# STAT 110
# INTRODUCTION TO
# DESCRIPTIVE STATISTICS

Fall, 2006

**Lecture Notes**

**Joshua M. Tebbs**

**Department of Statistics**

**The University of South Carolina**

# Contents

# 1 Where Do Data Come From?

Complementary reading from Moore and Notz: Chapter 1.

## 1.1 Introduction

*TERMINOLOGY*: **Statistics** is the development and application of methods to the collection, analysis, and interpretation of observed information (data) from planned investigations.

*SCOPE OF APPLICATION*: Statistical thinking can be used in all disciplines!! Consider the following examples:

- In a reliability study, tribologists aim to determine the main cause of failure in a jet engine assembly. Recently failures reported in the field have had an effect on customer confidence (for safety concerns).
- In a marketing project, store managers in Aiken, SC want to know which brand of coffee is most liked among the 18-24 year-old population.
- In a clinical trial, physicians on a Drug and Safety Monitoring Board want to determine which of two drugs is more effective for treating HIV in the early stages of the disease.
- In an automotive paint production process, a chemical reaction may produce a change in color depending on the temperature and stirring rate in the vessel where the reaction takes place. An engineer would like to understand how the color is affected by the temperature and stirring rate.
- The Human Relations Department at a major corporation wants to determine employee opinions about a prospective pension plan adjustment.

- In a public health study conducted in Sweden, researchers want to know whether or not smoking (i) causes lung cancer and/or (ii) is strongly linked to a particular social class.
- Shaquille O'Neill claims that he "can make free throws when they count." Is there any evidence to support his claim?
- Which professions boast the highest starting salaries for undergraduates? An advisor at USC is interested in collecting information which she can provide to her advisees.
- A physical therapist is trying to determine which type of therapy (conventional versus experimental) is better at helping his patients recover from serious wrist injuries.

## 1.2 Individuals, variables, and data

*TERMINOLOGY*: **Individuals** are simply the objects measured in a statistical problem. A **variable** is a characteristic that we would like to measure on individuals. The actual measurements recorded on individuals are called **data**.

Table 1.1: *Yield data (kg/plot) for three different fertilizers.*

| Fertilizer 1 | Fertilizer 2 | Fertilizer 3 |
|---|---|---|
| 64.8 | 56.5 | 65.8 |
| 60.5 | 53.8 | 73.2 |
| 63.4 | 59.4 | 59.5 |
| 48.2 | 61.1 | 66.3 |
| 55.5 | 58.8 | 70.2 |

**Example 1.1.** In an agricultural study in Kansas, researchers want to know which of three fertilizer compounds produces the highest wheat yield (in kg/plot). An experi-

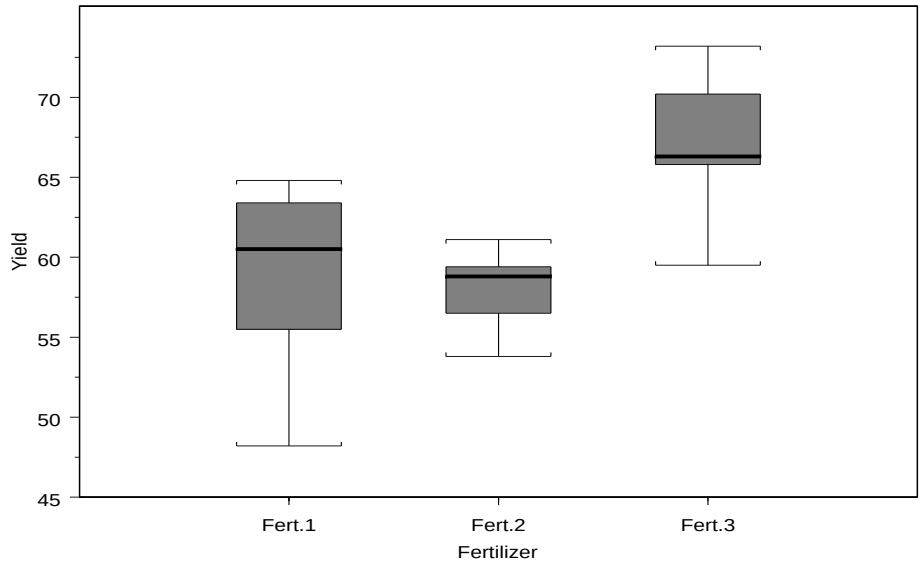



Figure 1.1: *Wheat experiment. Wheat yields for three different fertilizers.*

menter uses 15 plots of land. Each fertilizer is applied to 5 plots of land. After harvest, the resulting yield is measured. See Table 1.1 for the data in this investigation.

- **Individuals**: plots of land
- **Variables**: fertilizer type and yield (are there other variables of interest?)

**Example 1.2.** The HIV virus destroys the immune system; thus, individuals infected are susceptible to various infections which ultimately leads to death. Individuals can be infected with HIV when they come into contact with infected blood or other bodily fluids. In a large study conducted in Houston, 921 heterosexual males were recruited. These individuals were known intravenous drug users who were not receiving treatment for their addiction. See Table 1.2 for the data in this investigation.

- **Individuals**: male subjects
- **Variables**: HIV (positive/negative status), drug injected

Table 1.2: *Houston* HIV *study. Infectivity data for four different drug use behaviors.*

| Drug injected | Number of subjects | Number infected | Percentage |
|---|---|---|---|
| Heroin | 32 | 1 | 3.1% |
| Cocaine/Heroin | 198 | 11 | 5.6% |
| Cocaine | 476 | 44 | 9.2% |
| Cocaine/Amphetamines | 215 | 21 | 9.8% |

## 1.3 Observational studies

*TERMINOLOGY*: An **observational study** is an investigation carried out to observe individuals, but there is no attempt to influence the responses from those individuals. A **response** is a variable that measures the main outcome of the study. The purpose of an observational study is to describe some group or situation.

**Example 1.3.** Questioning individuals about drug use, sexual orientation, criminal activity, or other sensitive topics, is a difficult task. Individuals are not often willing to reveal such information for fear of social stigma, yet, for public-health and socioeconomic reasons, accurate information on such behaviors is often needed. In a telephone survey involving USC undergraduates, 700 individuals were asked to respond to the question:

*"Have you ever experimented with marijuana?"*

**Results:** Of these 700 individuals, 360 responded "yes," while 210 responded "no." There were 130 individuals which did not answer.

## 1.4 Populations and samples

*TERMINOLOGY*: In a statistical problem, the **population** is the entire group of individuals that we want to make some statement about. A **sample** is a part of the population that we actually observe.

QUESTION: In Examples 1.1, 1.2, and 1.3, what is the population? What is the sample?

*Example* 1.1. Population: Sample:

*Example* 1.2. Population: Sample:

*Example* 1.3. Population: Sample:

*NOTE*: Usually, we cannot examine all members of a population due to time, cost, and other constraints. So, we examine only a portion of the population and try to draw conclusions about the whole. This smaller portion is the sample. The process of generalizing the results in our sample to that of the entire population is known as **statistical inference**. We'll study this more formally later in the course.

*TERMINOLOGY*: A special type of sample, known as a **census**, is one where an attempt is made to canvas the entire population.

**Example 1.4.** The **Population Estimates Program** publishes total resident population estimates and demographic components of change (births, deaths, etc.) each year. The program also publishes the estimates by demographic characteristics (age, sex, race) for the nation, states, and counties. For more information, see `http://www.census.gov/`.

## 1.5 Experiments

*TERMINOLOGY*: An **experiment** is a planned study where individuals are subjected to **treatments**. An often important scientific question then becomes, "How do the treatments affect the individuals' responses?"

- Unlike observational studies, well-designed experiments can give us good evidence for **cause and effect**.
- Virtually all scientific research involves conducting well-designed experiments. Results from these experiments either support or refute a given hypothesis.

**Example 1.5.** Salmonella bacteria are widespread in human and animal populations, and there are over 2,000 known serotypes. The reported incidence of salmonella illnesses in humans is about 17 cases per each 100,000 persons. (**Source**: CDC. Preliminary FoodNet Data on the Incidence of Foodborne Illnesses, 2000). A food scientist wants to see how withholding feed from pigs prior to slaughter can reduce the size of gastrointestinal tract lacerations during the actual slaughtering process. This is an important issue since pigs infected with salmonellosis may contaminate the food supply through these lacerations (among other routes, including fecal matter and meat juices).

- **Individuals** = pigs
- **Population** = all market-bound pigs
- **Sample** = 45 pigs from 3 farms (15 per farm) assigned to three treatments:
  - Treatment 1: no food withheld prior to transport,
  - Treatment 2: food withheld 12 hours prior to transport, and
  - Treatment 3: food withheld 24 hours prior to transport.
- Data were measured on many variables, including body temperature prior to slaughter, weight prior to slaughter, treatment assignment, the farm from which each pig originated, number of lacerations recorded, and size of laceration (cm).
- **Boxplots** of the lacerations lengths (by treatment) are in Figure 1.2.

*DISCUSSION QUESTIONS*:

- How should we assign pigs to one of the three treatments?
- Why would one want to use animals from three farms?
- Why might body temperature or prior weight be of interest?

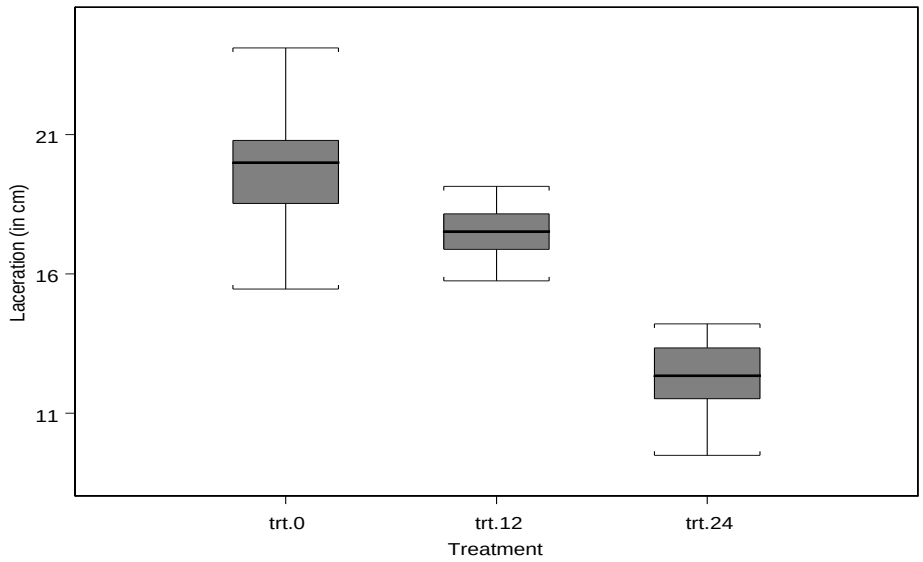



Figure 1.2: *Salmonella experiment. Laceration length for three treatments.*

*REMARK*: In agricultural, medical, and other experimental applications, the most common objective is to **compare** two or more treatments. In light of this, we will often talk about statistical inference in the context of comparing treatments in an experimental setting. For example, in the salmonella experiment, one goal is to compare the three withholding times (0 hours, 12 hours, and 24 hours).

- Since populations are usually large, the sample we observe is just one of many possible samples that are possible to observe. That is, samples may be similar, but they are by no means identical.
- *Because of this, there will always be a degree of uncertainty about the decisions that we make concerning the population of interest!!*
- We would like to make conclusions based on the data we observe, and, of course, we would like our conclusions to apply for the entire population of interest. This is the idea behind **statistical inference**.

# 2 Samples, Good and Bad

Complementary reading from Moore and Notz: Chapter 2.

## 2.1 Introduction

*BASIC IDEA*: An investigator usually wants to make a statement about a large group of individuals. This group is called the **population**.

- all eligible voters in SC
- all employees at a company
- all market-bound pigs
- all USC students
- all advanced-HIV patients.

*THE CATCH*: Because of cost and time considerations, it is not practical to measure each individual in the population. Instead, only a part of the population can ever be examined; this part is called the **sample**. *Hopefully, the sample is representative of the population.* So, how do we choose a sample from a population?

*TERMINOLOGY*: The way which we select a sample from the population is called the **sampling design**.

## 2.2 Poor sampling designs

*TERMINOLOGY*: If a sampling design provides samples that are not representative of the population from which the individuals were drawn, we say that the sampling design is **biased**. We now look at two biased sampling designs.

*VOLUNTARY RESPONSE SAMPLE*: A **voluntary response sample** (VRS) consists of people who choose themselves by responding to a general appeal (e.g., online polls, mail-in surveys, etc.). The problem with such samples is that people with strong opinions are usually the only ones which respond. This gives biased results!

**Example 2.1.** C-Span routinely provides viewers the opportunity to phone in comments about important political issues. Recently, callers were allowed to comment on the current state of affairs with Iran and North Korea. Do the callers' comments accurately reflect the views of the entire voting public?

**Example 2.2.** The 1936 presidential election proved to shape the future of opinion polling. The Literary Digest, the venerable magazine founded in 1890, had correctly predicted the outcomes of the 1916, 1920, 1924, 1928, and 1932 elections by conducting polls. These polls were a lucrative venture for the magazine: readers liked them; newspapers played them up; and each "ballot" included a subscription blank. The 1936 postal card poll claimed to have asked one forth of the nation's voters which candidate they intended to vote for. In Literary Digest's October 31 issue, based on more than 2,000,000 returned post cards, it issued its prediction:

> "Republican candidate Alfred Landon would win 57 percent of the popular vote."

**In the election, FDR won 62 percent of the popular vote in one of the biggest landslides in election history!** The failure of the Literary Digest poll can be attributed to bad sampling techniques.

- The sample itself was a voluntary response sample. Some people who received questionnaires responded, but most did not.
- 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. Meanwhile, George Gallup was beginning to use **random samples** and scientific methods, and the survey research industry was born.

*CONVENIENCE SAMPLE*: A **convenience sample** chooses individuals that are easiest to contact. The researcher makes no attempt, or only a limited attempt, to ensure that this sample is an accurate representation of some larger group or population.

- In general, the statistics community frowns on convenience samples.
- You will often have great difficulty in generalizing the results of a convenience sample to any population that has practical relevance.

**Example 2.3.** Interviewees standing by the door at Walmart taking surveys at 11.00am are excluding those individuals who do not frequent Walmart and those individuals which can not shop at that time.

**Example 2.4.** An interview on the street corner would exclude non-ambulatory patients. If the outcome measures are not strongly related to this factor, this may be alright. For example, an assessment of eye color is probably safe in this setting. But a street corner interview would be a disaster if you were measuring something like the degree of disability.

## 2.3 Simple random samples

*SIMPLE RANDOM SAMPLE*: A **simple random sample (SRS)** is a sampling design where each sample of size $n$ has an equal chance of being selected.

- We are choosing individuals so that our sample will hopefully be representative.
- Each individual in the population has an equal chance of being selected.
- We use the terms "random sample" and "simple random sample" interchangeably.

*CONCEPTUALIZATION*: Drawing random samples can be "thought of" as picking numbers at random from a hat. This is useful for small populations; not for large!

*IN PRACTICE*: Choosing random samples must involve the use of a **randomization mechanism**. This is simply a tool which can select individuals using chance.

*TABLE OF RANDOM DIGITS*: Table A (see pages 550-551 MN) contains a **Table of Random Digits**. We can use this as a randomization mechanism to help select simple random samples!

- The table consists of integers 0, 1, 2, ..., 9.
- The integers have no pattern to them (they were most likely generated by a computer program).
- Since they are randomly generated, we can exploit this to choose an SRS.

*USING THE TABLE*: We can use Table A to choose an SRS; we do this in three steps:

1. Assign a numerical label to every individual in the population. Be sure that all labels have the same number of digits.
2. Starting **anywhere** in Table A, choose a long string of numbers.
3. Reading from left to right (for ease), match the chosen numbers to those individuals in your list. Those that match make up the sample.

**Example 2.5.** Last semester, I taught a STAT 110 section with 50 students. I wanted to get student feedback on my teaching styles, but I did not have time to get feedback from every student in the class. I decided to take a random sample of $n = 5$ students from the class. Then, with each student in the sample, I would get feedback (in the form of a questionnaire) and then use this information to summarize the impressions of the entire class. Here is a list of students (for verisimilitude, I have changed the last names to preserve confidentiality):

Table 2.3: *Class list for STAT* 110*, Spring,* 2005.

| | | | |
|---|---|---|---|
| BROWN | 01 | CALDWELL | 26 |
| CAMPBELL | 02 | CAST | 27 |
| CLAYTON | 03 | CLINGER | 28 |
| COLUMBUS | 04 | COUCH | 29 |
| DEE | 05 | DEROGATIS | 30 |
| DOTSON, C. | 06 | DOTSON, F. | 31 |
| ELLIS | 07 | FINE | 32 |
| GAY | 08 | GRAHAM, J. | 33 |
| GRAHAM, S. | 09 | GROUND | 34 |
| HAGENHOFF | 10 | HAMMON | 35 |
| HANSON | 11 | HII | 36 |
| HIX | 12 | HOPKINS | 37 |
| IVORY | 13 | KALAMAR | 38 |
| KREGER | 14 | LOPEZ | 39 |
| MANNING | 15 | MARTIN | 40 |
| MAXWELL | 16 | MCCUISTION | 41 |
| MCLEAN | 17 | MCMENAMY | 42 |
| MEHARG | 18 | MELTON | 43 |
| MILLER | 19 | MOWREY | 44 |
| OGLESBY | 20 | OLIVER | 45 |
| PERRYMAN | 21 | REBER | 46 |
| RIGDON | 22 | ROBINSON | 47 |
| ROGERS, D. | 23 | ROGERS, J. | 48 |
| SAURO | 24 | SCHIEBER | 49 |
| SHUGHART | 25 | SNOW | 50 |

*SELECTING THE SAMPLE*: Starting at **Line 122** in Table A, we read the random digits across:

```
13873 81598 95052 90908.
```

With these digits, we can form the following two-digit numbers (reading left to right):

```
13, 87, 38, 15, 98, 95, 05, 29, 09, 08.
```

The first five that match our list are

```
13, 38, 15, 05, 29.
```

The five students which constitute the random sample are

| | |
|---|---|
| IVORY | 13 |
| KALAMAR | 38 |
| MANNING | 15 |
| DEE | 05 |
| COUCH | 29 |

EXERCISE: Note that if you start at a different line in Table A, you will get a different sample! Using a different line from the Table of Random Digits (you pick!), select another random sample of students from the list in Table 2.3. Write your sample below.

*SOFTWARE*: Computer programs (much like those used to design Table A) can also be used. The web site `www.randomizer.org` contains an applet that allows one to select a random sample directly. Using the randomizer at this site, I asked for a random string of numbers between 01 and 50; the string generated was 26, 49, 11, 46, and 23. The students that match these numbers would constitute my sample.

*DISCLAIMER*: Selecting simple random samples from very large populations is not as easy as it sounds! For example, the population size of USC students (just in Columbia) is around 28,000! In practice, researchers most likely would not

(i) identify exactly what the population is,

(ii) identify all of the members of the population (an impossible task, most likely), and

(iii) choose an SRS from this list.

*REALITY*: It is more likely that reasonable attempts will have been made to choose the individuals at random from those available. Hence, in theory, the sample obtained for analysis might not be "a true random sample;" however, in reality, the SRS model assumption might not be that far off.

## 2.4 Trusting samples

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

- Good design: SRS
- Bad designs: VRS, convenience

*MAIN POINT*: Voluntary response and convenience samples will rarely provide representative snapshot of the population. To avoid bias, the best design to use is an SRS.

*UNFORTUNATE REALITY*: Even if we use an SRS, we still may (due to chance luck of the draw) select individuals which are not representative of the population in some way. However, this is our best chance.

*SCRUTINIZING WHAT YOU READ*: When reading newspaper/online articles, the following phrases usually indicate that the results are biased, most likely due to poor sampling designs and/or poor data collection methods:

- "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 which responded."

**Example 2.6.** On February 5, 2005, the *New York Times* published an editorial article citing that 41 percent of biology teachers in Louisiana rejected the theory of evolution.

# 3 What Do Samples Tell Us?

Complementary reading from Moore and Notz: Chapter 3.

## 3.1 Parameters and statistics

**Example 3.1.** A Fox News poll, taken on June 29, 2006, reported the results from an SRS of $n = 900$ adults nationwide. Interviewers asked the following question:

> *"Do you approve or disapprove of the way George W. Bush is handling his job as president?"*

Of the 900 adults in the sample, 369 responded by stating they approve of the President's handling of his job. What does this information tell us about the population of US adults?

*TERMINOLOGY*: A **parameter** is a number that describes a **population**. Since it characterizes a population, and we can not contact every member in the population, a parameter is unknown. Parameters are fixed values (they are what they are, even if we don't know them!).

*TERMINOLOGY*: A **statistic** is a number that describes a **sample**. When we take a sample, we can compute the value of the statistic. Of course, different samples may produce different values of the statistic!

*IMPORTANT*: We use sample statistics to **estimate** population parameters.

**Example 3.2.** In Example 3.1, let

$$p = \text{proportion of US adults which approve of President Bush}$$

The value of $p$ is a parameter since it represents the population. The parameter $p$ is called the **population proportion**. To know $p$, we would have to poll every American

adult! Since this is not done, we do not get to know $p$. On the other hand, let

$\widehat{p}$ = proportion of individuals *in our sample* who approve of President Bush.

Since $\widehat{p}$ is computed from the sample of individuals, we call it the **sample proportion**. Recall that, in the sample, 369 of the 900 adults approved of the way President Bush is handling his job. Thus, the sample proportion is

$$\widehat{p} = \frac{369}{900} = 0.41,$$

or 41 percent. We might state that

*"Our sample results indicate that 41 percent of US adults favor the way that President Bush is handling his job."*

*SUMMARY*:

- The population proportion $p$ is **unknown** because it represents the entire population of US adults.
- The sample proportion $\widehat{p} = 0.41$ is computed from the sample of 900 US adults.
- We use $\widehat{p} = 0.41$ as an **estimate** of $p$. That is, we use the sample proportion to estimate the population proportion!

**Example 3.3.** A Columbia-based health club wants to estimate the proportion of Columbia residents who enjoy running as a means of cardiovascular exercise. Define

$p$ = proportion of Columbia residents who enjoy running as a means of exercise.

Since $p$ denotes the proportion of all Columbia residents, it is a **parameter** which describes this population.

*SAMPLING DESIGN*: We decide to take an SRS of $n = 100$ Columbia residents. Recall that we have two values:

$p$ = the true proportion of Columbia residents who enjoy running (unknown)

$\widehat{p}$ = the proportion of residents who enjoy running observed in our sample.

In our SRS of $n = 100$ Columbia residents, 19 said that they enjoy running as a means of exercise. The sample proportion is

$$\widehat{p} = \frac{19}{100} = 0.19,$$

or 19 percent.

*HYPOTHETICALLY*: Suppose now that I take another SRS of Columbia residents of size $n = 100$ and 23 of them said that they enjoy running as a means of exercise. From this sample, our estimate of $p$ is

$$\widehat{p} = \frac{23}{100} = 0.23,$$

or 23 percent. That the two samples gave different estimates should not be surprising (the samples most likely included different people). *Statistics' values vary from sample to sample because of this very fact.* On the other hand, the value of $p$, the population proportion, does not change!

*OF INTEREST*: In light of the fact that statistics' values will change from sample to sample, it is natural to want to know what would happen if we **repeated** the sampling procedure many times. Suppose that I selected **50 different simple random samples**, each of size $n = 100$. Here are the results:

```
0.19 0.23 0.20 0.20 0.16 0.25 0.21 0.17 0.14 0.26
0.18 0.25 0.18 0.21 0.20 0.21 0.20 0.18 0.22 0.19
0.19 0.22 0.23 0.17 0.26 0.21 0.19 0.19 0.20 0.10
0.18 0.21 0.20 0.23 0.23 0.26 0.18 0.18 0.16 0.24
0.27 0.18 0.19 0.26 0.25 0.24 0.10 0.18 0.18 0.25
```

*SUMMARY*: We have just produced 50 different sample proportions $\widehat{p}$, each one computed from an SRS of size $n = 100$. Based on these values of $\widehat{p}$, obtained from repeated sampling from our population, consider the following questions:

- What do you think the **true value** of $p$ is close to?
- Do you think that $p$ is equal to 0.75? Why or why not?
- Notice how there is **variability** in the different values of $\widehat{p}$. Why do think that is?

## 3.2 Bias and variability

*TERMINOLOGY*: **Bias** is consistent, repeated deviation of a sample statistic from a population parameter (in the same direction) when we take repeated samples.

*TERMINOLOGY*: **Variability** describes how spread out the values of the sample statistic are when we take repeated samples.

*IMPORTANT*: A good sampling design has both small bias and small variability!

- To reduce bias, use simple random sampling (SRS). Simple random sampling produces estimates which are **unbiased**. That is, the values of the statistic neither overestimate nor underestimate the value of the population parameter.
- To reduce variability in an SRS, use a larger sample. The larger the sample size, the smaller the variability!

*SUMMARY*: Large random samples almost always give an estimate that is close to the truth (i.e., that is close to the value of the population parameter).

*WARNING*: We can reduce the variability in our estimates by taking larger samples; however, **this only applies to simple random samples!** Larger sample sizes do not reduce variability when using bad sampling designs (e.g., convenience samples and VRS).

## 3.3 Margin of error

**Example 3.4.** During the week of 8/10/01, Gallup/CNN/USA conducted a poll asking an SRS of 1000 Americans whether they approve of President Bush's performance as President. The approval rating was 57 percent. In their next poll conducted during the week of 9/21/01, Gallup/CNN/USA conducted the same poll asking an SRS of 1000 Americans whether they approve of President Bush's performance as President. The approval rating was 90 percent. Why was there such a big difference in approval rating?

*COMMON STATEMENT*: When reported in the news, good news reporters would convey this in the following manner:

- "... in this poll, the approval rating for President Bush is 57% plus or minus 3%."
- "... in this poll, the approval rating for President Bush is 90% plus or minus 3%."
- What is this plus or minus 3%??? Where does it come from?

*FACT OF LIFE*: An SRS is a good sampling design. It produces unbiased estimates of population parameters and provides small variability as well. However, even with an SRS, a sample statistic will **never** estimate the population parameter exactly. There will **always** be a degree of uncertainty. This is true because we never sample the whole population! *Because we will never estimate the population parameter exactly, we would like to quantify the uncertainty in our estimate; we do this by using the margin of error.*

*MARGIN OF ERROR*: The **margin of error** is a value that quantifies the uncertainty in our estimate. A measure of how close we believe the sample proportion $\widehat{p}$ is to the population proportion $p$ is given by the margin of error.

*INTERPRETATION*: In Example 3.4, the margin of error is given by the value "3%." This is the value that quantifies the uncertainty in our estimates

$$\widehat{p} = 0.57 \ \text{(before 9/11)} \qquad \text{and} \qquad \widehat{p} = 0.90 \ \text{(after 9/11)}.$$

*INTERPRETATION*: We interpret these statements as

- "We are **95 percent confident** that the true approval rating for President is between 54% and 60%." (before 9/11)
- "We are **95 percent confident** that the true approval rating for President is between 87% and 93%." (after 9/11)

*NOTE*: These statements describe the **entire population** of US adults!

*COMPUTING THE MARGIN OF ERROR*: With an SRS, we use the sample proportion $\widehat{p}$ to estimate the true population proportion $p$. The margin of error associated with $\widehat{p}$, using 95 percent confidence, is approximately equal to $1/\sqrt{n}$; that is,

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

**Example 3.5.** In Example 3.4, Gallup/CNN/USA used an SRS of size $n = 1000$ (before 9/11). The margin of error associated with the estimate $\widehat{p}$ is equal to

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

Recall that Gallup/CNN/USA announced a margin of error of 3%. This is where that figure comes from!

*IMPORTANT QUESTION*: What does the term "95 percent confident" really mean?

## 3.4 Confidence statements

*ANSWER*: The term "95 percent confident" has to do with the notion of **repeated sampling**:

- If we took many samples using the same method, 95 percent of the time we would get a result within the margin of error.

- 95 percent of time, we will be close to the truth (i.e., close to the true $p$); that is, our estimate $\widehat{p}$ will be inside the margin of error.
- 5 percent of the time, our estimate $\widehat{p}$ will "miss" by more than the margin of error.
- Unfortunately, we will never get to know whether we "hit" or "miss." *This is just a fact of life!*

*FORMAL DEFINITION*: A **confidence statement** has two parts:

- A **margin of error**. This quantifies how close the sample statistic is to the population parameter.
- A **level of confidence**. This says what percentage of possible samples satisfy the margin of error.

*IMPORTANT POINTS*: These are points to remember. All of these statements assume that an SRS is being used.

1. The conclusion of a confidence statement applies to the population of interest, not the sample.
2. Our conclusion about the population is never completely certain. The level of confidence tells us how confident we are.
3. A survey can choose to use a confidence level other than 95 percent; however, it is very common to report the margin of error for 95 percent confidence.
4. To reduce the margin of error associated with an estimate, use a larger sample.
5. The margin of error does not depend on the size of the population, as long as the population is at least 100 times larger than the sample. We will often deal with "large populations," (e.g., all US adults, all cancer patients, all single mothers, etc.) so this will not be a major point of contention.

# 4 Sample Surveys in the Real World

Complementary reading from Moore and Notz: Chapter 4.

## 4.1 Introduction

*FACT*: Many people are hesitant to willingly participate in surveys, especially over the phone, in person, or through email.

**Example 4.1.** A political opinion poll (conducted over the phone) talks to 1000 people, chosen at random, announces its results, and announces a margin of error. However, let's look more closely at what it took to secure 1000 responses.

Table 4.4: *Breakdown of individuals asked to participate in a phone survey.*

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

- A total of 2879 individuals were needed to get 1000 responses!
- The **response rate** is only
$$\frac{1000}{2879} \approx 0.35$$
or 35 percent.
- **Nonresponse rate** is 65 percent!!

*MORAL*: Getting feedback from a simple random sample usually is not so simple. Human subjects are inherently difficult to deal with.

## 4.2 How sample surveys go wrong

*TWO TYPES OF ERROR*: There are two types of errors that are inherently part of all sample surveys: sampling errors and nonsampling errors.

*TERMINOLOGY*: **Sampling errors** are errors caused by the act of taking a sample from the population. They cause sample results to be different from the results of a census (i.e., if we were to sample the entire population).

- **Random sampling error**
  - the deviation between the sample statistic (estimate) and the population parameter
  - caused by chance in taking random samples
  - in an SRS, the margin of error quantifies this.
- **Bad sampling methods** (VRS, convenience)
- **Undercoverage**
  - this occurs when some groups in the population are left out of the process of choosing the sample
  - often as a result of a defective **sampling frame** (a complete list of all individuals in the population).

*SAMPLING FRAMES*: If the sampling frame leaves out certain classes of people, even random samples will cause biased results! Which groups of individuals are left out with the following frames?

- Telephone directory of Columbia
- List of USC undergrad email addresses
- List of registered voters

*TERMINOLOGY*: **Nonsampling errors** are errors not related to the act of selecting a sample from the population.

- Processing errors/data entry errors
- **Nonresponse**: the failure to obtain data from an individual selected for a sample (individual refuses to participate).
- **Response bias**
  - **Interviewer bias:** occurs when the interviewer knowingly/unknowingly sways responses
  - **Respondent bias:** occurs when respondent gives false replies. This is especially prevalent when subjects are asked sensitive questions.
- **Poorly worded questions**: The wording of questions almost always influences the answers.

*EXAMPLES OF POORLY WORDED QUESTIONS*:

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

A more appropriate question would leave out the part about "mounting casualties." No one likes casualties.

> *"Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?"*

When 22 percent of the sample said that it was "possible," the news media wondered how so many Americans could be uncertain. A much simpler version of this question was later asked and only 1 percent of the respondents said it was "possible."

*"Is our government providing too much money for welfare programs?"*

- 44 percent responded "yes."
- when the word "welfare" was replaced with "assistance to the poor," only 13 percent responded "yes."
- the word "welfare" carries with it a negative connotation.

## 4.3 Stratified random samples

*TERMINOLOGY*: A **probability sample** is a sample chosen by chance. There are two types of probability samples we will discuss:

- simple random sample (SRS)
- stratified random sample

*REVIEW*: We prefer the simple random sampling (SRS) method over methods such as voluntary response and convenience sampling because the latter methods are prone to bias. The SRS method (in the absence of all nonsampling errors) eliminates bias and produces small variability with large samples.

*INNATE DEFICIENCY*: With an SRS, because each sample of size $n$ has the same chance of being selected, we can not obtain information for separate "groups" of individuals (e.g., individuals of different gender, race, income class, religion, etc.).

*"Do you approve or disapprove of the way George W. Bush is handling the response to Hurricane Katrina?"*

*LEGITIMACY*: Might males/females respond to this question differently? Individuals of different races? Individuals of different income classes? Individuals of different religions? If so, then an SRS design would not be appropriate!

*TERMINOLOGY*: A **stratified random sampling design** is selected as follows:

1. First, divide the sampling frame into distinct groups or **strata**.
2. Take simple random samples (SRS) within each stratum and combine these to make up the complete sample.

**Example 4.2.** In a large HIV seroprevalence study conducted in Houston, 921 heterosexual males were recruited in a stratified fashion. These individuals were known intravenous drug users who were not receiving treatment for their addiction. See Table 4.5 for the data in this investigation. Here we see that individuals have been separated (**stratified**) by race.

Table 4.5: *Houston* HIV *study. Infectivity data for three different races.*

| Race | Number of subjects | Number infected | Percentage |
|---|---|---|---|
| Hispanic | 107 | 4 | 3.7% |
| white | 214 | 14 | 6.5% |
| black | 600 | 59 | 9.8% |

**Example 4.3.** In Example 4.2, what is the **margin of error** for each stratum?

- **Hispanic** group:

$$\textbf{margin of error} = \frac{1}{\sqrt{107}} \approx 0.10, \text{ or about 10 percent.}$$

- **White** group:

$$\textbf{margin of error} = \frac{1}{\sqrt{214}} \approx 0.07, \text{ or about 7 percent.}$$

- **Black** group:

$$\textbf{margin of error} = \frac{1}{\sqrt{600}} \approx 0.04, \text{ or about 4 percent.}$$

*IMPORTANT QUESTION*: Are there really true differences among the **population proportions** for the three groups? Answer this question by writing confidence statements for each race.

- **Hispanic** group:

- **White** group:

- **Black** group:

## 4.4 Questions to ask before you believe a poll

- Who carried out the survey?
- What was the population? To which group can I make a statement about?
- How was the sample selected?
- How large was the sample (do they give any insight as to the margin of error?)
- What was the response rate?
- How were the subjects contacted?
- When was the survey conducted?
- What were the exact questions asked?

*WARNING*: **Always be very skeptical** of anything you read which does not answer all of these questions explicitly!

# 5 Experiments, Good and Bad

Complementary reading from Moore and Notz: Chapter 5.

## 5.1 Terminology

Here is a quote from Moore and Notz (page 76):

> *"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*: A **response variable** is a variable that measures an outcome or result of a study.

*TERMINOLOGY*: An **explanatory variable** is a variable that we think explains or causes changes in the response variable.

*TERMINOLOGY*: The individuals studied in the experiment are called **subjects**.

*TERMINOLOGY*: A **treatment** is any specific experimental condition applied to the subjects.

## 5.2 Examples

**Example 5.1.** *Does aspirin reduce the rate of heart attacks?* The Physicians Health Study (PHS) was a large **double-blinded** experiment involving 22,000 male physicians. One of the goals of this study was to investigate whether taking aspirin reduces the risk of heart attacks. One group of about 11,000 took an aspirin every second day, while the rest of them took a **placebo** (i.e., a "sugar pill" designed to look like an aspirin).

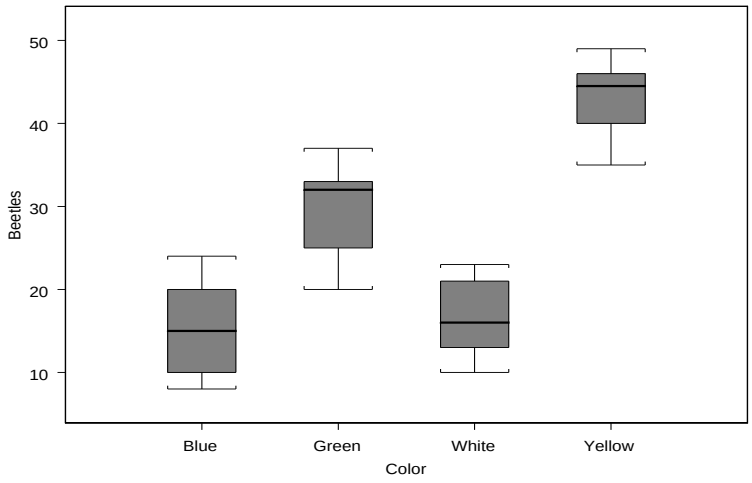



Figure 5.3: *Number of beetles caught by colored boards. Boxplots for four colors.*

- **Subjects**: Physicians (they receive the treatment)
- **Response variable**: Heart attack/not
- **Explanatory variable**: Drug (placebo/aspirin). Drug is also the treatment because it is the condition applied to the physicians.

*FINDINGS*: Daily low-dose aspirin decreased the risk of a first myocardial infarction by 44 percent over those that took placebo. For more information about the Physicians Health Study, see `http://phs.bwh.harvard.edu/index.html`.

*TERMINOLOGY*: In Example 5.1, we might call the group that received the placebo the **control group**. Using a control group gives a frame of reference for comparisons.

**Example 5.2.** *Are beetles attracted to bright colors?* To detect the presence of harmful beetles in farm fields, experimenters placed six boards of each of four colors at random locations in a field of oats and measured the number of cereal leaf beetles trapped. The boards were covered with sticky material to trap the beetles easily. Side-by-side boxplots appear in Figure 5.3.

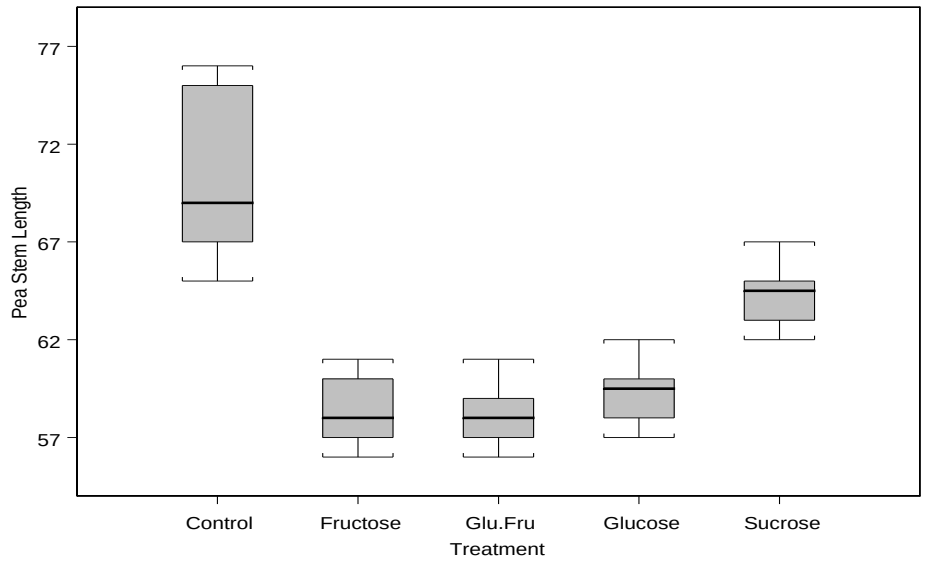



Figure 5.4: *Pea section data for five sugar treatments.*

- **Subjects**: Boards
- **Response variable**: Number of beetles trapped
- **Explanatory variable**: Color (blue, green, white, yellow). Color is also the treatment because it is the condition applied to the boards.

*NOTE*: In this experiment, we have no control group (what would that even be?) but we can **compare the colors** among themselves. In this case, the different colors act as controls for each other.

**Example 5.3.** *Does sugar affect the growth of pea stems?* The data in Table 5.6 are lengths of pea sections, in ocular units ($\times 0.114$ mm), grown in tissue culture with auxin (a plant hormone) present. The purpose of the experiment was to test the effects of the addition of various sugars on growth as measured by length. Pea plants were randomly assigned to one of five treatment groups:

Table 5.6: *Pea plant data.*

| | Control | 2% fru | 1%/1% g/f | 2% glu | 2% suc |
|---|---|---|---|---|---|
| | 75 | 58 | 58 | 57 | 62 |
| | 67 | 61 | 59 | 58 | 66 |
| | 70 | 56 | 58 | 60 | 65 |
| | 75 | 58 | 61 | 59 | 63 |
| | 65 | 57 | 57 | 62 | 64 |
| | 71 | 56 | 56 | 60 | 62 |
| | 67 | 61 | 58 | 60 | 65 |
| | 67 | 60 | 57 | 57 | 65 |
| | 76 | 57 | 57 | 59 | 62 |
| | 68 | 58 | 59 | 61 | 67 |
| Average | 70.1 | 58.2 | 58.0 | 59.3 | 64.1 |

1. control (no sugar added)
2. 2% fructose added
3. 1% glucose and 1% fructose added
4. 2% glucose added
5. 2% sucrose added

*DESIGN*: Ten observations were obtained for each group of plants. In all, 50 pea plants were used. How might we physically randomize the pea stems to the 5 treatments?

- **Subjects**: Pea stems
- **Response variable**: Growth
- **Explanatory variable**: Sugar content (control, fru, fru/glu, glu, suc). Sugar content is also the treatment because it is the condition applied to the pea stems.

## 5.3 How to experiment badly

*TERMINOLOGY*: A **lurking variable** is a variable that has an important effect on the relationship among the variables in the study but is not one of the explanatory variables studied.

*TERMINOLOGY*: 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.** In a clinical trial, physicians on a Drug and Safety Monitoring Board want to determine which of two drugs is more effective for treating HIV in its early stages. A sample 200 patients in the trial are randomly assigned to one of two treatment groups. After 6 weeks on treatment, the net CD4 count change is observed for each patient.

- Does gender have an effect on the CD4 count response? If it does, and we ignore it in the analysis of the data, then gender would be a **lurking variable**!
- If we give Drug 1 to only black subjects and Drug 2 to only white and Hispanic subjects, then race is **confounded** with the drug! That is, if we did see a difference between the drugs, we wouldn't know if it was because of the drugs or because of differences in the races!

## 5.4 Randomized comparative experiments

*EXAMPLES*: Examples 5.1-5.4 (notes) are all examples of **randomized comparative experiments**. In such experiments,

- **randomization** is used to assign subjects to treatments.
- two or more treatments are to be **compared**.

*GENERAL OUTLINE*: Here is the general outline of how a randomized comparative experiment is performed:

**General Outline of Experimentation**

Sample subjects from population

$\downarrow$

Randomize subjects to treatments

$\downarrow$

Record data on each subject

$\downarrow$

Analyze data

$\downarrow$

Make statement about the differences among treatments.

*IMPORTANCE*: **Randomization** produces groups of subjects that should be similar in all respects before we apply the treatments. Using impersonal chance to assign treatments tends to "average out" the effects of other (possibly confounding) variables.

*IMPORTANCE*: **Comparative design** ensures that influences other than the treatments operate equally on each group.

*RESULT*: In properly-designed randomized comparative experiments, differences in the response variable must be due to the effects of the treatments!

*GOAL*: Our goal as an investigator is to determine if these treatment differences are **real**. *That is, are we seeing differences in the response because the treatments are truly different, or, are we seeing treatment differences which could have arisen just by chance?*

*PREVIEW*: To answer the last question formally, we need to learn more about **statistical inference**.

## 5.5 Principles of experimental design

*IMPORTANT*: Here are the **three principles of experimental design** that we should keep in mind.

1. **Control** the effects of lurking variables on the response.
2. **Randomize** subjects to the different experimental treatments.
3. **Use enough subjects** in each group to reduce chance variation in the results. The more subjects we use, the more information we have!

**Example 5.5.** Ginkgo is gaining recognition as a brain tonic that enhances memory because of its positive effects on the vascular system, especially in the cerebellum. A new drug (Drug A, based on Ginkgo extract) has been developed to help people improve their memory. *Is this drug better than the current standard?* Suppose we tested each member of the treatment groups through a memory exercise (a list of 10 objects).

- **Situation #1:**
  - In Group A, 33% of the subjects remembered all items on the list.
  - In Group B, 32% of the subjects remembered all items on the list.

  Is this compelling evidence to show Drug A is better than Drug B?
- **Situation #2:**
  - In Group A, 59% of the subjects remembered all items on the list.
  - In Group B, 29% of the subjects remembered all items on the list.

  Is this compelling evidence to show Drug A is better than Drug B?

*TERMINOLOGY*: An observed effect so large that it would rarely occur by chance is called **statistically significant**.

# 6 Experiments in the Real World

Complementary reading from Moore and Notz: Chapter 6.

## 6.1 Equal treatment

*PREVAILING THEME*: The logic behind **randomized comparative experiments** is that all subjects are treated alike except for the treatments. Unequal treatment has the potential to cause bias!

**Example 6.1.** A group of psychologists carried out an experiment to determine if certain stimuli (e.g., shocks, etc.) had an effect on attitude and personality in college students. One hundred college freshmen were recruited; 20 students to each of 5 different stimulus groups. To conduct the experiment quickly, 5 different rooms were used; each stimulus was administered in a different room. Is there potential for unequal treatment?

*BLINDING*: When the subjects in an experiment are humans, it is good to use **blinding** (i.e., subjects do not know the identity of the treatment they have received). Many studies incorporate **double blinding**, where neither the subjects nor the researchers know which treatment has been given to subjects. This guards against researcher bias.

*REMARK*: Blinding and double blinding are very important in **clinical trials**!

- *The patient.* If the patient knows he/she is receiving a new treatment, this may confer a psychological benefit. The degree of this benefit depends on the type of disease and the nature of the treatments. Whether it is asthma, cancer, or heart disease, the manner in which patients are informed of therapy can have a profound effect in subsequent performance.
- *The treatment team.* Patients known to be receiving a new treatment may get treated differently than those on a standard treatment. Such differences in ancillary care may affect the eventual responses.

- *The evaluator*. It is especially important that the individuals evaluating response be objective. A physician who has some pre-conceived ideas of how a new treatment might work may introduce bias in his/her evaluation if they are aware of the patient's treatment. Those analyzing the data from the trial must also be objective.

*PLACEBOS*: Blinding treatments takes a great deal of care and planning. If we are giving pills, for example, we must make sure the pills are of the same size, color, taste, texture, etc. It has been well documented that there is a **placebo effect**; i.e., an apparent favorable/unfavorable response to a non-treatment!

- 42 percent of balding men maintain/increased amount of hair when taking an innocuous chemical composition 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!

*NOTE*: Although the principles of blinding are sound, they are sometimes not feasible. For example, if we are comparing surgery versus chemotherapy in a cancer clinical trial, there is no way to blind anyone.

## 6.2 Problems

*UNDERCOVERAGE*: **Undercoverage** is also an issue with experiments; e.g.,

- a clinical trial which excludes poor patients, those without insurance, etc.
- an agricultural experiment excluding pigs with certain diets, etc.
- a psychology experiment excluding engineering majors, etc.

*GOAL*: We want our conclusions to apply to an appropriate population of interest (e.g., USC students, market-bound pigs, American adults with lung cancer). If certain groups are not represented, our conclusions will be biased (or, at least, limited).

*NON-ADHERENCE*: This occurs when human subjects (mostly) 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.

*DROPOUTS*: When dealing with human subjects, some people decide that they no longer want to participate in the study. This may be due to undesirable side effects, lack of interest, or some other factor. **Dropouts** are subjects who begin the experiment but do not complete it.

- How should we handle dropouts? Simply discard their data? Did they drop out for a reason?
- Dropouts cause bias because we don't get to see how these subjects truly respond to the treatment.

*LACK OF REALISM*: **Lack of realism** occurs when we can not generalize our experimental results to that of a larger population. For example,

- using only plots of land with excellent soil composition properties
- using patients for which the disease has not progressed
- center-brake light example (see page 101-102, MN)
- psychological experiments with an irritating stimulus; patients know the experiment will be over soon.

**Example 6.2.** A new method has been developed to combat arthritis. Subjects can take a new drug (high versus low dosage) and at the same time receive massage therapy (traditional versus acupuncture). Researchers believe that efficacy should depend on the dose of the drug and the massage therapy. With 100 patients, explain how you would design an appropriate experiment to address the researchers' questions. What are the four possible treatment groups? What would you measure on each individual?

## 6.3 Experimental designs

### 6.3.1 Completely randomized designs

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

**Example 6.3.** One of the important characteristics in the pulp and paper industry is the brightness of the final product. In an experiment, engineers wanted to compare four bleaching chemicals and their effects on paper brightness (measured on a scale from 0-10). The engineers used 20 pulp specimens (selected at random). The 20 specimens were randomly assigned to the chemicals, 5 specimens to each chemical.

- **Subjects/Individuals**: Pulp specimens
- **Response variable**: Brightness rating
- **Treatment**: Bleaching chemical (also, an explanatory variable)

*NOTE*: What makes this a completely randomized design is the fact that individuals (pulp specimens) were **allocated at random** among all treatments (chemicals). How might we carry out the actual randomization? Could we use Table A? How?

### 6.3.2 Randomized block design

*TERMINOLOGY*: In a **randomized block design**, random assignment of subjects to treatments is carried out separately within each block. In the language of experiments, a **block** is a collection of individuals that are thought to be "more alike" in some way.

*ASIDE*: A block in an experiment may be thought of as a stratum in a sample survey (the same idea: individuals within a block/strata share inherent characteristics; e.g., gender, race, major, farm, etc.).

**Example 6.4.** An entomologist wants to determine if two preparations of a virus would produce different effects on tobacco plants. Consider the following experimental design:

- He took 10 leaves from each of 4 plots of land (so that there are 40 leaves in the experiment).
- For each plot, he randomly assigned 5 leaves to Preparation 1 and 5 leaves to Preparation 2.

*NOTE*: Randomization in this experiment is **restricted**; leaves were randomly assigned to treatments within each plot. In this experiment, the researcher wants to compare the two preparations. This experiment is an example of a **randomized block design** where

- **Subjects/Individuals**: Tobacco leaves
- **Response variable**: Amount of leaf damage
- **Treatment**: Preparations (also, an explanatory variable)
- **Block**: Plot of land (also, an explanatory variable)

*IMPORTANT*: Suppose that our researcher used a completely randomized design and simply randomized 20 leaves to each preparation. In this situation, he would have lost the ability to account for the possible differences among plots! The effects of the different preparations would be **confounded** with the differences in plots.

*MAIN POINT*: If there are differences among the blocks of subjects, then a completely randomized design will be worse when compared to the randomized block design. The former design ignores the differences among blocks; the latter design acknowledges it!

**Example 6.5.** In a clinical trial, physicians want to determine which of two drugs is more effective in treating HIV in its early stages. Patients in the trial are randomly assigned to one of two treatment (drug) groups. After 6 weeks on treatment, the net CD4 count change is observed for each patient. Consider the following three designs:

- **Design 1**: assign all the males to Drug 1, and assign all the females to Drug 2.
- **Design 2**: ignoring gender, randomize each individual to one of the two drugs.
- **Design 3**: randomly assign drugs within gender; that is, randomly assign the two drugs within the male group, and do the same for the female group.

*DISCUSSION*: Design 1 would be awful since if we observed a difference (a *significant* difference) between the CD4 counts for the two groups, we would have no way of knowing whether or not it was from the treatments (i.e., drugs) or from the differences in genders. In this situation, the drugs are **confounded** with gender! Design 2 (a completely randomized design) is much better than Design 1, but we still might not observe the differences due to drugs since differences in genders may not "average out" between the groups. However, Design 3 acknowledges that there may be an effect due to the different genders and incorporates that into the randomization. Design 3 is an example of an **randomized block design** with

- **Subjects/Individuals**: Patients with early-stage HIV
- **Response variable**: CD4 count
- **Treatment**: Drugs (also, an explanatory variable)
- **Block**: Gender (also, an explanatory variable).

### 6.3.3 Matched pairs design

**Example 6.6.** High blood pressure or hypertension is the most common cardiovascular disease and one of the greatest public health problems, affecting more than 60 million Americans. It directly contributes to the deaths of at least 250,000 people per year in the United States. Researchers are studying a certain stimulus thought to produce an increase in mean systolic blood pressure (SBP) in middle-aged men.

*DESIGN ONE*: Take a random sample of men and randomly assign each man to receive the stimulus (or not) using a **completely randomized design**. Here, the two groups of men can be thought of as "independent" since the groups contain different men.

*DESIGN TWO*: Consider an alternative design, the so-called **matched-pairs design**.

- Rather than assigning men to receive one treatment or the other (stimulus/no stimulus), obtain a response from each man under both treatments!
- That is, obtain a random sample of middle-aged men and take two readings on each man, one with the stimulus and one without the stimulus.
- Because readings of each type are taken on the same man, the difference between two readings on a given man should be less variable than the difference between a stimulus-response on one man and an no-stimulus-response on a different man.
- The man-to-man (i.e., subject-to-subject) variation inherent in the latter difference is not present in the difference between readings taken on the same subject!

*ADVANTAGE OF MATCHED PAIRS*: In general, by obtaining a pair of measurements on a single individual (e.g., man, rat, pig, plot, tobacco leaf, etc.), where one of the measurements corresponds to treatment 1 and the other measurement corresponds to treatment 2, you eliminate the subject-to-subject variability. Thus, you may compare the treatments (e.g., stimulus/no stimulus, ration A/ration B, etc.) under more **homogeneous** conditions where only variation within individuals is present (that is, the variation arising from the difference in treatments).

*A NOTE ON RANDOMIZATION*: In matched-pairs experiments, it is common practice, when possible, to **randomize** the order in which treatments are assigned. This may eliminate "common patterns" (that may confound our ability to determine a treatment effect) from always following, say, treatment A with treatment B. In practice, the experimenter could flip a fair coin to determine which treatment is applied first. If there are **carry-over effects** that may be present, these would have to be dealt with accordingly.

# 7 Data Ethics

Complementary reading from Moore and Notz: Chapter 7.

## 7.1 Introduction

*REMARK*: Usually, the goal of an experiment or observational study is to argue that a **hypothesis** is valid; e.g.,

- "Our new Drug B is superior to Drug A."
- "This brand of fertilizer is better than the competing fertilizers."
- "Democrats are less likely to oppose Roe v. Wade than Republicans."
- "Women in abusive relationships are more likely to suffer from depression."
- "For HIV patients, those without dental insurance are more likely to have unmet dental needs than those with private insurance."

*FACT*: Hypotheses may or may not be true, and the data from our investigations are used as evidence for or against them. If the data contradict a specific hypothesis put forth by an investigator, he/she might be viewed negatively. Thus, there is much at stake for the investigator in a statistical analysis!

*REMARK*: Because the investigation carries a lot with it (e.g., money, reputation, jobs, etc.), some investigators may try to falsify the data or use unethical practices; e.g.,

- remove records/individuals from the study; only retain those data which are favorable to one's hypothesis
- change one's hypothesis (or assumptions) during the study
- using randomization methods inappropriately; "cheating" on the randomization.

## 7.2 Ethical studies

*DATA ETHICS*: In large-scale studies which involve human subjects (e.g., clinical trials, public surveys, etc.),

- the group/organization which carries out the study must have an **institutional review board** (IRB) that reviews all studies (in advance and during) in order to protect the participants from harm.
- All individuals in the study must give their **informed consent** to participate.
- All individual data are to be kept **confidential**. Only statistical summaries (e.g., means, proportions, etc.) for groups of subject may be made public.

*INSTITUTIONAL REVIEW BOARDS*: In 1974, the U.S. Congress established the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research as part of the National Research Act. The Act required the establishment of the IRB for all research funded in whole, or in part, 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 (FDA).

- IRBs must have at least five members with expertise relevant to safeguarding the rights and welfare of subjects.
- At least one should be a scientist, one a non-scientist, and at least one must be unaffiliated with the institution/organization.
- The IRB should be made up of individuals with diverse racial, gender, and cultural backgrounds.

*NOTES*: IRBs approve human research studies that meet specific prerequisites.

- The risks to the study participants are minimized.

- The selection of study patients is equitable.
- Informed consent is obtained and documented for each participant.
- The privacy of the participants and confidentiality of the data are protected.

*TERMINOLOGY*: **Data confidentiality** is the protection of information provided by respondents and the assurance that information about individual respondents cannot be derived from the statistics reported.

*NOTE*: **Data anonymity** is the protection of the individual's data from all persons, even those that are involved with the study.

## 7.3 Randomized response

*HYPOTHETICALLY*: What would you do if someone asked you the following question in a survey:

*"Have you ever used illegal drugs for non-medicinal purposes?"*

*REALITY*: Questioning individuals about sexual orientation, criminal activity, abortion, or other sensitive topics, is a difficult task. The **randomized-response technique** can be an effective survey method to find such estimates because individual anonymity is preserved. Suppose that Tim is to be interviewed. Here is how it works:

- The interviewer presents Tim with two questions:
  - Question 1: "Have you ever smoked marijuana for non-medicinal purposes?"
  - Question 2: "Is the last digit of your SSN odd?"
- The interviewer tells Tim to roll a die, but not to reveal the outcome of the die to the interviewer.

  - If Tim rolls a 1, 2, 3, or 4, he is to answer yes/no to Question 1.
  - If Tim rolls a 5 or 6, he is to answer yes/no to Question 2.

- Tim then hands his yes/no response back to the interviewer.
- The interviewer records the yes/no response, but does not know which question Tim has answered.

*ESTIMATE*: If an SRS of $n$ individuals uses this strategy, we can find an estimate of $p$, the proportion of individuals who have used illegal drugs for non-medical purposes! **For the technique that we just outlined**, the formula for the estimate of $p$ is

$$\frac{3x}{2n} - \frac{1}{4},$$

where $n$ is the number of individuals and $x$ is the number of "yes" responses. Simple probability calculations (STAT 201) can be used to show that this formula is correct.

*NOTE*: The formula that we learned in Chapter 3 for margin of error; recall

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

does **not** apply to the randomized response technique! Thus, we can not make confidence statements for the population using this formula.

## 7.4 More about clinical trials

*TERMINOLOGY*: **Clinical trials** are experiments that study the effectiveness of medical treatments on actual patients. A clinical trial in the clearest method of determining whether an intervention has a postulated effect. It is easy for **anecdotal information** about the benefit of a therapy to be accepted and become standard. The consequences of not conducting appropriate trials can be serious and costly.

- Laetrile was rumored to be a "wonder drug" for cancer patients even though there was no evidence of biological activity. People were convinced that there was a

conspiracy in the medical community to keep the treatment from them! The NIH finally conducted a formal trial and demonstrated a lack of efficacy.

*REALITY*: A clinical trial involves **human subjects**. As such, we must be aware of ethical issues in the design and conduct of such experiments. Some ethical issues that need to be considered include the following:

- No superior alternative intervention is available for each subject.
- **Equipoise**. There should be genuine uncertainty about which trial intervention might be superior for each individual subject before a physician is willing to allow their patient to participate in such a trial.
- Exclude patients for whom risk/benefit ratio is likely to be too unfavorable.
- Exclude patients who have a very high risk of toxicity; e.g., if a drug may worsen an existing problem.
- Other reasons for exclusion (e.g., pregnant women if possibility of harmful effect to fetus, too sick to benefit, prognosis is good without intervention, etc.).

*AXIOM*: Clinical trials are ethical in the setting of uncertainty.

*THE FIRST CLINICAL TRIAL?*: In 1753, the following experiment was detailed by James Lind, a naval surgeon in the Mediterranean:

> "*I took 12 patients in the scurvy abroad the Salisbury at sea. The cases were as similar as I could have them.....they lay together in one place.....and had one common diet to them all.....To two of them were given a quart of cider a day, to two an elixir of vitriol, to two vinegar, to two oranges and lemons, to two a course of sea water, and to the remaining two the bigness of nutmeg. The most sudden and visible good effects were perceived from the use of oranges and lemons, one of those who had taken them being at the end of six days fit for duty.....the other was appointed nurse to the sick.*"

## 7.5 An unethical investigation

**Background:** Your thyroid gland is a small, butterfly-shaped gland located just below your Adam's apple. The thyroid produces hormones that affect your body's metabolism and energy level. Thyroid problems are among the most common medical conditions but, because their symptoms often appear gradually, they are commonly misdiagnosed. *Synthroid* is a thyroid hormone supplement used to treat people who do not produce enough thyroid hormone on their own. Synthroid helps to reduce the symptoms of low thyroid hormone such as weight gain, sensitivity to cold, lack of energy, and dry skin.

- made by Knoll Pharmaceuticals; one of the top 10 prescribed drugs in 1995.
- Knoll's enormous success with Synthroid has been entirely dependent on its continuing ability to convince users that the drug is worth the extra cost.
- this the company has done brilliantly for decades, despite any real proof of Synthroids superiority.

**Study:**

- A clinical pharmacist at UC San Francisco, named Betty Dong, published a limited study that strongly suggested Synthroid would beat out its competitors in a blinded randomized trial with three other competing drugs.
- In 1987, the company approached Dong, offering her the full $250,000 needed to pay for such a long and complex study.

**Result:**

- The study backfired on the company. To the surprise of nearly everyone, including Dong, the results suggested that Synthroid was no more or less effective than three much cheaper competitors!

**Catch:**

- As the study's sponsor, the company had not only been able to design the protocols of the drug trial; it also had exclusive access to the prepublished results of the study as well as final approval over whether the study could ever be made public.

**Actions:**

- With the results so threatening to its marketing efforts, the company set out to thwart the study.
  - Delayed publication in a scientific journal by many years, effectively destroying the relevance of its data.
  - The company preemptively published the UCSF data in a lesser-known journal with a different (much friendlier) conclusion.
  - The company waged a massive PR campaign against the real study written by Dong et al. in the *Journal of the American Medical Association*.

**Ramifications:**

- A massive class-action lawsuit followed the publication of Dong's *JAMA* paper, alleging on behalf of all Synthroid users that Knoll had defrauded them of hundreds of millions of dollars in inflated costs.
- The company has offered to settle for a sum close to $100 million, which would be the largest cash settlement for a class-action suit of its kind in history.

**Damage:**

- Even with such a fantastic price to pay, one can only conclude that in the end Knoll has benefited tremendously from its brash interference in the academic research process. One hundred million dollars is a small fraction of the profits the company made from Synthroid during the years it was suppressing the study. By its ability to taint Dong's study with controversy over the years, Knoll was able to nullify any would-be effect.

# 8 Measuring

Complementary reading from Moore and Notz: Chapter 8.

## 8.1 Introduction

*TERMINOLOGY*: We **measure** a property of a person or thing when we assign a number to represent the property. We use an **instrument** to make a measurement. Measurements will have **units** attached to them.

*EXAMPLES*:

- To measure the weight of pregnant mothers, we use a standard scale (instrument); our measurements are recorded in pounds; e.g., 156.3 lbs.
- To measure time off from work due to sickness for USC employees, we use a payroll database; our measurements are recorded in the number of days; e.g., 5.5 days.
- To measure the diameter of a plasma-punched hole in a metal slab, we use a hand-held caliper; our measurements are recorded in centimeters; e.g., 12.45 cm.
- To measure the scholastic aptitude of entering freshmen, we use SAT scores; our measurements are recorded in points; e.g., 1120 points.
- In a Senegalese study involving female prostitutes, to measure the rate of HIV infecteds per 1000 people, we use an enzyme-linked immunosorbant assay (ELISA) for each subject; our measurements are recorded as a rate; e.g., 94.7 infecteds/1000.

*DIFFICULTIES*: Some properties are difficult to measure!

- severity of depression
- amount of leisure time spent (what counts as leisure?)

- student performance in college
- unemployment rate
- intelligence
- worker motivation

*VALIDITY*: A variable is a **valid** measure of a property if it is relevant or appropriate as a representation of that property.

## 8.2 Rates

**Example 8.1.** The following table lists the number of sports-related injuries treated in U.S. hospital emergency rooms in 2001, along with an estimate of the number of participants (in thousands) in the sports:

| Sport | Injuries | # partic. | rate | Sport | Injuries | # partic. | rate |
|---|---|---|---|---|---|---|---|
| Basketball | 646,678 | 26,200 | 24.7 | Fishing | 84,115 | 47,000 | 1.8 |
| Bicycle riding | 600,649 | 54,000 | 11.1 | Horse riding | 71,490 | 10,100 | 7.1 |
| Base/softball | 459,542 | 36,100 | 12.7 | Skateboarding | 56,435 | 8,000 | 7.1 |
| Football | 453,684 | 13,300 | 34.1 | Ice hockey | 54,601 | 1,800 | 30.3 |
| Soccer | 150,449 | 10,000 | 15.0 | Golf | 38,626 | 24,700 | 1.6 |
| Swimming | 130,362 | 66,200 | 2.0 | Tennis | 29,936 | 16,700 | 1.8 |
| Volleyball | 129,839 | 22,600 | 5.7 | Ice skating | 29,047 | 7,900 | 3.7 |
| Roller skating | 113,150 | 26,500 | 4.3 | Water skiing | 26,633 | 9,000 | 3.0 |
| Weightlifting | 86,398 | 39,200 | 2.2 | Bowling | 25,417 | 40,400 | 0.6 |

QUESTION: If one uses the **number of injuries** as a measure of the hazardousness of a sport, which sport is more hazardous between bicycle riding and football? between soccer and ice hockey? between swimming and skateboarding?

*CALCULATIONS*: Here is how injury rate was computed:

$$\text{rate} = \frac{\#\text{ injuries}}{\#\text{ participants}}$$

Since the number of participants is recorded in 1000's, this rate denotes the **injury rate per 1000 participants**!

QUESTION: In terms of the **injury rate per 1000 participants**, which sport is more hazardous between bicycle riding and football? between soccer and ice hockey? between swimming and skateboarding?

*MORAL*: Understand what it is being measured!! The injury rate per 1000 participants is a more valid measure of a sports hazardousness than the number of injuries. The latter variable excludes how many people participate in these sports!!

*TERMINOLOGY*: Often, a **rate** (a fraction, proportion, or percent) at which something occurs is a more valid measure than a simple **count** of occurrences.

## 8.3 Predictive ability

*TERMINOLOGY*: A measurement of a property has **predictive validity** if it can be used to predict success on tasks that are related to the property being measured. How would you rate the predictive validity in the following situations?

- Mammogram image $\longrightarrow$ breast cancer?
- SAT score $\longrightarrow$ success in college?
- Sexual practices $\longrightarrow$ HIV positive?
- Gun control $\longrightarrow$ increase in crime rate?
- Number of fights with partner $\longrightarrow$ problems with depression?

## 8.4 Measurement error

**Example 8.2.** In 30 seconds, read the following passage and count the number of "F's."

> "THE NECESSITY OF TRAINING HANDS FOR FIRST-CLASS FARMS IN THE FATHERLY HANDLING OF FRIENDLY FARMS LIVESTOCK IS FOREMOST IN THE MINDS OF 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 OWNERS OF THE FARMS FEEL THEY SHOULD CARRY ON WITH THE FAMILY TRADITION OF TRAINING FARM HANDS OF FIRST-CLASS FARMS IN THE FATHERLY HANDLING OF FARM LIVESTOCK BECAUSE THEY BELIEVE IT IS THE BASIS OF GOOD FUNDAMENTAL FARM EQUIPMENT."

Number of F's on 1st reading: _____

Number of F's on 2nd reading: _____

*TERMINOLOGY*: A measurement process has **bias** if it systematically overstates or understates the true value of the property it measures.

*TERMINOLOGY*: A measurement process has **random error** if repeated measurements on the same individual give different results. If random error is small, the measurement process is **reliable**. If a measurement process is not reliable, then our data are not either!

*CONCEPTUALIZATION*: The measurements that we take on any individual obey the following conceptual formula:

$$\text{measured value} = \text{true value} + \text{bias} + \text{random error}.$$

*MORAL*: There is no such thing as a perfectly reliable measurement instrument; i.e., there will always be random error! However, we would like to use an instrument which does not have bias.

*WEIGHT-SCALE ANALOGY*: See pages 141-142 MN.

## 8.5 Likert scales

*TERMINOLOGY*: A **Likert scale** (named after Rensis Likert; invented 1932) is a rating scale measuring the strength of agreement with a clear statement. Often, a Likert scale is used in the form of a questionnaire used to gauge attitudes or reactions. Response choices almost always have "roughly equal intervals" with respect to agreement.

**Example 8.3.** The following statement appeared recently in an employee satisfaction survey:

*My job provides the opportunity for independent action.*

| Strongly Disagree | Disagree | Slightly Disagree | Undecided | Slightly Agree | Agree | Strongly Agree |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |

*NOTE*: The Likert scale is a commonly-used measurement scale, but there are a few issues that always arise with using it; e.g.,

- How many response choices should be included?
- Which number should be the "midpoint label?" Should it be included?
  - Subjects will often choose the middle category when it is offered.
- How strong is the statement? Statements that are too mild may generate some "yea-saying;" harsh statements might generate stronger opposing views.
  - *"Physicians generally ignore what patients say."*
  - *"Sometimes, physicians do not pay as much attention as they should to patients' comments."*
  - *"Once in a while, physicians might forget or miss something that a patient has told them."*

# 9 Do the Numbers Make Sense?

Complementary reading from Moore and Notz: Chapter 9.

*UNFORTUNATE REALITY*: Statistics has a public relations problem, and people often perceive statistics as a "fishy science." In this chapter, we'll talk about some reasons why.

*SIX GOOD QUESTIONS*: Here are 6 questions to consider when looking at a report of statistics. Answer these questions before you believe what is being said.

- What didn't they tell us?
- Are the numbers consistent with each other?
- Are the numbers plausible?
- Are the numbers too good to be true?
- Is the arithmetic right?
- Is there a hidden agenda?

## 9.1 What didn't they tell us?

**Example 9.1.** A news report on snowstorms says, *"A winter storm spread snow across the area, causing 28 minor traffic accidents."* What aren't they telling us? How many traffic accidents happen on days with good weather? Did the snowstorm actually prevent more accidents?

**Example 9.2.** *Universities and their average SAT scores.* Northeastern University leaves out international and "remedial" students! What does this do to the average?

**Example 9.3.** *Chicago Cubs.* When comparing numbers over time, you can slant the comparison by choosing your starting point. Suppose the track record for the Chicago

Cubs over a 10 game sequence is L, L, L, L, L, W, W, W, W, L (L=lose, W=win). Here are two statements, both of which are correct!

1. "'The Cubs are hobbling into tonight's game, having lost 6 of their past 10 games."
2. "The Cubs are the division's hottest team, having won 4 of their last 5 games."

**Example 9.4.** *"In a survey, four out of five dentists agree that...."*

## 9.2 Are the numbers consistent with each other?

**Example 9.5.** A researcher was performing an experiment on 6 sets of 20 animals each. He reported 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 particular experiment in mind? Exactly how can one get 53 percent successes out of 20 animals?

## 9.3 Are the numbers plausible?

**Example 9.6.** The (very reputable) journal *Science* reported a California field produces 750,000 melons per acre. Well, there are 43,560 square feet in an acre, so this means

$$\text{melons/square foot } = \frac{750,000}{43,560} = 17.2.$$

*Does this even make sense?* The editor later reported that the correct figure was about 11,000 melons per acre.

## 9.4 Are the numbers too good to be true?

*MISLEADING THE PUBLIC*: Writers and reporters will often report only those cases/statistics which are favorable to their point of view. Newspaper editors usually

have no idea what is valid and what is not.

*"Torture numbers, and they'll confess to anything."*

**Example 9.7.** Do 41 percent of Louisiana biology teachers really reject evolution? The *New York Times* editors believed (or wanted to believe) so; see the recent *Chance* article.

## 9.5 Is the arithmetic right?

**Example 9.8.** In a paper submitted by a student for a class project (at a public university in California), the student mentions the following: "Who uses the gym? Here are some relevant statistics: 42% Men; 40% Women; 18% Other."

**Example 9.9.** Poll results for the survey question: "Do you approve of the current bond measure? Results: 52% Yes; 44% No; 15% Undecided."

*PERCENTAGE CHANGE*: A statistic often quoted is the **percentage change**. This quantity is found by

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

**Example 9.10.** Monthly sales were \$21.58 million in November and \$23.90 million in December. What is the percentage change in sales?

$$\begin{aligned} \text{percent change} &= \frac{23.90 - 21.58}{21.58} \times 100\% \\ &= 10.8\% \end{aligned}$$

That is, their was a 10.8% increase in sales between November and December.

*NOTE*: When reporting percent changes, those who lie with statistics might use the divisor that makes their point. To make a percent change appear bigger, they will use the smaller divisor. To make the percent change appear smaller, they will use the larger

divisor. In the last example, if we erroneously compute the percentage change as

$$\begin{aligned} \text{percent change} &= \frac{23.90 - 21.58}{23.90} \times 100\% \\ &= 9.7\%, \end{aligned}$$

the result looks less than the real percentage change of 10.8%!

*NOTE*: 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.

*NOTE*: Nothing can decrease by more than 100%. If a quantity has lost 100% of its value, it has lost it's entire value! A statement like *"this mouthwash reduces plaque on teeth by 200%"* makes no sense.

**Example 9.11.** 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."

QUESTION: What is the distance from Charleston to Seattle? (about 3,000 miles).

## 9.6 Is there a hidden agenda?

*MY OPINION*: Too many people use statistics to "drive their agenda" and lack the patience and perspicacity to understand what the data are actually saying.

**Example 9.12.** The following excerpt recently appeared the news:

> *"It is estimated that disposable diapers account for less than 2% of the trash in today's landfills. In contrast, beverage containers, third-class mail and yard wastes are estimated to account for about 21% of the trash in landfills. Given this, in your opinion, would it be fair to ban disposable diapers?"*

Is it fair to compare the 2 and 21 percent statistics (provided, of course, that these are even accurate)? Why or why not?

## 9.7 Top 10 List: Favorite statistics quotes

10. *"Statistics class is a lot like church—many attend but few understand."*

9. *"People can come up with statistics to prove anything...forty percent of all people know that."*

8. *"There are two kinds of statistics, the kind you look up, and the kind you make up."*

7. *"The statistics on sanity are that one out of every four Americans is suffering from some form of mental illness. So, think of your three best friends. If they're okay, then it's you."*

6. *"Statistics are like lampposts: they are good to lean on, but they don't shed much light."*

5. *"Like other occult techniques of divination, the statistical method has a private jargon deliberately contrived to obscure its methods from non-practitioners."*

4. *"Smoking is one of the leading causes of statistics."*

3. *"A statistician is a man who believes figures don't lie, but admits that under analysis some of them won't stand up either."*

2. *"If you want to inspire confidence, give plenty of statistics. It does not matter that they should be accurate, or even intelligible, as long as there is enough of them."*

1. *"There are three kinds of lies: lies, damned lies, and statistics."*

# 10 Graphs, Good and Bad

Complementary reading from Moore and Notz: Chapter 10.

## 10.1 Types of variables

*RECALL*: A **variable** is a characteristic (e.g., temperature, age, race, CD4 count, growth, education level, etc.) that we would like to measure on individuals. The actual measurements recorded on individuals in the sample are called **data**.

*TWO TYPES*: **Quantitative** variables have measurements (data) on a numerical scale. **Categorical** variables have measurements (data) where the values simply indicate group membership.

**Example 10.1.** Which of the following variables are quantitative in nature? Which are categorical?

- IKEA-Atlanta daily sales (measured in $1000's)
- store location (Baltimore, Atlanta, Houston, Detroit, etc.)
- CD4 cell count
- yield (bushels/acre)
- payment times (in days)
- payment times (late/not late)
- age
- advertising medium (ratio/TV/internet)
- number of cigarettes smoked per day
- smoking status (yes/no).

## 10.2 Graphs for categorical variables

*TERMINOLOGY*: The **distribution** of a variable tells us what values the variable takes and how often it takes these values.

*IMPORTANT*: Presenting data effectively is an important part of any statistical analysis. How we display data distributions depends on the type of variable(s) or data that we are dealing with.

- **Categorical**: pie charts, bar charts, tables
- **Quantitative**: stemplots, boxplots, histograms, timeplots

*UNDERLYING THEMES*: Remember that the data we collect may be best viewed as a **sample** from a larger population of individuals. In this light, we have two primary goals in this section:

- learn how to summarize and to display the **sample** information, and
- start thinking about how we might use this information to learn about the underlying **population** distribution.

**Example 10.2.** HIV infection has spread rapidly in Asia since 1989, partly because blood and plasma donations are not screened regularly before transfusion. To study this, researchers collected data from a sample of 1390 individuals from villages in rural eastern China between 1990 and 1994 (these individuals were likely to donate plasma for financial reasons). One of the variables studied was **education level**. This was measured as a categorical variable with three categories (levels): illiterate, primary, and secondary.

*TABLES*: I think the easiest way to portray the distribution of a categorical variable is to use a **table** of counts and/or percents. A table for the education data collected in Example 10.2 is given in Table 10.7.

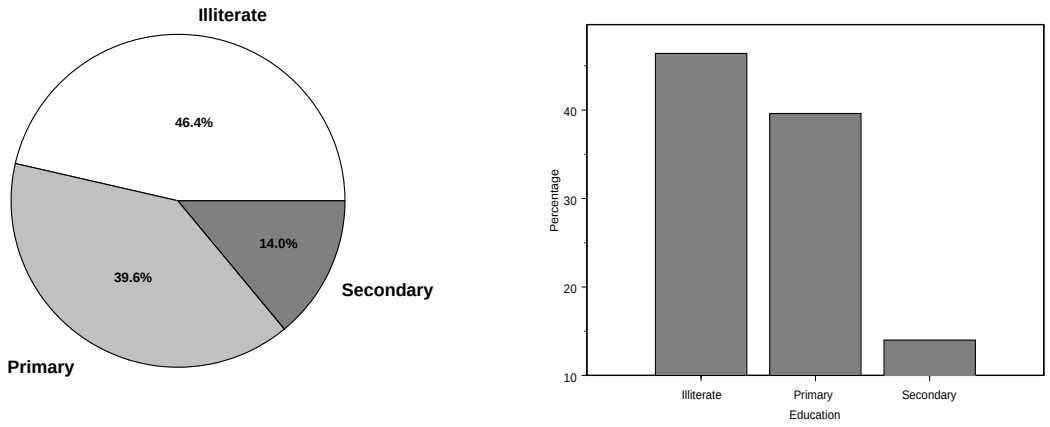
Table 10.7: *Education level for plasma donors in rural eastern China between* 1990-1994.

| Education level | Count | Percentage |
|---|---|---|
| Illiterate | 645 | 46.4 |
| Primary | 550 | 39.6 |
| Secondary | 195 | 14.0 |
| Total | 1390 | 100.0 |

*REMARK*: Including percentages in the table (in addition to the raw counts) is helpful for interpretation. Most of us can understand percentages easily. Furthermore, percentages put numbers like "645" into perspective; e.g., 645 out of what?

*INFERENCE*: These data are from a sample of rural villagers in eastern China. From these data, what might we be able to say about the entire population of individuals?

*PIE CHARTS AND BAR GRAPHS*: **Pie charts** and **bar graphs** are appropriate for categorical data but, unlike tables, are more visual in nature.




(a) Pie chart. (b) Bar graph.

Figure 10.5: *Education level for plasma donors in rural eastern China between* 1990-1994.

*QUESTIONS*: Which display do you like better? Is there anything dissatisfying with the bar graph?

## 10.3 Line graphs

*LONGITUDINAL DATA*: In many applications, especially in business, data are observed **over time**. Data that are observed over time are sometimes called **longitudinal data**. More often, longitudinal data are quantitative (but they need not be). Examples of longitudinal data include monthly sales, daily temperatures, hourly stock prices, etc.

*TERMINOLOGY*: If it is the longitudinal aspect of the data that you wish to examine, you need to use a graphical display which exploits this aspect. A **line graph** (or **time plot**) of a variable plots each observation against the time at which it was measured. To construct a line graph, simply plot the individual values (on the vertical axis) versus time (on the horizontal). Individual values are then connected with lines.

**Example 10.3.** The Foster Brewing Company is largely responsible for the development of packaged beers in Australia. In fact, some of the first canned American beers were first produced in Australia in the early 1950s. In the last 50 years, improved engineering techniques have led to larger vessels and improved productivity. The data are available online at

`http://www.maths.soton.ac.uk/teaching/units/math6011/mwhdata/Beer2.htm`

are the monthly beer sales data in Australia from January 1991 to September 1996. A time plot of the data appears in Figure 10.6.

*INTERPRETING TIME PLOTS*: When I look at time plots, I usually look for two things in particular:

- Increasing or decreasing **trends**. Is there a general shift over time upward or downward? Is it slight or notably apparent?
- Evidence of **seasonal effects**. Are there repeated patterns at regular intervals? If there is, what is most likely to have produced this pattern?

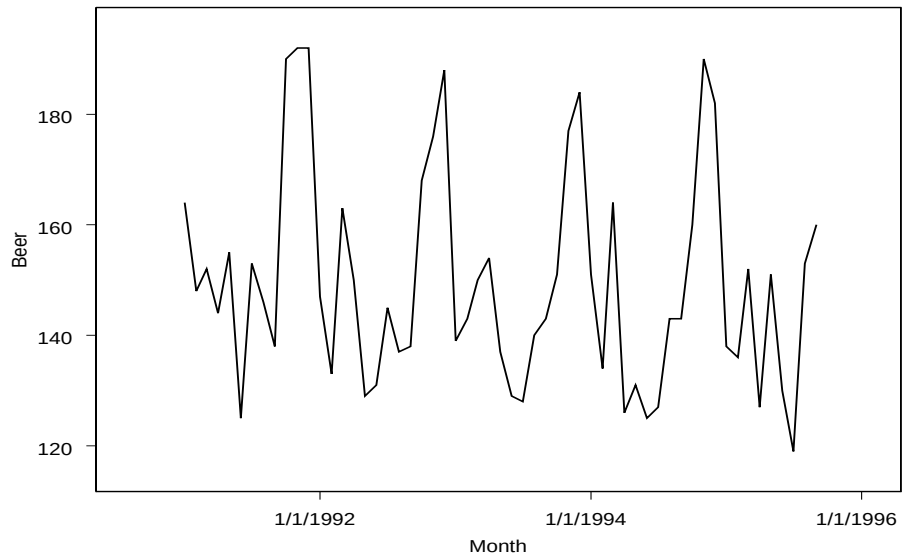



Figure 10.6: *Australian beer sales from January* 1991 *to September* 1996.

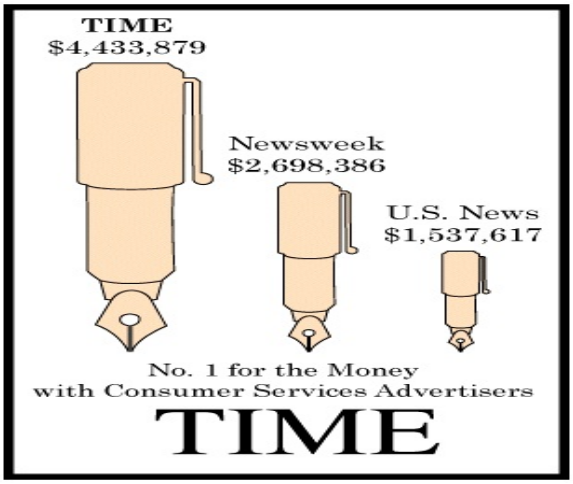
*USEFULNESS*: Analyzing longitudinal data is important for **forecasting** or **prediction**. For example, can we forecast the next two years of beer sales? Why might this information be important?

## 10.4 Bad/misleading graphs

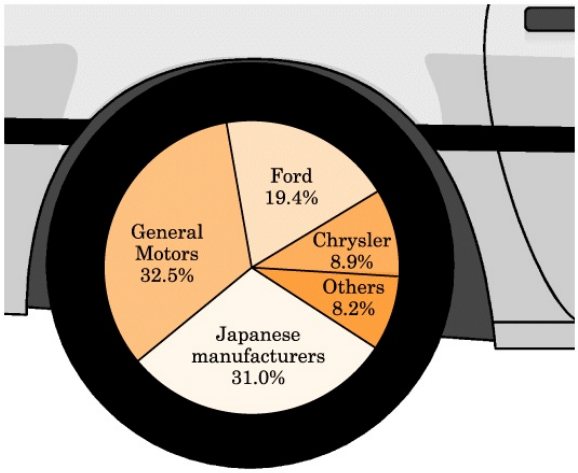
*MISCONCEPTION*: Statistics is a boring topic so graphs should be "souped up" to attract readers.

*CONSEQUENCE*: This practice leads to graphs that are misleading and confusing.

- Graphical displays that display **chartjunk** hinder the main message; readers address the design and not the substance!
- For examples of chartjunk, pick up any copy of *USA Today* (they are the worst). See also the next page!!

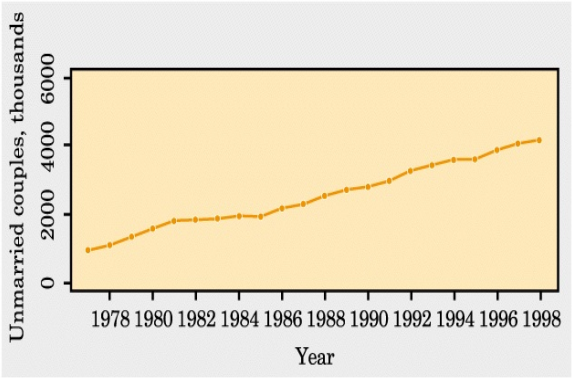



(a) Advertising figures.

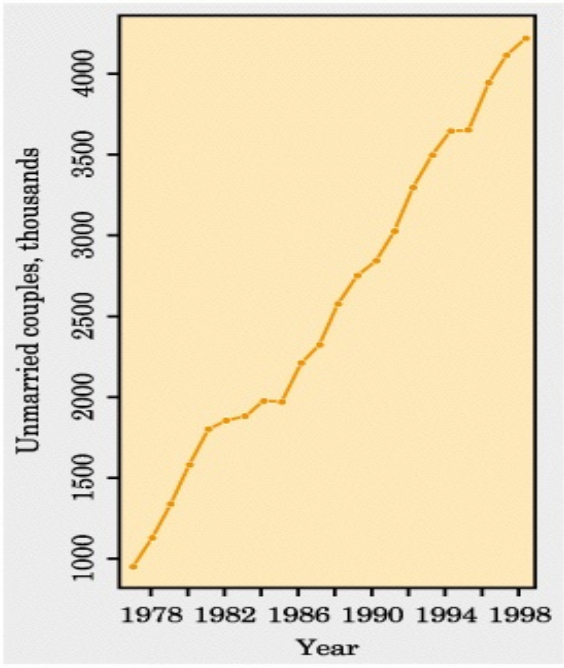



(b) Car sales.

Figure 10.7: *Examples of chartjunk.*

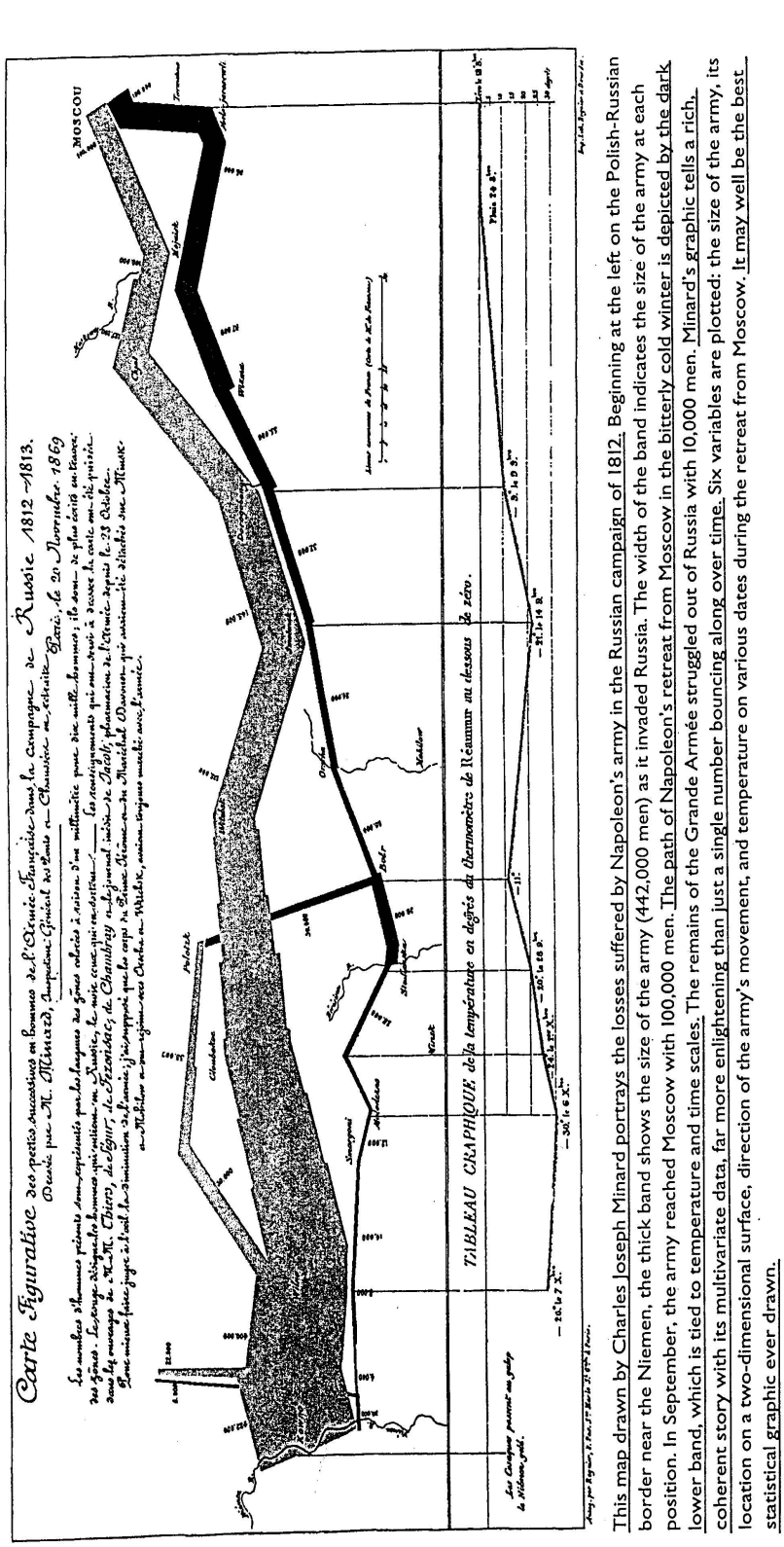



(a) Scale: 0-6000 thousand.




(b) Scale: 1000-5000 thousand.

Figure 10.8: *Number of unmarried couples; illustrates the effects of using different scales.*




This map drawn by Charles Joseph Minard portrays the losses suffered by Napoleon's army in the Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Niemen, the thick band shows the size of the army (442,000 men) as it invaded Russia. The width of the band indicates the size of the army at each position. In September, the army reached Moscow with 100,000 men. The path of Napoleon's retreat from Moscow in the bitterly cold winter is depicted by the dark, lower band, which is tied to temperature and time scales. The remains of the Grande Armée struggled out of Russia with 10,000 men. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. It may well be the best statistical graphic ever drawn.

Figure 10.9: *The best graphical display ever made?*

# 11 Displaying Distributions with Graphs

Complementary reading from Moore and Notz: Chapter 11.

## 11.1 Introduction

*REMARK*: In the last chapter, we discussed graphical displays for categorical data (**pie charts** and **bar graphs**). In this chapter, we focus primarily on graphical displays for use with quantitative data. In particular, we discuss

- Histograms
- Stem plots

*NOTE*: Both of these displays will help us graphically portray the **distribution** of quantitative data.

## 11.2 Histograms

**Example 11.1.** Monitoring the shelf life of a product from production to consumption by the consumer is essential to ensure the quality of a product. A sample of $n = 25$ cans of a beverage were used in an industrial experiment that examined the beverage's shelf life, measured in days (clearly, shelf life is a **quantitative variable**). The data collected are given in Table 11.8.

Table 11.8: *Beverage shelf life data.*

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 262 | 188 | 234 | 203 | 212 | 212 | 301 | 225 | 241 | 211 | 231 | 227 | 217 |
| 252 | 206 | 281 | 251 | 219 | 268 | 231 | 279 | 243 | 241 | 290 | 249 | |

*TERMINOLOGY*: In order to construct a histogram, we first construct a frequency table. A **frequency table** simply summarizes quantitative data in a tabular form. Included are two things:

- **class intervals**: intervals of real numbers
- **frequencies**: how many observations fall in each interval.

Table 11.9: *Frequency table for the shelf life data in Example* 11.1.

| Class Interval | Frequency |
|---|---|
| $175 \leq$ days $< 200$ | 1 |
| $200 \leq$ days $< 225$ | 7 |
| $225 \leq$ days $< 250$ | 9 |
| $250 \leq$ days $< 275$ | 4 |
| $275 \leq$ days $< 300$ | 3 |
| $300 \leq$ days $< 325$ | 1 |

*NOTE*: *There is no one right way to choose the class intervals!*

- Different choices will lead to different-looking histograms!
- Class intervals should all have the same width!
- Make sure that all data values are included!
- The number of class intervals should be large enough that not all observations fall in one or two intervals, but small enough so that we don't have each observation belonging to its own interval.

*HISTOGRAMS*: To construct a **histogram**, all you do is plot the frequencies on the vertical axis and the class intervals on the horizontal axis. The histogram for the shelf life data in Example 11.1 is given in Figure 11.10.

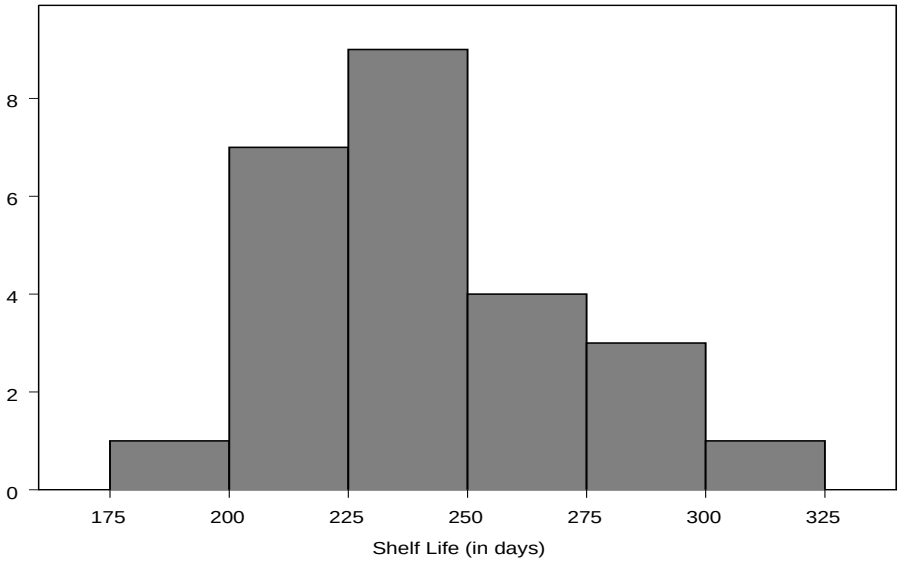



Figure 11.10: *Histogram for the shelf life data in Example* 11.1.

*GOALS*:

- The goal of a graphical display is to provide a **visual impression** of the characteristics of the data from a sample. The hope is that the characteristics of the **sample** are a likely indication of the characteristics of the population from which it was drawn.
- When we examine histograms (or stem plots), we will be always be interested in the overall pattern and any striking deviations from that pattern; in particular:
  - **center** of the distribution of data
  - **spread** (variation) in the distribution of data
  - **shape**: is the distribution symmetric or skewed? **Symmetric distributions** are those whose left and right-hand sides look like mirror images of one another (perfect symmetry is a rarity in real life).
  - the presence of **outliers** (i.e., "unusual observations").

*REMARK*: Here is how we would interpret the distribution of the shelf life data.

- **Center:** The center looks to be around 240 days.
- **Spread:** There is quite a bit of variability in the shelf lives; the lives ranging from 188 to 301 days.
- **Shape:** The shape of the distribution looks **approximately symmetric** (or maybe slightly skewed right).
- **Outliers:** There does not appear to be any gross outliers.

*NOTES*:

- *We can use the histogram to estimate what percentage of cans have shelf life in a certain range of interest.* Suppose that the experimenter believed "most shelf lives should be larger than 250 days." From the distribution, we see that this probably is **not** true if these data are representative of the population of shelf lives.
- We can also associate the percentage of lives in a certain interval as being proportional to the **area** under the histogram in that interval.
- For example, are more cans likely to have shelf lives of 200-225 days or 300-325 days? We can estimate these percentages by looking at the data graphically.

*TERMINOLOGY*: If a distribution is not symmetric, it is said to be **skewed**.

- A distribution is **skewed to the right** if there is a long right tail.
- A distribution is **skewed to the left** if there is a long left tail.

*SKEWED DISTRIBUTIONS*: Skewed distributions occur naturally in many applications. Not all distributions are symmetric or approximately symmetric! In Figure 11.11, the left distribution is **skewed right**; the right distribution is **skewed left**.

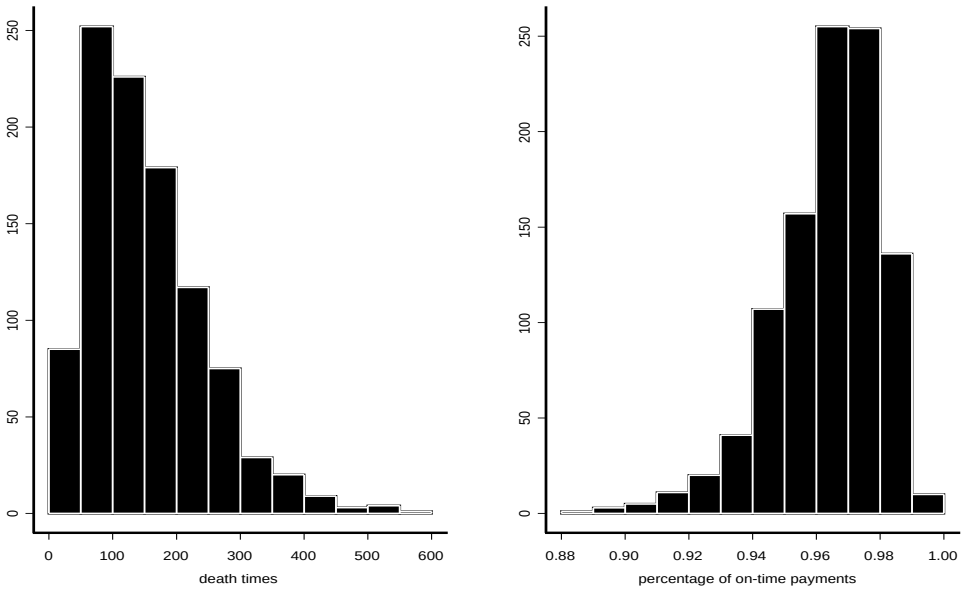



Figure 11.11: *Left: Death times for rats treated with a toxin. Right: Percentage of monthly on-time payments.*

## 11.3 Stem plots

*STEM PLOTS*: **Stem plots** provide a display of the distribution while retaining the numerical values themselves. The idea is to separate each data value into a **stem** and a **leaf**. Stem plots work well with small-to-moderate-sized data sets, say, 15 to 50 observations.

*DIRECTIONS*: To create a stem plot for this data (see textbook for full instructions).

1. Separate each of the observations into 2 parts: the rightmost digit (the leaf) and the remaining digits (the stem). **Rounding** may be necessary.

2. Vertically list the levels of the stem in ascending order (smallest on top) and draw a vertical line to its right.

3. Write each leaf to the right of the appropriate stem. For a given stem, the leaves should be in ascending order from left to right.

Table 11.10: *Stem plot for the shelf life data in Example* 11.1.

| | |
|---|---|
| **18** | 8 |
| **19** | |
| **20** | 3 6 |
| **21** | 1 2 2 7 9 |
| **22** | 5 7 |
| **23** | 1 1 4 |
| **24** | 1 1 3 9 |
| **25** | 1 2 |
| **26** | 2 8 |
| **27** | 9 |
| **28** | 1 |
| **29** | 0 |
| **30** | 1 |

*SHELF LIFE DATA*: The stem plot for the shelf life data in Example 11.1 appears in Table 11.10. In this plot, the **units digit** is the leaf; the **tens and hundreds digits** form the stem.

EXERCISE: Here are final exam scores from a course I taught in Summer, 2003 (Introduction to Business Statistics) at Oklahoma State. Make a **histogram** and a **stem plot** for these data.

93 92 88 79 63 55 90 81 87 79 87 58 66 80 77 59

81 81 78 79 68 76 79 76 73 81 78 64 71 47 70 65

- For the histogram, use the class intervals, 40-50, 50-60, 60-70, 70-80, 80-90, and 90-100.
- For the stem plot, use the units digit as the leaf; use the tens digit as the stem.
- Describe the distribution of these scores (talk about things like center, variability, shape, presence of outliers, etc.).

# 12 Describing Distributions with Numbers

Complementary reading from Moore and Notz: Chapter 12.

*REMARK*: In the last chapter, our main goal was to describe quantitative data distributions **graphically**. We now wish to do this **numerically**. For the remainder of the course, we will adopt the following notation to describe a sample of **quantitative** data.

$$n = \text{number of observations in sample} \qquad x = \text{variable of interest}$$

$$x_1, x_2, ..., x_n = \text{the } n \text{ data values in our sample}$$

*PREVAILING THEME*: Whenever we examine a sample of data, our goal is to **numerically summarize** the distribution of the sample and get an idea of these same notions for the population from which the sample was drawn.

## 12.1 Median, quartiles, and boxplots

### 12.1.1 Median

**Example 12.1.** In Example 11.1 (notes), we looked at an engineering process which produced the following quantitative data (beverage shelf lives, measured in days).

Table 12.11: *Beverage shelf life data.*

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 262 | 188 | 234 | 203 | 212 | 212 | 301 | 225 | 241 | 211 | 231 | 227 | 217 |
| 252 | 206 | 281 | 251 | 219 | 268 | 231 | 279 | 243 | 241 | 290 | 249 | |

*GOAL*: Find the **median** for these data. Informally, the median is the middle ordered value (when the data have been ordered from low to high).

*FORMALLY*: The **median** is the midpoint of a distribution; i.e., the observation such that half of observations are smaller and the other half are larger.

*COMPUTATION*: Here are the steps we take for computing the median, $M$.

1. Order the data from low to high.
2. Compute the **median position location**; i.e., compute

$$\text{median position location} = \frac{n+1}{2}.$$

3. The median of the data is given by the ordered value in this position.
   - If $n$ is odd, the median position location will be a whole number, say, 8; in this case, the median would be the 8th ordered observation.
   - If $n$ is even, the median position location will be a fraction, say, 8.5; in this case, the median would be the (arithmetic) average of the 8th and 9th ordered observations.

**Example 12.2.** Here are the final exam scores for a class that I taught last semester:

```
96   92   84   77   74   84   80   74
```

Here, $n = 8$ (the number of observations). **First**, we order the data from low to high:

```
74   74   77   80   84   84   92   96
```

The **median position location** is

$$\text{median position location} = \frac{n+1}{2} = \frac{8+1}{2} = 4.5.$$

Thus, the median is the average of the 4th and 5th ordered values; i.e., the median is

$$M = \frac{80+84}{2} = 82.$$

*REMARK*: Note that exactly half of the scores are below 82, and half of the scores are above 82. This is what the median does; it is the value which "halves" the data.

**Example 12.1** (continued): In the last chapter, we constructed the stem plot for the beverage shelf life data. What is the median shelf life?

Table 12.12: *Stem plot for the shelf life data in Example* 12.1.

| | |
|---|---|
| **18** | 8 |
| **19** | |
| **20** | 3 6 |
| **21** | 1 2 2 7 9 |
| **22** | 5 7 |
| **23** | 1 1 4 |
| **24** | 1 1 3 9 |
| **25** | 1 2 |
| **26** | 2 8 |
| **27** | 9 |
| **28** | 1 |
| **29** | 0 |
| **30** | 1 |

### 12.1.2 Quartiles

*QUARTILES*: We have seen that the median $M$ is the value which "halves" the data (the lower half and the upper half). Informally, the first quartile is the median of the lower half; similarly, the third quartile is the median of the upper half.

*FORMALLY*: To calculate the **quartiles**, we do the following:

1. Arrange data from low to high and calculate the median $M$.
   - Identify the **lower half** of the data ($M$ excluded)
   - Identify the **upper half** of the data ($M$ excluded)
2. The **first quartile**, $Q_1$, is the median of the lower half of the data.
3. The **third quartile**, $Q_3$ is the median of the upper half of the data.

**Example 12.1** (continued): The median of the beverage shelf life data is

$$M = 234.$$

The **lower half** of the beverage data is

188 203 206 211 212 212 217 219 225 227 231 231

(note how we have excluded the median $M = 234$ in presenting the lower half). The **first quartile** $Q_1$ is the median of this lower half; i.e.,

$$Q_1 = \frac{212 + 217}{2} = 214.5.$$

The **upper half** of the beverage data is

241 241 243 249 251 252 262 268 279 281 290 301

(note how we have excluded the median $M = 234$ in presenting the upper half). The **third quartile** $Q_1$ is the median of this upper half; i.e.,

$$Q_3 = \frac{252 + 262}{2} = 257.$$

### 12.1.3 Five Number Summary and boxplots

*TERMINOLOGY*: The **Five Number Summary** of a data distribution consists of the following statistics:

1. **Min**, the smallest observation
2. $Q_1$
3. $M$
4. $Q_3$
5. **Max**, the largest observation.

*REMARK*: The Five Number Summary offers a helpful **numerical description** of a data set. The median is a measure of center; the Min and Max summarize the smallest and largest observations (so you can get an idea of how spread out the distribution is); the difference in the quartiles gives you an idea of how spread out the middle 50 percent of the distribution is.

**Example 12.1** (continued). For the beverage shelf life data in Example 12.1, the Five Number Summary is

$$\text{Min} = 188 \qquad Q_1 = 214.5 \qquad M = 234 \qquad Q_3 = 257 \qquad \text{Max} = 301$$

*TERMINOLOGY*: A **boxplot** is a graph of the Five Number Summary!

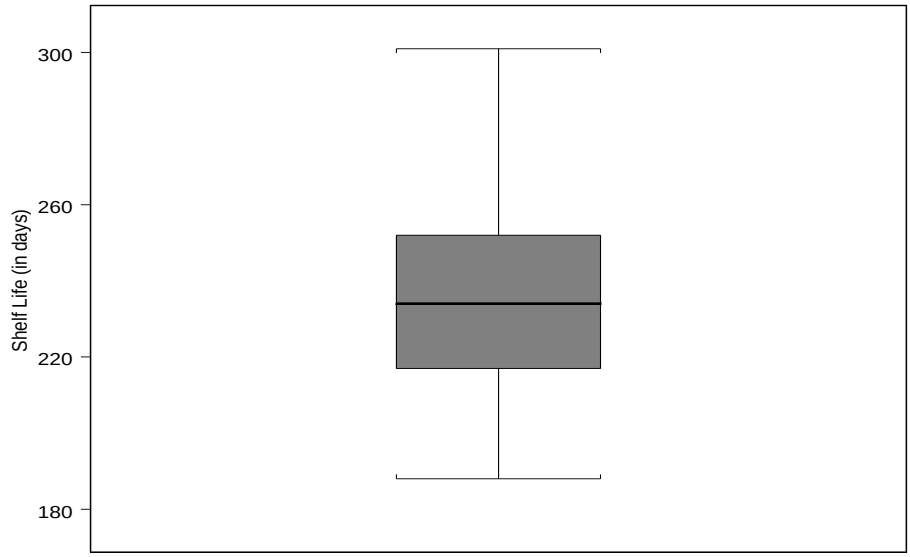



Figure 12.12: *Box plot for shelf life data in Example* 12.1.

*REMARK*: As a practicing statistician, I am a big fan of boxplots, especially when they are used to compare different data distributions. We have seen examples of **side-by-side boxplots** in earlier chapters (see also next page).

QUESTION: What would the boxplot look like for a distribution which was perfectly symmetric? skewed right? skewed left?

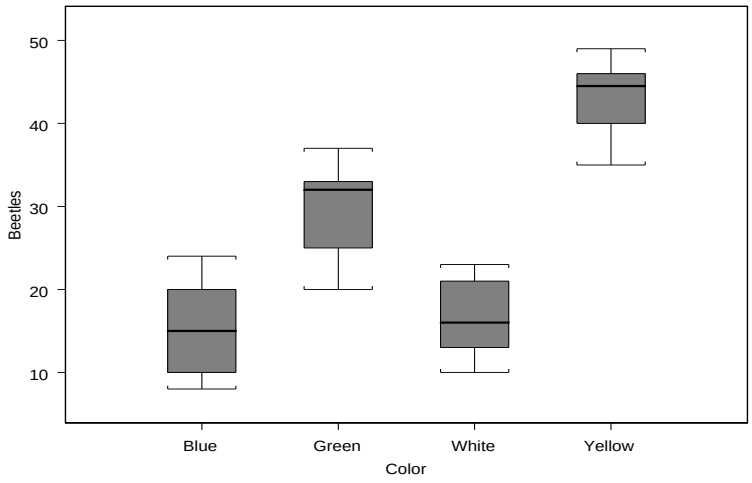



Figure 12.13: *Number of beetles caught by colored boards. Boxplots for four colors.*

## 12.2 Measures of center

*TERMINOLOGY*: With a sample of observations $x_1, x_2, ..., x_n$, the **sample mean** is defined as

$$\begin{aligned} \overline{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{1}{n}\sum x_i. \end{aligned}$$

That is, the sample mean is just the arithmetic average of the $n$ values $x_1, x_2, ..., x_n$. The symbol $\overline{x}$ is pronounced "$x$-bar," and is common notation. Physically, we can envision $\overline{x}$ as the balancing point on the histogram for the data.

*SIGMA NOTATION*: The symbol

$$\sum$$

(the capital Greek letter "sigma") simply means "add." This is convenient shorthand notation that we will use for the remainder of the course.

**Example 12.3.** With our beverage shelf life data from Example 12.1, the sum of the

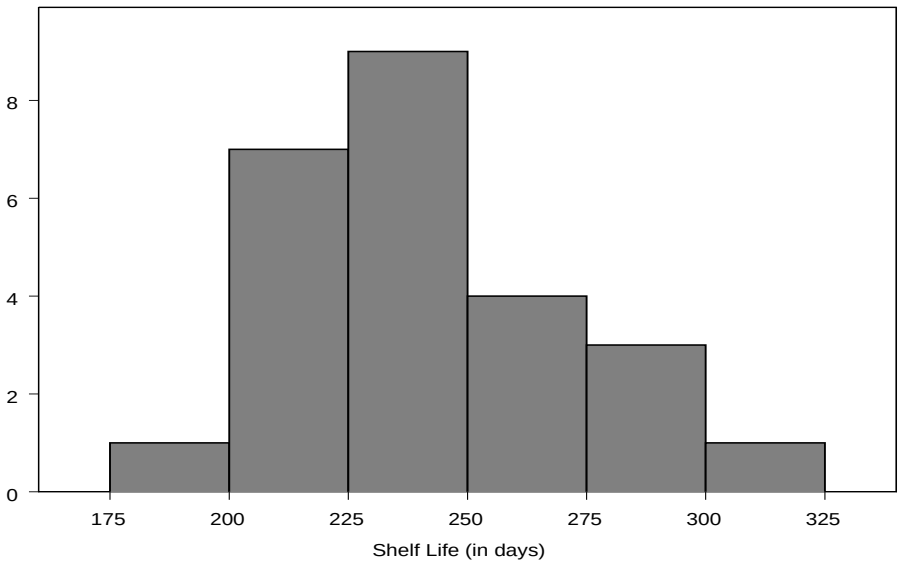



Figure 12.14: *Histogram for the shelf life data in Example* 12.1.

data is

$$\sum x_i = x_1 + x_2 + \cdots + x_{25} = 262 + 188 + \cdots + 249 = 5974,$$

and the sample mean of the 25 shelf lives is given by

$$\overline{x} = \frac{1}{25}\sum x_i = \frac{1}{25}(5974) = 238.96 \text{ days.}$$

Note that the mean and median are fairly "close," but not equal; i.e., $\overline{x} = 238.96$ and $M = 234$. Both statistics measure the **center** of a distribution.

*COMPARING THE MEAN AND MEDIAN*: The mean is a measure that can be heavily influenced by **outliers**.

- Unusually high data observations will tend to increase the mean, while unusually low data observations will tend to decrease the mean. One or two outliers will generally not affect the median!
- Sometimes we say that the median is generally **robust** to outliers.

**Example 12.4.** In a manufacturing process involving the production of automotive paint, a specimen is taken from the filling floor and sent to the chemistry lab for analysis (chemists check for color, texture, viscosity, etc.). Here is a sample of $n = 10$ elapsed times (in hours) for the chemistry lab to complete the analysis, after receiving a specimen from filling:

```
2.5   1.8   0.8   3.2   2.1   2.0   2.5   26.9   2.8   1.7
```

The ordered data are

```
0.8   1.7   1.8   2.0   2.1   2.5   2.5   2.8   3.2   26.9
```

*CALCULATIONS*: The **median** is equal to

$$M = \frac{2.1 + 2.5}{2} = 2.3 \text{ hours.}$$

The **mean** is equal to

$$\overline{x} = \frac{1}{10}\sum x_i = \frac{1}{10}(46.3) = 4.63 \text{ hours.}$$

*COMPARISON*: Note the large difference in the two measures of center! The sample mean suggests that the "normal" waiting time is greater than 4 hours (half of an 8-hour shift!). The sample median suggests that the "normal" waiting time is just a shade over 2 hours. These statistics convey very different interpretations. The question should be, *"What happened to cause the* 26.9 *observation?"*

EXERCISE: Recalculate the mean and median in Example 12.4, but exclude the outlier. Which statistic's value changes more?

*MORAL*: If there are distinct outliers in your data set, then the median should be used as a measure of center instead of the mean.

*MORAL*: The source of outliers should always be investigated! The presence of outliers suggests the presence of **assignable-cause variation** (rather than just chance variation). If we know what this assignable cause is, we can take the necessary steps to eliminate it.

*OBSERVATION*: The following results hold true for data distribution shapes:

- if a data distribution is **perfectly symmetric**, the median and mean will be equal.
- if a data distribution is **skewed right**, the mean will be greater than the median.
- if a data distribution is **skewed left**, the mean will be less than the median.

## 12.3 Measures of spread

*OBSERVATION*: Two data sets could have the same mean, but values may be spread about the mean value differently. For example, consider the two data sets:

| | | | | |
|---|---|---|---|---|
| 24 | 25 | 26 | 27 | 28 |
| 6 | 16 | 26 | 36 | 46 |

Both data sets have $\overline{x} = 26$. However, the second data set has values that are much more spread out about 26. The first data set has values that are much more compact about 26. That is, **variation** in the data is different for the two data sets.

### 12.3.1 Range and interquartile range

*RANGE*: An easy way to assess the variation in a data set is to compute the **range**, which we denote by $R$. The range is the largest value minus the smallest value; i.e.,

$$R = \text{Max} - \text{Min}.$$

For example, the range for the first data set above is $28 - 24 = 4$, while the range for the second is $46 - 6 = 40$.

*DRAWBACKS*: The range is very sensitive to outliers since it only uses the extreme observations!

*INTERQUARTILE RANGE*: The **interquartile range**, $IQR$, measures the spread in the center half of the data; it is the difference between the first and third quartiles; i.e.,

$$IQR = Q_3 - Q_1.$$

*NOTE*: This measure of spread is more resistant to outliers since it does not use the extreme observations. Because of this, the $IQR$ can be very useful for describing the spread in **skewed distributions**.

### 12.3.2 Variance and standard deviation

*TERMINOLOGY*: The **sample variance** of the data $x_1, x_2, ..., x_n$ is denoted by $s^2$ and is given by

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

*RATIONALE*: We are trying to measure how far observations deviate from the sample mean $\overline{x}$. To do this, a natural quantity to examine would be each observation's **deviation from the mean**, i.e., $x_i - \overline{x}$. However, one can show that

$$\sum(x_i - \overline{x}) = 0;$$

that is, the positive deviations and negative deviations "cancel each other out" when you add them!

*REMEDY*: Devise a measure that maintains the magnitude of each deviation but ignores their signs. Squaring each deviation achieves this goal. The quantity

$$\sum(x_i - \overline{x})^2$$

is called the **sum of squared deviations**. Dividing $\sum(x_i - \overline{x})^2$ by $(n-1)$ leaves (approximately) an average of the $n$ squared deviations. This is the sample variance $s^2$.

*TERMINOLOGY*: The **sample standard deviation** is the positive square root of the variance; i.e.,

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

**Example 12.5.** Keeping plants healthy requires an understanding of the organisms and agents that cause disease as well as an understanding of how plants grow and are affected by disease. In an experiment studying disease transmission in insects, $x$ denotes the number of insects per plot. A sample of $n = 5$ plots is observed yielding $x_1 = 5$, $x_2 = 7$, $x_3 = 4$, $x_4 = 9$ and $x_5 = 5$. What is the sample standard deviation?

SOLUTION. Here are the steps that I take when computing variance and standard deviation "by hand." Note that the sum of the data is

$$\sum x_i = 5 + 7 + 4 + 9 + 5 = 30.$$

Thus, the sample mean (we need this first!) is

$$\overline{x} = \frac{1}{5}(30) = 6 \text{ insects.}$$

Now, we form the following table:

| Observation | Mean | Deviation from mean | Squared deviation from mean |
|---|---|---|---|
| $x_i$ | $\overline{x}$ | $x_i - \overline{x}$ | $(x_i - \overline{x})^2$ |
| 5 | 6 | $5 - 6 = -1$ | $(-1)^2 = 1$ |
| 7 | 6 | $7 - 6 = 1$ | $1^2 = 1$ |
| 4 | 6 | $4 - 6 = -2$ | $(-2)^2 = 4$ |
| 9 | 6 | $9 - 6 = 3$ | $3^2 = 9$ |
| 5 | 6 | $5 - 6 = -1$ | $(-1)^2 = 1$ |

The **sum of squared deviations** is equal to

$$\sum (x_i - \overline{x})^2 = 1 + 1 + 4 + 9 + 1 = 16.$$

The **sample variance** is

$$s^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2 = \frac{1}{4}(16) = 4.$$

The **sample standard deviation** is

$$s = \sqrt{s^2} = \sqrt{4} = 2.$$

*FACTS ABOUT THE STANDARD DEVIATION AND VARIANCE*:

1. The larger the value of $s$ ($s^2$), the more variation in the data $x_1, x_2, ..., x_n$; i.e., **the larger the spread** in the distribution of data.
2. $s \geq 0$ ($s^2 \geq 0$); i.e., these values can not be negative!
3. If $s = 0$ ($s^2 = 0$), then $x_1 = x_2 = \cdots = x_n$. That is, all the data values are equal (there is zero spread).
4. $s$ and $s^2$, in general, are heavily influenced by outliers. To be precise, outliers drive up the values of $s$ and $s^2$.
5. $s$ is measured in the **original units** of the data; $s^2$ is measured in (units)$^2$. This is an advantage of using the sample standard deviation.

EXERCISE: Here are the number of points scored by LA Lakers basketball player Kobe Bryant over an $n = 11$ game span in the 1999-2000 season:

```
15   22   18   22   30   31   13   26   29   18   18
```

Find the sample mean and sample standard deviation for these data.

## 12.4 Review

*NOTES*: In this chapter, we have talked about different statistics to summarize (numerically) the notion of **center** and **spread** in a quantitative data distribution.

*HOWEVER*: When you are doing **exploratory data analysis**, simply computing the values of these statistics (e.g., mean, standard deviation, etc.) is not enough. If you only look at these numerical values, you are missing the picture that the distribution paints for you. *Always plot your data distribution first before doing any computations!*

# 13 Normal Distributions

Complementary reading from Moore and Notz: Chapter 13.

## 13.1 Introduction

*RECALL*: We now have examined graphical displays for quantitative data (e.g., boxplots, histograms, stem plots); we use these displays to portray **empirical distributions** (i.e., distributions of data). For any quantitative data distribution, we know to always investigate

- the **shape** of the distribution (e.g., symmetric, skewed, etc.)
- the **center** of the distribution (numerically summarize with mean or median)
- the **spread** of the distribution (numerically summarize with standard deviation, range, IQR)
- the presence of unusual observations and deviations (e.g., **outliers**, etc.).

**Example 13.1.** Low infant birth weight is a fairly common problem that expecting mothers experience. In Larimer County, CO, a group of researchers studied mothers aged from 25 to 34 years during 1999 to 2003. During this time, 5242 male birth weights were recorded. The data appear in a **relative frequency histogram** in Figure 13.15. A relative frequency histogram is a special histogram. Percentages are plotted on the vertical axis (instead of counts like we did in Chapter 11). *Plotting percentages does not alter the shape of the histogram.*

*ANALYSIS*:

- The distribution of male birth weights in Larimer County looks very **symmetric**.
- The **center** of the distribution appears to be close to 8 lbs (or slightly less).

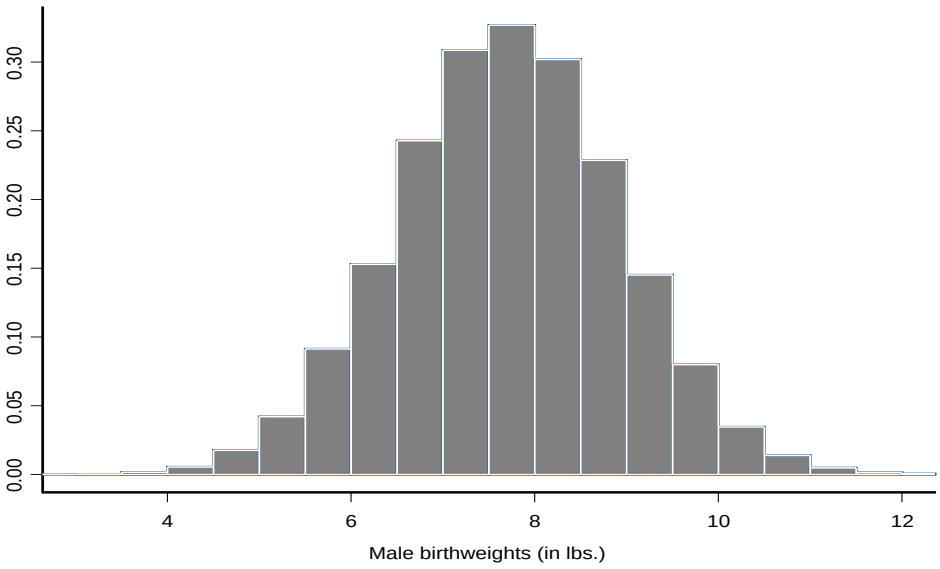



Figure 13.15: *Distribution of male birth weights in Larimer County, CO.*

- The **spread** in the distribution is apparent, ranging from around 4 lbs (or slightly less) to 12 lbs (or slightly more).
- The distribution looks very symmetric with **no real evidence of outliers** (e.g., there aren't any babies weighing 16 pounds!).

## 13.2 Density curves

*OBSERVATION*: Sometimes the overall pattern of a large number of observations is so "regular" that we can describe the pattern by a smooth curve. We call this curve a density curve. *A density curve describes the overall pattern of a distribution.*

*TERMINOLOGY*: One can think of a **density curve** as a "smooth curve approximation" to a histogram of data. It is sometimes convenient to think about them as **theoretical models** for the variable of interest (in Example 13.1, male birth weights

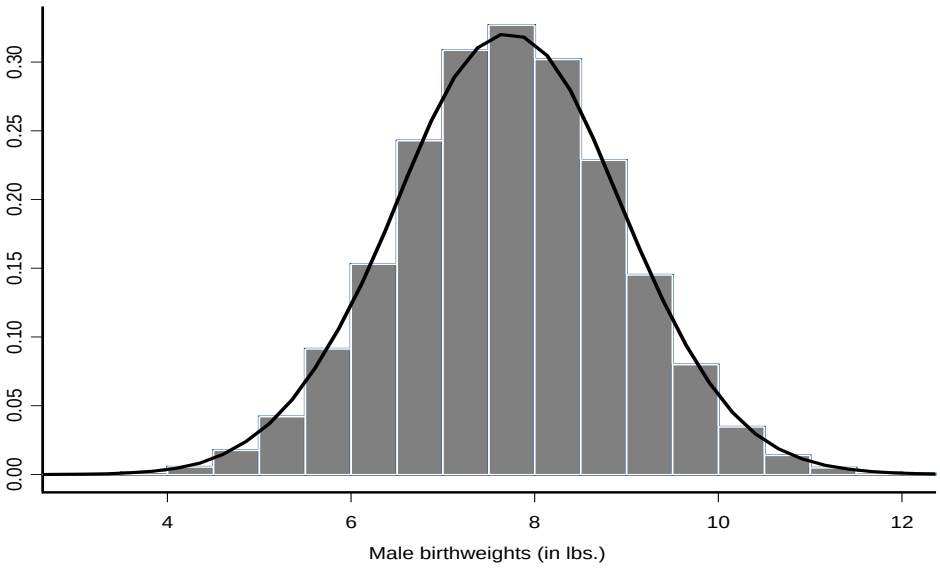



Figure 13.16: *Distribution of male birth weights in Larimer County, CO. A density curve has been superimposed over the relative frequency histogram.*

for Larimer County, CO). In Figure 13.16, I have superimposed a density curve over the relative frequency histogram.

*PROPERTIES*: In general, a **density curve** associated with a quantitative variable (e.g., male birth weights, etc.) is a curve with the following properties:

1. the curve is non-negative
2. the area under the curve is 1
3. the area under the curve between two values $a$ and $b$ represents the **proportion** of observations that fall in that range.

*NOTE*: We display density curves in a way so that the proportion of observations in any region can be found by **finding the area under the curve**. To do this, we rescale the vertical axis so that the total area under the curve is equal to 1.

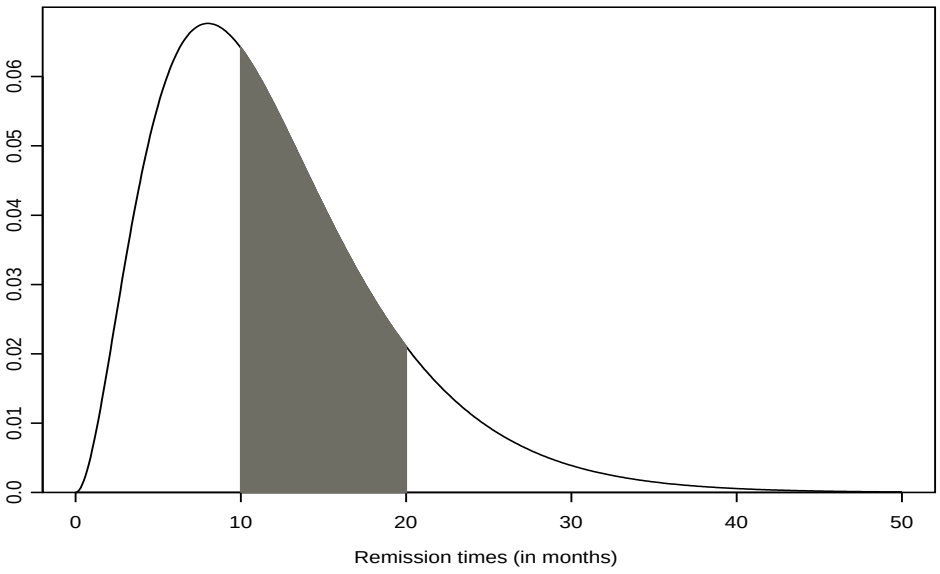



Figure 13.17: *Density curve for the remission times of leukemia patients.*

**Example 13.1** (continued). For the density curve in Figure 13.16,

- Approximate the proportion of male newborns that weigh **between** 7 and 9 lbs. Shade this area under the curve.
- Approximate the proportion of male newborns that weigh **less than** 5.5 lbs (i.e., are classified as low birthweight). Shade this area under the curve.

**Example 13.2.** Researchers have learned that the best way to treat patients with acute lymphoid leukemia is to administer large doses of several chemotherapeutic drugs over a short period of time. The density curve in Figure 13.17 represents the remission time for a certain group of leukemia patients (note how the vertical axis has been scaled). The shaded area represents the **proportion** of remission times between 10 and 20 months.

*FACT*: Density curves come in many different shapes and sizes! We can still characterize a density curve as **symmetric** or **skewed**. See Figure 13.18.

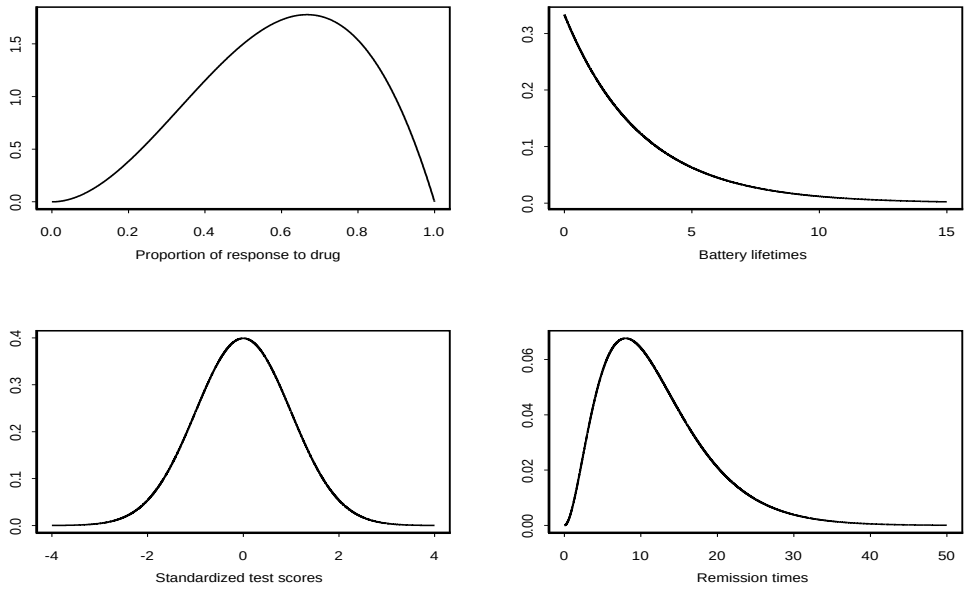



Figure 13.18: *Four density curves.*

## 13.3 Measuring the center and spread for density curves

*MAIN POINT*: The notion of **center** and **spread** is the same for density curves as before when we were discussing histograms. However, because we are now talking about theoretical models (instead of raw data), we change our notation to reflect this. In particular,

- the **mean** for a density curve is denoted by $\mu$.
- the **standard deviation** for a density curve is denoted by $\sigma$.

*SUMMARY*: Here is a review of the notation that we have adopted so far in the course. This is standard notation.

| | Observed data | Density curve |
|---|---|---|
| Mean | $\overline{x}$ | $\mu$ |
| Standard deviation | $s$ | $\sigma$ |

*CONNECTION*:

- The **sample values** $\overline{x}$ and $s$ are computed from observed data. These are statistics!
- The **population values** $\mu$ and $\sigma$ are theoretical values. These are parameters!

*COMPARING MEAN AND MEDIAN*: The **mean** for a density curve $\mu$ can be thought of as a "balance point" for the distribution. On the other hand, the **median** for a density curve is the point that divides the area under the curve in half. The relationship between the mean and median is the same as it was before; namely,

- if a density curve is perfectly symmetric, the median and mean will be **equal**.
- if a density curve is skewed right, the mean will be **greater** than the median.
- if a density curve is skewed left, the mean will be **less** than the median.

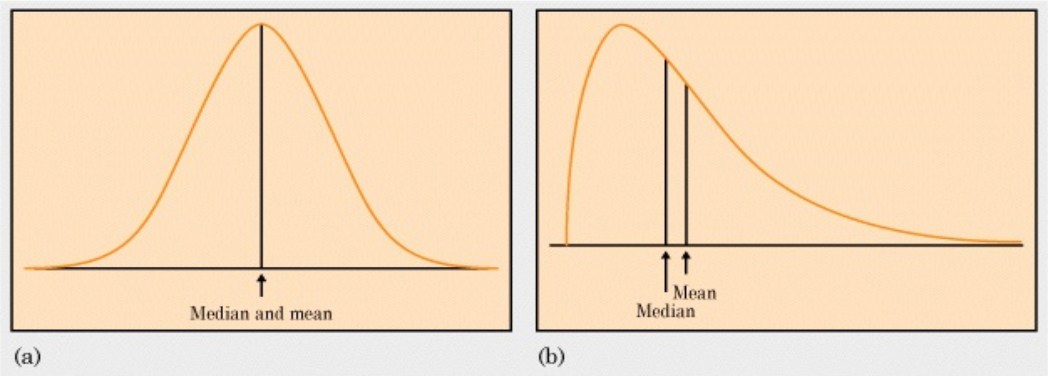



Figure 13.19: *Comparing the mean and median for density curves.*

## 13.4 Normal distributions

*NORMAL DENSITY CURVES*: The most famous (and important) family of density curves is the normal family or **normal distributions**. A normal distribution is also sometimes called the "bell-shaped distribution" or "bell-shaped curve."

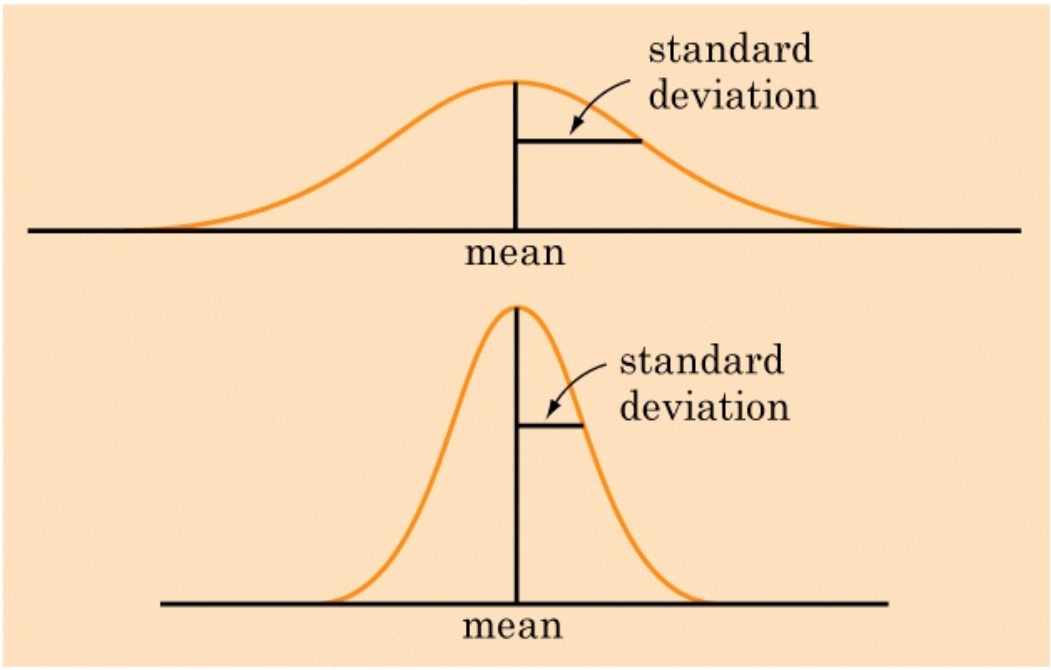



Figure 13.20: *Two normal curves. The standard deviation fixes the spread of the curve.*

*DESCRIPTION*: *There are many normal curves!* To describe a particular normal curve, we need 2 pieces of information:

- the mean $\mu$
- the standard deviation, $\sigma$.

*FACTS*: Here are some properties for the normal family of density curves:

- the curves are **symmetric** (mean and median are equal)
- area under the a normal density curve is 1
- "mound shaped" and **unimodal**; i.e., there is only one "peak"
- change of curvature points at $\mu \pm \sigma$.

*SHORT-HAND NOTATION FOR NORMAL DISTRIBUTIONS*: Because we will mention normal distributions often, a short-hand notation is useful. *We abbreviate the normal*

*distribution with mean $\mu$ and standard deviation $\sigma$ by*

$$\mathcal{N}(\mu, \sigma).$$

To denote that a quantitative variable $X$ follows a normal distribution with mean $\mu$ and standard deviation $\sigma$, we write $X \sim \mathcal{N}(\mu, \sigma)$. The shorthand symbol "$\sim$" is read "is distributed as."

### 13.4.1 Empirical Rule

*EMPIRICAL RULE*: For the normal family of distributions, approximately

- 68 percent of the observations will be within **one** standard deviation of the mean,
- 95 percent of the observations will be within **two** standard deviation of the mean, and
- 99.7 percent (or almost all) of the observations will be within **three** standard deviations of the mean.

This is also known as the **68-95-99.7 Rule**.

**Example 13.3.** The variable $X$ denotes the tomato yields (measured in bushels/acre) in an agricultural experiment for a certain region. In Figure 13.22, we have a normal density curve with $\mu = 25$ bushels/acre and $\sigma = 5$ bushels/acre; this is the density curve for $X$. We write $X \sim \mathcal{N}(25, 5)$. The **Empirical Rule** says that

- about 68 percent of the observations (yields) will be within $25 \pm 5$, or between 20 and 30 bushels/acre.
- about 95 percent of the observations (yields) will be within $25 \pm 10$, or between 15 and 35 bushels/acre.

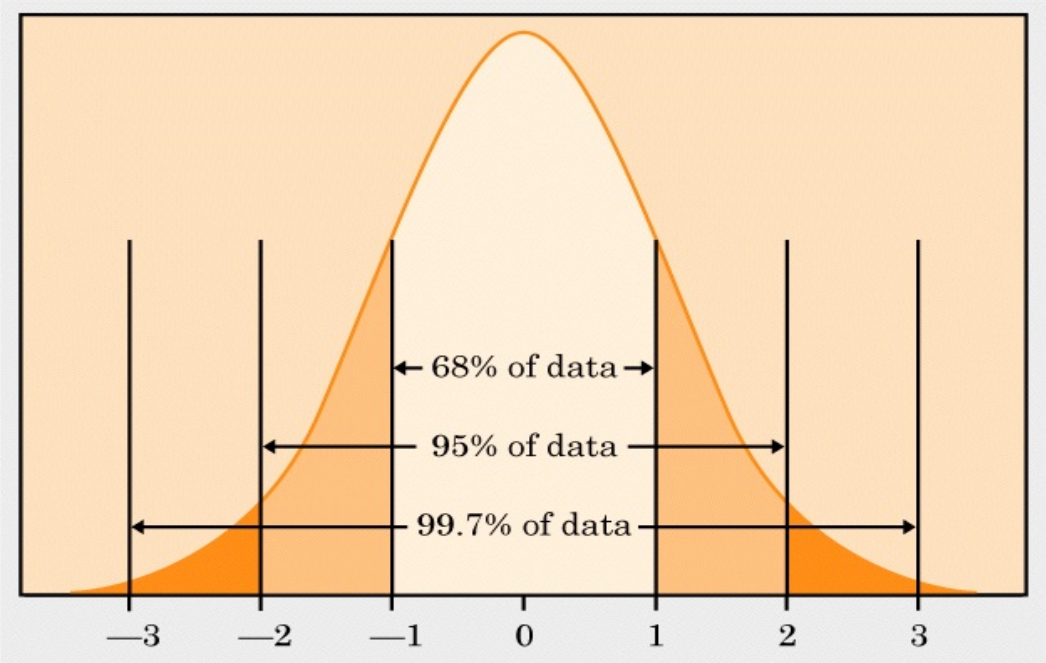



Figure 13.21: *The Empirical Rule for normal distributions. The normal distribution depicted here has* $\mu = 0$ *and* $\sigma = 1$.

- about 99.7 percent of the observations (yields) will be within $25 \pm 15$, or between 10 and 40 bushels/acre.

EXERCISE: Refer to Figure 13.22, a normal distribution with mean $\mu = 25$ and standard deviation $\sigma = 5$. This is our theoretical model for the distribution of tomato yields. Using the Empirical Rule,

(a) what proportion of tomato yields will be **less than** 20 bushels per acre?

(b) what proportion of tomato yields will be **greater than** 35 bushels per acre?

(c) what proportion of tomato yields will be **between** 10 and 35 bushels per acre?

(d) what proportion of tomato yields will be **between** 15 and 30 bushels per acre?

For each question, draw the corresponding picture (with appropriate shading). Make sure that your picture is well-labeled.

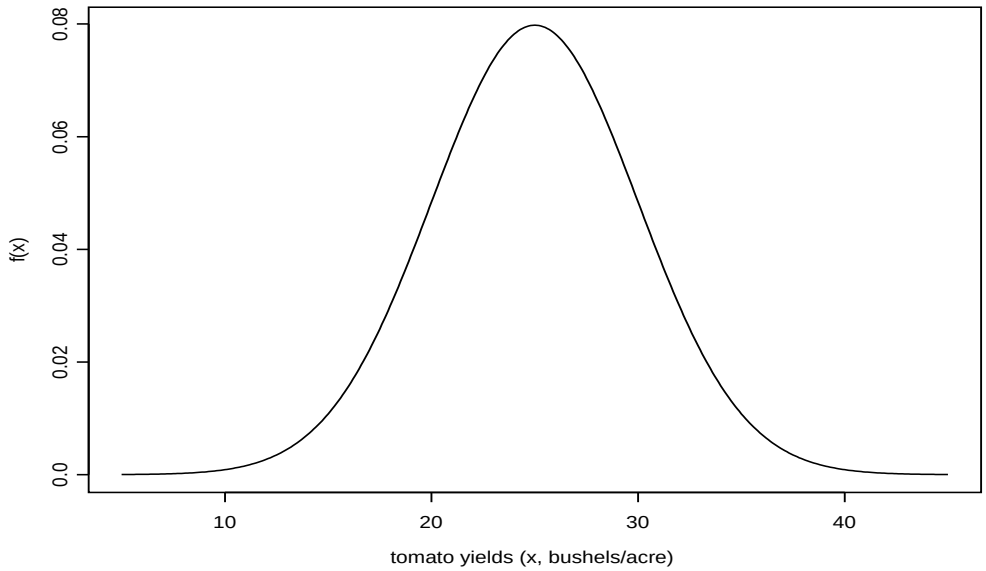



Figure 13.22: *A normal density curve with mean $\mu = 25$ and standard deviation $\sigma = 5$. A model for tomato yields.*

### 13.4.2 Standardization

*TERMINOLOGY*: The **standard normal distribution** is a normal density curve with mean 0 and standard deviation 1. This density curve is abbreviated $\mathcal{N}(0, 1)$.

*STANDARDIZED SCORES*: If $x$ is an observation from a $\mathcal{N}(\mu, \sigma)$ density curve, then the **standard score** of $x$ is

$$z = \frac{x - \mu}{\sigma}.$$

This is also called a **standardized value** or $z$**-score**.

*FACTS ABOUT STANDARDIZED VALUES*:

- Standardized values follow a standard normal distribution!
- Standardized values are **unitless** (i.e., they have no units attached to them).

- A standardized value $z$ indicates how many standard deviations an observation $x$ falls above or below the mean $\mu$.
  - If the standardized value $z < 0$, then $x$ is **below** the mean $\mu$.
  - If the standardized value $z > 0$, then $x$ is **above** the mean $\mu$.

*NOVELTY*: Standardized scores are helpful because they allow us to compare observations on a unitless scale. The following example illustrates this.

**Example 13.4.** College entrance exams, such as the SAT and ACT, are important in determining whether or not a potential undergraduate student is admitted. Historical evidence suggests that

- SAT scores are normally distributed with mean 1000 and standard deviation 180.
- ACT scores are normally distributed with mean 21 and standard deviation 5.

(a) Draw sketches of these two distributions. Label the scales as accurately as possible.

(b) Suppose that Joe scores 1180 on the SAT and Jane scores 30 on the ACT. Mark these values on your sketches. Who has done better compared to their peers? Answer this question by computing each student's standardized score.

(c) Suppose that Matt scores 910 on the SAT and Maria scores 15 on the ACT. Mark these values on your sketches. Who has done better compared to their peers?

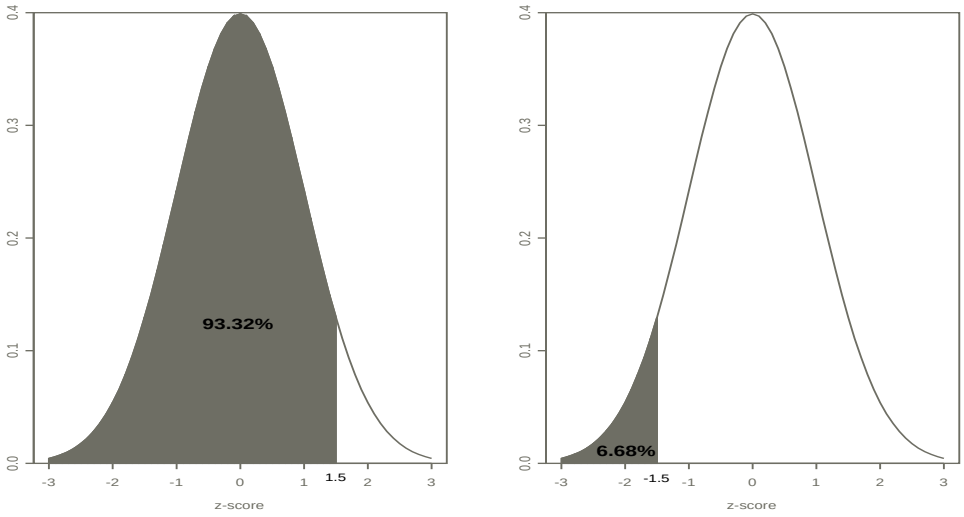



Figure 13.23: *Standard normal density curve. Left: The* 93.32*nd percentile is* $z = 1.5$. *Right: The* 6.68*th percentile is* $z = -1.5$.

### 13.4.3 Percentiles

*TERMINOLOGY*: The $c$**th percentile** of a distribution is a value such that $c$ percent of the observations lie below it, and the rest lie above. Here, $c$ is a value between 0 and 100. We have a table that provides percentiles for the standard normal distribution! This is the table on the next page (in the text, it is Table B, page 552).

*SPECIAL PERCENTILES*: We have already discussed some "special" percentiles; in particular,

- the **median** is the 50th percentile.
- the **first quartile**, $Q_1$, is the 25th percentile.
- the **third quartile**, $Q_3$, is the 75th percentile.

QUESTION: What are these values for the standard normal density curve? See next page.

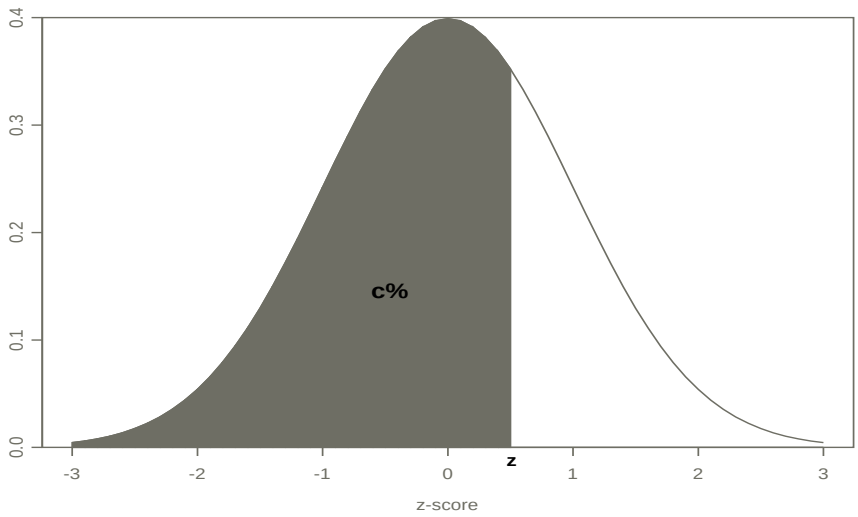



Table 13.13: *Percentiles of the standard normal distribution.*

| Standard score ($z$) | Percentile ($c$) | Standard score ($z$) | Percentile ($c$) | Standard score ($z$) | Percentile ($c$) |
|---|---|---|---|---|---|
| -3.4 | 0.03 | -1.1 | 13.57 | 1.2 | 88.49 |
| -3.3 | 0.05 | -1.0 | 15.87 | 1.3 | 90.32 |
| -3.2 | 0.07 | -0.9 | 18.41 | 1.4 | 91.92 |
| -3.1 | 0.10 | -0.8 | 21.19 | 1.5 | 93.32 |
| -3.0 | 0.13 | -0.7 | 24.20 | 1.6 | 94.52 |
| -2.9 | 0.19 | -0.6 | 27.42 | 1.7 | 95.54 |
| -2.8 | 0.26 | -0.5 | 30.85 | 1.8 | 96.41 |
| -2.7 | 0.35 | -0.4 | 34.46 | 1.9 | 97.13 |
| -2.6 | 0.47 | -0.3 | 38.21 | 2.0 | 97.73 |
| -2.5 | 0.62 | -0.2 | 42.07 | 2.1 | 98.21 |
| -2.4 | 0.82 | -0.1 | 46.02 | 2.2 | 98.61 |
| -2.3 | 1.07 | 0.0 | 50.00 | 2.3 | 98.93 |
| -2.2 | 1.39 | 0.1 | 53.98 | 2.4 | 99.18 |
| -2.1 | 1.79 | 0.2 | 57.93 | 2.5 | 99.38 |
| -2.0 | 2.27 | 0.3 | 61.79 | 2.6 | 99.53 |
| -1.9 | 2.87 | 0.4 | 65.54 | 2.7 | 99.65 |
| -1.8 | 3.59 | 0.5 | 69.15 | 2.8 | 99.74 |
| -1.7 | 4.46 | 0.6 | 72.58 | 2.9 | 99.81 |
| -1.6 | 5.48 | 0.7 | 75.80 | 3.0 | 99.87 |
| -1.5 | 6.68 | 0.8 | 78.81 | 3.1 | 99.90 |
| -1.4 | 8.08 | 0.9 | 81.59 | 3.2 | 99.93 |
| -1.3 | 9.68 | 1.0 | 84.13 | 3.3 | 99.95 |
| -1.2 | 11.51 | 1.1 | 86.43 | 3.4 | 99.97 |

*FINDING PERCENTILES FOR ANY NORMAL DENSITY CURVE*: We have seen that the standardized value $z$ is related to an observation $x$ in the following way:

$$z = \frac{x - \mu}{\sigma}.$$

If we solve this equation for the observation $x$, we get

$$x = \mu + z\sigma$$

This equation allows us to find percentiles for any normal density curve. All we have to know is $\mu$, $\sigma$, and $z$.

**Example 13.5.** Historical evidence suggests that SAT scores are normally distributed with mean 1000 and standard deviation 180. What score do you have to make to be in the **top 10 percent**?

SOLUTION. We would like to find the 90th percentile of a $\mathcal{N}(1000, 180)$ density curve (draw this density curve below). For $c = 90$, we have that

$$z \approx 1.3.$$

We already know that $\mu = 1000$ and $\sigma = 180$. Thus, solving for $x$, we obtain

$$x = \mu + z\sigma = 1000 + 1.3(180) = 1000 + 234 = 1234.$$

If we score a 1234 (or higher), then we would be in the top 10 percent.

QUESTIONS: Historical evidence suggests that SAT scores are normally distributed with mean 1000 and standard deviation 180.

(a) What SAT score do you have to make to be in the **top 30 percent**?

(b) Ten percent of all SAT scores will be **below** which value?

(c) What is the **interquartile range** of the SAT score distribution? Recall that $IQR = Q_3 - Q_1$.

# 14 Describing Relationships: Scatterplots and Correlation

Complementary reading from Moore and Notz: Chapter 14.

## 14.1 Introduction

*OVERVIEW*: A problem that often arises in the social sciences, economics, industrial applications, and biomedical settings is that of investigating the mathematical **relationship** between two variables.

*EXAMPLES*:

- amount of alcohol in the body (BAL) versus body temperature (degrees C)
- weekly fuel consumption (degree days) versus house size (square feet)
- amount of fertilizer (pounds/acre) applied versus yield (kg/acre)
- sales ($1000s) versus marketing expenditures ($1000s)
- HIV status (yes/no) versus education level (e.g., primary, secondary, college, etc.)
- gender (M/F) versus promotion (yes/no). Are promotion rates different?
- helium-filled footballs versus air-filled footballs. Is there a difference?
- Remission time (in days) versus treatment (e.g., surgery/chemotherapy/both)

**Example 14.1.** Many fishes have a lateral line system enabling them to experience *mechanoreception*, the ability to sense physical contact on the surface of the skin or movement of the surrounding environment, such as sound waves in air or water. In an experiment to study this, researchers subjected fish to electrical impulses. The frequency (number per second) of electrical impulses (EI) emitted from $n = 7$ fish measured at several temperatures (measured in Celcius); the data are listed in Table 14.14.

Table 14.14: *Fish electrical impulse data.*

| Temperature ($x$) | Frequency ($y$) | Temperature ($x$) | Frequency ($y$) |
|---|---|---|---|
| 20 | 224 | 27 | 301 |
| 22 | 252 | 28 | 306 |
| 23 | 267 | 30 | 318 |
| 25 | 287 | | |

## 14.2 Scatterplots

*TERMINOLOGY*: A **scatterplot** is a graphical display which 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.
- Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

*NOTE*: Scatterplots give a **visual impression** of how the two variables behave together.

**Example 14.1** (continued). The scatterplot for the fish electrical impulse data is given in Figure 14.24. It is clear that these variables are strongly related. As the water temperature increases, there is a tendency for the frequency of impulses to increase as well.

*VARIABLES*: In studies where there are two variables under investigation (e.g., temperature, EI frequency), it is common that one desires to study how one variable is affected by the other. In some problems, it makes sense to focus on the behavior of one variable and, in particular, determine how another variable influences it. In a scientific investigation, the variable that is of primary interest is the **response variable**. An **explanatory variable** explains or causes changes in the response variable.

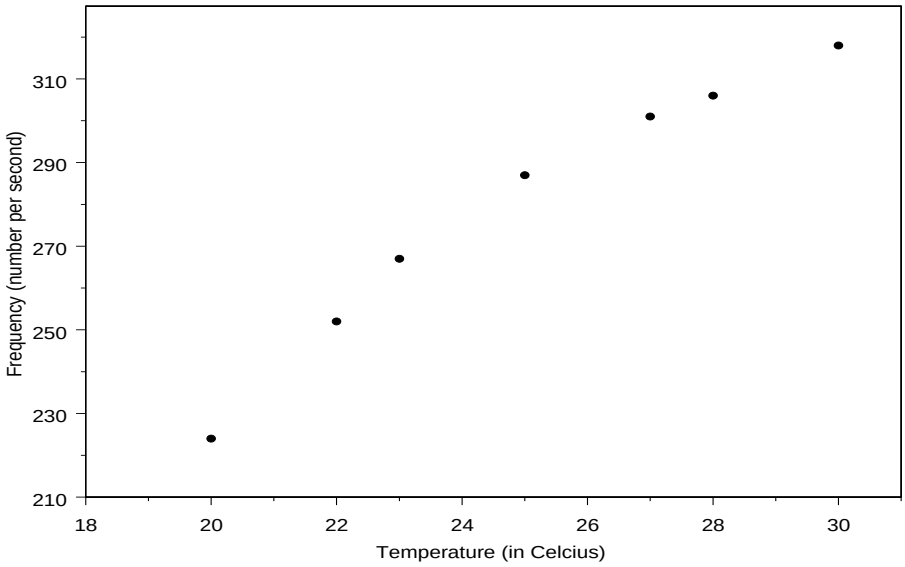



Figure 14.24: *Scatterplot for EI frequency at different temperatures.*

*NOTATION*: As a reminder, we usually denote the response variable by $y$ and the explanatory variable by $x$. In Example 14.1, EI frequency ($y$) is the response variable, and temperature ($x$) is the explanatory variable.

*INTERPRETING SCATTERPLOTS*: It is important to describe the overall pattern of a scatterplot by examining the following:

- **form**; are there linear or curved relationships or different clusters of observations?
- **direction**; are the two variables positively related or negatively related?
- **strength**; how strong is the relationship? Strong? Weak? Mild?
- the presence of **outliers**.

*LINEAR RELATIONSHIPS*: If the form of the scatterplot looks to resemble a straight-line trend, we say that the relationship between the variables is **linear**.

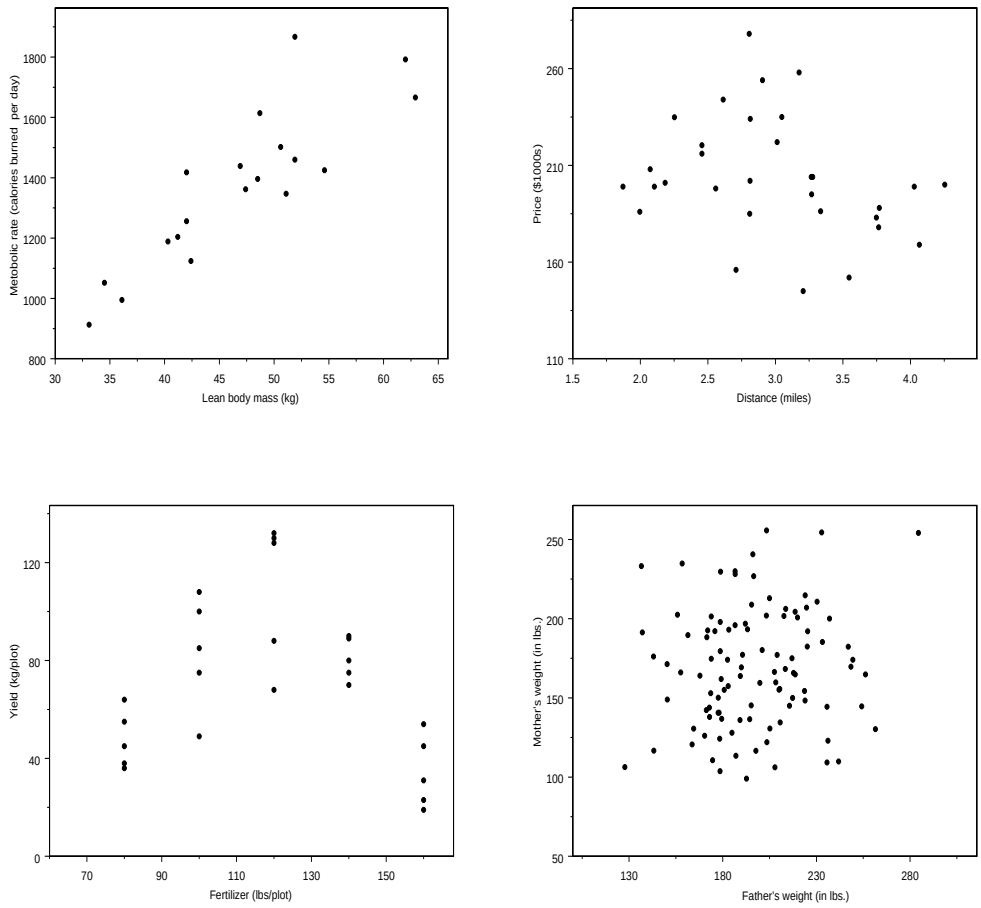



Figure 14.25: *Upper left: positive linear relationship. Upper right: mild linear negative relationship. Lower left: curved relationship. Lower right: random scatter.*

*TERMINOLOGY*: Two variables are **positively related** if they tend to increase together. They are **negatively related** if an increase in one is associated with a decrease in the other.

**Example 14.2.** Consider a situation where interest focuses on two different methods of calculating the age of a tree. One way is by counting tree rings. This is considered to be very accurate, but requires sacrificing the tree. Another way is by a carbon-dating process. Suppose that data are obtained for $n = 50$ trees on age by the counting method ($x$) and age by carbon dating ($y$), both measured in years. A scatterplot of these data is given in Figure 14.26.

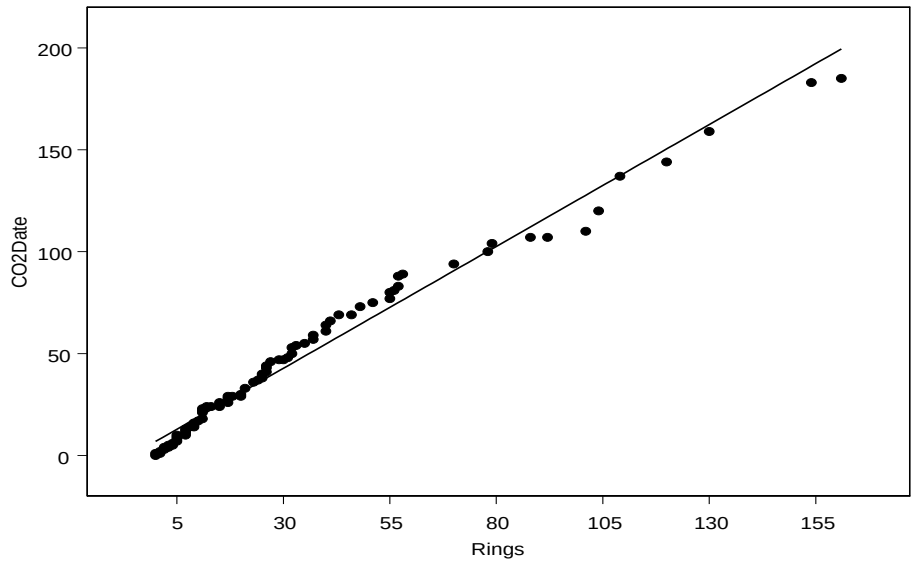



Figure 14.26: *Tree ages measured by two different methods. A straight line has been superimposed to emphasize the linear trend.*

*INTERPRETATION*:

- **Form**: There is **linear** relationship between the two aging methods.
- **Direction**: The two aging methods are **positively related**.
- **Strength**: The relationship between the two aging methods is **strong** (this is reassuring; this suggests that both methods are similar in assessing tree age; i.e., the carbon dating method could be used in place of the ring-counting method).
- **Outliers**: There is not a strong indication of outliers being present here.

**Example 14.3.** Two identical footballs, one air-filled and one helium-filled, were used outdoors on a windless day at Ohio State University's athletic complex. Each football was kicked 39 times and the two footballs were alternated with each kick. The experimenter recorded the distance traveled by each ball. The distances are recorded in yards.

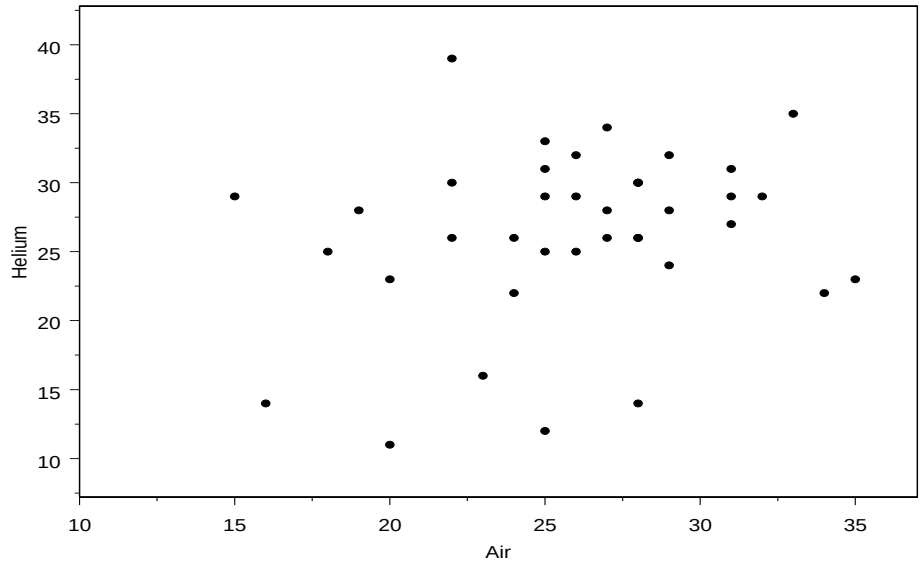



Figure 14.27: *Ohio State football data. Helium versus air football distances.*

*INTERPRETATION*:

- **Form**: There does not look to be a linear relationship between the two types of footballs (although it slightly linear).
- **Direction**: It is hard to discern whether or not the relationship is positive or negative (although it is slightly positive).
- **Strength**: This relationship is not strong ("weak" at best).
- **Outliers**: A few outliers (perhaps). This is not data from Ohio State's 2006 kicker!

## 14.3 Correlation

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

*TERMINOLOGY*: 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, denote by $x_i$ and $y_i$ the two measurements for the $i$th individual. The correlation is computed by the following formula:

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

where $\overline{x}$ and $\overline{y}$ are the sample means and $s_x$ and $s_y$ are the sample standard deviations. *It is often best to use statistical software to compute the correlation.* You should note that the terms

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

are the **sample standardized values** of $x_i$ and $y_i$, respectively.

*HAND-CALCULATION*: See Examples 2 and 3, pages 273-6 (MN).

## 14.4 Understanding correlation

*MY PHILOSOPHY*: Knowing how to compute the correlation $r$ "by hand" is not very important at this level. It is **much more important** that you understand how the correlation measures the association between two variables.

*PROPERTIES OF THE CORRELATION*:

1. The correlation $r$ is a **unitless number**; that is, there are no units attached to it (e.g., dollars, mm, etc.). This also means that you could change the units of your data (e.g., inches to cm) and this would not affect the value of $r$.

2. It also makes no difference what you call $x$ and what you call $y$; the correlation will be the same. That is, the correlation $r$ **ignores** the distinction between explanatory and response variables.

3. The correlation $r$ **always** satisfies $-1 \leq r \leq 1$.

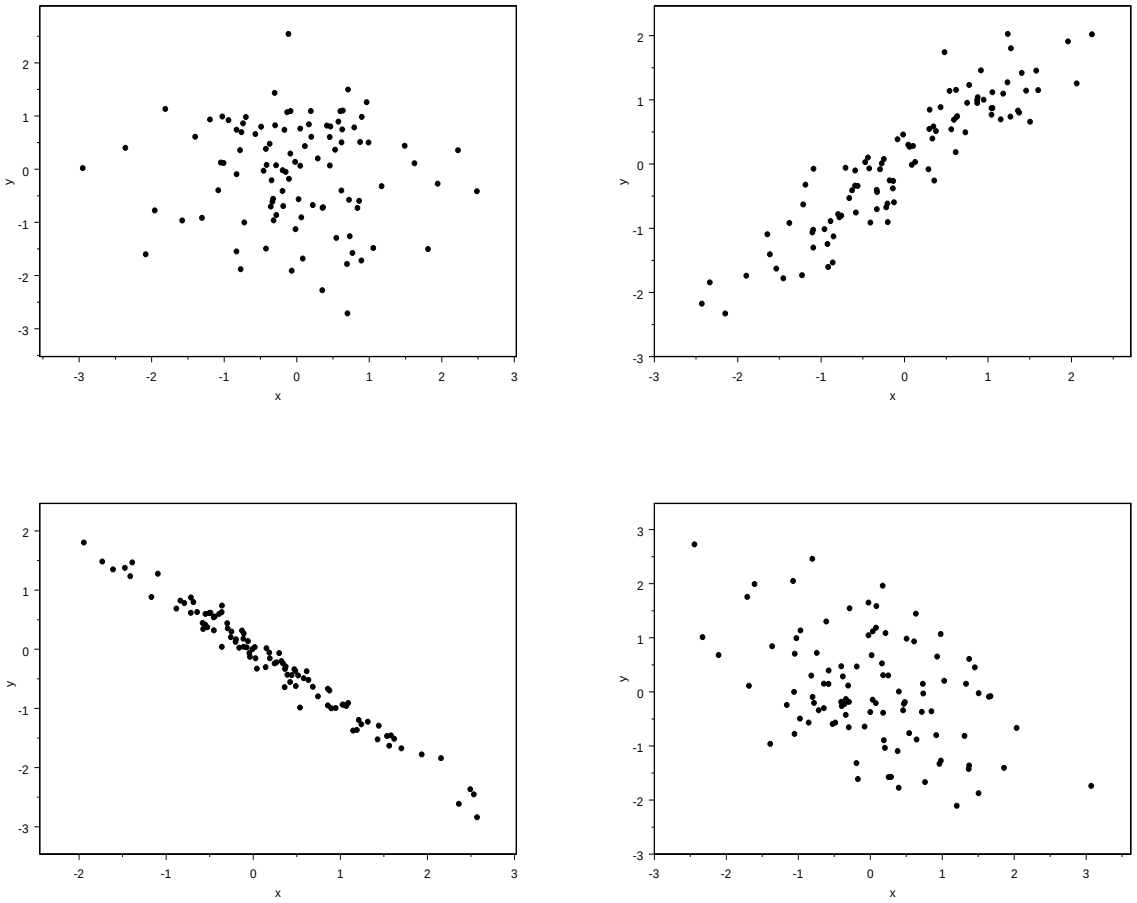


Figure 14.28: *Four scatterplots using data generated from Minitab. Upper left:* $r = 0$. *Upper right:* $r = 0.9$. *Lower left:* $r = -0.99$. *Lower right:* $r = -0.5$.

4. If $r = 1$, then all data lie on a straight line with **positive** slope. If $r = -1$, then all the data lie on a straight line with **negative** slope.

5. When $0 < r < 1$, there is a tendency for the values to vary together in a positive way (i.e., a positive linear relationship). When $-1 < r < 0$ there is a tendency for the values to vary together in a negative way (i.e., a negative linear relationship).

6. Values of $r$ close to zero (e.g., $-0.25 < r < 0.25$) indicate very weak linear association between $x$ and $y$. If $r = 0$, then there is **no linear relationship** present in the data.

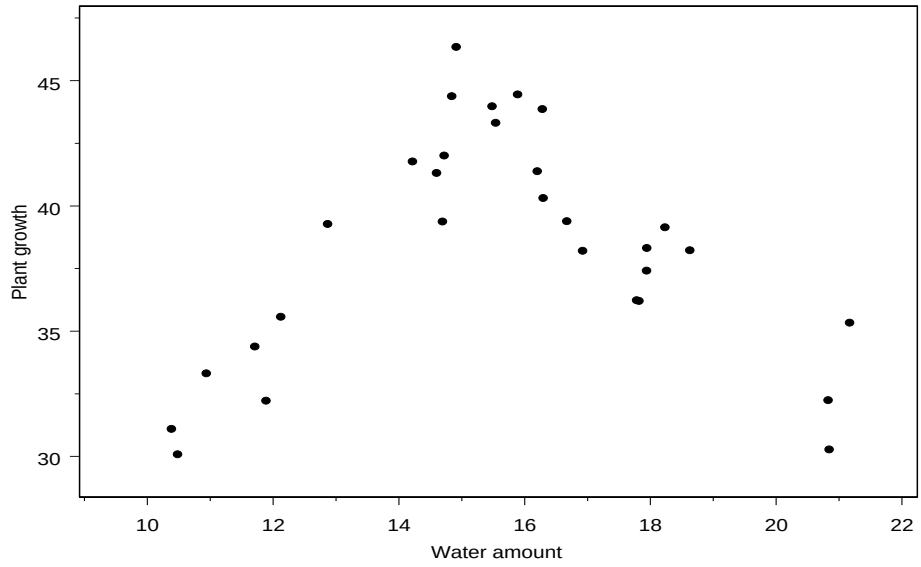



Figure 14.29: *Plant growth versus water amount.*

7. **The correlation only measures linear relationships!** It does not describe a curved relationship, no matter how strong that relationship is! Thus, we could have two variables $x$ and $y$ that are very strongly related, but the correlation still be close to zero! This would occur if the variables are related **quadratically**. See Example 14.3.

8. The value of $r$ could be highly affected by **outliers**. This makes sense since sample means and sample standard deviations are affected by outliers (and these values are required to compute $r$).

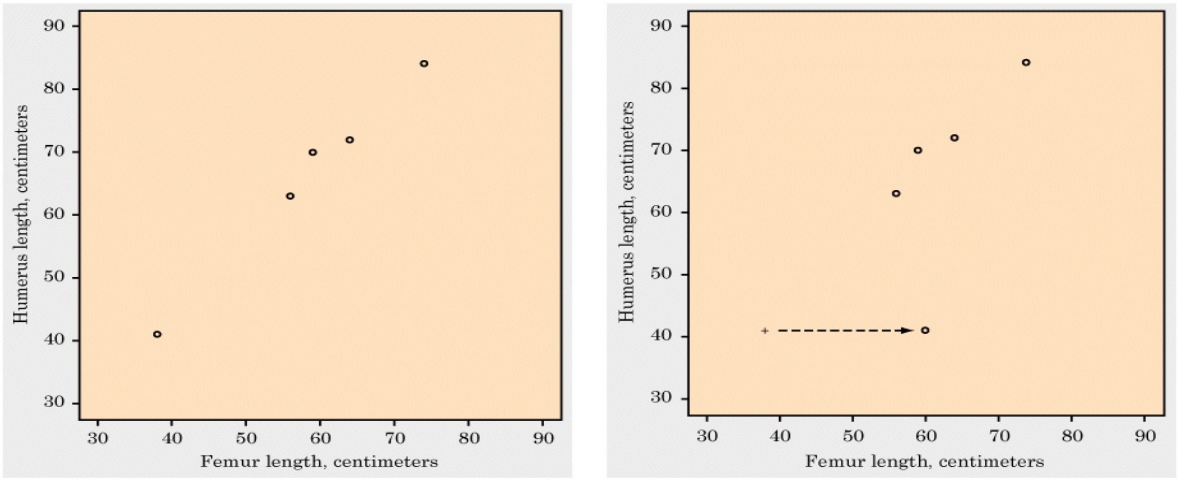
**Example 14.3** (continued). The correlation between air and helium is $r = 0.231$. This suggests a weak positive association.

**Example 14.4.** Researchers are trying to understand the relationship between the amount of water applied to plots (measured in cm) and total plant growth (measured in cm). A sample of $n = 30$ plots is taken from different parts of a field. The data from the

sample are given in Figure 14.29. Using statistical software, the correlation between plant growth and water amount is computed to be $r = 0.088$. This is an example where the two variables under investigation (water amount and plant growth) have a very strong relationship, but the correlation is near zero. This occurs because the relationship is not linear; rather, it is **quadratic**. An investigator that did not plot these data and only looked at $r$ could be lead astray and conclude that these variables were not related!

*MORAL*: Always plot your data first in a scatterplot to detect possible nonlinear relationships. *The correlation does not measure nonlinear relationships!*

**Example 14.5.** As we pointed out, outliers can drastically change the value of the correlation. To illustrate this, consider the following two scatterplots. Depicted here are the lengths of the femur (leg bone) and the humerus (upper arm bone) for five fossils for extinct animals (archaeopteryx). The data are given in Example 2 (page 273, MN).




(a) No outlier. (b) Outlier.

Figure 14.30: *Scatterplots of the lengths of two bones in* 5 *fossil specimens of the archaeopteryx. This shows the effect of outliers on the correlation. Left:* $r = 0.994$. *Right:* $r = 0.640$.

*MORAL*: Always plot your data first in a scatterplot to detect possible outliers; these unusual observations can have a large effect on the correlation.

# 15 Describing Relationships: Regression, Prediction, and Causation

Complementary reading from Moore and Notz: Chapter 15.

## 15.1 Introduction

*GOAL*: If a scatterplot shows a **straight-line relationship** between two quantitative variables, we would like to summarize this overall pattern by "fitting" a line to the data.

**Example 15.1.** In Example 14.1 (notes), we looked at the relationship between the temperature and EI frequency for fish (the scatterplot is below). As before, we have plotted temperature ($x$) on the horizontal axis and EI frequency ($y$) on the vertical.

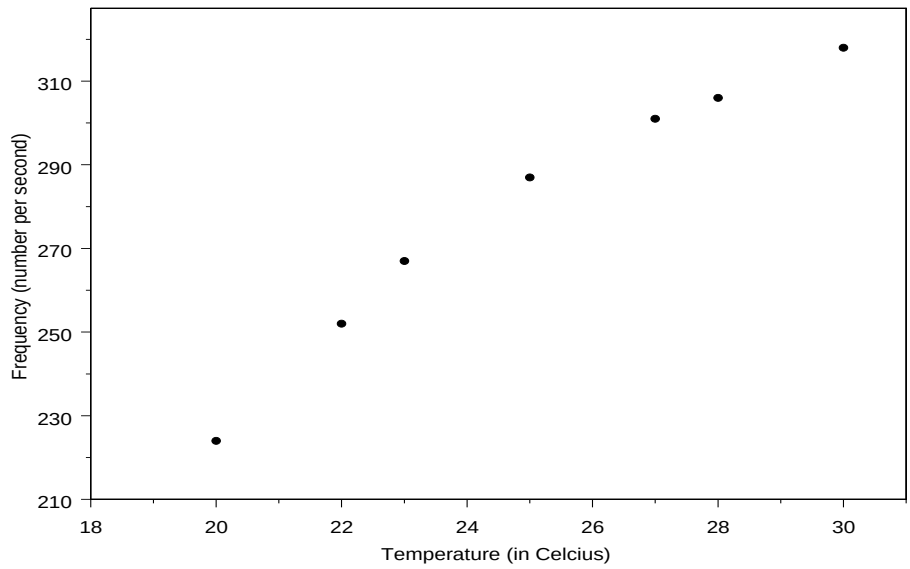



Figure 15.31: *Scatterplot for EI frequency at different temperatures.*

*ASSESSMENT*: There is a strong **positive** straight-line relationship between these two variables (temperature and EI frequency).

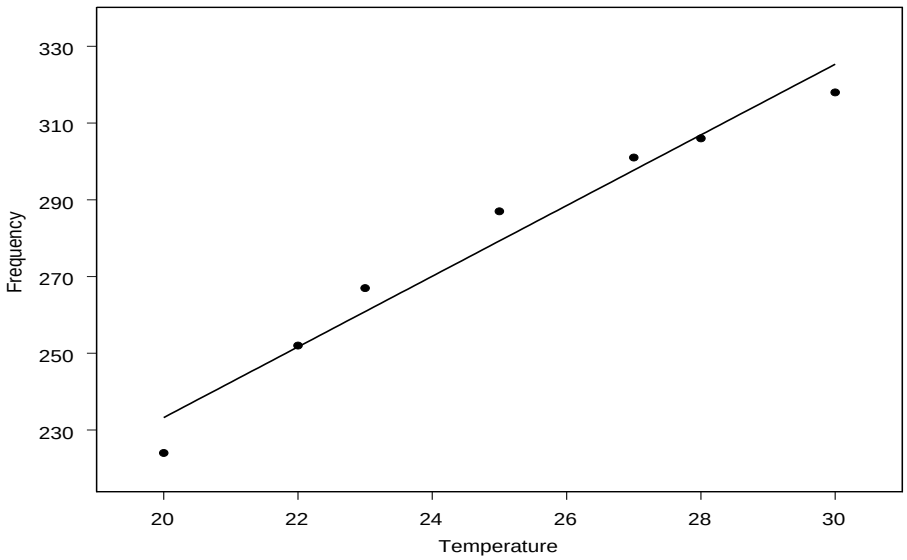



Figure 15.32: *EI frequency at different temperatures with a straight-line fit.*

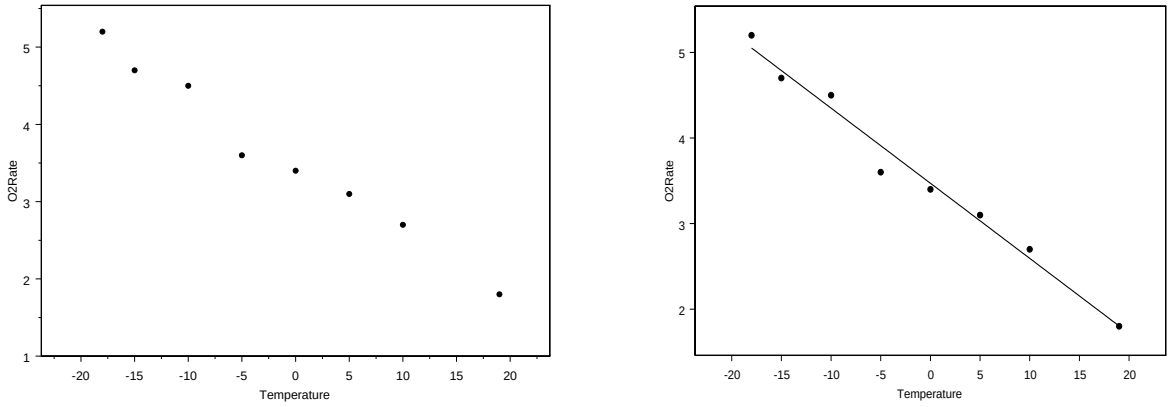
*TERMINOLOGY*: A **regression line** is a straight line that describes how a response variable $y$ changes as an explanatory variable $x$ changes.

*NOTE*: With real data (like above), it will be generally impossible to find a line that "goes through all the points." There will always be some observations which fall above the line; some below. More on this momentarily...

**Example 15.2.** These data are rates of oxygen consumption ($y$) for birds measured at different temperatures ($x$). The temperatures were set by the investigator, and the $O_2$ rates (ml/g/hr) were observed for these particular temperatures. The scatterplot of the data appears in Figure 15.33(a). The regression line appears in Figure 15.33(b).

| $x$, (degrees Celcius) | $-18$ | $-15$ | $-10$ | $-5$ | 0 | 5 | 10 | 19 |
|---|---|---|---|---|---|---|---|---|
| $y$, (ml/g/hr) | 5.2 | 4.7 | 4.5 | 3.6 | 3.4 | 3.1 | 2.7 | 1.8 |

*ASSESSMENT*: There is strong **negative** straight-line relationship between the two variables (temperature and oxygen consumption).




(a) Scatterplot. (b) With regression line.

Figure 15.33: *Bird oxygen rate data for different temperatures.*

*IDEA*: We often use a regression line to **predict** the value of the response variable $y$ for a given value of the explanatory variable $x$. For example, looking at the scatterplots (with straight-line fits),

- in Example 15.1, what would you predict the EI frequency to be when the temperature is 26 degrees C?
- in Example 15.2, what would you predict the $O_2$ consumption rate to be when the temperature is 2.5 degrees C?

## 15.2 Regression equations

*STRAIGHT LINE EQUATIONS*: Suppose that $y$ is the response variable (plotted on the vertical axis) and that $x$ is the explanatory variable (plotted on the horizontal axis). A **straight line** relating $y$ to $x$ has an equation of the form

$$y = a + bx,$$

where the constant $b$ represents the **slope** of the line and $a$ denotes the **$y$-intercept**.

*INTERPRETATION*: The **slope** of a regression line gives the amount by which the response $y$ changes when $x$ increases by one unit. The $y$**-intercept** gives the value of the response $y$ when $x = 0$.

*RECALL*: With real data, we will never be able to find a line which "fits perfectly;" i.e., that goes through all the points. *So, how do we find the equation of the regression line?*

*LEAST-SQUARES PRINCIPLE*: The **least-squares regression line** is the line which minimizes the sum of squared vertical distances of the data points from the line. There is only one line that does this!

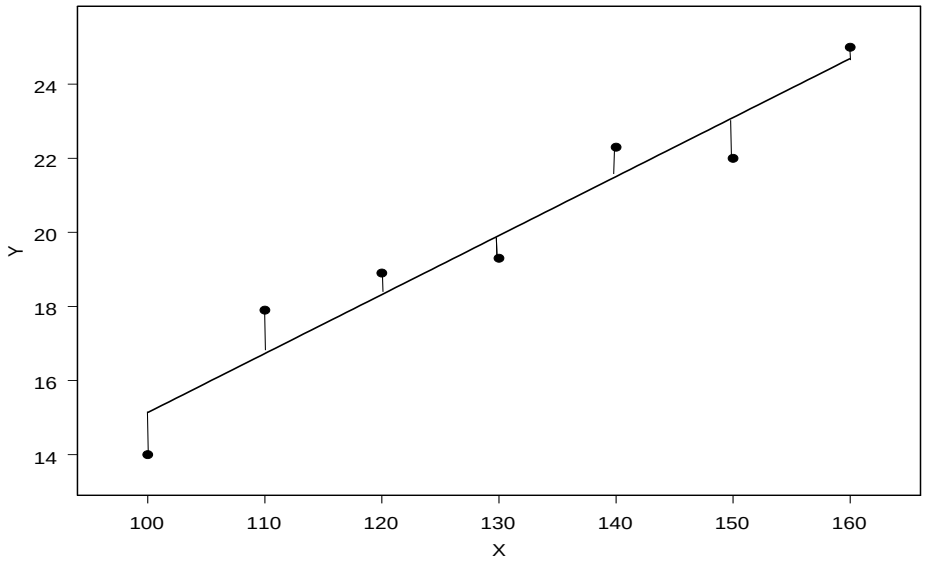



Figure 15.34: *A scatterplot with straight line and residuals associated with the straight line.*

*TERMINOLOGY*: Each observation has its own vertical distance from the regression line. We call these vertical distances **residuals**. See Figure 15.34.

*NOTE*: In this class, we will not give the formulas for computing the least-squares regression line; this is the job of a computer with appropriate software (e.g., Excel, Minitab). We will only focus on interpretation.

**Example 15.2** (continued). Using statistical software, I found the regression equation for the bird-oxygen rate data. It is

$$\texttt{O2Rate} = 3.47 - 0.0878\ \texttt{Temperature}.$$

*INTERPRETATION*:

- The slope $b = -0.0878$ is interpreted as follows: "for a one-unit (degree) increase in temperature, we would expect for the oxygen rate to **decrease** by 0.0878 ml/g/hr."
- The $y$-intercept $a = 3.47$ is interpreted as follows: "for a temperature of $x = 0$, we would expect the oxygen rate to be 3.47 ml/g/hr."

## 15.3 Prediction

**Example 15.3.** In Example 14.5 (notes), we examined the lengths (cm) of the femur (leg bone) and the humerus (upper arm bone) for five fossils for extinct animals (archaeopteryx). The data are given in Example 2 (page 273, MN). The equation of the least-squares regression line for these data is

$$\texttt{humerus} = -3.66 + 1.197\ \texttt{femur}.$$

In Figure 15.35, we see the scatterplot, the least-squares regression line superimposed.

QUESTION. What would you **predict** the humerus length to be for an animal whose femur length was 50 cm?

SOLUTION. Answering questions like this is easy. All we have to do is substitute "50" into the regression equation for femur ($x$) and solve for humerus ($y$). The calculation is

$$\texttt{humerus} = -3.66 + 1.197(50) = 56.2 \text{ cm};$$

that is, for a fossil with a femur 50 cm long, we would predict the humerus length to be 56.2 cm.

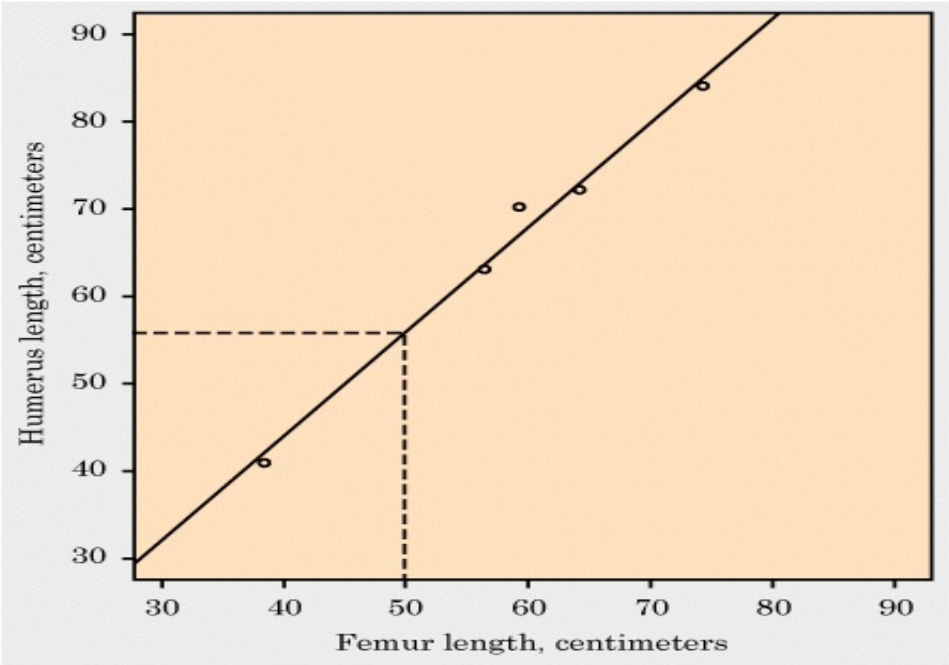



Figure 15.35: *Scatterplot of the lengths of two bones in* 5 *fossil specimens of the archaeopteryx. The least squares regression line is superimposed. The dotted line represents a prediction when* $x = 50$.

**Example 15.4.** For our bird-oxygen data in Example 15.2, suppose we wanted to predict a future oxygen consumption rate (for a new bird used in the experiment, say) when the temperature is set at $x = 2.5$ degrees. Using our regression equation, the prediction is

$$y = 3.47 - 0.0878(2.5) = 3.252 \text{ ml/g/hr}.$$

Thus, for a bird subjected to 2.5 degrees Celsius, we would expect its oxygen rate to be approximately 3.252 ml/g/hr.

*ADEQUACY*: Predictions are best when the model fits the data well. If our data are "well-fit" by a regression line, then our predictions are going to be good. If a straight line is not adequate for the data, then our predictions are not going to be very good.

*EXTRAPOLATION*: It is sometimes desired to make predictions based on the fit of the straight line for values of $x$ outside the range of $x$ values used in the original study. This is called **extrapolation**, and can be very dangerous. In order for our inferences to be

valid, we must believe that the straight line relationship holds for $x$ values **outside** the range where we have observed data.

## 15.4 Correlation and regression

*OBSERVATION*: Correlation and regression are both 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., $r = 0.84$.
- A regression line is a **mathematical equation** that describes the relationship.

Both correlation and regression are strongly affected by **outliers**; see Figure 15.36.

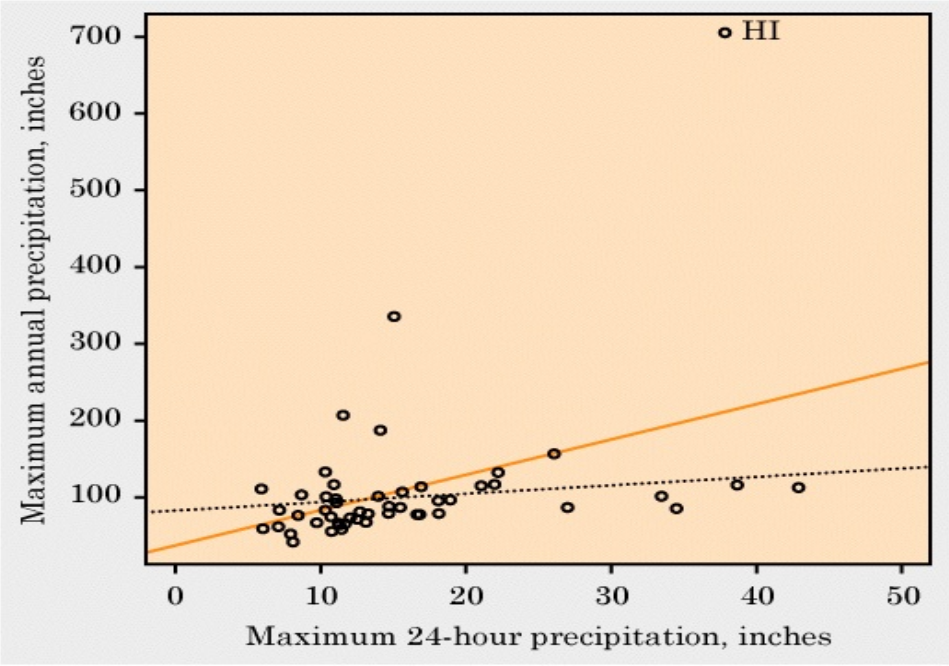



Figure 15.36: *Least-squares regression lines are strongly influenced by outliers. The solid line is based on all* 50 *states* ($r = 0.408$)*; the dotted line leaves out Hawaii* ($r = 0.195$).

*LINK*: The usefulness of the regression line for prediction depends on the strength of the straight-line association (which is measured by the correlation).

*SQUARE OF THE CORRELATION*: In a regression analysis, one way to measure how well a straight line fits the data is to compute the **square of the correlation** $r^2$. This statistic is interpreted as *the proportion of total variation in the data explained by the straight-line relationship with the explanatory variable.*

*NOTE*: Since $-1 \leq r \leq 1$, it must be the case that

$$0 \leq r^2 \leq 1.$$

Thus, an $r^2$ value "close" to 1 is often taken as evidence that the predictions made using the model are going to be adequate.

**Example 15.5.** With our bird-oxygen rate data from Example 15.2 I used statistical software to compute the correlation $r = -0.99$. The square of the correlation is

$$r^2 = (-0.99)^2 = 0.9801.$$

Thus, 98.01 percent of the variation in the oxygen rate data is explained temperature ($x$). This is a very high percentage! The other 1.99 percent is explained by other variables (not accounted for in our straight-line regression model). We could feel good about our predictions in this case.

## 15.5 Causation

*SMOKING*: Few would argue that cancer is not associated with persistent smoking. However, does smoking actually **cause** cancer? This is a more difficult question to answer. Here is a response I found at `http://www.quitsmokingsupport.com/questions.htm`.

> *"Yes. Tobacco smoke contains at least 43 carcinogenic (cancer-causing) substances. Smoking causes many kinds of cancer, not just lung cancer. Tobacco*

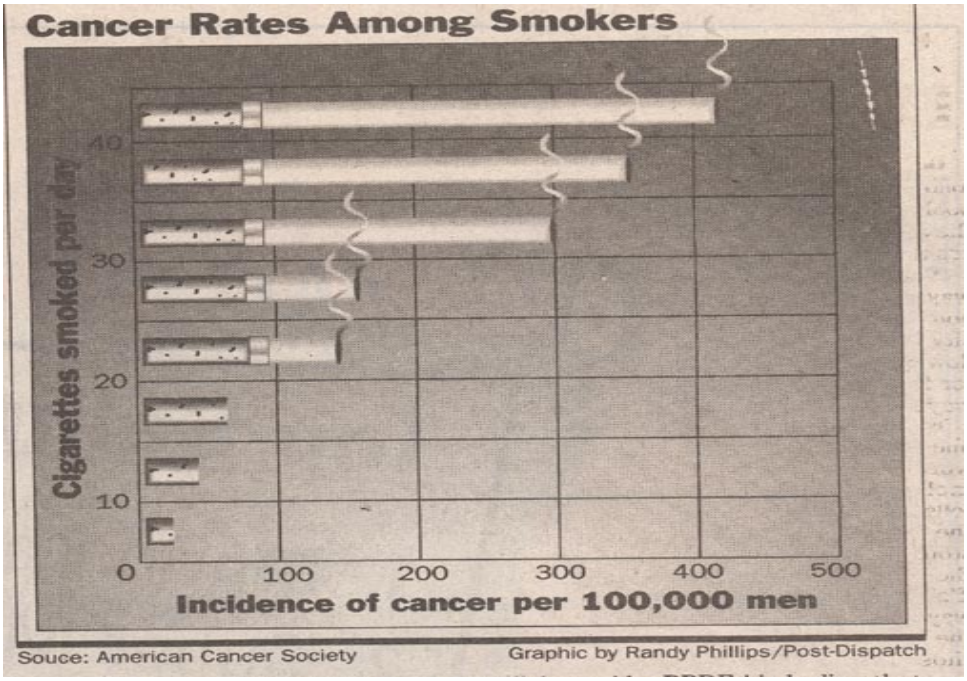



Figure 15.37: *A less-than-impressive graphical display which attempts to convey smoking behavior among men and its association with the incidence of cancer.*

> *use accounts for 30 percent, or one in three, of all cancer deaths in the United States. Smoking is responsible for almost 90 percent of lung cancers among men and more than 70 percent among women, about 83 percent overall."*

*ASSESSING CAUSATION*: Here are three important facts about statistical evidence for assessing causation:

1. A strong relationship between two variables does not necessarily mean that changes in one variable cause changes in the other.

2. The relationship between two variables is often influenced by other variables **lurking** in the background.

3. The best evidence for causation comes from **randomized comparative experiments**.

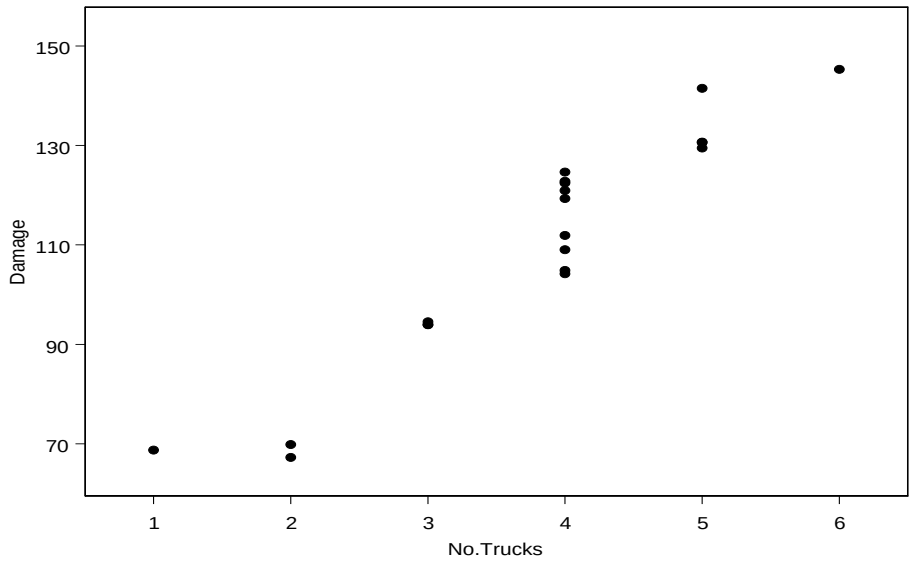



Figure 15.38: *Chicago fire damages* (\$1000$s$) *and the number of fire trucks.*

**Example 15.6.** A Chicago newspaper reported that "there is a strong correlation between the number of fire trucks ($x$) at a fire and the amount of damage ($y$, measured in \$1000's) that the fire does." Data from 20 recent fires in the Chicago area appear in Figure 15.38. From the plot, there appears to be a strong linear relationship between the number of fire fighters and the amount of damage. Few, however, would infer that the increase in the number of fire trucks actually **causes** the observed increase in damages!

*MORAL*: This phenomenon is the basis of the remark "*Correlation does not necessarily imply causation.*" An investigator should be aware of the temptation to infer causation in setting up a study, and be on the lookout for **lurking variables** that are actually the driving force behind observed results. In general, the best way to control the effects of lurking variables is to use a carefully **designed experiment**. Unfortunately, it may not be possible to perform such experiments when dealing with human subjects!

*REALITY*: In observational studies, it is very difficult to make **causal** statements. The best we can do is make statements documenting an association, and nothing more.

# 16 Thinking About Chance

Complementary reading from Moore and Notz (MN): Chapter 17.

**Note that we are not covering Chapter 16 (MN).**

## 16.1 Introduction

*CHANCE EVENTS IN REAL LIFE*: We are inundated everyday with numbers which quantify the **chance** of something occurring in the future (perhaps to us); e.g.,

- "The chance of winning the Powerball lottery is 1 in 56,000,000."
- "Duke is a 2:1 favorite to win the ACC men's basketball regular season championship."
- "For this cohort, subjects not wearing condoms are estimated to be 3 times more likely to contract an STD when compared to those subjects who wear condoms."
- "The chance of dying in a plane crash is 1 in 7,000,000."
- "The probability of transmission from aphid to host plant is 0.02."
- "As the course instructor, I give A's to the top 2 percent of the class."

## 16.2 Probability

*TERMINOLOGY*: We call a phenomenon **random** if individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions.

*TERMINOLOGY*: The **probability** of any outcome is the proportion of times the outcome would occur in a very long series of independent repetitions.

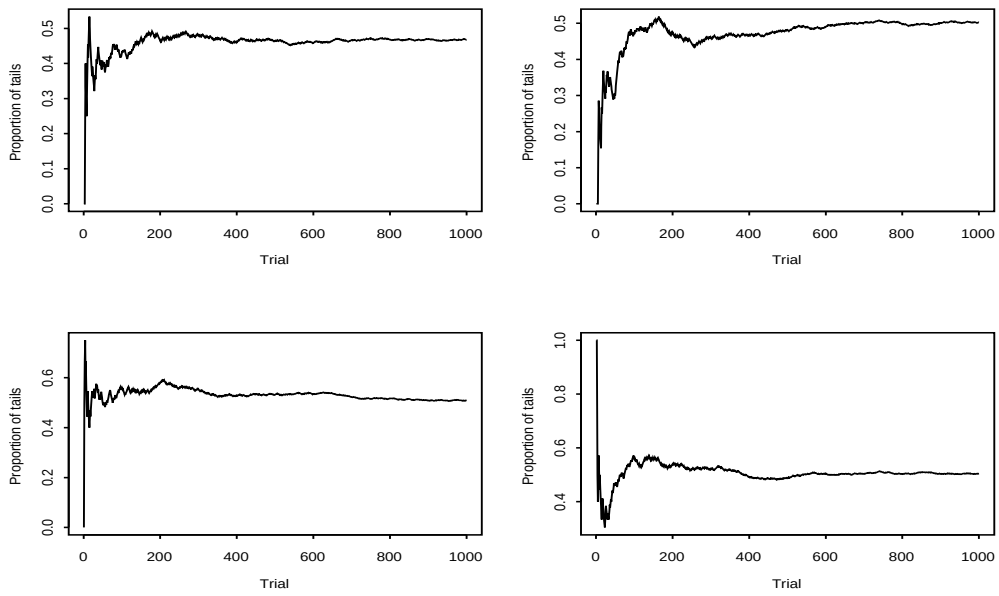



Figure 16.39: *The proportion of flips which result in a "tails"; each plot represents* $1,000$ *flips of a fair coin.*

*NOTE*: **A probability is a number between 0 and 1.**

- 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.

**Example 16.1.** *Flipping a coin.* Flip a coin. If it is unbiased, then each side (H, T) should occur with the same chance. And, over the long run (i.e., after many many flips), the proportion of "tails" which occur should be close to 1/2. However, the outcome on any one flip can not be predicted with certainty!

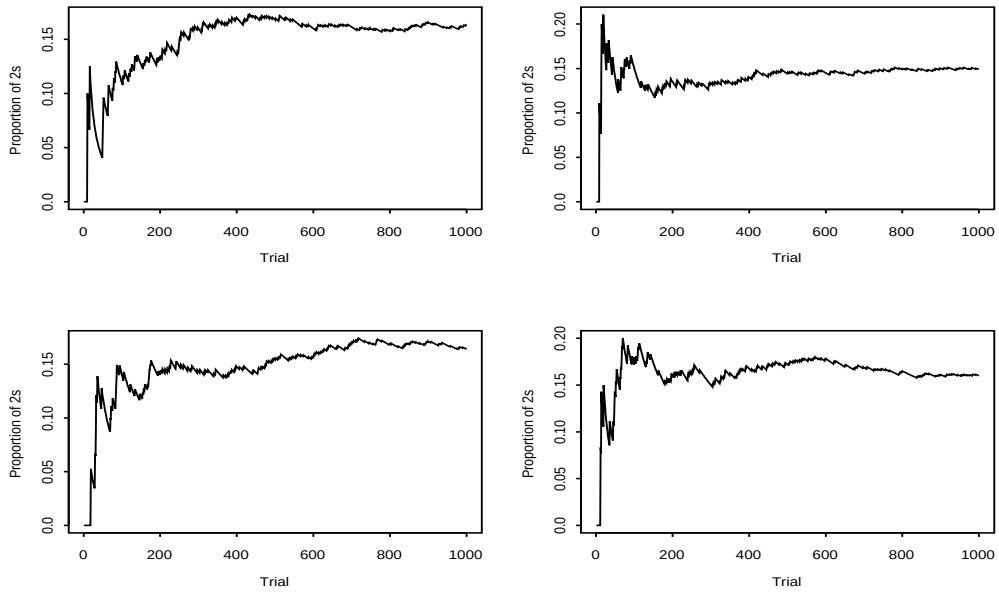



Figure 16.40: *The proportion of tosses which result in a "2"; each plot represents* $1,000$ *rolls of a fair die.*

**Example 16.2.** *Rolling a die.* Roll a die. If it is unbiased, then each side (1, 2, 3, 4, 5, 6) should occur with the same chance. And, over the long run (i.e., after many many rolls), the proportion of "2"'s which occur should be close to 1/6. But, it is impossible to know, with certainty, which face will appear on a particular roll.

*MORAL*: Chance behavior is unpredictable in the short run, but it has a regular and predictable pattern in the long run. Probability provides a language to describe the **long-term regularity** of random behavior.

## 16.3 Probability myths

*SHORT TERM REGULARITY*: Probability deals with long term regularity; not short term behavior. Consider each of the following situations where people try to place faith in the **myth of short-term regularity**.

- 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

  We walk to the table with this information. What should we bet on the next play?

- 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 (except in 1984), Harry would try to rally the crowd by making statements like the following:

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

- Empirical data suggests that the boys and girls are born at roughly the same rate (about 50/50). A couple's first three children are boys. What is the probability that the next borne will be a girl?

*SURPRISE EVENTS*: Certain events may be unlikely, but this doesn't mean that they can't occur! **Even outcomes with small probability occur!** Consider each of the following examples of "surprise events."

- *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).

- *Vioxx.* In Merck's VIGOR study, comparing Vioxx to naproxen (a nonsteroidal anti-inflammatory drug), there was a highly **statistically significant** five-fold increase in heart attacks in the overall rofecoxib group (0.5 percent) compared to the naproxen group (0.1 percent). This amounted to 20 heart attacks with rofecoxib (out of 4,047 patients) compared with 4 with naproxen (out of 4,029 patients). *If there is really no difference between the heart attack rates for the two drugs*, the chance this would occur (the 20-4 split in heart attacks) is approximately 0.00014, or about 1.4 in 1000.

- *Having the same birthday.* Chances are, there are two people sitting in this room with the same birthday. Does this sound surprising? Consider the following table of calculations:

Table 16.15: *The probability of having at least one common birthday in a group of $n$ people.*

| $n$ | Probability | $n$ | Probability |
|---|---|---|---|
| 2 | 0.0027397 | 28 | 0.6544615 |
| 4 | 0.0163559 | 30 | 0.7063162 |
| 6 | 0.0404625 | 32 | 0.7533475 |
| 8 | 0.0743353 | 34 | 0.7953169 |
| 10 | 0.1169482 | 36 | 0.8321821 |
| 12 | 0.1670248 | 38 | 0.8640678 |
| 14 | 0.2231025 | 40 | 0.8912318 |
| 16 | 0.2836040 | 50 | 0.9703736 |
| 18 | 0.3469114 | 60 | 0.9941227 |
| 20 | 0.4114384 | 70 | 0.9991596 |
| 22 | 0.4756953 | 80 | 0.9999143 |
| 24 | 0.5383443 | 90 | 0.9999938 |
| 26 | 0.5982408 | 100 | 0.9999997 |

## 16.4 Law of averages

*THE LAW OF AVERAGES*: As the number of repetitions increases (i.e., in the long run), the **sample proportion** $\widehat{p}$ of successes approaches the true probability of success $p$. We have seen this empirically in Examples 16.1 and 16.2.

*CONNECTION WITH SAMPLING*: In earlier chapters, we discussed the notion of using a **sample proportion** $\widehat{p}$ to estimate the **population proportion** $p$ using simple random sampling. The Law of Averages states that as the sample size $n$ increases, the sample proportion $\widehat{p}$ approaches the true population proportion $p$.

**Example 16.3.** A Fox News poll, taken on June 29, 2006, reported the results from an SRS of $n = 900$ adults nationwide. Interviewers asked the following question: *"Do you*

*approve or disapprove of the way George W. Bush is handling his job as president?"* Of the 900 adults in the sample, 369 responded by stating they approve of the President's handling of Katrina. The sample proportion of those who support President Bush is

$$\widehat{p} = \frac{369}{900} \approx 0.41.$$

Recall that the **margin of error** is approximately

$$\frac{1}{\sqrt{900}} \approx 0.033.$$

With $n = 2,000$ persons, our margin of error would have been 0.022. With $n = 5,000$ persons, the margin of error is 0.014. With $n = 15,000$, the margin of error is 0.008.

*MORAL*: As the sample size increases, the margin of error decreases; i.e., $\widehat{p}$ gets closer to the true $p$. This is precisely what the **Law of Averages** says should happen!

## 16.5 Personal probability assignment

**Example 16.4.** At this point in the class, what do you think your chance is of earning a 'B' or higher in this class? Express your answer as a **probability** (i.e., a number between 0 and 1).

*QUESTION*: The number you've written is a probability, but how does the notion of probability in this case differ from the notion of long-term regularity?

*TERMINOLOGY*: 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, hence, is often not scientific)
- Long term regularity interpretation: based on the notion of **repeated trials**; e.g., "what happens in the long term?"

# 17 Introduction to Statistical Inference and Sampling Distributions

Complementary reading from Moore and Notz (MN): Chapter 18 (only pages 373-377).

## 17.1 Introduction

*RECALL*: **Statistical inference** deals with drawing conclusions about a population on the basis of data from a sample.

- A Columbia-based health club wants to estimate the proportion of Columbia residents who enjoy running as a means of cardiovascular exercise. In a random sample of $n = 100$ residents, 23 percent (i.e., $\widehat{p} = 0.23$) said they enjoy running. What does this suggest about the population proportion $p$?
- *Does Dramamine reduce seasickness?* A random sample of $n = 500$ Navy servicemen received the drug to control seasickness while at sea, and 34 percent (i.e., $\widehat{p} = 0.34$) experienced some form of motion sickness. What does this suggest about the population proportion ($p$) of Navy sea servicemen who experience motion sickness?
- An education major would like to estimate the mean GPA for undergraduate pre-nursing students in South Carolina. A random sample of $n = 20$ students produces a sample mean $\overline{x} = 3.39$. What does this say about the population mean $\mu$ of all undergraduate pre-nursing students?
- In a large scale Phase III clinical trial involving patients with advanced lung cancer, the goal is to determine the efficacy of a new drug. Ultimately, we would like to know whether this drug will extend the life of lung cancer patients as compared to current available therapies. What do the results from the trial suggest about the population of individuals with advanced lung cancer?

- The National Collegiate Athletic Association (NCAA) requires colleges to report the graduation rates of their athletes. Here are the data from a Big 10 University Report. Student athletes admitted during 1989-1991 were monitored for six years. The (six-year) graduation rates were 37 of 45 female athletes and 58 of 102 male athletes. Is there a true (statistically significant) difference between the genders with respect to graduation rates?

*RECALL*: In Chapter 3, we talked about confidence statements; e.g.,

> *We are "very confident" that the proportion of all Navy sea servicemen who experience seasickness, while taking Dramamine, is between* 30% *and* 38%.

*RECALL*: A **confidence statement** has two parts:

- A **margin of error.** This quantifies how close the sample statistic is to the population parameter.
- A **level of confidence**. This says what percentage of possible samples satisfy the margin of error.

*NOTES*: The term "very confident" has to do with the notion of **repeated sampling**:

- In Chapter 3, we took the phrase "very confident" to mean "95 percent confident."
- That is, if we took many samples using the same method, 95 percent of the time we would get a result within the margin of error.
- 95 percent of time, we will be close to the truth (i.e., close to the true population parameter); that is, we will be inside the margin of error.
- 5 percent of the time, we will "miss" by more than the margin of error.

*TERMINOLOGY*: A **95 percent confidence interval** is an interval calculated from sample data by a process that is guaranteed to capture the true population parameter in 95 percent of all samples.

*OUR GOAL*: *We aspire to construct confidence intervals for population parameters.* These intervals help us make statements about the population of individuals under investigation (consider each of the examples stated at the beginning of this chapter).

## 17.2 Sampling distributions

**Example 17.1.** A Columbia-based health club wants to estimate the proportion of Columbia residents who enjoy running as a means of cardiovascular exercise.

$p$ = the true proportion of Columbia residents who enjoy running (unknown)

$\widehat{p}$ = the proportion of residents who enjoy running observed in our sample.

*SIMULATION*: We know that statistics' values will change from sample to sample. A natural question is *"What would happen if I took many samples?"* We can answer this question by using **simulation**. Simulation studies are commonly used by statisticians to determine "what if?" questions like this.

*SIMULATION*: Here is what happened when I simulated 100 different values of $\widehat{p}$ under the assumption that $p = 0.2$. Each sample proportion is based on an SRS of $n = 100$.

```
0.17 0.16 0.20 0.20 0.16 0.25 0.21 0.17 0.14 0.26 0.12 0.20 0.18
0.15 0.28 0.20 0.21 0.24 0.25 0.22 0.18 0.20 0.19 0.16 0.21 0.21
0.25 0.18 0.20 0.14 0.14 0.21 0.20 0.17 0.18 0.15 0.21 0.12 0.18
0.23 0.18 0.22 0.26 0.18 0.13 0.19 0.17 0.28 0.18 0.21 0.22 0.18
0.19 0.22 0.23 0.17 0.26 0.21 0.19 0.19 0.20 0.10 0.17 0.18 0.18
0.18 0.21 0.20 0.23 0.23 0.26 0.18 0.18 0.16 0.24 0.22 0.16 0.21
0.27 0.18 0.19 0.26 0.25 0.24 0.10 0.18 0.18 0.25 0.18 0.21 0.20
0.21 0.20 0.18 0.22 0.19 0.26 0.17 0.16 0.20
```

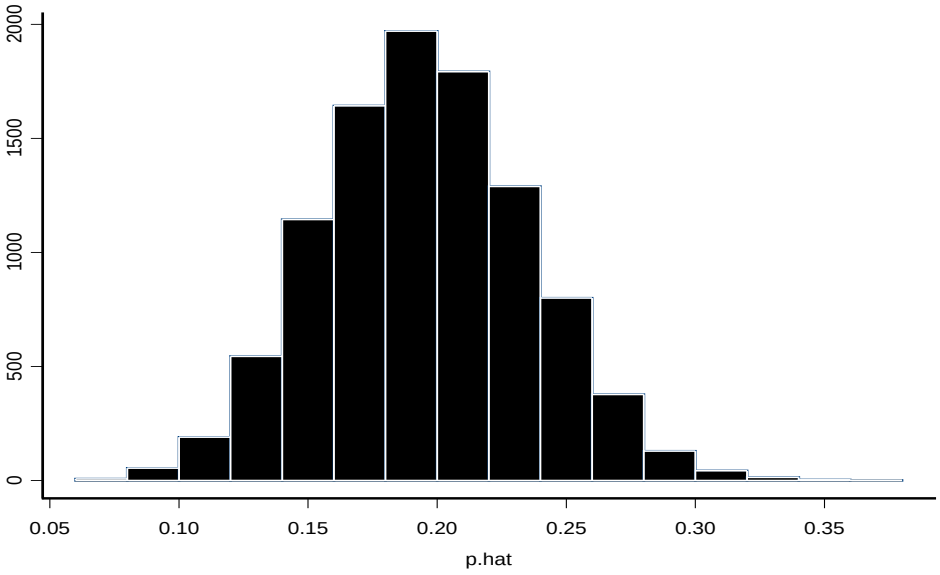



Figure 17.41: *Sampling distribution of the sample proportion* $\widehat{p}$ *when* $p = 0.2$.

*USING HISTOGRAMS*: Another natural question is, *"What does the distribution of* $\widehat{p}$ *values look like?"* To answer this, we can do the following:

1. Take a large number of samples assuming that $p = 0.2$, and calculate the sample proportion $\widehat{p}$ for each sample (we have done this above).
2. Make a **histogram** of the values of $\widehat{p}$.
3. Examine the distribution displayed in the histogram.

*RESULTS*: When I did this using 10,000 simulated random samples, each of size $n = 100$, I got the histogram in Figure 17.41.

*SAMPLING DISTRIBUTIONS*: The **sampling distribution** of a statistic is the distribution of values taken by the statistic in repeated sampling using samples of the same size. Figure 17.41 portrays the sampling distribution of the sample proportion $\widehat{p}$ when $n = 100$ and $p = 0.2$.

*REMARK*: Sampling distributions are important distributions in statistical inference. As with any distribution, we are interested in the following:

- **center** of the distribution
- **spread** (variation) in the distribution
- **shape**: is the distribution symmetric or skewed?
- the presence of **outliers**.

### 17.2.1 Unbiased estimators

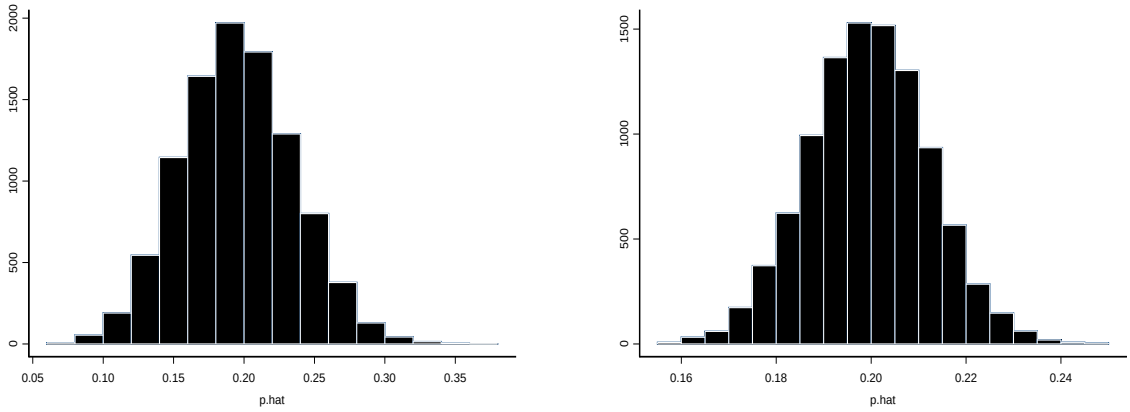
*TERMINOLOGY*: A statistic is said to be an **unbiased estimator** for a parameter if the mean of the statistic's sampling distribution is equal to the parameter. If the mean of the statistic's sampling distribution is not equal to this parameter, the statistic is called a **biased estimator**.

> *It should be clear that bias (or lack thereof) is a property that concerns the* ***center*** *of a sampling distribution.*

*MATHEMATICAL FACT*: If we use simple random sampling from a large population, the sample proportion $\widehat{p}$ is an **unbiased estimator** for the true population proportion $p$. This fact would be proven in a more mathematical course.

### 17.2.2 Variability

*TERMINOLOGY*: The **variability of a statistic** is described by the spread in the statistic's sampling distribution. This variability will always depend on the sample size $n$. In SRS designs, larger sample sizes correspond to smaller variability. This is desirable! We can increase our precision by taking larger samples!




(a) $n = 100$ (b) $n = 1000$

Figure 17.42: *Sampling distribution of the sample proportion when* $p = 0.2$. *Note how the variability associated with the sampling distribution based on* $n = 1000$ *observations is much smaller than that based on* $n = 100$ *observations.*

*IDEAL SITUATION*: Better statistics are those which have small (or no) bias and small variability.

- If a statistic is **unbiased** (or has very low bias) then we know that we are approximately "right on average."
- If we have a statistic with **small variability**, then we know there is not a large spread in that statistic's sampling distribution.
- The combination of "right on average" and "small variation" is the ideal case!

*SUMMARY*: To reduce bias, use random sampling. SRS designs produce unbiased estimates of population parameters. To reduce variation, use a larger sample size $n$. The variation in the sampling distribution decreases as $n$ increases.

*GOING FORWARD*: Now that we have discussed sampling distributions, we can move forward to the issue of **confidence intervals**.

# 18 Confidence Intervals for Proportions

Complementary reading from Moore and Notz (MN): Chapter 21 (only pages 425-436). **Note that we are not covering Chapters 19 and 20 (MN).**

## 18.1 Sampling distribution for the sample proportion

*GOAL*: Our goal is to come up with a formula we can use to compute a 95 percent confidence interval for $p$, the **population proportion**.

*RECALL*: In the last chapter, we used **simulation** to generate the **sampling distribution** of the sample proportion $\widehat{p}$ when $n = 100$ and $p = 0.2$; see Figure 18.43.

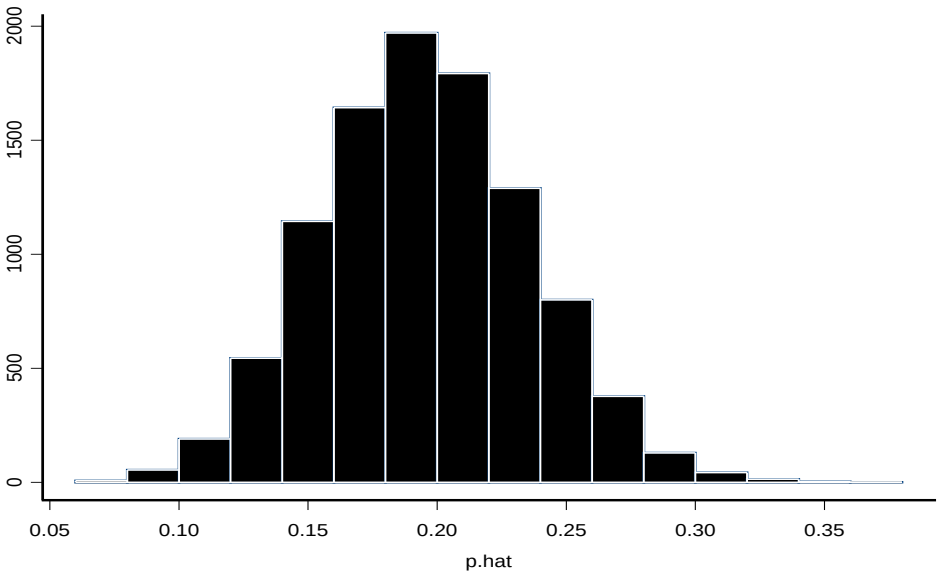



Figure 18.43: *Sampling distribution of the sample proportion $\widehat{p}$ when $p = 0.2$.*

*RECALL*: The **sampling distribution** of a statistic is the distribution of values taken by the statistic in repeated sampling using samples of the same size.

*DESCRIPTION*: In Figure 18.43, we make the following observations:

- The **center** of the sampling distribution looks to be right around $p = 0.2$.
- The **spread** of the sampling distribution ranges from about 0.05 to 0.30.
- The distribution **shape** is very symmetric; in fact, it looks like the histogram would be very well approximated by a normal density curve!
- There are no **outliers**.

*IMPORTANT RESULT*: Take an SRS of size $n$ from a large population of individuals, where $p$ denotes the population proportion ($p$ is an unknown **parameter**). Let $\widehat{p}$ denote the **sample proportion**; i.e.,

$$\widehat{p} = \frac{\text{number of sample successes}}{n}.$$

Note that $\widehat{p}$ is a **statistic** because it is computed from the sample. For large samples (i.e., for large $n$),

- The sampling distribution of $\widehat{p}$ is **approximately normal**.
- The **mean** of the sampling distribution is $p$; i.e., $\widehat{p}$ is an **unbiased** estimator of $p$.
- The **standard deviation** of the sampling distribution is

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

Putting this altogether, we have that

$$\widehat{p} \sim \mathcal{AN}\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

The symbol $\mathcal{AN}$ is read "approximately normal."

*REMARK*: This result is appropriate only when we have an **SRS** from a large population and the sample size $n$ is **large**. The larger the sample size, the better the normal density curve approximates the sampling distribution of $\widehat{p}$.

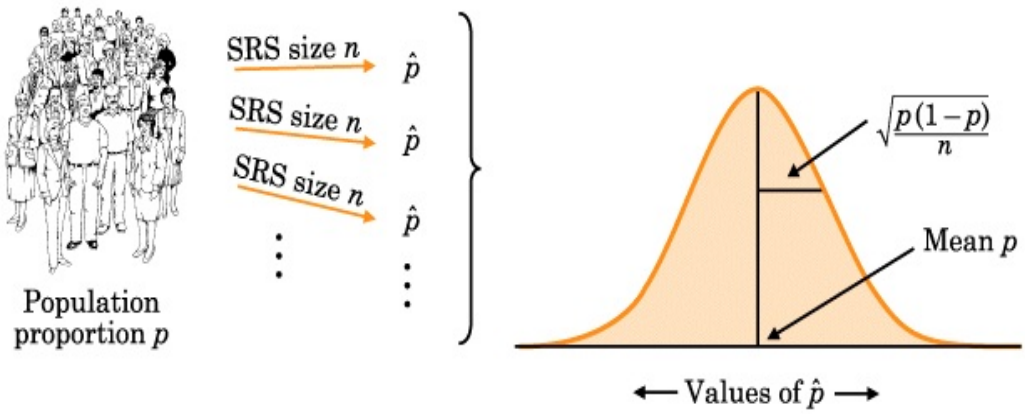



Figure 18.44: *The values of the sample proportion of successes $\widehat{p}$ follow an approximate normal distribution.*

## 18.2 Confidence intervals for a population proportion

**Example 18.1.** The Women's Interagency HIV Study (WIHS) is a longitudinal, multicenter study funded by the National Institutes of Health to investigate the effect of HIV infection in women. The WIHS collaborative study group includes 6 clinical consortia in Brooklyn, Chicago, Washington DC, San Francisco, Los Angeles, and Honolulu. An *American Journal of Public Health* article (published in April, 2000) reports that a total of 1,288 HIV-infected women were recruited in this study to examine the prevalence of childhood abuse. Of the 1,288 HIV positive women, a total of 399 reported that, in fact, they had been a victim of childhood abuse. Treating these 1,288 women as the **sample**, the value of the **sample proportion** of childhood abuse victims is

$$\widehat{p} = \frac{399}{1288} \approx 0.31,$$

This is an estimate of $p$, the **population proportion** of childhood abuse victims among HIV-infected women living in the United States. In this application, the sample size is quite large, so the normal approximation described earlier is probably very good.

*So, how do we construct the confidence interval for $p$?*

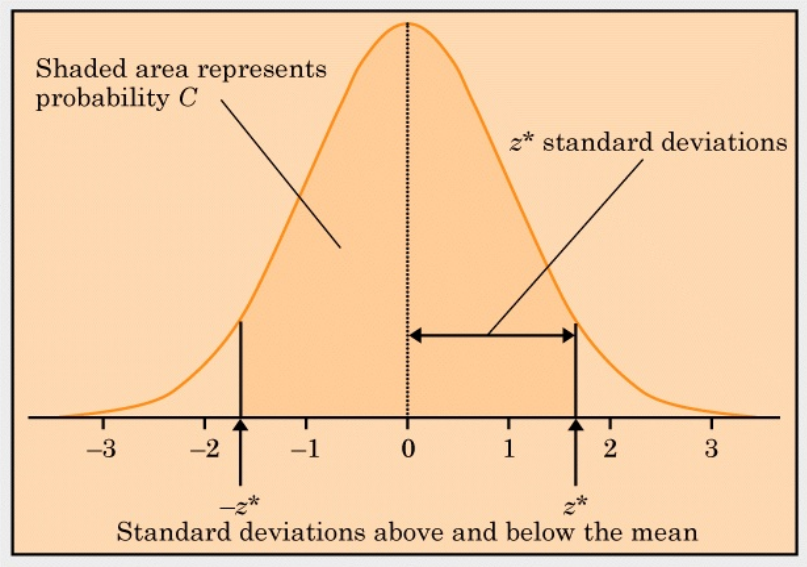



Figure 18.45: *Critical values $z^*$ from the standard normal distribution. The shaded region represents $C$ percent.*

*TERMINOLOGY*: A **level $C$ confidence interval** is an interval calculated from sample data by a process that is guaranteed to capture the true population parameter in $C$ percent of all samples.

- The parameter is $p$, the population proportion.
- $C$ is a percentage between 0 and 100.
- We usually take $C$ to be a large percentage; e.g., $C = 80$, $C = 90$, $C = 95$, and $C = 99$ are commonly used. Larger $C \Longrightarrow$ more confidence.

*MAIN RESULT*: Choose an SRS of size $n$ from a large population of individuals where $p$ denotes the **population proportion** of interest. Also, as usual, let $\widehat{p}$ denote the sample proportion. For large sample sizes, an approximate level $C$ confidence interval for $p$ 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 $C$. See Figure 18.45.

*INTERPRETATION*: When we compute this interval, we say that "we are $C$ percent confident that

$$\widehat{p} \pm z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},$$

contains the true population proportion $p$."

*FINDING* $z^*$: The value $z^*$ is a percentile of the standard normal distribution! Here are the values of $z^*$ for commonly used confidence levels.

| $C$ | 80 | 90 | 95 | 99 |
|---|---|---|---|---|
| $z^*$ | 1.28 | 1.64 | 1.96 | 2.58 |

*REMARKS*: Some very important points are worth noting.

- Note that, for $C = 95$ percent, $z^* = 1.96$ is very close to 2 (which is what we would expect from the **Empirical Rule**).
- As the level of confidence $C$ increases, the value of $z^*$ increases as well. This increases the margin of error, and thus, the length of the interval. This makes sense intuitively; namely, lengthier intervals are more likely (we're more confident) to capture the true population parameter $p$.
- In the long run, $C$ percent of our intervals will contain the true population proportion; see Figure 18.46.

**Example 18.1** (continued). Let's compute a 95 percent confidence interval for $p$, the true **population proportion** of childhood abuse victims among HIV-infected women living in the United States. The sample proportion of childhood abuse victims is

$$\widehat{p} = \frac{399}{1288} \approx 0.31.$$

The value of $z^* = 1.96$. The margin of error, $m$, is

$$m = z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 1.96 \sqrt{\frac{0.31(1-0.31)}{1288}} \approx 0.03.$$

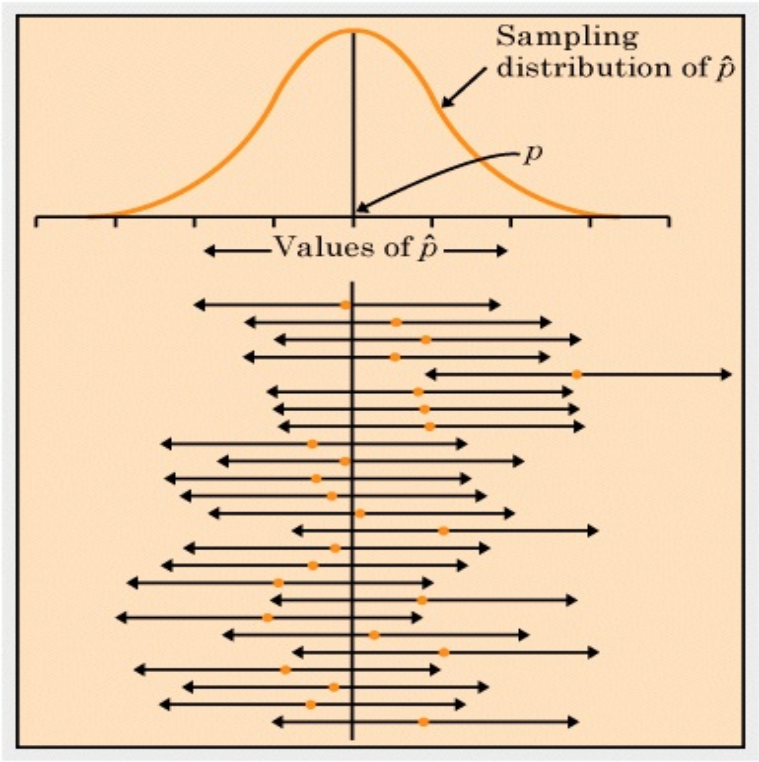



Figure 18.46: *Twenty five samples from the same population give* 95 *percent confidence intervals. In the long run,* 95 *percent of the intervals will contain the true population proportion* $p$*, marked by the vertical line.*

Thus, a 95 confidence interval for $p$ is

$$0.31 \pm 0.03 \Longrightarrow (0.28, 0.34).$$

*INTERPRETATION*: Based on our sample, we are 95 percent confident that the true proportion of childhood abuse victims among American HIV-positive women is between 0.28 and 0.34 (i.e., between 28 and 34 percent).

**Example 18.2.** Apple trees in a large orchard are sprayed in order to control moth injuries to apples growing on the trees. A random sample of $n = 1000$ apples is taken from the orchard, 150 of which are injured. Find a 95 percent confidence interval for $p$, the true population proportion of injured apples in the orchard.

SOLUTION. The **sample proportion** of injured apples is

$$\widehat{p} = \frac{150}{1000} = 0.15.$$

The value of $z^* = 1.96$. The margin of error, $m$, is

$$m = z^*\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 1.96\sqrt{\frac{0.15(1-0.15)}{1000}} \approx 0.02.$$

A 95 percent confidence interval for $p$, the population proportion of injured apples, is

$$0.15 \pm 0.02 \Longrightarrow (0.13, 0.17).$$

*INTERPRETATION*: Based on our sample, we are 95 percent confident that the true proportion of injured apples in this orchard is between 0.13 and 0.17 (i.e., between 13 percent and 17 percent).

### 18.2.1 A closer look at the confidence interval form

*CONFIDENCE INTERVAL FORM*: Let's look at the formula for a **95 percent** confidence interval for the population proportion $p$. This is given by

$$\widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

The form of this interval is ubiquitous in statistical inference; in particular, note how the interval is computed by taking

$$\text{estimate} \pm \text{margin of error}.$$

Here, the **margin of error** is approximately equal to

$$m \approx 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

*INSIGHT*: We know that

$$\begin{aligned} \widehat{p} \text{ is always between 0 and 1} &\Longrightarrow \widehat{p}(1-\widehat{p}) \text{ is always between 0 and 0.25} \\ &\Longrightarrow \sqrt{\widehat{p}(1-\widehat{p})} \text{ is always between 0 and 0.5} \\ &\Longrightarrow 2\sqrt{\widehat{p}(1-\widehat{p})} \text{ is always between 0 and 1} \end{aligned}$$

*DISCOVERY*: Using "1" as a conservative upper bound estimate for $2\sqrt{\widehat{p}(1-\widehat{p})}$ we get

$$m \approx 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \approx \sqrt{\frac{1}{n}}.$$

This should look familiar. This was our "quick formula" for the margin of error associated with $\widehat{p}$ back in Chapter 3!!!

*MOVING FORWARD*: From now on, we can be more exact when we make confidence statements regarding $p$ and use

$$\widehat{p} \pm z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

We no longer need to use the "quick formula" for margin of error as we did before.

**Example 18.3.** In a Phase III clinical trial involving $n = 1200$ subjects taking a new medication for sleep deprivation, 30 of them experienced some form of nausea. Using this sample information, find a 99 percent confidence interval for $p$, the population proportion of patients who will experience nausea. Interpret the interval.

SOLUTION. The value of the sample proportion of those experiencing nausea is

$$\widehat{p} = \frac{30}{1200} = 0.025.$$

The value of $z^* = 2.58$. The margin of error, $m$, is

$$m = z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 2.58\sqrt{\frac{0.025(1-0.025)}{1200}} \approx 0.012.$$

Thus, a 95 percent confidence interval for $p$ is given by

$$0.025 \pm 0.012 \Longrightarrow (0.013, 0.037).$$

*INTERPRETATION*: Based on our sample, we are 99 percent confident that $p$, the true proportion of subjects who will experience nausea from the medication, is between 0.013 and 0.037 (i.e., between 1.3 and 3.7 percent).

EXERCISE: Find a 90 percent confidence interval for $p$. What do you note about the length of this interval when compared to the 99 percent interval?

### 18.2.2 Choosing a sample size

*CHOOSING A SAMPLE SIZE*: Before one launches into a research investigation where data are to be collected, inevitably one starts with the simple question:

*"How many observations do I need?"*

The answer almost always depends on the resources available (e.g., time, money, space, etc.) and statistical issues like **confidence** and **margin of error**.

*RESULT*: To determine an appropriate sample size for estimating $p$ with a level $C$ confidence interval, we need to specify the **margin of error** that we desire; i.e.,

$$m = z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

We would like to solve this equation for $n$. However, note that $m$ depends on $\widehat{p}$, which, in turn, depends on $n$. This is a small problem, but we can overcome the problem by replacing $\widehat{p}$ with $p^*$, a **guess** for the value of $p$. Doing this, the last expression becomes

$$m = z^* \sqrt{\frac{p^*(1-p^*)}{n}},$$

and solving this equation for $n$, we get

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*).$$

This is the desired sample size to find a level $C$ confidence interval for $p$ with a prescribed margin of error equal to $m$.

**Example 18.4.** In a Phase II clinical trial, it is posited that the proportion of patients responding to a certain drug is $p^* = 0.4$. To engage in a larger Phase III trial, the researchers would like to know how many patients they should recruit into the study. Their resulting 95 percent confidence interval for $p$, the true population proportion of patients responding to the drug, should have a margin of error no greater than $m = 0.03$.

What **sample size** do they need for the Phase III trial?

SOLUTION. Here, we have $m = 0.03$, $p^* = 0.4$, and $z^* = 1.96$. The desired sample size is

$$\begin{aligned} n &= \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) \\ &= \left(\frac{1.96}{0.03}\right)^2 0.4(1-0.4) \approx 1024.43. \end{aligned}$$

Thus, their Phase III trial should recruit around 1025 patients.

EXERCISE: With the information from Example 18.4, determine the necessary sample size for a 99 percent confidence interval (still take $m = 0.03$ and $p^* = 0.4$). Is this larger or smaller than that corresponding to 95 percent? Why?

*CONSERVATIVE APPROACH*: If there is no sensible guess for $p$ available, use $p^* = 0.5$. In this situation, the resulting value for $n$ will be as large as possible. Put another way, using $p^* = 0.5$ gives the most **conservative** solution (i.e., the largest sample size).

**Example 18.5.** To gauge the public's opinion on the alleged "greenhouse effect" (i.e., the accumulation of $CO_2$ in the atmosphere caused by burning fossil fuels like oil, coal, and natural gas), a researcher would like to take a random sample of individuals. She would like to write a 95 percent confidence interval for $p$, the population proportion of individuals who believe the greenhouse effect is a serious problem. But, she would like her margin of error to be no greater than 0.01. How many individuals does she need to contact?

SOLUTION. Here, we have $m = 0.01$, $p^* = 0.5$ (no prior guess available), and $z^* = 1.96$. The desired sample size is

$$\begin{aligned} n &= \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) \\ &= \left(\frac{1.96}{0.01}\right)^2 0.5(1-0.5) = 9604. \end{aligned}$$

Thus, she would need to contact 9,604 individuals!

EXERCISE: With the information from Example 18.5, determine the necessary sample size for a 95 percent confidence interval with margin of error $m = 0.05$. Is this larger or smaller than the sample size computed with $m = 0.01$?

# 19 Confidence intervals for means

Complementary reading from Moore and Notz (MN): Chapter 21 (only pages 436-440).

## 19.1 Introduction

*REVIEW*: In the last chapter, we discussed sampling distributions and confidence intervals for **proportions**. Proportions summarize **categorical** data. We found out that, in certain instances, the **sample proportion** $\widehat{p}$ follows an approximate normal distribution. We used this fact to construct confidence intervals for $p$, the **population proportion**. For example, we might want to write a confidence interval for $p$, where $p$ denotes

- the true proportion of SC voters supporting Mark Sanford
- the true proportion of Iowa cattle suffering from some disease
- the true proportion of Roger Federer's successful first serves
- the true proportion of childhood abuse victims among American HIV+ women.

*NOW*: To summarize **quantitative** data, we use statistics like the sample mean, the sample standard deviation, etc. These statistics have sampling distributions too, and confidence intervals follow directly from these distributions. In this chapter, we pay attention to the sampling distribution of the **sample mean** $\overline{x}$. We will then use this distribution to construct level $C$ confidence intervals for the **population mean** $\mu$. For example, we might want to write a confidence interval for $\mu$, where $\mu$ denotes

- the true mean GPA for all USC undergraduate pre-nursing students
- the true mean monthly income among all university professors
- the true mean yield in an agricultural experiment
- the true mean IQ among all South Carolina high school seniors.

## 19.2 Sampling distribution of the sample mean $\overline{x}$, CLT

*FACTS*: Choose an SRS of size $n$ from a population in which individuals have mean $\mu$ and standard deviation $\sigma$. Let $\overline{x}$ denote the sample mean. Then,

- The sampling distribution of $\overline{x}$ is **approximately normal** when the sample size $n$ is large.
- The mean of the sampling distribution is equal to $\mu$; that is, $\overline{x}$ is an **unbiased estimator** of $\mu$.
- The standard deviation of the sampling distribution is $\sigma/\sqrt{n}$.

*SUMMARIZING*: Putting this all together, we have that

$$\overline{x} \sim \mathcal{AN}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

The symbol $\mathcal{AN}$ is read "approximately normal." The preceding list of facts comes from a very powerful result in theoretical statistics; this result is called the **Central Limit Theorem**. Succinctly put, this theorem says that "averages are approximately normal."

*REMARK*: Note that the **precision** with which $\overline{x}$ estimates $\mu$ improves as the sample size increases. This is true because the standard deviation of $\overline{x}$ gets smaller as $n$ gets larger! Thus, the sample mean $\overline{x}$ is a very good estimate for the population mean $\mu$; it has the desirable properties that a statistic should have; namely, **unbiasedness** and the potential to have **small variability** for large sample sizes.

**Example 19.1.** Animal scientists are interested in the proximate mechanisms animals use to guide movements. The ultimate bases for movements are related to animal adaptations to different environments and the development of behaviors that bring them to those environments. As part of a project, we are measuring the distance (in meters) that a banner-tailed kangaroo rat moves from its birth site to the first territorial vacancy. Suppose that the density curve for these distance measurements has mean $\mu = 10$, has standard deviation $\sigma = 3.2$, and is strongly skewed right; see Figure 19.47 (upper left).

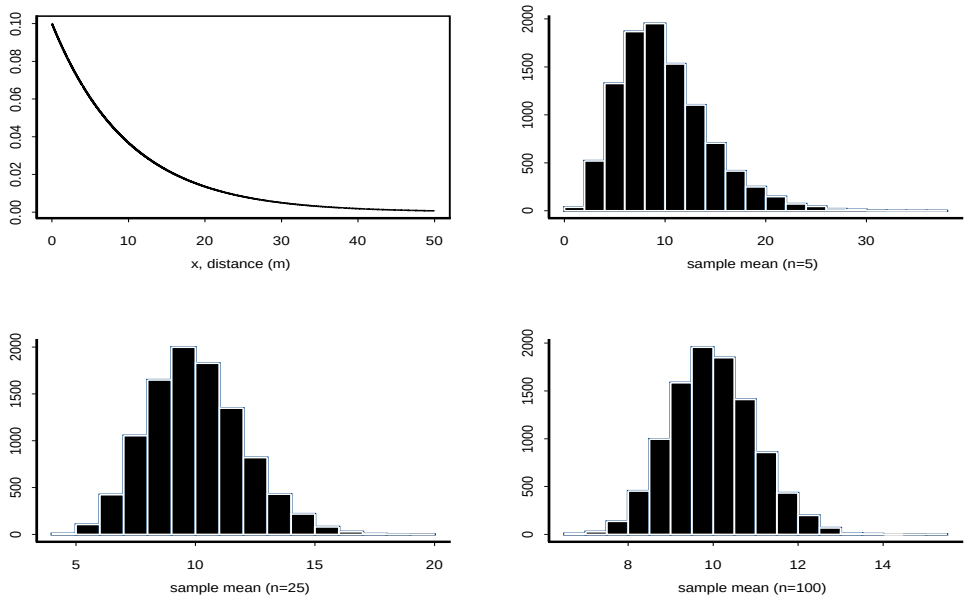



Figure 19.47: *Upper left: Population distribution of distance traveled to first territorial vacancy for banner-tailed kangaroo rats in Example* 19.1*; Upper right: Sampling distribution of* $\overline{x}$ *when* $n = 5$*; Lower left: Sampling distribution of* $\overline{x}$ *when* $n = 25$*; Lower right: Sampling distribution of* $\overline{x}$ *when* $n = 100$.

*INVESTIGATION*: Suppose that I take an SRS of $n$ banner-tailed kangaroo rats and record the distance traveled for each rat. How will the sample mean $\overline{x}$ behave in **repeated sampling** (i.e., what is the sampling distribution of $\overline{x}$)? To investigate this, we can use a **simulation study**.

- Generate 10,000 samples, each of size $n$, from this density curve with mean $\mu = 10$ and standard deviation $\sigma = 3.2$. These simulations are easy to do with a computer.
- Compute $\overline{x}$, the **sample mean**, for each sample.
- Plot all 10,000 sample means in a histogram. **This histogram approximates the true sampling distribution of** $\overline{x}$**.**

*OBSERVATIONS*: We first note that the density curve is heavily **skewed right**. However, from Figure 19.47, we note that

- the sampling distribution for $\overline{x}$, when $n = 5$, is still skewed right, but it already is taking a unimodal shape.
- the sampling distribution for $\overline{x}$, when $n = 25$, is almost symmetric! Sharp eyes might be able to detect a slight skew to the right.
- the sampling distribution for $\overline{x}$, when $n = 100$, is nearly perfectly normal in shape!

*What is going on here?*

*CENTRAL LIMIT THEOREM*: Draw an SRS of size $n$ from any density curve with mean $\mu$ and standard deviation $\sigma$. When $n$ is large, the sampling distribution of $\overline{x}$ is approximately normal with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$; that is,

$$\overline{x} \sim \mathcal{AN}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

The real novelty in this result is that the sample mean $\overline{x}$ will be approximately normal for large sample sizes, **even if the original density curve is not!**

**Example 19.2.** For one specific cultivar of potatoes, *Cherry Red*, an experiment was carried out using 40 plots of land. The plots were fairly identical in every way in terms of soil composition, amount of precipitation, etc. The density curve associated with the yields from last year's harvest was estimated to have mean $\mu = 158.2$ and standard deviation $\sigma = 14.9$ (bushels/plot). Suppose that this year's average yield (in the forty plots) was only $\overline{x} = 155.8$. Would you consider this to be necessarily unusual?

SOLUTION. We can answer this question by computing the **standardized value** of our observed sample mean $\overline{x} = 155.8$. Using last year's information, the standard deviation of the sampling distribution of $\overline{x}$ is

$$\frac{\sigma}{\sqrt{n}} = \frac{14.9}{\sqrt{40}} \approx 2.4.$$

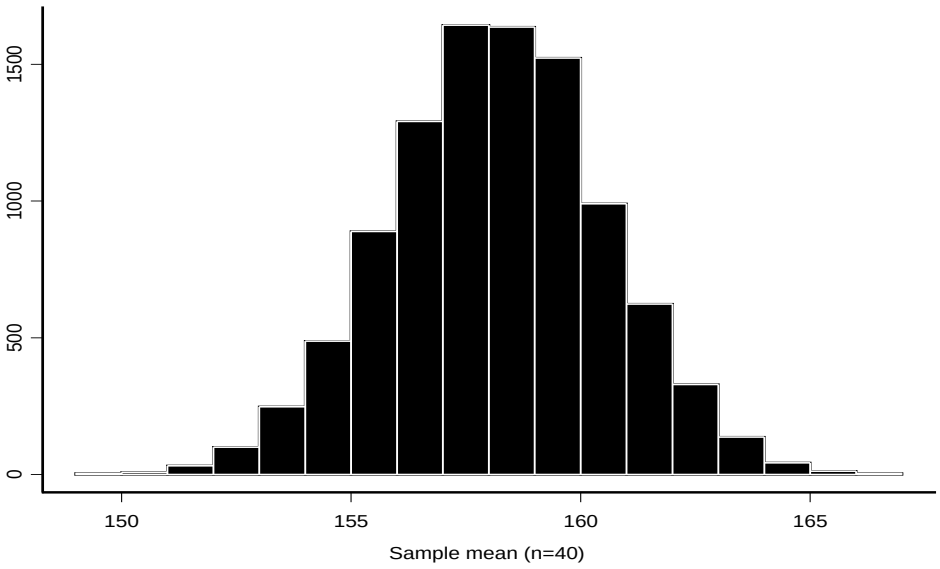



Figure 19.48: *Sampling distribution of the sample mean $\overline{x}$ when $n = 40$. This histogram represents the sampling distribution of $\overline{x}$ in Example* 19.2. *This sampling distribution was generated using a computer.*

From the Central Limit Theorem (CLT), we know that

$$\overline{x} \sim \mathcal{AN}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \text{ i.e., } \overline{x} \sim \mathcal{AN}\left(158.2, 2.4\right).$$

Thus, the **standardized value** of this year's sample mean $\overline{x} = 155.8$ is

$$\begin{aligned} z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} &= \frac{155.8 - 158.2}{2.4} \\ &= -1.0. \end{aligned}$$

This sample has produced a result which is just one standard deviation below last year's mean $\mu = 158.2$. *Is this year's results really all that unusual?* See Figure 19.48.

- What does it mean for a sample result to be **unusual**?
- How can we find a level $C$ **confidence interval** for the population mean $\mu$?

## 19.3 Confidence intervals for a population mean $\mu$

*MAIN RESULT*: Choose an SRS of size $n$ from a large population of individuals having mean $\mu$ and standard deviation $\sigma$.

- The **sample mean** is $\overline{x}$.
- The **sample standard deviation** is $s$.
- These are both statistics (estimates); they are computed from the sample data.

When $n$ is reasonably large, a **level $C$ confidence interval** for the population mean $\mu$ is

$$\overline{x} \pm \underbrace{z^* \left( \frac{s}{\sqrt{n}} \right)}_{\text{marg. err.}},$$

where $z^*$ is the critical value associated with the confidence level $C$. Popular critical values (and confidence levels) are below (these are unchanged from before). Recall how these critical values come from the standard normal distribution; see Figure 19.49.

| $C$ | 80 | 90 | 95 | 99 |
|---|---|---|---|---|
| $z^*$ | 1.28 | 1.64 | 1.96 | 2.58 |

*IMPORTANT NOTE*: Note that the particular form of this confidence interval is

**estimate $\pm$ margin of error**.

**Example 19.3.** An education major would like to estimate the mean GPA for undergraduate pre-nursing students at USC. An SRS of $n = 20$ students produced the following grade-point average data:

```
3.8  3.5  3.6  3.4  3.2  3.4  2.8  3.6  3.4  2.7
3.2  3.5  3.9  3.3  3.8  2.9  3.4  3.0  3.8  3.8
```

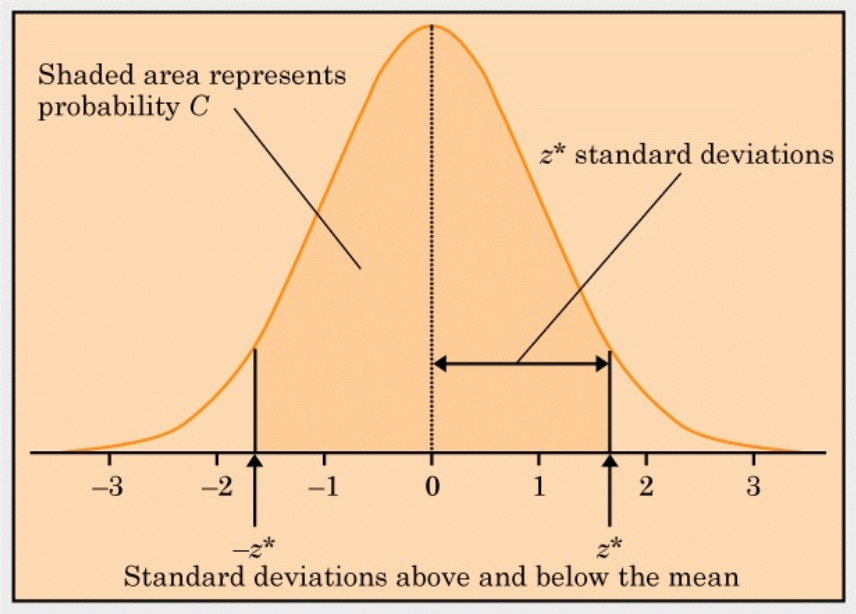



Figure 19.49: *Critical values $z^*$ from the standard normal distribution.*

*MINITAB*: I used Minitab (a statistics package) to calculate some summary statistics:

| Variable | N | Mean | StDev | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|---|
| GPA | 20 | 3.4 | 0.4 | 2.7 | 3.2 | 3.4 | 3.7 | 3.9 |

Here, we have that $n = 20$ (sample size), $\overline{x} = 3.4$ (sample mean), and $s = 0.4$ (sample standard deviation). A **95 percent** confidence interval for $\mu$, the population mean GPA is given by

$$\begin{aligned} \overline{x} \pm z^* \left(\frac{s}{\sqrt{n}}\right) &\implies 3.4 \pm 1.96 \left(\frac{0.4}{\sqrt{20}}\right) \\ &\implies 3.4 \pm 0.18 \\ &\implies (3.22, 3.58). \end{aligned}$$

Thus, we are 95 percent confident that the true mean GPA for USC undergraduate pre-nursing students is between 3.22 and 3.58.

EXERCISE: Using the data from Example 19.3, find an 80 percent confidence interval for the mean GPA $\mu$; also, find a 99 percent confidence interval. Interpret each interval!

SOLUTIONS:

- **80 percent**:

$$\begin{aligned} \overline{x} \pm z^* \left(\frac{s}{\sqrt{n}}\right) &\implies 3.4 \pm 1.28 \left(\frac{0.4}{\sqrt{20}}\right) \\ &\implies 3.4 \pm 0.11 \\ &\implies (3.29, 3.51). \end{aligned}$$

  We are 80 percent confident that the mean GPA for SC undergraduate pre-nursing students is between 3.29 and 3.51.

- **99 percent**:

$$\begin{aligned} \overline{x} \pm z^* \left(\frac{s}{\sqrt{n}}\right) &\implies 3.4 \pm 2.58 \left(\frac{0.4}{\sqrt{20}}\right) \\ &\implies 3.4 \pm 0.23 \\ &\implies (3.17, 3.63). \end{aligned}$$

  We are 99 percent confident that the mean GPA for SC undergraduate pre-nursing students is between 3.17 and 3.63.

*NOTE*: The shortest interval is the 80 percent interval. The largest is the 99 percent interval. **The higher the confidence level, the larger the interval will be.**

### 19.3.1 Interval length

*CONFIDENCE INTERVAL LENGTH*: The **length** of a level $C$ confidence interval is equal to twice the margin of error; i.e.,

$$\text{length} \approx 2z^* \left(\frac{\sigma}{\sqrt{n}}\right).$$

- As the confidence level increases, so does $z^* \implies$ interval length **increases**.
- For larger values of $\sigma \implies$ interval length **increases**.
- As the sample size $n$ increases $\implies$ interval length **decreases**.

### 19.3.2 Sample size determination

*CHOOSING SAMPLE SIZE*: Suppose that we would like to determine the sample size necessary to estimate $\mu$ with a level $C$ confidence interval. The **margin of error** associated with the confidence interval is approximately equal to

$$m = z^* \left( \frac{\sigma}{\sqrt{n}} \right).$$

This is an equation we can solve for $n$; in particular,

$$n = \left( \frac{z^* \sigma}{m} \right)^2.$$

*REALIZATION*: If we specify our confidence level $C$, an estimate of the population standard deviation $\sigma$ (a guess, most likely), and a desired margin of error $m$ that we are willing to "live with," we can find the sample size needed.

**Example 19.4.** In a biomedical experiment, we would like to estimate the mean remaining life of healthy rats ($\mu$, measured in days) that are given a high dose of a toxic substance. This may be done in an early phase clinical trial by researchers trying to find a maximum tolerable dose for humans. Suppose that we would like to write a 99 percent confidence interval for $\mu$ with a margin of error equal to $m = 2$ days. The researchers have provided a guess of $\sigma \approx 8$ days. How many rats should we use for the experiment?
SOLUTION: With a confidence level of $C = 99$ percent, our value of $z^*$ is

$$z^* = 2.58.$$

If the desired margin of error is $m = 2$ days, the sample size needed is

$$n = \left( \frac{2.58 \times 8}{2} \right)^2 \approx 106.5.$$

That is, we would need $n = 107$ rats to achieve these goals.

*CURIOSITY*: If collecting 107 rats is not feasible, we might think about weakening our requirements (after all, 99 percent confidence is very high, and the margin of error is tight). Suppose that we used a 90 percent confidence level instead with margin of error

$m = 5$ days. Then, the desired sample size would be

$$n = \left(\frac{1.64 \times 8}{5}\right)^2 \approx 6.9.$$

This is an easier experiment to carry out now (we need only 7 rats). However, we have paid a certain price: less confidence and less precision (in our confidence interval).

### 19.3.3 Warnings about confidence intervals

*CAUTIONS*: We need to be aware of some of the potential difficulties that arise when computing/presenting confidence intervals for a population mean:

- Data that we use to construct a confidence interval should be (or should be close to) a **simple random sample**. If the sampling design is biased, the results likely will be as well. Poor data collection techniques inundated with **nonsampling errors** will produce poor results too!
- Confidence interval formulas for $\mu$ are different if you use different probability sampling designs (e.g., **stratified samples**, etc.). These would be presented in an advanced course.
- **Outliers** almost always affect the analysis. You need to be careful in checking for them and deciding what to do about them.
- When the sample size $n$ is small, and when the population density curve is highly nonnormal (as in Example 19.1), the confidence interval formula we use for $\mu$ is probably a bad formula. This is true because the normal approximation to the sampling distribution for $\overline{x}$ is probably a bad approximation. The text offers an $n \geq 15$ guideline for most population distributions. That is, if your sample size is larger than or equal to 15, you're probably fine. **However, this is only a guideline.** If the underlying density curve is very skewed, this may or may not be a good guideline (you might need the sample size $n$ to be larger).
- Always plot your data to get an idea of skewness and outliers!