:

Should police officers be required to wear body cameras while on duty?

The survey found that 700 of the 1000 adults in the sample answered "Yes" to this question. The **sample proportion** of "Yes" responses is the statistic

$$\widehat{p} = \frac{700}{1000} = 0.70.$$

The **population proportion** p, which is the proportion of all American adults, is a parameter (it is unknown).

Law of Averages: In a SRS, the sample proportion $\widehat{p}$ will approach the population proportion p when the sample size n becomes larger. This should not be surprising. Recall our "quick" formula for the margin of error (at the 95% confidence level); i.e.,

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

where n is the sample size.

Q: What happens to the margin of error as the sample size n increases?
A: It gets smaller, meaning that $\widehat{p}$ becomes a more precise **estimate** of p.

Calculations:

$n = 1000 \implies$ margin of error $= 0.032$

$n = 10000 \implies$ margin of error $= 0.010$

$n = 100000 \implies$ margin of error $= 0.003$

$n = 1000000 \implies$ margin of error $= 0.001$.

Discussion: When the margin of error is very small, this means that the sample proportion $\widehat{p}$ and the population proportion p must be close to each other (with 95% confidence). The law of averages says that the sample proportion $\widehat{p}$ will become closer and closer to the population proportion p as the sample size increases.

- How close will it get? The sample proportion $\widehat{p}$ and the population proportion p will coincide exactly when we sample the entire population (i.e., when we perform a **census**).
- The last statement is true when there are no nonsampling errors (which can still occur in a census). Recall that the margin of error only quantifies the amount of random sampling error in a SRS.

17.4 Personal probabilities

Example 17.5. As of right now, what do you think your chance is of earning a "B" or better in this class? Express your answer as a probability; i.e., as a number between 0 and 1.

Discussion: The number you have written is a probability, but how does this probability differ from that which is based on the notion of long-term regularity?

Definition: A **personal probability** of an outcome is a number between 0 and 1 that expresses an individual's judgment of how likely an outcome is. It is not based on the notion of the long-term regularity of random behavior.

- Personal probability: **subjective**; based on personal opinion and possibly even emotion (and, hence, is often not scientific). This type of probability is often used in sports predictions, for example.
- Long term regularity: based on the notion of **repeated trials**; i.e., "what happens in the long term?"

17.5 Assessing risk

Discussion: Probabilities are often used to characterize **risk**. One can think of risk as being exposed to danger, harm, or even death. For example, based on data from the Insurance Information Institute, the probability of dying

- in a plane crash during one's lifetime (for someone born in 2013) is about 1/10000.
- in a car crash during one's lifetime (for someone born in 2013) is about 1/600.

In other words, one is

$$\frac{10000}{600} \approx 16.7$$

times more likely to die in a car crash! Why are some people scared to death of flying but yet have no problems driving? Based on these probabilities, shouldn't it be the other way around?

Another example: Heart disease accounts for about 600,000 deaths per year in the United States, and the chance you will die from heart disease is about 0.25 (i.e., 1 in 4). What about dying from terrorism? You are 35,079 times more likely to die from heart disease than from a terrorist attack.

Q: Why are so many people afraid of a terrorist attack? Shouldn't one be more scared of heart disease?

Discussion:

- We might feel safer about risk when it is under our control. For example, we can (at least to some degree) control our diet, our smoking habits, and our exercise frequency. We usually can not control some lunatic about ready to carry out a terrorist attack.
- Innumeracy may play a role. For example, very small numbers like "1 in 10000" or "1 in 1000000" may sound similar, but they are not. The first probability refers to something that is 100 times riskier than the second.

18 Probability Models

18.1 Introduction

Remark: Probability is the language we use to quantify how likely something is to occur. Outcomes with small probabilities are not likely to occur. Outcomes with large probabilities are. In this chapter, we consider probability models for

(a) a single random phenomenon

(b) sample proportions.

Our discussion of (b) leads directly to an in-depth discussion of **statistical inference** for a population proportion p (Chapter 21).

18.2 Probability rules

Terminology: A **probability model** consists of two parts:

- a list of possible outcomes
- a probability for each outcome.

Example 18.1. The following table describes a probability model for the weights of American males (aged 18 and over):

Outcome	Obese	Overweight	Healthy	Underweight
Proportion	0.24	0.43	0.32	0.01

This model lists four **outcomes**: obese, overweight, healthy, and underweight. The probabilities for each outcome are 0.24, 0.43, 0.32, and 0.01, respectively.

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 that either event will occur is the sum of the individual probabilities.

Q: What is the probability an American male (e.g., randomly selected from the population) is either overweight or obese?
A: We add the probabilities

$$P(\text{Overweight}) + P(\text{Obese}) = 0.43 + 0.24 = 0.67.$$

Example 18.2. The Centers for Disease Control and Prevention uses the following probability model to describe the acquisition of hepatitis C among American adults:

Outcome	Probability
IVDU	0.60
Unprotected sex	0.25
Transfusion-related	0.05
Occupational	0.03
Other/Unknown	???

Q: What is the correct entry for "Other/Unknown?"
A: The probabilities taken together must add to 1. Therefore,

$$0.60 + 0.25 + 0.05 + 0.03 + \text{???} = 1 \implies P(\text{Other/Unknown}) = 0.07.$$

Seven percent (7%) of the hepatitis C cases among American adults are from "Other" or "Unknown" causes.

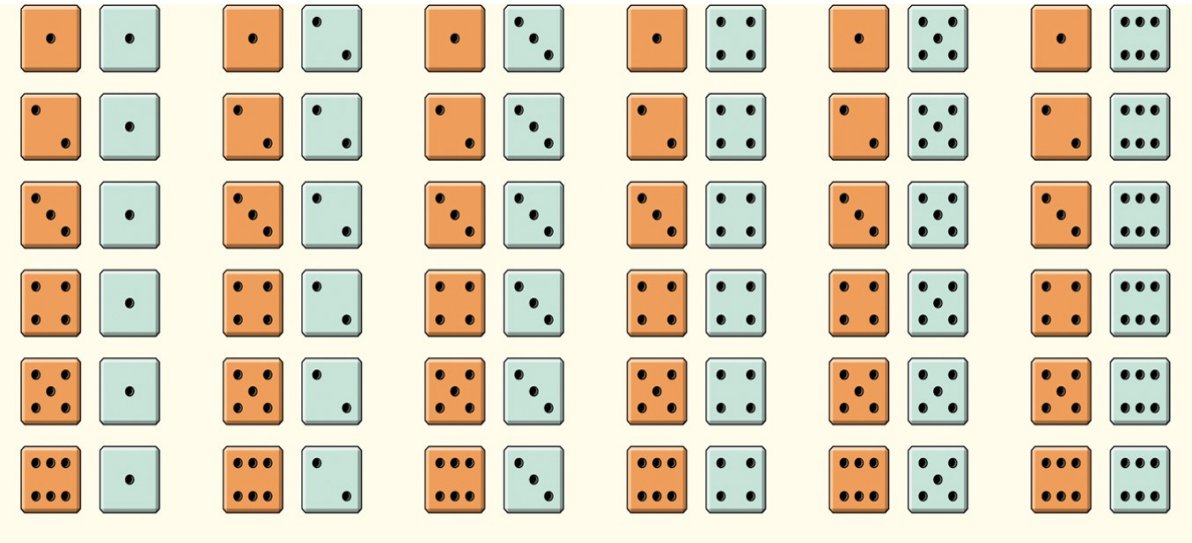


Figure 18.1: The 36 possible outcomes from rolling two dice.

Example 18.3. In the game of craps, two fair dice are rolled initially to start the game. Describe 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

$$P(7) + P(11) = \frac{6}{36} + \frac{2}{36} = \frac{8}{36} \approx 0.22.$$

18.3 Probability models for sampling

Example 18.4. During September 28-29, 2016, Rasmussen Reports conducted a national telephone and online survey using a SRS 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 that 290 of the 1000 adults in the sample answered "Yes" to this question. In other words, the sample proportion is

$$\widehat{p} = \frac{290}{1000} = 0.29 \ \ (\text{or } 29\%).$$

Discussion: We know that the sample proportion $\widehat{p}$ is calculated from the $n = 1000$ individuals who were in the SRS.

Q: What would have happened had Rasmussen conducted this poll again?
A: They would have gotten a different sample of 1000 American adults, and, as a result, a different sample proportion $\widehat{p}$.

- In other words, the sample proportion $\widehat{p}$ (a statistic) changes depending on what SRS is taken.
- It therefore makes sense to think about the following question: "What is a **probability model** for the sample proportion?" That is,
  - What is a list of the values the sample proportion $\widehat{p}$ can have?
  - What are the probabilities it take those values?

Discussion: To explore this, suppose the population proportion is $p = 0.30$; i.e., 30% of all American adults trust the media to accurately fact-check the candidates' comments. I used R to simulate 100 different values of $\widehat{p}$ under this assumption:

```
> 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
```

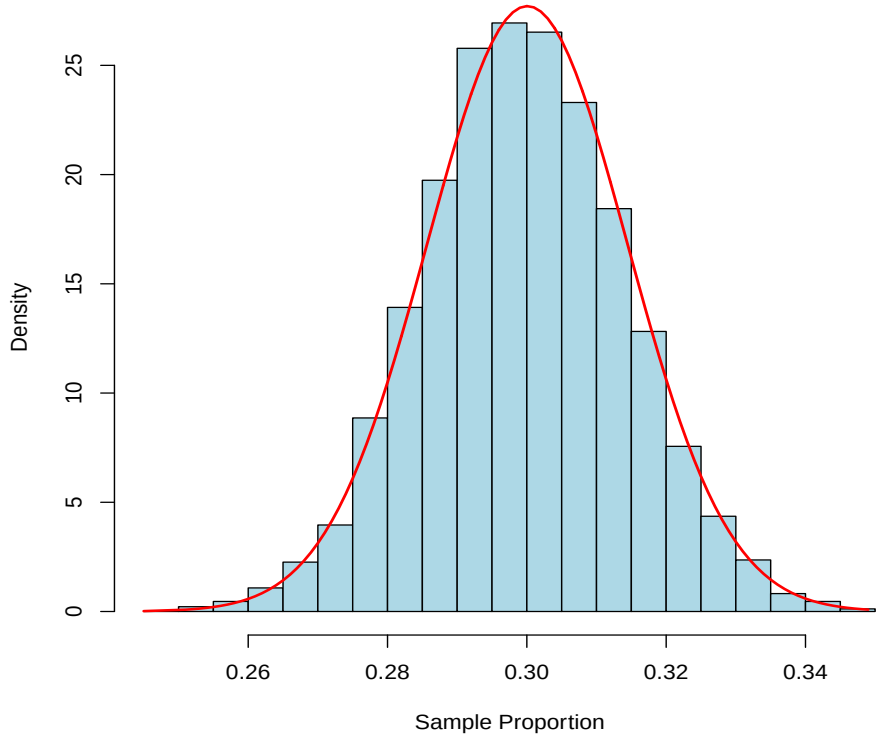


Figure 18.2: 10,000 simulated sample proportions $\widehat{p}$. Each one is based on a sample size of $n = 1000$ individuals. The population proportion is assumed to be $p = 0.30$. A normal density curve has been added.

- Most of these sample proportions are close to 0.30, but there is variation! For example, the largest $\widehat{p}$ simulated was 0.34. The smallest was 0.26.
- The possible values of $\widehat{p}$ follows a **normal distribution**, as shown in Figure 18.2.
  - The histogram in Figure 18.2 is based on 10,000 simulated values of $\widehat{p}$ (assuming $p = 0.30$). The density curve superimposed is a normal distribution.
- **Implication:** The value of the sample proportion $\widehat{p}$ (from a SRS) follows a normal distribution when the SRS contains a large number of individuals (like $n = 1000$).
- Because the normal distribution describes how the values of $\widehat{p}$ would change from sample to sample, we call this the **sampling distribution** of $\widehat{p}$.

Terminology: The **sampling distribution** of a statistic tells us (a) what values the statistic will have in repeated samples from the same population and (b) how often it takes those values.

- In other words, the term "sampling distribution" simply means "probability model for a statistic."

Result: Take a SRS of size n from a large population of individuals, where p denotes the population proportion. Let $\widehat{p}$ denote the sample proportion. For large samples (i.e., for large n),

- the sampling distribution of $\widehat{p}$ is represented by a normal density curve
- the **mean** of the sampling distribution is p
- the **standard deviation** of the sampling distribution is

$$\sqrt{\frac{p(1-p)}{n}}.$$

Calculations: Picking up where we left off in Example 18.4, we assumed the population proportion was $p = 0.30$.

- $p = 0.30$ is the **mean** of the sampling distribution of $\widehat{p}$.

Q: What is the **standard deviation** of the sampling distribution?
A: Use the formula above. With the sample size $n = 1000$, we have

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.30(1-0.30)}{1000}} \approx 0.015.$$

Using the 68-95-99.7% rule, we can say that

- 68% of the sample proportions $\widehat{p}$ should be within $(0.285, 0.315)$.
- 95% of the sample proportions $\widehat{p}$ should be within $(0.27, 0.33)$.
- 99.7% of the sample proportions $\widehat{p}$ should be within $(0.255, 0.345)$.

21 What is a Confidence Interval?

21.1 Introduction

Discussion: In this chapter, we present an introduction to confidence intervals. A **confidence interval** is an interval that contains a population parameter with a prescribed level of confidence.

- A confidence interval is calculated from a sample (usually from a simple random sample, SRS).
- Because a confidence interval is written to describe the potential values of a population parameter, using it is a form of **statistical inference**;
  - i.e., what does the information in the sample say about "what's going on" in the population?"
- Another common type of statistical inference procedure is a **hypothesis test** (which is covered in STAT 201).

Note: We will focus on two population parameters in this chapter:

- a population proportion p
  - This follows directly from our work in Chapter 18 and extends our **confidence statement** discussions in Chapter 3.
- a population mean μ.

21.2 Confidence interval for a population proportion p

Note: We begin by recalling our widely discussed example in Chapter 3.

Example 3.1. During July 24-25, 2016, Rasmussen Reports conducted a national telephone and online survey using a SRS of $n = 1000$ American adults. Each participant was asked:

Should police officers be required to wear body cameras while on duty?

The survey found that 700 of the 1000 adults in the sample answered "Yes" to this question.

- The **sample proportion** (a statistic) is

$$\widehat{p} = \frac{700}{1000} = 0.70 \quad (\text{or } 70\%).$$

- The **population proportion** (a parameter) describes all American adults (i.e., all 248 million of us). It is unknown.

For a 95% level of confidence, we used the following formula to quantify the margin of error:

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

where n is the sample size. In the Rasmussen example,

$$\frac{1}{\sqrt{1000}} \approx \frac{1}{31.62} \approx 0.0316 \quad (\text{that is, about } 3\%).$$

Confidence statement:

- "We are 95% confident that the proportion of American adults who agree police officers should wear body cameras while on duty is between 0.67 and 0.73 (i.e., between 67% and 73%)."

Discussion: In terms of proportions (not percents), this interval is calculated as

$$\begin{aligned} \widehat{p} \pm \textbf{margin of error} &\implies 0.70 \pm 0.03 \\ &\implies (0.67, 0.73). \end{aligned}$$

We are 95% confident that p (the population proportion) is between 0.67 and 0.73. The interval $(0.67, 0.73)$ is called a 95% confidence interval.

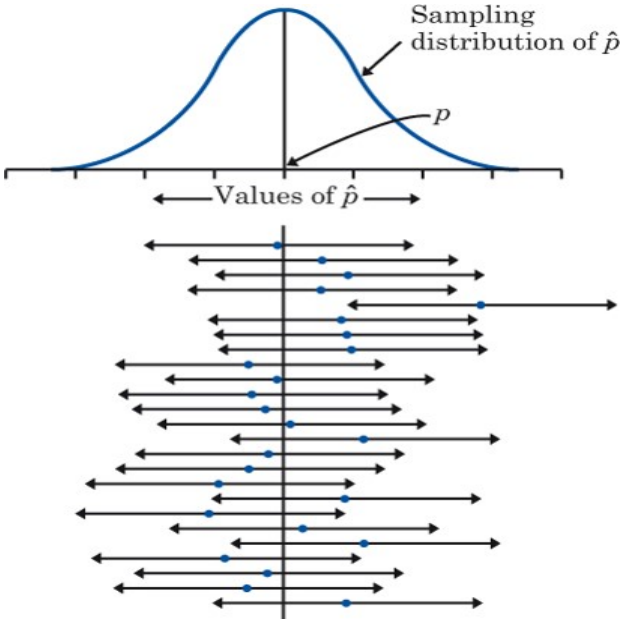


Figure 21.1: An illustration of 95% confidence intervals. **Interpretation:** In repeated sampling, 95% of these intervals will contain the population proportion p.

Terminology: A **95% confidence interval** is an interval that is guaranteed to capture the population parameter in 95% of all samples.

Remark: All the way back in Chapter 3, you were basically calculating confidence intervals for p (although you may not have realized it).

- 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.
- The quick formula for margin of error; i.e.,

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

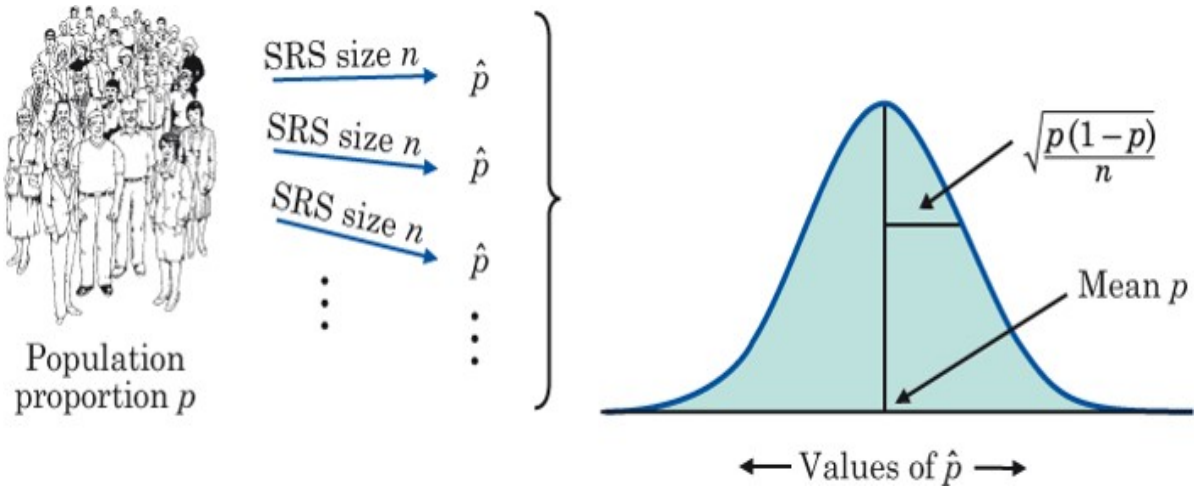
  is a conservative approximation for the true margin of error.
- We now aspire to make our confidence interval calculations for p more precise.

Discussion: We begin by recalling this important result (stated in Chapter 18):

Result: Take a SRS of size n from a large population of individuals, where p denotes the population proportion. Let $\widehat{p}$ denote the sample proportion. For large samples (i.e., for large n),

- the sampling distribution of $\widehat{p}$ is represented by a normal density curve
- the **mean** of the sampling distribution is p (i.e., $\widehat{p}$ is unbiased)
- the **standard deviation** of the sampling distribution is

$$\sqrt{\frac{p(1-p)}{n}}.$$



Note: Using the 68-95-99.7% rule, we know that 95% of all possible sample proportions $\widehat{p}$ will be within 2 standard deviations of the mean. In other words,

$$p \pm 2\sqrt{\frac{p(1-p)}{n}}$$

will contain 95% of all possible sample proportions $\widehat{p}$. Mathematically, this is the same as the following: "95% of the time, the interval

$$\widehat{p} \pm 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

will contain the population proportion p."

Result: Choose a SRS of size n from a large population with proportion p. A **95% confidence interval** for the population proportion p is

$$\widehat{p} \pm 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Q: How does this interval compare with our Chapter 3 calculation

$$\widehat{p} \pm \frac{1}{\sqrt{n}} \ ?$$

A: The first formula is more precise. The margin of error in the first formula is based on the sampling distribution of $\widehat{p}$. The second formula (Chapter 3) is OK to use as an approximation, but it is ultimately conservative (i.e., the interval is a little too wide).

Example 21.1. The Women's Interagency HIV Study was a study funded by the National Institutes of Health to investigate HIV infection in women living in the United States. Among 1,288 HIV positive women in the study (the sample), 399 self-reported having been abused as a child. The **sample proportion** of childhood abuse victims is therefore

$$\widehat{p} = \frac{399}{1288} = 0.31.$$

Q: What is the population here?

A: A reasonable answer is "all HIV positive women living in the United States." Based on 2014 CDC estimates, there are about 250,000 such women.

A 95% confidence interval for the population proportion of HIV infected women living in the United States who are childhood abuse victims is

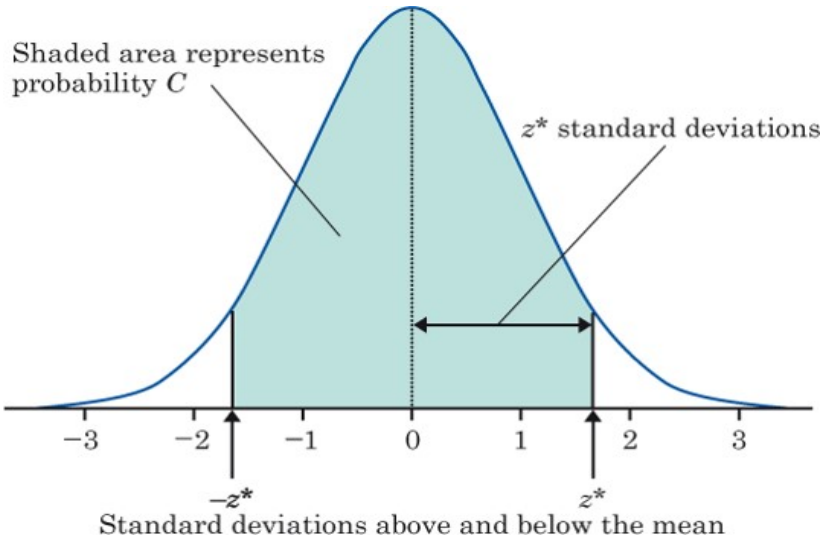
$$\begin{aligned} \widehat{p} \pm 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} &\implies 0.31 \pm 2\sqrt{\frac{0.31(1-0.31)}{1288}} \\ &\implies 0.31 \pm 0.03 \\ &\implies (0.28, 0.34). \end{aligned}$$

Interpretation: We are 95% confident that the population proportion of HIV infected women living in the United States who have been victims of childhood abuse is between 0.28 and 0.34 (i.e., between 28% and 34%).

Question: 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 is guaranteed to capture the population parameter in $C\%$ of all samples.

- C is a percentage between 0 and 100.
- We want C to be a large percentage;
  - e.g., $C = 80\%$, $C = 90\%$, $C = 95\%$, and $C = 99\%$ are 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

$$\widehat{p} \pm \underbrace{z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}_{\text{margin of error}},$$

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.

Remark: The critical value z^* is a percentile from the normal distribution with mean 0 and standard deviation 1. 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. PRAMS, the Pregnancy Risk Assessment Monitoring System, is a surveillance project of the Centers for Disease Control and Prevention and state health departments. In a recent PRAMS survey, 999 women who had given birth were asked about their smoking habits. Smoking during the last 3 months of pregnancy was reported by 125 of those sampled.

The **sample proportion** of women who smoked during the last 3 months of pregnancy is

$$\widehat{p} = \frac{125}{999} = 0.125.$$

Q: What is the population here?

A: A reasonable answer is "all recently pregnant women living in the United States."

A 95% confidence interval for the population proportion of recently pregnant women in the United States who smoked during the last 3 months of pregnancy is

$$\begin{aligned} \widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} &\implies 0.125 \pm 1.96\sqrt{\frac{0.125(1-0.125)}{999}} \\ &\implies 0.125 \pm 0.021 \\ &\implies (0.104, 0.146). \end{aligned}$$

Interpretation: We are 95% confident that the population proportion of recently pregnant women in the United States who smoked during the last 3 months of pregnancy is between 0.104 and 0.146 (i.e., between 10.4% and 14.6%).

Curiosity: What happens if we use other levels of confidence?

- A 90% confidence interval for the population proportion of recently pregnant women in the United States who smoked during the last 3 months of pregnancy is

$$\begin{aligned} \widehat{p} \pm 1.64\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} &\implies 0.125 \pm 1.64\sqrt{\frac{0.125(1-0.125)}{999}} \\ &\implies 0.125 \pm 0.017 \\ &\implies (0.108, 0.142) \quad - \textbf{shorter} \text{ than a } 95\% \text{ interval} \end{aligned}$$

- A 99% confidence interval for the population proportion of recently pregnant women in the United States who smoked during the last 3 months of pregnancy is

$$\begin{aligned} \widehat{p} \pm 2.58\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} &\implies 0.125 \pm 2.58\sqrt{\frac{0.125(1-0.125)}{999}} \\ &\implies 0.125 \pm 0.027 \\ &\implies (0.098, 0.152) \quad - \textbf{longer} \text{ than a } 95\% \text{ interval} \end{aligned}$$

- Both of these results make intuitive sense. The more confidence we require, the wider 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 here. The only difference is that now we are interested in a different population parameter.
- A population mean μ describes the center of a density curve; i.e., the curve that describes the distribution of a **quantitative variable** in the population.
- So, now we are interested in variables like IQ scores, BMI, birth weights (kg), incomes ($), arsenic levels (ppb), times to treatment failure (TTF, months), etc.

Central Limit Theorem: Take a SRS of size n from a large population of individuals, where μ denotes the population mean. Let $\overline{x}$ denote the sample mean. For large samples (i.e., for large n),

- the sampling distribution of $\overline{x}$ is represented by a normal density curve
- the **mean** of the sampling distribution is μ (i.e., $\overline{x}$ is unbiased)
- the **standard deviation** of the sampling distribution is

$$\frac{\sigma}{\sqrt{n}},$$

  where σ is the population standard deviation.

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

$$\overline{x} \pm z^* \left(\frac{s}{\sqrt{n}} \right).$$

- $\overline{x}$ is the sample mean and s is the sample standard deviation. These are calculated from the sample (i.e., they are statistics). Recall that R calculates these easily.
- 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 Example 12.2 (notes, pp 95), we examined arsenic concentration data (in parts per billion, ppb) for a random sample of $n = 102$ water wells in Texas. The histogram and boxplot of the sample data are shown in Figure 21.2. Find a 95% confidence interval for the population mean arsenic concentration μ and interpret it.

Q: What is the population here?

A: A reasonable answer is "all water wells in Texas."

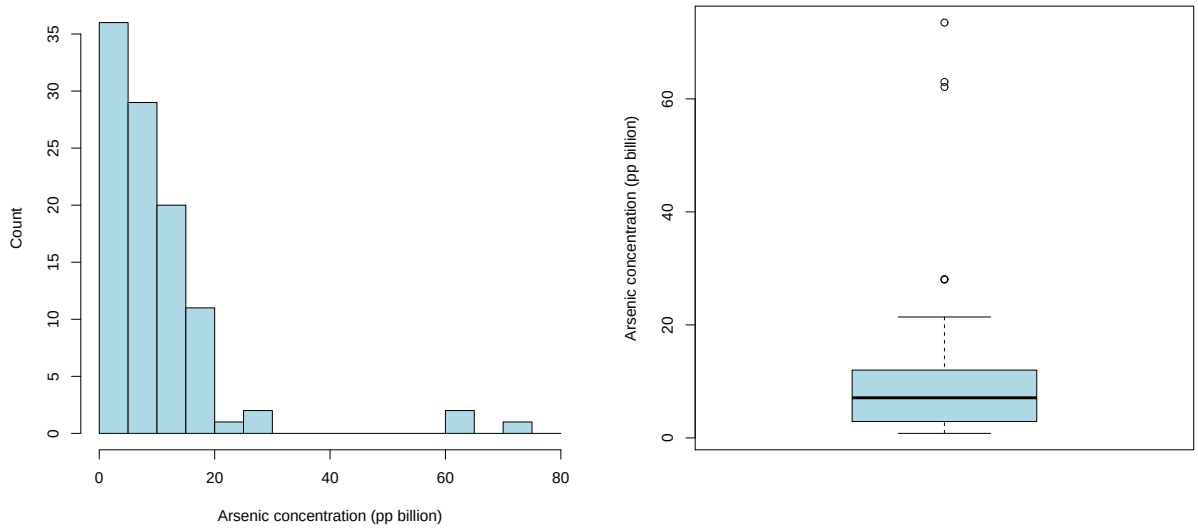


Figure 21.2: Arsenic data. Concentration of arsenic (in ppb) in ground water for a random sample of $n = 102$ wells in Texas.

SOLUTION: We use R to calculate the sample mean $\overline{x}$ and sample standard deviation s:

```
> mean(arsenic) # sample mean
[1] 9.7
> sd(arsenic) # sample standard deviation
[1] 11.5
```

A 95% confidence interval for the population mean arsenic concentration (among all water wells in Texas) is

$$\begin{aligned} \overline{x} \pm z^* \left(\frac{s}{\sqrt{n}}\right) &\implies 9.7 \pm 1.96 \left(\frac{11.5}{\sqrt{102}}\right) \\ &\implies 9.7 \pm 2.2 \\ &\implies (7.5, 11.9). \end{aligned}$$

Interpretation: We are 95% confident that the population mean arsenic concentration (among all water wells in Texas) is between 7.5 and 11.9 ppb.

Curiosity: What happens if we use other levels of confidence?

- A 90% confidence interval for the population mean arsenic concentration (among all water wells in Texas) is

$$\begin{aligned} \overline{x} \pm z^* \left(\frac{s}{\sqrt{n}}\right) &\implies 9.7 \pm 1.64 \left(\frac{11.5}{\sqrt{102}}\right) \\ &\implies 9.7 \pm 1.9 \\ &\implies (7.8, 11.6) \quad - \textbf{shorter} \text{ than a } 95\% \text{ interval} \end{aligned}$$

- A 99% confidence interval for the population mean arsenic concentration (among all water wells in Texas) is

$$\begin{aligned} \overline{x} \pm z^* \left(\frac{s}{\sqrt{n}}\right) &\implies 9.7 \pm 2.58 \left(\frac{11.5}{\sqrt{102}}\right) \\ &\implies 9.7 \pm 2.9 \\ &\implies (6.8, 12.6) \quad - \textbf{longer} \text{ than a } 95\% \text{ interval} \end{aligned}$$

Remarks:

- The confidence interval formula for μ is only applicable when the sample is a SRS.
- **Outliers** can have a large effect on a confidence interval for μ. This makes sense because both $\overline{x}$ and s can be heavily impacted by outliers.
- When the sample size n is small, the confidence interval formula we use for μ could be inappropriate.
  - Reason: The normal approximation to the sampling distribution for $\overline{x}$ may be a poor approximation.
  - The text offers an $n \geq 15$ guideline on when we can use the confidence interval formula. **However, this is only a guideline.**
  - If the underlying population density curve is very skewed, this may not be a good guideline (you might need a larger sample).
- Always plot your data first to get an idea of skewness and outliers!