

NORTHLAND COLLEGE

MTH107 – STATISTICAL ANALYSIS AND INTERPRETATION

Introduction to Statistical Analysis and Interpretation

Instructors:

Dr. Derek H. Ogle
Jodi Supanich

Department:

Mathematical Sciences

December 20, 2016

Contents

| | | |
|--------------|-------------------------------|-----|
| I | BEGINNINGS | |
| 1 | Why Statistics is Important | 2 |
| 2 | Foundational Definitions | 8 |
| 3 | Data Production | 19 |
| 4 | Getting Started with R | 33 |
| II | EXPLORATORY DATA ANALYSIS | |
| 5 | Univariate EDA - Quantitative | 47 |
| 6 | Univariate EDA - Categorical | 75 |
| 7 | Normal Distribution | 83 |
| 8 | Bivariate EDA - Quantitative | 97 |
| 9 | Bivariate EDA - Categorical | 115 |
| 10 | Linear Regression | 127 |
| III | INFERENCE CONCEPTS | |
| 11 | Probability Introduction | 153 |
| 12 | Sampling Distributions | 156 |
| 13 | Hypothesis Tests | 179 |
| 14 | Confidence Regions | 192 |
| IV | SPECIFIC HYPOTHESIS TESTS | |
| 15 | 1-Sample Z-Test | 206 |
| 16 | 1-Sample t-Test | 213 |
| 17 | 2-Sample t-Test | 222 |
| 18 | Chi-Square Tests | 235 |
| 19 | Goodness-of-Fit | 251 |
| APPENDIX | | |
| BIBLIOGRAPHY | | |
| INDEX | | |

Part I

Beginnings

CHAPTER 1

WHY STATISTICS IS IMPORTANT

Chapter Objectives:

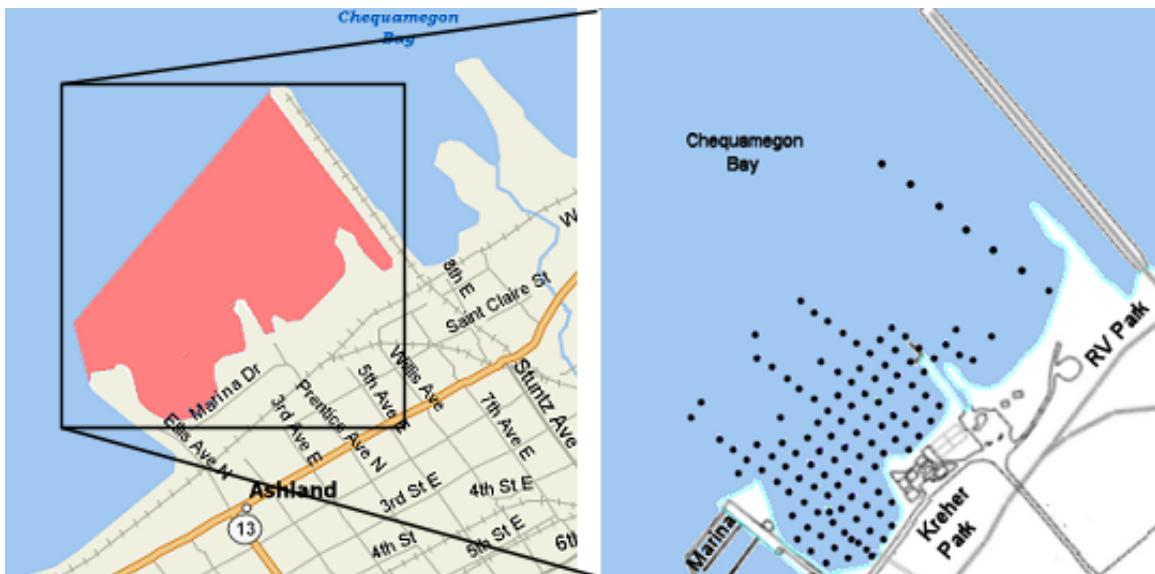
1. Describe the two major reasons why statistics is important for understanding populations.
2. Define natural and sampling variability.
3. Describe “difficulties” in making conclusions about population caused by sampling variability.
4. Define “statistics” (as a field of study).
5. Appreciate the importance of statistics in scientific inquiry.

Contents

| | | |
|-----|------------------------------------|---|
| 1.1 | Realities | 3 |
| 1.2 | Purpose of Statistics | 6 |
| 1.3 | Definition of Statistics | 6 |

1.1 Realities

THE CITY OF ASHLAND performed an investigation in the area of Kreher Park (Figure 1.1) when considering the possible expansion of an existing wastewater treatment facility in 1989. The discovery of contamination from what was believed to be creosote waste in the subsoils and ground water at Kreher Park prompted the city to abandon the project. A subsequent assessment by the Wisconsin Department of Natural Resources (WDNR) indicated elevated levels of hazardous substances in soil borings and ground water samples and in the sediments of Chequamegon Bay directly offshore of Kreher Park. In 1995 and 1999, the Northern States Power Company conducted investigations that further defined the area of contamination and confirmed the presence of specific contaminants associated with coal tar wastes. This site is now listed as a superfund site and is being given considerably more attention.¹



1.1 and that they were to present the results from these data at a public meeting. Further suppose that the WDNR scientist came to the meeting and showed a slide that simply had the list of lead concentration measurements on it (Table 1.1).² Is it easy to come to any conclusion about what these data mean from this type of presentation? Instead, suppose that the scientist came to the meeting with a simple plot of the frequency of concentrations present in the data and brief numerical summaries (Figure 1.2). With this presentation one can fairly easily see that the measurements were fairly symmetric with no obviously “weird” measurements and ranged from as low as $0.67 \mu\text{g} \cdot \text{m}^{-3}$ to as high as $1.36 \mu\text{g} \cdot \text{m}^{-3}$ with the measurements centered on approximately $1.0 \mu\text{g} \cdot \text{m}^{-3}$. These summaries will be discussed in detail in Chapter ???. However, at this point note that statistical methods are important for distilling or summarizing large quantities of data into graphs or numerical summaries from which it is much easier to identify characteristics of the data.

Table 1.1. Lead concentration measurements ($\mu\text{g} \cdot \text{m}^{-3}$) at each of 119 sites in Kreher Park superfund site.

```
[1] 0.91 1.03 0.87 1.24 1.05 0.88 1.07 1.11 1.09 0.95 1.23 1.06 0.91 0.67 1.17 0.99
[17] 1.00 1.14 1.12 1.09 1.14 1.12 1.01 0.70 1.09 0.99 0.98 0.78 0.93 1.06 1.20 0.98
[33] 1.06 0.99 0.79 0.94 0.94 0.99 1.17 1.11 0.98 0.96 1.10 1.08 0.90 0.89 1.05 1.12
[49] 0.98 1.13 1.06 0.91 1.05 0.83 1.21 1.30 0.94 0.84 1.09 0.98 1.36 0.99 1.10 1.00
[65] 0.89 1.03 0.73 1.22 1.02 1.33 1.07 0.89 1.09 0.86 0.81 1.04 0.93 1.00 1.01 0.91
[81] 0.91 0.98 1.18 0.77 1.09 1.05 1.16 0.95 1.06 1.04 0.92 1.18 1.17 1.11 1.24 1.08
[97] 0.81 0.91 0.82 0.93 0.91 1.01 0.86 1.02 0.90 1.27 1.11 1.14 1.06 1.25 0.90 0.93
[113] 1.21 0.90 0.97 0.94 0.95 0.96 1.07
```

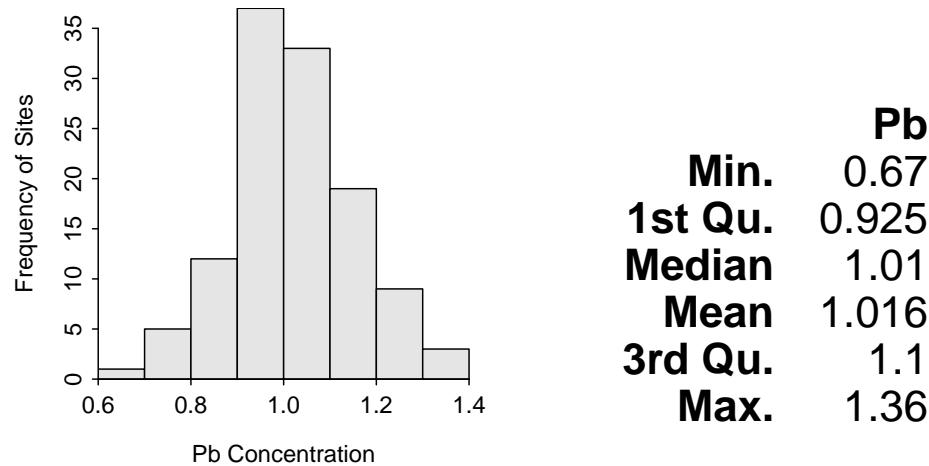


Figure 1.2. Histogram and summary statistics of lead concentration measurements ($\mu\text{g} \cdot \text{m}^{-3}$) at each of 119 sites in Kreher Park superfund site.

A critical question at this point is whether or not the results from the one sample of 119 sites perfectly represents what the results would be for the entire area. One way to consider this question is to examine the results obtained from another sample of 119 sites. The results from this second sample (Figure 1.3) are clearly, though not radically, different from the results of the first sample. Thus, it is seen that any one sample from a large area will not perfectly represent the area. Furthermore, it is observed that two different

²

These are hypothetical data for this site.

samples give two different results which will likely lead to two different, albeit generally only slightly different, conclusions.

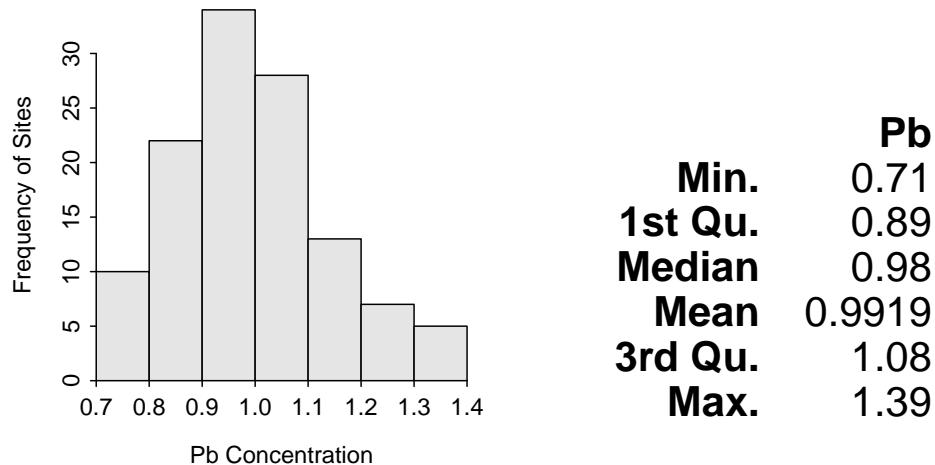


Figure 1.3. Histogram and summary statistics of lead concentration measurements ($\mu\text{g} \cdot \text{m}^{-3}$) at each of 119 sites (different from the sites shown in Figure 1.2) in Kreher Park superfund site.

The results of two different samples do not perfectly agree because each sample contains different individuals, and no two individuals (sites in this example) are exactly alike in every regard. The fact that no two individuals are exactly alike is **natural variability**, because of the “natural” differences that occur among individuals. The fact that the results from different samples are different is called **sampling variability**. If there was no natural variability, then there would be no sampling variability. If there was no sampling variability, then the field of statistics would not be needed because a sample (even of one individual) would perfectly represent the larger group of individuals. Thus, understanding variability is at the core of statistical practice. These concepts of natural and sampling variability will be revisited continuously throughout this course.

△ **Natural Variability:** The fact that no two individuals are exactly alike.

△ **Sampling Variability:** The fact that the results (i.e., statistics) from different samples are different.

All of this may be a bit unsettling! First, it was shown that an entire area or all of the individuals of interest cannot be examined. It was then shown that a sample of a subset of individuals did not perfectly represent all of the individuals. Furthermore, each sample is unique and will likely lead to a different conclusion. These are all real and difficult issues faced by the practicing scientist and considered by the informed consumer. However, the field of statistics is designed to “deal with” these issues such that the results from a relatively small subset of measurements can be used to make conclusions about the entire collection of measurements.

- ◊ Statistics provides methods for overcoming the difficulties caused by the requirement of sampling and the presence of sampling variability.

1.2 Purpose of Statistics

The field of statistics has two primary purposes, which were illustrated in the Kreher Park example. First, statistics provides methods to summarize large quantities of data into concise and informative numerical or graphical summaries. For example, it is easier to discern the general underlying structure of the lead measurements from the statistics and histograms presented in Figures 1.2 and 1.3 than it is from the full list of lead measurements in Table 1.1. Second, statistical methods allow inferences to be made about all individuals (i.e., a population) from a few individuals (i.e., a sample). Population and sample are defined more completely in Section ??.

- ◊ Statistics, as a field of study, is used to (1) summarize large quantities of data and (2) make inferences about populations from samples.

1.3 Definition of Statistics

With the Kreher Park example in mind, let's consider a definition of statistics. Statistics is the science of collecting, organizing, and interpreting numerical information or data (Moore and McCabe 1998). People (students and professionals) study statistics for a variety of reasons, including (Bluman 2002):

1. They must be able to read and understand the statistical studies performed in their field. To have this understanding they must be knowledgeable about the vocabulary, symbols, concepts, and statistical procedures used in these studies.
2. They may need to conduct research in their field. To accomplish this they must be able to design experiments and samples; collect, organize, analyze, and summarize data; and possibly make reliable predictions or forecasts for future use. They must also be able to communicate the results of the study.
3. They also need to be better consumers of statistical information.

The science of statistics permeates a wide variety of disciplines. Moore and McCabe (1998) state:

The study and collection of data are important in the work of many professions, so that training in the science of statistics is valuable preparation for a variety of careers. Each month, for example, government statistical offices release the latest numerical information on unemployment and inflation. Economists and financial advisers, as well as policy makers in government and business study these data in order to make informed decisions. Doctors must understand the origin and trustworthiness of the data that appear in medical journals if they are to offer their patients the most effective treatments. Politicians rely on data from polls of public opinion. Business decisions are based on market research data that reveal customer tastes. Farmers study data from field trials of new crop varieties. Engineers gather data on the quality and reliability of manufactured products. Most areas of academic study make use of numbers, and therefore also make use of the methods of statistics.

Δ **Statistics:** The science of collecting, organizing, and interpreting numerical information or data.

Review Exercises

- 1.1** There are 1499 lakes in Ashland, Bayfield, and Douglas counties of Wisconsin. However, only 605 of these are named. A random sample of named lakes from this population is extracted with the following R code:

```
> library(NCStats)
> data(ABDLakes)
> named <- filterD(ABDLakes,named)
> srsdf(named,n=50,vars=c("county","area"))
```

Use this code and some hand (or calculator) calculations to answer the questions below. [Answer](#)

- (a) Extract a sample of 50 lakes with the code above. Compare the size (area in acres) of the first two lakes. What do you observe? This is an example of what type of variability?
 - (b) Compute the proportion of lakes in your sample that are from Bayfield county.
 - (c) Extract another sample of 50 lakes. Compare the proportion of lakes that are from Bayfield county in this sample to the proportion from your first sample. What do you observe? This is an example of what type of variability?
 - (d) Of the named lakes in the three counties, 346 are from Bayfield county. Was the proportion of lakes from Bayfield County in either of your samples equal to the proportion of all named lakes that were from Bayfield County? Were you surprised? Why or why not?
-

CHAPTER 2

FOUNDATIONAL DEFINITIONS

Chapter Objectives:

1. Describe what an individual is.
2. Describe what a population and a sample are and how they differ.
3. Describe what a parameter and a statistic are and how they differ.
4. Describe how a population, parameter, sample, and statistic are related.
5. Identify the individual, variable(s), population, parameter(s), sample, and statistic(s) (IVPPSS) in a given situation.
6. Identify variable types in context.

Contents

| | | |
|-----|--------------------------------|----|
| 2.1 | IVPPSS Definitions | 9 |
| 2.2 | Performing an IVPPSS | 11 |
| 2.3 | Variable Types | 14 |

STATISTICAL INFERENCE IS THE PROCESS of forming conclusions about the unknown parameters of a population by computing statistics from the individuals in a sample. As you can imagine from this definition, it is important that you understand the difference between a population and a sample and a parameter and a statistic before you can understand and appreciate the process of making statistical inferences. Before identifying these items you must also identify the individual and variable(s) of interest. These six items must be explicitly identified at the beginning of any statistical analysis for that analysis to be conducted properly. Understanding and identifying these items is the focus of this section. Formal methods of inference are discussed beginning with Chapter 11.

Δ **Inference:** The process of forming conclusions about the unknown parameters of a population by computing statistics from the individuals in a sample.

Throughout this chapter we will identify the individual, variable, population, parameter, sample, and statistic (referred to as the **IVPPSS**) for the following hypothetical example. Assume that interest is in determining the mean (or average) length of the 1015 fish in Square Lake (Figure 2.1). Note that in “real life” we would not know how many fish are in this lake. However, for the purpose of illustrating the concepts of this chapter we will suppose that specific values of several variables for all 1015 fish in this lake are known.

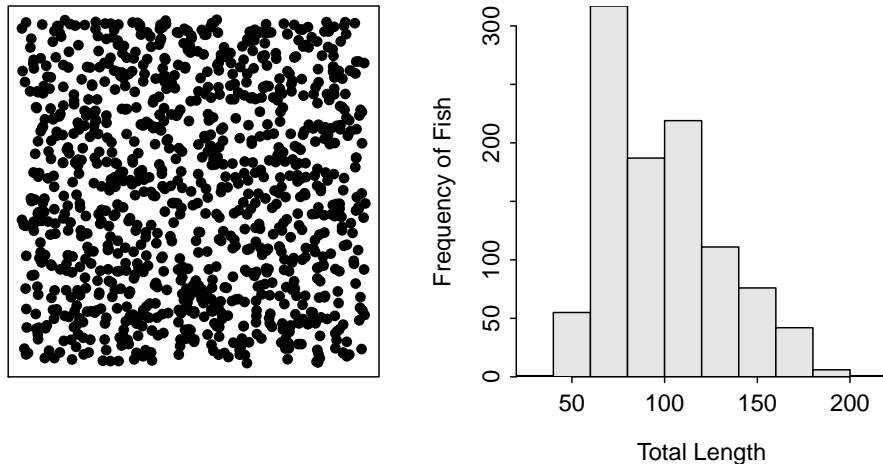


Figure 2.1. Schematic representation of individual fish (i.e., dots; **Left**) and histogram (**Right**) of the total length of the 1015 fish in Square Lake.

2.1 IVPPSS Definitions

The **individual** in a statistical analysis is one of the “items” to be examined by the researcher. Sometimes the individual is a person, but it may be an animal, a piece of wood, a site or location, a particular time, or an event. It is extremely important that you don’t always visualize a person when you use the word individual in a statistical context. Synonyms for individual are unit, experimental unit (usually used in experiments), sampling unit (usually used in observational studies), case, and subject (usually used in studies involving humans). The individual of interest in the Square Lake example is an individual fish, because the researcher will collect a set of fish and examine each fish individually.

Δ Individual: One of the items examined by the researcher.

◊ An individual is not necessarily a person.

The **variable** is the characteristic of interest about each individual. The variable is the information that the researcher records about each individual. The variable of interest in the Square Lake example is the length of each fish. Note that in most “real life” studies the researcher will be interested in more than one variable. For example, in this example, the researcher may also record the fish’s weight, sex, and age. Studies with one variable are called univariate studies, studies with two variables are bivariate studies, and studies with more than two variables are called multivariate studies.

Δ Variable: The characteristic of interest about each individual.

A **population** is the collection of ALL individuals of interest. Simply put, the population is all of the individuals. In the Square Lake example, the population is all 1015 fish in the lake. You should define the population as thoroughly as possible including all qualifiers as necessary. This example is simple because Square Lake is so well defined; however, as you will see in the chapter review exercises, the population is often only well-defined by your choice of descriptors.

Δ Population: The collection of ALL individuals of interest.

A **parameter** is a summary computed from ALL individuals in a population. The term for the particular summary is usually preceded by the word “population.” Parameters are ultimately what is of interest because interest is in all individuals in the population. However, in practice, parameters cannot be computed because the entire population cannot be “seen.” In this hypothetical example, all 1015 fish are accessible and the parameters are computed (Table 2.1).¹ As stated above, in this example, interest is in the population mean (or average) length of all fish in Square Lake, which is 98.06 mm.

Table 2.1. Summary parameters for the total length of all 1015 fish in the Square Lake population.

| n | mean | sd | min | Q1 | median | Q3 | max |
|------|-------|-------|-----|----|--------|-----|-----|
| 1015 | 98.06 | 31.49 | 39 | 72 | 93 | 117 | 203 |

Δ Parameter: A summary of all individuals in a population.

◊ Populations and parameters can generally not be “seen.”

The entire population cannot be “seen” in real life. Thus, the only alternative for learning something about the population is to examine a subset of the population. This subset is called a **sample**. The red dots in Figure 2.2 represent a random sample of 50 fish from Square Lake (note that the sample size is usually denoted by a lower-case n; i.e., n=50).

Δ Sample: A subset of the population examined by a researcher.

¹

We will discuss how to compute and interpret each of these values in later chapters.

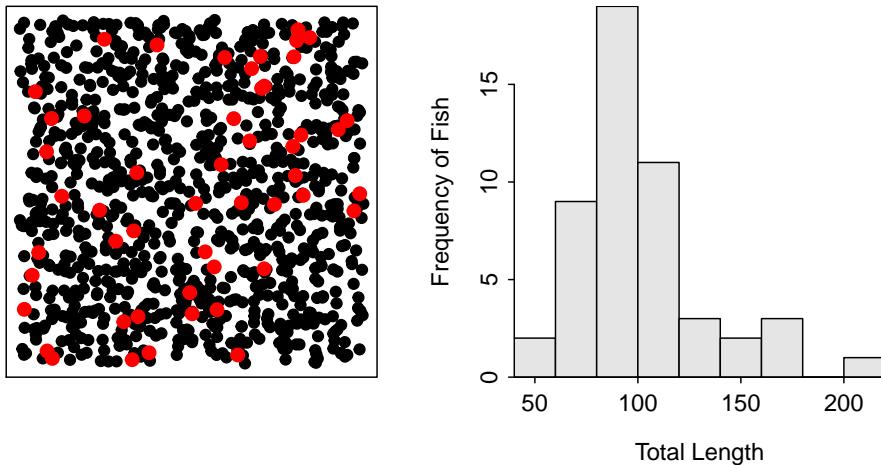


Figure 2.2. Schematic representation (**Left**) of a sample of 50 fish (i.e., red dots) from Square Lake and histogram (**Right**) of the total length of the 50 fish in this sample.

Summaries computed from individuals in a sample are called **statistics**. Specific names of statistics should be preceded by the word “sample.” The statistic of interest is always the same as the parameter of interest; i.e., the statistic describes the sample the same way that the parameter describes the population. For example, if interest is in the population mean, then the sample mean would be computed. Some statistics computed from this sample are shown in Table 2.2 and Figure 2.2. The sample mean of 100.04 mm is the best “guess” at the population mean. Not surprisingly from the discussion in Section ??, the sample mean does not perfectly equal the population mean.

Table 2.2. Summary statistics for the total length of a sample of 50 fish from the Square Lake population.

| n | mean | sd | min | Q1 | median | Q3 | max |
|----|--------|-------|-----|----|--------|-----|-----|
| 50 | 100.04 | 31.94 | 49 | 81 | 91 | 118 | 203 |

2.2 Performing an IVPPSS

In each statistical analysis it is important that you determine the Individual, Variable, Population, Parameter, Sample, and Statistics (IVPPSS). It is simplest to do this in this order – IVPPSS. First, determine what items you are actually going to look at; those are your individuals. Second, what are you going to record when you look at an individual; that is the variable. Third, the population is simply ALL of the individuals. Fourth, the parameter is a summary (e.g., mean or proportion) of the variable recorded from all of the individuals in the population.² Fifth, we realize that we cannot see all of the individuals in the population so we examined only a few; those few were the sample. Finally, the summary of the individuals in the sample is the statistic. The statistic has to be the same summary of the sample as the parameter was of the population.

When performing an IVPPSS, keep in mind that parameters describe populations (note that they both start with “p”) and statistics describe samples (note that they both start with “s”). This can also be looked

²

Again, parameters generally cannot be computed because all of the individuals in the population can generally not be seen. Thus, the parameter is largely conceptual.

at from another perspective. A sample is an estimate of the population and a statistic is an estimate of a parameter.

This process can be illustrated by performing an IVPPSS for the following situation:

*A University of New Hampshire graduate student (and Northland College alum) investigated habitat utilization by New England (*Sylvilagus transitionalis*) and Eastern (*Sylvilagus floridanus*) cottontail rabbits in Maine. In a preliminary portion of his research he determined the proportion of “rabbit patches” that were inhabited by New England cottontails. He examined 70 “patches” and found that 53 showed evidence of inhabitance by New England cottontails.*

- The individuals are the rabbit patches (i.e., a rabbit patch is the “item” being sampled and examined).
- The variable is “evidence for New England cottontails or not (yes or no)” (i.e., the characteristic of each rabbit patch that was recorded).
- The population is ALL rabbit patches in Maine.
- The parameter is the proportion of ALL rabbit patches in Maine that showed evidence for New England cottontails.³
- The sample is the 70 rabbit patches in Maine that were actually examined by the researcher.
- The statistic is the proportion of the 70 rabbit patches in Maine actually examined that showed evidence for New England cottontails.⁴

In some situations it may be easier to start by identifying the sample. From this, and through the realization that a sample is always “of the individuals”, it may be easier to identify the individual. This process is illustrated in the following example, with the items listed in the order identified rather than in the traditional IVPPSS order.

The Duluth, MN tourism board is interested in the average number of raptors seen per year at Hawk Ridge.⁵ To determine this value, they collected the total number of raptors seen in a sample of years from 1971-2003.

- The sample is the 32 years between 1971-2003 at Hawk Ridge.
- An individual is a year (because a “sample of years” was taken) at Hawk Ridge.
- The variable recorded was the number of raptors seen in one year.
- The population is ALL years (this is a bit ambiguous but may be thought of as all years that Hawk Ridge has existed).
- The parameter is the average number of raptors seen per year in ALL years.
- The statistic is the average number of raptors seen in the 1971-2003 sample of years.

Review Exercises

³

Note that this population and parameter cannot actually be calculated but it is what the researcher wants to know.

⁴

Note that this statistic is the same as the parameter; it is just computed on a different collection of individuals.

⁵

Information about Hawk Ridge is found [here](#).

- 2.1** My Dad owns 60 acres of timber (mostly Oak, Walnut, and Poplar) in Iowa. He wants to measure the mean diameter-at-breast-height (DBH) of the oak trees on his property. He measures the DBH of 75 randomly selected oak trees. Use this information to perform an IVPSS. [Answer](#)
- 2.2** I have a friend who wants to start a (fishing) bait store on the West end of Ashland. He wants to determine what proportion of Ashland residents who currently use the East end bait store would use a store in the West end if one existed. He sends out 5000 questionnaires and receives 2378 back from patrons of the East end store. Use this information to perform an IVPSS. [Answer](#)
- 2.3** I'm interested in developing a model to predict how many points an NBA starting basketball player scores. Therefore, I want to determine the relationship between points scored and height, speed (in the 40-yard dash), position, and minutes played. To identify this relationship I gather these data from 100 NBA starting basketball players. Use this information to perform an IVPSS. [Answer](#)
- 2.4** Pollsters wanted to determine the proportion of registered voters who approved of President Clinton's performance. They called 5000 randomly selected registered voters and ask 4123 of those (the rest weren't home, didn't answer, or hung up) "Do you approve of Pres. Clinton's performance?" Use this information to perform an IVPSS. [Answer](#)
- 2.5** You Might Be Interested To Know (YMBITK), the average level of mercury in newly-hatched goslings in the upper Midwest (MI, MN, ND, SD, WI). You obtained 20 goslings from resource agencies in each state. Use this information to perform an IVPSS. [Answer](#)
- 2.6** YMBITK, the proportion of NC students that think NC can become "the nation's leading environmental liberal arts college" in the next decade. You polled 124 students. Use this information to perform an IVPSS. [Answer](#)
- 2.7** YMBITK, the relationship between hours studied and GPA of students in the UW system (excluding UW-Madison). You interviewed 250 students from throughout the system. Use this information to perform an IVPSS. [Answer](#)
- 2.8** YMBITK, the average difference in salaries between the head coaches of men's and head coaches of women's basketball teams at Division I schools. You interviewed 73 head-coach pairs. Use this information to perform an IVPSS. [Answer](#)
- 2.9** YMBITK, the proportion of graduates from small private schools, who majored in Biology and who have been out of school for at least 5 years, that feel that statistics is an "important" course. You interviewed 1023 people. Use this information to perform an IVPSS. [Answer](#)
- 2.10** Scientist in Chivyrkui Bay on Lake Baikal (Owens and Pronin 2000) were interested, among other things, in determining the mean age of pike (*Esox lucius*) in the bay. They collected scales from 30 fish using gill nets and angling methods. Use this information to perform an IVPSS. [Answer](#)
- 2.11** The Eurasian ruffe is an exotic species of fish that is causing some alarm in fisheries biologists in the Great Lakes area (Maniak *et al.* 2000). A few of these biologists tested to see if a certain pheromone released by injured ruffe would repel other ruffe. If so, natural, or possibly synthetic, versions of this pheromone could be used to distract ruffe from areas in which they are causing damage. In their experiment, they observed

ruffe held in aquaria divided into four sections. They recorded what proportion of 24 randomly-selected ruffe caught in the St. Louis River Harbor, and then held in the aquaria, left a section when the chemical was added to that section. Use this information to perform an IVPPSS. [Answer](#)

2.2.1 Sampling Variability, Again

It is instructive to once again consider how the statistics might change when different samples are taken. Table 2.3 and Figure 2.3 show the results from three more samples of 50 fish from the Square Lake population. All four samples (including the sample shown in Table 2.2 and Figure 2.2) had means that were quite different from the known population mean of 98.06. All four histograms were similar in appearance but were slightly different in actual values. Similar conclusions could be made for the other parameters and statistics (e.g., compare the population and sample standard deviations, min, max, etc.). The concept illustrated here is that a sample will likely represent the population of interest, but there will be variability among samples. This variability between samples is called **sampling variability**, which is one of the most important concepts in statistics and will be discussed in great detail throughout this course.

Table 2.3. Summary statistics for the total length in three samples of 50 fish from the Square Lake population.

| n | mean | sd | min | Q1 | median | Q3 | max |
|----|--------|-------|-----|----|--------|-----|-----|
| 50 | 99.56 | 32.47 | 57 | 69 | 91 | 123 | 167 |
| 50 | 88.64 | 24.52 | 53 | 68 | 86 | 106 | 166 |
| 50 | 112.74 | 35.86 | 61 | 84 | 108 | 147 | 174 |

Δ **Sampling Variability:** The realization that no two samples are exactly alike. Thus, statistics computed from different samples will likely vary.

This example also illustrates another important statistical concept. Parameters are fixed values because populations don't change. If a population does change, then it is considered to be a different population. In the Square Lake example, if a fish is removed from the lake, then the lake would then be considered a different population of fish. Statistics, on the other hand, vary in value depending on the sample because each sample consists of different individuals and individuals vary (i.e., sampling variability exists because natural variability exists).

◊ Parameters are fixed in value, while statistics vary in value.

2.3 Variable Types

The type of statistic that can be calculated is dictated by the type of variable to be analyzed. For example, a sample mean (or average) can only be calculated for a quantitative variable (defined below). Thus, immediately after identifying the variable in the IVPPSS procedure you should identify the type of that variable.

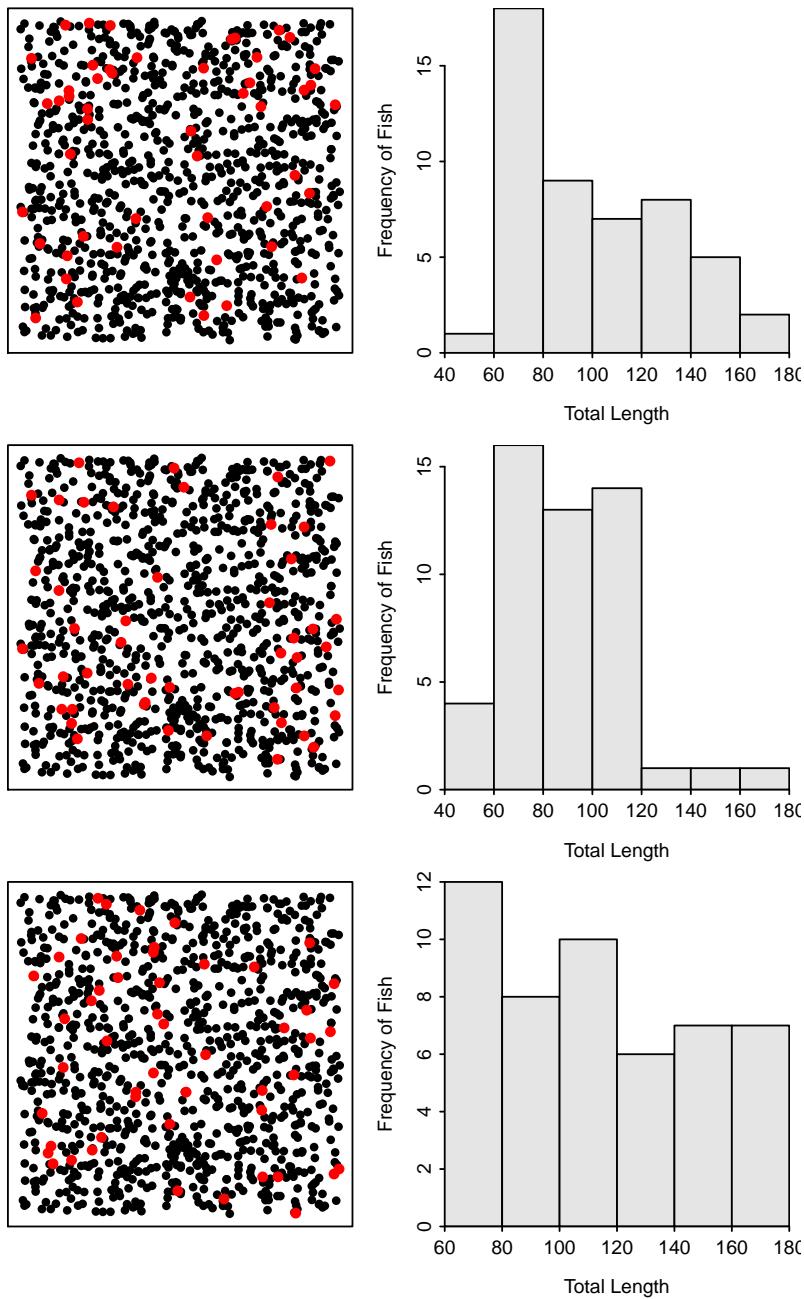


Figure 2.3. Schematic representation (**Left**) of three samples of 50 fish (i.e., red dots) from Square Lake and histograms (**Right**) of the total length of the 50 fish in each sample.

2.3.1 Variable Definitions

There are two main groups of variable types – quantitative and categorical (Figure 2.4). **Quantitative** variables are variables with numerical values for which it makes sense to do arithmetic operations (like adding or averaging). Synonyms for quantitative are measurement or numerical. **Categorical** variables

are variables that record to which group or category an individual belongs. Synonyms for categorical are qualitative or attribute. Within each main type of variable are two subgroups (Figure 2.4).

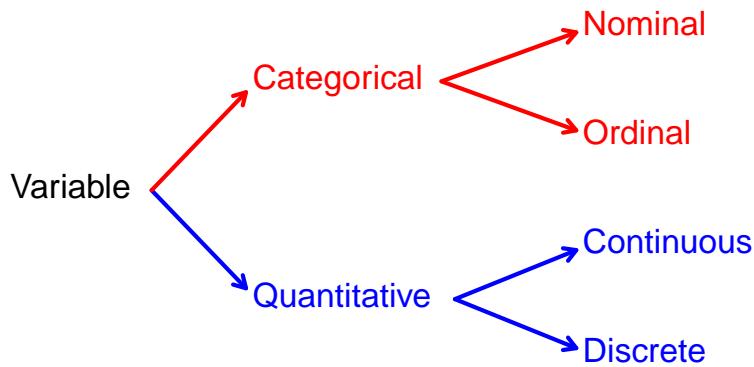


Figure 2.4. Schematic representation of the four types of variables.

The two types of quantitative variables are continuous and discrete variables. **Continuous** variables are quantitative variables that have an uncountable number of values. In other words, a potential value DOES exist between every pair of values of a continuous variable. **Discrete** variables are quantitative variables that have a countable number of values. Stated differently, a potential value DOES NOT exist between every pair of values of a discrete variable. Typically, but not always, discrete variables are counts of items.

Continuous and discrete variables are easily distinguished by determining if it is possible for a value to exist between every two values of the variable. For example, can there be between 2 and 3 ducks on a pond? No! Thus, the number of ducks is a discrete variable. Alternatively, can a duck weigh between 2 and 3 kg? Yes! Can it weigh between 2 and 2.1 kg? Yes! Can it weigh between 2 and 2.01 kg? Yes! You can see that this line of questions could continue forever; thus, duck weight is a continuous variable.

△ Discrete Variable: A quantitative variable that can assume a countable number of values.

△ Continuous Variable: A quantitative variable that can assume an uncountable number of values.

◊ A quantitative variable is continuous if a possible value exists between every two values of the variable; otherwise, it is discrete.

The two types of categorical variables are ordinal and nominal. **Ordinal** variables are categorical variables where a natural order or ranking exists among the categories. **Nominal** variables are categorical variables where no order or ranking exists among the categories.

Ordinal and nominal variables are easily distinguished by determining if the order of the categories matters. For example, suppose that a researcher recorded a subjective measure of condition (i.e., poor, average, excellent) and the species of each duck. Order matters with the condition variable – i.e., condition improves from the first (poor) to the last category (excellent) – and some reorderings of the categories would not

make sense – i.e., average, poor, excellent does not make sense. Thus, condition is an ordinal variable. In contrast, species (e.g., mallard, redhead, canvasback, and wood duck) is a nominal variable because there is no inherent order among the categories (i.e., any reordering of the categories also “makes sense”).

△ **Ordinal Variable:** A categorical variable for which a natural order exists among the categories.

△ **Nominal Variable:** A categorical variable for which a natural order DOES NOT exist among the categories.

◊ Remember that ordinal means that an order among the categories exists (note “ord” in both ordinal and order).

The following are some issues to consider when identifying the type of a variable:

1. The categories of a categorical variable are sometimes labeled with numbers. Don’t let this fool you into calling the variable quantitative.
2. Rankings, ratings, and preferences are ordinal (categorical) variables.
3. Counts of numbers are discrete (quantitative) variables.
4. Measurements are typically continuous (quantitative) variables.
5. Categorical variables that consist of only two levels or categories will be labeled as a nominal variable (because any order of the groups makes sense). This type of variable is also often called “binomial.”.
6. Do not confuse “what type of variable” (answer is one of “continuous”, “discrete”, “nominal”, or “ordinal”) with “what type of variability” (answer is “natural” or “sampling”) questions.

◊ “What type of variable is ...?” is a different question than “what type of variability is ...?” Be careful to note the word difference (i.e., “variable” versus “variability” when answering these questions.

Review Exercises

2.12 What type of variable is the number of ducks found at the “Hot Pond” every morning? [Answer](#)

2.13 What type of variable is the genotype (AA, Aa, aa) of a particular species of sunflower? [Answer](#)

2.14 What type of variable is the length of petals on individual flowers? [Answer](#)

2.15 What type of variable is the number of seeds produced by an individual sunflower? [Answer](#)

2.16 What type of variable is the “quality” of the seeds produced by an individual plant (“quality” is recorded as 1=poor, 2=low, 3=good, and 4=excellent)? [Answer](#)

- 2.17** What type of variable is student rankings (“Excellent”, “Very Good”, “Good”, “Fair”, “Poor”) of a professor’s abilities? [Answer](#)
- 2.18** What type of variable is whether an account is valid or invalid? [Answer](#)
- 2.19** What type of variable is the number of defects produced by a machine? [Answer](#)
- 2.20** What type of variable is the ounces of cola in a sample of 100 bottles? [Answer](#)
- 2.21** What type of variable is the sex of fish collected from a lake? [Answer](#)
- 2.22** What type of variable is the number of legs on frogs collected in Bayfield County? [Answer](#)
- 2.23** What type of variable is the frequency (mhz) of a bullfrog’s “croak”? [Answer](#)
- 2.24** What type of variable is the number of incorporated towns in a county? [Answer](#)
- 2.25** What type of variable is the qualitative size of least weasels (small, medium, large)? [Answer](#)
-

CHAPTER 3

DATA PRODUCTION

Chapter Objectives:

1. Identify major differences between data produced from experiments and observational studies.
2. Understand basic ideas of simple random experiments with one and two factors.
3. Describe the principles of experimental design.
4. Describe the principles of observational studies.
5. Understand basic ideas of designing simple observational studies, and
6. Explain the importance of randomization in both experiments and observational studies.

Contents

| | |
|--|----|
| 3.1 Experiments | 20 |
| 3.2 Observational Studies – Sampling | 28 |

STATISTICAL INFERENCE IS THE PROCESS of making conclusions about an entire population based on the results from the individuals in a single sample. To make conclusions about the larger population from a sample requires a sample that fairly represents the larger population. In this chapter two ways of producing data – (1) Experiments and (2) Observational Studies – are described. The information in this chapter is critical to the field of statistics and science in general, because if data are not properly collected, then inferences and proper conclusions cannot be made.

△ **Inference:** The process of forming conclusions about the unknown parameters of a population by computing statistics from the individuals in a sample.

- ◊ If data are not properly collected, then inferences cannot be made.

3.1 Experiments

An experiment deliberately imposes a condition, or treatment, on individuals in order to observe their response. In a properly designed experiment all variables that we are not interested in are held constant while the variable(s) that we are interested in are changed among treatments. As long as the experiment is designed properly (see below), tests for differences in the response variable among treatment groups can be made. If differences among treatment groups occur, then those differences are due either to the variable(s) that were deliberately changed or randomness (chance). Thus, strong cause-and-effect statements can be made from data collected with a carefully designed experiment.

- ◊ An experiment deliberately imposes a condition, or treatment, on individuals in order to observe their response.
- ◊ Strong cause-and-effect statements can be made from data collected with a carefully designed experiment.

3.1.1 Single-factor Experiments

A factor is the variable that is deliberately manipulated to determine its effect on the response variable. Sometimes the factor is called an explanatory variable because we are attempting to determine how it affects (or “explains”) the response variable. The simplest experiment is a single-factor experiment where the individuals are split into groups defined by the categories of a single factor variable.

For example, suppose that a research group wants to examine the effect of temperature on the growth rate of a strain of bacteria (i.e., total number of bacterial cells at the end of the experiment). The researchers will examine growth at only two temperatures in this simple experiment – 10°C and 15°C. In addition, they have prepared 120 agars (petri dishes with a material on which the bacteria can grow) that were inoculated with the bacteria and placed in a chamber where the temperature and all other environmental conditions (e.g., humidity, light) were controlled exactly. In this very simple experiment, temperature is the only factor as it is the only variable that is manipulated to different values to determine its impact on the growth rate response variable.

Δ Factor(s): The variable(s) that is (are) deliberately manipulated in the experiment to determine its effect on the response variable. Sometimes called the explanatory variable.

Δ Response: The variable observed in an experiment to identify the effect of the factors on it.

- ◊ In a single-factor experiment only one explanatory variable (i.e., factor) is allowed to vary; all other explanatory variables are held constant.

Levels are the number of categories of the factor variable. In this example, there are two levels – 10°C and 15°C. Treatments are the number of unique conditions that individuals in the experiment are exposed to. In single-factor experiments, the number of treatments is the same as the number of levels of the single factor. Thus, in this simple experiment, there are two treatments – 10°C and 15°C. Treatments will be discussed more thoroughly in the next section.

The number of replicates in an experiment is the number of individuals that will receive each treatment. In this example, the replicates are the number of inoculated agars that will receive each of the two temperature treatments. The number of replicates is determined by dividing the total number of available individuals (120) by the number of treatments (2). Thus, in this case, the number of replicates is 60 inoculated agars.

Δ Levels: The number of categories or groupings of the factor.

- ◊ In single-factor experiments, the number of treatments in the experiment equals the number of levels of the single factor.

Δ Replicates: The number of individuals in each treatment group.

- ◊ The number of replicates is determined by dividing the total number of available individuals by the number of treatments.

The agars used in this experiment will be randomly allocated to the two temperature treatments. All other variables – humidity, light, etc. – are kept the same for each group. At the end of two weeks, the total number of bacterial cells on each agar (i.e., the response variable) will be recorded and compared between the agars kept at both temperatures.¹ Any difference in average growth rates was either due to the different temperature treatments or randomness, because all other variables were the same between the two treatment groups.

- ◊ Differences among treatment groups are either caused by randomness (chance) or the factor.

We are not restricted to just two categories for our single factor. For example, we may be interested in more than two temperatures, say 10°C, 12.5°C, 15°C, and 17.5°C. There is still only one factor with this modification – temperature – but there are now four levels. Because this is still a single-factor experiment

¹

Methods for making this comparison are shown in Chapter ??.

there are still only as many treatment groups as levels of the factor (four in this case).

3.1.2 Multi-factor Experiments

More than one factor can be tested in one experiment. In fact, it can be shown that it is more efficient to have a properly designed but slightly more complex experiment where more than one factor is varied at a time than it is to use separate experiments in which only one factor is varied in each. However, before showing this benefit let's examine the definitions from the previous section in a multi-factor experiment.

Suppose that the previous experiment was modified to examine, in addition to the two temperatures from above, the effect of different relative humidity levels on the growth of this strain of bacteria. This modified experiment now has two factors – temperature (with two levels of 10°C or 15°C) and relative humidity (with four levels of 20%, 40%, 60%, and 80%). The number of combinations of all factors in the experiment is the number of treatments and is found by multiplying the levels of all factors (i.e., $2 \times 4 = 8$ in this case). The number of replicates in this experiment is now 15 (i.e., the total number of available agars divided by eight; $120/8$).

A quick drawing of the experimental design can be instructive (below). The drawing is a grid where the levels of one factor make up the rows and the levels of the other factor make up the columns of the grid. The number of rows and columns correspond to the levels of the two factors, respectively, whereas the number of cells in the grid is the number of treatments (numbered in this table to show eight treatments).

| | | Relative Humidity | | | | |
|--|--|-------------------|-----|-----|-----|---|
| | | 20% | 40% | 60% | 80% | |
| | | 10°C | 1 | 2 | 3 | 4 |
| | | 15°C | 5 | 6 | 7 | 8 |

△ **Treatments:** The number of combinations of all factors in the experiment.

◊ The number of treatments equals the product of the levels for each factor.

◊ The number of treatments is determined for the overall experiment, whereas the number of levels is determined for each factor.

The statistical analysis of a multi-factor experimental design such as this is more involved than what will be shown in this book. However, multi-factor experiments have many benefits. The benefits of a multi-factor experiment can be illustrated by comparing the design of a multi-factor experiment to separate single-factor experiments. Let's continue with the experiment to identify the effect of both temperature and relative humidity on the growth rate of a certain strain of bacteria. However, consider for the moment that (1) separate single-factor experiments will also be conducted to determine the effect of each factor and (2) we cannot use any of the individuals (i.e., agars) in more than one experiment.

To conduct the two separate experiments, randomly split the 120 available agars into two equally-sized groups of 60. The first 60 will be used in the first experiment concerning the effect of temperature on growth rate. In this experiment, the 60 individuals will be split into two groups of 30 individuals (for the two levels of temperature). The second 60 individuals will be used in the second experiment concerning relative humidity. In this experiment, the 60 individuals will be split into four groups of 15 individuals (for the four

levels of relative humidity). These separate single-factor experiments are summarized in the following tables (where the numbers in the cells represent the number of replicates in each treatment).

| Temperature | | Relative Humidity | | | |
|-------------|------|-------------------|-----|-----|-----|
| 10°C | 15°C | 20% | 40% | 60% | 80% |
| 30 | 30 | 15 | 15 | 15 | 15 |

Now reconsider the design where both factors were varied at once (the table below was modified to include the number of replicates in each treatment).

| | | Relative Humidity | | | |
|------|----|-------------------|-----|-----|-----|
| | | 20% | 40% | 60% | 80% |
| 10°C | | 15 | 15 | 15 | 15 |
| 15°C | 15 | 15 | 15 | 15 | 15 |

The key to examining the benefits of the multi-factor experiment is to determine the number of individuals that give “information” about (i.e., are exposed to) each factor. From the last table it is seen that all 120 individuals are exposed to one of the temperature levels with 60 individuals exposed to each level. In contrast, only 30 individuals were exposed to these levels in the single-factor experiment. In addition, all 120 individuals are exposed to one of the relative humidity levels with 30 individuals exposed to each level. Again, this is in contrast to the single-factor experiment where only 15 individuals were exposed to these levels. Thus, the first advantage of multi-factor experiments is that the available individuals are used more efficiently. In other words, more “information” (i.e., the responses of more individuals) is obtained from a multi-factor experiment than from combinations of single-factor experiments.²

- ◊ Multi-factor experiments use available individuals more efficiently; i.e., more “information” about the effect of the factors on the response is gained from the same number of individuals.

A properly designed multi-factor experiment also allows researchers to determine if multiple factors interact to impact an individual’s response. For example, consider the hypothetical results from this experiment in Figure 3.1.³ The effect of relative humidity is to increase the growth rate for those individuals at 10°C (black line) but to decrease the growth rate for those individuals at 15°C (blue line). That is, the effect of relative humidity differs depending on the level of temperature. When the effect of one factor differs depending on the level of the other factor, then the two factors are said to *interact*. Interactions cannot be determined from the two single-factor experiments because the same individuals are not exposed to levels of the two factors at the same time.

Multi-factor experiments are used to detect the presence or absence of interaction, not just the presence of it. The hypothetical results in Figure 3.2 show that the growth rate increases with increasing relative humidity at about the same rate for both temperatures. Thus, because the effect of relative humidity is the same for each temperature (and vice versa), there does not appear to be an interaction between the two factors. Again, this could not be determined from the separate single-factor experiments.

- ◊ Multi-factor experiments can be used to detect interactions between multiple factors.

²

The real importance of this advantage will become apparent when statistical power is introduced in Chapter ??.

³

The means of each treatment are plotted and connected with lines in this plot.

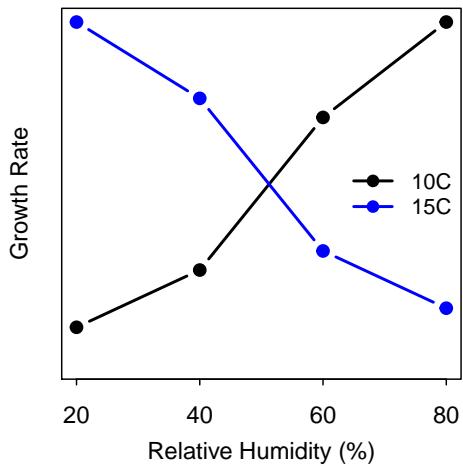


Figure 3.1. Mean growth rates in a two-factor experiment that depict an interaction effect.

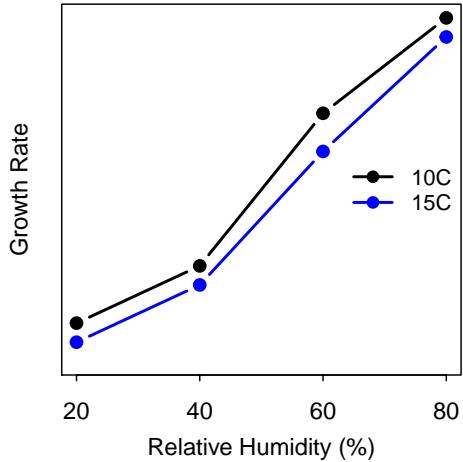


Figure 3.2. Mean growth rates in a two-factor experiment that depict no interaction effect.

3.1.3 Allocating Individuals

In the previous examples, each individual⁴ was allocated (i.e., placed into) treatment groups. Individuals should be allocated to treatment groups with a randomization procedure. Randomization will tend to even out differences among groups for variables not considered in the experiment. In other words, randomization should help assure that all groups are similar before the treatments are imposed. Thus, randomly allocating individuals to treatments removes any bias (foreseen or unforeseen) from entering the experiment.

In the single-factor experiment above – two treatments of temperature – there were 120 agars. To randomly allocate these individuals to the treatments, 60 pieces of paper marked with a “10” and 60 marked with a “15” could be placed into a hat. The pieces of paper would then be drawn for each agar and that agar would

⁴

When discussing experiments, an “individual” is often referred to as a “replicate” or an “experimental unit.”

receive the temperature found on the piece of paper. Alternatively, each agar could be assigned a unique number between 1 and 120 and pieces of paper with these numbers could be placed into the hat. The agars corresponding to the first 60 numbers drawn from the hat could then be placed into the first treatment with the agars for the next (or remaining) 60 numbers in the second treatment. This process is essentially the same as randomly ordering the 120 numbers. A random order of numbers is obtained with R by including the count of numbers as the only argument to `sample()`. For example, randomly ordering the numbers 1 through 120 is accomplished (and saved to an object) with

```
> ( ragars1 <- sample(120) )
[1]  80  30 100  90  21  68 104  79  64 106  98  16  73  91 107  1  60  54  26  99
[21] 108 111  31  47  57  92  5  58  37  50  34  88  41  66  65  29 110 113  4  75
[41]  93  23  49  97  35  84  74  7  15  39  70  94 114  14  71  20  33  67  86  8
[61]   6  28  52  48  13  18  63  72  69 120  55  83  42  3  77  82  38  22  96  43
[81]  56  89  78  17 112  44 103  46  59  85 109 115 118  87  32  62  51  95  24  40
[101] 119 102  19  27 116  36  2  12  45  53  11  76 117  61 105  9 101  25  81  10
```

Thus, the first five agars in the 10°C treatment are 80, 30, 100, 90, and 21. The first five agars in the 15°C treatment are 6, 28, 52, 48, and 13.

Now consider the modified experiment with two factors – temperature and relative humidity – with eight treatments containing 15 agars each. In this case, it is more efficient to save the sorted numbers into an object and then select the numbers in the first 15 positions, then the second 15 positions, etc. For example,

```
> ragars2 <- sample(120)
> ragars2[1:15]      # "grab" the first 15 numbers
[1]  61  82 103  31  66  81 105  40 104 106  5  9  71  36  8
> ragars2[16:30]     # "grab" the second 15 numbers, and so on
[1] 120   6  26  41  62 111  83  20  57  1  63  86  70  85  73
```

This design might be shown with the following table, where the numbers in each cell represent the first two agars selected to receive that treatment.⁵

| | | Relative Humidity | | | |
|------|-----|-------------------|-------------|-----------|------------|
| | | 20% | 40% | 60% | 80% |
| 10°C | 20% | 61,82,... | 120,6,... | 60,72,... | 89,49,... |
| | 40% | 78,10,... | 109,101,... | 22,2,... | 114,77,... |

◊ Individuals should be randomly allocated to treatments to remove bias.

3.1.4 Design Principles

There are many other methods of designing experiments and allocating individuals, including blocked designs, nested designs, etc., that are beyond the scope of this book. However, all experimental designs contain these three basic principles of experimental design.

⁵

Only the first two numbers are shown because of space constraints.

1. Control the effect of variables on the response variable by deliberately manipulating factors to certain levels and maintaining constancy among other confounding variables.
2. Randomize the allocation of individuals to treatments to eliminate bias in the experiment.
3. Replicate individuals (use many individuals) in the experiment to reduce chance variation in the results.

Proper control in an experiment allows the experiment to produce strong cause-and-effect statements; i.e., to state that an observed difference in the response variable was due to the levels of the factor or chance variation rather than some other variable (foreseen or unforeseen). Randomly allocating individuals to treatments removes any bias that may be included in the experiment. For example, if we don't randomly allocate the agars to the treatments, then it is possible that a set of all "poor" agars may end up in one treatment. In this case, any observed differences in the response may not be due to the levels of the factor but to the prior quality of the agars. Replication means that there should be more than one or a few individuals in each treatment. This reduces the effect of each individual on the overall results. For example, if there was one agar in each treatment then, even with random allocation, the effect of that treatment may be due to some inherent properties of that agar rather than the levels of the factors. Replication, along with randomization, helps assure that the groups of individuals in each treatment are as alike (homogeneous) as possible at the start of the experiment.

◊ Control, Randomization, and Replication are the three major principles of experimental design.

Review Exercises

- 3.1** While studying the foraging ecology of northern elephant seals, marine biologists from California observed the health of wild seals in fenced enclosures of two different water temperatures ($< 47^{\circ}\text{F}$ and $> 47^{\circ}\text{F}$) and compared these results to the health of domestic seals in two pools, with water temperatures analogous to the wild seals. The wild seals were allowed to eat what they wanted, but the domestic seals were fed a known diet. There were 20 wild seals and 20 domestic seals, each of which was randomly allocated to the two water temperatures (enclosures for the wild seals). Use this information to answer the questions below.

[Answer](#)

- (a) Construct a simple diagram to represent this experiment.
- (b) What is the response variable?
- (c) What are the factors (list all of them)?
- (d) How many levels are there (list in same order as factors in answer c)?
- (e) How many treatments are there?
- (f) How many replicates are there?

- 3.2** An agronomist is interested in the effect of plowing depth (10 cm, 17 cm, and 25 cm) and amount of applied fertilizer (none or 3 kg per acre) on the harvest of sugar beets. There are 36 nearly identical plots (fields) available for research. The agronomist has asked you to help design an experiment. Specifically, you are asked the questions below. [Answer](#)

- (a) What are the factors?
- (b) List the levels for each factor.

- (c) How many treatments?
- (d) How many replicates for each treatment?
- (e) Physically, what is a replicate in this case?
- (f)  Describe how you would allocate individuals to treatments. Show your R work.

3.3 Translocation is an important tool in modern wildlife management. Current techniques, however, result in the death of many translocated individuals shortly after release in their new homes. Researchers in France (Letty *et al.* 2000) simultaneously examined the use of tranquilization (tranquilized or not) and acclimatization pens (pens where an individual can “get used to” the new environment; used acclimatization pen or not) on the survival rate (survived or not) of translocated rabbits. Their experiment used a total of 64 European wild rabbits. Use this information to answer the questions below. [Answer](#)

- (a) Construct a diagram to represent this experiment.
- (b) What is the response variable?
- (c) What are the factors (list all of them)?
- (d) How many levels are there (list in same order as factors in answer c)?
- (e) How many treatments are there?
- (f) How many replicates are there?
- (g) What is an individual in this experiment?

3.4 In 1994, biologists studied the health of whitetail deer as it relates to eating habits. Sixty-four deer were randomly allocated into four groups. One group was to be kept on a deer farm and fed a strict diet. The other two groups would be sent to Channel Island off the coast of Alaska. One of the Channel Island groups would be restricted to browsing in prairies to simulate farm fields. The second was to be restricted to browsing in hardwood forests. The third Channel Island group would be fed a strict diet on the island. The researchers literally followed these deer around for 9 months, recording what the deer ate as they moved. Urine was also collected to assess the health of the deer. Use this information to answer the questions below. [Answer](#)

- (a) What is the response variable?
- (b) What are the factors (list all of them)?
- (c) How many levels are there (list in same order as factors in answer b)?
- (d) How many treatments are there?
- (e) How many replicates are there?
- (f) What is an individual in this experiment?

3.5 A chemical engineer is designing the production process for a new product. The chemical reaction that produces the product may have higher or lower yield, depending on the temperature and stirring rate in the vessel in which the reaction takes place. The engineer decides to investigate the effect on yield of two temperatures (50C and 60C) and three stirring rates (60, 90, and 120 rpm). A new vessel should be used for each production and only 30 vessels exist. Help the engineer set up this experiment by answering the questions below. [Answer](#)

- (a) What are the factors (list all of them)?
- (b) How many levels are there (list in same order as factors in answer a)?
- (c) How many treatments are there?
- (d) What is the response variable?
- (e) How many replicates are there?
- (f) Physically, what is a replicate (i.e., not a number)?
- (g)  Identify the individuals for each treatment. Show your R work.

(h) Use a simple table to diagram the experimental setup.

3.6

A student is designing an experiment to determine the simultaneous effects of calcium in the diet and regular exercise on blood pressure. In this experiment, some subjects will be given a calcium supplement pill and some will be given a placebo sugar pill. In addition, some subjects will be required to perform aerobic exercises once a day, whereas others will not. The researcher has 32 male subjects available that are as similar as possible (similar ages, weights, initial blood pressures, etc.). Help the student design this experiment by answering the questions below. [Answer](#)

- (a) What are the factors (list all of them)?
 - (b) How many levels are there (list in same order as factors in answer a)?
 - (c) How many treatments are there?
 - (d) What is the response variable?
 - (e) How many replicates are there?
 - (f) Physically, what is a replicate (i.e., not a number)?
 - (g)  Identify the individuals for each treatment. Show your R work.
 - (h) Use a simple table to diagram the experimental setup.
-

3.2 Observational Studies – Sampling

In observational studies the researcher has no control over any of the variables observed for an individual. The researcher simply observes individuals, disturbing them as little as possible, trying to get a “picture” of the population. Observational studies cannot be used to make cause-and-effect statements, because all variables that may affect the outcome may not have been measured or specifically controlled. Thus, any observed difference among groups may be caused by the variables measured, some other unmeasured variables, or chance (randomness).

Consider the following as an example of the problems that can occur when all variables are not measured in a study. For many years scientists thought that the brains of females weighed less than that of males. They used this finding to support all kinds of ideas about sex-based differences in learning ability. However, these earlier researchers failed to measure body weight, which has since been found to be strongly related to brain weight in both males and females. After controlling for the effect of differences in body weights, there was no difference in brain weights between the sexes. Thus, many sexist ideas persisted for years because cause-and-effect statements were inferred from data where all variables were not recorded.

◊ Strong cause-and-effect statements CANNOT be made from data collected as part of observational studies.

In observational studies, it is important to understand to what population inferences will be made.⁶ To make useful inferences from a sample, the sample must be an unbiased representation of the population. In other words, it must not systematically represent or favor certain individuals or outcomes.

⁶

Thus, it is very important to first perform an IVPPS as discussed in Section ??.

For example, consider that you want to determine the mean length of all fish in a particular lake (e.g., Square Lake as introduced in Section ??). Using a net with large mesh, such that only large fish are caught, would produce a biased sample because interest is in all fish not just the large fish in Square Lake. Setting the nets near spawning beds (i.e., only adult fish) would also produce a biased sample. In both instances, a sample would be collected from a population other than the population in which interest lies. This is why it is important that the population is specified and a sample selection technique is used that selects individuals from that population.

- ◊ It is important to understand what the population is before considering how to take a sample.

3.2.1 Types of Sampling Designs

Three common types of sampling schemes – voluntary response, convenience, and probability-based samples – are considered in this section. Voluntary response and convenience samples tend to produce biased samples, whereas probability-based samples, if done properly, will produce an unbiased sample.

A voluntary response sample is a sample of individuals that choose themselves for the sample by responding to a general appeal. An example of a voluntary response sample would be the group of people that respond to a general appeal placed in the school newspaper. If the population of interest in this sample was all students at the school, then this type of general appeal would likely produce a biased sample of students that (i) read the school newspaper, (ii) feel strongly about the topic, or (iii) both.

A convenience sample is a sample of individuals who are easiest to reach for the researcher. An example of a convenience sample might be a situation where a researcher queried only those students in a particular class. This sample is “convenient” because the individuals are easy to gather. However, if the population of interest was all students at the school, then this type of sample would likely produce a biased sample of students that is likely (i) of one major or another, (ii) in one or two-years (e.g., Freshman or Sophomores), or (iii) both.

Δ **Voluntary Response Sample:** A sample of individuals that choose themselves for the sample by responding to a general appeal.

Δ **Convenience Sample:** A sample of individuals who are easiest to reach for the researcher.

- ◊ Voluntary response and convenience samples often produces a biased sample.

In probability-based sampling, each member of the population has a known chance of being selected for the sample. The simplest form of probability-based sample is the **Simple Random Sample** (SRS) where each member of the population has the same chance of being selected. Proper selection of an SRS requires each member of the population to be assigned a unique number. The individuals in the SRS are then selected by choosing random numbers and collecting the individuals that those numbers correspond to.

For example, an auditor may need to select a sample of 30 financial transactions from all transactions of a particular bank during the previous month. Because each transaction is numbered, the auditor may know that there was a total of 1112 transactions at this bank during the previous month (i.e., the population).

The auditor would then number each transaction from 1 to 1112 (this is likely already done in this case), randomly select 30 numbers (with no repeats) from between 1 and 1112, and then physically locate the 30 transactions that correspond to the 30 selected numbers. Those 30 transactions are the SRS. A random selection of numbers is selected by including the total number as the first argument and the number to select as the second argument to `sample()`. For example, 30 numbers from between 1 and 1112 is selected (and saved into an object) with

```
> ( raccts <- sample(1112,30) )
[1]  75 320 874 104 128 870 607 1091 1030 1053 1031 518 433 893 816 903
[17] 342 1016 136 580 670 376 576 1076 1034 365 492 189 409 66
```

Thus, the first five accounts to be selected would be the accounts corresponding to the numbers 75, 320, 874, 104, and 128.

There are other more complex types of probability-based samples – e.g., stratified samples and nested or multistage samples – that are beyond the scope of this book. However, the goal of these more complex types of samples is generally to impart more control into the sampling design.

Δ Probability-based Sample: A sample where each individual of the population has a known chance of being selected for the sample.

Δ Simple Random Sample: A probability-based sample where each individual of the population has the same chance of being selected for the sample. Usually abbreviated as SRS.

◊ To conduct a proper SRS each individual of the population must be able to be assigned a unique number.

If the population is such that a numerical label cannot be assigned to each individual, then the researcher must try to use a method of selection for which they feel each individual has an equal chance of being selected. Usually this means randomizing the technique rather than the individuals. In the fish example discussed on the previous page, the researcher may consider choosing random mesh sizes, random locations for placing the net, or random times for placing the net. Thus, in many real-life instances the researcher simply tries to use a method that is likely to produce an SRS or something very close to it.

◊ If a number cannot be assigned to each individual in the population, then the researcher should randomize the “technique” to assure as close to a random sample as possible.

Polls, campaign or otherwise, are examples of observational studies that you are probably familiar with. The following are links to sites that discuss various aspects of polling.

- How Polls are Conducted by Frank Newport, Lydia Saad, and David Moore, The Gallup Organization.
- Why Do Campaign Polls Zigzag So Much? by G.S. Wasserman, Purdue U.

3.2.2 Of What Value are Observational Studies?

In the last two sections it may have become apparent that properly designed experiments can lead to “cause-and-effect” statements, whereas observational studies (even properly designed) are unlikely to lead to such statements. Furthermore, in the last section, it was suggested that it is very difficult to take a proper probability-based sample because it is hard to assign a number to each individual in the population (precisely because entire populations are very difficult to “see”). So, do observational studies have any value? There are at least three reasons why observational studies are useful.

The scientific method begins with a scientist making an observation about a natural phenomenon. Observational studies may serve to provide such an observation. Alternatively, observational studies may be deployed after an observation has been made to see if that observation is “prevalent” and worthy of further investigation. Thus, observational studies may lead directly to hypotheses that form the basis of experiments.

Experiments are often conducted under very confined and controlled conditions so that the effect of one or more factors on the response variable can be identified. However, at the conclusion of an experiment it is often questioned whether a similar response would be observed “in nature” under much less controlled conditions. For example, one might determine that a certain fertilizer increases growth of a certain plant in the greenhouse, with consistent soil characteristics, temperatures, lighting, etc. However, it is a much different, and, perhaps, more interesting, question to determine if that fertilizer elicits the same response when applied to an actual field.

Finally, there are situations where conducting an experiment simply cannot be done, either for ethical, financial, size, or other constraints. For example, it is generally accepted that smoking causes cancer in humans even though an experiment where one group of people was forced to smoke while another was not allowed to smoke has not been conducted. Similarly, it is also very difficult to perform valid experiments on “ecosystems.” In these situations, an observational study is simply the best study allowable. Cause-and-effect statements are arrived at in these situations because observational studies can be conducted with some, though not absolute, control and control can be imparted mathematically into some analyses.⁷ In addition, a “preponderance of evidence” may be arrived at if enough observational studies point to the same conclusion.

⁷

These analyses are beyond the scope of this book, though.

Review Exercises

- 3.7** The National Institutes of Health (NIH) established the Women's Health Initiative (WHI) in 1991 to address the most common causes of death, disability and impaired quality of life in postmenopausal women. The WHI addressed cardiovascular disease, cancer, and osteoporosis. The WHI was a 15 year multi-million dollar endeavor, and one of the largest U.S. prevention studies of its kind. One aspect of the WHI enlisted 93,676 postmenopausal women between the ages of 50-79 from 40 Clinical Centers from throughout the United States (see [this map](#)). The women were not asked to take any medication or change their health habits. The health of OS participants was tracked over an average of eight years through asking the women to fill out periodic health forms. What type of study is this? [Answer](#)
- 3.8** The U.S. Department of Transportation sponsored a study to determine the transportation patterns and motivations for driving among offenders before, during, and after suspension of their driver's license for an alcohol-related offense (more information [here](#).). For each subject their travel patterns were examined for two four-hour periods during the last month of the suspension period (one observation Monday-Thursday 6 - 10 a.m. and the other observation Friday or Saturday evening 6 - 10 p.m.; the actual days were randomly selected). These observation periods were selected to include a time period when the subject would likely be traveling to work and a time period when the subject would likely be traveling for personal, recreational, or social reasons. Similar examinations were conducted at least one month after drivers had had their license reinstated. These post-suspension observations were to be conducted for each subject at the same times of day and days of the week as the during-suspension observations. What type of study is this? Why?
[Answer](#)
- 3.9** I have noticed that the needles of white pine trees near major highways are brown. I hypothesized that this may be caused by increased levels of carbon monoxide (CO; ppm) and salt (ppt) near the roads. I am considering two studies to test this hypothesis. First, at two types of sites – near highways and far from highways – I will count the number of trees that are mostly brown and measure levels of CO and salt. Second, I will determine the effect of CO and salt levels by growing randomly-selected nearly-identical seedlings in pots that only differ in the levels of CO and salt – 0 and 5 ppm CO and 0 and 4 ppt salt (NOTE: the 0 levels correspond to normal background levels). I have 20 trees to allocate to the different treatments.
[Answer](#)
- (a) Use a diagram to clearly depict the experimental situation described above.
 - (b)  Write the numeric label for each individual in the appropriate place on your diagram. Show your R work.
 - (c) In the experiment, which treatment is considered a control? Why?
 - (d) Which study will provide a definitive answer to my stated hypothesis? Explain why!
-

CHAPTER 4

GETTING STARTED WITH R

Chapter Objectives:

1. Understand the difference between R expressions and assignments.
2. Understand the different types of data that can be stored in R.
3. Understand the different types of data structures used in R.
4. Be able to enter data into R data.frames.
5. Be able to isolate individual variables and individuals in R.
6. Be able to create data.frames that are subsets of larger data.frames.
7. Understand how homework assignments should be formatted.

Contents

| | | |
|-----|------------------------------------|----|
| 4.1 | Setting Up R and Helpers | 34 |
| 4.2 | Working With R Basics | 34 |
| 4.3 | Working With Data | 36 |

4.1 Setting Up R and Helpers

Detailed methods for downloading, installing, and configuring R, RStudio, and NCStats on your personal computer are given on the [Resources page of the course website](#).

4.2 Working With R Basics

4.2.1 Saving Results

Results are not saved in R or RStudio. Rather, “scripts” of successful R commands are saved and, then, if the analysis needs to be re-done, the entire set of commands is opened in RStudio and run again. When writing a report, all tabular and graphical output should be copied from RStudio and pasted into your report document. This document will serve as your analysis report and can be modified to include answers to questions, references to the tables and graphs, etc.¹ All data that is not a simple vector (see Section 4.3.4) should be entered into R through text files (see Section 4.3.2).

R does allow one to save a “workspace”, though I urge you not to do that. Rather, save your “good” commands in a script and save your “good” results in a report document; do not save the workspace.

- ◊ Do NOT save the workspace in R.

4.2.2 Expressions and Assignments

Expressions in R are mathematical “equations” that are evaluated by R with a result seen immediately. An example of an expression in R is

```
> 5+log(7)-pi  
[1] 3.804317
```

where `log()` and `pi` are built-in functions used to compute the natural log and find the value of π , respectively. Expressions in R are like using a calculator where the result is shown, but not saved for subsequent analyses. In addition, expressions in R follow the same order of operations and use of parentheses as expressions entered into your calculator.

- ◊ The results of expressions in R are temporary unless the result is assigned to an object.

Results from an expression are typically saved for further computations by assigning the results of the expression to an object with the assignment operator (i.e., `<-`). The general form for saving the result of an expression into an object is `object <- expression`. The result of the expression will not be seen unless the object name is subsequently typed into R (but see below). For example, the result of the previous expression is saved into an object called `x` and then viewed with

```
> x <- 5+log(7)-pi  
> x  
[1] 3.804317
```

¹

Specifics for how to format homework assignments are on the course syllabus

A short-cut for both assigning and printing the result of an expression is to surround the entire command in parentheses. For example, the following assigns the result of the expression to `y` and prints the result.²

```
> ( y <- 15*exp(2) )
[1] 110.8358
> # note that the next line is now not needed
> y
[1] 110.8358
```

- ◊ The convention of surrounding commands in parentheses to both assign and print the results will be used extensively in this book to save space.

An object can be named whatever you want, with the exception that it cannot start with a number, contain a space, or be the name of a reserved word or function in R (e.g., `pi` or `log`). Furthermore, object names should be short and simple enough that you can remember what is contained in the object. It is also good practice to view the object immediately after making the assignment to make sure that it contains results that seem appropriate.

◊ In general, computational results should be assigned to an object.

◊ Type the name of the object after making the assignment to confirm the results.

Review Exercises

4.1  Compute the value of $\frac{3}{7} + \frac{1}{2}$. [Answer](#)

4.2  Compute the value of $\pi * 3.7^2$. [Answer](#)

4.3  Assign the value of 3.7 to `r`. [Answer](#)

4.4  Compute the value of πr^2 using the value of `r` assigned in the previous problem. [Answer](#)

4.5  Assign the value 1.2 to `r` and then re-evaluate πr^2 . [Answer](#)

²

Note that the spaces after the opening parentheses and before the closing parentheses are not needed. However, they make the code more legible.

4.2.3 Functions and Arguments

R contains many “programs,” or functions, to perform particular tasks. A function is “called” by typing the function name followed by open and closed parentheses. Arguments, which the function will use to perform its task, are contained within the parentheses. The `log()` function, used in the previous section, is an example of a function. The name of the function is `log` and the argument, the number for which to compute the natural log, is contained within the parentheses following the function name. Many other functions will be described below and in subsequent chapters.

△ **Function:** An R program that performs a particular task.

△ **Argument:** A “directive” that is provided to a function. Arguments are contained within parentheses that follow the function name.

◊ Regular curved parentheses have two primary uses in R: (1) to control order of operations in expressions (as with a calculator) and (2) to contain the arguments sent to a function.

4.3 Working With Data

4.3.1 Data Types

Data in R will be designated as an integer (whole numbers), numeric (non-integer numerical values), character (strings), factor (group membership), or logical (TRUE/FALSE). The type of data largely dictates the types of analyses that can be performed. Data types will be discussed in more detail as needed. Note, however, that the **factor** data type is a special case of the character data type, where the specific items describe to which group an individual belongs. This description allows certain types of analyses to be performed in later chapters.

△ **Factor:** A special type of variable that identifies the group to which an individual belongs.

4.3.2 Entering Data

For real data (i.e., several variables from many individuals) it is most efficient to enter data into a comma-separated values (CSV) file and then import that file into R. Creating a CSV file with Microsoft Excel is described below, though there are other ways to create CSV files (see [FAQs on class webpage](#)). This explanation assumes that you have a basic understanding of Excel (or other spreadsheet softwares).

◊ Realistic datasets are most efficiently entered into a comma-separated values (CSV) file in preparation for importing into R.

The spreadsheet should be organized with variables in columns and individuals in rows, with the exception that the first row should contain variable names. The example spreadsheet below shows the length (cm),

weight (kg), and capture location data for a small sample of Black Bears.

| | A | B | C |
|----|-----------|-----------|----------|
| 1 | length.cm | weight.kg | loc |
| 2 | 139 | 110 | Bayfield |
| 3 | 138 | 60 | Bayfield |
| 4 | 139 | 90 | Bayfield |
| 5 | 120.5 | 60 | Bayfield |
| 6 | 149 | 85 | Bayfield |
| 7 | 141 | 100 | Ashland |
| 8 | 141 | 95 | Ashland |
| 9 | 150 | 85 | Douglas |
| 10 | 166 | 155 | Douglas |
| 11 | 151.5 | 140 | Douglas |
| 12 | 129.5 | 105 | Douglas |
| 13 | 150 | 110 | Douglas |

△ **Data Frame:** A two-dimensional organization of variables (as columns) recorded on multiple individuals (as rows) where the columns may be of different data types.

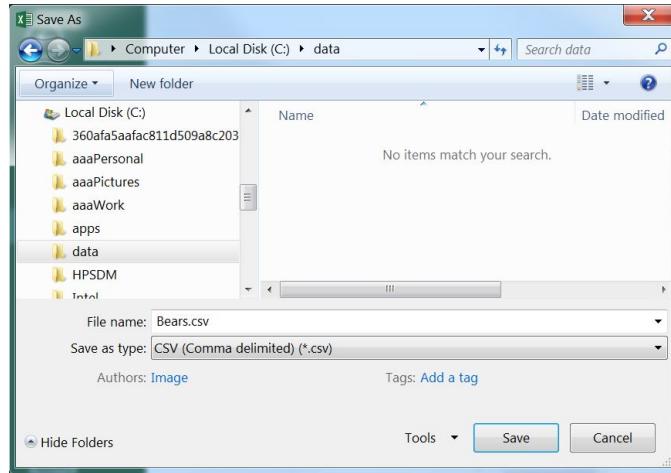
◊ A **data.frame** may have columns of different data types (e.g., numeric and character) whereas a **vector** must have columns all of the same data type.

◊ The columns of a **data.frame** correspond to variables and the rows of a **data.frame** correspond to individuals.

Variable names must NOT contain spaces. For example, don't use *total length* or *length (cm)*. If you feel the need to have longer variable names, then separate the parts with a period (e.g., *length.cm*) or an underscore (e.g., *length_cm*). Furthermore, numerical measurements should NOT include units (e.g., don't use *7 cm*). Finally, for categorical data, make sure that all categories are consistent (e.g., do not have a column that contains both *bayfield* and *Bayfield*).

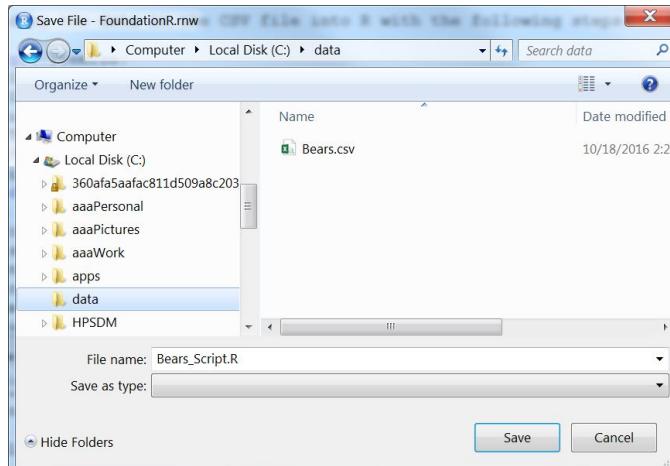
◊ **Variable names and data should not contain spaces. The Error in scan error message usually indicates that you have spaces in the variable names or data.**

The spreadsheet is saved as a CSV file by selecting the **File..Save As** menu item, which will produce the dialog box below. In this dialog box, change **Save as type** to **CSV (Comma delimited) (*.csv)** (you may have to scroll down), provide a file name (don't have any periods in the name besides for ".csv", which you should not have to type), select a location to save the file (don't forget this location!!), and press **Save**. Two "warning" dialog boxes may then appear – select **OK** for the first and **YES** for the second. You can now close the spreadsheet file (you may be asked to save changes – you should say **No**).



The following steps should be used to load the data in the CSV file into R.

- Open RStudio.
- Open a new script by selecting the **File**, **New File**, **R Script** menu items.
- Type **library(NCStats)** in the new script.
- Save this script by selecting the **File**, **Save** menu items. In the ensuing dialog box, navigate to the **exact same directory** where you saved the data, type a name for the file in the **File name:** box (**do not use a period in this name!!**), and press **Save**.



- Set the working directory (tell R where the file is) with the **Session**, **Set Working Directory . . .**, **To Source File Location** menu items in RStudio. RStudio will print an appropriate **setwd()** command to the console. Copy this command from the console to the second line in your script.³ For example, I stored the file created above in the **C:/data** directory, so that RStudio will create this **setwd("C:/data")**.
- The CSV file is read into R by including the name of the file (in quotes) in **read.csv()**. For example, "**Bears.csv**" is read into R and stored into an object called **bears** with **bears <- read.csv("Bears.csv")**.
- One should check the data in this object as described in Section 4.3.3 below

³

Doing this will eliminate the need to manually select the menu options every time you want to run this script.

- ◊ Data stored in an external CSV file is read into R with `read.csv()`.

It is important that each row of the `data.frame` correspond to one individual. This is critically important when data are recorded for two different groups (e.g., for a two-sample t-test; see Section ??). For example, consider the following data of methyl mercury levels recorded in mussels from a location labeled as “impacted” and a separate location labeled as “reference.”

```
impacted  0.011  0.054  0.056  0.095  0.051  0.077
reference 0.031  0.040  0.029  0.066  0.018  0.042  0.044
```

To follow the “one individual per row” rule, these data are entered in stacked format where the “reference” data are stacked underneath the “impacted” data and a column is used to indicate to which group the individuals belong. For example, the Excel file for data entry would look like the following

| | A | B |
|----|-----------|-------|
| 1 | loc | merc |
| 2 | impacted | 0.011 |
| 3 | impacted | 0.054 |
| 4 | impacted | 0.056 |
| 5 | impacted | 0.095 |
| 6 | impacted | 0.051 |
| 7 | impacted | 0.077 |
| 8 | reference | 0.031 |
| 9 | reference | 0.04 |
| 10 | reference | 0.029 |
| 11 | reference | 0.066 |
| 12 | reference | 0.018 |
| 13 | reference | 0.042 |
| 14 | reference | 0.044 |

- ◊ All data files should be constructed so that there is only data from one individual in each row.

Alternative Forms of Getting Data

Some of the data files that you will use are provided on the [Data for MTH107](#) resource page of the class webpage. In these cases, the data should be downloaded from the webpage and saved in the same directory or folder as your analysis script. The downloaded file is then read into R in the same manner as described previously (i.e., set the working directory with `setwd()` and use `read.csv()`).

A few data files used in this book are supplied with R or the NCStats package. These files are loaded with `data()`. For example, the `iris` data file is loaded into R with

```
> data(iris)
```

4.3.3 Working With Data Frames

Viewing a Data Frame

Many users are disoriented in RStudio because they cannot “see” their data in the same way that they see it in a spreadsheet program. There are, however, several options for viewing your data. First, you can type the name of the data.frame object to see its entire contents.

```
> bears
  length.cm weight.kg      loc
1     139.0       110 Bayfield
2     138.0        60 Bayfield
3     139.0        90 Bayfield
4     120.5        60 Bayfield
5     149.0        85 Bayfield
6     141.0       100 Ashland
7     141.0        95 Ashland
8     150.0        85 Douglas
9     166.0       155 Douglas
10    151.5       140 Douglas
11    129.5       105 Douglas
12    150.0       110 Douglas
```

Typing the name is adequate for small data.frames, but not useful for large data.frames. The entire data.frame is opened in a separate window by double-clicking on the name of the data.frame in the **Environment** tab of RStudio. Alternatively, the first and last three rows are viewed by including the data.frame object in **headtail()**.

```
> headtail(bears)
  length.cm weight.kg      loc
1     139.0       110 Bayfield
2     138.0        60 Bayfield
3     139.0        90 Bayfield
10    151.5       140 Douglas
11    129.5       105 Douglas
12    150.0       110 Douglas
```

In addition to viewing the contents of the data.frame, it is often useful to examine the structure of the data.frame returned from **str()**. In this example, it is seen that three variables were recorded on 12 individuals. The first variables – *length.cm* and *weight.kg* – are numerical measurements made on the bears. The last variable – *loc* – is a factor variable that records the capture location for each bear.

```
> str(bears)
'data.frame': 12 obs. of  3 variables:
 $ length.cm: num  139 138 139 120 149 ...
 $ weight.kg: int  110 60 90 60 85 100 95 85 155 140 ...
 $ loc       : Factor w/ 3 levels "Ashland","Bayfield",...: 2 2 2 2 2 1 1 3 3 3 ...
```

The levels of the *loc* variable may be seen by including this variable (with the data.frame name) as the argument to **levels()**.

```
> levels(bears$loc)
[1] "Ashland"  "Bayfield" "Douglas"
```

In the previous example, the `$` notation was used to identify a particular variable (i.e., `loc`) from within a particular data.frame (`bears`). Think of variables as being nested inside of a data.frame and, thus, to access the variable you must first identify the data.frame in which it exists and then the name of the variable. The `$` simply separates the data.frame from the variable.

```
> bears$length.cm
[1] 139.0 138.0 139.0 120.5 149.0 141.0 141.0 150.0 166.0 151.5 129.5 150.0
> bears$loc
[1] Bayfield Bayfield Bayfield Bayfield Bayfield Ashland Ashland Douglas Douglas
[10] Douglas Douglas Douglas
Levels: Ashland Bayfield Douglas
```

- ◊ A dollar sign is ONLY used in R to separate the name of a data.frame from the name of a variable within that data.frame.

Selecting Individuals

In some instances, it may be important to select or exclude an individual from a data.frame. Data.frames are two-dimensional objects that are indexed by a row and a column, in that order. Positions within a data.frame are selected within paired square brackets. For example, the item in the third row and second column of `bears` is selected with

```
> bears[3,2]
[1] 90
```

- ◊ Identifying the position of an item in an object is the ONLY time that square brackets are used in R.

An entire row or column may be selected by omitting the other dimension. For example, one could select the entire second column with `bears[,2]`, but this is also the `weight.kg` variable and is better selected, as shown above, with `bears$weight.kg`. As a better example, the entire third row is selected with (note that the column designation was omitted)

```
> bears[3,]
  length.cm weight.kg      loc
3       139        90 Bayfield
```

More than one row can be selected by first combining the row indices together with `c()` (more about `c()` in Section 4.3.4). For example, the third, fifth, and eighth rows are selected with (again, the column index is omitted)

```
> bears[c(3,5,8),]
  length.cm weight.kg      loc
```

```

3      139      90 Bayfield
5      149      85 Bayfield
8      150      85 Douglas

```

Finally, rows can be excluded by preceding the row indices with a negative sign.

```

> bears[-c(3,5,8,10,12),]
  length.cm weight.kg      loc
1     139.0       110 Bayfield
2     138.0        60 Bayfield
4     120.5        60 Bayfield
6     141.0       100 Ashland
7     141.0        95 Ashland
9     166.0       155 Douglas
11    129.5       105 Douglas

```

Filtering a data.frame

It is common to create a new data.frame that contains only some of the individuals from an existing data.frame. For example, a researcher may want a data.frame of only bears captured in Bayfield County or bears that weighed more than 100 kg. The process of creating the newer, smaller data.frame is called filtering (or subsetting) and is accomplished with `filterD()`. The `filterD()` function requires the original data.frame as the first argument and a condition statement as the second argument. The condition statement is used to either include or exclude individuals from the original data.frame. Condition statements consist of the name of a variable in the original data.frame, a comparison operator, and a comparison value (Table 4.1). The result from `filterD()` should be assigned to an object, which is then the name of the new data.frame.

◊ The `filterD()` function is used to create a new data.frame that consists of individuals selected by some criterion from an existing data.frame.

Table 4.1. Condition operators used in `filterD()` and their results. Note that *variable* generically represents a variable in the original data.frame and *value* is a generic value or level. Both *variable* and *value* would be replaced with specific items.

| Condition Operator | Individuals Returned from Original Data Frame |
|------------------------------------|--|
| <code>variable == value</code> | all individual that are equal to the given value |
| <code>variable != value</code> | all individuals that are NOT equal to the given value |
| <code>variable > value</code> | all individuals that are greater than the given value |
| <code>variable >= value</code> | all individuals that are greater than or equal to the given value |
| <code>variable < value</code> | all individuals that are less than the given value |
| <code>variable <= value</code> | all individuals that are less than or equal to the given value |
| <code>condition , condition</code> | all individuals that meet both conditions |
| <code>condition condition</code> | all individuals that meet one or both conditions ⁴ |

The following are examples of new data.frames created from `bears`. The name of the new data.frame (i.e., object to the left of the assignment operator) can be any valid object name. As demonstrated below, the new

`data.frame` (or its structure) should be examined after each filtering to ensure that the `data.frame` actually contains the items that you desire.

- Only individuals from *Bayfield* county.

```
> bf <- filterD(bears, loc=="Bayfield")
> bf
  length.cm weight.kg      loc
1     139.0      110 Bayfield
2     138.0       60 Bayfield
3     139.0       90 Bayfield
4    120.5       60 Bayfield
5     149.0       85 Bayfield
```

- Individuals from both *Bayfield* and *Ashland* counties.

```
> bfash <- filterD(bears, loc %in% c("Bayfield", "Ashland"))
> bfash
  length.cm weight.kg      loc
1     139.0      110 Bayfield
2     138.0       60 Bayfield
3     139.0       90 Bayfield
4    120.5       60 Bayfield
5     149.0       85 Bayfield
6     141.0      100 Ashland
7     141.0       95 Ashland
```

- Individuals with a weight greater than 100 kg.

```
> gt100 <- filterD(bears, weight.kg>100)
> gt100
  length.cm weight.kg      loc
1     139.0      110 Bayfield
2     166.0      155 Douglas
3     151.5      140 Douglas
4     129.5      105 Douglas
5     150.0      110 Douglas
```

- Individuals from *Douglas* County that weigh at least 150 kg.

```
> do150 <- filterD(bears, loc=="Douglas", weight.kg>=150)
> do150
  length.cm weight.kg      loc
1       166      155 Douglas
```

◊ View or “structure” the `data.frame` created from using `filterD()` to assure that it contains data.

Review Exercises

4.6

 Two students at Seattle Community College made biometric measurements on 25 Douglas fir (*Pseudotsuga menziesii*) trees in the lowlands of western Washington. The variables recorded in the [DougFirBiometrics.csv](#) file are a unique tree identifier (*tree*), the observer's name (*observer*; either "Ingrid" or "Dylan"), the circumference at breast height (meters; *circ*), the height to the eye of the observer (meters; *eyeht*), the horizontal distance between observer and tree (meters; *horizdist*), the angle between observer and top of tree (degrees; *angle*), and the estimated height of tree (meters; *height*) using right-angle trigonometry.

[Answer](#)

- Read this data file into an object called *DF*.
- Examine the structure of this *data.frame*.
- Show all measurements made on the third tree. [Do not do this manually; use R code.]
- Show all estimated tree heights.
- Show the estimated tree height for the fifth tree.
- Show all measurements for all trees measured by "Ingrid". [HINT: use filtering.]
- Show all estimated tree heights for all trees measured by "Dylan". [HINT: use filtering.]
- Show all measurements for tree heights less than 10 m. [HINT: use filtering.]
- Show all measurements for tree heights greater than 10 m and circumference less than 1 m. [HINT: use filtering.]

4.3.4 Vectors

Data.frames are the primary structure in which to store real data. However, much simpler situations that don't require a *data.frame* may arise. In R, items of the same data type (Section 4.3.1) are stored in a one-dimensional "list" called a *vector*. Vectors are usually displayed in one row (with many columns), but they may also be thought of as a single column (with many rows). Items are entered into a vector with *c()*, where the individual arguments are specific numbers, characters, or logical values.⁵ Items for a vector of characters must be contained within paired quotes.

```
> ( v <- c(1,2,5) )
[1] 1 2 5
> ( y <- c("Iowa","Minnesota","Wisconsin") )
[1] "Iowa"      "Minnesota" "Wisconsin"
```

Δ Vector: A one-dimensional list of items of the same data type. The primary information storage unit in R.

Single variables from a *data.frame* are vectors. Vectors that are not extracted from a *data.frame* will only be used in this course for very simple lists of items, usually as arguments in a function.

⁵

Note that *c* comes from the word "concatenate."

```
> bears$length.cm  
[1] 139.0 138.0 139.0 120.5 149.0 141.0 141.0 150.0 166.0 151.5 129.5 150.0
```

- ◊ The columns of a data.frame are accessed with the name of the data.frame, a dollar sign, and then the name of the variable – i.e., generically, `dataframe$varname`.

Review Exercises

4.7  Create a vector called `h` that contains nine heights of people. [Answer](#)

4.8  Create a vector called `w` that contains nine weights of people. [Answer](#)

4.9  Create a vector called `hc` that contains nine hair colors of people. [Answer](#)

4.10  Create a vector called `m` that contains nine logical values (=TRUE if male). [Answer](#)

4.11  Using the vectors from the previous questions, [Answer](#)

- (a) ... create the largest possible data.frame (use `data.frame()`).
- (b) ... identify the height of the third individual of this data.frame.
- (c) ... identify the hair color for the sixth individual of this data.frame.

Part II

Exploratory Data Analysis

CHAPTER 5

UNIVARIATE EDA - QUANTITATIVE

Chapter Objectives:

1. Construct histograms with quantitative data,
2. Use graphs to describe the shape of a distribution, and
3. Use graphs to describe outliers in a distribution.
4. Calculate summary statistics for measuring the center of quantitative data,
5. Calculate summary statistics for measuring the dispersion of quantitative data,
6. Describe the underlying differences in how the different statistics measure center and dispersion,
7. Identify which summary statistics are appropriate in a given situation, and
8. Construct an appropriate overall numerical summary.

Contents

| | | |
|-----|-----------------------------------|----|
| 5.1 | Items to Describe | 48 |
| 5.2 | Histograms | 49 |
| 5.3 | Interpreting Shape | 52 |
| 5.4 | Interpreting Outliers | 54 |
| 5.5 | Measures of Center | 55 |
| 5.6 | Measures of Dispersion | 61 |
| 5.7 | Overall Summaries | 66 |
| 5.8 | Multiple Groups | 70 |
| 5.9 | Example Interpretations | 72 |

ONCE DATA HAVE BEEN COLLECTED (Chapter 3), it is important to develop a “feel” for the data, to identify what types of values each variable takes, and to determine if there are any “issues” in the data. This first step in a statistical analysis is called EXPLORATORY DATA ANALYSIS (EDA). We will begin by examining the distribution of each variable by itself, called a univariate EDA, and then examine pairs of variables, called a bivariate EDA (see Chapters 8 and 9). Additionally, the methods employed differ for quantitative and categorical variables. Quantitative variables are the focus of this chapter, whereas categorical variables are the focus of Chapter 6.

5.1 Items to Describe

A univariate EDA for a quantitative variable is concerned with describing the distribution of the values for that variable; i.e., describing what values occurred and how often those values occurred. Specifically, the distribution is described by four specific attributes:

1. **shape** of the distribution,
2. presence of **outliers**,
3. **center** of the distribution, and
4. **dispersion** or spread of the distribution.

Graphs are best for identifying shape and the presence of outliers and for getting a general feel for the center and dispersion of the data. However, numerical summaries are better for identifying the center and dispersion of the data.

- ◊ Shape, center, dispersion, and outliers are described for each quantitative variable.
- ◊ Shape and outliers are described from graphs; center and dispersion with numerical summaries.

Three primary data sets will be explored throughout this chapter.

- Measurements of water consumption in one hour by mice (Table 5.1).¹
- Richter scale recordings for 15 major earthquakes (Table 5.2).
- The number of days of ice cover at ice gauge station 9004 in Lake Superior (data originally from the [National Snow and Ice Data Center](#) and found in the [LakeSuperiorIce.csv](#) data file on the class webpage²). The days variable is the total number of days of ice cover at this site for nearly every ice season from 1955-56 to 1996-97 (three years were missing). These data are loaded into *LSI* below.

```
> LSI <- read.csv("data/LakeSuperiorIce.csv")
```

Table 5.1. Amount of water consumed (in ml) in one hour by a sample of mice.

| | | | | | | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 10.6 | 14.1 | 13.7 | 15.2 | 15.4 | 12.5 | 12.9 | 14.3 | 13.0 | 16.6 | 11.5 | 9.4 | 16.5 | 13.7 | 14.7 |
| 12.6 | 12.0 | 14.0 | 10.0 | 18.2 | 18.4 | 17.4 | 11.1 | 15.8 | 15.8 | 16.6 | 11.4 | 17.0 | 13.6 | 13.5 |

¹

See Section 4.3.2 for how to enter these data into R.

²

See Section 4.3.2 for more description of how to access these data.

Table 5.2. Richter scale recordings for 15 major earthquakes.

| | | | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 5.5 | 6.3 | 6.5 | 6.5 | 6.8 | 6.8 | 6.9 | 7.1 | 7.3 | 7.3 | 7.7 | 7.7 | 7.7 | 7.8 | 8.1 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|

5.2 Histograms

5.2.1 General Construction

A histogram is a plot of the frequency of occurrence of individuals (y-axis) in classes of values of the variable (x-axis). The steps for constructing a histogram from raw data are:

1. Create categorical classes of values for the variable of interest,
2. Count the frequency of individuals in each class,
3. Construct a graph template with values of the variable on the x-axis and frequency of individuals on the y-axis, and
4. Draw bars on the graph that are as wide as the class of values and as tall as the frequency of individuals.

These steps are illustrated below with the mouse water consumption data.

The easiest way to create a list of classes of the water consumption data by hand is to find the difference between the maximum and minimum value in the data set and divide this value by the number of desired bars. The number of bars is usually a “nice” number near eight to ten. This number is typically rounded up to make classes that are easy to work with. The range of values in this example is $18.4 - 9.4 = 9.0$. A “nice” value between eight and ten to divide the range by in this example is nine. Thus, the classes of data should be one unit wide. To make the histogram as easy as possible to construct by hand it is best to start the classes at 9 mm (Table 5.3).

Table 5.3. Frequency table of mouse consumption values in one-unit classes.

| Class | Frequency |
|-----------|-----------|
| 9.0– 9.9 | 1 |
| 10.0–10.9 | 2 |
| 11.0–11.9 | 3 |
| 12.0–12.9 | 4 |
| 13.0–13.9 | 5 |
| 14.0–14.9 | 4 |
| 15.0–15.9 | 4 |
| 16.0–16.9 | 3 |
| 17.0–17.9 | 2 |
| 18.0–18.9 | 2 |

The number of individuals with a value of the variable in each class are called frequencies and are shown in the second column of Table 5.3.

The plot is prepared with values of the classes forming the x-axis and the frequencies forming the y-axis (Figure 5.1-Left). The first bar added to this skeleton plot has the bottom-left corner at 9 and the bottom-right corner at 10 on the x-axis, and a height equal to the frequency of individuals in the 9 to 9.9 class (Figure 5.1-Center). A second bar is then added with the bottom-left corner at 10 and the bottom-right corner at 11 on the x-axis, and a height equal to the frequency of individuals in the 10 to 10.9 class (Figure 5.1-Right). This process is continued with the remaining classes until the full histogram is constructed (Figure 5.2).

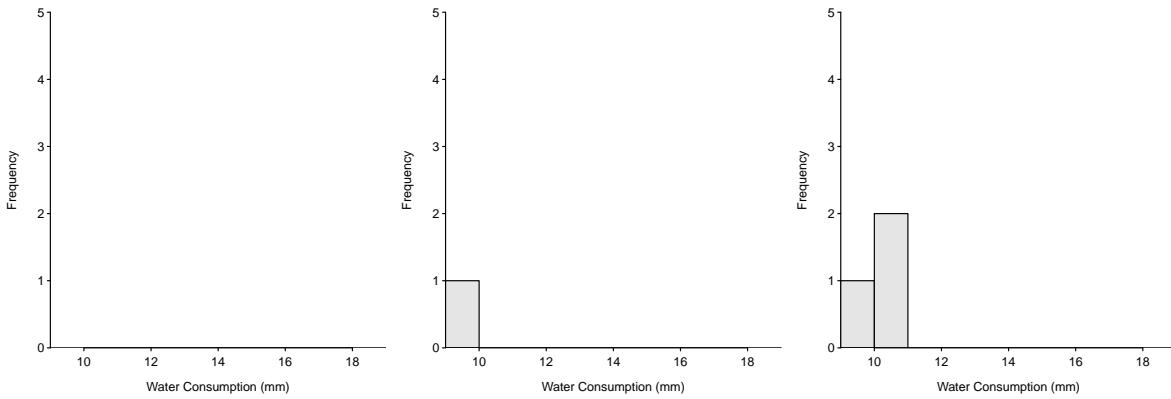


Figure 5.1. Steps illustrating the development of a histogram.

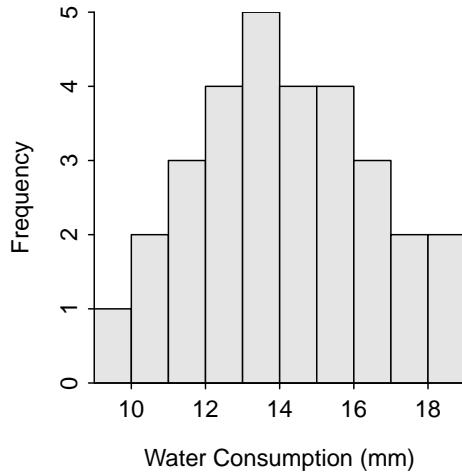


Figure 5.2. Histogram of water consumption (mm) by mice.

Ideally eight to ten classes (i.e., bars) are used to construct a histogram. Too many or too few bars make it difficult to identify the shape and may lead to different interpretations. A dramatic example of the effect of changing the number of classes is seen in histograms of the length of eruptions for the Old Faithful geyser (Figure 5.3).

5.2.2 Histograms in R

A simple (by default) histogram is constructed with `hist()` using a one-sided formula of the form `~quant`, where `quant` generically represents the quantitative variable, and the corresponding data frame in `data=`. The x-axis label may be improved from the default value by including a label in `xlab=`³. The width of the classes may be controlled by including a class width in `w=`.⁴

³

`xlab=` is for the “x-axis label.”

⁴

The endpoints for the classes may also be set by giving a vector of endpoints to `breaks=`.

Figure 5.3. Hisgoram of length (minutes) of eruptions for Old Faitful geyser with varying number of classes.

```
> hist(~days,data=LSI,xlab="Days of Ice Cover")      # Left
> hist(~days,data=LSI,xlab="Days of Ice Cover",w=20) # Right
```

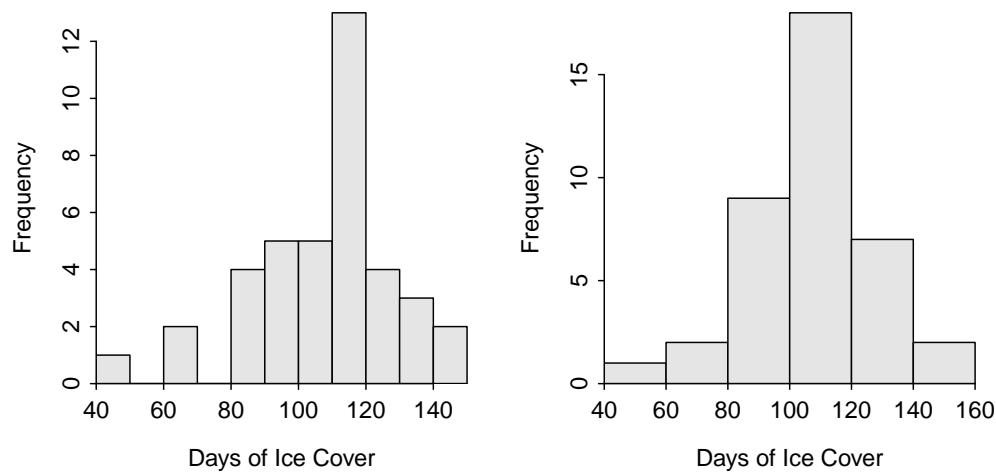


Figure 5.4. Histograms of the duration of ice cover at ice gauge 9004 in Lake Superior using the default class widths (Left) and widths of 20 days (Right).

- ◊ The default histogram should be modified by properly labeling the x-axis and possibly changing the class width.

Review Exercises

5.1 Histograms are constructed from what type of variables? [Answer](#)

5.2 What type of values are plotted on the x-axis of a histogram? [Answer](#)

5.3 What type of values are plotted on the y-axis of a histogram? [Answer](#)

5.4 What is the ideal number of bars on a histogram? [Answer](#)

5.5 The table below contains the concentrations (International Units per liter) of creatine phosphokinase (an enzyme related to muscle and brain functions) in 36 male volunteers. Construct a histogram from these data. [HINT: Enter these data into a CSV file and load into R as described in Section 4.3.2.] [Answer](#)

121 82 100 151 68 58 95 145 64 119 104 110 113 118 203 62 83 67
201 101 163 84 57 139 60 78 94 93 92 110 25 123 70 48 95 42

5.6 The table below contains the carbon monoxide levels (ppm) arising from one of the stacks for an oil refinery northeast of San Francisco between April 16 and May 16, 1993. The measurements were submitted as evidence for establishing a baseline to the Bay Area Air Quality Management District (BAAQMD).⁵ Construct a histogram from these data. [HINT: Enter these data into a CSV file and load into R as described in Section 4.3.2.] [Answer](#)

30 30 34 36 37 38 40 42 43 43 45 52 55 58 58 58
59 63 63 71 75 85 86 86 99 102 102 141 153 261 21

5.3 Interpreting Shape

A histogram has two tails – a left-tail for smaller or more negative values and a right-tail for larger or more positive values. The relative appearance of these two tails is used to identify three different shapes of distributions – symmetric, left-skewed, and right-skewed. If the left- and right-tail of a histogram are equal in shape (length and height), then the distribution is said to be **symmetric**. Perfectly symmetric distributions rarely occur in “real-life.” Therefore, if the left- and right-tail are approximately equal in shape, then the distribution is **approximately symmetric**. If the left-tail of the histogram is stretched out or, alternatively, the left-tail is longer and flatter than the right-tail, then the distribution is negatively- or **left-skewed**. If the right-tail of the histogram is stretched out or, alternatively, the right-tail is longer and flatter than the left-tail, then the distribution is positively- or **right-skewed**. The type of skew is defined by the longer tail; a longer right-tail means the distribution is right-skewed and a longer left-tail means it is left-skewed. Examples of each shape are shown in Figure 5.5.

⁵

BAAQMD personnel had also made nine independent measurements of the carbon monoxide from this same stack over the period from September 11, 1990, to March 30, 1993, (which are not shown).

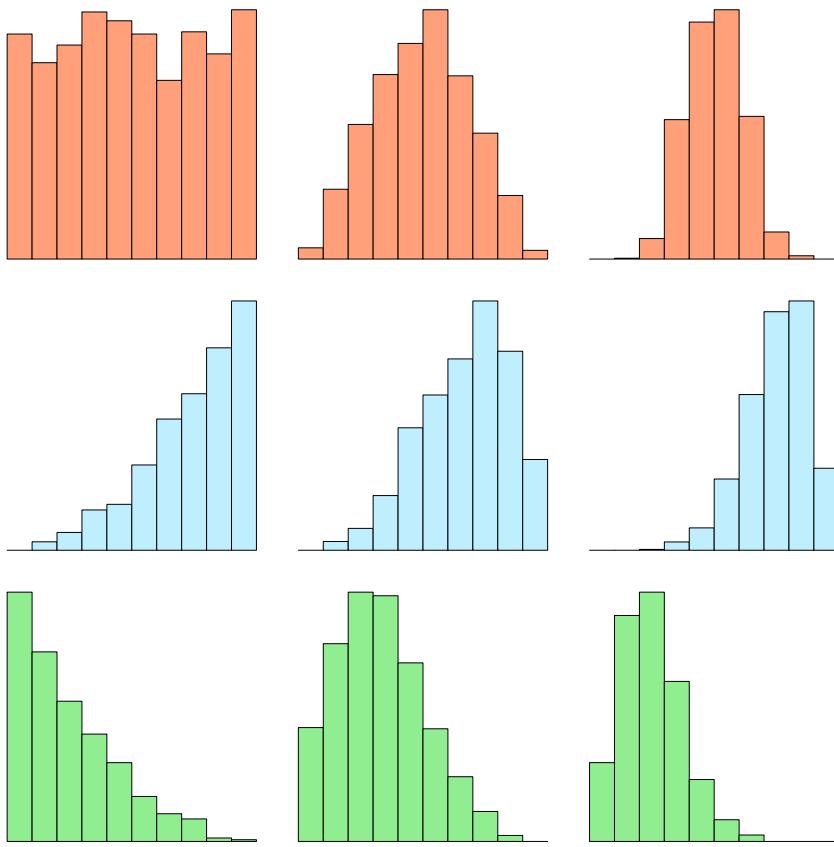


Figure 5.5. Examples of approximately symmetric (top, red), left-skewed (middle, blue), and right-skewed (bottom, green) histograms. Note that the axes labels were removed to focus attention on the shape of the histograms. Each histogram was constructed from $n=1000$ individuals and the x-axis range is from 0 to 1.

Δ Symmetric: The left- and right-tail of a distribution are nearly the same in length and height.

Δ Left-skewed: The left-tail of a distribution is longer or more drawn out than the right-tail.

Δ Right-skewed: The right-tail of a distribution is longer or more drawn out than the left-tail.

◊ The longer tail defines the type of a skewed distribution.

You will find in practice that these labels form a continuum. For example, a perfectly symmetric distribution is rare. However, in the many cases of an asymmetric distribution, it is a fine line between calling the shape approximately symmetric or one of the skewed distributions.

◊ Symmetric, left-skewed, and right-skewed descriptors are guides; many “real” distributions will not fall neatly into these categories.

5.4 Interpreting Outliers

An outlier is an individual in a sample whose value is widely separated from the main cluster of values in the sample. On histograms, outliers appear as bars that are separated from the main cluster of bars by “white space” or areas with no bars (Figure 5.6). In general, outliers must be on the margins of the histogram, should be separated by one or two missing bars, and should only be one or two individuals.

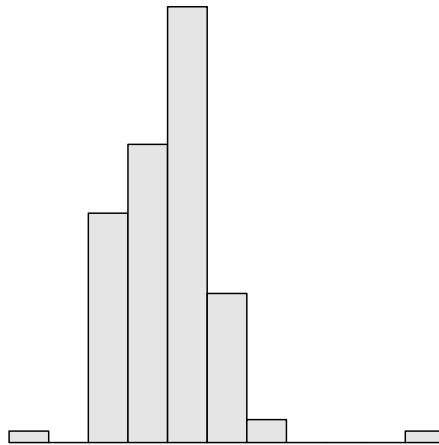


Figure 5.6. Example histogram with an outlier to the right.

Δ Outlier: An individual in a sample whose value is widely separated from the main cluster of values in the sample.

◊ In general, don't consider a group of more than two individuals as outliers even if they are separated from the main cluster of individuals.

Outliers are not necessarily evil. Sometimes outliers do occur as a result of human error in the sampling process. If this is the case, then the individual's value should be corrected or removed. Sometimes outliers are individuals that are not part of the population of interest – e.g., an adult animal that was sampled when only immature animals were being considered. In this case, the individual's value should be removed from the sample. Sometimes outliers are simply part of the population. In this case you should generally not remove the individual's value; in fact you may wish to highlight its existence as an interesting observation. In any case, it is important that you plot the distribution of your data with a histogram to determine if outliers are present or not.

Don't let outliers completely influence how you define the shape of a distribution. For example, if the main cluster of values is approximately symmetric and there is one outlier to the right of the main cluster (as illustrated in Figure 5.6), DON'T call the distribution right-skewed. You should describe this distribution as approximately symmetric with an outlier to the right.

◊ Not all outliers warrant removal from your sample.

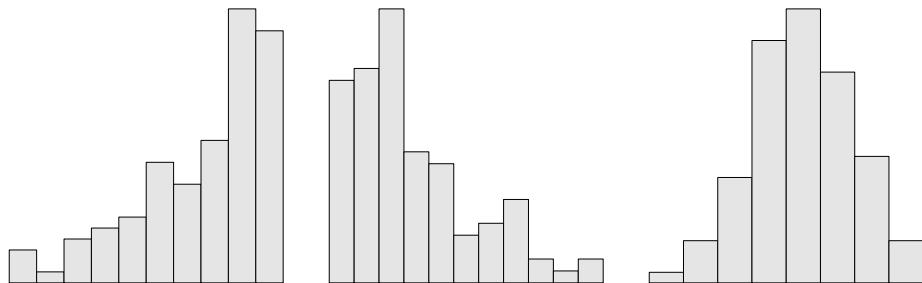
◊ Don't let outliers completely influence how you define the shape of a distribution.

Review Exercises

5.7 What is a distribution with a long left-tail called? [Answer](#)

5.8 What is a distribution with a long right-tail called? [Answer](#)

5.9 What is the shape of the distribution on the left below? [Answer](#)



5.10 What is the shape of the distribution in the center above? [Answer](#)

5.11 What is the shape of the distribution on the right above? [Answer](#)

5.12 Comment on the shape and presence of outliers in Figure 1.2. [Answer](#)

5.5 Measures of Center

There are three common methods to measure the center of a distribution: the mode, median, and mean. The median and mean are the most widely used methods. The choice of which method to use depends, in part, on the shape of the distribution, the presence of outliers, and your purpose.

The modes, medians, and means computed in this section are summary statistics – i.e., they are computations from individuals in a sample. Thus, they should specifically be called the sample mode, sample median, and sample mean. The mode, median, and mean can also be computed from every individual in the population, if it is known. The computed values would then be parameters and should be called the population

mode, population median, and population mean. See Section ?? for clarification on the differences between populations and samples and parameters and statistics.

◊ Three measures of the center of a distribution are the mode, median, and mean.

◊ Measures of center computed from individuals in a sample are preceded by the word “sample”; those computed from all individuals in a population are preceded by the word “population.”

5.5.1 Mode

The mode is the value that occurs most often in a data set. If the variable is continuous, then the modal class is the class of values that occurs most often in a data set. In other words, it is the class that forms the peak of a distribution. For example, in the mouse water consumption data (Figure 5.2) the modal class is the 13.0-13.9 class. Some data sets may have what looks like two “humps.” Each “hump” is considered a mode and the distribution is said to be **bimodal**.

Δ **Mode:** The value or class of values that occurs most often in a data set.

Δ **Bimodal:** The shape of a distribution with two peaks or “humps.”

5.5.2 Median

The median is the value of the individual in the position that splits the **ordered** list of individuals into two **equal-sized** halves. In other words, if the data are ordered, half the values will be smaller than the median and half will be larger.

The process for finding the median consists of three steps.⁶

1. Order the data from smallest to largest.
2. Find the “middle **position**” (mp) with $mp = \frac{n+1}{2}$.
3. If mp is an integer (i.e., no decimal), then the median is the value of the individual in that position. If mp is not an integer, then the median is the average of the value immediately below and the value immediately above the mp .

As an example, the ordered mouse water consumption data from Table 5.1 are,

| | | | | | | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 9.4 | 10.0 | 10.6 | 11.1 | 11.4 | 11.5 | 12.0 | 12.5 | 12.6 | 12.9 | 13.0 | 13.5 | 13.6 | 13.7 | 13.7 |
| 14.0 | 14.1 | 14.3 | 14.7 | 15.2 | 15.4 | 15.8 | 15.8 | 16.5 | 16.6 | 16.6 | 17.0 | 17.4 | 18.2 | 18.4 |

Because $n = 30$, the $mp = \frac{30+1}{2} = 15.5$. The mp is not an integer so the median is the average of the values

⁶

Most computer programs use a more sophisticated algorithm for computing the median and, thus, will produce different results than what will result from applying these steps.

in the 15th and 16th ordered positions (i.e., the two actual positions closest to mp). Thus, the median water consumption in this sample of mice is $\frac{13.7+14.0}{2} = 13.85$ mm.

As another example, consider finding the median of the Richter Scale magnitude recorded for fifteen major earthquakes (Table 5.2; note that these data were already ordered). Because $n = 15$, the $mp = \frac{15+1}{2} = 8$. The mp is an integer (i.e., no decimals) so the median is the value of the individual that is in the 8th ordered position, which is 7.1.

Δ Median: The midpoint of the data, i.e., the value of the individual in the position that splits the ordered list of individuals into two equal-sized halves.

5.5.3 Mean

The mean is the arithmetic average of the data. The sample mean is denoted by \bar{x} and the population mean is denoted by μ . If the measurement of the generic variable x on the i th individual is denoted as x_i , then the sample mean is computed with these two steps,

1. Sum (i.e., add together) all of the values – $\sum_{i=1}^n x_i$.
2. Divide by the number of individuals in the sample – n .

or more succinctly summarized with this equation,

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (5.5.1)$$

For example, the sample mean of the mouse consumption data is computed as follows:

$$\bar{x} = \frac{9.4 + 10.0 + 10.6 + 11.1 + 11.4 + 11.5 + \dots + 16.6 + 16.6 + 17.0 + 17.4 + 18.2}{30} = \frac{421.2}{30} = 14.04$$

Δ Mean: The center of gravity or balance point of the data, i.e., the sum of the data divided by the number of individuals

5.5.4 Measures of Center in R

The mean and median (along with other measures) are calculated in R with `Summarize()` using a one-side formula of the form `~quant`, where `quant` generically represents the quantitative variable, and the `data=` argument. The number of digits after the decimal place may be controlled with `digits=`.

```
> Summarize(~days,data=LSI,digits=2)
      n    valid   mean      sd     min      Q1    median      Q3      max
  42.00  39.00 107.85  21.59  48.00  97.00 114.00 118.00 146.00
```

From this it is seen that the sample mean is 107.85 days and the sample median is 114.00 days.

Review Exercises

- 5.13**  The following values are the maximum gauge heights of the Bois Brule River in Brule, WI from 10-25Feb05.⁷ Compute the mean and median of these data both “by hand” and with R. [HINT: Enter these data into a CSV file and load into R as described in Section 4.3.2.] [Answer](#)

1.56 1.54 1.54 1.57 1.58 1.61 1.60 1.69
1.99 2.11 1.98 1.76 1.69 1.99 1.86 1.53

- 5.14**  The following values are the population density (number of people per acre of land) for 15 randomly selected Wisconsin counties.⁸ Compute the mean and median of these data both “by hand” and with R. [HINT: Enter these data into a CSV file and load into R as described in Section 4.3.2.] [Answer](#)

429.0 67.8 52.1 97.4 57.9 354.9 16.2 19.1
127.0 27.6 10.2 54.6 28.8 30.1 20.2

- 5.15**  Compute the mean and median of the creatine phosphate data in Exercise 5.5. [Answer](#)

- 5.16**  Compute the mean and median of the carbon monoxide data in Exercise 5.6. [Answer](#)

5.5.5 Comparing the Median and Mean

The mean and median measure center in different ways. The median is concerned with the **position** of the value rather than the value itself (recall how it is calculated). The mean, on the other hand, is the value such that the sum of the distances from it to all points smaller than it is the same as the sum of the distances from it to all points greater than it. The mean is very much concerned about the **values** for each individual, as the values are used to find the “distance” from the mean.

- ◊ The actual values of the data (beyond ordering the data) are not considered when calculating the median; whereas the actual values are very much considered when calculating the mean.

A plot of the Richter scale data against the corresponding ordered individual number is shown in Figure 5.7-Left.⁹ The median (blue line) is found by locating the middle position on the individual number axis and

⁷

Data collected from [USGS](#).

⁸

Data collected from [U.S. census](#).

⁹

This is a rather non-standard graph but it is useful for comparing how the mean and median measure the center of the data.

then finding the corresponding Richter scale value (move right until the point is intercepted and then move down to the x-axis). The vertical blue line represents the median, and it can be seen that it has the same **number** of individuals (i.e., points) below it as above it. In contrast, the mean finds the Richter scale value that has the same total distance to values below it as total distance to values above it. In other words, the mean is the vertical red line so that the total **length** of the horizontal dashed red lines is the same to the left as it is to the right. Thus, the median balances the number of individuals above and below the median, whereas the mean balances the difference in values above and below the mean.

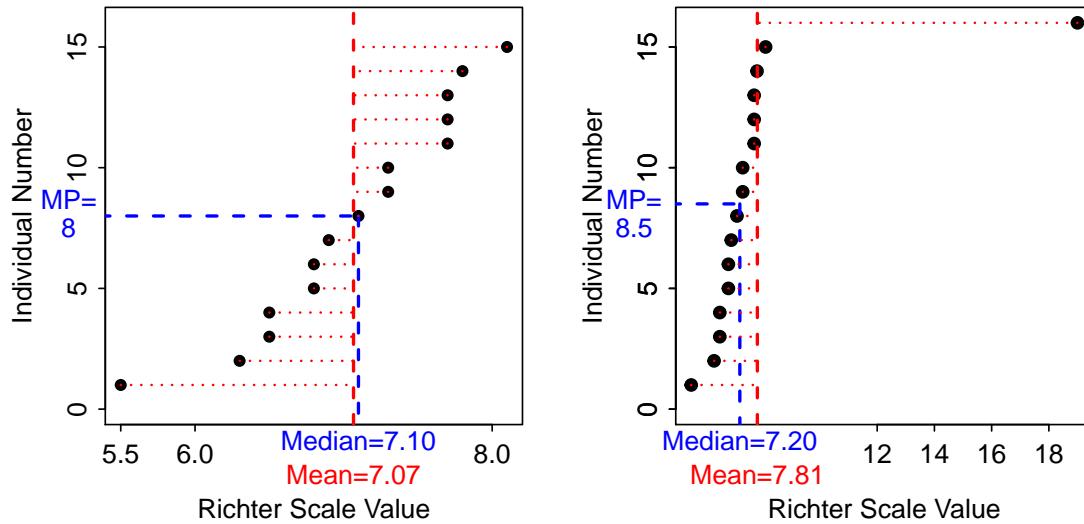


Figure 5.7. Plot of the individual number versus Richter scale values for the original earthquake data (**Left**) and the earthquake data with an extreme outlier (**Right**). The median value is shown as a blue vertical line and the mean value is shown as a red vertical line. Differences between each individual value and the mean value are shown with horizontal red lines.

- ◊ The mean balances the distance to individuals above and below the mean. The median balances the number of individuals above and below the median.
- ◊ The sum of all differences between individual values and the mean (as properly calculated) equals zero.

The mean and median differ in their sensitivity to outliers (Figure 5.7-Right). For example, suppose that an incredible earthquake with a Richter Scale value of 19.0 was added to the earthquake data set. With this additional individual, the median increases from 7.1 to 7.2, but the mean increases from 7.1 to 7.8. The outlier affects the value of the mean more than it affects the value of the median because of the way that each statistic measures center. The mean will be pulled towards an outlier because it must “put” many values on the “side” of the mean away from the outlier so that the sum of the differences to the larger values and the sum of the differences to the smaller values will be equal. Thus, the outlier in this example creates a large difference to the right of the mean so the mean has to “move” to the right to make this difference smaller, move more individuals to the left side of the mean, and increase the differences of individuals to the left of the mean to balance this one large individual. The median on the other hand will simply “put” one more individual on the side opposite of the outlier because it balances the number of individuals on each side of

it. Thus, the median has to move very little to the right to accomplish this balance.

- ◊ The mean is more sensitive (i.e., changes more) to outliers than the median; it will be “pulled” towards the outlier more than the median.

The shape of the distribution, even if outliers are not present, also has an effect on the values of the mean and median as depicted in Figure 5.8. If a distribution is perfectly symmetric, then the median and mean (along with the mode) will be identical. If the distribution is approximately symmetric, then the median and mean will be approximately equal. If the distribution is right-skewed, then the mean will be greater than the median. Finally, if the distribution is left-skewed, then the mean will be less than the median.

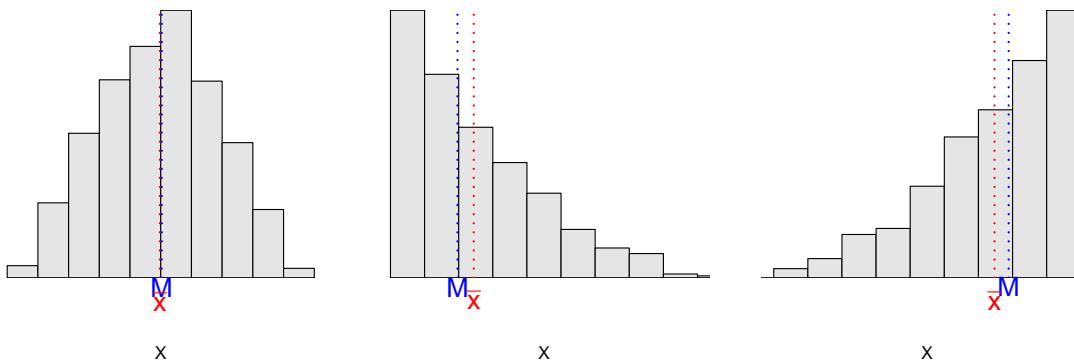


Figure 5.8. Three differently shaped histograms with vertical lines superimposed at the median (M ; blue lines) and the mean (\bar{x} ; red lines).

- ◊ The mean and median are equal for symmetric distributions.

- ◊ The mean is pulled towards the long tail of a skewed distribution. Thus, the mean is greater than the median for right-skewed distributions and the mean is less than the median for left-skewed distributions.

As shown above, the mean and median measure center in different ways. The question now becomes “which measure of center is better?” The median is a “better” measure of center when outliers are present. In addition, the median gives a better measure of a typical individual when the data are skewed. Thus, in this course use the median when outliers are present or the distribution of the data is skewed. If the distribution is symmetric, then the purpose of the analysis will dictate which measure of center is “better.” However, in this course, use the mean when the data are symmetric or, at least, not strongly skewed.

- ◊ Describe center with the median if outliers are present or the data are skewed; use the mean if the data are symmetric and no outliers are present.

Review Exercises

5.17 Is the mean less than, approximately equal to, or greater than the median for the distribution shown in Exercise 5.9? [Answer](#)

5.18 Is the mean less than, approximately equal to, or greater than the median for the distribution shown in Exercise 5.10? [Answer](#)

5.19 Is the mean less than, approximately equal to, or greater than the median for the distribution shown in Exercise 5.11? [Answer](#)

5.20 Is the mean divided by the median less than 1, equal to 1, or greater than 1 for a symmetric distribution? [Answer](#)

5.21 From your calculation of the mean and median in Review Exercise 5.13 do you expect the histogram to be left-skewed, approximately symmetric, or right-skewed? [Answer](#)

5.22 From your calculation of the mean and median in Review Exercise 5.14 do you expect the histogram to be left-skewed, approximately symmetric, or right-skewed? [Answer](#)

5.6 Measures of Dispersion

There are three common measures of the dispersion of a distribution: the range, inter-quartile range (IQR), and standard deviation. The standard deviation is the most widely used method. The choice of which method to use depends, however, on what statistic you chose as the measure of center (which, as described in Section 5.5.5, depends on the shape of the distribution, presence of outliers, and your purpose).

The range, IQR, and standard deviation computed in this section are summary statistics – i.e., they are computations from individuals in a sample. Thus, they should all be preceded with the word “sample.” See Section ?? for clarification on the differences between populations and samples and parameters and statistics.

◊ Three measures of the dispersion of a distribution are the range, inter-quartile range (IQR), and standard deviation.

◊ Measures of dispersion computed from individuals in a sample are preceded by the word “sample”; those computed from all individuals in a population are preceded by the word “population.”

5.6.1 Range

The range is the difference between the maximum and minimum values in the data. It is a measure of the ultimate dispersion or spread of the data. The range in the mouse consumption data (Table 5.1) is $18.4 - 9.4 = 9.0$.

The range should never be used by itself as a measure of dispersion. The range is extremely sensitive to outliers and is best used only to show all possible values present in the data. The range (as strictly defined) also suffers from a lack of information. For example, what does a range of 9 mean? It can have a completely different interpretation if it came from the values of 1 to 10 or if it came from the values of 1000 to 1009. Thus, the range is more instructive if presented as both the maximum and minimum value rather than just the range.

Δ Range: The difference between the maximum and minimum value in a data set.

◊ Never use the range by itself as a measure of dispersion.

5.6.2 IQR

Quartiles are the values for the three individuals that divide ordered data into four (approximately) equal parts. Finding the three quartiles consists of finding the median, splitting the data into two equal parts at the median, and then finding the medians of the two halves.¹⁰ A concern in this process is that the median is NOT part of either half if there is an odd number of individuals. These steps are summarized as,

1. Order the data from smallest to largest.
2. Find the median – this is the second quartile (Q2).
3. Split the data into two halves at the median. If n is odd (so that the median is one of the observed values), then the median is not part of either half.¹¹
4. Find the median of the lower half of data – this is the 1st quartile (Q1).
5. Find the median of the upper half of data – this is the third quartile (Q3).

These calculations are illustrated with the earthquake data (Table 5.2). Recall from above (Section 5.5.2) that the median ($=7.1$) is in the eighth position of the ordered data. The value in the eighth position will not be included in either half. Thus, the two halves of the data are 5.5 6.3 6.5 6.5 6.8 6.8 6.9 and 7.3 7.3 7.7 7.7 7.7 8.1. Each half contains seven individuals so the middle position for each half is $mp = \frac{7+1}{2} = 4$. Thus, the median for each half will be the individual in the fourth position. Therefore, the median of the first half is $Q1 = 6.5$ and the median of the second half is $Q3 = 7.7$.

As another example, consider the quartiles of the mouse consumption data (the median was computed in Section 5.5.2). Because $n = 30$ is even, the halves of the data split naturally with 15 individuals in each half. Therefore, the $mp = \frac{15+1}{2} = 8$ and the median of each half is the value of the individual in the eighth position. Thus, $Q1 = 12.5$ and $Q3 = 15.8$. In summary, the first, second, and third quartiles for

¹⁰

You should review how a median is computed before proceeding with this section.

¹¹

Some authors will put the median into both halves when n is odd. The difference between the two methods is minimal for large n .

the mouse water consumption data are 12.5, 13.85, and 15.8, respectively. These three values separate the ordered individuals into approximately four equally-sized groups – those with values less than 12.5, with values between 12.5 and 13.85, with values between 13.85 and 15.8, and with values greater than 15.8.

△ Quartiles: The values that divide the ordered data into quarters.

The interquartile range is the difference between the third quartile (Q3) and the first quartile (Q1), namely Q3-Q1. The IQR for the mouse consumption data is, thus, $15.8 - 12.5 = 3.3$. Intuitively, the IQR can be thought of as the “range of the middle half of the data.” The IQR is favored over the range because it is not sensitive to outliers (*you should convince yourself that this is true*). As with the range, however, the IQR suffers from a lack of information. Thus, you should always present the IQR by presenting both Q1 and Q3 rather than the difference between the two. Finally, the IQR should be chosen as the measure of dispersion when the median is used as the measure of center because they are conceptually related (both rely on position rather than actual value). Thus, the IQR is used if outliers are present or the data are skewed.

△ Inter-Quartile Range (IQR): The difference between the third (Q3) and first (Q1) quartiles.

◊ The IQR should be used as the measure of dispersion only if the median is chosen as the measure of center.

5.6.3 Standard Deviation

The sample standard deviation, denoted by s , can be thought of as “the average difference between the observed values and the mean.”¹² The standard deviation is computed with these six steps:

1. Compute the sample mean (i.e., \bar{x}).
2. For each value (x_i), find the difference between the value and the mean, namely $x_i - \bar{x}$.
3. Square each difference, namely $(x_i - \bar{x})^2$.
4. Add together all the squared differences.
5. Divide this sum by $n - 1$. [*Stopping here gives the sample variance, s^2 .*]
6. Square root the result from the previous step to get s .

These steps are neatly summarized with

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (5.6.1)$$

The calculation of the standard deviation of the earthquake data (Table 5.2) is facilitated with the calculations shown in Table 5.4. In Table 5.4, note that \bar{x} is equal to the sum of the “Value” column divided by $n = 15$ (i.e., $\bar{x} = 7.07$). Step 2 of the calculation of s appears in the column labeled as “Diff” as this column contains

¹²

This statement is not strictly correct as will likely become obvious. However, this is an acceptable general interpretation of the standard deviation.

each observed value minus the calculated \bar{x} . Step 3 of the calculation of s appears in the “Diff²” column as this column contains the square of the previously calculated differences. Step 4 is the sum of the “Diff²” column. The sample variance, i.e., the result of Step 5 in the calculation, is equal to this sum divided by $n - 1 = 14$ or $\frac{6.773}{14} = 0.484$. Finally, the sample standard deviation is the square root of the sample variance or $s = \sqrt{0.484} = 0.696$. Thus, on average, each earthquake is approximately 0.7 units on the Richter scale different than the average earthquake in this data set.

Table 5.4. Table showing an efficient calculation of the standard deviation of the earthquake data.

| Indiv i | Value x_i | Diff $x_i - \bar{x}$ | Diff ² $(x_i - \bar{x})^2$ |
|------------|----------------|-------------------------|--|
| 1 | 5.5 | -1.57 | 2.454 |
| 2 | 6.3 | -0.77 | 0.588 |
| 3 | 6.5 | -0.57 | 0.321 |
| 4 | 6.5 | -0.57 | 0.321 |
| 5 | 6.8 | -0.27 | 0.071 |
| 6 | 6.8 | -0.27 | 0.071 |
| 7 | 6.9 | -0.17 | 0.028 |
| 8 | 7.1 | 0.03 | 0.001 |
| 9 | 7.3 | 0.23 | 0.054 |
| 10 | 7.3 | 0.23 | 0.054 |
| 11 | 7.7 | 0.63 | 0.401 |
| 12 | 7.7 | 0.63 | 0.401 |
| 13 | 7.7 | 0.63 | 0.401 |
| 14 | 7.8 | 0.73 | 0.538 |
| 15 | 8.1 | 1.03 | 1.068 |
| Sum | 106 | 0 | 6.773 |

△ **Standard Deviation:** “Essentially” the average deviation or difference of each individual from the mean.

◊ In the standard deviation calculations don’t forget to take the square root of the variance.

There are three characteristics of the standard deviation that you should be aware of:

1. $s \geq 0$ ($s=0$ only if there is no dispersion; i.e., all values are the same).
2. s is strongly influenced by outliers.
3. s is inflated for skewed distributions (similar to the mean).

The final two characteristics are a result of the standard deviation being computed from the **values**, rather than the position, of the individuals (as is the mean). The argument here is the same as it was for the mean. In fact, it should be obvious that the mean and standard deviation are conceptually linked (i.e., they both require the actual values and the mean is within the standard deviation calculation).

◊ The standard deviation should be used as the measure of dispersion only if the mean is chosen as the measure of center.

At the beginning of this section, the standard deviation was defined as “essentially the average difference between the values and the mean.” **Essentially** was emphasized because the formula for the standard

deviation does not simply add together the differences and divide by n as this definition would imply. Notice in Table 5.4 that the sum of the differences from the mean is 0. This will be the case for all standard deviation calculations using the correct mean, because the mean balances the distance to individuals below the mean with the distance of individuals above the mean (review Section 5.5). Thus, the mean difference will always be zero. This “problem” was corrected by squaring the differences before summing them. To get back to the original units, the squaring is later “reversed” by the square root. So, more accurately, the standard deviation is the square root of the average squared difference between the values and the mean. Therefore, the original definition of the standard deviation is strictly incorrect; however, it works well as a practical definition of the meaning of the standard deviation.

- ◊ Use the fact that the sum of all differences from the mean equals zero as a check of your standard deviation calculation.

Furthermore, you should note that the mean is the value that will minimize the value of the standard deviation – i.e., putting any other value besides the mean into the standard deviation equation will result in a larger standard deviation value.

Finally, why is the sum of the squared differences divided by $n - 1$, rather than n , in the standard deviation calculation? Recall (from Section ??) that statistics are meant to estimate parameters. The sample standard deviation is supposed to estimate the population standard deviation (σ). Theorists have shown that if we divide by n , s will consistently underestimate σ . Thus, s calculated in this way would be a biased estimator of σ . Theorists have found, though, that dividing by $n - 1$ will cause s to be an unbiased estimator of σ . Being unbiased is generally good – it means that on average our statistic estimates our parameter (this concept is discussed in more detail in Chapter 12).

5.6.4 Measures of Dispersion in R

The minimum, maximum, Q1, Q3, and standard deviation are calculated with `Summarize()` as described previously for the mean and median.

```
> Summarize(~days,data=LSI,digits=2)
  n  nvalid    mean      sd      min      Q1   median      Q3      max
  42.00  39.00 107.85  21.59  48.00  97.00 114.00 118.00 146.00
```

Thus, $s = 21.59$, the IQR is from $Q1 = 97.00$ to $Q3 = 118.00$, and the range is from 48.00 to 146.00.

Review Exercises

- 5.23** Compute the range, IQR, and standard deviation for the maximum gauge heights of the Bois Brule River in Brule, WI from Exercise 5.13 both “by hand” and with R. [Answer](#)
- 5.24** Compute the range, IQR, and standard deviation for the population density of Wisconsin counties from Exercise 5.14 both “by hand” and with R. [Answer](#)
- 5.25** Compute the range, IQR, and standard deviation of the creatine phosphate data in Exercise 5.5. [Answer](#)
- 5.26** Compute the the range, IQR, and standard deviation of the carbon monoxide data in Exercise 5.6. [Answer](#)

5.7 Overall Summaries

The relationships among measures of center and dispersion must be considered in order to create an overall numerical summary of the data. From the previous section it should be obvious that the standard deviation and mean are conceptually linked as are the median and IQR. Indeed, the linked measure of center must be computed first in both dispersion measure calculations. Thus, if it is decided that the mean will be used to measure center, then the standard deviation must be used to measure dispersion. Similarly, if the median is used to measure center, then the IQR must be used to measure dispersion.¹³

5.7.1 Boxplots

The median, range, and IQR form the **five-number summary**. Specifically, the five-number summary consists of the minimum value, Q1, median, Q3, and maximum value. The five-number summary for the mouse consumption data is 48.0, 97.0, 114.0, 118.0, and 146.0 (all values computed in the previous section).

The five-number summary is typically displayed as a graph called a box-and-whisker or **boxplot**. A traditional boxplot consists of a horizontal line at the median, horizontal lines at Q1 and Q3 that are connected with vertical lines to form a box, and then vertical lines (or whiskers) from Q1 to the minimum value and from Q3 to the maximum value. Modern boxplots have been modified to allow easier detection of outliers. Many of these modern boxplots construct the “box” in the same manner as the traditional boxplot. However, the upper whisker extends from Q3 to the last observed value that is within 1.5 IQRs of Q3 and the lower whisker extends from Q1 to the last observed value that is within 1.5 IQRs of Q1. Observed values outside of the whiskers are termed “outliers” by this algorithm and are typically plotted with circles or asterisks. If no individuals are deemed “outliers” by this algorithm, then the two forms of boxplots will be the same. A boxplot of the mouse consumption data is shown in Figure 5.9.

△ **Boxplot:** Generally, a graphical depiction of the five-number summary.

The relative length from the median to Q1 and the median to Q3 (i.e., the position of the median line in the box) can give an indication of the shape of a distribution. If the distribution is left-skewed (i.e., the left tail or lesser-valued individuals is spread out; Figure 5.10-Right), then the “dispersion” of the individuals in the second quarter of the ordered data (i.e., Q1 - median) will be greater than the “dispersion” of the individuals in the third quarter (i.e., median - Q3). In contrast, if the distribution is right-skewed (i.e., the right tail or larger-valued individuals is spread out; Figure 5.10-Middle), then the “dispersion” of the individuals in the third quarter of the ordered data will be greater than the “dispersion” of the individuals in the second quarter. Thus, if the distribution is right-skewed, then the median will be closer to Q1 than to Q3. If the distribution is left-skewed, then the median will be closer to Q3 than to Q1. If the distribution is approximately symmetric (Figure 5.10-Left), then the median will be in the middle of the box (and the middle of the whiskers).

¹³

Recall that the range will never be used by itself.

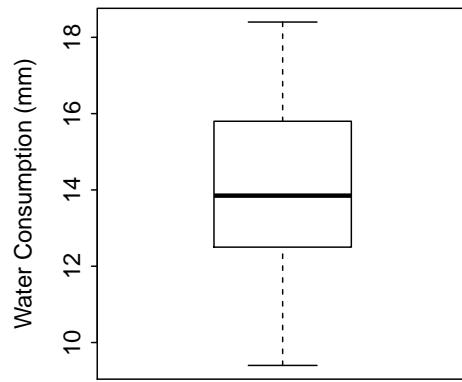


Figure 5.9. Boxplot of the mouse consumption data.

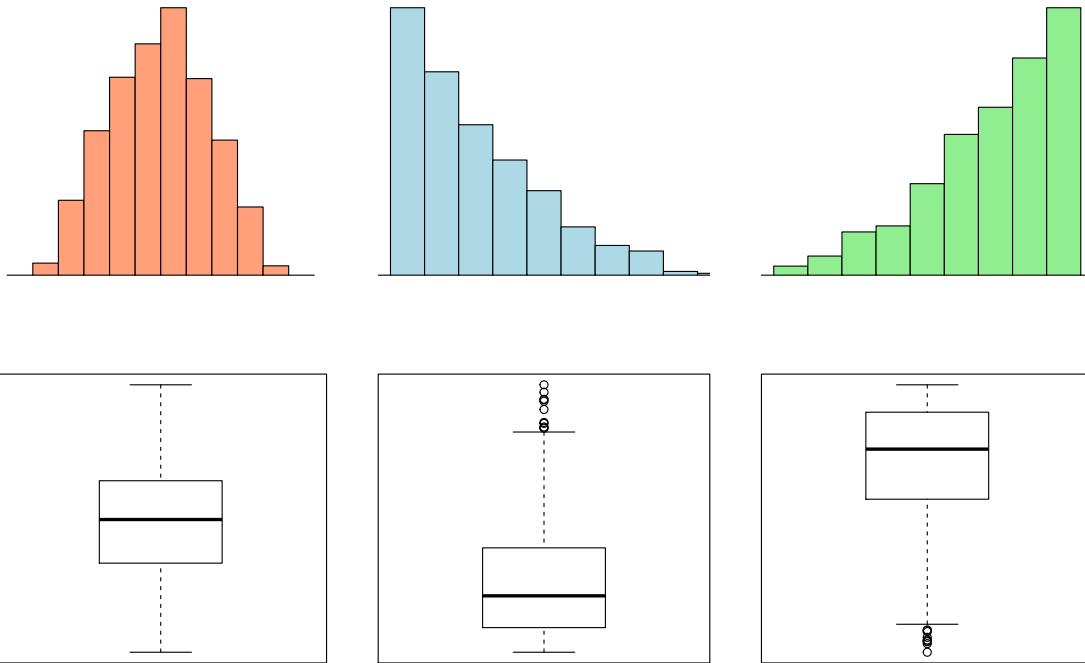


Figure 5.10. Histograms and boxplots for several different shapes of distributions.

- ◊ If a distribution is right-skewed, then the median will be closer to Q₁ than to Q₃. If the distribution is left-skewed, then the median will be closer to Q₃ than to Q₁.

- ◊ Even though shape can be described from a boxplot, it is always easier to describe shape from a histogram.

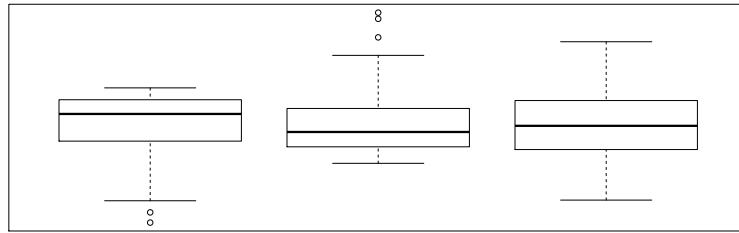
A boxplot is constructed in R with `boxplot()`. This function requires only the name of the quantitative variable as the first argument although the x- and y-axes are labeled with `xlab=` and `ylab=`, respectively.

Review Exercises

5.27 What is the five-number summary for the maximum gauge heights of the Bois Brule River in Brule, WI from Exercise 5.13. [Answer](#)

5.28  Construct a boxplot for the population density of Wisconsin counties from Exercise 5.14. [Answer](#)

5.29 What is the shape of the left boxplot below? [Answer](#)



5.30 What is the shape of the middle boxplot above? [Answer](#)

5.31 What is the shape of the right boxplot above? [Answer](#)

5.32 If the distribution is skewed left, which measure should you generally use to measure center? [Answer](#)

5.33 Which measure of center should you generally use for a right-skewed distribution? [Answer](#)

5.34 Which measure of center should you generally use for a symmetric distribution? [Answer](#)

5.35 Which measure of dispersion should you generally use for a symmetric distribution? [Answer](#)

5.36 Which measure of dispersion should you generally use for a left-skewed distribution? [Answer](#)

5.37 Which measure of dispersion should you generally use for a right-skewed distribution? [Answer](#)

5.38 Is Q3-Q2 less than, approximately equal to, or greater than Q2-Q1 if the data are left-skewed? [Answer](#)

5.39 What is the shape of the distribution if Q3-Q2 is greater than Q2-Q1? [Answer](#)

5.8 Multiple Groups

It is common to conduct a univariate EDA for a quantitative variable separately for multiple groups of individuals. In these cases it is beneficial to have a function that will efficiently construct a histogram and compute summary statistics for the quantitative variable separated by the levels of a factor variable. Separate histograms are constructed with `hist()` if the first argument is a “formula” of the type `quant~group` where `quant` represents the quantitative response variable of interest and `group` represents the factor variable that indicates to which group the individual belongs. The data frame that contains `quant` and `groups` is given to `data=`. Summary statistics are separated by group by supplying the same formula and `data=` arguments to `Summarize()`.

As an example, suppose that you want to examine the average annual days of ice for each decade (using the LSI data). One might expect to use the `days~decade` formula except that the `decade` variable is not a factor.¹⁴ A new variable, called `fdecade`, that is a factored version of `decade` is created by including `decade` in `factor()` as follows,

```
> LSI$fdecade <- factor(LSI$decade)
> str(LSI)
'data.frame': 42 obs. of 5 variables:
 $ season : int 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 ...
 $ decade : int 1950 1950 1950 1950 1950 1960 1960 1960 1960 1960 ...
 $ temp   : num 22.9 23 25.7 20 24.8 ...
 $ days    : int 87 137 106 97 105 118 118 136 91 NA ...
 $ fdecade: Factor w/ 5 levels "1950","1960",...
```

Histograms (Figure 5.11) and summary statistics separated by decade are then constructed with

```
> hist(days~fdecade,data=LSI,ylab="Days of Ice Cover",w=20)
> Summarize(days~fdecade,data=LSI,digits=2)
  fdecade n  nvalid   mean     sd min   Q1 median     Q3 max
1    1950  5      5 106.40 18.73  87  97.00 105.0 106.0 137
2    1960 10     10 113.12 14.80  91 104.20 116.0 119.8 136
3    1970 10     10 115.50 19.19  82 105.80 115.0 124.0 146
4    1980 10     10 103.80 24.88  48  90.25 116.0 118.0 123
5    1990  7      6  96.00 28.53  62  72.00 100.5 114.0 132
```

The desired grouping variable may already be a factor in some data frames. It was not in this case because the values in the `decade` variable were numerical. Thus, most grouping variables will not require modification with `factor()`. Furthermore, note that any variable modification should be stored into a new variable (e.g., `fdecade`) so as not to alter the data in the original variable.

Side-by-side boxplots (Figure 5.12) are an alternative to separated histograms and are constructed by including the same formula and `data=` arguments to `boxplot()`,

```
> boxplot(days~fdecade,data=LSI,ylab="Days of Ice Cover",xlab="Decade")
```

¹⁴

It is not a factor because the data in that variable looks numeric to R.

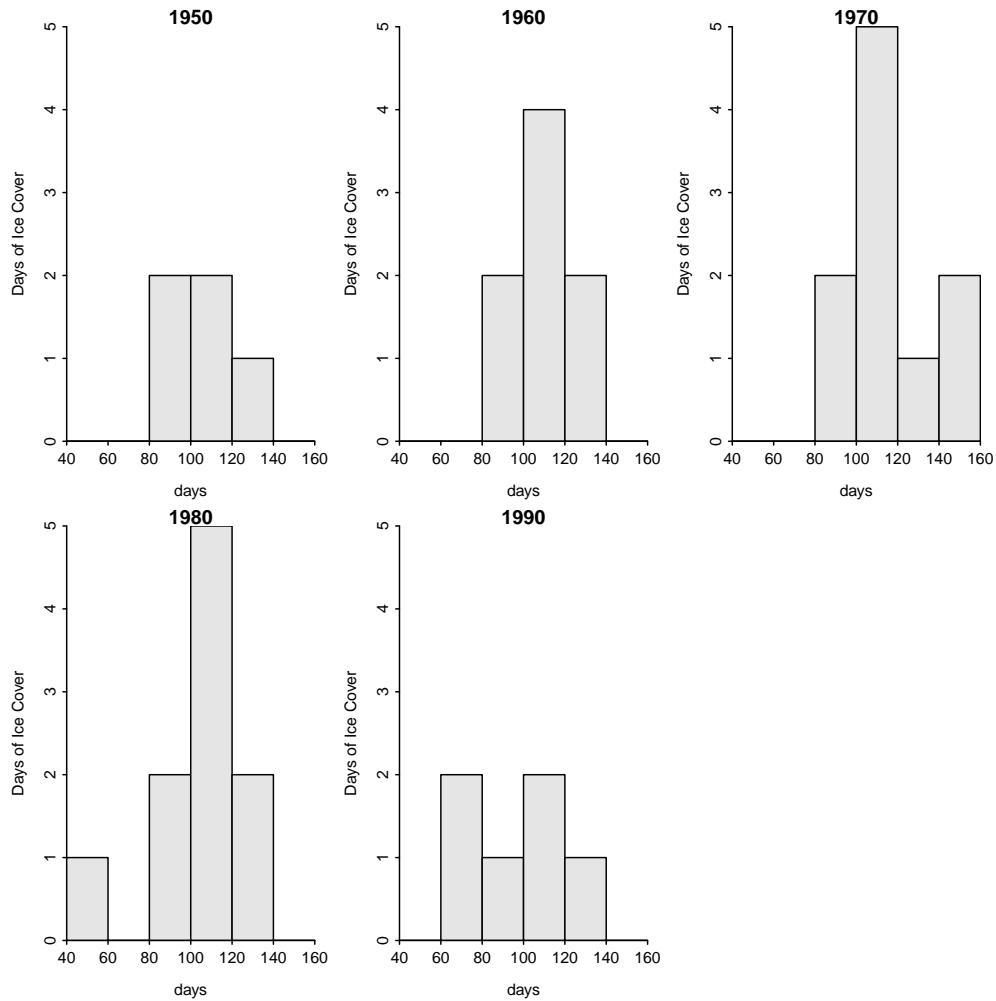


Figure 5.11. Histograms of the duration of ice cover at ice gauge 9004 in Lake Superior by each decade.

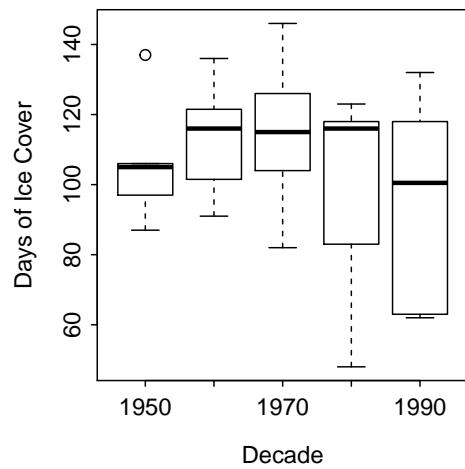


Figure 5.12. Boxplot of the duration of ice cover at ice gauge 9004 in Lake Superior by each decade.

Review Exercises

5.40

 Arsenic concentrations were measured in the well water and in the toe nails of 21 people with home wells. Also recorded were the person's age, sex, and qualitative measurements of usage for drinking and cooking. The data are found in [Arsenic.csv](#). Load these data into R to answer the questions below.

[Answer](#)

- (a) Construct a univariate EDA for the well water measurements.
 - (b) Construct a univariate EDA for the measurements of arsenic in the toe nails.
 - (c) Construct a univariate EDA for the toe nail arsenic levels separated by levels of drinking water usage.
-

5.9 Example Interpretations

While most of the previous sections focused on how to construct various graphs and numerical summaries, the most important aspect of this chapter is that you can make appropriate interpretations for an EDA from the summary results. For quantitative data, an appropriate EDA consists of identifying the shape, center, dispersion, and outliers for the variable. For categorical data, an appropriate EDA consists of identifying the major characteristics among the categories. Below, I model properly constructed EDAs for the mouse consumption data and two new data sets.

Mouse Consumption Example

Construct a proper EDA for the following situation and data – ‘The following measurements (Table 5.1) are of the consumption of water in one hour by mice in a laboratory setting.’

Mouse water consumption appears to be approximately symmetric without any outliers present (Figure 5.2). The center of the distribution is best measured by the mean, which is 14.05 ml (Table 5.5). The range of water consumption by the mice in the sample is from 9.4 to 18.4 ml while the dispersion as measured by the standard deviation is 2.41 ml (Table 5.5). I chose to use the mean and standard deviation because the data were symmetric with no outliers. [NOTE: 1) the use of units, 2) the reference to the figure and table, 3) the labeling of the figure and table, 4) I did not present or discuss the median and IQR because I chose to use the mean and standard deviation, 5) I did not use the range alone as a measure of dispersion, 6) I explained why I used the mean and standard deviation rather than the median and IQR, and 7) I provided the R code used.]

Table 5.5. Descriptive statistics of mouse water consumption.

| n | mean | sd | min | Q1 | median | Q3 | max |
|-------|-------|------|------|-------|--------|-------|-------|
| 30.00 | 14.05 | 2.41 | 9.40 | 12.52 | 13.85 | 15.80 | 18.40 |

R commands:

```
> setwd("c:/data/")
> mc <- read.csv("MouseData.csv")
> str(mc)
> Summarize(~consump,data=mc,digits=2)
> hist(~consump,data=mc,xlab="Water Consumption (mm)")
```

Crayfish Temperature Selection

Peck (1985) examined the temperature selection of dominant and subdominant crayfish (*Orconectes virilis*) together in an artificial stream. The temperature ($^{\circ}\text{C}$) selection by the dominant crayfish in the presence of subdominant crayfish in these experiments was recorded below. Thoroughly describe all aspects of the distribution of selected temperatures, using appropriate numerical summaries where needed.

| | | | | | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 30 | 26 | 26 | 26 | 25 | 25 | 25 | 25 | 25 | 24 | 24 | 24 | 24 | 24 | 24 | 23 |
| 23 | 23 | 23 | 22 | 22 | 22 | 22 | 21 | 21 | 21 | 20 | 20 | 19 | 19 | 18 | 16 |

The shape of the data is slightly left-skewed (Figure 5.13) with a possible weak outlier at the maximum value of 30°C (Table 5.6). The center is best measured by the median, which is 23°C (Table 5.6) and the dispersion is best measured by the IQR, which is from 21 to 25°C (Table 5.6). I used the median and IQR because of the (combined) skewed shape and outlier present.

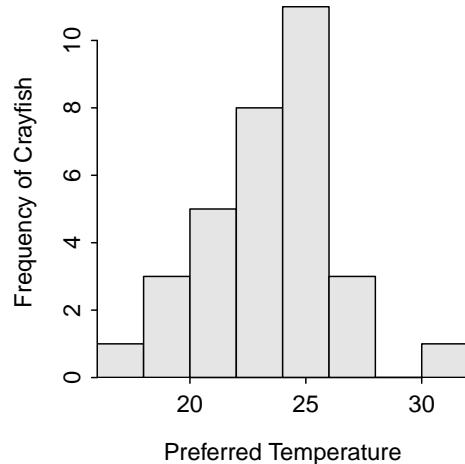


Figure 5.13. Histogram of crayfish temperature preferences.

Table 5.6. Descriptive statistics of crayfish temperature preferences.

| n | mean | sd | min | Q1 | median | Q3 | max |
|-------|-------|------|-------|-------|--------|-------|-------|
| 32.00 | 22.88 | 2.79 | 16.00 | 21.00 | 23.00 | 25.00 | 30.00 |

R commands:

```
> setwd("c:/data/")
> cray <- read.csv("Crayfish.csv")
> str(cray)
> hist(~temp,data=cray,xlab="Preferred Temperature",ylab="Frequency of Crayfish",w=2)
> Summarize(~temp,data=cray,digits=2)
```

Review Exercises

- 5.41**  Construct a proper EDA for the creatine phosphokinase data presented in Exercise 5.5. Make sure to defend your choice of numerical summaries. [Answer](#)
- 5.42**  The Dow Jones Travel Index tracks the cost of hotel and car-rental rates in 20 major cities. For its May 7, 1996, survey the following rates were given for the 20 cities: 152, 180, 167, 119, 115, 113, 119, 135, 140, 126, 114, 133, 205, 104, 149, 124, 127, 161, 106, and 179. Thoroughly describe the distribution of these data. [Note: You can use fewer than the ideal number of bars on your histogram because the sample size is so small in this situation.] [Answer](#)
- 5.43**  The data in [Zoo2.csv](#) contains the physical size (in acres) of a sample of zoos from around the United States. Perform a univariate EDA on the *size* variable. [Answer](#)
-

CHAPTER 6

UNIVARIATE EDA - CATEGORICAL

Chapter Objectives:

1. Construct frequency and percentage tables with categorical data.
2. Construct bar-charts with categorical data, and
3. Use tables and graphs to describe the categorical data.

Contents

| | | |
|-----|-----------------------------------|----|
| 6.1 | Items to Describe | 76 |
| 6.2 | Summary Tables | 76 |
| 6.3 | Bar Plots | 78 |
| 6.4 | Example Interpretations | 81 |

6.1 Items to Describe

Interpreting summaries of a single categorical variable is more intuitive and less defined than that for quantitative data. Specifically, one DOES NOT describe shape, center, dispersion, and outliers for categorical data. In this chapter, methods to construct tables and graphs for categorical data are described and the interpretation of the results demonstrated. These concepts are illustrated with data recorded about MTH107 students in the Winter 2010 semester. Whether or not a student was required to take the course for a subset of individuals is shown in Table 6.1.

Table 6.1. Whether (Y) or not (N) MTH107 was required for a subset of individuals in MTH107 in Winter 2010.

| Individual | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|------------|---|---|---|---|---|---|---|---|
| Required | Y | N | N | Y | Y | Y | N | Y |

6.2 Summary Tables

A simple method to summarize categorical data is to count the number of individuals in each category (or level) of the categorical variable. These counts are called frequencies and the resulting table (Table 6.2) is called a frequency table. From this table, it is seen that there were five students that were required and three that were not required to take MTH107.

Table 6.2. Frequency table for whether MTH107 was required (Y) or not (N) for a subset of individuals in MTH107 in Winter 2010.

| Required | Freq |
|----------|------|
| Y | 5 |
| N | 3 |

- ◊ Frequency tables show the total count or number of individuals in each category (or level) of a categorical variable.

The remainder of this chapter will use the results from the entire class rather than the subset used above to illustrate computing a frequency table. The frequency tables of individuals by sex and by year-in-school for the entire class is shown in Table 6.3.

Table 6.3. Frequency tables for whether (Y) or not (N) MTH107 was required (Left) and year-in-school (Right) of all individuals in MTH107 in Winter 2010.

| Required | Freq | Year | Freq |
|----------|------|------|------|
| Y | 38 | Fr | 19 |
| N | 30 | So | 12 |

| | |
|----|----|
| Jr | 29 |
| Sr | 9 |

Many times, frequency tables are modified by showing the percentage, rather than the frequency, of individuals in each category. These modified tables are called **percentage tables**. Percentage tables are constructed from frequency tables by dividing the frequency of individuals in each category by the total number of individuals examined (n) and then multiplying by 100. For example, the percentage tables for both whether or not MTH107 was required and year-in-school (Table 6.4) for students in MTH107 is constructed from Table 6.3 by dividing the value in each cell by 68, the total number of students in the class, and then multiplying by 100. From this it is seen that 55.9% of students were required to take the course and 13.2% were seniors.

Table 6.4. Percentage tables for whether (Y) or not (N) MTH107 was required (Left) and year-in-school (Right) of all individuals in MTH107 in Winter 2000.

| Required | Perc | Year | Perc |
|----------|------|------|------|
| Y | 55.9 | Fr | 27.9 |
| N | 44.1 | So | 17.6 |
| | | Jr | 42.6 |
| | | Sr | 13.2 |

- ◊ Percentage tables show the percentage of all individuals in each category (or level) of a categorical variable.

6.2.1 Tables in R

The General Sociological Survey (GSS) is a very large survey that has been administered 25 times since 1972. The basic purpose of the GSS is to gather data on contemporary American society in order to monitor and explain trends in attitudes, behaviors, and attributes. One question that was asked in a recent GSS was “How often do you make a special effort to sort glass or cans or plastic or papers and so on for recycling?” These data are found in the *recycle* variable of the [GSSEnviroQues.csv](#) file.

```
> GSS <- read.csv("data/GSSEnviroQues.csv")
> str(GSS)
'data.frame': 3539 obs. of  2 variables:
 $ recycle: Factor w/ 5 levels "Always","Never",...: 1 1 1 1 1 1 1 1 1 ...
 $ tempgen: Factor w/ 5 levels "Extremely","Not",...: 1 1 1 1 1 1 1 1 1 ...
> levels(GSS$recycle)
[1] "Always"      "Never"       "Not Avail"    "Often"        "Sometimes"
```

These results show the five levels in the *recycle* factor variable. Note that the levels appear to be out of order because the default for a factor variable is to alphabetically order the levels. The levels should be in the order “Always”, “Often”, “Sometimes”, “Never”, and “Not Avail” to follow the natural order of this ordinal variable. The order of a factor variable is controlled by including the ordered level names within a vector sent to the `levels=` argument in `factor()`. The level names in this vector must be exactly as they appear in the original variable and they must be contained within quotes. The order of the levels of *recycle* is changed and stored in a new variable in *GSS* with

```
> GSS$freqcycle <- factor(GSS$recycle,levels=c("Always","Often","Sometimes","Never","Not Avail"))
> levels(GSS$freqcycle)
[1] "Always"      "Often"       "Sometimes"   "Never"       "Not Avail"
```

The advantage of correcting this order is that when the summary table is made, the order will follow the natural order of the variable rather than the alphabetical order.

◊ The order of the levels of a factor are controlled with the `levels=` argument in the `factor()` function.

◊ When changing the order of the levels with the `levels=` argument, the level names must be contained within quotes and they must be spelled exactly as they were spelled in the original variable.

The frequency table of a single categorical variable is computed with `xtabs()` where the first argument is a one-sided formula of the form `~var` with the corresponding data.frame in `data=`. The result from `xtabs()` should be assigned to an object for further processing. For example, the frequency table is produced, stored in the `tabRecycle` object, and displayed with

```
> ( tabRecycle <- xtabs(~frecycle,data=GSS) )
frecycle
  Always    Often Sometimes    Never Not Avail
  1289      850       823      448      129
```

Thus, for example, 1289 respondents answered “Always” to the recycling question.

The percentage table is computed in R by including the saved frequency table as the first argument to `percTable()`.¹ The number of digits of output is controlled with `digits=`.

```
> percTable(tabRecycle,digits=1)
frecycle
  Always    Often Sometimes    Never Not Avail      Sum
  36.4     24.0      23.3     12.7      3.6     100.0
```

Thus, for example, 36.4% of respondents answered “Always” to the recycling question.

6.3 Bar Plots

Bar plots, or bar charts, are used to display the frequency or percentage of individuals in each level of a categorical variable. Bar plots look similar to histograms in that they have the frequency of individuals on the y-axis. However, category labels rather than quantitative values are plotted on the x-axis. In addition, bars on a bar plot are usually not connected in order to highlight the categorical nature of the data. A bar plot for whether or not individuals were required to take MTH107 is shown in Figure 6.1-Left. This bar plot is not very helpful because there were only two categories (i.e., it does not add much to the frequency table). However, bar plots make it easier to compare the number of individuals in each of several categories as in Figure 6.1-Right.

¹

Thus, `xtabs()` must be completed and saved to an object before `percTable()`.

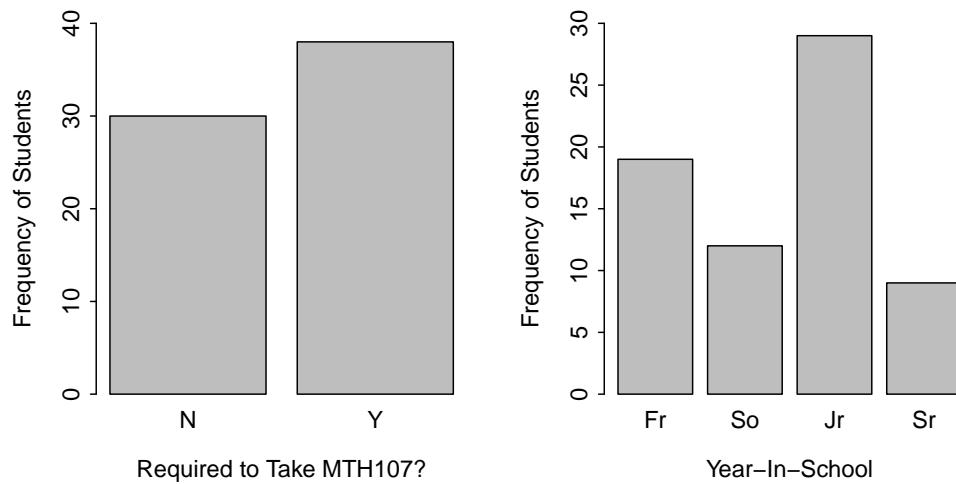


Figure 6.1. Bar charts of the frequency of individuals in MTH107 during Winter 2010 by whether or not they were required to take MTH107 (**Left**) and year-in-school (**Right**).

- ◊ Bar charts are used to display the frequency of individuals in the categories of a categorical variable. Histograms are used to display the frequency of individuals in classes created from quantitative variables.

Shape, center, dispersion, and outliers are not described for categorical data because the data is not numerical and, if nominal, no order exists. In general, the major characteristics of the table or graph are described from an intuitive basis. For example, there were more males than females in the Winter 2010 MTH107 class and mostly juniors and Freshmen.

- ◊ Do not describe shape, center, dispersion, and outliers for a categorical variable.

6.3.1 Bar Plots in R

A bar plot is produced by submitting the `xtabs()` object as the first argument to `barplot()`. The x- and y-axes may be explicitly labeled with `xlab=` and `ylab=`, respectively. The bar plot for the recycling data (Figure 6.2) is produced with

```
> barplot(tabRecycle, ylab="Frequency", xlab="Recycle Response")
```

Review Exercises

- 6.1** Use the [Arsenic.csv](#) data in Exercise 5.40 to answer the questions below. [Answer](#)

- (a) Construct a univariate EDA for the assessment of drinking water usage.

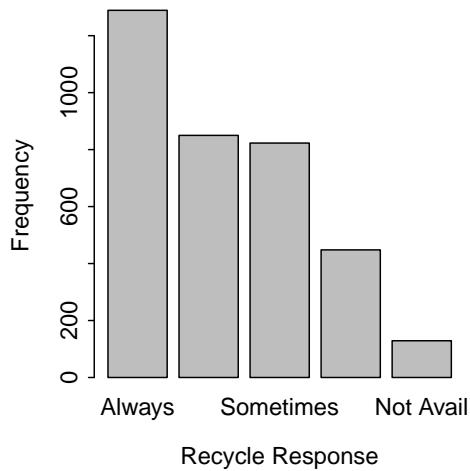


Figure 6.2. Bar chart of the frequency of responses to the recycling question on the GSS.

- (b) Construct a univariate EDA for the assessment of cooking water usage.

6.2 The Environmental Protection Agency (EPA) commissioned the Gallup Organization to conduct a nationwide telephone survey of 1000 households during August and September of 2002 regarding consumer knowledge and satisfaction with drinking water quality. Of the 1000 respondents surveyed, 751 knew that their drinking water came from a public or commercial water supplier. Of these 751 respondents, the following percentages knew precisely where that water was derived:

| Ground-water | Lake/Reservoir | River | Multiple Sources | Don't Know | Refused Answer |
|--------------|----------------|-------|------------------|------------|----------------|
| 15.9% | 29.2% | 9.6% | 15.7% | 29.4% | 0.2% |

Use these data to answer the questions below. [Answer](#)

- (a) Construct a frequency table of these data (note percentages above were rounded).
 (b) Write a brief conclusion derived from these data.

6.3 A neighborhood in Honolulu conducted a survey to determine if residents participated in the curbside recycling program. One question on their survey was, "How much has curbside recycling reduced your regular refuse? 0%, 25%, 50%, 75%, 100%, or 'too early to tell'?" The individual responses for the returned surveys are shown below with letters corresponding to the category choices offered (e.g., A=0%, B=25%, and so on).

C, C, B, B, B, C, E, B, B, C, B, C, C, C, E, B, B, B,
 C, B, B, C, C, B, C, B, B, C, B, B, B, C, E, B,
 E, B, B, C, C, B, B, E, B, C, C, B, B, C, B, B, B, B

Use these data to answer the questions below. [Answer](#)

- (a) Construct a frequency table of these data.
 (b) Construct a percentage table of these data.
 (c) Write a brief conclusion derived from these data.

- 6.4**  Students in a senior level environmental studies class at Rice University conducted a voluntary response survey regarding water usage by their peers. They received returned surveys from a total 130 students. One question on their survey was, “On average, for how many minutes do you let the water run each time you take a shower? 0-5, 6-10, 11-15, or over 15 minutes?” The individual responses for this survey are shown below with letters corresponding to the category choices offered (e.g., A=“0-5”, B=“6-10”, and so on).

[Answer](#)

D, C, B, B, C, C, B, C, C, C, B, D, B, C, C, B, C, D, D,
 B, C, C, A, B, C, C, A, C, C, D, A, C, C, B, B, B, B, C,
 D, B, D, B, C, B, C, C, D, C, B, B, D, C, B, C, B, B, C, C,
 B, C, B, C, B, B, C, D, B, C, D, C, B, C, D, C, C, B, C, B,
 D, B, B, D, B, C, B, C, B, C, D, D, C, D, B, B, C, B, C,
 A, A, B, C, B, C, D, D, C, B, D, C, C, C, A, C, D, B, C,
 B, B, D, C, B, B, A, B, C, B

Use these data to answer the questions below.

- Construct a frequency table of these data.
- Construct a percentage table of these data.
- Write a brief conclusion derived from these data.

6.4 Example Interpretations

While most of the previous sections focused on how to construct various graphs and numerical summaries, the most important aspect of this chapter is that you can make appropriate interpretations for an EDA from the summary results. For categorical data, an appropriate EDA consists of identifying the major characteristics among the categories.

6.4.1 Mixture Seed Count

A bag of seeds was purchased for seeding a recently constructed wetland. The purchaser wanted to determine if the percentage of seeds in four broad categories – “grasses”, “sedges”, “wildflowers”, and “legumes” – was similar to what the seed manufacturer advertised. The purchaser examined a 0.25-lb sample of seeds from the bag and recorded the results in [WetlandSeeds.csv](#). Use these data to describe the distribution of seed counts into the four broad categories.

The majority of seeds were either sedge or grass with sedge being more than twice as abundant as grass (Table 6.5; Figure 6.3). Very few legumes or wildflowers were found in the sample.

Table 6.5. Percentage distribution of wetland seeds by type.

| grass | legume | sedge | wildflower | Sum |
|-------|--------|-------|------------|-------|
| 27.9 | 1.6 | 64.5 | 6.0 | 100.0 |

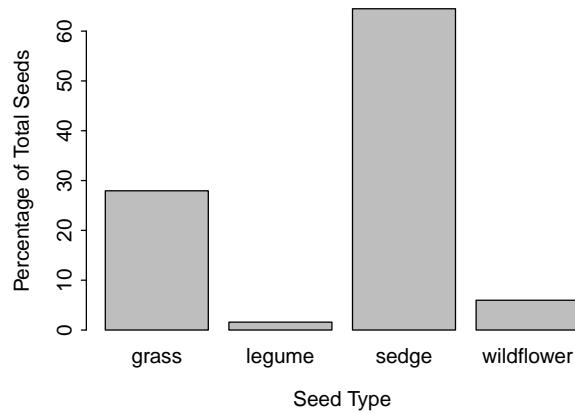


Figure 6.3. Barplot of the percentage of wetland seeds by type.

R commands:

```
> ws <- read.csv("data/WetlandSeeds.csv")
> str(ws)
> wtbl <- xtabs(~type,data=ws)
> percTable(wtbl,digits=1)
> barplot(wtbl[,-5],ylab="Percentage of Total Seeds",xlab="Seed Type")
```

Review Exercises

- 6.5** The data in [Zoo1.csv](#) contains a list of animals found in several different zoos. In addition, each animal was classified into broad “type” categories (“mammal”, “bird”, and “amph/rep”). Perform a univariate EDA on the *type* variable. [Answer](#)
-

CHAPTER 7

NORMAL DISTRIBUTION

Chapter Objectives:

1. Describe what a normal distribution looks like and what parameters control its shape.
2. Describe simple properties describing the distribution of individuals on a normal distribution.
3. Compute the proportion of individuals with a particular set of values from a normal distribution (“forward” calculations).
4. Compute the range of values for a certain proportion of individuals from a normal distribution (“reverse” calculations).

Contents

| | | |
|-----|--|----|
| 7.1 | Characteristics | 84 |
| 7.2 | Simple Areas Under the Curve | 85 |
| 7.3 | Forward Calculations | 88 |
| 7.4 | Reverse Calculations | 91 |
| 7.5 | Standardization and Z-Scores | 95 |

A MODEL FOR THE DISTRIBUTION of a single quantitative variable can be visualized by “fitting” a smooth curve to a histogram, removing the histogram, and using the remaining curve as a model for the distribution of the entire population of individuals. This process is illustrated with the set of three figures shown in Figure 7.1. The underlying histogram was computed from the individuals in a very large sample. The smooth red curve was drawn over the histogram and then removed to serve as a model for the distribution of the entire population. If the smooth curve follows a known distribution, then certain calculations will be greatly simplified.

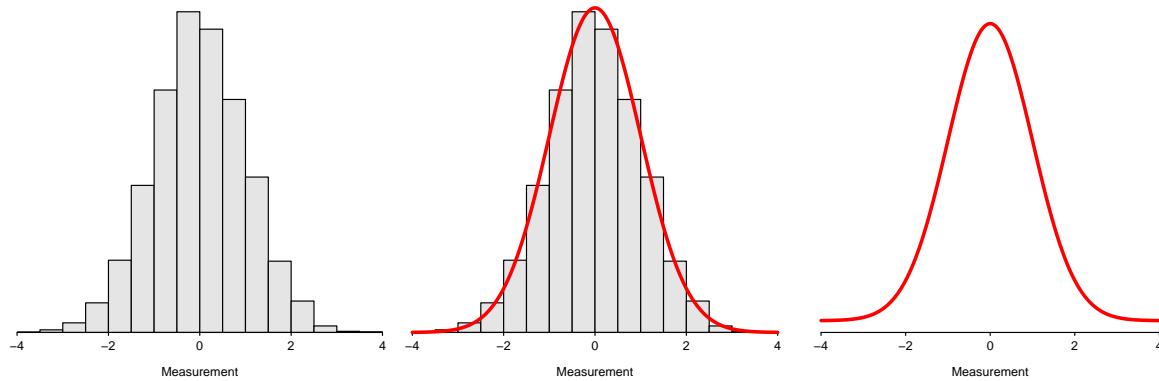


Figure 7.1. Depiction of fitting a smooth curve to a histogram and then removing the histogram to leave the smooth curve model.

The normal distribution is one of the most important distributions in statistics because it serves as a model for the distribution of individuals in many natural situations and the distribution of statistics from repeated samplings (i.e., sampling distributions¹). The use of a normal distribution model to make certain calculations will be demonstrated in this chapter.

7.1 Characteristics

The normal distribution is the common bell-shaped curve that you are probably familiar with (Figure 7.1–Right). Normal distributions are abstractions of reality that are meant to represent all of the individuals in a population. The height of the curve at a value of x is found with

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad (7.1.1)$$

which has the two parameters² – the population mean, μ , and the population standard deviation, σ . The mean, μ , controls the center and the standard deviation, σ , controls the dispersion of the normal distribution (Figure 7.2).

◇ It is not important that you remember the equation for the height of a normal distribution; you need to remember, though, that the exact position and width of the normal distribution is controlled only by the values of μ and σ .

¹

See Chapter 12.

²

The e and π are the usual numerical constants.

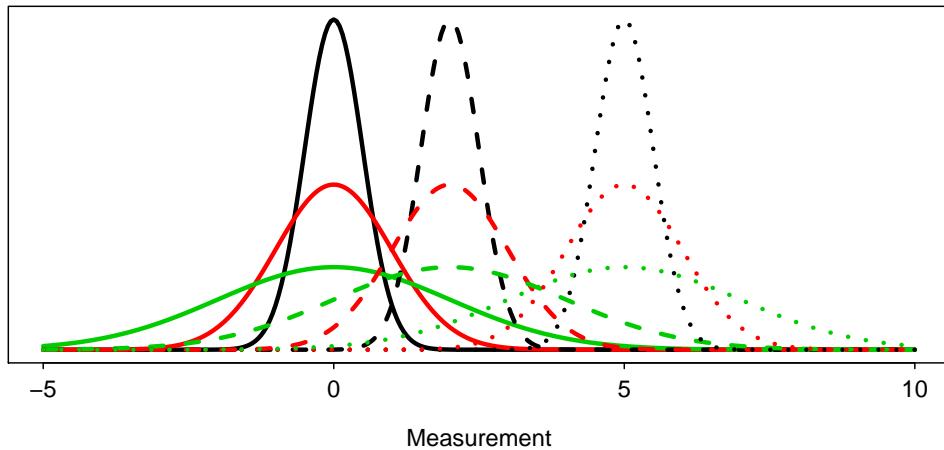


Figure 7.2. Multiple normal distributions. Distributions with the same line type have the same value of μ . Distributions with the same color have the same value of σ . Values of μ are 0 (solid), 2 (dashed), and 5 (dotted). Values of σ are 0.5 (black), 1 (red), and 2 (green).

There are an infinite number of normal distributions because there are an infinite number of combinations of μ and σ . However, each normal distribution will

1. be bell-shaped and symmetric,
2. have a center at μ ,
3. have inflection points at $\mu \pm \sigma$, and
4. have a total area under the curve equal to 1.

◊ All normal distributions are bell-shaped. The center and dispersion of each, though, is dictated by the values of μ and σ , respectively.

If a generic variable X follows a normal distribution with a mean of μ and a standard deviation of σ , then it is said that $X \sim N(\mu, \sigma)$. For example, if the heights of students (H) follows a normal distribution with a μ of 66 and a σ of 3, then it is said that $H \sim N(66, 3)$. As another example, $Z \sim N(0, 1)$ means that the variable Z follows a normal distribution with a mean of $\mu=0$ and a standard deviation of $\sigma=1$.

◊ A generic variable X that is normally distributed with a mean of μ and standard deviation of σ is abbreviated as $X \sim N(\mu, \sigma)$.

7.2 Simple Areas Under the Curve

A common statistical problem is to determine the proportion of individuals that have values of the variable between two numbers. For example, you might be faced with determining the proportion of all sites that have lead concentrations between 1.2 and $1.5 \mu\text{g} \cdot \text{m}^{-3}$, the proportion of students that scored higher than 700 on the SAT, or the proportion of least weasels that are shorter than 150 mm. Before considering these more realistic situations we will explore the calculations for the generic variable X shown in Figure 7.3.

Let's consider finding the proportion of individuals in a *sample* with values between 0 and 2. A histogram can be used to answer this question because it is about the individuals in a sample (Figure 7.3-Left). In this case, the proportion of individuals with values between 0 and 2 is computed by dividing the number of individuals in the red shaded bars by the total number of individuals in the histogram. The analogous computation on the superimposed smooth curve is to find the area under the curve between 0 and 2 (Figure 7.3-Right). The area under the curve is a “proportion of the total” because, as stated above, the area under the entire curve is equal to 1. The actual calculations on the normal curve will be shown in the following sections. However, at this point, note that the calculation of an area on a normal curve is analogous to summing the number of individuals in the appropriate classes of the histogram and dividing by n .

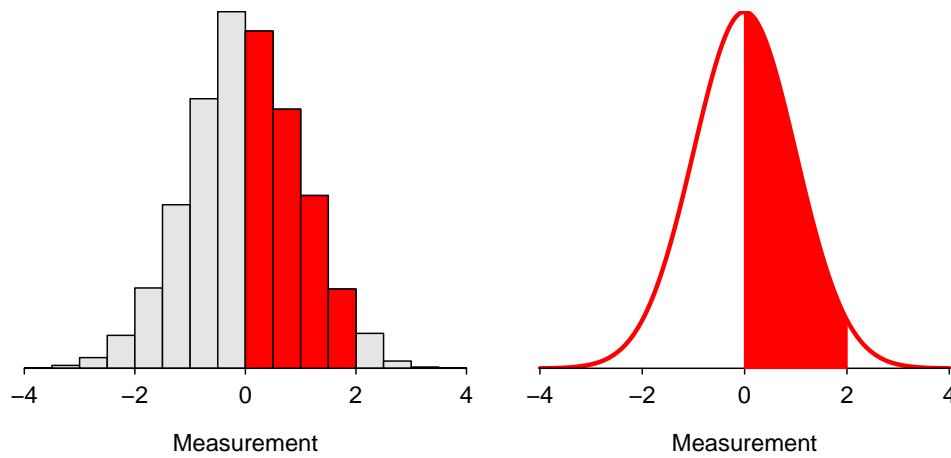


Figure 7.3. Depiction of finding the proportion of individuals between 0 and 2 on a histogram (**Left**) and on a standard normal distribution (**Right**).

- ◊ The proportion of individuals between two values of a variable that is normally distributed is found by finding the area under the normal distribution between those two values.

The 68-95-99.7 Rule³ states that 68% of the individuals that follow a normal distribution will have values between $\mu - 1\sigma$ and $\mu + 1\sigma$, 95% will be between $\mu - 2\sigma$ and $\mu + 2\sigma$, and 99.7% will be between $\mu - 3\sigma$ and $\mu + 3\sigma$ (Figure 7.4). The 68-95-99.7 Rule is true no matter what μ and σ are as long as the distribution is normal. For example, if $A \sim N(3, 1)$, then 68% of the individuals will fall between 2 (i.e., $3-1*1$) and 4 (i.e., $3+1*1$) and 99.7% will fall between 0 (i.e., $3-3*1$) and 6 (i.e., $3+3*1$). Alternatively, if $B \sim N(9, 3)$, then 68% of the individuals will fall between 6 (i.e., $9-1*3$) and 12 (i.e., $9+1*3$) and 95% will be between 3 (i.e., $9-2*3$) and 15 (i.e., $9+2*3$). Similar calculations can be made for any normal distribution.

Δ 68-95-99.7 Rule: For all normal distributions 68% of the individuals will be between $\mu \pm 1\sigma$, 95% will be between $\mu \pm 2\sigma$, and 99.7% will be between $\mu \pm 3\sigma$

The 68-95-99.7 Rule is used to find areas under the normal curve as long as the value of interest is an **integer** number of standard deviations away from the mean. For example, the proportion of individuals that have

³

Other authors call this the “Empirical Rule.”

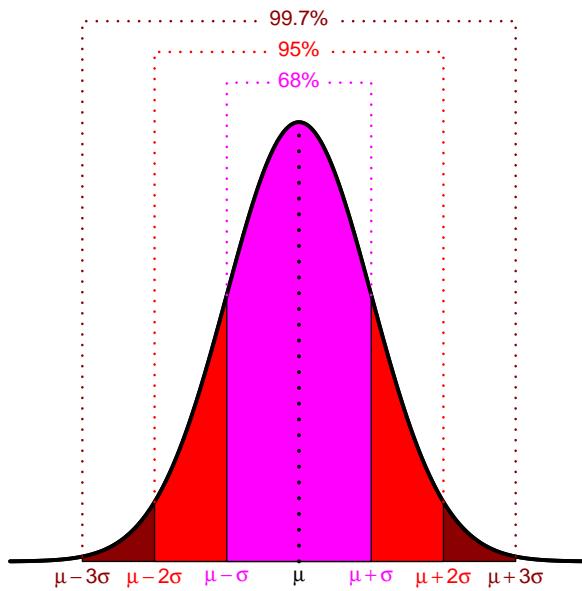


Figure 7.4. Depiction of the 68-95-99.7 (or Empirical) Rule on a normal distribution.

a value of A greater than 5 is found by first realizing that 95% of the individuals on this distribution fall between 1 and 5. By subtraction this means that 5% of the individuals must be less than 1 **AND** greater than 5. Finally, because of the symmetry of normal distributions, the same proportion of individuals must be less than 1 as are greater than 5. Thus, half of 5%, or 2.5%, of the individuals have a value of A greater than 5.

◊ The 68-95-99.7 Rule can only be used for questions involving integer standard deviations away from the mean.

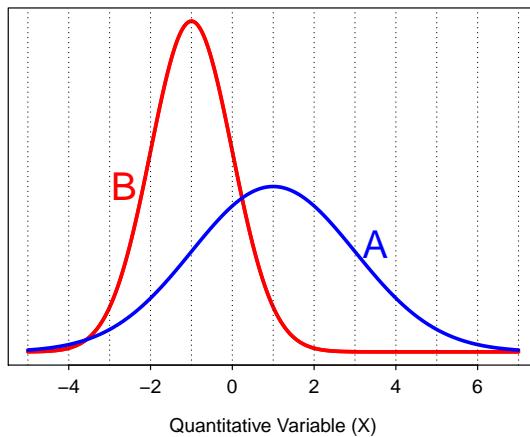
Review Exercises

- 7.1** On any normal distribution, what percentage of the individuals is within $\pm 1\sigma$ of μ ? [Answer](#)
- 7.2** On any normal distribution, what percentage of the individuals are greater than $\mu + \sigma$? [Answer](#)
- 7.3** On any normal distribution, what percentage of the individuals are greater than $\mu - 2\sigma$? [Answer](#)
- 7.4** On any normal distribution, what percentage of the individuals are between $\mu - 2\sigma$ and $\mu + 1\sigma$? [Answer](#)
- 7.5** On a $N(-1,1)$ distribution, what percentage of the individuals are negative? [Answer](#)
- 7.6** On a $N(100,20)$ distribution, what percentage of the individuals are less than 80? [Answer](#)

7.7 On a $N(-20,100)$ distribution, what percentage of the individuals are greater than 80? [Answer](#)

7.8 Identify the mean and standard deviation for each population on the graph below (HINT: "eyeball" integers).

[Answer](#)



7.3 More Complex Areas (Forward Calculations)

Areas under the curve relative to a non-integer number of standard deviations away from the mean used to be found via a calculation and examination of a so-called standard normal table. With the advent of computers and cheap software these areas are now found simply with the aid of computer software like R. The area under a normal curve relative to a particular value is computed in R with `distrib()`. This function requires the *particular value* as the first argument, the mean of the normal distribution in the `mean=` argument, and the standard deviation of the normal distribution in the `sd=` argument. The `distrib()` function defaults to finding the area under the curve to the **left of** the particular value but it can find the area under the curve to the right of the particular value by including the `lower.tail=FALSE` argument.

For example, suppose that the heights of a population of students, represented by H , is known to be $H \sim N(66, 3)$. Thus, the proportion of students in this population that have a height less than 71 inches is computed with (the results are shown below and in Figure 7.5),

```
> ( distrib(71,mean=66,sd=3) )
[1] 0.9522096
```

Thus, approximately 95.2% of the students in this population have a height less than 71 inches. The proportion of students in this population that have a height *greater* than 68 inches is computed by including the `lower.tail=FALSE` argument as follows (with the results shown below and in Figure 7.6),

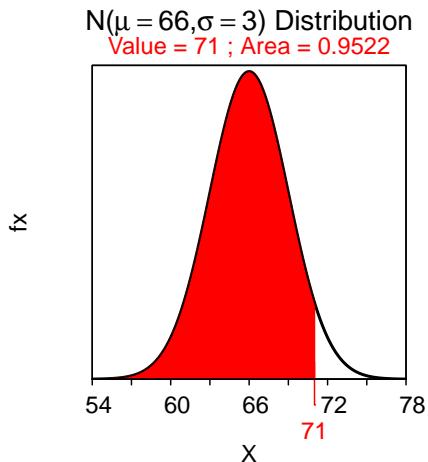


Figure 7.5. Calculation of the proportion of individuals on a $N(66, 3)$ with a value less than 71.

```
> ( distrib(68,mean=66,sd=3,lower.tail=FALSE) )
[1] 0.2524925
```

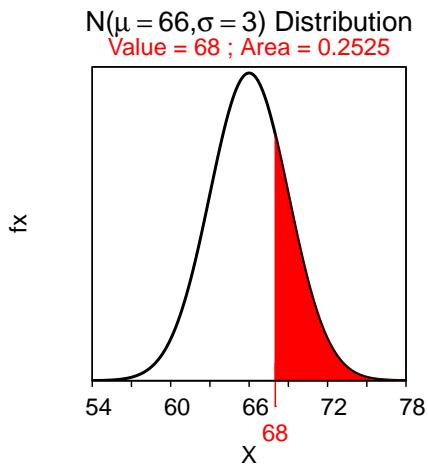


Figure 7.6. Calculation of the proportion of individuals on a $N(66, 3)$ with a value greater than 68.

Thus, approximately 25.2% of the students in this population have a height greater than 68 inches.

- ◊ The area greater than a particular value is found by including the `lower.tail=FALSE` argument in `distrib()`.

Finding the area between two particular values is a bit more work. To answer “between”-type problems, the area less than the smaller of the two values is subtracted from the area less than the larger of the two values. This is illustrated by noting that two values split the area under the normal curve into three parts – A, B,

and C (Figure 7.7). The area between the two values is B. The area to the left of the larger value corresponds to the combined area of A and B (i.e., A+B). The area to the left of the smaller value corresponds to the area A. Thus, subtracting the latter from the former leaves the “in-between” area B (i.e., (A+B)-A = B).

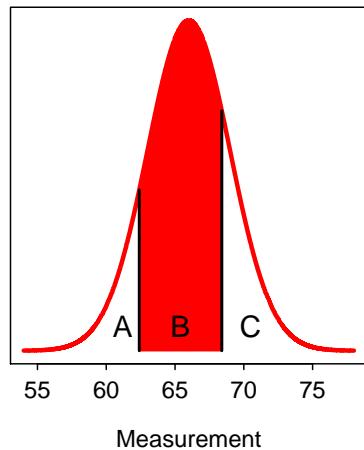


Figure 7.7. Schematic representation of how to find the area between two Z values.

For example, the area between 62 and 70 inches of height is found with (with intermediate calculations shown in Figure 7.8)

```
[1] 0.9087888
> ( A <- distrib(62,mean=66,sd=3) )
[1] 0.09121122
> AB-A
[1] 0.8175776
```

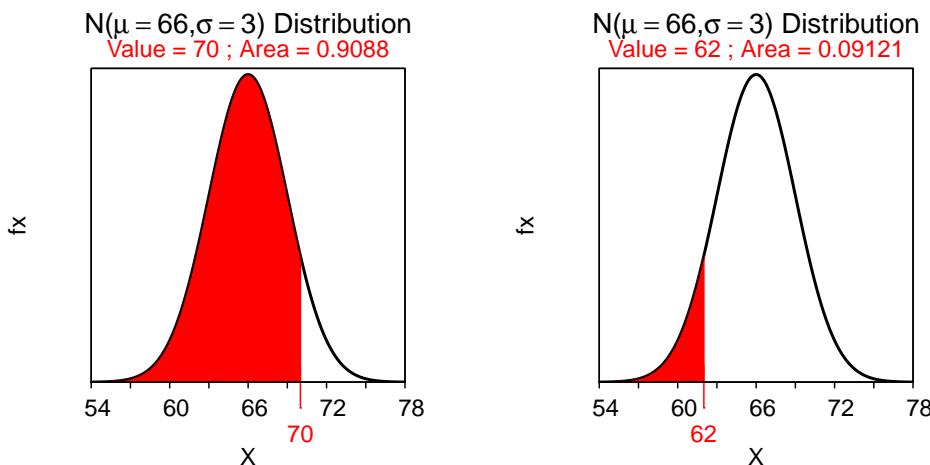


Figure 7.8. Calculation of the areas less than 70 inches (Left) and 62 inches (Right).

Thus, 81.8% of students in this population have a height between 62 and 70 inches.

- ◊ The area between two values is found by subtracting the area less than the smaller value from the area less than the larger value.

Review Exercises

- 7.9** If $X \sim N(0, 1)$, then what is the percentage of $X < 0.11$? [Answer](#)
- 7.10** If $X \sim N(0, 1)$, then what is the percentage of $X > -0.11$? [Answer](#)
- 7.11** If $X \sim N(0, 1)$, then what is the percentage of $-1.45 < X < 1.11$? [Answer](#)
- 7.12** If $Y \sim N(70, 6)$, then what is the percentage of $Y > 75$? [Answer](#)
- 7.13** If $Y \sim N(70, 6)$, then what is the percentage of $Y < 63$? [Answer](#)
- 7.14** If $Y \sim N(70, 6)$, then what is the percentage of $62.3 < Y < 72.9$? [Answer](#)

7.4 Values from Areas (Reverse Calculations)

Another important calculation with normal distributions is finding the value or values of X with a given proportion of individuals less than, greater than, or between. For example, it may be necessary to find the test score such that 90% (or 0.90 as a proportion) of the students scored lower. In contrast to the calculations in the previous section, the value of X is given and a proportion of individuals (or area) is asked for, the calculations in this section give a proportion and ask for a value of X . These types of questions have been dubbed “reverse” normal distribution questions to contrast them with the questions in the previous section.

Reverse questions are also answered with `distrib()`. Again, the first argument must be the value of interest – a proportion (or area) in these questions – and the mean and standard deviation are given in `mean=` and `sd=`, respectively. However, the question is treated as a “reverse” question when the `type="q"` argument is supplied.⁴ Thus, the height that has 20% of all individuals shorter (Figure 7.9) is computed with

```
> ( distrib(0.20,mean=66,sd=3,type="q") )
[1] 63.47514
```

⁴

“q” stands for quantile.

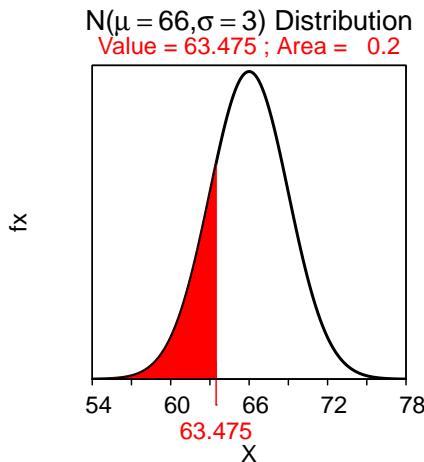


Figure 7.9. Calculation of the height with 20% of all students shorter.

Thus, 20% of the population of students is shorter than 63.5 inches. “Greater than” reverse questions are computed by including the `lower.tail=FALSE` argument. For example, the top 10% of heights (Figure 7.10) is found with

```
> ( distrib(0.10,mean=66,sd=3,type="q",lower.tail=FALSE) )
[1] 69.84465
```

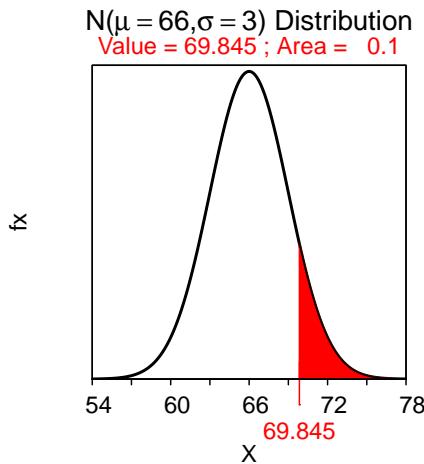


Figure 7.10. Calculation of the height with 10% of all students taller.

Thus, 10% of the population of students is taller than 69.8 inches.

“Between” questions can only be easily handled if the question is looking for endpoint values that are symmetric about μ . In other words, the question must ask for the two values that contain the “most common” proportion of individuals. For example, suppose that you were asked to find the most common 80% of heights. This type of question is handled by converting this “symmetric between” question into two

“less than” questions. For example, in Figure 7.11 the area D is the symmetric area of interest. If D is 0.80, then C+E must be 0.20.⁵ Because D is symmetric about μ , C and E must both equal 0.10. Thus, the lower bound on D is the value that has 10% of all values smaller. Similarly, because the combined area of C and D is 0.90, the upper bound on D is the value that has 90% of all values smaller. This question has now been converted from a “symmetric between” to two “less than” questions that can be answered exactly as shown above. For example, the two heights that have a symmetric 80% of individuals between them are 62.2 and 69.8 as computed with

```
> ( distrib(0.10,mean=66,sd=3,type="q") )
[1] 62.15535
> ( distrib(0.90,mean=66,sd=3,type="q") )
[1] 69.84465
```

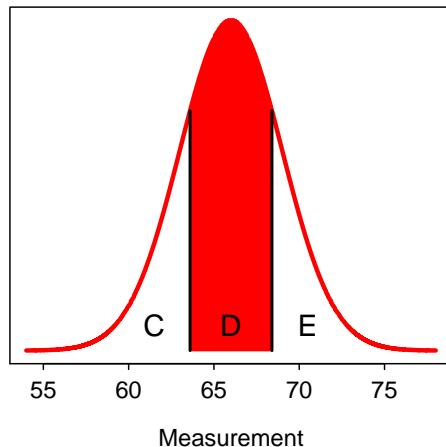


Figure 7.11. Depiction of areas in a reverse between type normal distribution question.

Review Exercises

7.15 If $Y \sim N(70, 6)$, then what is Y such that the area to the left of it is 0.3? [Answer](#)

7.16 If $Y \sim N(70, 6)$, then what is Y such that the area to the right of it is 0.4? [Answer](#)

7.17 If $Y \sim N(70, 6)$, then what are the Ys such that the area between them is 0.5? [Answer](#)

⁵

Because all three areas must sum to 1.

7.4.1 Distinguish Calculation Types

It is critical to be able to distinguish between the two main types of calculations made with normal distributions. The first type of calculation is a “forward” calculation where the area or proportion of individuals relative to a value of the variable must be found. The second type of calculation is a “reverse” calculation where the value of the variable relative to a particular area is calculated.

Distinguishing between these two types of calculations a matter of deciding if (i) the value of the variable is known and the proportion (or area) is to be found or (ii) if the proportion (or area) is known and the value of the variable is to be found. Therefore, distinguishing between the calculation types is as simple as identifying what is known and what must be found. If the value of the variable is known but not the proportion or area, then a forward calculation is used. If the area or proportion is known, then a reverse calculation to find the value of the variable is used.

Review Exercises

- 7.18** The age at which “traditional” students graduate from college is $N(22.1, 1.1)$. Use this information to answer the questions below. [Answer](#)

- (a) What percentage of the students graduate by the age of 21?
- (b) What percentage of the students graduate after age 24?
- (c) What is the age range for the middle 95% of the students?
- (d) What is the age at which 90% of the students have graduated?

- 7.19** We know, from years of study of black bears, that the population distributions for head length is $N(13.7, 1.9)$, neck girth is $N(20.9, 4.8)$, and body length is $N(60.0, 10)$. All other variables measured on black bears cannot be described by a normal distribution. Use this information to answer the questions below. [Answer](#)

- (a) What is the percentage of bears between 45" and 65" in body length?
- (b) What is the percentage of bears that weighs more than 200 lbs?

- 7.20** The brain weights of short-tailed shrews (*Blarina brevicauda*) is normally distributed with a mean of 0.14 grams and a standard deviation of 0.04 grams. Use this information to answer the questions below. [Answer](#)

- (a) What percentage of shrews have a brain weight less than 0.09 grams?
- (b) What percentage of shrews have a brain weight between 0.09 and 0.17 grams?
- (c) What is the brain weight such that 30% of all shrews have a larger brain weight?

- 7.21** The distribution of arrival times for the BART bus at Northland is normally distributed with a mean of 0 and standard deviation of 3, where negative values indicate early arrivals (i.e., before the scheduled time) and positive values indicate late arrivals. Use this information to answer the questions below. [Answer](#)

- (a) What percentage of the arrivals are more than 5 minutes late?
- (b) What percentage of the arrivals are more than 4 minutes early?
- (c) What percentage of the arrivals are between 4 minutes early and 4 minutes late?
- (d) What is the arrival time such that 25% of all arrival times are later than that time?
- (e) What are the most common 60% of arrival times?

- (f) What kind of variable is arrival time?

7.22 Researchers on Storfosna Is., Norway wanted to examine reproductive habits of roe deer *Capreolus capreolus* in the northern extremities (Andersen and Linnell 2000). The researchers observed how many fawns were born to each of 149 female, sexually mature roe deer between the years 1991 and 1994. The mean number of fawns from each deer was 2.235 with a standard deviation of 0.460. Use this information to answer the questions below. [Answer](#)

- (a) What percentage of does have less than 2 fawns.
- (b) What percentage of does have more than 3 fawns.
- (c) What percentage of does have between 1 and 3 fawns.
- (d) What is the number of fawns such that only 7.6% of the does have fewer fawns?
- (e) What is the number of fawns such that only 4.2% of the does have more fawns?
- (f) What is the most common 87% of number of fawns born per doe?

7.23 I recently investigated the efficacy of becoming a commercial crayfisherman (crayfish = crawfish = crawdad) on the lake I live on. With carefully constructed samples I concluded that the size of crayfish was $N(93, 8)$. The market for crayfish resides in Sweden. Swedes prefer (hence, will only buy) crayfish that are between 90 and 110 mm long (< 90 are too small to deal with and > 110 taste bad). Use this information to answer the questions below. [Answer](#)

- (a) How many acceptably-sized crayfish could I send to market, if I could catch approximately 50,000 crayfish? [HINT: compute the proportion of preferably-sized crayfish first.]
 - (b) If I could find an alternative market for the larger (> 110) crayfish, how many could I send to it (again assume that I could catch 50,000 crayfish)?
-

7.5 Standardization and Z-Scores

The relative magnitude that an individual differs from the mean is often better expressed as the number of standard deviations that the individual is away from the mean. For example, if heights are $N(66, 3)$ and an individual's height is 59 inches, then it is said that the individual is seven inches shorter than average. Can this be considered to be a large or a small difference? However, it can also be said that the individual is $\frac{-7}{3} = -2.33$ standard deviations below the mean. One can conclude that the height of this individual is relatively rare because it is known⁶ that very few individuals are more than two standard deviations away from the mean.

Values are “standardized” by changing the original scale (inches in this example) to one that counts the number of standard deviations (i.e., σ) that the value is away from the center of the distribution (i.e., μ). For example, with the height variable (i.e., $N(66,3)$), 69 inches is one standard deviation above the mean and, thus, corresponds to 1 on the standardized scale. Similarly, 60 inches is two standard deviations below the mean and corresponds to -2 on the standardized scale. Finally, 67.5 inches on the original scale is one half standard deviation above the mean and corresponds to 0.5 on the standardized scale.

⁶

From the 68-95-99.7% Rule.

The process of computing the number of standard deviations that an individual is away from the mean is called **standardizing**. Standardizing is accomplished with the generic formula,

$$Z = \frac{\text{"value" - "center"}}{\text{"dispersion"}} \quad (7.5.1)$$

or with the more specific formula,

$$Z = \frac{x - \mu}{\sigma} \quad (7.5.2)$$

The more general of these two formulae, i.e., (7.5.1), is preferred over the specific formula because it will work in all later applications. Using the height example again, the standardized value of an individual with a height of 59 inches is $z = \frac{59-66}{3} = -2.33$. Thus, this individual's height is 2.33 standard deviations below the average height in the population.

The standardized value (Z) that is the result from (7.5.1) follows a $N(0, 1)$. Thus, the $N(0, 1)$ is called the “standard normal distribution.” The relationship between X and Z is one-to-one meaning that each value of X converts to one and only one value of Z . This means that the area to the left of X on a $N(\mu, \sigma)$ is the same as the area to the left of $Z = \frac{x-\mu}{\sigma}$ on a $N(0, 1)$. This one-to-one relationship is illustrated in Figure 7.12 using the individual with a height of 59 inches which resulted in $Z = -2.33$.

- ◊ The standardized scale (i.e., z-scores) represents the number of standard deviations that a value is from the mean.

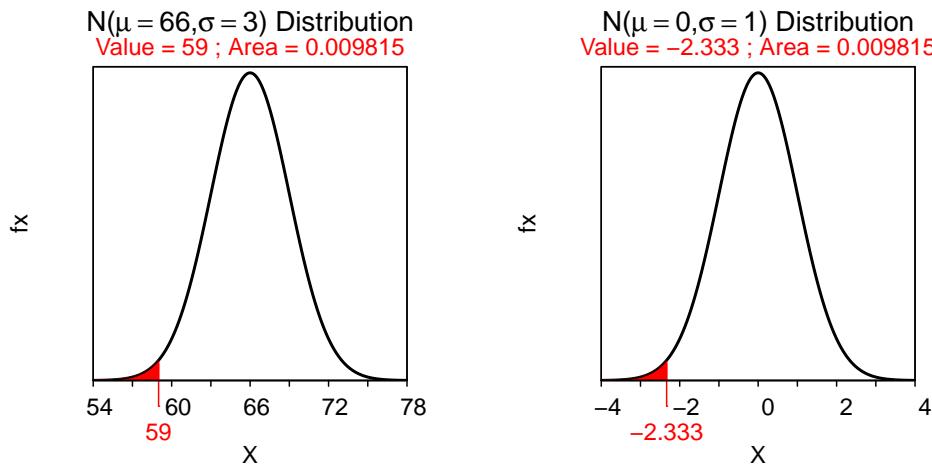


Figure 7.12. Plots depicting the area to the left of 59 on a $N(66, 3)$ (extbfLeft) and the area to the right of the corresponding Z-score of $Z = -2.33$ on a $N(0, 1)$ (extbfRight). Note that the x-axis scales are different.

CHAPTER 8

BIVARIATE EDA - QUANTITATIVE

Chapter Objectives:

1. Describe what bivariate data is.
2. Distinguish between response and explanatory variables.
3. Construct scatterplots of bivariate quantitative data.
4. Describe bivariate relationships with interpretations from scatterplots.
5. Describe how the correlation coefficient is calculated.
6. Use the correlation coefficient to describe the strength (and association) of the relationship between two quantitative variables.

Contents

| | | |
|-----|-------------------------------|-----|
| 8.1 | Scatterplots | 98 |
| 8.2 | Items to Describe I | 101 |
| 8.3 | Correlation | 105 |

BIVARIATE DATA OCCURS WHEN TWO variables have been measured on the same individuals. For example, you may measure (i) the height and weight of students in class, (ii) depth and area of a lake, (iii) gender and age of welfare recipients, or (iv) number of mice in a field and the biomass of legumes in the field. The meaning of bivariate is easy to remember if you split the word into its roots – bi, or “two”, and variate, or “variables.”

△ **Bivariate:** Data where two variables have been measured on the same individuals.

This chapter is focused on describing the bivariate relationship between two quantitative variables. Bivariate relationships between two categorical variables is described in Chapter 9

Data on the *weight* (lbs) and highway miles per gallon (stored as *HMPG*) for 93 cars from the 1993 model year will be used as an example throughout this section (data from Lock (1993)). Ultimately, the relationship between highway MPG and the weight of a car will be examined. These are bivariate data because measurements of both variables (*HMPG* and *Weight*) are recorded for each individual (i.e., a car). The data are stored in [93cars.csv](#). The following commands read the data into R and lists the *HMPG* and *weight* values for six randomly selected cars¹.

```
> cars93 <- read.csv("data/93cars.csv")
> headtail(cars93,which=c("HMPG","Weight"))
   HMPG Weight
1     31    2705
2     25    3560
3     26    3375
91    25    2810
92    28    2985
93    28    3245
```

8.1 Scatterplots

Scatterplots are used to display and identify the relationship between **TWO quantitative** variables. Scatterplots are what most people think of when they hear “plot the data.” The correct construction of a scatterplot usually requires that one of the variables be identified as a response variable and the other as an explanatory variable. The **response variable** is the variable that one is interested in explaining something about (i.e., variability) or in making future predictions about. Synonyms for response variable are dependent variable or predicted variable. The **explanatory variable** is the variable that may help explain or allow one to predict the response variable. Synonyms for explanatory variable are independent variable or predictor variable.

◇ **Both variables must be quantitative to construct a scatterplot.**

△ **Response Variable:** The variable that we are interested in explaining or predicting. Synonyms are “dependent” or “predicted” variable.

¹

The vector in the second argument to `headtail()` is used to show only the two variables of interest.

Δ Explanatory Variable: The variable that we think may explain or allow us to predict the response variable. Synonyms are “independent” or “predictor” variable.

In the car data, the weight of the car may help explain the number of highway MPG of the car (e.g., a hypothesis might be that the heavier the car, the lower the MPG will be). Thus, the number of highway MPG is the response variable because it is the variable of primary interest and the variable we are trying to explain. The explanatory variable is the weight of the car as it will be used to explain the number of highway MPG.

Deciding which variable is the response variable often depends on the context of the situation. In the first example of bivariate data given above, the response variable may be weight if we are interested in predicting weight from height or it may be height if we are interested in predicting height from weight². The researcher (you) will identify the context of the problem. In the four previous examples, the response and explanatory variables are as follows (followed by context notes):

- R = weight, E = height [want to predict weight (hard to measure) from height (easy to measure)].
- R = area, E = depth [area is hard to measure, depth is easy].
- CAN’T DO, both variables are categorical.
- R = number of mice in a field, E = biomass of legumes in the field [hypothesized that higher biomass leads to more mice].

◊ Which variable is the response variable depends on the context of the problem or the researcher’s needs (i.e., which variable is being explained or predicted).

A scatterplot is a graph of points where each point simultaneously represents the values of both the response and explanatory variable. The value of the explanatory variable gives the x-coordinate and the value of the response variable gives the y-coordinate of the point plotted for an individual. For the first individual of the **cars93** data, a point would be placed at x (*Weight*) = 2705 and y (*HMPG*) = 31. For the second individual, a point would be placed at x=3560 and y=25. The scatterplot for all individuals in the data file is shown in Figure 8.1.

◊ Response variables are plotted on the y-axis and explanatory variables are plotted on the x-axis.

8.1.1 Scatterplots in R

Scatterplots are constructed in R with `plot()`. This function requires a model formula as the first argument³ followed by the dataframe name in `data=`. This model formula is of the form `Y~X` where Y and X are vectors of quantitative data to be plotted on the y- and x-axes, respectively. As with histograms, the x- and y-axis labels are modified with the `xlab=` and `ylab=` arguments. The scatterplot of highway MPG versus car weight (Figure 8.2) was created with

²

The latter is usually not the case, though.

³

This function can also take the vector of x-axis data as its first argument followed by a vector of y-axis data as its second argument. The formula notation is preferred for ease of transferability to other functions.

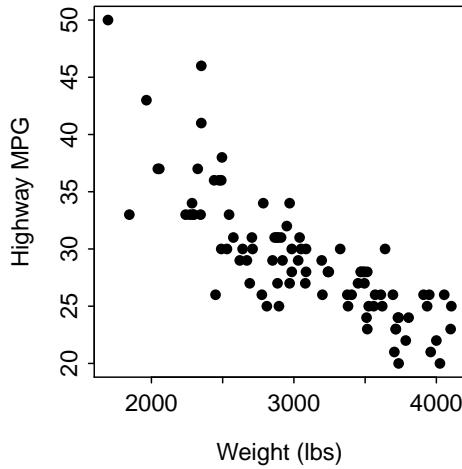


Figure 8.1. Scatterplot between the highway MPG and weight of cars manufactured in 1993.

```
> plot(HMPG~Weight,data=cars93,ylab="Highway MPG",xlab="Weight (lbs)")
```

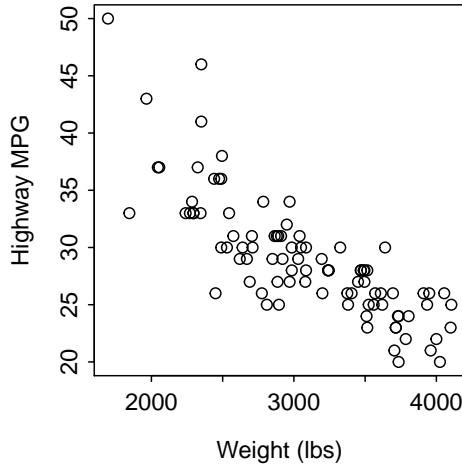


Figure 8.2. Scatterplot between the highway MPG and weight of cars manufactured in 1993 (using R default values)

The character plotted at each point can be changed with the `pch=` argument⁴. This argument defaults to a value of 1 which is an open-circle. Numerical values used to represent other plotting characters are shown in Figure 8.3. For example, the scatterplot shown in (Figure 8.1) was created with

```
> plot(HMPG~Weight,data=cars93,ylab="Highway MPG",xlab="Weight (lbs)",pch=16)
```

⁴

This argument is short for “plotting character”.

| | | | | |
|------|------|------|------|------|
| □ 0 | ○ 1 | △ 2 | + 3 | × 4 |
| ◇ 5 | ▽ 6 | ⊗ 7 | * 8 | ◊ 9 |
| ⊕ 10 | ⊗ 11 | 田 12 | ⊗ 13 | □ 14 |
| ■ 15 | ● 16 | ▲ 17 | ◆ 18 | ● 19 |
| ● 20 | ○ 21 | □ 22 | ◇ 23 | △ 24 |
| | | | | ▽ 25 |

Figure 8.3. Plotting characters available in R and their numerical codes.

A scatterplot with different plotting characters for individuals from different groups is obtained with the `col=` argument or `pch=` coupled with the `as.numeric()` argument and a factor variable. The `as.numeric()` argument converts a factor variable to a numeric variable where the first factor is listed with a “1”, the second with a “2”, and so on. If the result from `as.numeric()` is assigned to the `pch=` argument, then these numbers will serve to identify different plotting characters for the different levels of the factor variable. Of course, this plot should have a legend to this plot, which is added with `legend()`. This function requires a position for the legend as the first argument, names for the levels in the `legend=` argument, and the same `pch=` argument as used in `plot()` except that the data frame from which the factor comes from must be explicitly stated. For example, the plot of highway MPG versus weight separated by the type of the vehicle (Figure 8.4) is constructed with

```
> plot(HMPG~Weight,data=cars93,ylab="Highway MPG",xlab="Weight (lbs)",pch=as.numeric(Type))
> legend("topright",legend=levels(cars93$Type),pch=1:length(levels(cars93$Type)),cex=0.75)
```

8.2 Items to Describe I

Four characteristics should be described when exploring bivariate data with a scatterplot,

1. **Direction** of the relationship, or the association between the variables.
2. **Form** of the relationship.
3. **Strength** of the relationship.
4. Presence or absence of **outliers**.

All four of these items can be described from the examination of a scatterplot. It should be noted, though, that the strength of the relationship is best described with the correlation coefficient (see Section 8.3).

Association is a general statement about the direction of the relationship. Three general statements of association are used – positive, negative, and none. A positive association is when the scatterplot resembles an increasing function – i.e., increases from lower-left to upper-right (Figure 8.5-Right). For a positive association, most of the individuals are simultaneously above average or below average for both of the variables. A negative association is when the scatterplot looks like a decreasing function – i.e., decreases from upper-left to lower-right (Figure 8.5-Left). For a negative association, most of the individuals are simultaneously above average for one variable and below average for the other variable. No association is

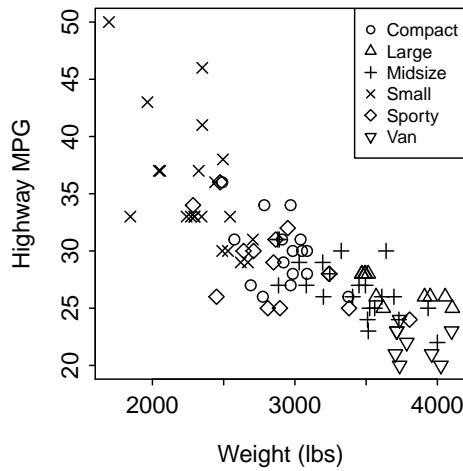


Figure 8.4. Scatterplot of highway MPG versus weight separated by the six types of vehicle.

when the scatterplot looks like a flat horizontal line or a “shotgun blast” of points (Figure 8.5-Middle). For no association, there are no tendencies for individuals to be above or below average for one variable and above or below average for the other.

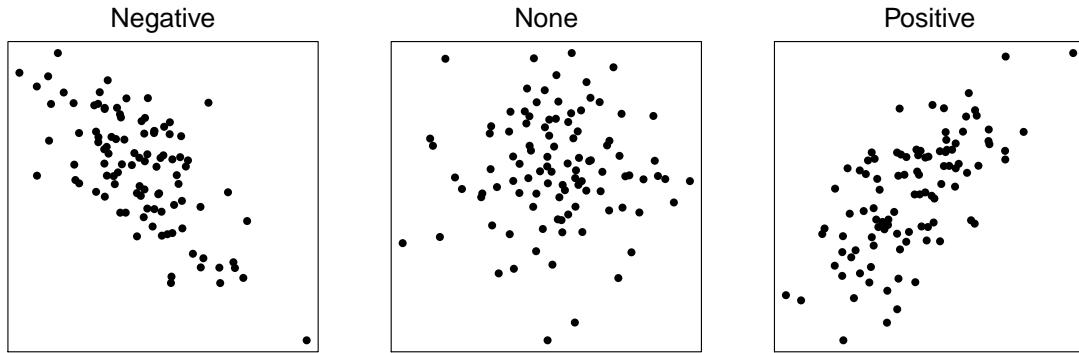


Figure 8.5. Depiction of three types of association present in scatterplots.

△ **Positive Association:** Most of the individuals are either above average or below average for both of the variables.

△ **Negative Association:** Most of the individuals are above average for one variable and below average for the other variable.

△ **No Association:** There are no tendencies for individuals to be above or below average for one variable and above or below average for the other.

For the purposes of this introductory text, form will be defined very generally with only two types considered – straight line and curved. The positive and negative association scatterplots in Figure 8.5 are two examples

of a straight line shape. In general, the form should be obviously curved before describing it as curved.

Strength is a summary of how closely the points cluster about the general form of the relationship. For example, in a straight-line form it would be how closely the points cluster around the line. Strength is difficult to define from a scatterplot because it is a relative term. The general idea of strength is depicted in Figure 8.6. However, an objective numerical measure – the correlation coefficient – will be defined in Section 8.3.

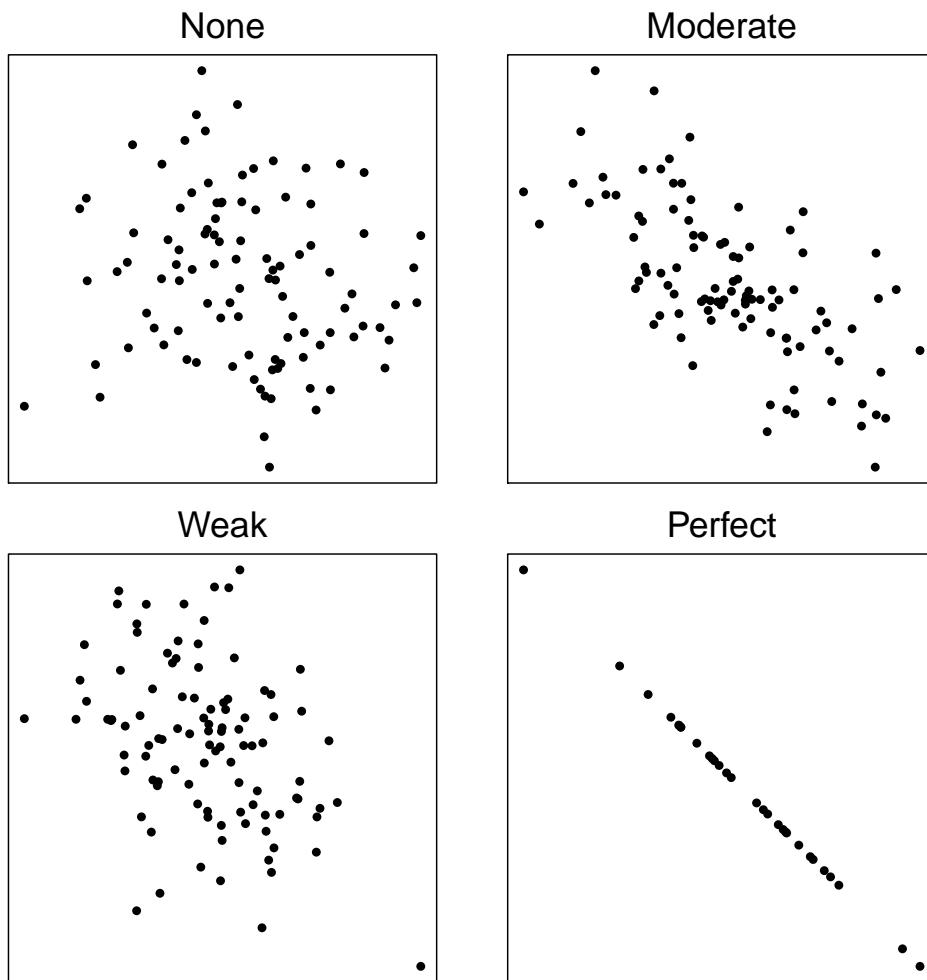


Figure 8.6. Scatterplots depicting four relatives types of strength.

△ **Strength:** How closely the points cluster about the general form of the relationship.

◊ **Strength can only be subjectively described from a scatterplot; use the correlation coefficient to be more objective.**

Outliers are points that are far removed from the main cluster of points. Keep in mind (as always) that just because a point is an outlier doesn't mean it is wrong.

The relationship between highway MPG and the weight of cars (Figure 8.1) appears to be negative, primarily linear (although I see a very slight concavity), and moderately strong. The three points at (2400,46), (2500,27), and (1800,33) might be considered SLIGHT outliers (these are not far enough removed for me to consider them outliers, but some people may).

A general conclusion that could be made from these results is that as the weight of the cars increases, the highway MPG attained by the car decreases in a linear fashion. While this conclusion is correct, it is also very carefully worded. We must be very careful to not state that increasing the weight of the car CAUSES a decrease in MPG. We cannot attribute cause because these data come from an observational study and because several other important variables were not considered in the analysis. For example, the scatterplot in Figure 8.7, coded for different numbers of cylinders in the car's engine, indicates that the number of cylinders may be inversely related to the highway MPG and positively related to the weight of the car. So, does the weight of the car, the number of cylinders, or both, significantly explain the decrease in highway MPG?

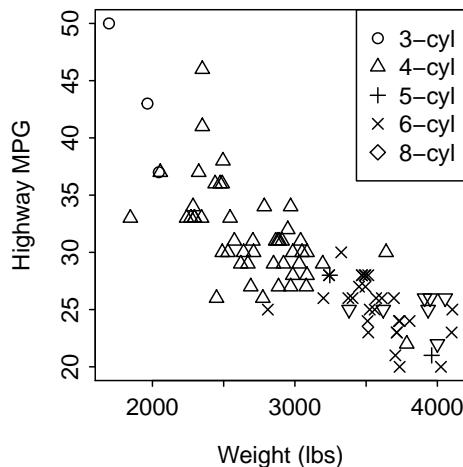


Figure 8.7. Scatterplot between the highway MPG and weight of cars manufactured in 1993 separated by number of cylinders.

Review Exercises

- 8.1** Researchers in Northern Wisconsin wanted to explain the role of the whitetail deer as a keystone herbivore (Waller and Alverson 1997). As a part of their analysis, they examined the relationship between the mean number of hemlock saplings on 14×21 m sections of a woodlot and a browsing index (a complicated measurement that gives the amount of food a deer has been eating in a given area). Use the data in the table below to make a scatterplot of the mean number of hemlock saplings versus the browsing index and describe the bivariate relationship from it. [Answer](#)

| | |
|---------------------------|---|
| mean no. hemlock saplings | 0.95 2.89 2.97 3.94 4.74 5.10 6.64 7.13 |
| browse index | 0.31 0.35 0.49 0.50 0.61 0.63 0.86 0.90 |

8.3 Correlation

The sample correlation coefficient, abbreviated as r , is calculated with

$$r = \frac{\sum_{i=1}^n \left[\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \right]}{n - 1} \quad (8.3.1)$$

where s_x and s_y are the sample standard deviations⁵ for the explanatory and response variable, respectively. The formulas in the two sets of parentheses in the numerator are standardized values⁶; thus, the value in each parenthesis is called the standardized x or standardized y, respectively. Using this terminology, the formula for the correlation coefficient reduces to these steps:

1. For each individual, standardize x and standardize y.
2. For each individual, find the product of the standardized x and standardized y.
3. Sum all of the products from step 2.
4. Divide the sum from step 3 by $n-1$.

◊ The sample correlation coefficient is abbreviated with r . The population correlation coefficient is abbreviated with ρ .

The table below illustrates these calculations for the first five individuals in the **cars93** data set (the five individuals are treated as if they are the entire sample). In the table note that the “i” column is an index for each individual, the x_i and y_i columns are the observed values of the two variables for individual i , \bar{x} was computed by dividing the sum of the x_i column by n , s_x was computed by dividing the sum of the $(x_i - \bar{x})^2$ column by $n - 1$ and taking the square root, and the “std x” column is the standardized x values found by dividing the value in the $x_i - \bar{x}$ column by s_x . Similar calculations were made for the y variable. The final correlation coefficient is the sum of the last column divided by $n - 1$. Thus, the correlation between car weight and highway mpg for these five cars is -0.54.

| | | HMPG | Weight | | | | | | | |
|-----|-------|-------|-----------------|-----------------|---------------------|---------------------|--------|--------|------------------|--|
| i | y_i | x_i | $y_i - \bar{y}$ | $x_i - \bar{x}$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x})^2$ | std. y | std. x | (std. y)(std. x) | |
| 1 | 31 | 2705 | 3.4 | -632 | 11.56 | 399424 | 1.26 | -1.71 | -2.15 | |
| 2 | 25 | 3560 | -2.6 | 223 | 6.76 | 49729 | -0.96 | 0.6 | -0.58 | |
| 3 | 26 | 3375 | -1.6 | 38 | 2.56 | 1444 | -0.59 | 0.1 | -0.06 | |
| 4 | 26 | 3405 | -1.6 | 68 | 2.56 | 4624 | -0.59 | 0.18 | -0.11 | |
| 5 | 30 | 3640 | 2.4 | 303 | 5.76 | 91809 | 0.89 | 0.82 | 0.73 | |
| sum | 138 | 16685 | 0 | 0 | 29.2 | 547030 | 0 | 0 | -2.17 | |

You should note that there are easier formulas for calculating the correlation coefficient than that illustrated above. However, the formula and method above illustrates some intuitive concepts to be discussed next.

⁵

See Section 5.6.3 for a review of standard deviations.

⁶

See Section 7.5 for a review of standardized values.

The correlation coefficient is a measure of both association and strength. The sign of r indicates the direction or association between the two variables. A positive r means a positive association and a negative r means a negative association. The absolute value of r (i.e., the value ignoring the sign) is an indicator of the strength of relationship. Absolute values nearer 1 are stronger relationships. Each of these concepts is discussed further in the following paragraphs.

A positive association occurs when both variables measured on an individual tend to be above or below average together. To illustrate this concept, examine the scatterplot in Figure 8.8-Left that has superimposed lines at the means of both the x and y variables. The sign of a standardized value for a measurement larger than the mean is positive, because the difference between the larger observed value and the mean is positive. With similar reasoning, the sign of the standardized value for a measurement smaller than the mean is negative.

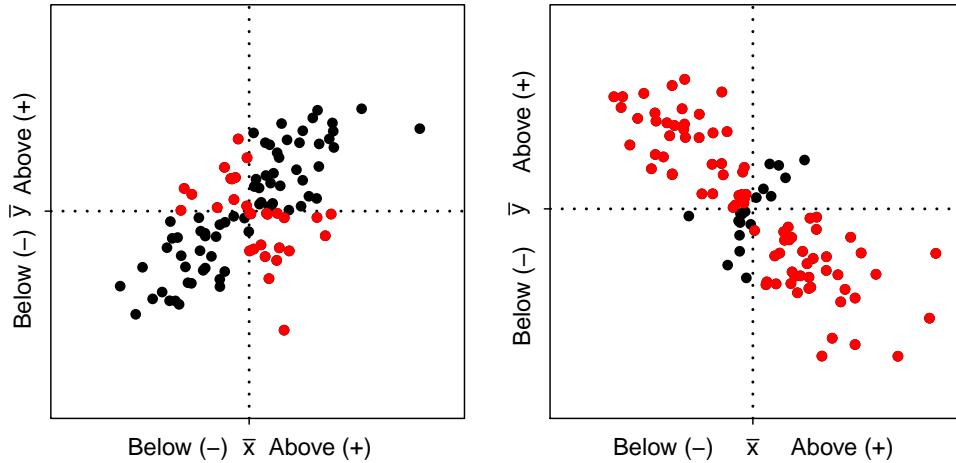


Figure 8.8. Scatterplot with mean lines superimposed and the signs of standardized values for both x and y shown for a positive (Left) and negative (Right) association.

Now consider the product of standardized x 's and y 's in each quadrant of Figure 8.8-Left. In the quadrant that corresponds to above average for both standardized values (i.e., both positive signs) the product is positive (denoted by black dots). In the quadrant that corresponds to below average for both standardized values the product is also positive. In the other two quadrants the product is negative (denoted by red dots). From Figure 8.8-Left it is seen that, for a positive association, the numerator of the correlation coefficient is the sum of many positive products of standardized x 's and y 's (black dots) and few negative products (red dots). Thus, the numerator is positive. The denominator (recall it is $n-1$) is always positive. Thus, the correlation for a positive association is positive.

A negative association is examined in the same manner (Figure 8.8-Right). The signs of the products in the quadrants are the same as described above. With the negative association, the numerator is the sum of many negative numbers (red dots) and a few positive numbers (black dots). Thus, the numerator is negative. Therefore, the correlation for a negative association is negative.

- ◊ The correlation coefficient is positive for positive associations and negative for negative associations.

Correlations range from -1 to 1. Absolute values of r equal to 1 indicate a perfect correlation; i.e., all points fall exactly on a line. A correlation of 0 indicates no association. Thus, absolute values of r near 1 indicate strong relationships and those near 0 are weak. The range of correlation values and a few scatterplots illustrating how the strength and direction of the relationship between two variables changes along this scale is illustrated in Figure 8.9. The categorizations in Table 8.1 can be used as a rough guideline for categorizing the strength of a relationship between two variables.

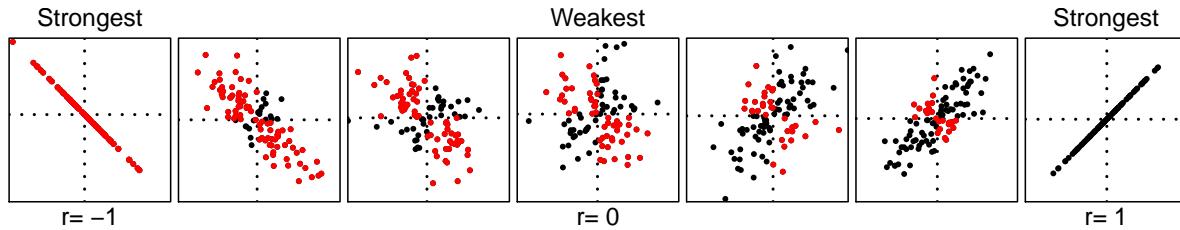


Figure 8.9. Scatterplots along the continuum of r values.

Table 8.1. Classifications of strength of relationship for absolute values of r by type of study.

| Strength of Relationship | Uncontrolled/ Observational | Controlled/ Experimental |
|--------------------------|--------------------------------|-----------------------------|
| Strong | > 0.8 | > 0.95 |
| Moderate | > 0.6 | > 0.9 |
| Weak | > 0.4 | > 0.8 |

◊ **Absolute values of correlation coefficients nearer one are stronger.**

You should practice estimating the strength and direction of a relationship by simply looking at scatterplots. Johnson and Kuby (2000) provide a five step graphical procedure for estimating the correlation coefficient from a scatterplot:

1. Place two lines on the scatterplot that are parallel to the direction of the relationship and as close together as possible while still containing all of the points.
2. Visualize a rectangular region that is bounded by the two lines from step 1 and has ends just beyond the points on the scatterplot.
3. Estimate how many times longer the rectangle is than it is wide; call this value k . An easy way to do this is to mentally mark off and then count squares in the rectangle.
4. Estimate $|r| = 1 - \frac{1}{k}$.
5. Assign a sign to r based on the direction of the association.

The Johnson and Kuby (2000) procedure for estimating r is illustrated in Figure 8.10. In the right-most figure it appears that the length of the rectangle is about 3 times as long as the width. This corresponds to an estimated correlation coefficient⁷ of $1 - \frac{1}{3} = 0.67$. Note that the actual correlations is 0.708.

⁷

r is positive because the relationship has a positive association.

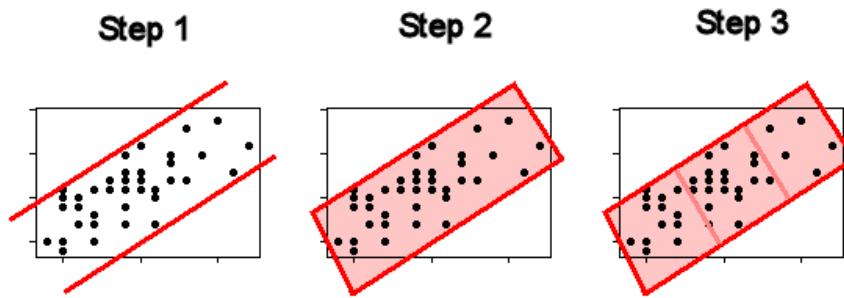


Figure 8.10. Depiction of the steps for estimating r from Johnson and Kuby (2000).

The following are important characteristics of correlation coefficients:

- The variables must be quantitative (i.e., if you should not make a scatterplot, then don't calculate r).
- Correlation only measures the strength of relationships that are linear (i.e., if the form of the relationship is not linear, then the correlation calculation is meaningless – MORAL, graph your data).
- The units that the variables are measured in do not matter (i.e., if you are looking at height and weight you will calculate the same r if the data were measured in inches and lbs, inches and kg, m and kg, cm and kg, cm and inches, etc.). This is because of the standardization of the two variables in the calculations.
- The distinction between response and explanatory variables is not needed. That is, the correlation of GPA and ACT scores is the same as the correlation of ACT scores and GPA.
- Correlations are between -1 and 1.
- Correlations are strongly affected by outliers (simply, because both the mean and standard deviation, used in the calculation of r , are strongly affected by outliers).
- Correlation is not causation – just because a strong correlation is observed it doesn't mean that the explanatory variable caused the response variable (an exception may be in carefully designed experimental studies).

◊ The word “correlation” is often mis-used in everyday language. This word is used only when discussing the actual correlation coefficient (i.e., r). When discussing the association between two variables, one should use the word “relationship” rather than “correlation” (e.g., “What is the relationship between age and rate of cancer?”).

8.3.1 Correlations in R

The correlation coefficient (r) between two quantitative variables is computed with `cor()`. When only two variables are considered, `cor()` requires only two arguments – vectors containing the two quantitative variables. Note that these two vectors must be of the same length. For example, the correlation between highway MPG and weight of the car is found with

```
> cor(cars93$HMPG, cars93$Weight)
[1] -0.8106581
```

The correlation coefficient can be simultaneously computed among many pairs of quantitative variables found in a data frame if that data frame is the only argument sent to the `cor()` function. As noted above r is

calculated only with quantitative data. Thus, all variables in the data frame must be quantitative. For example, if one wants to find the correlations between each pair of highway MPG, size of the fuel tank, length, and weight of cars in the `cars93` data frame, then these variables must be isolated and assigned to a new data frame as follows,

```
> cars93a <- cars93[,c("HMPG", "FuelTank", "Length", "Weight")]
> str(cars93a)
'data.frame': 93 obs. of 4 variables:
 $ HMPG    : int  31 25 26 30 31 28 25 27 25 ...
 $ FuelTank: num  13.2 18 16.9 21.1 21.1 16.4 18 23 18.8 18 ...
 $ Length   : int  177 195 180 193 186 189 200 216 198 206 ...
 $ Weight   : int  2705 3560 3375 3405 3640 2880 3470 4105 3495 3620 ...
```

In some instances, the data frame may contain some missing values (i.e., data that was not recorded). The individuals with missing pieces of data are efficiently removed when computing the correlation coefficient by including the `use="pairwise.complete.obs"` argument to `cor()`. Thus, the correlations between these four variables is obtained with

```
> cor(cars93a, use="pairwise.complete.obs")
      HMPG   FuelTank   Length   Weight
HMPG  1.0000000 -0.7860386 -0.5428974 -0.8106581
FuelTank -0.7860386  1.0000000  0.6904612  0.8940181
Length  -0.5428974  0.6904612  1.0000000  0.8062743
Weight  -0.8106581  0.8940181  0.8062743  1.0000000
```

These results are a so-called correlation matrix where each cell in the matrix represents the r between the variables that label the corresponding row and column. Thus, the correlation between highway MPG and size of the fuel tank is -0.79. The correlation matrix has all 1s on the main diagonal because the correlation between a variable and itself is always 1 (i.e., a perfect relationship). In addition, the matrix is symmetric about the main diagonal because the correlation between X and Y is the same as the correlation between Y and X .

- ◊ If the vector submitted to `cor()` has missing data, then the individuals with missing data should be excluded by including the `use="pairwise.complete.obs"` argument in `cor()`.

A visual that corresponds with the correlation matrix is the scatterplot matrix. A scatterplot matrix is a graphic that contains scatterplots of all possible pairs of variables in one plot (Figure 8.11). Each subplot in the scatterplot matrix is a scatterplot with the variable listed in the same column on the x-axis and the variable listed in the same row on the y-axis. For example, the scatterplot in the upper-right corner of Figure 8.11 has highway MPG on the y-axis and car weight on the x-axis. A scatterplot matrix is constructed in R by submitting the “reduced” data frame to `pairs()`. For example, the scatterplot matrix in Figure 8.11 was constructed with

```
> pairs(cars93a)
```

8.3.2 Example - Highway MPG and Weight

The following overall bivariate summary for the relationship between highway MPG and weight is made from the analyses in the previous sections. The relationship between highway MPG and the weight of cars (Figure

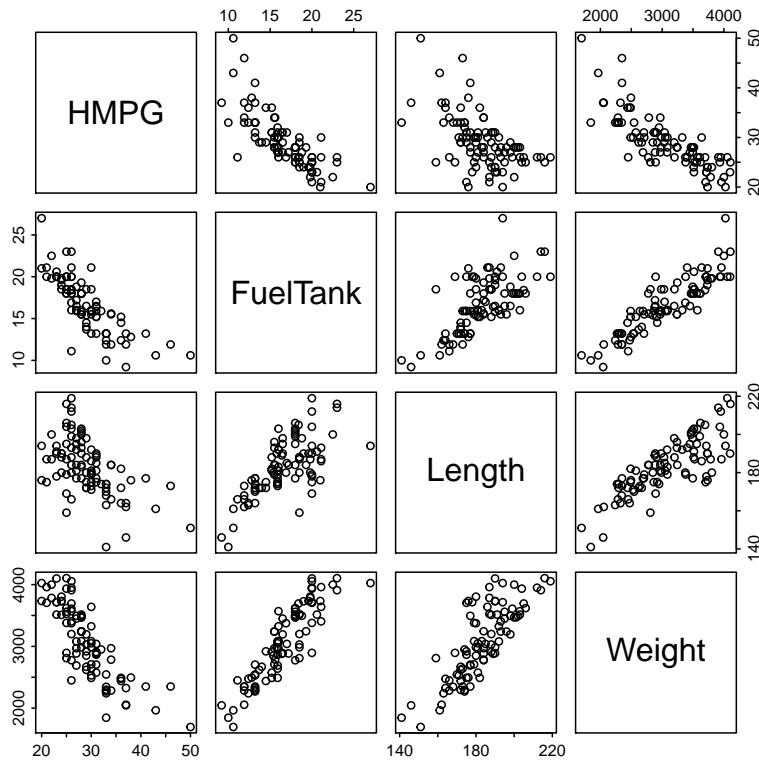


Figure 8.11. Scatterplot matrix of the highway MPG, fuel tank size, length, and weight of cars.

8.1) appears to be negative, primarily linear (although I see a very slight concavity), and moderately strong with a correlation of -0.79. The three points at (2400,46), (2500,27), and (1800,33) might be considered SLIGHT outliers (these are not far enough removed for me to consider them outliers, but some people may).

8.3.3 Example - State Energy Usage

A 2001 report from the [Energy Information Administration](#) of the Department of Energy details the total consumption of a variety of energy sources by state in 2001. Construct a proper EDA for the relationship between total petroleum and coal consumption (in trillions of BTU).

The relationship between total petroleum and coal consumption is generally positive, linear, weak, with two outliers at total petroleum levels greater than 3000 trillions of BTU (Figure 8.12-Left). I did not compute a correlation coefficient because of the outliers. The two outliers were Texas and California. After removing them from the data set the relationship is clearly positive, linear, weak ($r = 0.53$), with no additional outliers (Figure 8.12-Right).

The relationship between the consumption of petroleum and coal exhibits two outliers – Texas and Oklahoma. The relationship not considering these two data points appears to be positive, linear, and weak ($r=0.53$).

This example illustrates a few key points in the description of a bivariate EDA. First, the descriptions of association, strength, and form should not be influenced by the presence of outliers. In other words, describe

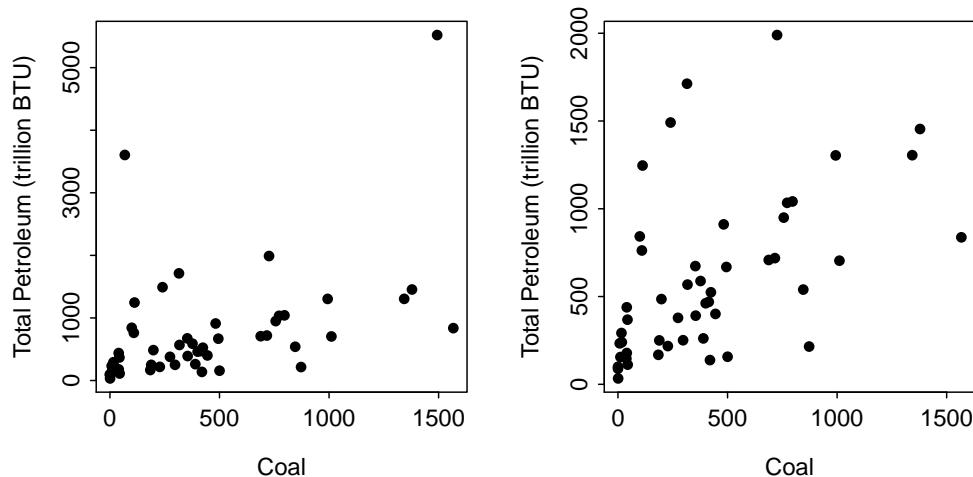


Figure 8.12. Scatterplot of the total consumption of petroleum versus the consumption of coal (in trillions of BTU) by all 50 states and the District of Columbia. The points shown in the left with total petroleum values greater than 3000 trillion BTU are deleted in the right plot.

association, strength, and form ignoring any outliers present in the data. If you don't have the ability to compute r without the outliers (e.g., you are just given r for the entire data set), then **DO NOT** report r because it is too strongly influenced by the outliers. Second, the form of weak relationships is difficult to describe because, by definition in a weak relationship, there is very little clustering to a form. As a rule-of-thumb, if the scatterplot does not have an obvious curvature to it, then it is described as linear by default.

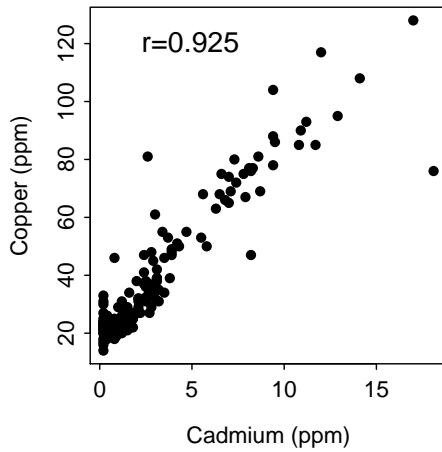
◇ Outliers should not influence the descriptions of association, strength, and form.

Review Exercises

- 8.2** Calculate the correlation coefficient between the mean number of hemlock saplings and deer browse index given in Review Exercise 8.1. [Answer](#)
- 8.3** The concentration of cadmium and copper in the topsoil of 115 15mX15m plots along the river Meuse in the village Stein in New Zealand was recorded by van Rijn and Rikken⁸. Use the scatterplot below to describe the bivariate relationship between these two variables. [Answer](#)

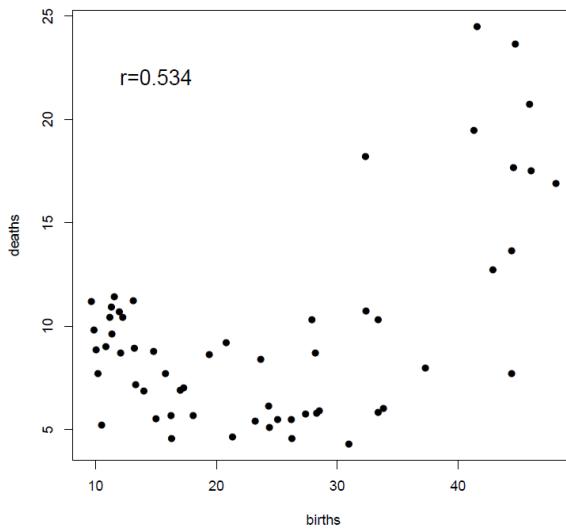
⁸

These data are available in `data(meuse)` of the `sp` package.



- 8.4** Ten variables were measured on 57 countries and reported in the International Vital Statistics (1996). A scatterplot of the birth and death rates is shown below. Write a brief description of this bivariate relationship.

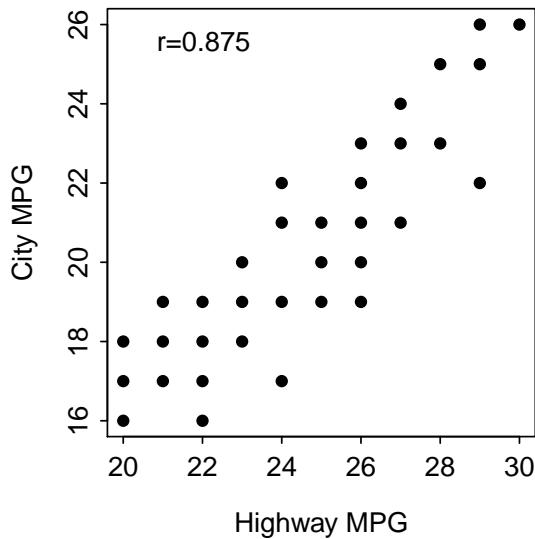
[Answer](#)



- 8.5** Allen *et al.* (1997) investigated the impact of the density of red-imported fire ants (RIFA) on the recruitment of white-tailed deer fawns (an index of does to fawns). A modified version of their data is recorded in [RIFA.csv](#). Use this information to write a brief description of this bivariate relationship.

[Answer](#)

- 8.6** Researchers at Chevrolet attempted to determine the relationship between gas mileage (MPG) of Luminas in the city (CITY) and on the highway (HIGHWAY). Their results are shown below. Use this information to write a brief description of this bivariate relationship. [Answer](#)

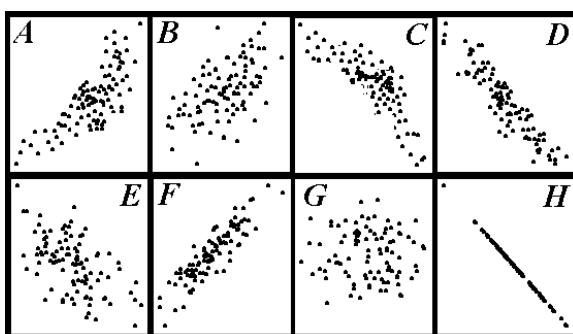


- 8.7** Mladenoff *et al.* (1997) estimated the territory size (km^2) of wolf (*Canis lupus*) packs and the density of whitetail deer (number/ km^2 ; *Odocoileus virginianus*) in the same areas in northern Wisconsin. Their data is recorded in [Wolves2.csv](#). Load these data into R and generate results to write a brief description of this bivariate relationship. [Answer](#)

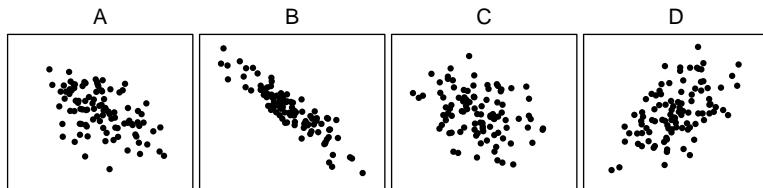
- 8.8** The Park Management team of Kejimkujik National Park, Nova Scotia examined the relationship between the length and weight of yellow perch (*Perca flavescens*) captured from Grafton Lake in the park in 2000 following the removal of a dam (Brylinsky 2001). Their data is stored in [PerchGL.csv](#). Load these data into R, isolate just the results from 2000 (i.e., use `filterD()`), and generate results to describe this bivariate relationship. [Answer](#)

- 8.9** It has been said that you can roughly estimate the temperature from the number of cricket chirps heard. To determine if this relationship existed, an entomologist recorded the number of chirps in a 15-second interval by crickets held at different temperatures. The researcher's data is recorded in [Chirps.csv](#). Load these data into R and generate results to write a brief description of this bivariate relationship. [Answer](#)

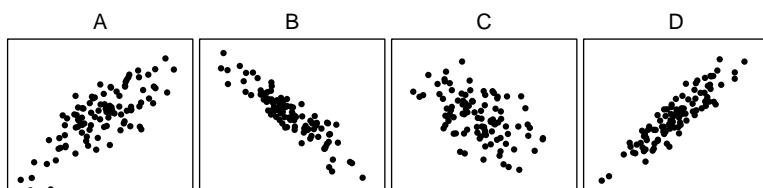
- 8.10** Five of the scatterplots below correspond to the following correlation coefficients — 0.89, -0.48, -0.92, 0.56, 0.00. Identify the scatterplot that each correlation corresponds to. Some scatterplots will not be used. [Answer](#)



8.11 Order the following graphs from (i) lowest to highest value of r and (ii) weakest to strongest. [Answer](#)



8.12 Order the following graphs from (i) lowest to highest value of r and (ii) weakest to strongest. [Answer](#)



CHAPTER 9

BIVARIATE EDA - CATEGORICAL

Chapter Objectives:

1. Describe what bivariate data is.
2. Distinguish between response and explanatory variables.
3. Construct two-way contingency tables from raw data.
4. Identify marginal distributions.
5. Construct row-, column-, and table-percentage tables from two-way tables.
6. Interpret two-way contingency tables.

Contents

| | | |
|-----|-----------------------------|-----|
| 9.1 | Frequency Tables | 117 |
| 9.2 | Percentage Tables | 118 |
| 9.3 | Which Table? | 121 |
| 9.4 | Tables in R | 123 |

TWO-WAY FREQUENCY TABLES summarize two categorical variables recorded simultaneously on the same individual by displaying the categories of the first variable as rows and the categories of the second variable as columns. Each cell in this table contains a count of the number of individuals that were in the corresponding categories of each variable. Frequency tables are often converted to percentage tables for ease of summarization and comparison between populations. This chapter explores the construction and interpretation of these types of tables.

Throughout this chapter we will consider the following data from the General Sociological Survey (GSS). Two questions asked to 3955 respondents for the GSS were,

- What is your highest degree earned? [choices – “less than high school diploma”, “high school diploma”, “junior college”, “bachelors”, or “graduate”; labeled as *degree*]
- How willing would you be to accept cuts in your standard of living in order to protect the environment? [choices – “very willing”, “fairly willing”, “neither willing nor unwilling”, “not very willing”, or “not at all willing”; labeled as *grnsol*]

These data, stored in [GSSWill2Pay.csv](#), are loaded into R, with the structure and a random six individuals were observed with

```
> gss <- read.csv("data/GSSWill2Pay.csv")
> str(gss)
'data.frame': 3955 obs. of 2 variables:
 $ degree: Factor w/ 5 levels "BS","grad","HS",...: 5 5 5 5 5 5 5 5 5 ...
 $ grnsol: Factor w/ 5 levels "neither","un",...: 4 4 4 4 4 4 4 4 4 ...
> headtail(gss)
      degree grnsol
1      lthS    vwill
2      lthS    vwill
3      lthS    vwill
3953   grad     vun
3954   grad     vun
3955   grad     vun
```

The *degree* and *grnsol* variables are both *ordinal* categorical variables. By default R orders the levels for a factor variable alphabetically. Alphabetically is not typically the “natural order” of the levels. Thus, it is important to identify the order of levels for each ordinal variable by submitting the variable name to `levels()`. For example, the levels for the *degree* and *grnsol* variable are seen with

```
> levels(gss$degree)
[1] "BS"    "grad"  "HS"    "JC"    "lthS"
> levels(gss$grnsol)
[1] "neither" "un"    "vun"   "vwill" "will"
```

It is seen that these ordinal variables do not have the levels in their proper order. The correct order is obtained by using `factor()` where the first argument is the name of the factor variable and a vector of correctly ordered levels is included in the `levels=` argument. The result of `factor()` should be set to a new variable in the data frame. For example, two new factor variables – *fdegree* and *fgrnsol* – with the correct order of levels are created with

```
> gss$fdegree <- factor(gss$degree, levels=c("ltHS", "HS", "JC", "BS", "grad"))
> gss$fgrnsol <- factor(gss$grnsol, levels=c("vwill", "will", "neither", "un", "vun"))
```

Now the levels of these two new variables are in the correct order as seen with

```
> levels(gss$fdegree)
[1] "ltHS" "HS"   "JC"   "BS"   "grad"
> levels(gss$fgrnsol)
[1] "vwill" "will" "neither" "un"   "vun"
```

If the variables had been nominal or if the natural order of levels were the same as the alphabetical order, then the use of `factor()` illustrated above would not have been needed.

- ◊ Levels for a factor variable are ordered alphabetically by default in R. You may need to use `factor()` with `levels=` to control the order of levels if the factor variable is ordinal.

9.1 Frequency Tables

A common method of summarizing bivariate categorical data is to count the number of individuals that have each combination of the first and second categorical variables. For example, how many respondents had less than a HS degree and were very willing, how many had a high school degree and were very willing, and so on. The count of the number of individuals of each combination is called a frequency. A two-way frequency table offers an efficient way to display these frequencies (Table 9.1). Thus, for example, 40 of the respondents had less than a high school degree and were very willing to take a cut in their standard of living to protect the environment. Similarly, 542 respondents had a high school degree and were willing to cut their standard of living.

Table 9.1. Frequency table of respondent's highest completed degree and willingness to cut their standard of living to protect the environment.

| | vwill | will | neither | un | vun | Sum |
|------|-------|------|---------|------|-----|------|
| ltHS | 40 | 145 | 132 | 151 | 178 | 646 |
| HS | 87 | 542 | 512 | 557 | 392 | 2090 |
| JC | 15 | 61 | 64 | 54 | 44 | 238 |
| BS | 42 | 199 | 179 | 187 | 75 | 682 |
| grad | 24 | 104 | 83 | 64 | 24 | 299 |
| Sum | 208 | 1051 | 970 | 1013 | 713 | 3955 |

A two-way frequency table is usually augmented with a column of row totals and a row of column totals (as in Table 9.1). This row and column is called the marginal row and the marginal column, respectively. Each marginal total represents the distribution of one of the categorical variables while ignoring the other categorical variable. The total column represents the distribution of the row variable; in this case, the highest degree completed. Thus, in this case the total column represents the number of respondents according to their highest degree completed. Similarly, the total row represents the distribution of the column variable. In this case, the total row represents the number of respondents according to their willingness to cut their standard of living to protect the environment. Thus, for example there were 238 respondents whose highest complete degree was junior college and there were 713 respondents who were very unwilling to cut their standard of living to protect the environment.

Review Exercises

- 9.1** A group of marine biologists from California wanted to study the foraging ecology of northern elephant seals off the California coast (Le Boeuf *et al.* 2000). Part of their analysis required that they record, for each observed seal, the month that it was observed and the sex of the seal. Their results from 47 seals are listed below. Construct a two-way frequency table, including the marginal totals, of these data with month as the column variable. [Answer](#)

| indiv | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Mon | Jun | Jul | Aug | |
| Sex | M | M | M | M | M | M | M | M | M | M | M | M | M | M | M | |
| indiv | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
| Mon | Aug | Jun | Jun | Jun | Jun | Jun |
| Sex | M | M | M | M | M | M | M | M | M | M | M | F | F | F | F | F |
| indiv | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | |
| Mon | Jul | Jul | Aug | |
| Sex | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | |

9.2 Percentage Tables

Two-way frequency tables are often converted to two-way percentage tables to allow for ease of comparison between levels of the variables and also between samples. For example, it is difficult to determine from Table 9.1 if respondents with a high school degree are more likely to be very willing to cut their standard of living than respondents with a graduate degree, because there are approximately seven times as many respondents with a high school degree in the sample. This comparison is easily made, however, if the frequencies are converted to percentages. Three types of percentage tables are constructed from a raw frequency table. These are discussed in the next sections.

9.2.1 Row-Percentage Table

A **row-percentage table** is computed by dividing the value in the SAME cell on the original frequency table by the TOTAL value in the same row on the original frequency table and multiplying by 100. For example, the value in the “vwill” column and “ltHS” row on the row-percentage table (Table 9.2) is computed by dividing the value in the “vwill” column and “ltHS” row on the original frequency table (i.e., 40; Table 9.1) by the value in the “Sum” column and the “ltHS” row on the original frequency table (i.e., 646) and multiplying by 100.

The value in each cell of a row-percentage table is the percentage OF ALL individuals with the characteristic of that row that also have the characteristic of that column. For example, 6.2% of the respondents with less than a high school degree are very willing to cut their standard of living to protect the environment.

Table 9.2. Row-percentage table of respondent's highest completed degree and willingness to cut their standard of living to protect the environment.

| | vwill | will | neither | un | vun | Sum |
|------|-------|------|---------|------|------|-------|
| ltHS | 6.2 | 22.4 | 20.4 | 23.4 | 27.6 | 100.0 |
| HS | 4.2 | 25.9 | 24.5 | 26.7 | 18.8 | 100.0 |
| JC | 6.3 | 25.6 | 26.9 | 22.7 | 18.5 | 100.0 |
| BS | 6.2 | 29.2 | 26.2 | 27.4 | 11.0 | 100.0 |
| grad | 8.0 | 34.8 | 27.8 | 21.4 | 8.0 | 100.0 |

This needs to be read very closely and literally. OF THE RESPONDENTS WITH LESS THAN A HIGH SCHOOL DEGREE, not of all respondents, 6.2% were very willing to cut their standard of living.

- ◊ Each value in a row-percentage table is computed by dividing the value in the SAME cell on the original frequency table by the TOTAL value in the same row on the original frequency table and multiplying by 100.
- ◊ The value in each cell of a row-percentage table is the percentage OF ALL individuals with the characteristic of that row that also have the characteristic of that column.

9.2.2 Column-Percentage Table

A **column-percentage table** is computed by dividing the value in the SAME cell on the original frequency table by the TOTAL value in the same column on the original frequency table and multiplying by 100. For example, the value in the “vwill” column and “ltHS” row on the column-percentage table (Table 9.3) is computed by dividing the value in the “vwill” column and “ltHS” row on the original frequency table (i.e., 40; Table 9.1) by the value in the “Sum” row and the “vwill” column on the original frequency table (i.e., 208) and multiplying by 100.

Table 9.3. Column-percentage table of respondent's highest completed degree and willingness to cut their standard of living to protect the environment.

| | vwill | will | neither | un | vun |
|------|-------|-------|---------|-------|-------|
| ltHS | 19.2 | 13.8 | 13.6 | 14.9 | 25.0 |
| HS | 41.8 | 51.6 | 52.8 | 55.0 | 55.0 |
| JC | 7.2 | 5.8 | 6.6 | 5.3 | 6.2 |
| BS | 20.2 | 18.9 | 18.5 | 18.5 | 10.5 |
| grad | 11.5 | 9.9 | 8.6 | 6.3 | 3.4 |
| Sum | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

The value in each cell of a column-percentage table is the percentage OF ALL individuals with the characteristic of that column that also have the characteristic of that row. For example, 19.2% of respondents who were very willing to cut their standard of living had less than a high school degree. Again, this is a very literal statement. OF THE RESPONDENTS WHO WERE VERY WILLING TO CUT THEIR STANDARD OF LIVING, not of all respondents, 19.2% had less than a high school degree.

- ◊ Each value in a column-percentage table is computed by dividing the value in the SAME cell on the original frequency table by the TOTAL value in the same column on the original frequency table and multiplying by 100.
- ◊ The value in each cell of a column-percentage table is the percentage OF ALL individuals with the characteristic of that column that also have the characteristic of that row.

9.2.3 Table-Percentage Table

Each value in a **table-percentage table** is computed by dividing the value in the SAME cell on the original frequency table by the TOTAL number of ALL individuals in the original frequency table and multiplying by 100. For example, the value in the “vwill” column and “ltHS” row on the table-percentage table (Table 9.4) is computed by dividing the value in the “vwill” column and “ltHS” row on the original frequency table (i.e., 40; Table 9.1) by the value in the “Sum” row and the “Sum” column on the original frequency table (i.e., 3955) and multiplying by 100.

Table 9.4. Table-percentage table of respondent's highest completed degree and willingness to cut their standard of living to protect the environment.

| | vwill | will | neither | un | vun | Sum |
|------|-------|------|---------|------|------|-------|
| ltHS | 1.0 | 3.7 | 3.3 | 3.8 | 4.5 | 16.3 |
| HS | 2.2 | 13.7 | 12.9 | 14.1 | 9.9 | 52.8 |
| JC | 0.4 | 1.5 | 1.6 | 1.4 | 1.1 | 6.0 |
| BS | 1.1 | 5.0 | 4.5 | 4.7 | 1.9 | 17.2 |
| grad | 0.6 | 2.6 | 2.1 | 1.6 | 0.6 | 7.6 |
| Sum | 5.3 | 26.6 | 24.5 | 25.6 | 18.0 | 100.0 |

The value in each cell of a table-percentage table is the percentage OF ALL individuals that have the characteristic of that column AND that row. For example, 1.0% of ALL respondents had less than a high school degree AND were very willing to cut their standard of living to protect the environment. Compare this interpretation to the interpretations from the row and column-percentage tables above. This interpretation does refer to all respondents.

- ◊ Each value in a table-percentage table is computed by dividing the value in the SAME cell on the original frequency table by the TOTAL number of ALL individuals in the original frequency table and multiplying by 100.
- ◊ The value in each cell of a table-percentage table is the percentage OF ALL individuals that have the characteristic of that column AND that row.

Review Exercises

- 9.2 Construct a row-, column-, and table-percentage table from the frequency table for the seal data in Review

Exercise 9.1. [Answer](#)

9.3 Which Table?

Determining which table to use comes from applying one simple rule and practicing with several tables. The rule stems from determining if the question restricts the frame of reference to a particular level or category of one of the variables. If the question does restrict to a particular level, then either the row or column-percentage table that similarly restricts the frame of reference must be used. If a restriction to a particular level does not appear to be made, then the table-percentage table is used.

For example, consider the question – “What percentage of respondents with a bachelors degree were very unwilling to cut their standard of living to protect the environment?” This question restricts the frame of reference to respondents with bachelors degrees (i.e., “... of respondents with a bachelors degree ...”). Thus, the answer is restricted to the “BS” row of the frequency table. The ROW-percentage table restricts the original table to the row levels and is, thus, used to answer this question. Therefore, 11.0% of respondents with bachelors degrees were very unwilling to cut their standard of living to protect the environment (Table 9.2).

Now consider the question – “What percentage of all respondents had a high school degree and were very willing to cut their standard of living?” This question does not restrict the frame of reference because it refers to “... of all respondents ...”. Therefore, from the table-percentage table (Table 9.4), 2.2% of respondents had a high school degree and were very willing to cut their standard of living.

Also consider this question – “What percentage of respondents who were neither willing nor unwilling to cut their standard of living had graduate degrees?” This question restricts the frame of reference to respondents who were neither willing nor unwilling to cut their standard of living and, thus, restricts the question to the “neither” column of the original frequency table. Thus, the answer will come from the COLUMN-percentage table. Therefore, 8.6% of respondents who were neither willing nor unwilling to cut their standard of living had graduate degrees (Table 9.3).

Finally, consider this question – “What percentage of all respondents were very willing to cut their standard of living to help the environment?” This question has no restrictions on the frame of reference so the table-percentage table should be used. In addition, this question is only concerned with only one of the two variables in the two-way table; thus, the answer will come from one of the marginal distributions. Therefore, 208 out of all 3955 respondents, or 5.3%, were very willing to cut their standard of living to help the environment.

- ◊ To determine which percentage table to use determine what type of restriction, if any, has been placed on the frame of reference for the question.

- ◊ If a question does not refer to one of the two variables, then the answer will generally come from the marginal distribution of the other variable.

It should be noted that if one of the two categorical variables is determined to be a response variable, then

this variable is usually used to define the columns and the row-percentage table becomes the main table of interest. In this example, the “willingness to cut” would be considered the response variable and it was then, appropriately, placed as the column variable in the two-way table. Thus, the questions answered from the row-percentage table (i.e., “Of respondents with a certain degree ...”) make “more sense” than the questions answered from the column-percentage table (i.e., “Of respondents with a certain willingness ...”).

- ◊ The response variable is typically used to define the columns of the two-way table.

Review Exercises

9.3 Use the frequency and percentage tables for the seal data constructed in Review Exercises 9.1 and 9.2 to answer the questions below. [Answer](#)

- What percentage of elephant seals were male?
- What percentage of male elephant seals were observed in July?
- What percentage of elephant seals were observed in August?
- What percentage of elephant seals were females observed in July?

9.4 Weitz (1979) conducted a survey of general and family practitioners, pediatricians, and obstetrician-gynecologists in the cities of Phoenix and Tucson, Arizona. In one part of the study, each physician was classified according to religion and whether they supported genetic counseling for parents or not. A summary of their responses for Jewish, Protestant, and Catholic physicians is shown in the table below. Use these results to answer the questions below. [Answer](#)

- What percentage of Jewish physicians support genetic counseling?
- What percentage of Catholic physicians don't support genetic counseling?
- What percentage of all physicians surveyed were Protestant?
- What percentage of those physicians not supporting genetic counseling were Catholic?
- What percentage of all physicians supported genetic counseling?

| | Jewish | Protestant | Catholic |
|---------------|--------|------------|----------|
| Support | 21 | 36 | 10 |
| Don't Support | 26 | 142 | 52 |

9.5 The two-way table below depicts the results of an observational study concerned with the timing (i.e., month) of death for young herring gulls (after fledging) in three locations. Each cell in the table is the number of dead herring gulls in each month-location combination. Use the table to answer the questions below. [Answer](#)

- What percentage of the gulls that died in New Jersey died in July?
- What percentage of all gulls died in July?
- What percentage of all gulls died in September and in The Netherlands?

| Month | Location | | | Total |
|-------|------------|-------------|---------|-------|
| | New Jersey | Netherlands | England | |
| Jul | 4 | 4 | 10 | 18 |
| Aug | 7 | 28 | 60 | 95 |
| Sep | 19 | 130 | 89 | 238 |
| Oct | 9 | 150 | 39 | 198 |
| Nov | 2 | 61 | 31 | 94 |
| Dec | 1 | 32 | 12 | 45 |
| Total | 42 | 405 | 241 | 688 |

- 9.6** In an attempt to study rainfall patterns in West Africa caused by El Nino weather events, Nicholson and Kim (1997) constructed a two-way table that relates the number of days rainfall that occurred each month to the amount of rain in inches that fell on those days (categorized as less than 1 inch and more than 1 inch). Use the modified version of their table below to answer the questions further below. [Answer](#)

| | Jun | Jul | Aug |
|-------|-----|-----|-----|
| <1 in | 7 | 11 | 20 |
| >1 in | 5 | 9 | 10 |

- (a) How many days did it rain in July?
- (b) In the months of June and August, how many days did it rain more than 1 inch?
- (c) What percentage of rainy days in August had less than 1 inch of precipitation?
- (d) If there are 31 days in July, on what percentage of those days did it rain?
- (e) What percentage of rainy days did more than 1 inch of rain fall?
- (f) What percentage of rainy days were in June?

9.4 Tables in R

Two-way frequency tables are constructed in R with `xtabs()`. The first argument is a formula of the form `~rowvar+colvar` and the corresponding data.frame must be included in `data=`. The result of `xtabs()` should be assigned to an object. For example, using the GSS data,

```
> (tbl1 <- xtabs(~fdegree+fgrnsol,data=gss) )
      fgrnsol
fdegree vwill will neither un vun
  1tHS    40 145     132 151 178
    HS    87 542     512 557 392
    JC    15  61      64  54  44
    BS    42 199     179 187  75
    grad   24 104     83  64  24
```

Totals are added to the margins of the table with `addMargins()`. The first argument to `addMargins()` must be an object previously constructed with `xtabs()`. For example,

```
> addMargins(tbl1)
  fgrnsol
fdegree vwill will neither un vun Sum
  lHS    40 145     132 151 178 646
  HS     87 542     512 557 392 2090
  JC     15  61      64  54  44 238
  BS     42 199     179 187  75 682
  grad   24 104      83  64  24 299
  Sum    208 1051    970 1013 713 3955
```

Percentage tables are constructed by submitting the previously saved `xtabs()` object to `percTable()`. If only a previously saved table is given to `percTable()` then a table-percentage table is constructed. The number of digits of the output is controlled with `digits=`. For example, the table-percentage table is constructed with

```
> percTable(tbl1,digits=1)
  fgrnsol
fdegree vwill will neither un vun Sum
  lHS    1.0 3.7     3.3 3.8 4.5 16.3
  HS     2.2 13.7    12.9 14.1 9.9 52.8
  JC     0.4 1.5     1.6 1.4 1.1 6.0
  BS     1.1 5.0     4.5 4.7 1.9 17.2
  grad   0.6 2.6     2.1 1.6 0.6 7.5
  Sum    5.3 26.5    24.4 25.6 18.0 99.8
```

Row- and column-percentage tables are constructed by including the `margin=` argument to `percTable()`. If `margin=1` is used, then a row-percentage table is constructed. For example, the row-percentage table is constructed with

```
> percTable(tbl1,margin=1,digits=1)
  fgrnsol
fdegree vwill will neither un vun Sum
  lHS    6.2 22.4    20.4 23.4 27.6 100.0
  HS     4.2 25.9    24.5 26.7 18.8 100.1
  JC     6.3 25.6    26.9 22.7 18.5 100.0
  BS     6.2 29.2    26.2 27.4 11.0 100.0
  grad   8.0 34.8    27.8 21.4  8.0 100.0
```

If `margin=2` is used, then a column-percentage table is constructed. For example, the column-percentage table is constructed with

```
> percTable(tbl1,margin=2,digits=1)
  fgrnsol
fdegree vwill will neither un vun
  lHS   19.2 13.8    13.6 14.9 25.0
  HS    41.8 51.6    52.8 55.0 55.0
  JC    7.2  5.8     6.6  5.3  6.2
  BS   20.2 18.9    18.5 18.5 10.5
  grad 11.5  9.9     8.6  6.3  3.4
  Sum  99.9 100.0   100.1 100.0 100.1
```

- ◊ The table submitted as the first argument to percTable() must be a raw frequency table WITHOUT margin totals added to it.

Review Exercises

9.7  Using the data provided in Review Exercise 9.1. Construct a two-way frequency table, including the marginal totals, of these data with month as the column variable. [Answer](#)

9.8 Construct a row-, column-, and table-percentage table from the frequency table for the seal data in Review Exercise 9.7. [Answer](#)

9.9 Use the [Arsenic.csv](#) data introduced in Review Exercise 5.40 to construct a bivariate EDA for the drinking and usage variables. [Answer](#)

9.10 In the General Social Survey (GSS), two questions were asked – “How often do you make a special effort to sort glass or cans or plastic or papers and so on for recycling?” and “In general, do you think that a rise in the world’s temperature caused by the greenhouse effect, is extremely likely, very likely, somewhat likely, not very likely, or not at all likely?”. Both of these variables are recorded in the [GSSEnviroQues.csv](#) file. Use these data to answer the questions below. [Answer](#)

- What percentage of all respondents recycle often and feel that it is very likely that the greenhouse effect has caused the rise in world’s temperature?
- What percentage of those respondents that recycle often feel that it is very likely that the greenhouse effect has caused the rise in world’s temperature?
- What percentage of those respondents that think it is very likely that the greenhouse effect has caused the rise in world’s temperature also recycle often?
- What percentage of all respondents recycle often?
- What percentage of all respondents think it is very likely that the greenhouse effect has caused the rise in world’s temperature?

9.11  The data in Zoo1.csv contains a list of animals found in several different zoos¹. In addition, each animal was classified into broad “type” categories (“mammal”, “bird”, and “amph/rep”). The researchers that collected these data wanted to examine if the distribution of broad animal types differed among zoos. Use these data to answer the questions below. [Answer](#)

- What is the “response” variable in this analysis?
- What percentage of all animals were birds?
- What percentage of animals in the Minnesota zoo were birds?
- What percentage of animals in the Chicago zoo were amphibians/reptiles?
- What percentage of animals were in the Chicago zoo?
- What percentage of birds were in the Minnesota zoo?

¹

These data are stored in a “comma separated values” (CSV) file rather than a “tab delimited text” file. Thus, these data must be loaded into R with `read.csv()` rather than `read.csv()`. The arguments to `read.csv()` are the same as `read.csv()`.

CHAPTER 10

LINEAR REGRESSION

Chapter Objectives:

1. Describe the purposes of regression.
2. Describe the criteria used to determine the best-fit line to a set of bivariate data.
3. Describe the assumptions surrounding the best-fit criteria.
4. Identify the response and explanatory variables.
5. Describe the equation of a line and what the slope and intercept “mean.”
6. Make appropriate predictions using the best-fit line.
7. Describe the meaning of the coefficient of determination.

Contents

| | |
|---|-----|
| 10.1 Response and Explanatory Variables | 128 |
| 10.2 Slope & Intercept | 129 |
| 10.3 Predictions and Residuals | 131 |
| 10.4 Best-fit Criteria | 134 |
| 10.5 Assumptions | 136 |
| 10.6 Coefficient of Determination | 136 |
| 10.7 Examples I | 139 |
| 10.8 Regression in R | 144 |
| 10.9 Examples II | 146 |

LINEAR REGRESSION ANALYSIS IS USED TO MODEL THE RELATIONSHIP between two quantitative variables for two related purposes – (i) explaining variability in the response variable and (ii) predicting future values of the response variable. Examples include predicting ...

- ... the future sales of a product from its price.
- ... family expenditures on recreation from family income.
- ... an animal's food consumption in relation to ambient temperature.
- ... a person's score on a German assessment test based on how many years the person studied German.

◊ Explaining variability of and predicting future values of response variables are the two goals of regression.

Exact predictions cannot be made because of natural variability. For example, two people with the same intake of mercury (from consumption of fish) will not have the same level of mercury in their blood stream (e.g., observe the two individuals in Figure 10.1 that had intakes of 580 ug HG/day). Thus, the best that can be accomplished is to predict the average or expected value for a person with a particular intake value. This will be accomplished by finding the line that best “fits” the points on a scatterplot of the data. Finding and using that “best-fit” line is the topic of this chapter.

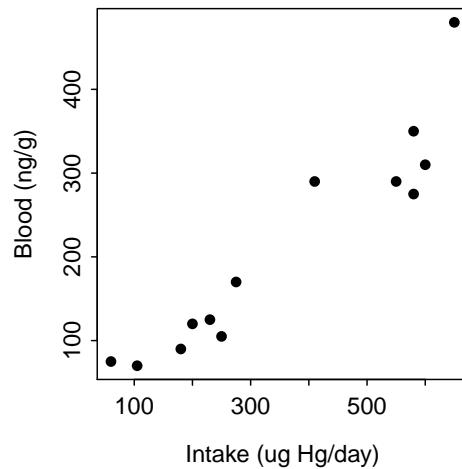


Figure 10.1. Scatterplot between the intake of mercury in fish and the mercury in the blood stream of individuals.

10.1 Response and Explanatory Variables

The variable to be predicted and the variable used to make the prediction must be identified before conducting a regression analysis. The variable to be predicted is the **response variable**¹. The variable used to make the prediction is called the **explanatory variable**². In the examples mentioned above, future sales, family

¹

A synonym is dependent variable.

²

expenditures on recreation, the animal's food consumption, and score on the assessment test are the response variables and product price, family income, temperature, and years studying German are the explanatory variables, respectively. When a scatterplot is made in a regression analysis, the response variable must be placed on the y-axis (and, thus, the explanatory variable on the x-axis)³.

Δ Response Variable: The variable that will be explained or predicted.

Δ Explanatory Variable: The variable that may explain or be used to predict the response variable.

Review Exercises

10.1 Dudgeon (2000), while describing the features of major tropical Asian rivers, examined the relationship between the length (km) and drainage area (km^2) of 11 waterways. In particular he wanted to determine if a model could be produced that would allow the drainage area of the river to be predicted from the length of the river. Identify the response and explanatory variables. Explain your choices. [Answer](#)

10.2 Researchers collected data on 56 normal births at a Wellington, New Zealand hospital. They were interested in determining if the weight of the newborn child (labeled as *BirthWt*) could be predicted by knowing the mothers age (labeled as *Age*). Identify the response and explanatory variables. Explain your choices.

[Answer](#)

10.2 Slope & Intercept

The equation of a line is commonly known as,

$$y = mx + b$$

where both x and y are variables, m represents the slope of the line, and b represents the y-intercept⁴. It is important that you can look at the equation of a line and identify the response variable, explanatory variable, slope, and intercept. The response variable will always appear on one side of the equation (usually the left) by itself. The value or symbol that is multiplied by the explanatory variable (e.g., x) is the slope and the value or symbol by itself is the intercept. For example, without any further explanation, consider the following equation of a line,

$$\text{blood} = 3.501 + 0.579 * \text{intake}$$

A synonym is independent variable.

³

This is the same as what was done in Chapter ??.

⁴

Hereafter, simply called the "intercept."

From this it is seen that *blood* is the response variable, *intake* is the explanatory variable, 0.579 is the slope (it is multiplied by the explanatory variable), and 3.501 is the intercept (it is not multiplied by anything in the equation). The same conclusions are reached even if the equation is written as

$$\text{blood} = 0.579 * \text{intake} + 3.501$$

- ◊ In the equation of a line, the slope is always multiplied by the explanatory variable and the intercept is always by itself.

In addition to being able to identify the slope and intercept of a line you also need to be able to interpret these values. Most students define the slope as “rise over run” and the intercept as “where the line crosses the y-axis.” These “definitions” are very loose geometric representations. For our purposes, the slope and intercept must be more strictly defined.

To define the slope, first think of “plugging” two values of intake into the equation discussed above. For example, if *intake* = 100, then *blood* = 3.501 + 0.579 * 100=61.40 and if *intake* is one unit larger (i.e., *intake* = 101), then *blood* = 3.501 + 0.579 * 101=61.98. The difference between these two values is 61.98-61.40=0.579. Thus, it is seen that the slope is the change in value of the response variable for a single unit change in the value of the explanatory variable (Figure 10.2). That is, mercury in the blood changes 0.579 units for a single unit change in mercury intake. So, if an individual increases mercury intake by one unit, then mercury in the blood will increase by 0.579 units, on average. Alternatively, if one individual has one more unit of mercury intake than another individual, then the first individual will have, on average, 0.579 more mercury units in the blood.

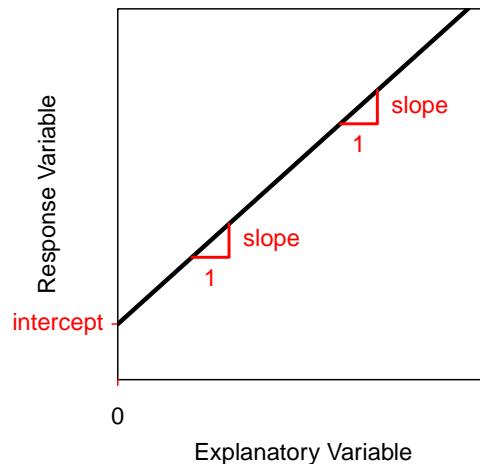


Figure 10.2. Schematic representation of the meaning of the intercept and slope in a linear equation.

To define the intercept, first “plug” *intake* = 0 into the equation discussed above. Thus, if *intake* = 0, then *blood* = 3.501 + 0.579 * 0=3.501. Thus, it is seen that the intercept is the value of the response variable when the explanatory variable is equal to zero (Figure 10.2). In this example, the average mercury in the blood for an individual with no mercury intake is 3.501 units. Many times, as is true with this example, the interpretation of the intercept will be nonsensical. This is because $x = 0$ will likely be outside the range of the data collected and, perhaps, outside the range of possible data that could be collected.

The equation of the line is a model for the relationship depicted in a scatterplot. Thus, the interpretations for the slope and intercept represent the *average* change or the *average* response variable. Thus, whenever a slope or intercept is being interpreted it must be noted that the result is an *average* or *on average*.

Δ Slope: The change in value of the response variable for a unit change in the value of the explanatory variable.

Δ Intercept: The value of the response variable when the explanatory variable is equal to zero.

Review Exercises

10.3 The research described in Review Exercise 10.1 identified the best-fit line equation as $Area = -159131 + 314.229Length$. [Answer](#)

- (a) What is the response variable?
- (b) Interpret the value of the slope in terms of the variables of this problem.
- (c) Interpret the value of the intercept in terms of the variables of this problem.
- (d) If one river was 10 km longer than another river, then how much more area would you expect it to drain?

10.4 The research described in Review Exercise 10.2 computed the following regression results: $BirthWt = 2054 + 51.7Age$. [Answer](#)

- (a) What is the explanatory variable?
 - (b) Interpret the value of the slope in terms of the variables of this problem.
 - (c) Interpret the value of the intercept in terms of the variables of this problem.
 - (d) Assume that a mother had a child when she was 20 and when she was 25. On average, how much more or less would you expect, based on these findings, the second child to weigh compared to the first child?
-

10.3 Predictions and Residuals

Once a best-fit line has been identified (criteria for doing so is discussed in Section 10.4), the equation of the line can be used to predict the average value of the response variable for individuals with a particular value of the explanatory variable. For example, the best-fit line for the mercury data shown in Figure 10.1 is

$$blood = 3.501 + 0.579 * intake$$

Thus, the predicated average level of mercury in the blood for an individual that consumed 240 ug HG/day is found with

$$blood = 3.501 + 0.579 * 240 = 142.461$$

Similarly, the predicted average level of mercury in the blood for an individual that consumed 575 ug HG/day is found with

$$\text{blood} = 3.501 + 0.579 * 575 = 336.426$$

Graphically, a prediction can be visualized by finding the value of the explanatory variable on the x-axis, drawing a vertical line until the best-fit line is intercepted, and then drawing a horizontal line over to the y-axis and reading off the value of the response variable (Figure 10.3).

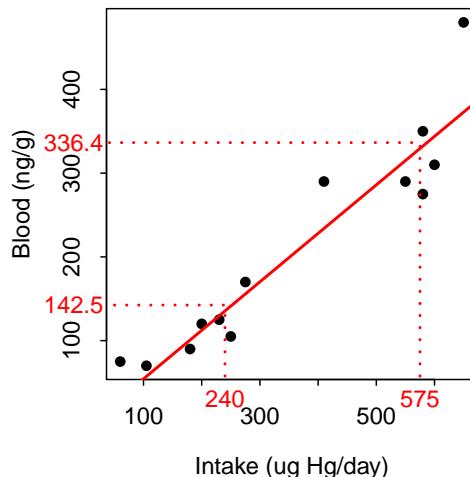


Figure 10.3. Scatterplot between the intake of mercury in fish and the mercury in the blood stream of individuals with superimposed best-fit regression line illustrating predictions for two values of mercury intake.

Δ Predicted Value: The value of y on the best-fit line at the observed value of x ; abbreviated as \hat{y}_i for the i th individual.

◊ The predicted value of the response variable at a given value of the explanatory variable is found by “plugging” the value of the explanatory variable into the equation of the line.

When predicting values of the response variable, it is important to not extrapolate beyond the range of the data. In particular, you should try not to make predictions with values of the explanatory variable that are outside the range of values of the explanatory variable that were used to find the best-fit line. An excellent example would be to consider the height “data” collected during the early parts of a human’s life (say the first ten years). During these early years there is likely a good fit between height (the response variable) and age. Using this relationship to predict an individual’s height at age 40 would likely result in a ridiculous answer (i.e., probably over ten feet). The problem is that the best-fit line only “works” for the first ten years of our life; i.e., it is not known if the same linear relationship exists outside that range of years. In fact, with human heights, it is generally known that growth first slows, eventually quits, and may, at very old ages, actually decline. Thus, the linear relationship found early in life does not hold for later years. Critical mistakes can be made when using a linear relationship to extrapolate outside the range of the data.

◊ When making predictions of the response variable, do not extrapolate beyond the range of the data.

The predicted value is a “best-guess” for an individual based on the best-fit line. The actual value for any individual is likely to be different from this predicted value. The **residual** is a measure of how “far off” the prediction is from what is actually observed for an individual. Specifically, the residual for an individual is computed by subtracting the predicted value (given the individual’s observed value of the explanatory variable) from the individual’s observed value of the response variable, or

$$\text{residual} = \text{observed response} - \text{predicted response}$$

For example, consider an individual that has an observed intake of 650 and an observed level of mercury in the blood of 480. The predicted level of mercury in the blood for this individual is

$$\text{blood} = 3.501 + 0.579 * 650 = 379.851$$

The residual for this individual is then $480 - 379.851 = 100.149$. This positive residual indicates that the observed value is approximately 100 units greater than would be expected based on the best-fit line⁵. As a second example, consider a second individual with an observed intake of 250 and an observed level of mercury in the blood of 105. The predicted value for this individual is

$$\text{blood} = 3.501 + 0.579 * 250 = 148.251$$

The residual for this individual is then $105 - 148.251 = -43.251$. This negative residual indicates that the observed value is approximately 43 units less than would be expected based on the best-fit. Both of these residuals are visualized as the vertical distances between an individual’s point and the corresponding point on the best-fit line in Figure 10.4.

Δ Residual: The vertical difference between the observed and predicted values of the response variable for an individual; computed as the difference between the observed and predicted values of the response.

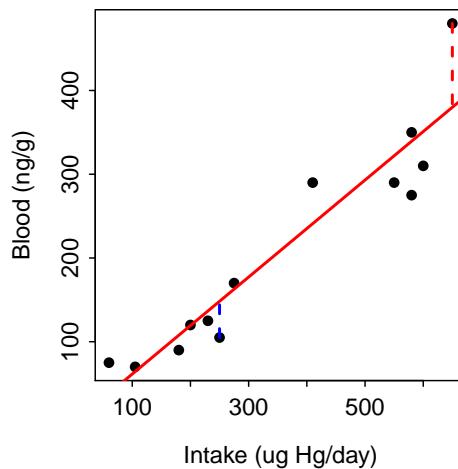


Figure 10.4. Scatterplot between the intake of mercury in fish and the mercury in the blood stream of individuals with superimposed best-fit regression line illustrating the residuals for two individuals.

⁵

In other words, the observed value is “above” the line.

Review Exercises

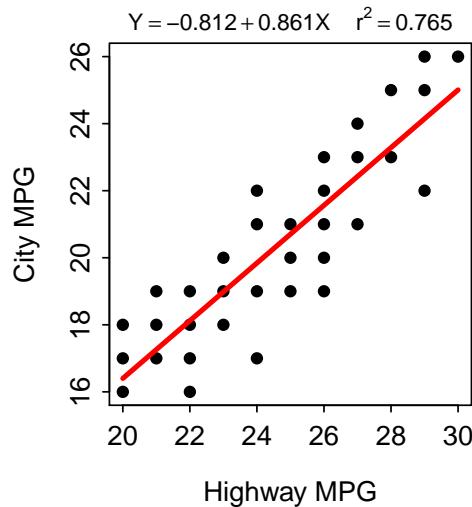
10.5 Use the results described in Review Exercise 10.3 to answer the questions below. [Answer](#)

- (a) Predict the drainage area for a river 3500 km long.
- (b) Calculate the residual if the river above (3500 km length) had a drainage area of 1,000,150 km².
- (c) Predict the drainage area for a river 7500 km long.

10.6 Use the results described in Review Exercise 10.4 to answer the questions below. [Answer](#)

- (a) Predict the weight of a child born to a 30-year-old mother.
- (b) A 30-year-old mother had a child that weighed 3550 g. Find the residual for that mother.
- (c) Predict the weight of a child born to an 18-year-old mother.

10.7 Researchers at Chevrolet attempted to determine the relationship between gas mileage (MPG) of Luminas in the city (CITY) and on the highway (HIGHWAY). Results of their analysis is shown below. [Answer](#)



- (a) Predict the city mpg for a Lumina that gets 25 mpg on the highway.
- (b) Predict the highway mpg for a Lumina that gets 25 mpg in the city.
- (c) Predict the city mpg for a Lumina that gets 40 mpg on the highway.
- (d) What is the residual for a Lumina that gets 25 mpg on the highway and 20 in the city?

10.4 Best-fit Criteria

An infinite number of lines can be placed on a graph. Obviously, many of those lines do not adequately describe the data. In contrast, many of the lines will appear, to our eye, to adequately describe the data. So

as not to rely on the subjectivity of our eye, a specific procedure will be employed to derive a quantifiable and objective measure of which line best “fits” the data. This procedure is called **least-squares regression** and is discussed in this section.

Residuals are a measure of how far away an individual is from a candidate best-fit line. All residuals computed from all individuals in a data set is a measure of how far away all of the individuals are from the candidate best-fit line. Thus, the residuals for all individuals can be used to identify the best-fit line. The residual sum-of-squares (RSS) is defined as the sum of all squared residuals. The least-squares criterion says that the “best-fit” line is the line out of all possible lines that has the minimum RSS. Only one candidate line will have the minimum RSS and the least-squares method defines this as the “best-fit” line to the data Figure 10.5.

Figure 10.5. An animation illustrating how the residual sum-of-squares (RSS) for a series of candidate lines (red lines) is minimized at the best-fit line (green line).

Δ **Residual sum-of-squares:** The sum of all squared residuals; abbreviated as RSS.

◊ **The least-squares criterion is that the “best-fit” line is the line of all possible lines with the minimum RSS.**

The presentation so far implies that all possible lines must be “fit” to the data and the one with the minimum RSS is chosen as the “best-fit” line. As there are an infinite number of possible lines, this would be impossible to do. Theoretical statisticians have shown that the application of the least-squares criterion always produces a best-fit line with a slope given by

$$\text{slope} = r \frac{s_y}{s_x}$$

and an intercept given by

$$\text{intercept} = \bar{y} - \text{slope} * \bar{x}$$

So, one can apply these formulas to the bivariate data to find the best-fit line but these formulas effectively find the best-fit line by minimizing the RSS from all possible lines.

10.5 Assumptions

The least-squares method to find a best-fit line only works appropriately if each of several assumptions about the data has been met. The five assumptions of least-squares regression are,

1. A line describes the data.
2. Homoscedasticity.
3. Normally distributed residuals at a given x .
4. Independent residuals at a given x .
5. The explanatory variable is measured without error.

While all five assumptions of linear regression are important, only the first two are vital when the best-fit line is being used primarily as a descriptive model for data⁶. Description is the primary goal of linear regression used in this book and, thus, we will focus on the first two assumptions, which are required for adequate description.

The first assumption appears to be obvious – if a line does not represent the data, then don’t try to fit a line to it! Violations of this assumption are evident by a non-linear or curving form in the scatterplot. The second assumption, homoscedasticity, states that the variability about the line is the same at all values of the explanatory variable. In other words, the dispersion of the data about the line must be the same everywhere along the entire line. Violations of this assumption are generally evident by a “funnel-shaped” dispersion of points from left-to-right on a scatterplot. Violations of these assumptions are often evident on so-called “fitted-line plots” – i.e., scatterplots with the best-fit line superimposed (Figure 10.6)⁷. If the points look more-or-less like random scatter around the best-fit line, then neither the linearity nor the homoscedasticity assumption has been violated.

In this text, if an assumption has been violated, then one should not continue to interpret the linear regression. However, in many instances, an assumption violation can be “corrected” by transforming one or both variables to a different scale. Transformations are not discussed in this book.

◊ If the regression assumptions are not met, then the regression results should not be interpreted.

10.6 Coefficient of Determination

The coefficient of determination, abbreviated as r^2 , is the proportion of the total variability in the response variable that is explained away by knowing the explanatory variable and the best-fit model. The r^2 can take values between 0 and 1⁸. In simple linear regression⁹ r^2 is literally the square of r , the correlation coefficient¹⁰.

⁶

In contrast to using the model to make inferences about a population model.

⁷

Residual plots, not discussed in this text, are another plot that often times is used to better assess assumption violations.

⁸

It is common for r^2 to be presented as a percentage.

⁹

Simple linear regression is the fitting of a model with a single explanatory variable and is the only model considered in this chapter and this book.

¹⁰

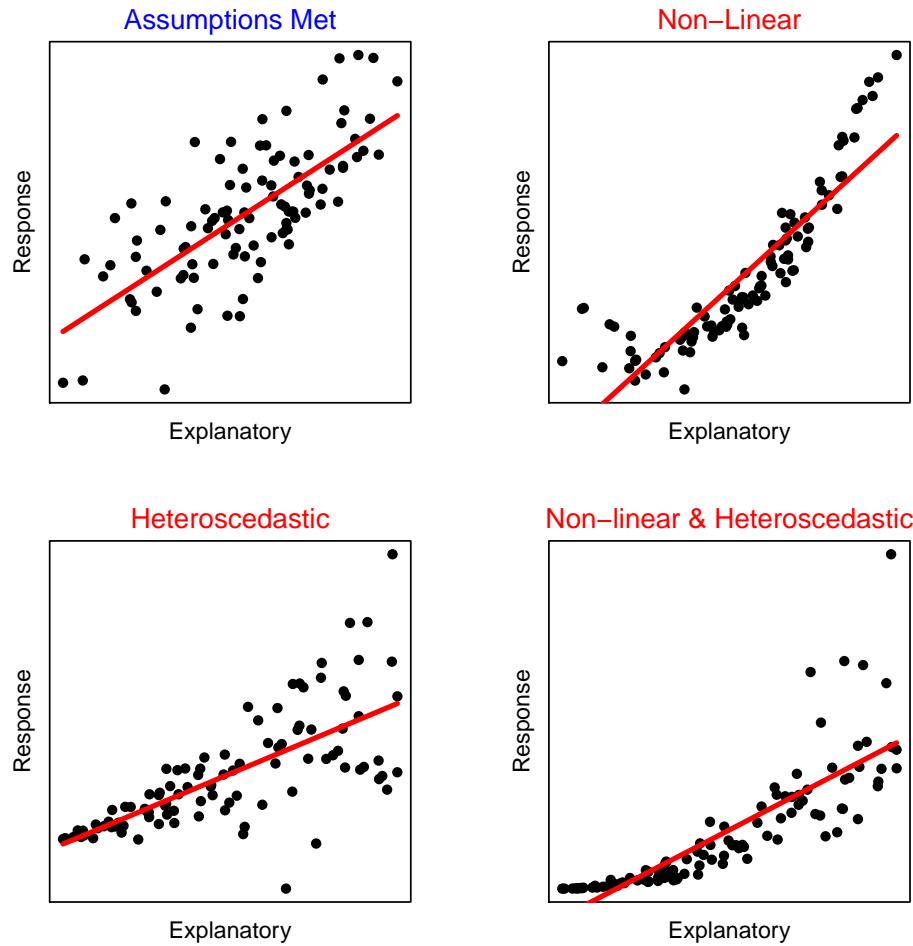


Figure 10.6. Fitted-line plots illustrating when the regression assumptions are met (upper-left) and three common assumption violations.

△ **Coefficient of Determination:** The proportion of the total variability in the response variable that is explained away by knowing the explanatory variable and the best-fit model; abbreviated as r^2 .

◊ r^2 can take values between 0 and 1.

The meaning of r^2 can be examined by considering predictions of the response variable with and without knowledge of the value of the explanatory variable. First, consider predicting the value of a particular response variable without any information about the explanatory variable. In this case, the best prediction for the value of the response variable is to use the sample mean of the response variable (represented by the dashed blue horizontal line in Figure 10.7). However, because of natural variability, not all individuals will have this value. Thus, the prediction might be “bracketed” by saying that the individual will be between the observed minimum and maximum values (solid blue horizontal lines). Loosely speaking, this range can

See Section 8.3 for a review of the correlation coefficient.

be thought of as the “total variability in the response variable” (blue box).

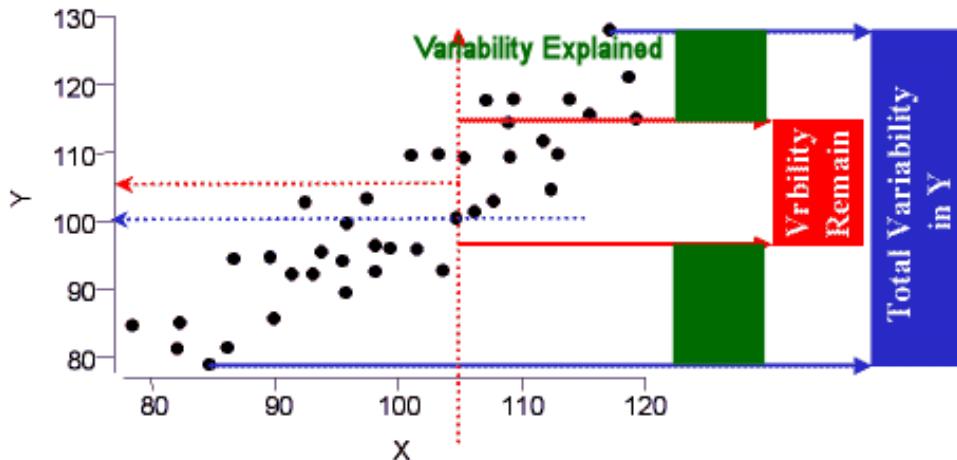


Figure 10.7. Fitted line plot with visual representations of variabilities explained and unexplained. A full explanation is in the text.

Suppose now that interest is in predicting the value of the response variable for an individual with a known value of the explanatory variable (at the dashed vertical red line in Figure 10.7). The predicted value for this individual is the value of the response variable at the corresponding point on the best-fit line (dashed horizontal red line). Again, because of natural variability, not all individuals with this value of the explanatory variable will have this exact value of the response variable. However, the prediction is now “bracketed” by the minimum and maximum value of the response variable **ONLY** for those individuals with the particular value of the explanatory variable (solid red horizontal lines). Loosely speaking, this range can be thought of as the “variability in the response variable remaining after knowing the value of the explanatory variable” (red box). This is the variability in the response variable that remains even after knowing the value of the explanatory variable or the variability in the response variable that cannot be explained away (by the explanatory variable).

The portion of the total variability in the response variable that was explained away consists of all the values of the response variable that would no longer be entertained as possible predictions once the value of the explanatory variable is known (green box in Figure 10.7). Now, by the definition of r^2 , the computation of r^2 can be visualized as the area of the green box divided by the area of the blue box. This calculation does not depend on which value of the explanatory variable is chosen as long as the data are evenly distributed around the line (i.e., homoscedastic – see Section 10.5).

If the variability explained away (the green box in Figure 10.7) approaches the total variability in the response variable (the blue box), then r^2 approaches 1. This will happen only if the variability about the line approaches zero. In contrast, the variability explained (the green box) will approach zero if the slope is zero (i.e., there is no relationship between the response and explanatory variables). Thus, values of r^2 also indicate the strength of the relationship; values near 1 are stronger than values near 0. Values near 1 also mean that predictions will be fairly accurate – i.e., there is little variability remaining after knowing the explanatory variable.

- ◊ A value of r^2 near 1 represents a strong relationship between the response and explanatory variables that will lead to accurate predictions.

10.7 Examples I

There are twelve questions that are commonly asked relative to linear regression results. These twelve questions are listed below with some hints about things to remember when answering some of the questions. An example of these questions in context is then provided.

1. What is the response variable? *Identify which variable is to be predicted or explained, which variable is dependent on another variable, which would be hardest to measure, or which is on the y-axis.*
2. What is the explanatory variable? *The remaining variable after identifying the response variable.*
3. Comment on linearity and homoscedasticity. *Examine fitted-line plot for curvature (i.e., non-linearity) or a funnel-shape (i.e., heteroscedasticity).*
4. What is the equation of the best-fit line? *In the generic equation of the line ($y = mx + b$) replace y with the name of the response variable, x with the name of the explanatory variable, m with the value of the slope, and b with the value of the intercept.*
5. Interpret the value of the slope. *Comment on how the response variable changes by slope amount for each one unit change of the explanatory variable, on average.*
6. Interpret the value of the intercept. *Comment on how the response variable equals the slope, on average, if the explanatory variable is zero.*
7. Make a prediction given a value of the explanatory variable. *Plug the given value of the explanatory variable into the equation of the best-fit line.*
8. Compute a residual given values of both the explanatory and response variables. *Make a prediction (see previous question) and then subtract this value from the observed value of the response.*
9. Identify an extrapolation in the context of a prediction problem. *Examine the x-axis scale on the fitted-line plot and do not make predictions outside of the plotted range.*
10. What is the proportion of variability in the response variable explained by knowing the value of the explanatory variable? *This is r^2 .*
11. What is the correlation coefficient? *This is the square root of r^2 . Make sure to put a negative sign on the result if the slope is negative.*
12. How much does the response variable change if the explanatory variable changes by X units? *This is an alternative to asking for an interpretation of the slope. If the explanatory variable changes by X units, then the response variable will change by X*slope units, on average.*

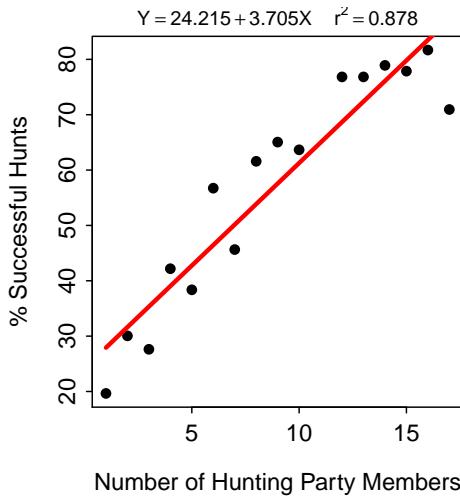
All answers should refer to the variables of the problem – thus, “y”, “x”, “response”, or “explanatory” should not be in any part of any answer. The questions about the slope, intercept, and predictions need to explicitly identify that the answer is an “average” or “on average.”

Chimp Hunting Parties

*Stanford (1996) gathered data to determine if the size of the hunting party (number of individuals hunting together) affected the hunting success of the party (number of hunts that resulted in a kill) for wild chimpanzees (*Pan troglodytes*) at Gombe. The results of their analysis for 17 hunting parties is shown in the figure below¹¹. Use these results to answer the questions below.*

¹¹

These data came from [Chimp.csv](#).



Q: What is the response variable?

A: The response variable is the percent of successful hunts because the authors are attempting to see if success depends on hunting party size. In addition, the percent of successful hunts is shown on the y-axis.

Q: What is the explanatory variable?

A: The explanatory variable is the size of the hunting party.

Q: In terms of the variables of the problem, what is the equation of the best-fit line?

A: The equation of the best-fit line for this problem is % Success of Hunt = 24.215 + 3.705*Number of Hunting Party Members.

Q: Interpret the value of the slope in terms of the variables of the problem.

A: The slope indicates that for every increase of one member to the hunting party the percent of successful hunts increases by 3.705, on average.

Q: Interpret the value of the intercept in terms of the variables of the problem.

A: The intercept indicates that a hunting party with no members will have a percent of successful hunts of 24.215, on average.

Q: What is the predicted hunt success if the hunting party consists of 20 chimpanzees?

A: The predicted hunt success for parties with 20 individuals is an extrapolation, because 20 is outside the range of the number of members observed on the fitted-line plot.

Q: What is the predicted hunt success if the hunting party consists of 12 chimpanzees?

A: The predicted hunt success for parties with 12 individuals is $24.215 + 3.705*12 = 68.7\%$.

Q: What is the residual if the hunt success for 10 individuals is 50%?

A: The residual in this case is $50 - (24.215 + 3.705*10) = 50 - 61.3 = -11.3$. Therefore, it appears that the success of this hunting party is 11.3% lower than average for this size of hunting party.

Q: What proportion of the variability in hunting success is explained by knowing the size of the hunting party?

A: The proportion of the variability in hunting success that is explained by knowing the size of the hunting party is $r^2=0.88$.

Q: What is the correlation between hunting success and size of hunting party?

A: The correlation between hunting success and size of hunting party is $r = 0.94$.

Q: How much does hunt success decrease, on average, if there are two fewer individuals in the party?

A: If the hunting party has two fewer members, then the hunting success would decrease by 7.4% (i.e., -2×3.705), on average.

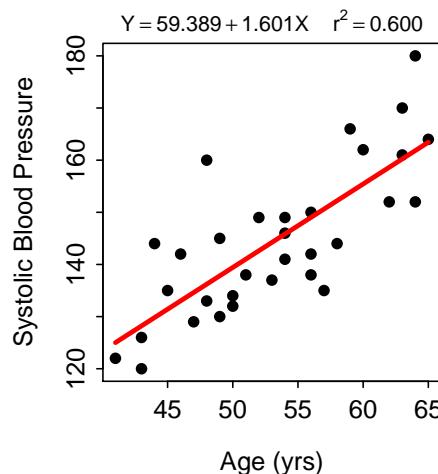
Q: Does any aspect of this regression concern you (i.e., consider the regression assumptions)?

A: The data appear to be very slightly curved but there is no evidence of a funnel-shape. Thus, the data may be slightly non-linear but they appear homoscedastic.

◊ All interpretations should be “in terms of the variables of the problem” rather than the generic terms of x , y , response variable, and explanatory variable.

Review Exercises

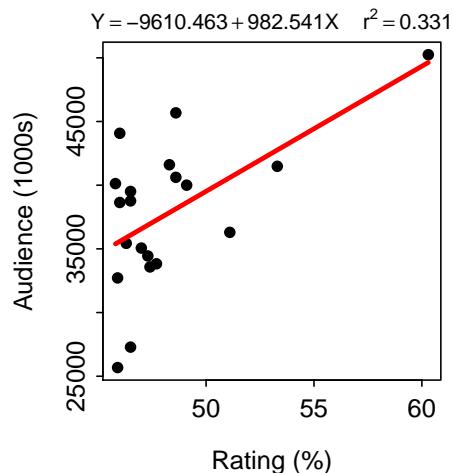
- 10.8** The age (in years) and systolic blood pressure were measured for 32 white males over the age of 40. The researchers wanted to determine if systolic blood pressure increased with increasing age. Thus, they computed the regression depicted in the fitted-line plot below. Use these results to answer the questions below. [Answer](#)



- (a) Which is the explanatory variable?
- (b) Which is the response variable?
- (c) In terms of the variables of this problem, what is the equation of the best-fit line?
- (d) In terms of the variables of this problems, interpret the value of the intercept.
- (e) In terms of the variables of this problems, interpret the value of the slope.
- (f) If male A is 3 years younger than male B, how much difference do you expect to see in their systolic blood pressures?
- (g) What is the predicted systolic blood pressure for a 70-year-old male?
- (h) What is the residual for a 50-year-old male with a SBP of 131?
- (i) What is the correlation coefficient between Age and SBP?
- (j) What proportion of the variability in SBP is explained by knowing the person's AGE?
- (k) What is the predicted systolic blood pressure for a 55-year-old male?

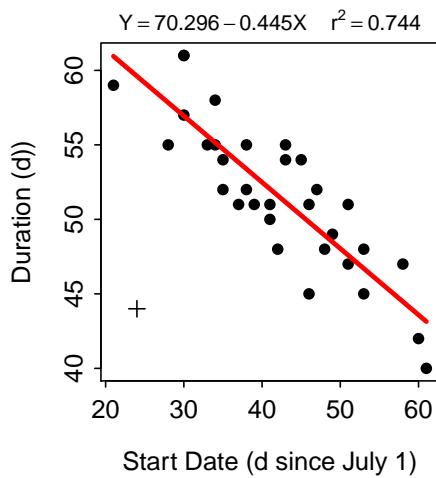
10.9 There are at least two ways that special TV programs could be rated, and both are of interest to advertisers – the estimated size of the audience and the percentage of TV-owning households that tuned into the program. Use the results below for the 20 all-time top-rated programs to determine if the estimated size of the audience can be predicted from the percentage of TV-owning households tuned into the program.

[Answer](#)



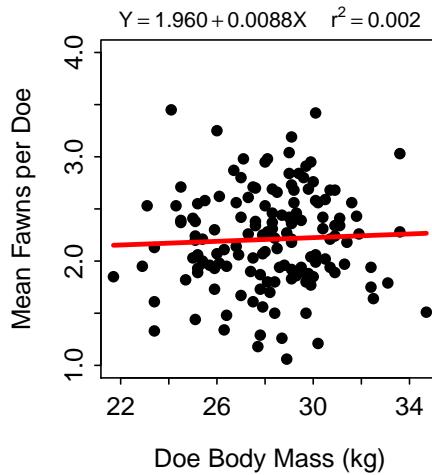
- (a) What did the researchers consider the response variable to be?
- (b) What is the equation of the best-fit line in terms of the variables of the problem?
- (c) Interpret the value of the slope in terms of the variables of the problem.
- (d) What is the predicted audience size for a show with a rating of 40.1%?
- (e) What is the residual for a show with a rating of 55 and an audience size (1000s) of 40000?
- (f) What proportion of the variability in audience size is explained by known the rating percentage?
- (g) What is the correlation between audience size and rating percentage?
- (h) What are two things that bother you about this analysis as it is presented here? Be specific!

10.10 Vega Rivera *et al.* (1998) examined the relationship between the duration of molt and the date of molt start (measured in days since July 1) for wood thrush (*Hylocichla mustelina*). A recreation of their results is shown below (note that the outlier marked by a "+" in the scatterplot was ignored in the calculation of the best-fit line). Use these results to answer the questions below. [Answer](#)



- (a) What is the explanatory variable?
- (b) What is the response variable?
- (c) In terms of the variables of this problem, what is the equation of the best-fit line?
- (d) In terms of the variables of this problem, interpret the value of the slope.
- (e) In terms of the variables of this problem, interpret the value of the intercept.
- (f) What is the predicted molt duration if the molt starts on September 10 (71 d since July 1)?
- (g) What is the residual if molt duration is 48 d and the start date is Aug. 12 (43 d since July 1)?
- (h) What is the correlation between molt duration and molt start date?
- (i) What proportion of the variability in molt duration is explained when molt start date is 37?
- (j) What proportion of the variability in molt duration is explained when molt start date is 57?
- (k) What would happen to the value of the slope if the outlier was NOT ignored?

10.11 Wildlife ecologists in Texas wanted to determine if the number of fawns born to each doe could be explained by the doe's body mass (Ginnett and Young 2000). As part of their study, the researchers recorded the mean number of fawns born to a doe (over a period of time) and the body mass of the doe (kg). Use the results in the following graph to explain the relationship to answer the questions below. [Answer](#)



- (a) Which is the explanatory variable?
 - (b) Which is the response variable?
 - (c) Express the equation of the best-fit line in terms of the variables of the problem.
 - (d) Interpret the slope of the best-fit line in terms of the variables of the problem.
 - (e) If a doe weighed 45 kg, how many fawns on average would you expect her to have?
 - (f) If a doe weighing 32 kg gave birth to an average of 1.9 fawns, what is the residual for this doe?
 - (g) What is the correlation coefficient between mean number of fawns born and doe body mass?
 - (h) How much of the variability in the mean number of fawns born is explained by knowing the body mass of does?
 - (i) If body mass increases by 5 kg, how many more fawns can you expect that doe have?
 - (j) Do you have any concerns about the strength of this relationship?
-

10.8 Regression in R

The linear regression model is fit to two quantitative variables with `lm()`. The first argument to `lm()` is a model formula of the form `response~explanatory` where `response` is a vector containing the response variable and `explanatory` is a vector containing the explanatory variable. These two vectors must be found in the data frame that is sent to the `data=` argument. The results of `lm()` should be assigned to an object so that that object can be submitted to other functions to extract specific results.

◊ The model formula used in the scatterplot and the linear model should be the same.

The regression is fit to the mercury in the blood and the intake of mercury data by first loading and viewing the structure of the data with

```
> setwd('c:/data/')
> merc <- read.csv("Mercury.csv")

> str(merc)
'data.frame': 13 obs. of 2 variables:
 $ intake: num 180 200 230 410 600 550 275 580 580 105 ...
 $ blood : num 90 120 125 290 310 290 170 275 350 70 ...
```

The appropriate formula and `data=` arguments are then submitted to `lm()`, with the result assigned to an object and the results printed¹², with

```
> ( lm1 <- lm(blood~intake,data=merc) )
Coefficients:
(Intercept)      intake
3.5007        0.5791
```

From this it is seen that the intercept is 3.501 and the slope is 0.579. A fitted-line plot (i.e., a scatterplot with the best-fit line superimposed) is constructed by submitting the `lm` object to `fitPlot()`. For example, the fitted-line plot shown in Figure 10.8 was constructed with

```
> fitPlot(lm1)
```

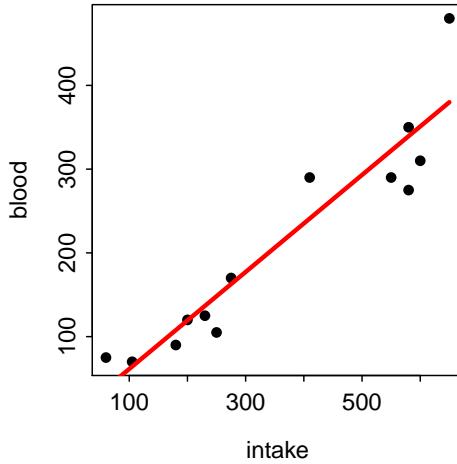


Figure 10.8. Fitted-line plots for the regression of mercury in the blood on mercury intake.

Predicted values from the linear regression is obtained with `predict()`. The `predict()` function requires the saved `lm` object as its first argument. The second argument is a data frame constructed with `data.frame()` that contains the **EXACT** name of the explanatory variable as it appeared in `lm()` set equal to the value of the explanatory at which the prediction should be made. For example, the predicted amount of mercury in the blood for an intake of 240 μg per day is obtained with

¹²

Alternatively, the coefficients are obtained by submitting the `lm` object to `coef()` as such `coef(lm1)`.

```
> predict(lm1,data.frame(intake=240))
   1
142.4895
```

Thus, a fish with an intake of 240 μg of mercury per day will have 142.5 mg per g of mercury in the blood, on average.

- ◊ The name of the explanatory variable used in the `predict()` function must be exactly the same as it appears in the original data frame.

The coefficient of determination is computed by submitting the object saved from `lm()` to `rSquared()`. In addition, the value can be rounded to a certain level of digits by using the `digits=` argument. For example, the r^2 value for this example is computed with

```
> rSquared(lm1,digits=3)
[1] 0.884
```

Thus, 88.4% of the variability in mercury in the blood is explained by knowing the amount of mercury at intake.

10.9 Examples II

Car Weight and MPG

In Chapter ??, an EDA for the relationship between *HMPG* (the highway miles per gallon) and *Weight* (lbs) of 93 cars from the 1993 model year was performed. This relationship will be explored further here as an example of a complete regression analysis. In this analysis, the regression output will be examined within the context of answering the ten typical questions. These data are read into R with

```
> cars93 <- read.csv("data/93cars.csv")
```

The linear regression model is fit, coefficients extracted, fitted-line plot constructed, and coefficient of determination extracted with

```
> ( lm2 <- lm(HMPG~Weight,data=cars93) )
Coefficients:
(Intercept)      Weight
 51.601365     -0.007327
> fitPlot(lm2,ylab="Highway MPG")
> rSquared(lm2,digits=3)
[1] 0.657
```

The simple linear regression model appears to fit the data moderately well as the fitted-line plot (Figure 10.9) shows only a very slight curvature and only very slight heteroscedasticity¹³. The sample slope is -0.00733,

¹³

In advanced statistics books, objective measures for determining whether there is significant curvature or heteroscedasticity in the data are used. In this book, we will only be concerned with whether there is strong evidence of curvature or heteroscedasticity. There does not seem to be either here.

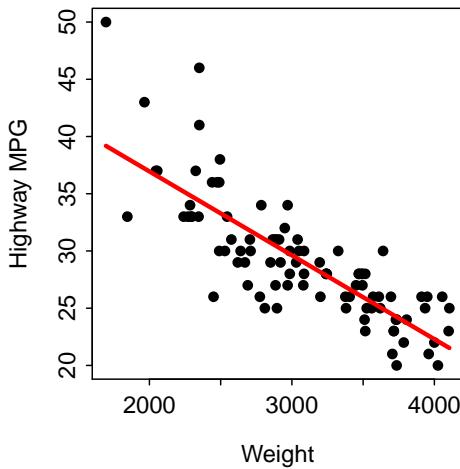


Figure 10.9. Fitted line plot of the regression of highway MPG on weight of 93 cars from 1993.

the sample intercept is 51.60137, and the coefficient of determination is 0.657.

Q: What is the response variable?

A: The response variable in this analysis is the highway MPG, because that is the variable that we are trying to learn about or explain the variability of.

Q: What is the explanatory variable?

A: The explanatory variable in this analysis is the weight of the car (by process of elimination).

Q: In terms of the variables of the problem, what is the equation of the best-fit line?

A: The equation of the best-fit line for this problem is $\text{HMPG} = 51.60137 - 0.00733\text{Weight}$ ¹⁴.

Q: Interpret the value of the slope in terms of the variables of the problem.

A: The slope indicates that for every increase of one pound of car weight the highway MPG decreases by -0.007, on average¹⁵.

Q: Interpret the value of the intercept in terms of the variables of the problem.

A: The intercept indicates that a car with 0 weight will have a highway MPG value of 51.6, on average¹⁶.

Q: What is the predicted highway MPG for a car that weighs 3100 lbs?

¹⁴

The phrase “in terms of the variables of the problem” means to substitute the actual response variable for y and the actual explanatory variable for x . Your answer should not contain references to “y”, “x”, “response variable”, or “explanatory variable” but to the actual variable names.

¹⁵

The words “on average” are important in this and other answers as the best-fit line predicts the average response for a given value of the explanatory variable not the actual value for an individual which will vary due to natural variability.

¹⁶

This is the correct interpretation of the intercept. However, the interpretation of the intercept will often be nonsensical because it is likely an extrapolation (i.e., note that no value of car weight is near 0).

A: The predicted highway MPG for a car that weighs 3100 lbs is $51.60137 - 0.00733(3100) = 28.9$ MPG¹⁷. Alternatively, this value is computed in R with

```
> predict(lm2,data.frame(Weight=3100))
   1
28.88748
```

Q: What is the predicted highway MPG for a car that weighs 5100 lbs?

A: The predicted highway MPG for a car that weighs 5100 lbs should not be computed with the results of this regression, because 5100 lbs is outside the domain of the data (Figure 10.9).

Q: What is the residual for a car that weights 3500 lbs and has a highway MPG of 24?

A: The predicted highway MPG for a car that weighs 3500 lbs is $51.60137 - 0.00733(3500) = 26.0$. Thus, the residual for this car is $24 - 26.0 = -2.0$. Alternatively, this is computed in R with

```
> 24-predict(lm2,data.frame(Weight=3500))
   1
-1.956658
```

Therefore, it appears that this car gets 2.0 MPG less than an average car with the same weight.

Q: What proportion of the variability in highway MPG is explained by knowing the weight of the car?

A: The proportion of the variability in highway MPG that is explained by knowing the weight of the car is $r^2=0.66$.

Q: What is the correlation between highway MPG and car weight?

A: The correlation between highway MPG and car weight is $r = -0.81$ ¹⁸.

Review Exercises

10.12  Wang and Finch (1997) hypothesized that larger willow flycatchers (*Empidonax traillii*) migrated up the Middle Rio Grande River earlier than small willow flycatchers. To test this hypothesis they captured flycatchers on several days during their migration and measured the wing length (mm; an index of overall body size) of each bird. They recorded the date that the bird was captured as a Julian date (days since Jan. 1). The results of their study are found in [Flycatcher.csv](#). Load these data into R and produce results that can be used to answer the questions below. [Answer](#)

- (a) What is the explanatory variable?
- (b) What is the response variable?
- (c) In terms of the variables of this problem, what is the equation of the best-fit line?
- (d) In terms of the variables of this problem, interpret the value of the intercept.
- (e) In terms of the variables of this problem, interpret the value of the slope.

¹⁷

The given weight is plugged into the best-fit line equation for the weight variable, because the weight of the car is within the domain of the data.

¹⁸

Remember to put a negative sign in front of your result from taking the square root of r^2 , because the relationship between highway MPG and weight is negative.

- (f) How much different do you expect the wing length to be ten days later?
- (g) What is the predicted wing length on day 180?
- (h) What is the residual for a bird with wing length 66.5 on day 151?
- (i) What proportion of the variability in wing length is explained by knowing the date?
- (j) What is the correlation coefficient between wing length and date?
- (k) Comment on the assumptions of the linear regression.

10.13  Carroll (1975) examined the relationship between per capita consumption of animal fat (g/day; AnimFat1) and age-adjusted death rate from breast cancer (AgeAdjDe) for 39 countries. Her goal was to determine if variability in the breast cancer death rate could be explained by the amount of fat consumed. The data for their study are found in [CancerFat.csv](#). Load these data into R and produce results that can be used to answer the questions below. [Answer](#)

- (a) Which variable is the response variable?
- (b) What is an individual in this study?
- (c) In terms of the variables of this problem, what is the equation of the best-fit line?
- (d) In terms of the variables of this problems, interpret the value of the slope.
- (e) If country A consumes 4 g/day less animal fat than country B, how much different will the predicted age adjusted death rate due to breast cancer be for country A?
- (f) What is the predicted age adjusted death rate due to breast cancer for a country that consumes 170 g/day of animal fat?
- (g) What is the residual for a country that consumes 90 g/d of animal fat and has an age adjusted death rate due to breast cancer of 14.5?
- (h) What is the correlation coefficient between the age adjusted death rate and the intake of animal fat?
- (i) How much of the variability in a country's age adjusted death rate due to breast cancer is explained by knowing the value of its animal fat intake?
- (j) Can it be said that an increase in intake of animal fat is the cause for an increase in the age adjusted death rate due to breast cancer? Why or why not?

10.14  Allen et al. (1997) investigated the impact of the density of red-imported fire ants (*Solenopsis invicta*; RIFA) on the recruitment of white-tailed deer (*Odocoileus virginianus*) fawns (an index of does to fawns). A modified version of their results are found in [RIFA.csv](#). Load these data into R and produce results that can be used to answer the questions below. [Answer](#)

- (a) What is the response variable?
- (b) What is the explanatory variable?
- (c) In terms of the variables of this problem, what is the equation of the best-fit line?
- (d) In terms of the variables of this problem, interpret the value of the slope.
- (e) If the RIFA index increases by 500, how much different do you expect fawn recruitment to be?
- (f) What is the predicted fawn recruitment when the RIFA index is 1700?
- (g) What is the residual when the RIFA index is 2700 and fawn recruitment is 0.3?
- (h) What is the correlation coefficient between RIFA and fawn recruitment?
- (i) What proportion of the variability in fawn recruitment is explained by knowing the RIFA index?
- (j) Comment on the assumptions in this regression.

10.15  All incoming freshmen are required to take a math assessment test to determine which math classes they should take. Sometimes pre-registering students will register before taking the assessment. To make the best possible course choices for these students, the adviser would like to predict their assessment score (ASSESS) based on their math ACT scores (ACT). The ACT score and assessment score from 72 freshmen

from 2003 are stored in [NCAssess.csv](#). Load these data into R and produce results that can be used to answer the questions below. [Answer](#)

- What is the explanatory variable?
- In terms of ACT and Assessment test scores, what does the value of the slope mean?
- Mary Lamb had an ACT score of 40. Predict her assessment score.
- John Tukey had an ACT score of 19. Predict his assessment score.
- John Tukey actually scored a 15 on his assessment test. Calculate his residual?
- What proportion of the variability in assessment score is explained by knowing the ACT score?
- What are the two most important assumptions in a regression analysis. Are these violated for this data set? Why or why not?
- Do you think that these results provide a useful predictor of math assessment scores in cases where those scores are not available but ACT scores are? Explain.

10.16  Suit and Bauer (1990) examined DNA indices obtained from fresh and frozen tissue samples with the goal of determining if fresh values could be predicted from frozen values. The data for their study are found in [DNA.csv](#). Load these data into R and produce results that can be used to answer the questions below.

Note that one outlier should be excluded from the analysis. [Answer](#)

- What did the researchers consider as the response variable?
- What is the equation of the best-fit line in terms of the variables of the problem?
- Interpret the value of the slope in terms of the variables of the problem.
- What is the predicted fresh index if the frozen index is 4.05?
- What is the residual for a fresh index of 2.1 and a frozen index of 2.2?
- What proportion of the variability in the fresh index is explained by knowing the frozen index?
- What is the correlation between the fresh and frozen indices?
- What are the two major assumptions of regression and do they look like they've been met with these data (be specific!)?

10.17  Wildlife ecologist in Texas wanted to determine if the amount of precipitation could explain some of the variability observed in the number of fawns born to each doe (Ginnett and Young 2000). Because Texas has many different climatic regions, the state was broken down into eight precipitation zones, and the mean precipitation for each zone over a period of five years was calculated. Furthermore, the researchers measured the mean number of fawns born per 100 does for each of these five years. The data for their study are found in [deer1.csv](#). Load these data into R and produce results that can be used to answer the questions below. [Answer](#)

- Express the equation of the best-fit line in terms of the variables of the problem.
- Interpret the slope of the best-fit line in terms of the variables.
- If the mean precipitation in an area were 1500 mm, how many fawns per 100 does would you expect?
- If a precipitation zone has a mean precipitation of 1050 mm and an average of 37 fawns per 100 does, what is the residual of this zone?
- What is the correlation coefficient between mean no. of fawns per 100 does and mean precipitation?
- What proportion of the variability in the mean number of fawns per 100 does is explained by knowing the mean precipitation?
- If the average amount of precipitation increases by 100 mm, how many more fawns per 100 does would you expect to be born?

10.18  It has been said that temperature can be estimated from the number of cricket chirps heard. To determine if this relationship existed, an entomologist recorded the number of chirps in a 15-second interval

by crickets held at different temperatures. The data for their study are found in [Chirps.csv](#). Load these data into R and produce results that can be used to answer the questions below.

[Answer](#)

- (a) What is the response variable?
 - (b) What is the explanatory variable?
 - (c) In terms of the variables of this problem, what is the equation of the best-fit line?
 - (d) In terms of the variables of this problem, interpret the value of the slope.
 - (e) If the number of chirps increases by 5, then how much different do you expect temperature to be?
 - (f) If you hear 18 chirps during the day and 15 chirps at night, then how much different is the temperature, on average?
 - (g) What is the residual when you hear 12 chirps and the temperature is 65 F?
 - (h) What is the correlation coefficient between temperature and the number of chirps?
 - (i) What proportion of the variability in temperature is explained by knowing the number of chirps?
 - (j) Construct a residual plot and use it to interpret the validity of regression assumptions.
-

Part III

Inference Concepts

CHAPTER 11

PROBABILITY INTRODUCTION

Chapter Objectives:

1. Identify the two major assumptions for computing basic probabilities.
2. Calculate basic probabilities in discrete item cases.
3. Calculate basic probabilities for continuous variables that follow a normal distribution.

PROBABILITY is the “language” used by statisticians to describe the proportion of time that a random event will occur. The language of probability is at the center of statistical inference. Thus, an understanding of probability is critical to understanding inferential statistical methods (see Chapter ??). The level of understanding of probability required to understand the basic statistical methods found in most basic inferential methods, including all of those in this book, is not great. Thus, this chapter provides a very short, example-based, introduction to probability. Remember that this is a very basic introduction that will suffice for the topics of this book; a deeper understanding of probability is required to understand more complex inferential methods.

The most basic forms of probability assume that the items in any probability calculation was selected completely randomly. In other words, simple probability calculations require that each item, whether that item is an individual or an entire sample, has the same chance of being selected. Thus, in simple intuitive examples it will be stated that the “box of balls was thoroughly mixed” and more realistic examples will require randomization¹. In its most basic form, probability requires randomization!

- ◊ Individuals must be randomly selected from the population or samples must be produced randomly for the concept of probability to work accurately.

¹

See sections in Chapter 3 for methods of selecting or allocating random individuals.

If every individual has the same chance of being selected, then the probability of an event is equal to the proportion of items in the event out of the entire population. In other words, the probability is the number of items in the event divided by the total number of items in the population. For example, the probability of selecting a red ball from a thoroughly mixed box containing 15 red and 10 blue balls is equal to $\frac{15}{25} = 0.6$ (i.e., 15 individuals (“balls”) in the event (“red”) divided by the total number of individuals (“all balls in the box”). Similarly, the probability of randomly selecting a woman from a room containing 20 women and 30 men is $0.4 (= \frac{20}{50})$. In both of these examples, the calculation can be considered a probability because (i) individuals were randomly selected and (ii) a proportion of a total was computed.

- ◊ If every item has the same chance of being selected, then the probability of observing an item with a certain characteristic is the proportion of items in the entire population that have that characteristic.

The two previous examples are rather simple examples where the selection is made from a small, discrete number of items. Probabilities can also be computed for continuous variables if the distribution of that variable for the entire population is known. For example, the probability that a random individual is greater than 71 inches tall can be calculated if the distribution of heights for all individuals in the population is known. Of course, information about the population is typically difficult to know. However, the normal distributions models from Chapter 7 can be used in many situations to provide a model of this population distribution. For example, an example in Chapter 7 assumed that the heights of a population was $N(66, 3)$ and found that the proportion of individuals in the population taller than 71 inches tall was 0.0478². This result can be considered a probability because this calculation finds the proportion of all individuals of interest in the entire population and the question was about a randomly selected individual.

- ◊ The calculations from the normal distribution made in Chapter 7 are probability calculations as long as the individuals are randomly selected.

A theory that explains the distribution of statistics computed from all possible samples from a population will be developed in Chapter 12. This distribution will be used to compute the probability of observing a particular range of statistics in a random sample of individuals. This technique is the basis for making statistical inferences about the population in Chapter ??.

Review Exercises

11.1 A coin purse contains 17 nickels and 15 dimes. Use this to answer the questions below. [Answer](#)

- What is the probability of randomly selecting a nickel from this purse?
- What is the probability of randomly selecting a dime from this purse?
- What is the probability of randomly selecting a dime from this purse assuming that two nickels and three dimes have already been removed?

11.2 A very small green house contains 10 tomato, 12 pea, and 8 cauliflower plants. Use this to answer the questions below. [Answer](#)

²

This value is computed with `distrib(71,mean=66,sd=3,lower.tail=FALSE)`.

- (a) What is the probability of randomly selecting a tomato plant from this greenhouse?
- (b) What is the probability of randomly selecting a cauliflower plant from this greenhouse?
- (c) What is the probability of randomly selecting a pea plant from this greenhouse assuming that all tomato plants had died and were removed from the greenhouse?

11.3  Suppose that the length of all needles on a particularly large pine tree is known to be normally distributed with a mean of 75 mm and a standard deviation of 8 mm. Use this to answer the questions below.

[Answer](#)

- (a) What is the probability that a randomly selected needle is between 70 and 80 mm long?
 - (b) What is the probability that a randomly selected needle is longer than 90 mm?
 - (c) What is the probability that a randomly selected needle is less than 50 mm long?
-

CHAPTER 12

SAMPLING DISTRIBUTIONS

Chapter Objectives:

1. Describe the concept of sampling variability.
2. Describe why sampling variability must be dealt with to make inferences.
3. Describe what a sampling distribution represents.
4. Identify how a sampling distribution differs from a population distribution.
5. Describe what a standard error is.
6. Identify how a standard error differs from a standard deviation.
7. Describe how and why sampling distributions are simulated.
8. Explain the concepts of precision, accuracy, and bias as it relates to statistics and parameters.
9. Describe the theoretical distribution of the sampling distribution of the sample means.
10. Gain some belief that the theoretical distribution actually represents the sampling distribution of the sample means.
11. Use the sampling distribution of sample means to compute the probability of particular sets of means.

Contents

| | |
|---|-----|
| 12.1 Definition and Characteristics | 157 |
| 12.2 Simulating | 162 |
| 12.3 Accuracy and Precision | 165 |
| 12.4 Central Limit Theorem | 168 |

STATISTICAL INFERENCE IS THE PROCESS of making a conclusion about the parameter of a population based on the statistic computed from a sample. This process is made more difficult by the fact that a statistic is a random variable; i.e., the exact value of the statistic depends on the individuals in the sample from which it was computed. For example, recall from Section ?? that the mean length of fish differed among the four samples of fish that were “taken” from Square Lake. Thus, to be able to make conclusions about the population from the sample, the distribution of the statistic computed from all possible samples must be understood. In other words, to adequately take sampling variability into account in making inferences, the shape, center, and dispersion of the statistic as if all possible samples had been taken must be understood.¹ In this chapter, the distribution of statistics from all possible samples is explored and generalizations used to make inferences are identified. In subsequent chapters, this information along with results from a single sample, will be used to make specific inferences about the population.

- ◊ Making statistical inferences requires a consideration of sampling variability.

12.1 Definition and Characteristics

A **Sampling distribution** is the distribution of the values of a particular statistic computed from all possible samples of the same size from the same population. It is important to note, though, that only one sample is ever taken from a population. However, this discussion of sampling distributions and all subsequent theories related to statistical inference are based on repeated samplings from the same population. As these theories are developed, we will consider taking multiple samples; however, after the theories have been developed, then only one sample will be taken with the theory then being applied to those results.

- Δ **Sampling Distribution:** The distribution of the values of a particular statistic computed from all possible samples of the same size from the same population.

Sampling distributions are integral to making statistical inferences about parameters in a population but can really only be computed for very small populations. Thus, to illustrate the concept of a sampling distribution, consider a population of six students that have scored 6, 6, 4, 5, 7 and 8 points, respectively, on an 8-point quiz. The mean of this population is $\mu = 6.000$ points and the standard deviation is $\sigma = 1.414$ points. Suppose that every sample of size $n = 2$ is extracted from this population and the sample mean is computed for each sample (Table 12.1).² The histogram of these 15 means is the sampling distribution of the sample mean from samples of $n = 2$ from this population (Figure 12.1).³

The mean and standard deviation of the 15 sample means are measures of center and dispersion for the sampling distribution. The mean and standard deviation of the 15 sample means are 6.000 and 0.845, respectively. The standard deviation of the statistics (i.e., the dispersion of the sampling distribution) is generally referred to as the **standard error of the statistic** (abbreviated as SE_{stat}). This new terminology is used to help keep the dispersion of the sampling distribution separate from the dispersion of the individuals in the population, which is measured by the standard deviation. Thus, the standard deviation of all possible sample means is generally referred to as the standard error of the sample means, or SE. The SE in this

¹

See Section ?? for a review of sampling variability.

²

These samples are found by putting the values into a vector with `vals <- c(6,6,4,5,7,8)` and then using `combn(vals,2)`. The means are found with `mns <- as.numeric(combn(vals,2,mean))`.

³

The histogram is constructed with `hist(~mns,w=0.5)`.

Table 12.1. All possible samples of $n = 2$ and the corresponding sample mean from the simple population of quiz scores.

| Scores | Mean |
|--------|------|--------|------|--------|------|--------|------|--------|------|
| 6,6 | 6.0 | 6,7 | 6.5 | 6,5 | 5.5 | 4,5 | 4.5 | 5,7 | 6.0 |
| 6,4 | 5.0 | 6,8 | 7 | 6,7 | 6.5 | 4,7 | 5.5 | 5,8 | 6.5 |
| 6,5 | 5.5 | 6,4 | 5 | 6,8 | 7.0 | 4,8 | 6.0 | 7,8 | 7.5 |

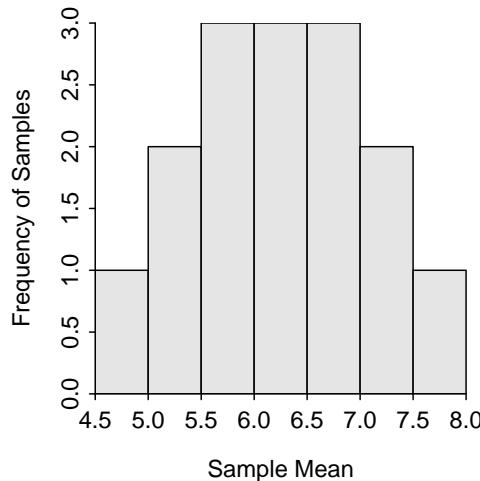


Figure 12.1. Sampling distribution of mean quiz scores from samples of $n = 2$ from the simple population of quiz scores.

example is thus 0.845.

Δ Standard Error: The numerical measure of dispersion used for sampling distributions – i.e., measures the dispersion among statistics from all possible samples.

This simple example illustrates three major concepts concerning sampling distributions. First, the sampling distribution of the statistic will more closely resemble a normal distribution than the original population distribution (unless, of course, the population distribution was a normal distribution).

Second, the center (i.e., mean) of the sampling distribution of a statistic will equal the parameter that the statistic was intended to estimate (e.g., a sample mean is intended to be an estimate of the population mean). In this example, it is seen that the mean of all possible sample means from this population ($= 6.0$ points) is equal to the mean of the original population ($\mu = 6.0$ points). A statistic is said to be **unbiased** if the center (mean) of its sampling distribution equals the parameter it was intended to estimate. This example illustrates that the sample mean is an unbiased estimator of the population mean.

Δ Unbiased Statistic: A statistic in which the center of its sampling distribution equals the parameter it is intended to estimate.

◊ Most statistics encountered in this book are unbiased.

Third, the standard error of the statistic will be less than the standard deviation of the original population. In other words, the dispersion of statistics is less than the dispersion of individuals in the population. For example, the dispersion of individuals in the population is $\sigma = 1.414$ points whereas the dispersion of statistics from all possible samples is $SE_{\bar{x}} = 0.845$ points.

- ◊ The sampling distribution will be more normal than the original population distribution.
- ◊ The mean of the statistics in a sampling distribution will (generally) equal the parameter that the statistic was intended to estimate.
- ◊ The dispersion of the sampling distribution will be less than the dispersion of the original population distribution.

Review Exercises

12.1 Use the simple population of quiz scores from the previous section (i.e., 6, 6, 4, 5, 7, and 8) to answer the questions below. [Answer](#)

- (a) Construct a table similar to Table 12.1 that shows the values and the mean of those values for all possible samples of size $n = 4$. Note: there are 15 such samples.
- (b) Construct a histogram of the means from all possible samples. Describe its general shape.
- (c) Compute the mean of the means from all possible samples. How does this compare to the mean of all six individuals in the population?
- (d) Compute the standard error of the means from all possible samples. How does this compare to the standard deviation of all six individuals in the population? How does this compare to the standard error of the means of all possible samples of $n = 2$ shown in Table 12.1 and for all possible samples of $n = 3$ shown in Table 12.2 (later in this chapter)? Can you make a general statement about how the standard error of the means is related to the size of the sample used to construct the means?

12.2 Suppose the individuals in a simple population have the following “values” for a simple binomial categorical variable – Y, Y, N, Y, Y, N, and N. Use this to answer the questions below. [Answer](#)

- (a) Construct a table similar to Table 12.1 that shows the “values” of the individuals and the proportion of “yeses” for all possible samples of size $n = 3$. Note: there are 35 such samples.
- (b) Construct a histogram of the proportions from all possible samples. Describe its general shape.
- (c) Construct the mean of the proportions from all possible samples. How does this compare to the proportion of “yeses” for all seven individuals in the population?
- (d) Construct the standard error of the proportions from all possible samples.

12.1.1 Critical Distinction

There are generally three distributions that are considered in statistical analyses. In this chapter, the sampling distribution or the distribution of a statistic computed from all possible samples of the same size from the same population was introduced. The key words in this definition are “distribution of a statistic.” In addition, the population distribution, which is the distribution of all individuals in a population, was introduced in Chapter 7. The key words in this definition are “distribution of individuals.” The last distribution used in statistics is the sample distribution which is the distribution of all individuals in a sample. This is also a distribution of individuals but only of those in a sample. The histograms of Chapter ?? were examples of sample distributions. As we continue to introduce inferential statistics, it is critically important that you distinguish between the population and sampling distributions. Always keep in mind that one (population) is the distribution of individuals and the other (sampling) is the distribution of statistics.

Just as importantly, you must remember that a standard error measures the dispersion among statistics (i.e., sampling variability) whereas a standard deviation measures dispersion among individuals (i.e., natural variability). Specifically, the population standard deviation measures dispersion among all individuals in the population and the sample standard deviation measures the dispersion of all individuals in a sample. In contrast, the standard error measures the dispersion among statistics computed from all possible samples. The population standard deviation is the dispersion on a population distribution whereas the standard error is the dispersion on a sampling distribution.

- ◊ Sampling distributions represent the distribution of statistics from all possible samples, whereas population distributions represent the distribution of all individuals in a population.
- ◊ Standard error measures dispersion among statistics, whereas standard deviation measures dispersion among individuals.
- ◊ Standard error measures sampling variability, whereas the standard deviation measures natural variability.

Review Exercises

12.3 What type of distribution is the distribution of blood serum level for every individual in a population?

[Answer](#)

12.4 What type of distribution is the distribution of mean cholesterol level computed from all possible samples of $n = 15$ patients for a clinic? [Answer](#)

12.5 What type of distribution is the distribution of water discharge amounts for Bay City Creek for every day in 2005 assuming that all days in 2005 was the population of interest? [Answer](#)

12.6 What type of distribution is the distribution of water discharge amounts for Bay City Creek for every day in

2005 if the population of interest is all days in the 21st century? [Answer](#)

12.7 What type of distribution is the proportion of days where the water discharge from Bay City Creek is near negligible calculated from all samples of $n = 30$ days. [Answer](#)

12.8 On average, the mean length of $n = 30$ cicadas is 2.9 mm away from the overall average. Is this a standard deviation or a standard error? [Answer](#)

12.9 On average, the number of litter items found along the Escarpment Trail in the Porcupine Mountains on a single day is 12 items different than the overall mean. Is this a standard deviation or a standard error? [Answer](#)

12.1.2 Dependencies

The sampling distribution of sample means from samples of $n = 2$ from the population of quizzes was shown above. The sampling distribution will look different if means from any other sample size are computed instead. For example, the samples and means from each sample of $n = 3$ is shown in Table 12.2. The mean of these means is 6.000, the standard error is 0.592, and the sampling distribution is symmetric, perhaps approximately normal (Figure 12.2). The three major characteristics of sampling distributions noted in Section 12.1 are still true: the sampling distribution is still more normal than the original population, the sample mean is still unbiased (i.e., the mean of the means is equal to μ), and the standard error is smaller than the standard deviation of the original population. However, also take note that the standard error of the sample mean is smaller from samples of $n = 3$ than from $n = 2$.⁴

Table 12.2. All possible samples of $n = 3$ and the corresponding sample means from the simple population of quiz scores.

| Scores | Mean |
|--------|------|--------|------|--------|------|--------|------|--------|------|
| 6,6,4 | 5.3 | 6,6,5 | 5.7 | 6,6,7 | 6.3 | 6,6,8 | 6.7 | 4,5,7 | 5.3 |
| 6,4,5 | 5.0 | 6,4,7 | 5.7 | 6,4,8 | 6.0 | 6,5,7 | 6.0 | 4,5,8 | 5.7 |
| 6,5,8 | 6.3 | 6,7,8 | 7.0 | 6,4,5 | 5.0 | 6,4,7 | 5.7 | 4,7,8 | 6.3 |
| 6,4,8 | 6.0 | 6,5,7 | 6.0 | 6,5,8 | 6.3 | 6,7,8 | 7.0 | 5,7,8 | 6.7 |

◊ Sampling distributions differ for samples of different sizes. In particular the distribution will be “more” normal and the standard error will be smaller as sample size increases.

The sampling distribution will also be different if the statistic changes; e.g., if the sample median rather than sample mean is computed in each sample. Before showing the results of each sample, note that the population median (i.e., the median of the individuals in the population — 6, 6, 4, 5, 7, and 8) is 6.0 points. The sample median from each sample is shown in Table 12.3 and the actual sampling distribution is shown in Figure 12.3. Note that the sampling distribution of the sample medians is still “more” normal

⁴

One should also look at the results from $n = 4$ in Review Exercise 12.1.

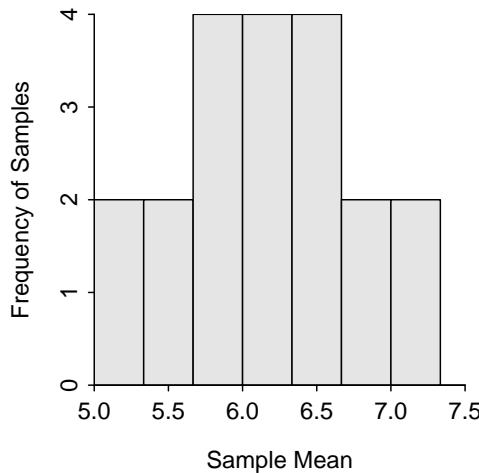


Figure 12.2. Sampling distribution of mean quiz scores from samples of $n = 3$ from the simple population of quiz scores.

than the original population distribution, the mean of the sample medians ($=6.000$ points) still equals the parameter (population median) that the sample median is intended to estimate (thus the sample median is also unbiased), and this sampling distribution differs from the sampling distribution of sample means from samples of $n = 3$.

Table 12.3. All possible samples of $n = 3$ and the corresponding sample medians from the simple population of quiz scores.

| Scores | Median |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 6,6,4 | 6 | 6,6,5 | 6 | 6,6,7 | 6 | 6,6,8 | 6 | 4,5,7 | 5 |
| 6,4,5 | 5 | 6,4,7 | 6 | 6,4,8 | 6 | 6,5,7 | 6 | 4,5,8 | 5 |
| 6,5,8 | 6 | 6,7,8 | 7 | 6,4,5 | 5 | 6,4,7 | 6 | 4,7,8 | 7 |
| 6,4,8 | 6 | 6,5,7 | 6 | 6,5,8 | 6 | 6,7,8 | 7 | 5,7,8 | 7 |

- ◊ Sampling distributions for different statistics are different.

These examples bring up another important point. You must be very specific when naming a sampling distribution. For example, the first sampling distribution in this chapter should be described as the “sampling distribution of sample means from samples of size $n=2$.” This last example should be described as the “sampling distribution of sample medians from samples of size $n=3$.” Doing this with each distribution will reinforce the point that sampling distributions depend on the sample size and the statistic calculated.

- ◊ Each sampling distribution should be specifically labeled with the statistic calculated and the sample size of the samples.

12.2 Simulating

In Section 12.1, exact sampling distribution for statistics computed from very small samples taken from a small population were shown. Exact sampling distributions are very difficult to show for even moderate

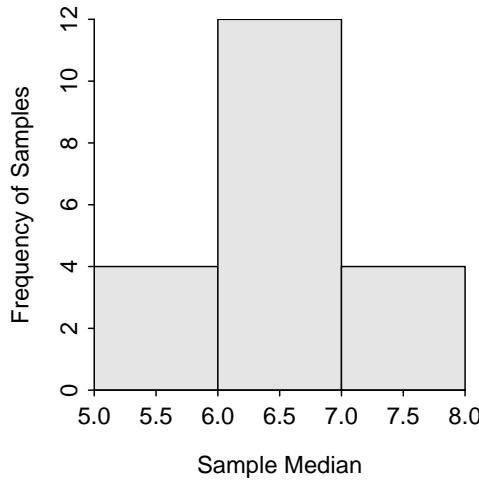


Figure 12.3. Sampling distribution of median quiz scores from samples of $n = 3$ from the simple population of quiz scores.

sample sizes from moderately-sized populations. For example, there are 15504 unique samples of $n = 5$ from a population of 20 individuals. How are sampling distributions examined in these larger cases?

There are two ways to examine sampling distributions in situations with large sample and large population sizes. First, the computer can take many (hundreds or thousands) samples and compute the statistic for each. These statistics can then be summarized to give an indication of what the actual sampling distribution would look like. This process is called “simulating a sampling distribution” and is the subject of this section. Second, theorems exist that describe the specifics of sampling distributions under certain conditions. One such theorem is described in Section 12.4. These theorems will be relied upon in subsequent chapters.

- ◊ The approximate shape of sampling distributions from large samples or large populations can be obtained from (1) theorems or (2) computer simulations.

Sampling distributions are simulated by drawing many (hundreds or thousands) samples from a population, computing the statistic of interest for each sample, and constructing a histogram of these statistics (Figure 12.4). The computer is very helpful with this simulation; however, keep in mind that the computer is basically following the same process as used in Section 12.1 with the exception that not every sample is viewed.

- ◊ Sampling distributions can be simulated by drawing many samples from a population, computing the statistic of interest for each sample, and constructing a histogram of the values of the statistic.

To illustrate the simulation of a sampling distribution, let’s return to the Square Lake fish population explored in Section ???. Recall that this is a hypothetical population with 1015 fish, a population distribution shown in Figure 2.1, and parameters shown in Table 2.1. Further recall that four samples of $n = 50$ were removed from this population and summarized in Table 2.2 and Table 2.3. Suppose, that an additional 996 samples of $n = 50$ were extracted in exactly the same way as the first four, the sample mean was computed in each sample, and the 1000 sample means were collected to form the histogram in Figure 12.5. This histogram is

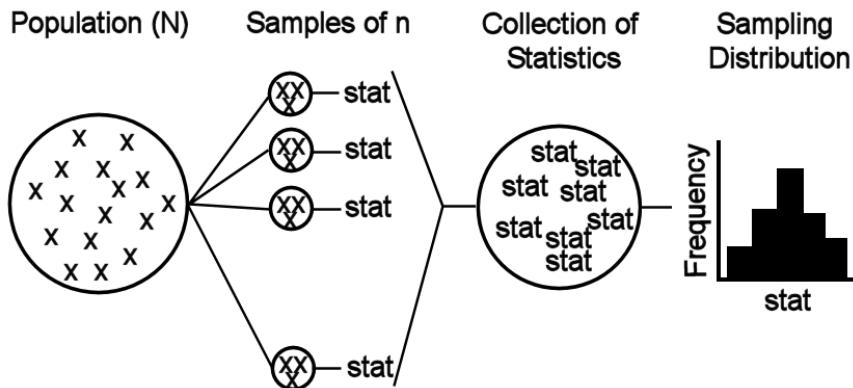


Figure 12.4. Schematic representation of the process for simulating sampling distributions.

a simulated sampling distribution of sample means because it represents the distribution of sample means from 1000, rather than all possible, samples.

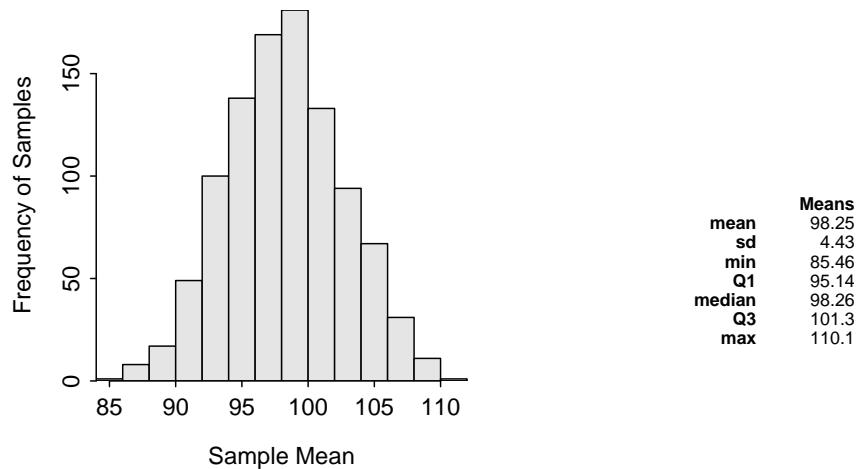


Figure 12.5. Histogram (Left) and summary statistics (Right) from 1000 sample mean total lengths computed from samples of $n = 50$ from the Square Lake fish population.

As with the actual sampling distributions discussed previously, three characteristics (shape, center, and dispersion) are examined with simulated sampling distributions. First, this sampling distribution looks at least approximately normally distributed. Second, the mean of the 1000 means in the sampling distribution ($=98.25$) is approximately equal to the mean of the original 1015 fish in Square Lake ($=98.06$). These two values are not exactly the same because the simulated sampling distribution was constructed from only a “few” samples rather than all possible samples. If all possible samples had been taken, then these values would be exactly equal, as was shown in Section 12.1. Third, the standard error of the sample means ($=4.43$) is much less than the standard deviation of individuals in the original population ($=31.49$). So, within reasonable approximation, the concepts identified with actual sampling distributions also appear to hold for simulated sampling distributions.

As before, computing a different statistic on each sample results in a different sampling distribution. This is illustrated by comparing the sampling distributions of a variety of statistics from the same 1000 samples

of size $n=50$ taken above (Figure 12.6).

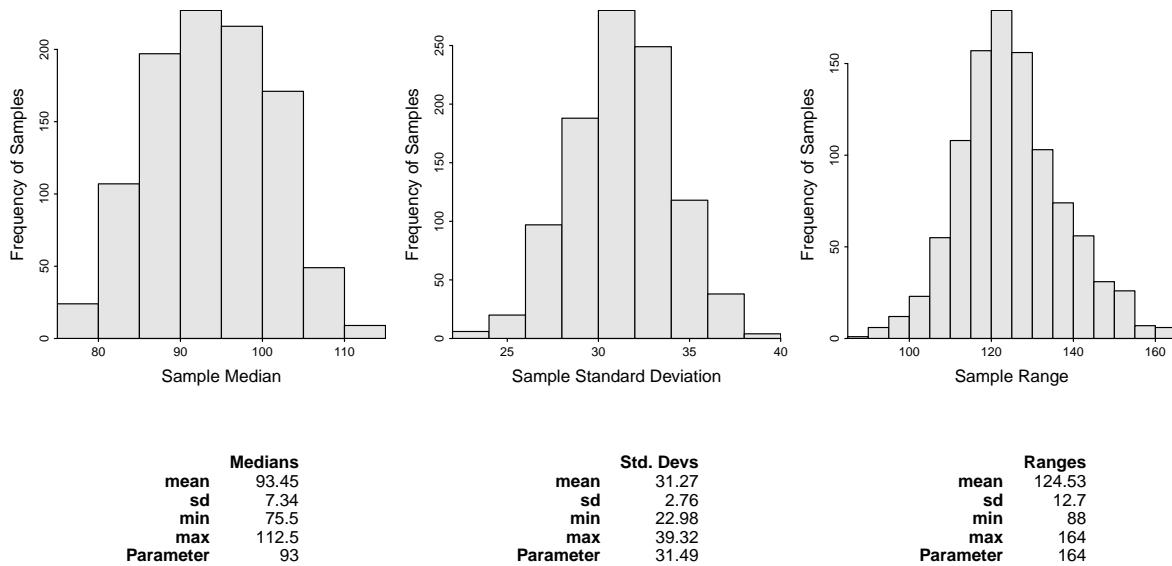


Figure 12.6. Histograms from 1000 sample median (Left), standard deviation (Center), and range (Right) of total lengths computed from samples of $n = 50$ from the Square Lake fish population. Note that the value in the parameter row is the value computed from the entire population.

Simulating a sampling distribution by taking many samples of the same size from a population is powerful for two reasons. First, it reinforces the ideas of sampling variability – i.e., each sample results in a slightly different statistic. Second, the entire concept of inferential statistics is based on theoretical sampling distributions. Simulating sampling distributions will allow us to check this theory and better visualize the theoretical concepts. From this chapter forward, though, you must remember that we will simulate sampling distributions mainly as a check of theoretical concepts. In real life, only one sample is taken from the population and the theory is used to identify the specifics of the sampling distribution.

- ◊ Simulating sampling distributions is a powerful tool for checking the theory concerning sampling distributions; however, in “real-life” only one sample from the population is needed.

12.3 Accuracy and Precision

Accuracy and **precision** are often used to describe characteristics of a sampling distribution. Accuracy refers to how closely a statistic estimates the intended parameter. If, **on average**, a statistic is approximately equal to the parameter it was intended to estimate, then the statistic is considered **accurate**. Unbiased statistics are also accurate statistics. Precision refers to the repeatability of a statistic. A statistic is considered to be **precise** if multiple samples produce “nearly alike” statistics. The standard error of a statistic is a measure of precision; i.e., a high SE means low precision and a low SE means high precision.

The concepts of accuracy and precision are illustrated in Figure 12.7. The targets in Figure 12.7 provide an intuitive interpretation of accuracy and precision, whereas the sampling distributions (i.e., histograms)

are what statisticians look at to identify accuracy and precision. Targets in which the blue plus (i.e., mean of the means) is close to the bullseye are considered accurate or unbiased. Similarly, sampling distributions where the observed center (i.e., blue vertical line) is very close to the actual parameter (i.e., black tick labeled with a “T”) are considered accurate or unbiased. Targets in which the red dots are closely clustered are considered precise. Similarly, sampling distributions that exhibit little variability (low dispersion) are considered precise.

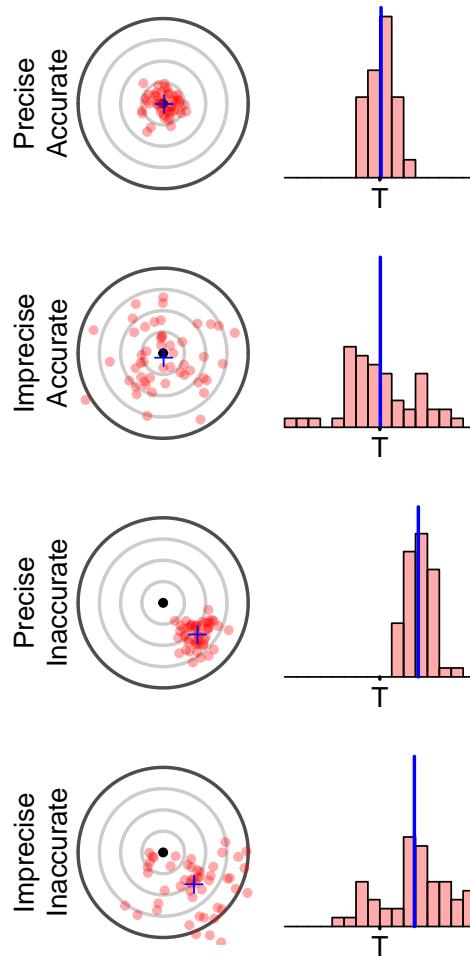


Figure 12.7. Model used to demonstrate accuracy, precision, and bias. The center of each target (i.e., the bullseye) and the point marked with a “T” (for “truth”) represent the parameter of interest. Each dot on the target represents a statistic computed from a single sample and, thus, the many red dots on each target represent repeated samplings from the same population. The center of the samples (analogous to the center of the sampling distribution) is denoted by a blue plus-sign on the target and a blue vertical line on the histogram. The target concept is modified from Ratti and Garton (1994).

Δ Accuracy: The tendency of a statistic to come close to the parameter it was intended to estimate.

Δ **Precision:** The tendency to have values clustered closely together. Precision is related inversely to the dispersion of the sampling distribution – the smaller the dispersion, the greater the precision.

Review Exercises

12.10 Suppose that it is known that a population has $\mu=10$. Use this to answer the questions below. [Answer](#)

- (a) Which is more accurate – four samples with means of 9,10,11, and 9 or means of 6,8,7, and 9?
 - (b) Which is more accurate – four samples with means of 6,14,8, and 12 or means of 8,7,9, and 8?
 - (c) Which is more precise – four samples with means of 7,14,8, and 11 or means of 7,7,9, and 8?
 - (d) How would you judge the accuracy and precision of four samples with means of 2,8,12, and 18?
 - (e) How would you judge the accuracy and precision of four samples with means of 9,10,11, and 10?
 - (f) How would you judge the accuracy and precision of four samples with means of 1,7,8, and 19?
-

12.4 Central Limit Theorem

The sampling distribution of the sample mean was examined in the previous sections by taking all possible samples from a small population (Section 12.1) or taking a large number of samples from a large population (Section 12.2). In both instances, it was observed that the sampling distribution of *the sample mean* was approximately normally distributed, centered on the true mean of the population, and had a standard error that was smaller than the standard deviation of the population and decreased as n increased. In this section, the Central Limit Theorem (CLT) is introduced and explored as a method for identifying the specific characteristics of the sampling distribution of the sample mean without going through the process of extracting multiple samples from the population.

The CLT specifically addresses the shape, center, and dispersion of the sampling distribution of the sample means by stating that $\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ as long as n is “large.” A “large” sample size is defined as,

1. $n \geq 30$,
2. $n \geq 15$ and the population distribution is not strongly skewed, **or**
3. the population distribution is normally distributed.

A major consequence of this definition of “large” is that the sampling distribution of \bar{x} should be normally distributed **no matter what the shape of the population distribution is** as long as the means are computed from samples of $n \geq 30$. The CLT also suggests that \bar{x} is unbiased and that the formula for the $SE_{\bar{x}}$ is $\frac{\sigma}{\sqrt{n}}$ regardless of the size of n . In other words, n impacts the shape of the sampling distribution of the sample means, but not the center or formula for computing the dispersion.

Δ Central Limit Theorem: If a variable x has a population distribution with a mean, μ , and a standard deviation, σ , then the sampling distribution of the sample means (\bar{x}) from random samples of size n , will have a mean equal to μ , a standard error equal to $\frac{\sigma}{\sqrt{n}}$, and a shape that will tend to be normal as n becomes “large.”

12.4.1 Exploring CLT

Characteristics of the sampling distribution of \bar{x} can be explored with the hypothetical population of 1015 lengths of fish in Square Lake. Recall from Section ?? that the population distribution (Figure 2.1) and several parameters are known (Table 2.1) and that the sampling distribution from $n = 50$ was previously examined (Figure 12.5). The effect of changing n on the sampling distribution of sample means can be explored in a similar manner and then examined to determine if the specifics of the CLT appear to be true (Figure 12.8).⁵

Several observations can be made relative to the CLT from the results shown in Figure 12.8. First, the sampling distribution is approximately normal even for very small sample sizes. This happened because the population distribution is only very slightly skewed (Figure 2.1). If the population distribution had been decidedly not normal, then the sampling distributions would only be approximately normally distributed for larger values of n (see next paragraph). Second, the centers (i.e., means) of all simulated sampling distributions are approximately equal to the known $\mu = 98.06$, regardless of sample size. Third, the dispersion of the sampling distributions (i.e., the standard error of the means) becomes smaller with increasing n . Furthermore, the standard errors from the simulated results closely matched the SE expected from the CLT (i.e., $\frac{34.19}{\sqrt{n}}$).

To illustrate that the CLT is not true just for the hypothetical Square Lake population, Figures 12.9 and 12.10 show similar results for samples from a uniform (i.e., rectangular) and a strongly right-skewed exponential distribution, respectively.⁶ For each figure, note how (1) each distribution becomes more “normal” as n increases, (2) the sampling distributions of sample means from the uniform distribution become normal faster (i.e., at a smaller n), (3) each sampling distribution remains centered on approximately the same value for all values of n (approximately 0.5 for the uniform and 1 for the skewed population distribution), (4) each sampling distribution becomes narrower as n increases (i.e., SE gets smaller), and (5) the observed SE is approximately equal to the SE expected from the CLT.

⁵

The effect of changing n can also be explored with the animation in Appendix ??.

⁶

Sampling distributions from theoretical population distributions are further explored with `cltSim()`.

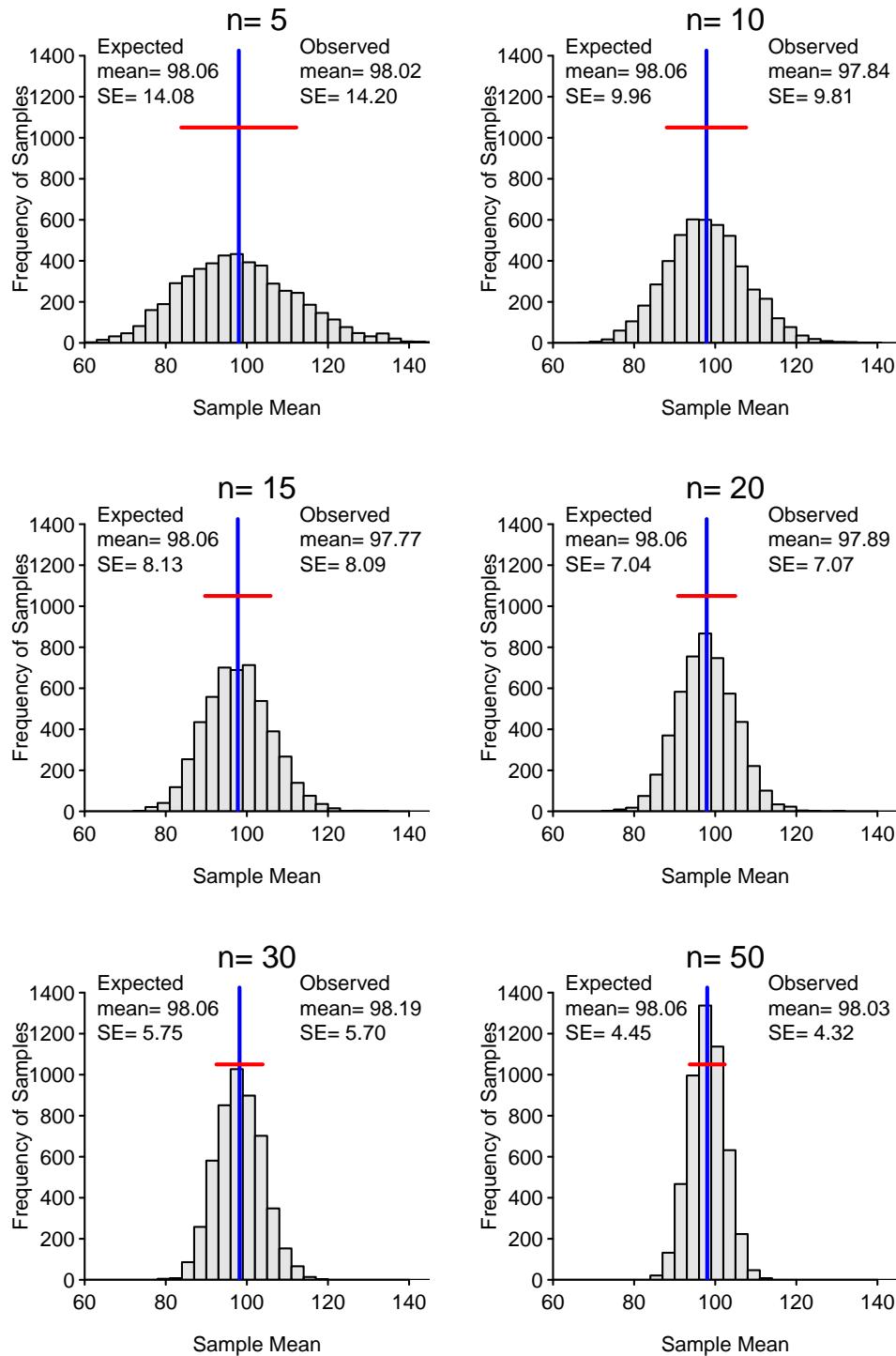


Figure 12.8. Sampling distribution of the sample mean TL simulated from 5000 samples of six different sample sizes extracted from the Square Lake fish population. The vertical blue line is the mean of the 5000 means and the horizontal red line represents $\pm 1\text{SE}$ from the mean.

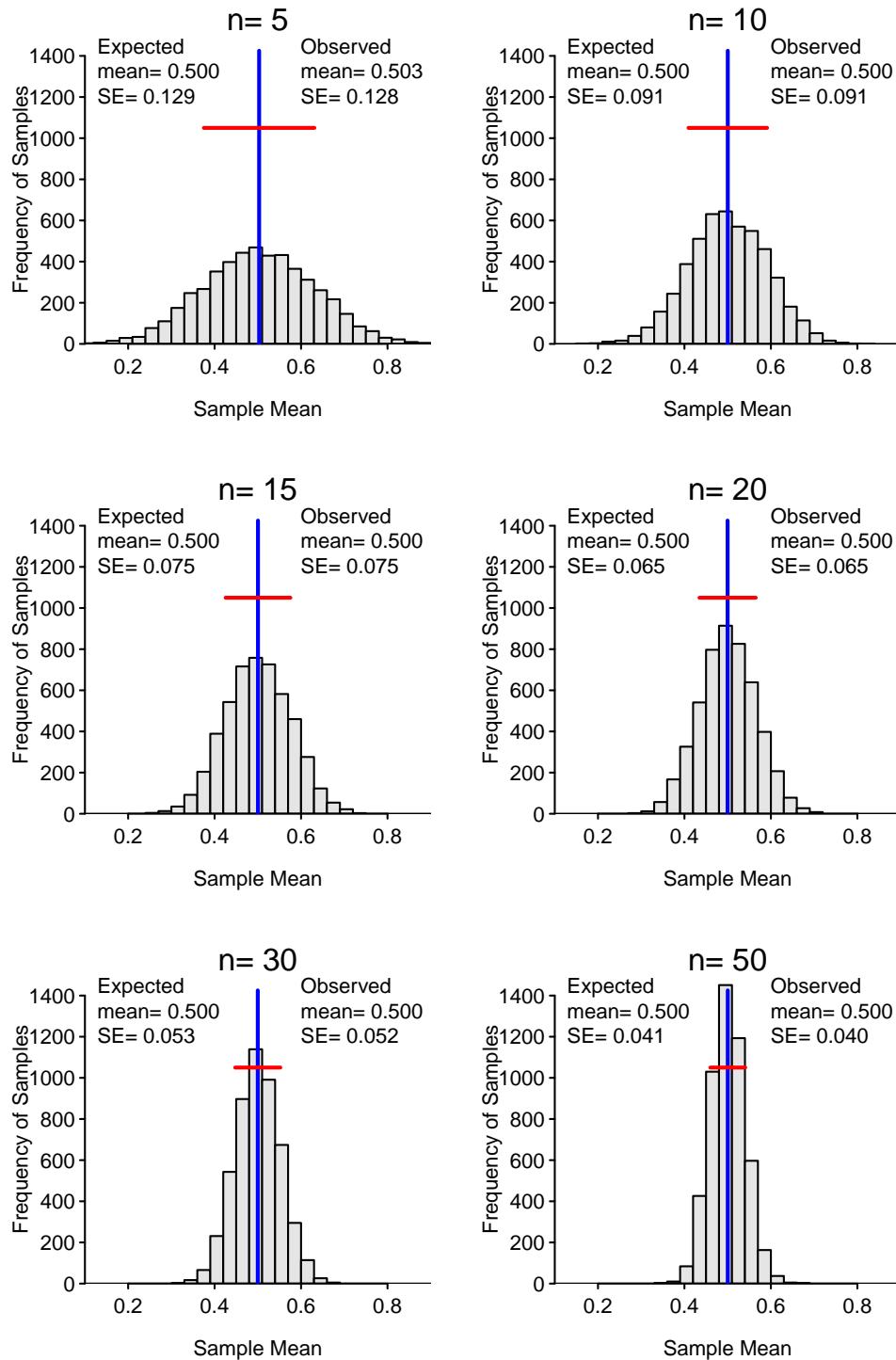


Figure 12.9. Sampling distribution of the sample mean simulated from 5000 samples of six different sample sizes extracted from a uniform population distribution. The vertical blue line is the mean of the 5000 means and the horizontal red line represents $\pm 1SE$ from the mean.

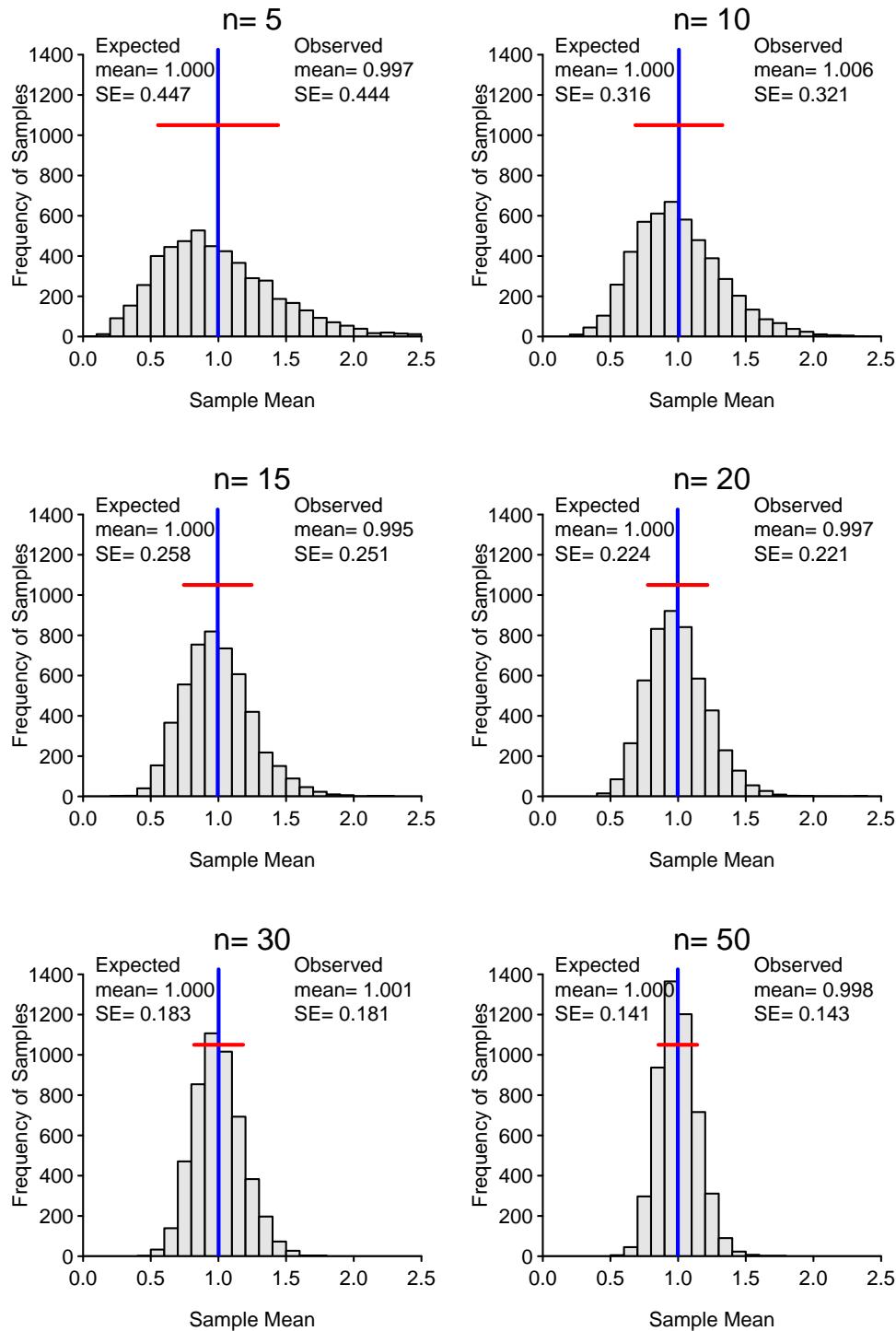


Figure 12.10. Sampling distribution of the sample mean simulated from 5000 samples of six different sample sizes extracted from an exponential population distribution ($\lambda = 1$). The vertical blue line is the mean of the 5000 means and the horizontal red line represents $\pm 1\text{SE}$ from the mean.

Review Exercises

12.11 Assume that the population distribution is $\sim N(100, 20)$ and you take samples of $n = 50$. [Answer](#)

- (a) What shape would you expect the sampling distribution of the sample means to be?
- (b) What do you expect the center of the sampling distribution of the sample means to equal?
- (c) What do you expect the standard deviation of the sampling distribution of the sample means to equal?
- (d) What do you expect the standard error of \bar{x} to equal?

12.12 Assume that the population distribution is skewed to the right with $\mu = 500$ and $\sigma = 60$. Further suppose that samples of $n = 100$ are taken. [Answer](#)

- (a) What shape would you expect the sampling distribution of the sample means to be?
- (b) What do you expect the center of the sampling distribution of the sample means to equal?
- (c) What do you expect the standard deviation of the sampling distribution of the means to equal?
- (d) What do you expect the standard error of \bar{x} to equal?

12.13 Assume that the population distribution is slightly skewed to the right with $\mu = 500$ and $\sigma = 60$. Further suppose that samples of $n = 20$ are taken. [Answer](#)

- (a) What shape would you expect the sampling distribution of the sample means to be?
- (b) What do you expect the center of the sampling distribution of the sample means to equal?
- (c) What do you expect the standard deviation of the sampling distribution of the means to equal?
- (d) What do you expect the standard error of \bar{x} to equal?

12.4.2 Probability Calculations

If the sample size is large enough (either outright or relative to the shape of the population distribution), then the CLT states that the sampling distribution of sample means is approximately normally distributed. If the sampling distribution is normal, then the methods from Chapter 7 may be used. Thus, if the sampling distribution of the sample means is normally distributed, then questions such as “what is the probability of observing a sample mean of less than 95 mm from a sample of size 50 from Square Lake?” can be answered. In other words, questions about **statistics** can be answered.

The question above is answered by first recalling that, for the length of all fish in Square Lake, $\mu = 98.06$ and $\sigma = 31.49$. Because the sample size in the question is greater than 30, the CLT says that the distribution of the sample means from samples of $n = 50$ is $\bar{x} \sim N(98.06, \frac{34.19}{\sqrt{50}})$ or $\bar{x} \sim N(98.06, 4.84)$. Thus, the answer to the question is as simple as computing the area less than 95 on a $N(98.06, 4.84)$ distribution. The proportion of samples of $n = 50$ from Square Lake with an $\bar{x} < 95$ mm is 0.2634 (Figure 12.11) as computed with⁷

⁷

Notice that the standard error of \bar{x} is put into the `sd=` argument of `distrib()`. Recall that a standard error really is a standard deviation, it is just named differently (see Section 12.1). R has no way of knowing whether the question is about an individual or a statistic; it requires the dispersion in either case and calls both of them `sd=`.

```
> ( distrib(95,mean=98.06,sd=34.19/sqrt(50)) )
[1] 0.2634127
```

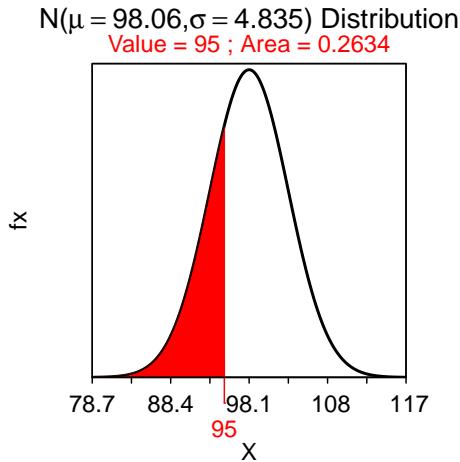


Figure 12.11. Proportion of sample means less than 95 mm on a $N(98.06, 4.84)$ distribution.

- ◊ Calculating the probability of a set of means is as simple as computing areas on a normal distribution as long as the assumptions of the CLT hold true (i.e., n is large enough).

Consider another question – “what is the probability, in a sample of size $n = 40$ from Square Lake, of observing a sample mean of more than 95 mm?” At first glance it may appear that this question can be answered from the work done for the previous question. However, the sample sizes differ between the two questions and, because the sampling distribution depends on the sample size, a different sampling distribution is used here. Because $n > 30$ the sampling distribution will be $\bar{x} \sim N(98.06, \frac{34.19}{\sqrt{40}})$ or $\bar{x} \sim N(98.06, 5.41)$ (Note the different value of the SE). Thus, the answer to this question is the area to the right of 95 on a $N(98.06, 5.41)$ or 0.7143 (Figure 12.12) as computed with

```
> ( distrib(95,mean=98.06,sd=34.19/sqrt(40),lower.tail=FALSE) )
[1] 0.714319
```

- ◊ Always check what sample size is being used – if the sample size changes, then the sampling distribution changes.

Consider one more Square Lake example – “What is the probability that a fish will have a length less than 85 mm?” Note that this question is about an individual, not a statistic from a sample as in the previous questions. Thus, the population distribution and NOT the sampling distribution is the appropriate distribution to use for this question. However, this question cannot be answered because the population distribution is not known to be normally distributed. Two points are illustrated with this example. First, if the question refers to an individual, then the population distribution is used; however, if the question refers to a statistic computed from a sample, then a sampling distribution is used. Second, no matter which

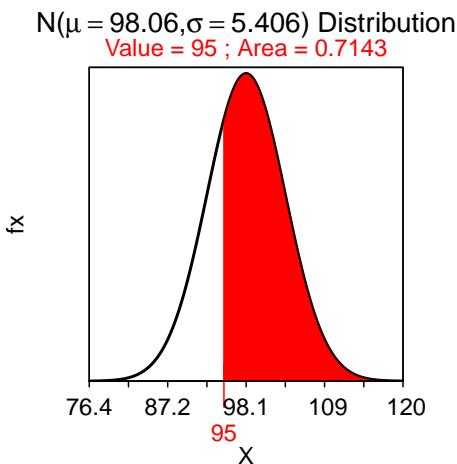


Figure 12.12. Proportion of sample means greater than 95 mm on a $N(98.06, 5, 41)$ distribution.

distribution is used, if the distribution is not known to be normal, then the probability cannot be computed (at least with the techniques in this book).

- ◊ If the question refers to individuals, then use the population distribution. If the question refers to a statistic computed from a sample, then use a sampling distribution.

12.4.3 Calculation Considerations

Suppose that it is known that an Isle Royale loon spends a mean of 13% (this is μ) of its time preening during the summer months. The distribution of time spent preening is strongly right-skewed with a standard deviation of 3% (this is σ). Now consider this question – “What is the probability, in a sample of size 10, of observing a mean time preening of less than 10%?” This is a question about a mean; thus, a sampling distribution should be used. However, because the population distribution is strongly skewed and the sample size is < 30 , the sampling distribution is not known to follow a normal distribution. Therefore, the calculations required to answer this question cannot be made. You cannot make normal distribution calculations on a distribution that is not known to be normal.

- ◊ If the distribution needed to answer a question is not normal, then normal distribution calculations cannot be used to answer the question. The proper answer to the question in this case is to say “I cannot compute the probability because the required distribution is not known to be normal.”

One issue that you have probably noticed while doing these calculations is that they can only be made if the mean, standard deviation, and shape (if $n < 30$) of the population is known. However, remember from Section ?? that one of the main reasons why statistics is important is because the population usually cannot be “seen.” So it should feel “uncomfortable” to assume that so much is known about the population. The only appropriate response to this concern is that we are building towards being able to make inferences with statements based on probabilities that take into account sampling variability. To make these probabilistic

statements we need the skills and concepts about sampling distributions that we have just learned. These skills, while they aren't generally used to answer the questions that we have just practiced on (and will continue to practice on), will be used to make inferences. The goal is to learn these calculations thoroughly here, with these simple questions, so that you can focus on the subtle concepts involved in making inferences in later chapters.

Review Exercises

12.14  Assume that it is known that the distribution of time spent hunting (hours) by an individual Minnesota moose (*Alces alces*) hunter is approximately symmetric in shape with a mean of 40 hours and a standard deviation of 15 hours. Use this information to answer the questions below. [Answer](#)

- (a) Describe what an individual is in this problem.
- (b) List the variable or variables in this problem and identify the type of variable for each.
- (c) What is the probability that a hunter will spend more than 55 hrs hunting moose?
- (d) What is the probability that the average hours spent hunting by a sample of 25 hunters is greater than 48 hrs?

12.15  Facilities management is interested in the mean relative weight (= actual weight / predicted weight; W_r) of fish in the portion of Bay City Creek that runs through the Northland campus. For each question below assume that W_r for fish in the population is $\sim N(1, 0.2)$. [Answer](#)

- (a) What is the population of interest (be very specific)?
- (b) What is the parameter of interest?
- (c) What is the value of the parameter of interest?
- (d) What statistic should be computed to estimate this parameter?
- (e) We can take a random sample of either 25 or 36 fish. Which sample, if either, would tend to produce the most accurate statistic? Why?
- (f) Which sample ($n = 25$ or $= 36$), if either, would tend to produce the most precise statistic? Why?
- (g) What is the exact distribution of the statistic for the n you chose to produce the most precise estimate?
- (h) A mean W_r under 0.95 is indicative of a stressed population. What is the probability of observing a mean W_r that is indicative of a stressed population in Bay City Creek? Use your chosen sample size (here and in the next two questions).
- (i) What are the lower and upper bounds for the most common 95% of W_r values?
- (j) What is the range for the most common 90% of mean W_r values?

12.16  The WI Department of Natural Resources is examining the amount of domestic corn consumed by raccoons per week. Assume that the amount eaten is slightly right-skewed, with a mean of 8 kg, and a standard deviation of 2 kg. [Answer](#)

- (a) What is the probability that a raccoon consumes more than 13 kg per week?
- (b) What is the probability that a sample of 25 raccoons have a mean corn consumption of more than 10 kg per week?
- (c) What is the probability that a sample of 60 raccoons have a TOTAL corn consumption of more than 510 kg per week?

12.17  Suppose that it is known that the number of yards gained per game for the primary running back on a National Football League team is slightly left-skewed with a mean of 82 yards and a standard deviation of 26 yards. [Answer](#)

- (a) What is the probability that a running back will gain more than 100 yards in a single game?
- (b) What is the probability that a running back will average more than 100 yards per game in a 16-game season?
- (c) What is the probability that a running back will average between 70 and 90 yards per game in a 16-game season?

- (d) What is the probability that a running back will average more than 70 yards per game over two 16-game seasons?
- (e) What is the top 25% of yards gained by a running back in a single game?
- (f) What is the top 5% of mean yards gained by a running back in a 16-game season?

12.18  Suppose that the average annual rate of return for a wide array of available stocks is approximately normally distributed with a mean of 4.2% with a standard deviation of 4.9%. [Answer](#)

- (a) What is the probability that five randomly selected stocks produce a positive average rate of return?
- (b) What is the probability that a randomly selected stock produces a positive rate of return?
- (c) What is the probability that ten randomly selected stocks produce a less than 2% average rate of return?
- (d) The top 10% of stocks produce what rate of return?
- (e) The top 10% of random samples of 10 stocks produce what average rate of return?

12.19  Renner (1970) examined the content of hydroxymethylfurfural (HMF) in honey. HMF is an organic compound derived from cellulose without the use of fermentation and is a potential “carbon-neutral” source for fuels. This study found that the distribution of HMF in honey was extremely strongly right-skewed with a mean of 9.5 g/kg and a standard deviation of 13.5 g/kg. [Answer](#)

- (a) What is the probability that one kg of honey have more than 20 g of HMF?
- (b) What is the probability that 20 samples of one kg of honey have an average of more than 20 g of HMF?
- (c) What is the probability that 50 samples of one kg of honey have an average of less than 10 g of HMF?
- (d) What are the 20% least common average amounts of HMF in 50 samples of one kg of honey?

12.20  Allanson (1992) examined the size of farms in England in 1939 and 1989. He found the distribution of farm sizes in 1989 to be very right-skewed with a mean of 65.13 ha and a standard deviation of 108.71 ha. [Answer](#)

- (a) What are the 10% most common sizes of farms in England?
- (b) What are the 10% most common average sizes in samples of 60 farms from England?
- (c) What is the probability that the average size of 60 farms from England is less than 50 ha?
- (d) What is the probability that a farm from England is greater than 50 ha?

12.21  Janzen and Morjan (2002) examined the size of male and female painted turtles (*Chrysemys picta*) at hatching. They found in a sample of 77 turtles that size at hatching was very slightly right-skewed with a mean of 4.46 g with a standard deviation of 0.13 g. Assume that the results of this sample extend to the population to answer the questions below. [Answer](#)

- (a) What is the probability that a turtle will hatch in more than 7 days?
- (b) What is the probability that a sample of 20 turtles will have an average number of days until hatching that is greater than 4.5 days?
- (c) What is the probability that a sample of 50 turtles will have an average number of days until hatching that is greater than 4.5 days?
- (d) What is the mean number of days until hatching such that 20% of samples of 50 turtles have a smaller mean?
- (e) What are the most common 80% of times to hatching?

CHAPTER 13

HYPOTHESIS TESTS

Chapter Objectives:

1. Describe the relationship between the scientific method and statistical hypothesis testing.
2. Properly construct statistical hypotheses.
3. Describe the concept underlying significance testing.
4. Describe possible errors in statistical decision making.

Contents

| | |
|---|-----|
| 13.1 Hypothesis Testing and the Scientific Method | 180 |
| 13.2 Statistical Hypotheses | 180 |
| 13.3 Test Statistics and Effect Sizes | 186 |
| 13.4 Hypothesis Testing Concept Summary | 186 |
| 13.5 Errors and Power | 188 |

A STATISTIC IS, BECAUSE OF SAMPLING VARIABILITY, an imperfect estimate of the unknown parameter. Thus conclusions about the parameter from the statistics could be wrong and should include a measure of the precision of the estimate. There are two possible calculations using the results of a single sample that recognize this imperfection and allow conclusions about parameters. First, a researcher may form an *a priori* hypothesis about the parameter and then use the information in the sample to make a judgment about the “correctness” of this hypothesis. Second, a researcher may form, from the information in the sample, a range of values that is likely to contain the parameter. The methods of the first calculation are called *hypothesis testing* and are the subject of this chapter. The second method consists of constructing a *confidence region*, which is introduced in Chapter 14. Specific applications of these two techniques are described in Chapters 15-19.

13.1 Hypothesis Testing and the Scientific Method

In its simplest form, the scientific method has four steps:

1. Observation and description of a natural phenomenon.
2. Formulation of a hypothesis to explain the phenomenon.
3. Use of the hypothesis to predict new observations.
4. Performance of experimental tests of the predictions.

If the predictions do not match the results of the experiment, then the hypothesis is rejected and an alternative hypothesis is proposed. If the predictions do closely match the results of the experiment, then belief in the hypothesis is gained, but the hypothesis will likely be subjected to further scrutiny.

Statistical hypothesis testing is key to performing the scientific method and, in fact, closely follows the scientific method in concept. Statistical hypothesis testing begins by taking a research hypothesis and formulating it into competing statistical hypotheses. The null hypothesis is used to make a prediction about a parameter of interest. Data is then collected and statistical methods are used to determine whether the observed statistic closely matches the predictions made from the null hypothesis or not. Probability is used in the measure of matching and sampling variability is taken into account. This process and the theory underlying statistical hypothesis testing is explained in the following sections

13.2 Statistical Hypotheses

The hypotheses in a research study are classified into two types: (1) research hypothesis and (2) statistical hypotheses. A research hypothesis is usually a general statement, stated in prose, about the phenomenon that the researcher is interested in investigating. The research hypothesis must be transferred into statistical hypotheses that are more mathematical and, thus, more easily subjected to statistical methods.

Δ **Research Hypothesis:** A very generally stated hypothesis about the phenomenon of interest.

Δ **Statistical Hypothesis:** A specific mathematically stated hypothesis about the phenomenon of interest.

There are two types of statistical hypotheses: (1) the null hypothesis and (2) the alternative hypothesis. The **null hypothesis**, typically abbreviated with H_0 , is a specific statement of no difference between a parameter and a specific value or between two parameters. Because H_0 is always the “no difference” situation it always contains the equals sign. The **alternative hypothesis**, usually abbreviated as H_A always states that there is a difference of some sort between a parameter and a specific value or between two parameters. The exact type of difference comes from the researcher’s question of interest and results in the use of a less than ($<$), greater than ($>$), or not equals (\neq) sign. The phenomenon described in the research hypothesis is most often mathematically formalized in the H_A .

Δ **Null Hypothesis:** A statistical hypothesis that states specifically that there is no difference between a parameter and a specific value or between two parameters; typically abbreviated with H_0 .

◊ Null hypotheses always represent the “no difference” situation and, thus, always contain an equals sign.

Δ **Alternative Hypothesis:** A statistical hypothesis that states a specific difference between a parameter and a specific value or between two parameters; typically abbreviated with H_A .

◊ Alternative hypotheses always represent some sort of difference and, thus, always contain one of these three directional symbols (\neq , $>$, and $<$).

The relationships between the research, null, and alternative hypotheses are illustrated with the following examples (the research hypothesis is listed first):

1. A medical researcher wants to determine if a new medicine has any undesirable side effects, particularly does it change the patients’ mean pulse rate.
 - $H_A : \mu \neq 82$ and $H_0 : \mu = 82$ (where μ represents the mean pulse rate and 82 is the “known” mean pulse rate in the population under study; thus, this alternative hypothesis represents a change from the “normal” pulse rate).
2. A chemist has invented an additive to automobile batteries that is intended to extend the average life of the battery.
 - $H_A : \mu > 36$ and $H_0 : \mu = 36$ (where μ represents the mean battery life and 36 is the “known” mean battery life of batteries without the new additive; thus, this alternative hypothesis represents an extension of the current battery life).
3. An engineer wants to determine if a new type of insulation will reduce the average heating costs of a typical house.
 - $H_A : \mu < 78$ and $H_0 : \mu = 78$ (where μ represents the mean monthly heating bill and 78 is the “known” mean monthly heating bill without the new insulation; thus, this alternative hypothesis represents a decline in heating bills from the previous “normal” amount).

◊ There are always two competing statistical hypotheses – null and alternative hypotheses.

The sign used in the alternative hypothesis comes directly from the wording of the research hypothesis (Table 13.1). An alternative hypothesis that contains the \neq sign is called a **two-tailed alternative** as the value can

be “not equal” to another value in two ways; i.e., less than or greater than the value. Alternative hypotheses with the $<$ or the $>$ signs are called **one-tailed alternatives**. The null hypothesis is easily constructed from the alternative hypothesis by replacing the sign in the alternative hypothesis with an equals sign.

Table 13.1. Common words that indicate which sign to use in the alternative hypothesis.

| $>$ | $<$ | \neq |
|------------------|------------------|--------------------|
| is greater than | is less than | is not equal to |
| is more than | is below | is different from |
| is larger than | is lower than | has changed from |
| is longer than | is shorter than | is not the same as |
| is bigger than | is smaller than | |
| is better than | is reduced from | |
| is at least | is at most | |
| is not less than | is not more than | |

- ◊ The “not-equals” alternative is called a two-tailed alternative, whereas the other two alternative hypotheses are called one-tailed alternatives.

Review Exercises

- 13.1** A researcher is investigating the mean growth of a certain cactus under a variety of environmental conditions. Under the current environmental conditions, he hypothesizes that mean growth is no more than 4 cm. What is H_0 and H_A in this situation? [Answer](#)
- 13.2** Machowiak *et al.* (1992) critically examined the belief that the mean body temperature differed from 98.6°F by measuring the body temperatures of 93 healthy humans. What is H_0 and H_A in this situation? [Answer](#)
- 13.3** A study by Cheshire *et al.* (1994) reported on six patients with chronic myofascial pain syndrome. The authors were examining the hypothesis that the mean pain length was greater than 2.5 years. What is H_0 and H_A in this situation? [Answer](#)

13.2.1 Hypothesis Testing Concept

Statistical hypothesis testing begins by using the null hypothesis to make a prediction of what value one should expect for the mean in a sample. So, for the Square Lake example, if $H_0 : \mu = 105$ and $H_A : \mu < 105$, then one would expect, if the null hypothesis is true, that the observed mean in a sample would be 105. If sampling variability did not exist and the observed sample mean was NOT equal to 105, then the prediction based on the null hypothesis would not be supported and the conclusion would be that the null hypothesis is incorrect. In other words, one could conclude that the population mean was not equal to 105.

Of course, sampling variability does exist and its existence complicates matters. The simple interpretation of not supporting H_0 because the observed sample mean did not equal the hypothesized population mean canNOT be made because, with sampling variability, one would not expect a statistic to exactly equal the parameter in the population from which the sample was extracted. For example, even if the null hypothesis was correct, then one would not expect, with sampling variability, the observed sample mean to exactly equal 105; rather, one would expect the observed sample mean to be **reasonably** close to 105.

Thus, hypothesis testing is a procedure for determining if the difference between the observed statistic and the expected statistic based on the null hypothesis is “large” **relative to sampling variability**. For example, the standard error of \bar{x} in samples of $n = 50$ in the Square Lake example is equal to $\frac{\sigma}{\sqrt{n}} = \frac{31.5}{\sqrt{50}} = 4.45$. Thus, with this amount of sampling variability, an observed sample mean of 103 would be considered reasonably close to 105 and one would have more belief in $H_0 : \mu = 105$. However, an observed sample mean of 70 would be considered further away from 105 than one would expect based on sampling variability alone and the belief in $H_0 : \mu = 105$ would lessen.

While the above procedure is intuitively appealing, it loses some of its objectivity when the examples chosen (i.e., samples means of 103 and 70) are not as extremely close or distant from the null hypothesized value (e.g., what would the “conclusion” be if the observed sample mean was 97?). A first step in creating a more objective decision criteria is to compute a value called the “p-value.” A p-value is defined as the probability of the observed statistic or a value of the statistic more extreme assuming that the null hypothesis is true. A more detailed description of the p-value is warranted given the centrality of the p-value to making conclusions about statistical hypotheses.

Δ p-value: The probability of the observed statistic or a value of the statistic more extreme assuming the null hypothesis is true.

The meaning of the phrase “or more extreme” is derived from the sign in H_A (Figure 13.1). If H_A is the “less than” situation, then “or more extreme” means “less than” or “shade to the left” for the probability calculation. The “greater than” situation is defined similarly but would result in shading to the “right.” In the “not equals” situation, “or more extreme” means further into the tail AND the exact same size of tail on the other side of the distribution. It should be clear from Figure 13.1 why the “less than” and “greater than” alternatives are called one-tailed hypotheses and the “not equals” alternative is called a two-tailed hypothesis.

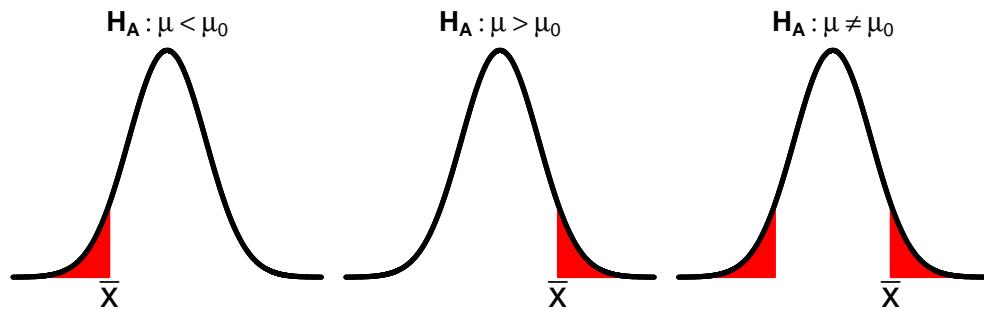


Figure 13.1. Depiction of the meaning of “or more extreme” in the calculation of p-values for the three possible alternative hypotheses.

The “assuming that the null hypothesis is true” phrase is used to define a μ for the sampling distribution on which the p-value will be calculated. This sampling distribution is called the **null distribution** because it

depends on the value of μ in the null hypothesis. One must remember that the null distribution represents the distribution of all possible sample means assuming that the null hypothesis is true; it does NOT represent the actual sample means¹. The null distribution in the Square Lake example is thus $\bar{x} \sim N(105, 4.45)$ because $n = 50 > 30$, $H_0 : \mu = 105$, and $SE = \frac{31.49}{\sqrt{50}} = 4.45$.

The p-value is computed with a “forward” normal distribution calculation on the null sampling distribution. For example, suppose that a sample mean of 100 was observed in a sample of $n = 50$ from Square Lake (as it was in Table 2.2). The definition of the p-value in this particular instance would be “the probability of observing $\bar{x} = 100$ or a smaller value assuming that $\mu = 105$.” The answer to this definition is computed by finding the area to the left of 100 on a $N(105, 4.45)$ null distribution. This is the exact same type of calculation that was made on sampling distributions in Section 12.4.2. Thus, a p-value of $p = 0.1308$ is computed and visualized (Figure 13.2) with

```
> ( distrib(100,mean=105,sd=31.49/sqrt(50)) )
[1] 0.1307722
```

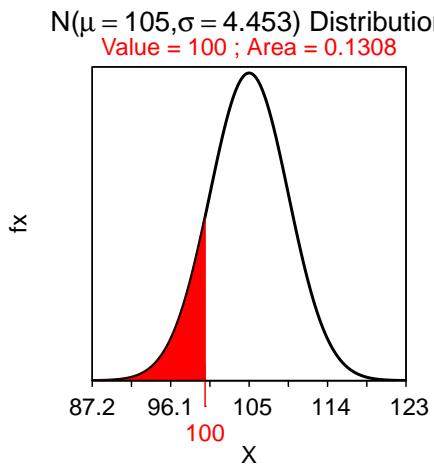


Figure 13.2. Depiction of the p-value for the Square Lake example where $\bar{x} = 100$ and $H_A : \mu < 105$.

Interpreting the p-value requires critically thinking about the p-value definition and how it is calculated. Small p-values appear when the observed statistic is “far” from the value expected from the null hypothesis. In this case there is a small probability of seeing the observed statistic ASSUMING that H_0 is true. Thus, the assumption is likely wrong and H_0 is likely incorrect. In contrast, large p-values appear when the observed statistic is close to the null hypothesized value suggesting that the assumption about H_0 may be correct.

- ◊ Small p-values are evidence against the null hypothesis.

The p-value serves as a numerical measure on which to base a conclusion about H_0 . To do this objectively requires an objective definition of what it means to be a “small” or “large” p-value. Statisticians use a cut-off value, called the rejection criterion which is symbolized as α , such that p-values less than α are considered

¹

Of course, unless the null hypothesis happens to be perfectly true.

small and would result in the rejection of H_0 as a viable hypothesis. The value of α is typically small, usually set at 0.05, although $\alpha = 0.01$ and $\alpha = 0.10$ are also commonly used.

$\Delta \alpha$: A predetermined rejection criterion value used in hypothesis testing. This value sets the “cutoff” for determining whether it was reasonable to have seen the observed statistic or not assuming the null hypothesis is true.

◊ Typical values of α are 0.01, 0.05, and 0.10.

The choice of α is made by the person conducting the hypothesis test and is based on how much evidence a researcher demands before rejecting H_0 . Smaller values of α require a larger difference between the observed statistic and the null hypothesized value and, thus, require “more evidence” of a difference for the H_0 to be rejected. For example, if a rejection of the null hypothesis will be heavily scrutinized by regulatory agencies, then the researcher may want to be very sure before claiming a difference and should then set α at a smaller value, say $\alpha = 0.01$. The actual choice for α MUST be made before collecting any data and canNOT be changed once the data has been collected. In other words, once the data are in hand, a researcher cannot lower or raise α to achieve a desired outcome regarding H_0 .

◊ The value of the rejection criterion (α) is set by the researcher BEFORE data is collected.

◊ Set α to lower values to make it more difficult to reject H_0 .

The null hypothesis in the Square Lake example was not rejected because the calculated p-value (i.e., $p = 0.1308$) is larger than any of the common values of α . Thus, the conclusion in this example is that it is possible that the mean of the entire population is equal to 105 and it is not likely that the population mean is less than 105. In other words, observing a sample mean of 100 is likely to happen based on random sampling variability alone and it is unlikely that the null hypothesized value is incorrect.

Review Exercises

13.4 Compute the p-value and make a decision about H_0 with the following information – $\alpha = 0.10$, $H_0 : \mu = 10$, $H_A : \mu > 10$, $\sigma = 5$, $n = 25$, and $\bar{x} = 12.1$. [Answer](#)

13.5 Compute the p-value and make a decision about H_0 with the following information – $\alpha = 0.05$, $H_0 : \mu = 50$, $H_A : \mu < 50$, $\sigma = 20$, $n = 50$, and $\bar{x} = 43.8$. [Answer](#)

13.6 Compute the p-value and make a decision about H_0 with the following information – $\alpha = 0.01$, $H_0 : \mu = 100$, $H_A : \mu \neq 100$, $\sigma = 15$, $n = 100$, and $\bar{x} = 98$. [Answer](#)

13.7 Describe why we must formally go through the steps of a hypothesis test to conclude that $\mu > 11$ when we observe $\bar{x} = 12.1$. [Answer](#)

13.3 Test Statistics and Effect Sizes

Instead of just reporting the observed statistic and the resulting p-value, it may be of interest to a researcher to know how “far” the observed statistic was from the hypothesized value of the parameter. This simplistic value is easily calculated with

$$\text{Observed Statistic} - \text{Hypothesized Parameter}$$

where “Hypothesized Parameter” represents the specific value in H_0 . However, the meaning of this value is difficult to interpret without an understanding of the standard error of the statistic. For example, a difference of 10 between the observed statistic and the hypothesized parameter seems “very different” if the standard error is 1 but does not seem “different” if the standard error is 100. Thus, it is common practice to standardize this difference by dividing by the standard error of the statistic. This measure of distance is called a *test statistic* and is generalized with

$$\text{Test Statistic} = \frac{\text{Observed Statistic} - \text{Hypothesized Parameter}}{SE_{\text{Statistic}}} \quad (13.3.1)$$

Thus, the test statistic (13.3.1) measures how many standard errors the observed statistic is away from the hypothesized parameter. Relatively large values are indicative of a difference that is likely not due to randomness (i.e., sampling variability) and suggest a rejection of the null hypothesis. There are other forms for calculating test statistics², but all test statistics retain the general idea of scaling the difference between what was observed and what was expected from the null hypothesis in terms of sampling variability. Even though there is a one-to-one relationship between a test statistic and a p-value, a test statistic is often reported with a hypothesis test to give another feel for the magnitude of the difference between what was observed and what was predicted.

- ◊ A test statistic measures how many standard errors the observed statistic is away from the hypothesized parameter.

13.4 Hypothesis Testing Concept Summary

In summary, hypothesis tests are statistically examined with the following procedure.

1. Construct null and alternative statistical hypotheses from the research hypothesis.
2. Construct an expected value of the statistic based on the null hypothesis (i.e., assume that the null hypothesis is true).
3. Calculate an observed statistic from the individuals in a sample.

²

A few of which we will see in this class.

4. Compare the difference between the observed statistic and the expected statistic based on the null hypothesis in relation to sampling variability (i.e., calculate a test statistic and p-value).
5. Use the p – value to determine if this difference is “large” or not .
 - If this difference is “large” (i.e., p – value $< \alpha$), then reject the null hypothesis.
 - If this difference is not “large” (i.e., p – value $> \alpha$), then “Do Not Reject” the null hypothesis.

When the p -value $> \alpha$, suggesting little evidence against the null hypothesis, you must be very careful to say “do not reject H_0 ” rather than “accept H_0 as true.” There are two reasons for this very specific choice of words.

First, there are several other possible values, besides the specific value of the null hypothesis, that would lead to “do not reject” conclusions. For example, if a null hypothesized value of 105 was not rejected, then values of 104.99, 104.98, etc. would also likely not be rejected³ So, we don’t want to say that we “accept” a particular hypothesized value when we know many other values would also be “accepted.”

Second, the null hypothesis is almost always not true. Consider the null hypothesis of the Square Lake example (i.e., “that the mean length is 105”). The mean length of fish in Square Lake is undoubtedly not exactly equal to 105. It may be 104.9, 105.01, or some other more disparate value. The point is that the specific value of the hypothesis is likely never true, especially for a continuous variable. The problem is that it takes large amounts of data to be able to distinguish means that are very close to the true population mean (i.e., it is difficult to distinguish between 104.9 and 105 when sampling variability is present). Very often we will not take a sample size large enough to distinguish these subtle differences. Thus, we will say that we “do not reject H_0 ” because there simply was not enough data to reject it.

Review Exercises

- 13.8** The managers of a wastewater treatment plant monitored the amount of biological oxygen demand (BOD; lbs/day) in the effluent of the plant each month from January 1991 to October 2000. The managers would need to take corrective actions if the average BOD over this time period was significantly greater than 2200 lbs/day at a 10% rejection level. Previous studies indicated that the standard deviation was 1200 lbs/day. Summary statistics from their sample of days is given below. Use this information to answer the questions below. [Answer](#)

| n | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|-----|------|---------|--------|------|---------|------|
| 118 | 630 | 1600 | 2240 | 2504 | 3193 | 6023 |

- (a) What are the null and alternative hypotheses?
- (b) What is the test statistic?
- (c) Compute the p-value.
- (d) Use the p-value to make a decision about H_0 .
- (e) What does this mean for the managers of the plant (i.e., will they need to take action)? Explain!

- 13.9** Admissions representatives at the University of Minnesota medical school were concerned that the average grade point average of applicants in non-science courses had dropped below 3.7. A sample of 40 of the most recent applicants indicated that the mean was 3.60. Information from the Association of American Medical Colleges suggested that the overall standard deviation was 0.35. Use this information, and an $\alpha = 0.05$, to answer the questions below. [Answer](#)

³

In fact, for example, the values in a 95% confidence interval – see next section – represent all possible hypothesized values that would not be rejected with a two-tailed alternative hypothesis using $\alpha = 0.05$.

- (a) What are the null and alternative hypotheses?
 - (b) What is the test statistic?
 - (c)  Compute the p-value.
 - (d) Use the p-value to make a decision about H_0 .
 - (e) Was the representatives concern about the average gpa of applicants warranted? Explain!
-

13.5 Errors and Power

The goal of hypothesis testing is to make a decision about H_0 . Unfortunately, because of sampling variability, there is always a risk of making an incorrect decision. Two types of incorrect decisions can be made (Table 13.2). A Type I error occurs when a true H_0 is falsely rejected. In other words, even if H_0 is true, there is a chance that a rare sample will occur and H_0 will be deemed incorrect. The probability of making a Type I error is set when α is chosen. A Type II error occurs when a false H_0 is not rejected. The probability of a Type II error is denoted by β .

Table 13.2. Types of decisions that can be made from a hypothesis test.

| | | Decision from Data | |
|------------------------|-------|--------------------|------------|
| | | Reject | Not Reject |
| Truth About Population | H_0 | Type I | Correct |
| | H_A | Correct | Type II |

Δ Type I error: Rejecting H_0 when H_0 was actually true. Probability of Type I error is α .

Δ Type II error: Not rejecting H_0 when H_0 was actually false. Probability of Type II error is β .

The decision in the Square Lake example above produced a Type II error because $H_0 : \mu = 105$ was not rejected even though we know that $\mu = 98.06$ (Table 2.1). Unfortunately, in real life, it will never be known exactly when a Type I or a Type II error has been made because the true μ is not known. However, it is known that a Type I error will be made $100\alpha\%$ of the time. The probability of a type II error (β), though, is never known because this probability depends on the true μ . Decisions can be made, however, that affect the magnitude of β (discussed below with power).

A concept that is very closely related to decision-making errors is the idea of **power**. Power is the probability of correctly rejecting a false H_0 . In other words, it is the probability of detecting a difference from the hypothesized value if a difference really exists. That is, power is used to demonstrate how sensitive a hypothesis test is for identifying a difference. High power related to a H_0 that is not rejected implies that the H_0 really should not have been rejected. Conversely, low power related to a H_0 that was not rejected implies that the test was very unlikely to detect a difference, so not rejecting H_0 is not surprising nor particularly conclusive.

Δ **Power:** The probability of correctly rejecting H_0 when H_0 was actually false.

Power is equal to $1 - \beta$ and, thus, like β it cannot be computed directly. However, a researcher can make decisions that will positively affect power (Figure 13.3, Appendix ??). For example, a researcher can positively impact power by increasing α or by increasing n . Increasing n is a more beneficial alternative as it does not result in an increase in Type I errors as increasing α would. In addition, power decreases as the difference between the hypothesized mean (μ_0) and the actual mean (μ_A) decreases. This means that the ability to detect increasingly smaller differences decreases. In addition, power decreases with an increasing amount of natural variability (i.e., σ). In other words, the ability to detect a difference decreases with increasing amounts of variability among individuals. A researcher cannot control the difference between μ_0 and μ_A or the value of σ . However, it is important to know that if a situation with a “large” amount of variability is encountered or the difference to be detected is small, the researcher will need to increase n to gain power.

◊ Power = $1-\beta$.

◊ Power will increase as the difference between the actual and hypothesized value of the parameter increases.

◊ Power will increase as the standard error of the statistic decreases. Thus, power increases as the sample size increases.

◊ Power will increase as the α level increases.

Power cannot usually be calculated because the actual mean (μ_A) is usually not known. However, in the Square Lake example, μ_A is known and power can be calculated in four steps:

1. Draw the sampling distribution assuming the H_0 is true (called the null distribution).
 - The null distribution is $N(105, \frac{31.49}{\sqrt{50}})$ because $H_0 : \mu = 105$, $\sigma = 31.49$, and $n = 50$.
2. Find the rejection region borders (based on α and H_A) in terms of the value of the statistic (a “reverse” calculation on the null distribution).
 - The rejection region is delineated by the \bar{x} that has $\alpha = 0.10$ to the left (because H_A is a “less than”) of it. The Z with 0.10 to the left of it is -1.282. Thus, the \bar{x} on the null distribution with 0.10 to the left of it is $-1.282 \frac{31.49}{\sqrt{50}} + 105 = 99.2908$.
3. Draw the sampling distribution corresponding to the “actual” parameter value (SE is the same as that for the null distribution).
 - The actual μ is 98.06. Thus, the actual sampling distribution is $N(98.06, \frac{31.49}{\sqrt{50}})$.
4. Compute the portion of the “actual” sampling distribution in the REJECTION region of null distribution (i.e., a “forward” calculation on the actual distribution).
 - This computation is to find the area to the left of 99.2908 on $N(98.06, \frac{31.49}{\sqrt{50}})$. The corresponding Z is $\frac{99.2908 - 98.06}{\frac{31.49}{\sqrt{50}}} = 0.29$. The area to the left of this Z is 0.6141.

Thus, the power to detect a $\mu_A = 98.06$ was 0.6141 (the value in Figure ?? is slightly more accurate because the approximate Z table was not used). This means that only about 61% of the time will the

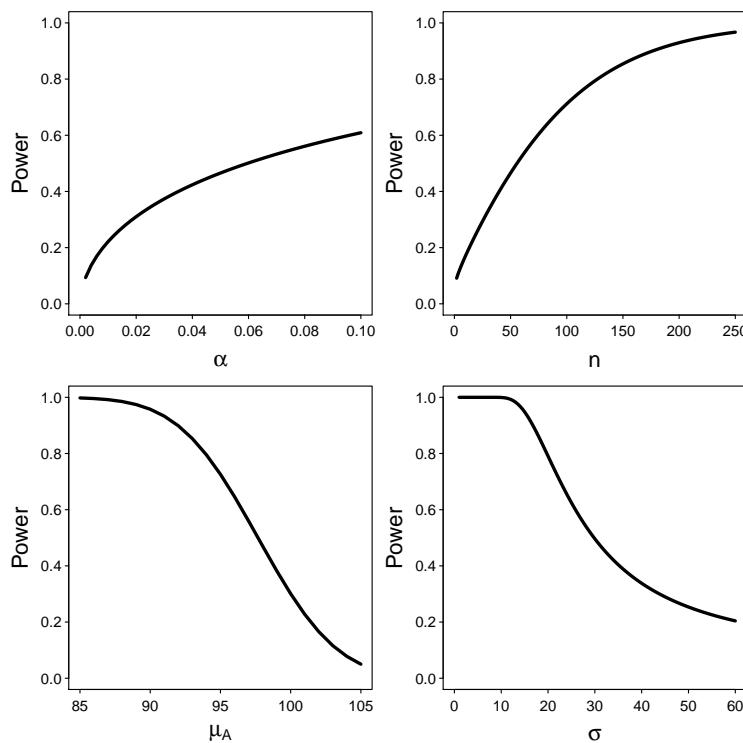


Figure 13.3. The relationship between one-tailed (lower) power and α , n , actual mean (μ_A), and σ . In all situations where the variable does not vary, $\mu_0 = 105$, $\mu_A = 98.06$, $\sigma = 31.49$, $n = 50$, and $\alpha = 0.05$.

false $H_0 : \mu = 105$ be correctly rejected. Thus, it is not too surprising that we did not reject H_0 in this example. If n could be doubled to 100, however, the power to correctly reject $H_0 : \mu = 105$ would increase to approximately 0.82 (Figure 13.3).

Review Exercises

13.10 What is β if power=0.875? [Answer](#)

13.11 For a constant sample size, σ , and difference between the hypothesized and actual means, what happens to power, if α is increased? [Answer](#)

13.12 For a constant α , σ , and difference between the hypothesized and actual means, what happens to power, if the sample size increases? [Answer](#)

13.13 For a constant α , σ , and sample size, what happens to power if the difference between the hypothesized and actual means increases? [Answer](#)

13.14 For a constant sample size, σ , and difference between the hypothesized and actual means, what happens to β , if α is increased? [Answer](#)

- 13.15** For a constant α , σ , and difference between the hypothesized and actual means, what happens to β , if the sample size is increased? [Answer](#)
- 13.16** For a constant α , σ , and sample size, what happens to β if the difference between the hypothesized and actual means increases? [Answer](#)
- 13.17** Describe a real-life situation where you think that making a Type II error would be much more “costly” than making a Type I error. Completely describe the situation at hand and what Type I and a Type II errors mean in terms of the situation you describe. [Answer](#)
- 13.18** Compute power given the following information: $\alpha = 0.05$, $H_0 : \mu = 50$, $H_A : \mu < 50$, $\sigma = 20$, $n = 50$, and $\mu_A = 45$. [Answer](#)
- 13.19** Compute power given the following information: $\alpha = 0.10$, $H_0 : \mu = 10$, $H_A : \mu > 10$, $\sigma = 5$, $n = 25$, and $\mu_A = 12$. [Answer](#)
- 13.20** Compute power given the following information: $\alpha = 0.01$, $H_0 : \mu = 75$, $H_A : \mu > 75$, $\sigma = 15$, $n = 30$, and $\mu_A = 82$. [Answer](#)
-

CHAPTER 14

CONFIDENCE REGIONS

Chapter Objectives:

1. Describe the concept underlying confidence intervals.
2. Construct confidence intervals for parameters.
3. Use the confidence interval formula to estimate desired sample sizes.

Contents

| | |
|--|-----|
| 14.1 Confidence Concept | 193 |
| 14.2 Constructing Confidence Regions | 196 |
| 14.3 Inference Type Relationship | 200 |
| 14.4 Precision and Sample Size | 201 |

THE FINAL RESULT FROM A HYPOTHESIS TEST (Chapter ??) can be somewhat uneventful – i.e., the conclusion is either that the parameter may be equal to or that the parameter is different from the hypothesized value¹. If the parameter is thought to be different from the hypothesized value we might go as far as to say that our best guess at the parameter is the value of our observed statistic. However, as has been seen many times, a statistic is, because of sampling variability, an imperfect estimate of the unknown parameter. Thus, this imperfection can be recognized by computing, from the results of a sample, a range of values that is likely to contain the parameter. This range is called a confidence region for the unknown parameter. For example, we may make a statement such as this – “Our best guess for the true population mean length of fish in Square lake is the sample mean of 98.5 mm; however, we are 95% confident that the mean of ALL fish in the lake is between 95.9 and 101.1 mm.” This last statement is the interpretation of a confidence interval and is important because it acknowledges sampling variability in the inferential statement. In this section, the concept, calculation, and interpretation of confidence regions will be explored.

14.1 Confidence Concept

A complete understanding of what it means to be “95% confident” requires examining multiple samples from a population in much the same way as how the concept of sampling variability was explored in Chapter 12. For the sake of simplicity in this exploration, the discussion here will be restricted to a confidence interval (CI) where a range, bounded on both ends, is computed. In addition, a 95%, rather than a more general value, CI will be used. General methods for constructing confidence regions of different types with different levels of confidence will be discussed thoroughly in the next section. These simplifying restrictions and the unrealistic idea that population values are known are made here only so that the **concept** of confidence intervals can be explored more easily.

Define a 95% CI for μ as $\bar{x} \pm 2SE_{\bar{x}}$. In addition, as concern rests with whether a CI contains μ or not, recall that $\mu=98.06$ and $\sigma=31.49$ for the Square Lake population (Table 2.1). Further recall from Table 2.2 that the first sample of $n=50$ from the Square Lake population resulted in $\bar{x} = 100.04$. Using the CI formula above, the associated 95% CI is $100.04 \pm 2 \frac{31.49}{\sqrt{50}}$, 100.04 ± 8.91 , or $(91.13, 108.95)$. In this exploratory example μ is known and, thus, it can be said that this interval does indeed contain μ . In other words, this particular CI accomplishes what it was intended to do, i.e., provide a range that contains μ .

Despite the success observed in this first sample, not all confidence intervals will contain μ . For example, four out of 100 95% confidence intervals shown in Figure 14.1 did not contain μ . Thus, four times in these 100 samples the researcher would have concluded that μ was in an incorrect interval. The concept of “confidence” in confidence regions is related to determining how often these types of mistakes are made.

From the Central Limit Theorem, the sampling distribution of \bar{x} for samples of $n=50$ is $N(98.06, \frac{31.49}{\sqrt{50}})$ or $N(98.06, 4.45)$ for this known population. According to the 68-95-99.7% Rule, it is known that 95% of the sample means in this sampling distribution will be between $\mu \pm 2SE$ or, in this specific case, between $98.06 \pm 2(4.45)$. The sampling distribution and this range of expected sample means is shown at the top of Figure 14.1. In addition, the range of expected sample means is extended down through all of the CI lines in Figure 14.1. Note that any sample that produced a sample mean (solid dot on the CI line) inside of the expected range of sample means also produced a 95% CI that contained μ (i.e., blue CI line). Thus, because 95% of the sample means will be within the expected range of sample means, 95% of the 95% CIs will contain μ . So, “95% confident” means that 95% of all 95% CIs will contain the parameter and 5% will not. In other words, the mistake identified above will be made with 5% of all 95% confidence intervals.

¹

Depending on the H_A it may be known if the parameter is more or less than the hypothesized value.

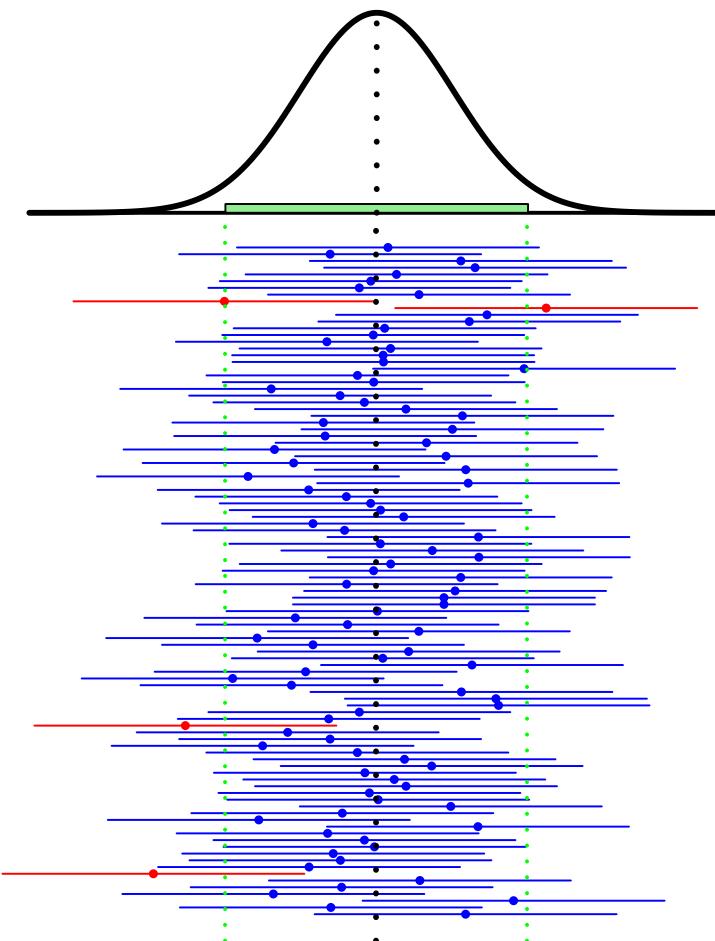


Figure 14.1. Sampling distribution of the sample mean (top) and 100 random 95% confidence intervals (horizontal lines) from samples of $n=50$ from the Square Lake population. Confidence intervals that do NOT contain $\mu=98.06$ are shown in red.

The specifics for constructing confidence regions with different levels of confidence will be described below. However, at this point, it should be noted that the number of CIs expected to contain the parameter of interest is set by the level of confidence used to construct the CI. For example, 80% of 80% CIs and 90% of 90% CIs will contain the parameter of interest. In either case, a particular CI either does or does not contain the interval and, in real-life, we will never know whether it does or does not (i.e., we won't know the value of the parameter). However, we do know that the technique (i.e., the construction of the CI) will “work” (i.e., contain the parameter) a set percentage of the time. To reiterate this point, examine the 100 90% CIs (Figure 14.2-Left) and 100 80% CIs (Figure 14.2-Right) for the Square Lake fish length data.

- ◊ The number of confidence intervals expected to contain the parameter of interest is set by the level of confidence used to construct the confidence interval.

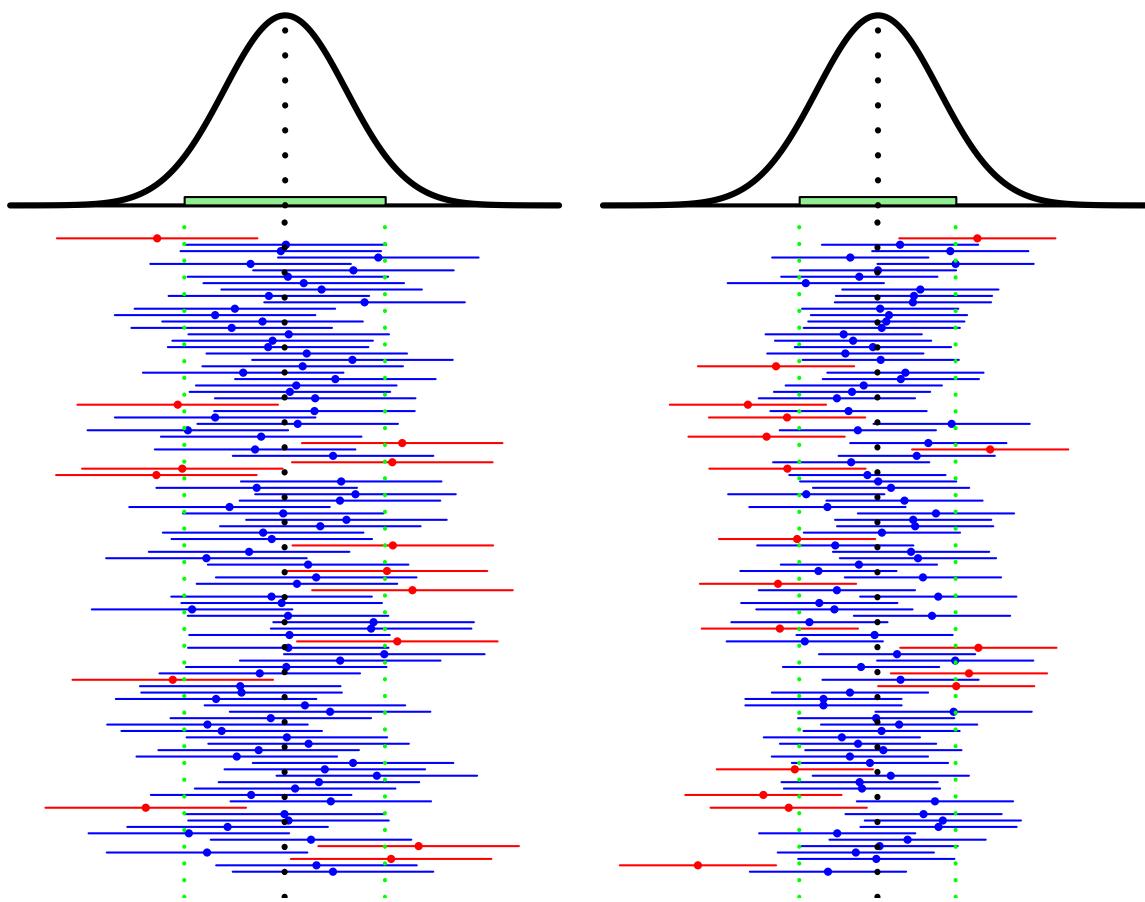


Figure 14.2. Sampling distribution of the sample mean (**tops**) and 100 random 90% (**Left**) and 80% (**Right**) confidence intervals (horizontal lines) from samples of $n=50$ from the Square Lake population. Confidence intervals that do NOT contain μ are shown in red.

One should consider the following subtleties when considering the concept of a confidence region,

- A CI is a random variable just like any other statistic. That is, each sample results in a different 95% CI (observe the CI lines on Figure 14.1) just like each sample results in a different \bar{x} (observe the dot on each CI line of Figure 14.1).
- Any one CI will either contain the parameter, μ in this case, or it will not. However, on average, 95% of 95% CIs will contain the parameter of interest and 5% will not. That is, if we could construct all possible 95% CIs, then 95% of all of those CIs would contain the parameter.
- The 95% CI is a technique that “works correctly” 95% of the time. In other words, 95% of all 95% CI “capture” the unknown value of the parameter.

Because of these subtleties confidence regions are often misinterpreted by novice (and even advanced) users of statistics. Some common misinterpretations are listed below with an explanation of the misinterpretation in parentheses. These misinterpretations should be studied, compared to the interpretations discussed above, and avoided.

1. “There is a 95% probability that the population mean is between the endpoints of the computed confidence interval.” [This is incorrect because the population mean is constant (not random) and it either is or is not in a particular computed interval, and it will never change whether it is or is not in

- that interval. The key point is that the confidence interval is random and the parameter is not.]*
2. “95% of all 95% confidence intervals will fall between the endpoints of the computed confidence interval.” [First, this is physically impossible at this point (i.e., using Z^*) because each confidence interval is the same width (if n and the level of confidence stay constant). Second, it is not important how many confidence intervals are contained in a confidence interval; interest is in whether the parameter is in the interval or not.]
 3. “There is a 95% probability that the sample mean is between the endpoints of the computed confidence interval.” [This is incorrect for the simple fact that confidence intervals are not used to estimate sample means (or, generally, statistics); they are used to estimate population means (or parameters). Furthermore, the sample mean has to be exactly in the middle of the confidence interval (see next section).]

◊ Care and specificity must be used when interpreting and describing confidence intervals.

◊ Confidence intervals are constructed for parameters, not statistics.

Review Exercises

14.1 True or False – A 95% confidence region can be constructed for \bar{x} ? [Answer](#)

14.2 True or False – A 95% confidence region can be constructed for the population median? [Answer](#)

14.3 True or False – A 95% confidence region can be constructed for σ ? [Answer](#)

14.4 Yes, No, Can't tell – I computed the following CI: (111.12, 123.32). Is the estimated parameter in this interval? [Answer](#)

14.5 Make this statement correct by replacing the “XXX” with a word – “I am 99% confident that the XXX of interest is within my confidence interval?” [Answer](#)

14.2 Constructing Confidence Regions

As alluded to previously, not all confidence regions are designed to contain the parameter 95% “of the time,” are intervals, or are computed to contain μ . Confidence regions can be constructed for any level of confidence, intervals or bounds, and for nearly all **parameters**.

The level of confidence (C) to use will be determined by the level of α chosen for the hypothesis test. Specifically, the level of confidence will be $100(1 - \alpha)\%$. For example, if one sets α at 0.05, then the level of

confidence should be 95% or if α is set at 0.01, then a 99% level of confidence should be used. From this, one can see that if α is decreased such that fewer Type I errors are made, then the confidence level will increase and more of the confidence regions will contain the parameter of interest (i.e., fewer errors). In this manner the proportion of Type I errors in the hypothesis testing framework is linked to the proportion of errors made from interpreting confidence regions.

- ◊ The level of confidence (C) is determined from α ; i.e., $C = 100(1 - \alpha)\%$.

The type of confidence region to be computed depends on the type of alternative hypothesis. If the alternative hypothesis is two-tailed (i.e., \neq), then the confidence region will be a bounded interval. In other words, two values will be computed such that the parameter of interest is expected, given a level of confidence, to be contained between those two values. These are the intervals discussed previously in Section 14.1. However, if the alternative hypothesis is one-tailed, then a so-called confidence bound is used. For example, if the alternative hypothesis is a “less than”, then interest lies in determining what is the “largest possible value” for the parameter rather than what is the range of possible values for the parameter (as would be obtained with a confidence interval). In other words, if the alternative hypothesis is a “less than”, then an upper confidence bound for the parameter is constructed. In contrast, if the alternative hypothesis is a “greater than”, then a lower confidence bound is constructed to estimate the “smallest possible value” for the parameter.

- ◊ A confidence interval should be constructed when a two-tailed H_A is used.

- ◊ A confidence bound should be constructed when a one-tailed H_A is used. If H_A is a “greater than”, then the smallest possible value of the parameter is sought and a lower bound is constructed. If H_A is a “less than”, then the largest possible value of the parameter is sought and an upper bound is constructed.

Fortunately, most² confidence regions follow the same basic form of,

$$\text{“Statistic”} (\pm \text{“margin of error”})$$

where “Statistic” represents whatever statistic is used to estimate the parameter and the \pm sign represents either $+$, $-$, or \pm (described below). For example, \bar{x} was used as the statistic in the previous example when confidence intervals were constructed to estimate μ . The margin of error generally has the form,

$$(\pm \text{“scaling factor”}) * SE_{statistic}$$

which makes the generic confidence interval formula,

$$\text{“Statistic”} (\pm \text{“scaling factor”}) * SE_{statistic}$$

The scaling factor serves a dual purpose – controls the width and type of the confidence region. The relative magnitude of the value controls the relative width of the region such that the parameter is contained in the

²

All that we will see in this class.

region at a rate according to the level of confidence. For example, in 99% confidence regions the scaling factor will be set such that 99% of the confidence regions will contain the parameter. The actual value of the scaling factor is computed from known sampling distributions. In the case, where σ is known (the situation considered here), the scaling factor is computed from a $N(0, 1)$ and is called a Z^* .

The sign of the scaling factor controls whether an interval, upper bound, or lower bound is computed. For example, if the alternative hypothesis is two-tailed, then two values of Z^* should be found such that an area equal to the level of confidence is contained between them (Figure 14.3-Left). The two values that delineate these boundaries will be the exact same value but with different signs because the $N(0, 1)$ distribution is symmetric about zero. Thus, a confidence interval is computed with a scaling factor of $\pm Z^*$.

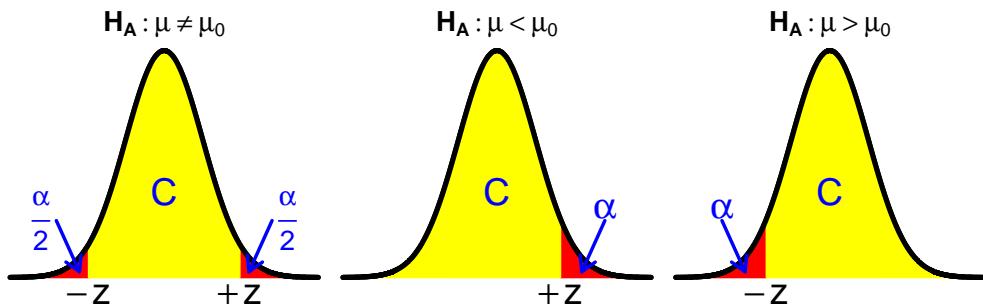


Figure 14.3. Depiction of the areas that define the z^* for creating confidence bounds of a parameter in a hypothesis test.

In contrast, if the alternative hypothesis is a “less than”, then an upper confidence bound is desired. In this case the Z^* is found such that it has an area equal to the level of confidence LESS THAN it (Figure 14.3-Middle). As the level of confidence will always be greater than 50%, this definition will produce a positive value of Z^* so that the scaling factor will be $+Z^*$. Similarly, if the alternative hypothesis is a “greater than”, then a lower confidence bound is desired and a value of Z^* with an area equal to the level of confidence GREATER THAN it should be found (Figure 14.3-Right). This will produce a negative value of Z^* so that the scaling factor will be $-Z^*$.

◊ Confidence intervals can be constructed for any level of confidence and for nearly every parameter.

◊ When finding Z^* for a confidence bound, the level of confidence always represents an area shaded in the same direction as the sign in H_A .

The following are three examples for calculating confidence regions.

1. For the Square Lake example, with $H_A : \mu < 105$ and $\alpha = 0.05$, a 95% upper confidence bound should be constructed. The corresponding $Z^* = 1.645$ is found with

```
> ( distrib(0.95,type="q") )
[1] 1.644854
```

Thus, with the summary information for a single sample of $n = 50$ shown in Table 2.2, the 95% upper confidence bound is $100.04 + 1.645 \frac{31.49}{\sqrt{50}}$, $100.04 + 7.33$, or 107.37 . Thus, one is 95% confident that the

true mean total length of all fish in Square Lake is less than 107.4 mm. By confident, it is meant that 95% of all 95% confidence regions will contain the true μ .

- Suppose that the mouse water consumption data from Table 5.1 was tested with $H_A : \mu \neq 10$ and $\alpha = 0.01$. In this case a 99% confidence interval should be constructed. The corresponding $Z^* = \pm 2.576$ is found with

```
> ( distrib(0.995,type="q") )
[1] 2.575829
```

Thus, assuming that $\sigma = 2$ ml and using the summary information computed in Section 5.5 the 99% confidence bound is $14.04 \pm 2.576 \frac{2}{\sqrt{30}}$, 14.04 ± 0.94 , or (13.10, 14.98). Thus, one is 99% confident that the true mean level of water consumption by all mice is between 13.1 and 15.0 ml. By confident, it is meant that 99% of all 99% confidence regions will contain the true μ .

- Suppose the third example hypothesis that started this chapter is being tested with an $\alpha = 0.10$. In this case a 90% upper confidence bound should be constructed. The corresponding $Z^* = +1.282$ is found with

```
> ( distrib(0.90,type="q") )
[1] 1.281552
```

Thus, assuming that $\sigma = 15$, $\bar{x} = 75$, and $n = 40$, the 90% confidence bound is $75 + 1.282 \frac{15}{\sqrt{40}}$, $75 + 3.04$, or 78.04. Thus, one is 90% confident that the true mean monthly heating bill for all houses is less than \$78.04. By confident, it is meant that 90% of all 90% confidence regions will contain the true μ .

Review Exercises

14.6 What is Z^* for a 99% confidence interval? [Answer](#)

14.7 What is Z^* for a 92% lower confidence bound? [Answer](#)

14.8 What is Z^* for a 90% upper confidence bound? [Answer](#)

14.9 What is Z^* for a 98% confidence interval? [Answer](#)

14.10 What is Z^* for a 95% lower confidence bound? [Answer](#)

14.11 What is Z^* for a 70% upper confidence bound? [Answer](#)

14.12 Construct and interpret (including describing what is meant by “confidence”) a proper confidence region for the mean BOD level presented in Review Exercise 13.8. [Answer](#)

14.13 Construct and interpret (including describing what is meant by “confidence”) a proper confidence region for the mean grade point average presented in Review Exercise 13.9. [Answer](#)

14.14  Construct and interpret (including describing what is meant by “confidence”) a proper confidence region if H_A is a “not equals” and $\alpha=0.05$ for the population mean gage height on the Bois Brule River presented in Review Exercise 5.13 assuming that the population standard deviation is 0.20 feet and the sampling distribution is approximately normal. [Answer](#)

14.15  Construct and interpret (including describing what is meant by “confidence”) a proper confidence region if H_A is a “less than” and $\alpha=0.10$ for the mean population density of all counties in Wisconsin using the data presented in Review Exercise 5.14 assuming that $\sigma = 125$ people/land acre and the sampling distribution is approximately normal. [Answer](#)

14.16  Construct and interpret (including describing what is meant by “confidence”) a proper confidence region if H_A is a “greater than” and $\alpha=0.05$ for the population mean creatine phosphokinase value using the data presented in Review Exercise 5.5 assuming that $\sigma = 40$. [Answer](#)

14.17  Hebblewhite (2000) reported the mean snow pack height (in cm) for Banff (data are below). These data were strongly right-skewed with a possible outlier at the maximum. Assume that it is known that $\sigma=15$ cm. (A) Compute a 99% confidence interval for the mean snow pack height. (B) In addition, comment on whether or not a confidence interval should be computed for these data (note: compute the CI in (A) regardless of your answer here). [Answer](#)

29.00, 45.51, 30.18, 45.83, 39.54, 80.39, 32.64, 32.89,
46.84, 45.79, 62.92, 67.24, 30.96, 46.08, 33.28

14.3 Hypothesis Tests and Confidence Region Relationship

The concept of confidence intervals can be visualized differently. This alternative view does not obfuscate the previous or subsequent discussions and, in fact, will strongly augment the hypothesis testing discussion of Section ???. This visualization, however, begins with a rather non-standard graphic where sample mean values that would be “reasonable to see” from a population with various possible values of μ are constructed. The construction and utility of this graphic will be illustrated below with the Square Lake fish example. With this example, consider that μ is unknown but that σ is known ($=31.49$), that samples of $n = 50$ are still used, and that 95% CIs will be computed.

As a first step, compute the most common 95% of sample means from a population assuming that $\mu = 70$. This is easily computed with $70 \pm 1.960 \frac{31.49}{\sqrt{50}}$, 70 ± 8.73 , or $(61.27, 78.73)$. This range is then plotted as a vertically oriented rectangle centered horizontally on $\mu = 70$ (left-most rectangle on Figure 14.4-Left). Then compute and plot the same range for a slightly larger assumed value of μ , say $\mu = 71$ (i.e., plot $(62.27, 78.73)$). Repeat these steps for sequentially larger values of μ until a plot similar to Figure 14.4 is constructed.

Before describing how this graphic is useful for understanding a confidence interval, consider very carefully what this graphic represents. The vertical rectangles represent the range of the most common 95% of sample means (values read from the y-axis) that will be produced for a particular population mean (value read from the x-axis). In a nutshell, each vertical line represents the sample means that are likely to be observed (y-axis values) from a population with a given population mean (x-axis).

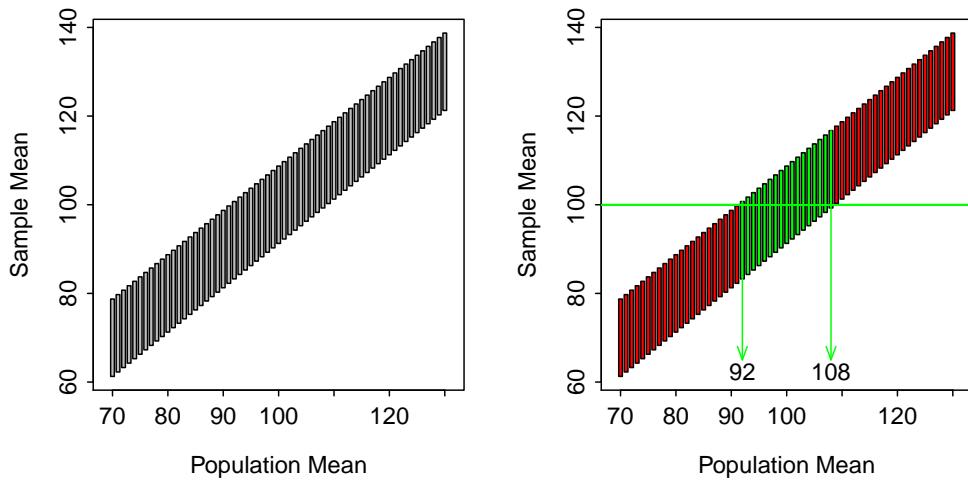


Figure 14.4. Range (95%) of sample means that would be produced by particular population means in the Square Lake fish length example (**Left**) and an illustration of the ranges intercepted by $\bar{x} = 100.0$ mm (**Right**).

Now suppose that the sample mean of 100.04 is observed (as in Table 2.2). Locate this value on the y-axis of Figure 14.4, draw a horizontal line across the graph at this value, and draw vertical lines down from where the horizontal line first enters and then leaves the band of possible sample means (Figure 14.4-Right). The values along the x-axis that the vertical lines intercept are approximations to the 95% confidence interval. The approximations are only as close as the intervals used to construct the rectangles (i.e., in this example intervals of 1.0 mm were used). However, the results from this graphical approach (i.e., (92, 108)) compare favorably to the results from using the CI formula (i.e., (91.27, 108.73)).

Surely, the CI formula discussed in Section 14.2 is a much quicker way to construct a 95% confidence interval. However, this graph illustrates a critical interpretation of confidence intervals. The confidence interval (or region, more generally) consists of the population means which would likely produce the observed sample mean. Thus, the values in a confidence interval represent population means that would be likely to produce the sample mean that was actually observed. Thus, the confidence region represents possible hypothesized population means that WOULD NOT BE rejected during hypothesis testing.

◊ The values in a confidence interval represent population means that were likely to have produced the sample mean that we actually observed.

14.4 Precision and Sample Size

The width of confidence intervals explains how precisely the parameter is estimated. For example, relatively narrow intervals represent relatively precise estimates of the parameter. From the general construction of confidence intervals it is seen that the width of a confidence interval is twice the margin of error. Thus, the width of a confidence interval depends on the margin of error which, in turn, depends on (1) the standard error and (2) the scaling factor. As either of these two items gets smaller (while holding the other constant), the width of the CI will get smaller.

- ◊ The width of a confidence interval is a measure of the precision of our estimate of the parameter.
- ◊ The width of a confidence interval depends on the standard error of the statistic and the scaling factor used.

A smaller standard error means that the estimate is more precise. More precise estimates are obtained only by increasing the sample size. A smaller standard deviation would also result in a smaller SE, but for most purposes the standard deviation is constant (i.e., a population has a standard deviation, we cannot make it smaller).

- ◊ Confidence intervals can be made narrower by increasing the sample size.

A smaller scaling factor is obtained by reducing the level of confidence. For example, a 90% confidence interval uses a $Z^* = 1.645$ whereas a 95% confidence interval uses a $Z^* = 1.960$ (as shown previously). Thus, narrower CIs can be constructed by decreasing the confidence level. However, there is a trade-off in reducing the level of confidence to make a narrower confidence interval because the number of confidence intervals not containing the parameter of interest will increase.

- ◊ Confidence intervals can be made (but should not be made) narrower by decreasing the confidence level of the interval.

The relationship between the precision of an estimate as reflected in the width of the confidence interval and the sample size provides a means for computing the same size required to estimate μ within $\pm m.e.$ units (i.e., margin-of-error) with C% confidence assuming that σ is known. A formula for determining the sample size given these constraints is derived by algebraically solving for n in the margin-of-error formula for the construction of a confidence interval for μ , i.e.,

$$\begin{aligned} m.e. &= z * \frac{\sigma}{\sqrt{n}} \\ \sqrt{n} &= \frac{z * \sigma}{m.e.} \\ n &= \left(\frac{z * \sigma}{m.e.} \right)^2 \end{aligned}$$

For example, suppose one wants to compute the sample size required to estimate the mean length of fish in Square Lake to within 5 mm with 90% confidence knowing that the population standard deviation is 34.91. First, define the symbols as $m.e.=5$, $\sigma=34.91$, and $Z^*=1.645$ (found previously for 90% confidence). Thus, $n = \left(\frac{1.645*34.91}{5} \right)^2 = 131.91$. Therefore, a sample of at least 132 fish from Square Lake should be taken. Note that sample size calculations are always rounded up to the next integer because rounding down will produce a sample size that does not meet the desired criteria (i.e., you need at least some fraction more to meet the desired criteria).

- ◊ Always round up to the next integer in sample size calculations.

The margin-of-error and confidence level in these calculations need to come from the researcher's beliefs in

how much error they can live with (i.e., chance that a confidence interval does not contain the parameter) and how precise their estimate of the mean needs to be. Values for σ are rarely known in practice (because it is a parameter) and estimates from preliminary studies, previous similar studies, similar populations, or wild guesses are often used instead. In practice, a researcher will often prepare a graph with varying values of σ (Figure 14.5) to make an informed decision of what sample size to choose.

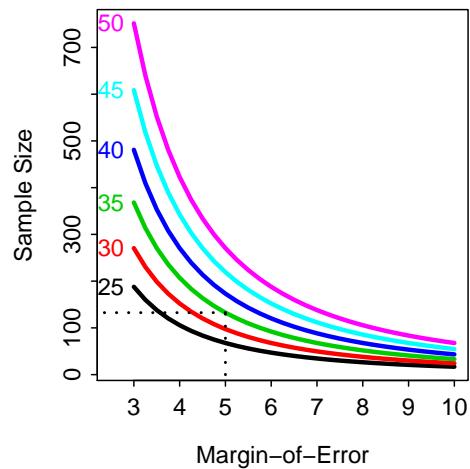


Figure 14.5. Desired sample size versus margin-of-error for constant values of σ (shown to the left of each line) and $C = 90$. The desired sample size for m.e.=5, $\sigma = 35$, and $C = 90$ is illustrated with the black dotted lines.

Review Exercises

- 14.18** If two populations have the same standard deviation and a sample of size 30 is taken from population A and a sample of size 50 from population B, which will have a narrower CI? [Answer](#)
- 14.19** If the same size of sample is taken from two populations, but Population C has a smaller standard deviation than Population D, which will have a narrower CI? [Answer](#)
- 14.20** From the same data, is a 95% or a 99% CI narrower? [Answer](#)
- 14.21** Describe how the margin of error will change as each of the following change (all others held constant): confidence level (C), z^* , n , σ , μ , and \bar{x} (in the case of CIs for μ). Make sure to explain your reasoning for each. [Answer](#)
- 14.22** Geographers measure the longest axis of pebbles to determine “grain” sizes. If the standard deviation of pebble long-axis length for a particular site is known to be 4 mm, how many pebbles must be measured in order to determine the average pebble length within 0.1 mm with 99% confidence? [Answer](#)
- 14.23** An investment group wants to start an Internet Service Provider (ISP) and, for their business plan and model, needs to estimate the average Internet usage of households. How many households must be randomly selected to be 95% sure that the sample mean is within 1 minute of the population mean? Assume that a previous survey of household usage had a standard deviation of 6.95 minutes. [Answer](#)
-

Part IV

Specific Hypothesis Tests

CHAPTER 15

1-SAMPLE Z-TEST

Chapter Objectives:

1. Properly construct statistical hypotheses.
2. Understand the specifics of a 1-Sample Z-Test.
3. Perform the 11-steps of a significance test in a 1-Sample Z-Test situation.

Contents

| | |
|---|-----|
| 15.1 11-Steps of Hypothesis Testing | 207 |
| 15.2 1-Sample Z-Test Specifics | 207 |
| 15.3 1-Sample Z-Test in R | 209 |

15.1 11-Steps of Hypothesis Testing

I have created an 11-step process to make sure that you complete all aspects important to statistical hypothesis testing. These steps are listed below and should be used for all hypothesis tests in ensuing chapters.

1. State the rejection criterion (α),
2. State the null and alternative hypotheses to be tested and define the parameter(s),
3. Identify the hypothesis test to use (e.g., 1-Sample t, 2-sample t, etc.) and explain why it is the test of choice,
4. Collect the data (describe how the data were collected and if randomization occurred),
5. Check all necessary assumptions (describe how you tested the validity),
6. Calculate the appropriate statistic(s),
7. Calculate the appropriate test statistic,
8. Calculate the p-value,
9. Reject/DNR H_0 ,
10. Summarize your findings in terms of the problem (do not use the word “reject”),
11. If H_0 was rejected, compute and interpret an appropriate confidence region for the parameter.

Two of these steps require amplifying discussion. First, the “collect the data” step (step 4) should be highlighted because the most important order in these 11 steps is that steps 1-3 **MUST** be completed before collecting the data and the remaining steps are performed after collecting the data. Second, when a null hypothesis is rejected it is implied that some difference exists between what was observed and what was expected. Following the detection of a difference it is important to clearly articulate the direction and magnitude of that difference. This is accomplished by computing an appropriate confidence region for the parameter (Step 11).

◊ The α , hypotheses, and test to use must be declared before data are collected.

◊ Confidence regions for a parameter are an appropriate component of an hypothesis testing procedure when the null hypothesis is rejected, because the confidence region clearly articulates the direction and magnitude of the difference.

15.2 1-Sample Z-Test Specifics

A 1-Sample Z-Test is the name for the procedure developed previously in this chapter. A 1-Sample Z-Test tests the null hypothesis that the population mean is equal to a specific value or, symbolically, $H_0 : \mu = \mu_0$ where μ_0 represents any specific value of the population mean. In this section, the specifics of a 1-Sample Z-Test are summarized and, in doing so, a framework to be used for subsequent hypothesis tests is developed. In addition, two full examples, using the 11-steps of any hypothesis test, will be completed.

A 1-Sample Z-Test is characterized by testing $H_0 : \mu = \mu_0$ in the situation when σ is known. The only test that can possibly be confused with a 1-Sample Z-Test is a 1-Sample t-Test (Chapter ??), which tests the same null hypothesis but in the situation where σ is unknown. The specifics of a 1-Sample Z-Test are identified in Table 15.1. The conceptual underpinnings of the 1-Sample Z-Test were discussed in great detail in previous sections of this chapter and in Chapter 12.

Table 15.1. Characteristics of a 1-Sample Z-Test.

- **Hypothesis:** $H_0 : \mu = \mu_0$
- **Statistic:** \bar{x}
- **Test Statistic:** $z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
- **Confidence Region:** $\bar{x}(\pm Z^*)\frac{\sigma}{\sqrt{n}}$
- **Assumptions:**
 1. σ is known
 2. $n > 30$, $n > 15$ and the **population** is not strongly skewed, OR the **population** is normally distributed.

15.2.1 Example - Intra-class Travel

Consider the following situation,

A dean is interested in the average amount of time it takes to get from one class to another. In particular, she wants to determine if it takes more than 10 minutes, on average, to go between classes. In an effort to test this hypothesis, she collects a random sample of 100 intra-class travel times and finds the mean to be 10.12 mins. Assume that it is known from previous studies that the distribution of intra-class times is symmetric with a standard deviation of 1.60 minutes. Use appropriate methods to test the dean's hypothesis with an $\alpha = 0.10$.

The 11-steps (Section 15.1) for completing a full hypothesis test for this example are as follows:

1. As stated, α should be set at 0.10.
 2. The null hypothesis will be about μ (mean time for all intra-class travel events) and it will be tested against a specific value, namely $\mu_0 = 10$. Thus, $H_0 : \mu = 10$ mins. The $H_A : \mu > 10$ mins (the dean is interested in seeing if the mean intra-class time is **more than** 10 mins).
 3. A 1-Sample Z-Test is required because a quantitative variable (intra-class travel time) was measured on individuals from one population, the population mean is compared to a specific value in the null hypothesis, and σ is known (given in the background).
 4. The data appear to be part of an observational study (the dean did not impart any conditions on the students) with a random selection of individuals.
 5. The sample size ($=100$) is much greater than 30, thus the test statistic computed below should reasonably follow a standard normal distribution. In addition, σ is known ($=1.60$ mins).
 6. The \bar{x} is the statistic of choice because the hypothesis is about μ . From the background information, the $\bar{x}=10.12$.
 7. The z test statistic is $\frac{10.12-10}{\frac{1.60}{\sqrt{100}}} = \frac{0.12}{0.16} = 0.75$.
 8. The p-value for this statistic is $p = 0.2266$ (Figure 15.1) as computed with
- ```
> (distrib(10.12,mean=10,sd=1.60/sqrt(100),lower.tail=FALSE))
[1] 0.2266274
```
9.  $H_0$  is not rejected because the  $p - value > \alpha = 0.10$ .
  10. It appears that the mean for **all** intra-class travel events is not more than 10 minutes.

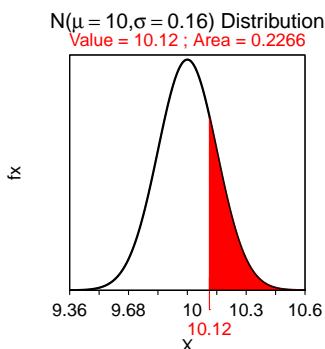


Figure 15.1. Depiction of the p-value for the intra-class travel example.

### Review Exercises

**15.1** A researcher is investigating the growth of a certain cactus under a variety of environmental conditions. He knows from previous research that the growth of this particular type of cactus is approximately normally distributed with a standard deviation of 1.40 cm. Under the current environmental conditions that he is investigating, however, he does not know the mean. He does hypothesize that it is no more than 4 cm. To test this hypothesis he used a preliminary sample of 10 randomly-selected cacti. He found the sample mean for these cacti to be 3.26 cm. Use this information to test his hypothesis with  $\alpha = 0.05$ . [Answer](#)

**15.2** Owens and Pronin (2000) studied the age and growth of pike in Chivyrkui Bay on Lake Baikal. They found that the length of the sample of 30 pike in Lake Baikal was slightly right-skewed with a mean of 656.1 mm. Suppose that a recent article in an outdoor magazine reported the average length of all pike in this lake to be 600 mm long. It is known from previous studies that the standard deviation of pike length is about 130 mm. Perform a test, using a 95% confidence level, to determine if the mean length of pike reported by the researchers significantly differs from that reported in the outdoor magazine. [Answer](#)

### 15.3 1-Sample Z-Test in R

The p-value in a 1-Sample Z-Test is computed from summary information using `distrib()` as shown in the previous discussion and example. However, if raw data exists it is more efficient to use `z.test()`<sup>1</sup>. This function requires a vector of the quantitative data as the first argument, the hypothesized value for  $\mu$  in the `mu=` argument, and a value of the known  $\sigma$  in the `sd=` argument. In addition, the type of alternative hypothesis is declared in the `alt=` argument. This argument requires a string (i.e., contained in quotes) of either "two.sided" (the default), "less", or "greater" corresponding to the "not equals", "less than", and "greater than" alternatives, respectively. Finally, a level of confidence is declared in the `conf.level=`

<sup>1</sup>

From the `TeachingDemos` package which is loaded with `NCStats`.

argument. This value must be a proportion (between 0 and 1) and defaults to 0.95. You should take note of the default values for the `alt=` and `conf.level=` arguments as these are what `z.test()` will use if these arguments are not specifically declared by you.

The results of `z.test()` should be assigned to an object. Typing the name of this object will produce output that shows, among other things, the calculated statistic ( $\bar{x}$ ), test statistic ( $Z$ ), p-value, and confidence region. In addition, the saved object is submitted to `plot()` to produce a visual representation of the test statistic and p-value. While the graphic from `plot()` does not provide any new information for the hypothesis test, it is highly recommended that you make the plot as a check of the p-value and your choice for the alternative hypothesis.

Use of `z.test()` and `plot()` are illustrated in the following example.

### 15.3.1 Body Temperature

Consider the following situation<sup>2</sup>,

*Machowiak et al. (1992) critically examined the belief that the mean body temperature is  $98.6^{\circ}\text{F}$  by measuring the body temperatures in a sample of healthy humans. Their data are found in `BodyTemp.csv`. Use these data, with a supposedly known  $\sigma = 0.63^{\circ}\text{F}$ , and an  $\alpha = 0.01$  to determine if the mean body temperature differs from  $98.6^{\circ}\text{F}$ .*

The 11-steps (Section 15.1) for completing a full hypothesis test for this example are as follows:

1. As stated,  $\alpha$  should be set at 0.01.
2. The null hypothesis will be about  $\mu$  and it will be tested against a specific value, namely  $\mu_0 = 98.6^{\circ}\text{F}$ . Thus,  $H_0 : \mu = 98.6^{\circ}\text{F}$ . The  $H_A : \mu \neq 98.6^{\circ}\text{F}$  (the researchers want to determine if the temperature is **different from**  $98.6^{\circ}\text{F}$ ).
3. A 1-Sample Z-Test is required because a quantitative variable (i.e., body temperature) was measured on individuals from one population, the population mean is compared to a specific value in the null hypothesis, and  $\sigma$  is known (given in the background).
4. The data appear to be part of an observational study although this is not made clear in the background information. There is also no evidence that randomization was used. The data were loaded into R from `BodyTemp.csv` and the results of the 1-Sample Z-Test were computed as follows,

```
> bt <- read.csv("data/BodyTemp.csv")
> headtail(bt)
 temp sex heart.rate
1 96.3 M 70
2 96.7 M 71
3 96.9 M 74
128 99.9 F 79
129 100.0 F 78
130 100.8 F 77
> (bt.z <- z.test(bt$temp,mu=98.6,sd=0.63,conf.level=0.99))
One Sample z-test with bt$temp
z = -6.3482, n = 130.000, Std. Dev. = 0.630, Std. Dev. of the sample mean =
```

<sup>2</sup>

There is an interesting discussion of studies of body temperature at [The Physics Factbook](#).

```

0.055, p-value = 2.178e-10
alternative hypothesis: true mean is not equal to 98.6
99 percent confidence interval:
98.10690 98.39156
sample estimates:
mean of bt$temp
98.24923
> plot(bt.z)

```

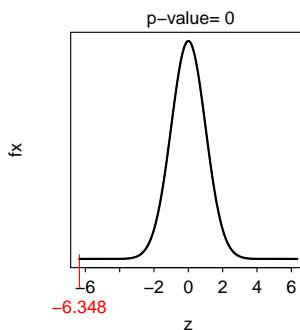


Figure 15.2. Depiction of the p-value for the body temperature example.

5. The sample size (130) is much greater than 30, thus the test statistic computed below should reasonably follow a standard normal distribution. In addition,  $\sigma$  is known ( $= 0.63^{\circ}\text{F}$ ).
6. The hypothesis is about  $\mu$ . Therefore, we want to calculate  $\bar{x}$ , which from the output above, is  $98.25^{\circ}\text{F}$ .
7. The z test statistic is -6.35.
8. The p-value for this value of the test statistic is  $p < 0.00005$  (Figure 15.2).
9. Reject  $H_0$  because  $p - value < \alpha = 0.01$ .
10. It appears that the mean body temperature of all humans is different from  $98.6^{\circ}\text{F}$ .
11. A **99% confidence interval** is warranted for this situation and is  $(98.11, 98.39)$ . Thus, one is 99% confident that the mean body temperature ( $\mu$ ) is actually between  $98.11$  and  $98.39^{\circ}\text{F}$ .

## Review Exercises

- 15.3** A study by Cheshire et al. (1994) reported on six patients with chronic myofascial pain syndrome (introduced in Review Exercise 13.3). The researchers determined the duration of pain for the six patients were 2.5, 2.7, 2.8, 2.8, 2.8, and 3.0. Test the hypothesis that the mean pain length was greater than 2.5 years at the 10% significance level. Assume that it is known that the distribution of duration of pain is normal with a standard deviation of 0.5 years. [Answer](#)

- 15.4** Suppose that it is known that cholesterol levels in women aged 21-40 in the U.S. has a mean of 190 mg/dl and standard deviation of 40 mg/dl. Suppose that we want to determine, at the 10% significance level, if the cholesterol level of Asian women is different from U.S. women as determined from 40 randomly selected Asian women aged 21-40 who had recently immigrated to the U.S. Assume that the Asian women

have the same standard deviation as the U.S. women population. The data from this sample are found in [Cholesterol.csv](#). [Answer](#)

---

---

---

# CHAPTER 16

---

## 1-SAMPLE T-TEST

**Chapter Objectives:**

1. Identify when a 1-sample t-Test is appropriate.
2. Perform the 11 steps of a significance test in a 1-sample t-Test situation.

---

### Contents

---

|                                          |     |
|------------------------------------------|-----|
| 16.1 t-distribution . . . . .            | 214 |
| 16.2 1-Sample t-Test Specifics . . . . . | 216 |
| 16.3 1-Sample t-Test in R . . . . .      | 218 |

---

**P**RIOR TO THIS CHAPTER, the hypothesis testing methods have required knowing  $\sigma$ . Of course,  $\sigma$  is a parameter that is seldom actually known, but is estimated by  $s$ , the sample standard deviation. When  $\sigma$  is replaced by  $s$ , the test statistic follows a Student's t rather than a standard normal (Z) distribution. In this chapter, the specifics of the t-distribution are described and a 1-sample t-Test for testing that the mean from one population equals a specific value is introduced. A 2-sample t-Test for comparing the means of two populations is introduced in Chapter 17.

## 16.1 t-distribution

A t-distribution is similar to a standard normal distribution (i.e.,  $N(0,1)$ ) in that it is centered on 0 and is bell shaped (Figure 16.1). The t-distribution differs from the standard normal distribution in that it is heavier in the tails, flatter near the center, and its exact dispersion is dictated by a quantity called the degrees-of-freedom (df). The t-distribution is “flatter and fatter” because of the uncertainty surrounding the use of the sample standard deviation in the standard error calculation<sup>1</sup>. The degrees of freedom are a function of the sample size and generally come from the denominator of the sample standard deviation calculation. As the degrees-of-freedom increase, the t-distribution becomes narrower and taller and approaches the shape and dispersion of the standard normal distribution (Figure 16.1).

Figure 16.1. Standard normal (black) and t-distributions (red) with varying degrees-of-freedom.

- ◊ A t-distribution is “wider” than a z-distribution because of the extra uncertainty from using  $s$  rather than  $\sigma$  in the test statistic calculation.

Proportional areas on a t-distribution are computed using `distrib()` in a manner similar to that described for a normal distribution in Chapter 7. To compute the area on a t-distribution the first argument to `distrib()` must be the value of  $t$ , the `distrib=` argument is required and is set equal to "t", and the `type=` argument is "p"<sup>2</sup>. In addition, the `df` (how to find the df will be discussed in subsequent sections) must also

<sup>1</sup>

Recall that the sample standard deviation is a statistic and is thus subject to sampling variability.

be provided in the `df=` argument. As before, `lower.tail=FALSE` is used to compute the upper tail area. For example, the area to the right of  $t = -1.456$  on a t-distribution with 9 df is 0.9103 (Figure 16.2) and is found with

```
> (distrib(-1.456,distrib="t",df=9,lower.tail=FALSE))
[1] 0.9103137
```

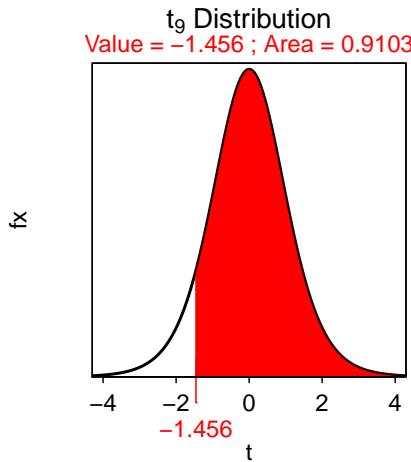


Figure 16.2. Depiction of the area to the right of  $t = -1.456$  on a t-distribution with 9 df.

Values of  $t$  with a certain area to the right or left can also be found with `distrib()`. In these cases, the first argument should be changed to the desired area and the `type=` argument must be set equal to "q". For example, the value of  $t$  with an area of 0.95 to the right on a t-distribution with 19 df is -1.729 (Figure 16.3) and is found with

```
> (distrib(0.95,distrib="t",type="q",df=19,lower.tail=FALSE))
[1] -1.729133
```

Of course, this last “reverse” calculation would be the  $t^*$  for a 95% lower confidence bound. This use will be illustrated in subsequent sections.

---

The `type=` argument defaults to "p" so it may be omitted.

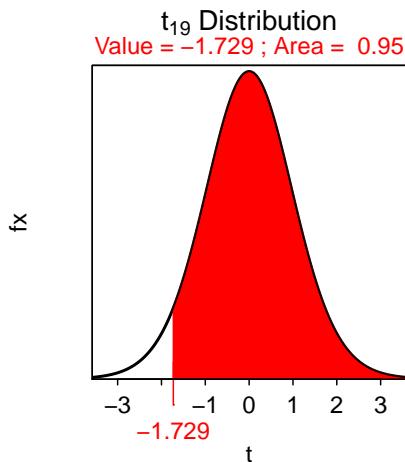


Figure 16.3. Depiction of the value of  $t$  with an area to the right of 0.95 on a  $t$ -distribution with 19 df.

### Review Exercises

**16.1** What is the p-value if  $H_A : \mu < 125$ ,  $t = -2.178$ , and  $df = 35$ ? [Answer](#)

**16.2** What is  $t^*$  for the previous question if  $\alpha = 0.05$ ? [Answer](#)

**16.3** What is the p-value if  $H_A : \mu > 125$ ,  $t = 1.856$ , and  $df = 81$ ? [Answer](#)

**16.4** What is  $t^*$  for the previous question if  $\alpha = 0.01$ ? [Answer](#)

**16.5** What is the p-value if  $H_A : \mu \neq 125$ ,  $t = -2.178$ , and  $df = 99$ ? [Answer](#)

**16.6** What is  $t^*$  for the previous question if  $\alpha = 0.10$ ? [Answer](#)

## 16.2 1-Sample t-Test Specifics

A 1-sample t-Test is very similar to a 1-sample z-test in that both tests test the same null hypothesis. The big difference, as discussed previously, is that when  $\sigma$  is unknown it is replaced by an estimate of  $\sigma$  (i.e.,  $s$ ), which causes the test statistic to become a  $t$ . Other aspects are similar between the two tests as shown in Table 16.1<sup>3</sup>.

<sup>3</sup>

Compare this table to Table 15.1.

Table 16.1. Characteristics of a 1-Sample t-Test.

- **Hypothesis:**  $H_0 : \mu = \mu_0$
- **Statistic:**  $\bar{x}$
- **Test Statistic:**  $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$
- **Confidence Region:**  $\bar{x}(\pm t^*)\frac{s}{\sqrt{n}}$
- **df:**  $n - 1$
- **Assumptions:**  $n > 40$ ,  $n > 15$  and **sample** is not strongly skewed, OR **sample** is normally distributed.

### 16.2.1 Example - Purchase Lot of Salmon?

Consider the following situation,

A fish wholesaler has a catch of several thousand salmon. A prospective buyer will buy the lot if it can be shown that the mean weight of all salmon is at least 19.9 lbs. A random selection of 50 salmon had a mean of 20.1 and a standard deviation of 0.76 lbs. Should the buyer accept the catch at the 5% level?

The 11-steps (Section 15.1) for completing a full hypothesis test for this example are as follows:

1. As stated,  $\alpha$  should be set at 0.05.
2. The null hypothesis will be about  $\mu$  and it will be tested against a specific value, namely  $\mu_0 = 19.9$  lbs. Thus,  $H_0 : \mu = 19.9$  lbs. The  $H_A : \mu > 19.9$  lbs (the buyer will buy the lot if the average weight is “at least” or, alternatively, “more than” 19.9 lbs).
3. A 1-sample t-Test is required because a quantitative variable (weight) was measured on individuals from one population (this lot of salmon), the population mean is compared to a specific value in the null hypothesis, and  $\sigma$  is UNKNOWN<sup>4</sup>.
4. The data appear to be part of an observational study with random selection.
5. The  $\sigma$  is unknown. The sample size is greater than 40; thus, the test statistic computed below should reasonably follow a t-distribution.
6. The statistic is  $\bar{x}$  (=20.1; from the background). In addition, the sample standard deviation is given as 0.76 lbs.
7. The test statistic is  $t = \frac{20.1 - 19.9}{\frac{0.76}{\sqrt{50}}} = \frac{0.2}{0.107} = 1.87$ . This test statistic has  $50 - 1 = 49$  df.
8. The p-value is  $p = 0.0337$  as calculated with

```
> (distrib(1.87,distrib="t",df=49,lower.tail=FALSE))
[1] 0.03373207
```

9. The  $H_0$  is rejected because the  $p - value < \alpha = 0.05$ .

<sup>4</sup>

If  $\sigma$  is given, then it will appear in the background information to the question and will be in a sentence that uses the words “population”, “assume that”, or “suppose that.”

10. The average weight of all salmon in this lot appears to be greater than 19.9 lbs; thus, the buyer should accept this lot of salmon.
11. A 95% lower confidence bound is warranted for this situation. The  $t^* = -1.677$  as computed with

```
> (distrib(0.95,distrib="t",type="q",df=49,lower.tail=FALSE))
[1] -1.676551
```

Thus,  $20.1 - 1.677 \frac{0.76}{\sqrt{50}}$  or  $20.1 - 0.18 = 19.92$ . Thus, one is 95% confident that the mean weight of all salmon in the lot is greater than 19.92 lbs.

## Review Exercises

**16.7** A general achievement test is standardized so that students should average 80 with a standard deviation of 5 (this is for the entire population not the population of students at the school described below). The superintendent at a school in a large district would like to show that her students averaged better than the 80 points. To test this, she had the test given to 32 randomly selected students from her school. The summary statistics for those 32 students are: mean=83.2, median=82.5, standard deviation=5.5, and IQR=7. Perform the appropriate hypothesis test for this superintendent at the 0.05 level. [Answer](#)

**16.8** The Northwestern University Placement center conducts random surveys on starting salaries of college graduates and publishes the results every year. The Dean of the College of Liberal Arts suggested to prospective students that graduates from the College would earn more than \$32000 as a starting salary on average. The results in the table below are from a part of the Placement Center's results for graduates of the College of Liberal Arts for the year just prior to the Dean's statements [Note that the measurements are in 1000s of dollars]. Use these results at the 10% level to determine the correctness of the Dean's statement. [Answer](#)

| n  | Min.  | 1st Qu. | Median | 3rd Qu. | Max.  | Mean   | StDev |
|----|-------|---------|--------|---------|-------|--------|-------|
| 42 | 29.30 | 31.30   | 32.50  | 33.80   | 36.80 | 32.511 | 1.713 |

## 16.3 1-Sample t-Test in R

The p-value in a 1-sample t-Test is computed from summary information using `distrib()`. However, if raw data exists it is more efficient to use `t.test()`. The arguments to this function are very similar to the arguments to `z.test()`. The `t.test()` function requires a vector of the quantitative data as the first argument and the hypothesized value for  $\mu$  in the `mu=` argument. In addition, the type of alternative hypothesis ("two.sided", "less", or "greater") is set in the `alt=` argument and a level of confidence is declared (as a proportion) in the `conf.level=` argument. As with `z.test()`, `t.test()` will default to a "not equals" alternative and a 95% confidence level. The results of `t.test()` should be assigned to an object with the results then seen by typing the name of that object and an illustrative plot of the p-value created by submitting that object to `plot()`. The use of `t.test()` is illustrated in the following example.

### 16.3.1 Example - Crab Body Temperature

Consider the following situation,

A marine biologist wanted to determine if the body temperature of crabs exposed to ambient air temperature would be different than the ambient air temperature. The biologist exposed a sample of 25 crabs to an air temperature of  $24.3^{\circ}\text{C}$  for several minutes and then measured the body temperature of each crab. The body temperatures for individual crabs is shown below. Perform a hypothesis test (at the  $\alpha = 0.01$ ) level to answer the biologist's question.

$22.9, 22.9, 23.3, 23.5, 23.9, 23.9, 24.0, 24.3, 24.5, 24.6, 24.6, 24.6, 24.8, 24.8,$   
 $25.1, 25.4, 25.4, 25.5, 25.5, 25.8, 26.1, 26.2, 26.3, 27.0, 27.3, 28.1$

The 11-steps (Section 15.1) for completing a full hypothesis test for this example are as follows:

1. As stated,  $\alpha$  should be set at 0.01.
2. The null hypothesis will be about  $\mu$  and it will be tested against a specific value, namely  $\mu_0 = 24.3^{\circ}\text{C}$ . Thus,  $H_0 : \mu = 24.3^{\circ}\text{C}$ . The  $H_A : \mu \neq 24.3^{\circ}\text{C}$  (the researcher is interested in identifying a difference).
3. A 1-sample t-Test is required because a quantitative variable (temperature) was measured on individuals from one population, the population mean is compared to a specific value in the null hypothesis, and  $\sigma$  is UNknown.
4. The data appear to be part of an experimental study (the temperature was controlled) with no suggestion of random selection of individuals. The data were entered into the `ct` vector in R with<sup>5</sup>

```
> ct <- c(22.9, 22.9, 23.3, 23.5, 23.9, 23.9, 24.0, 24.3, 24.5, 24.6, 24.6, 24.6, 24.8, 24.8,
 25.1, 25.4, 25.4, 25.5, 25.5, 25.8, 26.1, 26.2, 26.3, 27.0, 27.3, 28.1)
```

5. The  $\sigma$  is unknown. The sample size is not greater than 40 but it is greater than 15 and the distribution of values in the sample appears to be only slightly right-skewed (Figure 16.4). The histogram was constructed with

```
> hist(~ct, xlab="Crab Body Temp (C)")
```

Because both assumptions are adequately met, one can continue with the computation of the statistic, test statistic, p-value, and confidence region with

```
> (ct.t <- t.test(ct, mu=24.3, conf.level=0.99))
One Sample t-test with ct
t = 2.7128, df = 24, p-value = 0.01215
alternative hypothesis: true mean is not equal to 24.3
99 percent confidence interval:
 24.27741 25.77859
sample estimates:
mean of x
 25.028
> plot(ct.t)
```

<sup>5</sup>

These data may be more easily entered into a CSV file as described in Section 4.3.2 and then read into R with `read.csv()`.

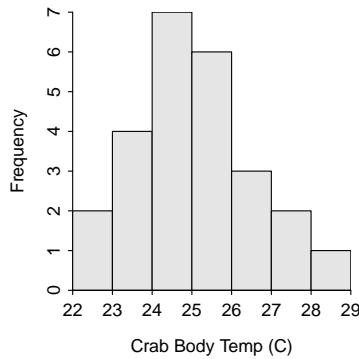


Figure 16.4. Histogram of the body temperatures of  $n=25$  crabs exposed to an ambient temperature of  $24.3^{\circ}\text{C}$ .

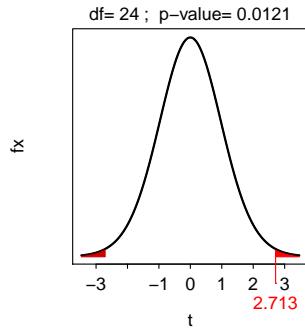


Figure 16.5. Depiction of the p-value for the crab body temperature example.

6. The statistic is  $\bar{x}=25.03^{\circ}\text{C}$ .
7. The test statistic is  $t=2.713$ . This test statistic has 24 df.
8. The p-value is  $p = 0.0121$  (Figure 16.5).
9. The  $H_0$  is not rejected because the  $p - value > \alpha = 0.01$ .
10. It appears that the average body temperature of the crabs is not different than the ambient temperature of  $24.3^{\circ}\text{C}$ .
11. *A confidence interval is not required as the  $H_0$  was not rejected.* However, this confidence interval shows that the true mean body temperature of the crabs is likely between  $24.28^{\circ}\text{C}$  and  $25.78^{\circ}\text{C}$ . Note that this interval contains  $\mu_0$  which is why  $H_0$  was not rejected.

## Review Exercises

- 16.9**  Fishing line is graded by the pounds (lbs) of pressure that it can withstand before breaking. For example, line that is rated as 6-lbs should not break for pressures under 6 lbs. Two physics students developed an apparatus for testing the breaking point of 2-foot sections of line to test the manufacturer's claim (i.e., determine if line rated at 6-lbs broke, on average, at pressures below 6 lbs). To test this, they measured the pounds of pressure it took 20 randomly selected 2-foot sections of line to break. Use their results shown below to test their hypothesis at the 10% level. [Answer](#)

6.1 5.3 5.5 4.9 6.2 6.5 5.7 5.5 4.7 6.2  
6.8 5.9 5.8 6.7 6.3 6.2 5.4 5.5 6.7 5.9

- 16.10**  Last year I planted 400 everbearing strawberry plants in my garden. The company I bought the plants from claimed that in the year following planting, each plant would produce an average of 12 berries. I was surprised by this claim and hypothesized that the plants would actually produce less than what the company said, on average. To test this claim, I counted the number of ripe berries produced for the entire season on 50 randomly selected plants. Use the data in [Strawberries.csv](#) to test the company's claim at the 10% level. [Answer](#)

- 16.11**  The toy industry rates toys regarding their ease for being put together in three categories: easy, moderate, and difficult. A toy is placed into the easy category if it takes 10 minutes or less to put the toy together, in the moderate category if it takes 20 minutes or less (and more than 10 minutes), and in the difficult category if it takes more than 20 minutes. A randomly selected group of 34 adults were asked to put together a new toy to determine which rating the toy should receive. The results from these 34 individuals are in [ToyTime.csv](#). Conduct a hypothesis test, at the 10% level, to determine whether the toy should receive the difficult rating. [Answer](#)

- 16.12**  One of the dominant uses of Madison area lakes is for boating. To develop a long-term data set on the temporal fluctuations and trends in such activity, the Long Term Ecological Research (LTER) project obtained records of boat traffic that passes through the locks at the head of the Yahara River on its stretch between Lake Mendota and Lake Monona. These data in [Yahara.csv](#) have been collected nearly daily from April through October since 1976. Use these data to determine, at the 5% level, if the mean total number of boats passing through the locks during the months of June, July, and August of 2005 is greater than 75. HINT: create a new data frame that contains just the data for this period (i.e., the data file contains more data than is needed for this question). I suggest that you do this in three separate steps – isolate 2005 data, isolate data for months after May (5), and then isolate data for months before September (9). [Answer](#)

- 16.13**  The golden rectangle is a rectangle with a length-to-width ratio of 1:1.618, or equivalently, a width-to-length ratio of 0.618:1 (See a description of the golden rectangle [here](#)). The golden rectangle is evident in several works by ancient Greeks and Egyptians. Anthropologists measured the width-to-length ratios of beaded rectangles used by the Shoshoni Indians of America to decorate their leather goods. Use their data<sup>6</sup> in [Shoshoni.csv](#) to determine, at the 5% level, if the golden rectangle is evident in the beadwork of the Shoshonis. [Answer](#)

---

<sup>6</sup>

This question and these data originated at [OzDASL](#).

---

---

# CHAPTER 17

---

## 2-SAMPLE T-TEST

**Chapter Objectives:**

1. Identify when a 2-Sample t-Test is appropriate.
2. Describe what a homogeneity of variance test is and why it is required within a 2-Sample t-Test.
3. Perform the 11 steps of a significance test in a 2-Sample t-Test situation.

---

### Contents

---

|                                            |     |
|--------------------------------------------|-----|
| 17.1 2-Sample t-Test Specifics . . . . .   | 223 |
| 17.2 Testing for Equal Variances . . . . . | 224 |
| 17.3 2-Sample t-Tests in R . . . . .       | 229 |

---

**T**HE T DISTRIBUTION and 1-Sample t-Test for comparing the mean of one population to a specific value were introduced in Chapter 16. In this chapter, a 2-sample t-Test for comparing the means of two populations is introduced.

While it is often useful to test whether a population mean is equal to a specific value, as was done with the 1-Sample Z-Test and 1-Sample t-Tests, there are many instances where interest is determining whether the means from two populations are equal. In these situations, one is usually trying to determine if a difference exists between the two population means. For example, is there a difference in income between males and females, in test scores between students from high- or low-income families, in percent body fat between raccoons from southern and northern Wisconsin, or in amount of milk produced between cows provided with a hormone and cows provided with a placebo. In all of these situations, two populations are being examined (males and females, students from high- and low-income families, raccoons from southern and northern Wisconsin, cows given a hormone and cows given a placebo) and interests is in determining if a difference in population means exists. The **2-sample t-Test** is used to make these determinations and is the subject of this section.

## 17.1 2-Sample t-Test Specifics

In a 2-sample t-Test, the null hypothesis is that the two population means are equal, i.e.,  $H_0 : \mu_1 = \mu_2$ . The null hypothesis can be rewritten as  $H_0 : \mu_1 - \mu_2 = 0$ , because the difference between two population means should be zero if the two population means are equal. With this new organization of the null hypothesis, one must think of finding a statistic that will be an estimate of  $\mu_1 - \mu_2$ , the hypothesized “parameter.” Analogous to using  $\bar{x}$  as an estimate of  $\mu$  in the 1-Sample t-Test,  $\bar{x}_1 - \bar{x}_2$  is an estimate of  $\mu_1 - \mu_2$ .

- ◊ The parameter in a 2-sample t-Test is the difference in population means ( $\mu_1 - \mu_2$ ). The corresponding statistic is the difference in sample means ( $\bar{x}_1 - \bar{x}_2$ ).

Now, when looking at the same “general” test statistic as used in the 1-Sample inferences – i.e., (13.3.1) as

$$\text{Test Statistic} = \frac{\text{Observed Statistic} - \text{Hypothesized Parameter}}{SE_{\text{Statistic}}}$$

it becomes apparent that an estimate of the standard error of  $\bar{x}_1 - \bar{x}_2$  (i.e., our statistic) is needed. Unfortunately, the calculation of this standard error depends on whether the two population variances are equal or not. When the variances are approximately equal (discussed in the next section), we first calculate a pooled estimate of the variance ( $s_p^2$ ) as a weighted average of the sample variances from the two samples, or

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- ◊ The  $s_p^2$  calculation can be “checked” by determining if the value of  $s_p^2$  is between  $s_1^2$  and  $s_2^2$  or if the value of  $\sqrt{s_p^2}$  is between  $s_1$  and  $s_2$ .

The standard error of  $\bar{x}_1 - \bar{x}_2$  is the square root of the product of the pooled variance and the sum of the

inverses of the sample sizes, or

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

The degrees-of-freedom for the 2-sample t-Test with equal variances still comes from the denominator of the variance (pooled in this case) calculation. Thus, the  $df = n_1 + n_2 - 2$ . These specifics are summarized in Table 17.1. The specifics for the 2-sample t-Test when the variances are unequal are not discussed in this book.

Table 17.1. Characteristics of a 2-sample t-Test with equal variances.

- **Hypothesis:**  $H_0 : \mu_1 - \mu_2 = 0$
- **Statistic:**  $\bar{x}_1 - \bar{x}_2$
- **Test Statistic:**  $t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$  where  $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ .
- **Confidence Region:**  $\bar{x}_1 - \bar{x}_2 (\pm t^*) \sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$
- **df:**  $n_1 + n_2 - 2$
- **Assumptions:**  $n_1 + n_2 > 40$ ,  $n_1 + n_2 - 2 > 15$  and **each sample** is not strongly skewed, OR **each sample** is normally distributed.

Many times the 2-sample t-Test will be used to test an alternative hypothesis of simply finding a difference between the two populations. However, if the null hypothesis is rejected in these instances (thus, identifying a significant difference between the two populations), then special care should be taken to specifically describe how the two populations differ. If the statistic is negative, then the mean of the first population is lower than the mean of the second population and, if the statistic is positive, then the mean of the first population is larger than the mean of the second population. The values of the confidence region should be used to identify how much larger or smaller the mean from one population is compared to the mean of the other population.

◊ Use the statistic and confidence region results to specifically determine which population has a larger or smaller mean when the null hypothesis of the 2-sample t-Test has been rejected in favor of the “not equals” alternative hypothesis.

## 17.2 Testing for Equal Variances

As noted above, the methods of a 2-sample t-Test differ depending on whether the population variances from the two populations are equal or not. This should present a problem to you because the population variances are parameters and are typically not known<sup>1</sup>. The question of whether these parameters are equal

<sup>1</sup>

Actually, the population variances don't have to be known exactly, it just needs to be known whether they are equal or not.

or not will be handled in the same manner as all other questions about a parameter or parameters have been handled – i.e., with a hypothesis test.

- ◊ A hypothesis test must be used to determine if two population variances are equal.

There are two hypothesis tests that are commonly used to test whether two (actually “two or more”) population variances are equal or not. The first is called Bartlett’s test and is used when it is known that two population distributions are normally distributed. The second test is called Levene’s test and is used for all continuous distributions, whether normal or not. Levene’s test will be used throughout this book as it is more general and a bit more conservative<sup>2</sup>.

- ◊ Use Levene’s test to test the hypothesis that two population variances are equal, because it does not require populations that are normally distributed.

The specifics of the Levene’s test will not be examined in detail in this book. Rather you will only need to know that the  $H_0 : \sigma_1^2 = \sigma_2^2$  is tested against the  $H_A : \sigma_1^2 \neq \sigma_2^2$ . The one-tailed alternatives are not considered with this test, nor are they of interest in this situation; i.e., one only needs to know if there is a difference in the population variances. Without knowing the full details of the Levene’s test, we will rely on computer software to compute the p-value. The p-value is interpreted as always – if the  $p\text{-value} < \alpha$ , then reject the  $H_0$  and conclude that the variances are unequal, if the  $p\text{-value} > \alpha$ , then do not reject the  $H_0$  and conclude that the variances are at least approximately equal.

- ◊ If the p-value from Levene’s test is less than  $\alpha$ , then reject the  $H_0$  and conclude that the variances are unequal.
- ◊ If the p-value from Levene’s test is greater than  $\alpha$ , then do not reject  $H_0$  and conclude that the variances are at least approximately equal.

### 17.2.1 Example - Corn and Fertilizers

Consider the following situation,

An agricultural researcher thought that corn plants grown in pots exposed to a certain type of synthetic fertilizer would grow taller than plants exposed to an organic fertilizer. To collect data to test this idea, he grew 50 corn plants in individual pots – 25 were treated with organic fertilizer and 25 were treated with synthetic fertilizer. Each pot contained soil from a well-mixed common source and was planted in the same greenhouse. Each plant was similar in all regards (similar genetics, age, etc.). Use the results (heights of individual plants) in Table 17.2 to test the researcher’s hypothesis (at the  $\alpha = 0.05$  level).

The 11-steps (Section 15.1) for completing a full hypothesis test for this example are as follows:

1. As stated,  $\alpha$  should be set at 0.05.

<sup>2</sup>

Levene’s test is more conservative because it does not require a normal distribution.

Table 17.2. Summary statistics and histogram of the corn plant height in two treatments.

|                | Synthetic | Organic |
|----------------|-----------|---------|
| means:         | 51.46     | 47.49   |
| SD:            | 5.975     | 6.721   |
| Levene's Test: | p=0.1341  |         |

2. Thus,  $H_0 : \mu_s - \mu_o = 0$  where  $s$  represents the synthetic and  $o$  represents the organic fertilizer (thus, positive numbers represent larger values for the synthetic fertilizer). The  $H_A : \mu_s - \mu_o > 0$  (representing the idea that the synthetic fertilizer will produce taller plants).
3. A 2-sample t-Test is required because a quantitative variable (height) was measured on two populations (synthetic and organic fertilizers) that were INdependent and two population means are being compared in the null hypothesis.
4. The data appear to be part of an experimental study (the researcher imposed the treatments on the plants) with no clear indication of random selection of plants or random allocation of plants to the two treatments.
5. The sample size ( $n_s + n_o = 50$ ) is  $> 40$ . Therefore, the test statistic computed below should reasonably follow a t-distribution with  $n_s + n_o - 2 = 48$  df. In addition, the two samples are independent as there does not appear to be any connection between pots. The two population variances appear to be equal because the p-value for Levene's test of the homogeneity of variance test (given as 0.1341) is "large" (i.e.,  $> 0.05$ ).
6. The statistic is  $\bar{x}_s - \bar{x}_o = 51.46 - 47.49 = 3.97$  (values from Table 17.2). The pooled sample variance is,

$$s_p^2 = \frac{(25-1)5.975^2 + (25-1)6.721^2}{25+25-2} = 40.44$$

The standard error of the statistic is,

$$SE_{\bar{x}_s - \bar{x}_o} = \sqrt{40.44 \left( \frac{1}{25} + \frac{1}{25} \right)} = 1.799$$

7. The test statistic is  $t = \frac{3.97-0}{1.799} = \frac{3.97}{1.799} = 2.207$ . This test statistic has  $25 + 25 - 2 = 48$  df.
  8. The p-value is  $p = 0.0161$ , as computed with
- ```
> ( distrib(2.207,distrib="t",df=48,lower.tail=FALSE) )
[1] 0.01606477
```
9. The H_0 is rejected because the $p-value < \alpha = 0.05$.
 10. The average height of the corn plants appears to be greater for the plants grown with the synthetic fertilizer than those plants grown with the organic fertilizer.
 11. A 95% confidence lower bound is warranted in this situation. The $t^* = -1.677$ as computed with

```
> ( distrib(0.95,distrib="t",type="q",df=48,lower.tail=FALSE) )
[1] -1.677224
```

Thus, $3.97 - 1.677 * 1.799$ or $3.97 - 3.02 = 0.95$. Thus, one is 95% confident that plants grown with synthetic fertilizer are more than 0.95 cm taller, on average, than plants grown with the organic fertilizer.

17.2.2 Example - Music and Anxiety

Consider the following situation,

An oral surgeon conducted an experiment to determine if background music decreased the anxiety level of patients during a tooth extraction. Over a one-month period, 32 patients had a tooth removed while listening to music and 36 had a tooth removed with no music to listen to. Each patient was given a questionnaire following the extraction. Answers to the questionnaire were converted to a numeric scale to measure the patient's level of anxiety (larger numbers mean greater anxiety). For those given background music, the mean anxiety level was 4.2 (with a standard deviation of 1.2), while the group without music had a mean of 5.9 (with a standard deviation of 1.9). The surgeon also reported a Levene's test p-value of 0.089. Test the surgeon's hypothesis using $\alpha = 0.05$.

The 11-steps (Section 15.1) for completing a full hypothesis test for this example are as follows:

1. As stated, α should be set at 0.05.
2. The $H_0 : \mu_w - \mu_o = 0$ where w represents “with” and o represents “without” the music (thus, negative numbers represent lower anxiety values in patients in the “with music” treatment). The $H_A : \mu_w - \mu_o < 0$ (representing lower levels of anxiety in patients in the “with music” treatment).
3. A 2-sample t-Test is required because a quantitative variable (anxiety level) was measured on two populations (music or no music) that were INdependent and two population means are being compared in the null hypothesis.
4. The data appear to be part of an observational study with $n_w = 32$ and $n_o = 36$. There is no obvious random selection or allocation in this study.
5. The sample size ($n_w + n_o = 68$) is > 40 . Therefore, the test statistic computed below should reasonably follow a t-distribution with $n_w + n_o - 2 = 66$ df. In addition, the two samples are independent because no one patient had any effect or impact on any other patient. The population variances appear to be equal between the two treatment groups because the p-value for Levene's test of the homogeneity of variance test (given as 0.089) is “large” (i.e., > 0.05).
6. The statistic is $\bar{x}_w - \bar{x}_o = 4.2 - 5.9 = -1.7$. The pooled sample variance is,

$$s_p^2 = \frac{(32-1)1.2^2 + (36-1)1.9^2}{32+36-2} = 2.59$$

The standard error of the statistic is,

$$SE_{\bar{x}_w - \bar{x}_o} = \sqrt{2.59 \left(\frac{1}{32} + \frac{1}{36} \right)} = 0.391$$

7. The test statistic is $t = \frac{-1.7-0}{0.391} = -4.348$. This test statistic has $32 + 36 - 2 = 66$ df.
8. The p-value is $p < 0.00005$, as computed with

```
> ( distrib(-4.348,distrib="t",df=66) )
[1] 2.43093e-05
```

9. The H_0 is rejected because the $p-value < \alpha = 0.05$.
10. The average anxiety level of the patients differed between when music was played and when it was not. In fact, it appears that the anxiety level was lower when the music was played.
11. A 95% upper confidence bound is warranted in this situation. The $t^* = 1.668$ as computed with

```
> ( distrib(0.95,distrib="t",type="q",df=66) )
[1] 1.668271
```

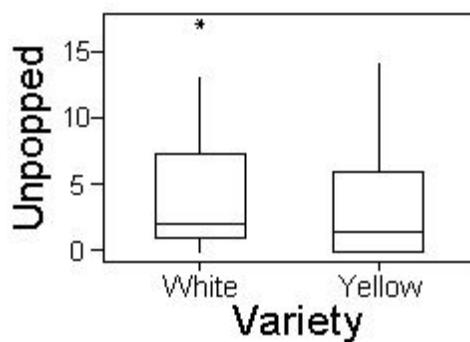
Thus, $-1.7 + 1.668 * 0.391$ or $-1.7 + 0.65 = -1.05$. Thus, one is 95% confident that the mean anxiety level is more than -1.05 points lower, on average, when music is played than when it is not.

Review Exercises

- 17.1** Erville Redenbacher wanted to see if the number of unpopped kernels differed between yellow and white varieties of his grandpa's famous popcorn. To test this, he would put 100 kernels of either white or yellow popcorn into a standard air popper, pop the corn until no "pops" were heard, and then count the number of unpopped kernels. He tested 30 randomly selected groups of 100 kernels for both white and yellow varieties (Erville is very thorough). Use the results below to test, at the 10% level, Erville's hypothesis. [Answer](#)

| Variable | N | Mean | Median | StDev | SE Mean |
|----------|----|-------|--------|-------|---------|
| White | 30 | 4.267 | 2.000 | 4.456 | 0.814 |
| Yellow | 30 | 3.567 | 1.500 | 4.485 | 0.819 |

Levene's Test -- P-Value = 0.972

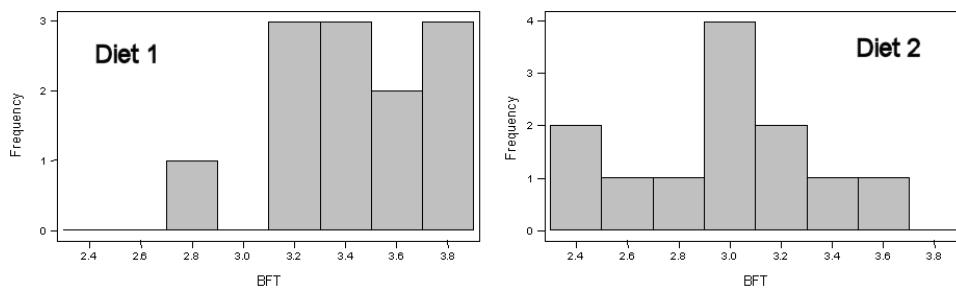


- 17.2** A study was performed in order to evaluate the effectiveness of two devices for improving the efficiency of gas home-heating systems. Energy consumption in houses was measured after one of the two devices was installed. The two devices were an electric vent damper (DampVent=Electric) and a thermally activated vent damper (DampVent=ThermAct). Energy consumption (in BTUs) was measured for a variety of houses fitted with the two devices. Compare, at the 10% level, the effectiveness of these two devices by determining if a difference exists in energy consumption between houses fitted with the devices. Note that Levene's test p-value is 0.996. [Answer](#)

| Variable | DampVent | N | Mean | Median | StDev | SE Mean | Minimum | Maximum | Q1 | Q3 |
|----------|----------|----|--------|--------|-------|---------|---------|---------|-------|--------|
| BTU.In | Electric | 40 | 9.908 | 9.590 | 3.020 | 0.477 | 4.000 | 18.260 | 7.885 | 11.555 |
| | ThermAct | 50 | 10.143 | 10.290 | 2.767 | 0.391 | 2.970 | 16.060 | 8.127 | 12.212 |

- 17.3** A pig diet manufacturer wants to determine if the backfat thickness differs between pigs raised on two different diets. Backfat thickness is an indicator of pork quality; smaller thicknesses mean better quality. A group of 24 pigs was randomly allocated to two groups which differed only in the diet received. Test the results from this experiment to see if a difference in backfat thickness is evident at the $\alpha = 0.05$ level. Note that Levene's test p-value is 0.532. [Answer](#)

| Var | Diet | N | Mean | Median | StDev | SE Mean | Min | Max |
|-----|------|----|-------|--------|-------|---------|------|------|
| BFT | 1 | 12 | 3.420 | 3.390 | 0.295 | 0.0850 | 2.87 | 3.87 |
| | 2 | 12 | 2.989 | 3.035 | 0.375 | 0.108 | 2.40 | 3.62 |



17.3 2-Sample t-Tests in R

17.3.1 Data Format

The data for a 2-sample t-Test must be entered in stacked format. In stacked format the measurements are in one vector and a label for which group the measurement was recorded from is in another vector. If both vectors are in a data frame³, then each row corresponds to the measurement and the group of a single individual. This is the general format in which most data is entered and in which most databases and R functions require the data.

The data for BOD measurements in either the inlet or outlet to an aquaculture facility are shown below. These data illustrate stacked data because each row corresponds to one individual and the columns are two variables defined on that individual with one variable being the measurement and the other variable being the group to which the individual belongs.

```
BOD      src
1 6.782  inlet
2 5.809  inlet
```

³

This will most likely be the case as this data will most likely be read from an external data file.

```
3 6.849 inlet
18 8.545 outlet
19 8.063 outlet
20 8.001 outlet
```

△ **Stacked Data:** Data where the quantitative measurements of two groups are “stacked” on top of each other and a second variable is used to record to which group the measurement belongs.

- ◊ Stacked data is the preferred format for 2-sample data, because each vector corresponds to a variable and each row corresponds to only one individual.

17.3.2 Levene’s Test

Before conducting a 2-sample t-Test, the assumption of equal variances must be tested with a Levene’s test. The Levene’s test is computed in R with stacked data using `levenesTest()`. The first argument to this function is a model formula of the type `response~factor` where `response` represents the vector containing the quantitative measurements and `factor` represents the vector containing the categorical grouping variable⁴. The data frame containing the variables in the formula must also be supplied in the `data=` argument of `levenesTest()`.

17.3.3 2-Sample t-Test

A 2-sample t-Test is computed in R with the same `t.test()` function used for the 1-Sample t-Test. The first argument to the `t.test()` function is a model formula of the exact same type sent to `levenesTest()`. In addition to this argument, the following arguments may be specified when conducting a 2-sample t-Test with `t.test()`

- `mu`: The specific value of the null hypothesis. In the 2-sample case this is the hypothesized difference among the population means. The default value is 0 and, thus, this argument does not usually have to be specified.
- `alt=`: A character string indicating the type of alternative hypothesis ("two.sided", "greater", or "less"). As previously, "two.sided" is the default.
- `conf.level=`: The level of confidence to be used when constructing the confidence interval for $\mu_1 - \mu_2$. As previously, 0.95 is the default.
- `var.equal=`: A logical value indicating whether the two population variances should be considered to be equal or not. If `var.equal=TRUE`, then the pooled sample variance is calculated and used in the standard error. The default value is to assume unequal variances; thus, this argument must be set to `TRUE` if the result from `levenesTest()` suggests that the variances are equal.

⁴

This is the same model formula introduced in Section 5.8 for summarizing multiple groups of data.

- ◊ The `var.equal=TRUE` argument must be used in the `t.test()` function if one is to assume equal variances. This is NOT the default setting in R.

It must be noted that R computes the difference in `t.test()` as the mean of the “first” level minus the mean of the “second” level where the default behavior orders the levels alphabetically. For example, if the two levels are `inlet` and `outlet`, then R will compute $\mu_{inlet} - \mu_{outlet}$. This may or may not be the order that you want to use. Thus, for example, if you wanted $\mu_{outlet} - \mu_{inlet}$, then you need to “manually” change the order of the levels with `factor()`. The `factor()` function requires the name of the categorical variable as its first argument. The order of the levels of this variable is explicitly set by setting the `levels=` argument of `factor()` equal to a vector of the level names in the desired order. For example, the order of the levels of the `src` variable in the `aqua` data frame is changed and stored in a new variable name in the data frame with

```
> aqua$src1 <- factor(aqua$src, levels=c("outlet", "inlet"))
> levels(aqua$src1)
[1] "outlet" "inlet"
```

Two things should be noted about the commands above. First, I “saved” the re-ordered factor variable in a new name (i.e., `src1`) in the data frame so as not to over-write the original data. This is prudent so that, in case you made a mistake, you can always retrieve your original data. Second, `levels()` is used to show the ordering of the levels of a factor variable.

17.3.4 Example - BOD in Aquaculture Water

Consider the following situation (which was examined in parts above),

An aquaculture farm takes water from a stream and returns it to the stream after it has circulated through the fish tanks. The owner is concerned that the water may contain heightened levels of organic matter when it is released into the stream after it has circulated in the tanks. He has taken steps to reduce this possibility, i.e., circulated the water rather quickly through the tanks, but is still concerned about the increase in organic material in the effluent. To determine if this is true, he takes samples of the water at the intake and, at other times, downstream from the outlet and measures the biological oxygen demand (BOD) as a measure of the organics in the effluent (a higher BOD at the outlet would imply that organics are taken up from the tanks). The farmers data are recorded in [BOD.csv](#). Test for any evidence (i.e., at the 10% level) of support for the farmers concern.

The 11-steps (Section 15.1) for completing a full hypothesis test for this example are as follows:

1. As stated, α should be set at 0.10.
2. The $H_0 : \mu_{outlet} - \mu_{inlet} = 0$ where `outlet` represents the outlet source and `inlet` represents the inlet source (thus, positive numbers represent larger values at the outlet implying that BOD is increasing in the water released from the facility). Thus, the $H_A : \mu_{outlet} - \mu_{inlet} > 0$ (which represents an increase in BOD in water released from the facility).
3. A 2-sample t-Test is required because a quantitative variable (BOD level) was measured on two populations (outlet or inlet) that were **IN**dependent and two population means are being compared in the null hypothesis.

4. The data appear to be part of an observational study with no obvious randomization. The data were loaded with

```
> aqua <- read.csv("data/BOD.csv")
> headtail(aqua)
      BOD      src
1 6.782  inlet
2 5.809  inlet
3 6.849  inlet
18 8.545 outlet
19 8.063 outlet
20 8.001 outlet
```

The order of the levels of the *src* variable were then changed to match the order of subtraction in the hypotheses with

```
> aqua$src1 <- factor(aqua$src,levels=c("outlet","inlet"))
> levels(aqua$src1)
[1] "outlet" "inlet"
```

5. The two samples are independent because there is no connection between specific measurements at the inlet and outlet (e.g., they were not taken at the same time). The combined sample size (20) is < 40 but > 15 . The histograms (Figure 17.1) are inconclusive about the shape because of the small sample sizes in each group. However, it appears that the *inlet* data is not strongly skewed whereas there is evidence that the *outlet* data is skewed. This result may invalidate the results of this hypothesis test but I will continue anyway. The histograms were constructed with

```
> hist(BOD~src1,data=aqua,main="",xlab="BOD Measurement")
```

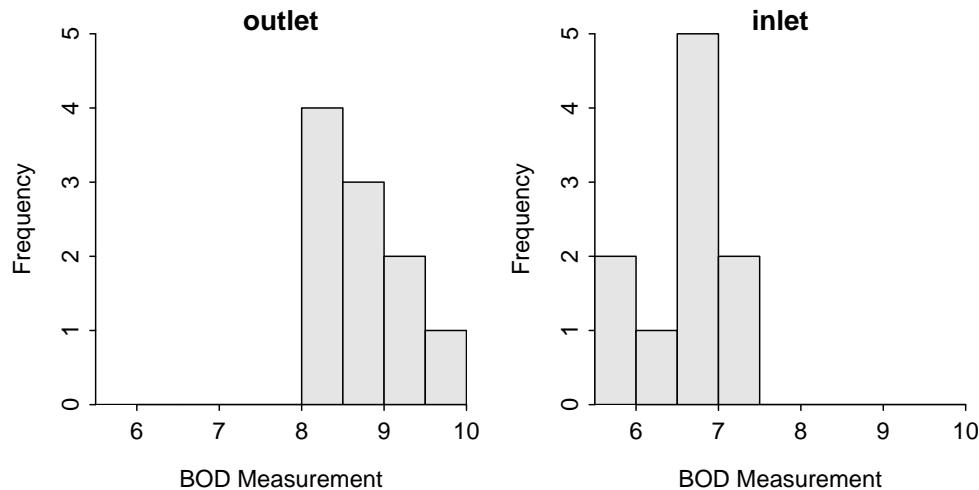


Figure 17.1. Histogram of the BOD measurements at the outlet and inlet of the aquaculture facility.

The variances appear to be equal because the Levene's test p-value ($p = 0.5913$) is larger than α . The Levene's test was computed with

```
> leveneTest(BOD~src1,data=aqua)
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group  1  0.2989 0.5913
      18
```

With the assumptions met (independence and equal variances) or nearly met (sample size), the 2-sample t-Test was conducted with

```
> ( aqua.t <- t.test(BOD~src1,data=aqua,var.equal=TRUE,alt="greater",conf.level=0.90) )
Two Sample t-test with BOD by src1
t = 8.994, df = 18, p-value = 2.224e-08
alternative hypothesis: true difference in means is greater than 0
90 percent confidence interval:
 1.732704      Inf
sample estimates:
mean in group outlet  mean in group inlet
     8.6873                  6.6538
> plot(aqua.t)
```

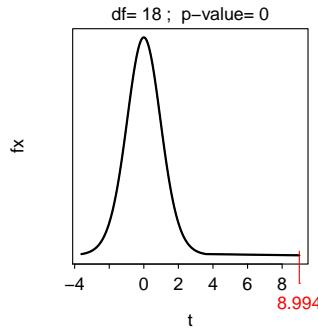


Figure 17.2. Depiction of the p-value in the 2-sample t-Test of BOD measurements in aquaculture example.

6. The group statistics are $\bar{x}_{outlet}=8.69$ and $\bar{x}_{inlet}=6.65$. Thus, the statistic is $8.69-6.65=2.03$.
7. The test statistic is $t=8.994$ with 18 df.
8. The p-value is $p < 0.00005$ (Figure 17.2).
9. The H_0 is rejected because the $p - value < \alpha$.
10. The average BOD is greater at the outlet than at the inlet to the aquaculture facility. Thus, it appears that the aquaculture facility adds to the oxygen demand of the water.
11. A 90% lower confidence bound is warranted in this situation and is 1.73. Thus, one is 90% confident that the BOD measurement at the outlet is AT LEAST 1.73 GREATER than the BOD measurement at the inlet.

Review Exercises

- 17.4**  A study⁵ examined the effectiveness of foil-lined milk cartons in reducing the “leakage” of dioxins from the carton to the milk (dioxins were found in milk cartons due to the bleaching process). The dioxin content (parts per thousand, ppt) in milk from 50 unlined and 50 lined cartons of milk were measured and recorded in [MilkCartons.csv](#). Use these data to determine, at the 1% level, if lining the cartons with foil significantly reduced the amount of dioxin in the milk. [Answer](#)
- 17.5**  The math department at the University of North Carolina is apparently noted for “giving out” low grades, relative to the rest of the school. To examine this, a random sample of the gpa for 22 math classes and 29 “other” university classes (from the last year) were examined. Use the data stored in [UNCGrades.csv](#) to determine if grades in math classes are significantly (at the 10% level) lower than grades in other classes. [Answer](#)
- 17.6**  A health commissioner needs to determine if the number of hours worked per week by medical interns differs between two cities. To examine this, the commissioner finds the mean number of hours worked by interns in the first city for a random sample of 13 weeks and the same for a random sample of 16 weeks from the second city. These results are found in [MedInternHrs](#). Use those results to determine if the hours worked by the interns differs, at the 10% level, between the two cities. [Answer](#)
- 17.7**  Agronomists are interested in determining conditions that increase the yield of crops. In one experiment 80 one-acre plots of corn were randomly divided into two groups of 40 plots each. An insecticide was used on each plot in one group and sterilized male individuals of an insect pest were released on each plot of the other group. The resulting yields were recorded in [CropYield.csv](#). Is there a difference, at the 10% level, in yield between the two treatments. [Answer](#)
- 17.8**  Templer’s Death Anxiety Scale (DAS) is a measure of an individual’s anxiety concerning death. Robbins (1990) examined 25 organ donors and 69 non-organ donors to determine if there was a difference in anxiety levels concerning death between these groups of people. The results are recorded in [DeathAnxiety.csv](#). Test Robbins’ researcher’s hypothesis at the 1% significance level. [Answer](#)

5

Data was recreated from Blaisdell 1998.

CHAPTER 18

CHI-SQUARE TESTS

Chapter Objectives:

1. Identify when a chi-square test is appropriate.
2. Perform the 11 steps of a significance test in a chi-square test situation.

Contents

| | |
|--|-----|
| 18.1 Chi-Square Distribution | 236 |
| 18.2 Chi-Square Test Specifics | 238 |
| 18.3 Chi-Square Test in R | 243 |

SITUATIONS WHERE A CATEGORICAL response variable is recorded would be summarized with a frequency or percentage table (see Section ?? and Section ??). The appropriate test statistic in these situations is a chi-square rather than a t. The chi-square test statistic follows a chi-square distribution which is introduced below. The rest of this chapter is dedicated to the general chi-square test where a categorical response is compared between two or more populations. The related goodness-of-fit test for a categorical response recorded for only one population is introduced in Chapter 19.

18.1 Chi-Square Distribution

A chi-square (χ^2) distribution is generally right-skewed (Figure 18.1). The exact shape of the χ^2 distribution depends on the degrees-of-freedom (df), where, as the df increase, the sharpness of the skew decreases (Figure 18.1).

Figure 18.1. χ^2 distributions with varying degrees-of-freedom.

In its simplest form the χ^2 distribution arises as a sampling distribution for the test statistic,

$$\chi^2 = \sum_{cells} \frac{(Observed - Expected)^2}{Expected}$$

where “Observed” and “Expected” represent the observed and expected frequencies of individuals in the cells of summary tables for categorical variables (see Section ?? and Section ??) and “cells” generically represents the number of cells in one of these tables. Thus, the χ^2 distribution arises naturally when the frequencies in two tables are being compared. Subsequent sections will demonstrate how this test statistic is used to compare a table of observed frequencies (i.e., from a sample) to a table of expected frequencies (i.e., from a null hypothesis).

Unlike with the other two distributions that we have seen (normal and t), the χ^2 distribution will always represent the two-tailed situation although the “two tails” will appear as one tail on the right side of the distribution. The simplest explanation for this characteristic is that the “squaring” in the calculation of the χ^2 test statistic results in what would be the “negative tail” being “folded over” onto what is the “positive tail” providing the appearance of only one tail. The result of this characteristic is that all area (i.e., probability) calculations on a χ^2 will pertain to two-tailed alternative hypotheses.

- ◊ Probability calculations on a χ^2 distribution always pertain to a two-tailed alternative hypothesis.

P-values are computed on a χ^2 distribution with `distrib()` very similarly to what was described for the normal (Section 15.2) and t distributions (Section 16.1). The first argument to this function is the value of the χ^2 test statistic, the `distrib=` argument must be set to "chisq", and the `type=` argument is "p"¹. In addition, the `df` (how to find the `df` will be discussed in subsequent sections) must also be provided in the `df=` argument. "More extreme" on a χ^2 distribution when computing p-values is always into the upper tail. Thus, all p-value calculations on a χ^2 must use `lower.tail=FALSE` in `distrib()`. For example, the area to the right of $\chi^2 = 6.456$ on a χ^2 distribution with 2 df is 0.0396 (Figure 18.2) and is found with

```
> ( distrib(6.456,distrib="chisq",df=2,lower.tail=FALSE) )
[1] 0.03963669
```

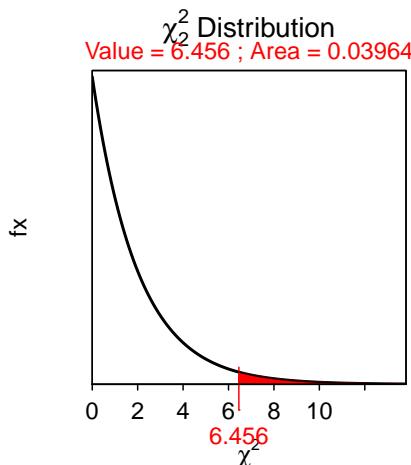


Figure 18.2. Depiction of the area to the right of $\chi^2 = 6.456$ on a χ^2 distribution with 2 df.

Review Exercises

18.1 What is the p-value if $\chi^2 = 10.25$ and $df = 3$? [Answer](#)

18.2 What is the p-value if $\chi^2 = 10.25$ and $df = 4$? [Answer](#)

18.3 What is the p-value if $\chi^2 = 10.25$ and $df = 6$? [Answer](#)

¹

The `type=` argument defaults to "p" so it may be omitted when computing a probability.

18.2 Chi-Square Test Specifics

It is common that one wants to compare the distribution of individuals into the levels of one categorical variable among multiple populations indexed by a second categorical variable. For example, suppose that researchers want to determine if the proportion of failing students differs between males and females, if the proportion of kids playing sports differs between kids from high- or low-income families, if the distribution of four major plant species differs between two locations, or if the distribution of responses to a five-choice question differs between respondents from neighboring counties. All of these questions require the collection of data for two categorical variables and making a comparison among two or more populations. Under these conditions, the goodness-of-fit test is inappropriate. However, the methods and concepts learned for a goodness-of-fit test are extended to what is called a chi-square test. The chi-square test is the subject of this section².

- ◊ This hypothesis test is called a chi-square test, NOT a chi-squareD test.

18.2.1 Hypotheses

The statistical hypotheses for a chi-square test are, in general, “wordy.” Before considering these hypotheses first let’s assume that a two-way contingency table (see Section ??) will be used to summarize the data where the rows will correspond to the levels that represent the separate populations and the columns correspond to the different levels of the response variable. In this organization, the null hypothesis basically says that the row percentage (or proportion) are all equal – i.e., “the percentage or proportional distribution of individuals into the levels of the response variable is the same for all populations.” The alternative hypothesis claims that there is some difference among the row percentages – i.e., “the percentage or proportional distribution of individuals into the levels of the response variable is NOT the same for all populations.”

In instances where there are only two levels of the categorical response variable the hypotheses may be slightly simpler. In these instances, the null hypothesis would be “the proportion of individuals in the level of interest is the same for all populations” whereas the alternative hypothesis is “the proportion of individuals in the level of interest is NOT the same for all populations.”

As one example (more will be shown below), consider the following situation,

An association of Christmas tree growers in Indiana sponsored a survey of Indiana households to help improve the marketing of Christmas trees. In telephone surveys of 421 households they found 160 households in rural areas and 261 households in urban areas. Of the rural households, 64 had a natural tree (as compared to an artificial tree). Of the urban households, 89 had a natural tree. Use these results to determine, at the 10% level, if the proportion of households with a natural tree differs between rural and urban households.

In this case there are two populations (rural and urban areas) and only two levels of the response variable (natural or artificial tree). Thus, the best way to write the hypotheses for this situation is,

$$\begin{aligned} H_0 &: \text{“the proportion of households with a natural tree is the same for urban and rural households”} \\ H_A &: \text{“the proportion of households with a natural tree is NOT the same for urban and rural households”} \end{aligned}$$

²

The chi-square test presented here is quite flexible and can be derived from different types of hypotheses than those described here. This section will only deal with this one type of chi-square test hypothesis.

18.2.2 Tables

As noted above, all two-way contingency tables used for a chi-square analysis will be organized such that the categorical response variable forms the columns and the variable that defines the populations forms the rows. With this organization, the row-percentage table becomes the table of primary interest in a chi-square test because it relates directly to the hypotheses described above. The question of a chi-square test then becomes one of determining whether each row of the row-proportions table is equal given sampling variability.

- ◊ In a chi-square test the categorical variable used to identify the population that an individual belongs to forms the rows of the summary two-way contingency table. Each chi-square test is then a test of whether or not each row in the row-percentage table is equivalent given sampling variability.

The observed raw data must be organized into a two-way observed table using the methods described in Section ???. For example, the Christmas tree data is summarized in a two-way table as shown in Table 18.1. The actual calculations for a chi-square test are performed on this observed table. However, the hypothesis test, as described above, is best viewed as a method for determining if each row in the row-percentage table is statistically equivalent or not. Thus, it is often useful for interpretation of the test results to examine the row-percentage table computed from the observed counts (Table 18.2).

Table 18.1. Contingency table showing the observed frequency of individuals in urban and rural households that have a natural or an artificial Christmas tree.

| Household | Tree Type | | |
|-----------|------------|------------|------------|
| | Natural | Artificial | |
| Urban | 89 | 172 | 261 |
| Rural | 64 | 96 | 160 |
| | 153 | 268 | 421 |

Table 18.2. Contingency table showing the observed (row) percentage of individuals within urban and rural households that have a natural or an artificial Christmas tree.

| Household | Tree Type | | |
|-----------|-------------|-------------|--------------|
| | Natural | Artificial | |
| Urban | 34.1 | 65.9 | 100.0 |
| Rural | 40.0 | 60.0 | 100.0 |
| | 36.3 | 63.7 | 100.0 |

As with all chi-square tests, a corresponding table of expected values, derived from the null hypothesis, must be constructed. Unlike with the goodness-of-fit test the expected table in a chi-square test does not obviously come from a theoretical distribution. The expected table in a chi-square test is derived from the margins of the observed table and is best seen through an illustrative example.

In the Christmas tree example, the null hypothesis states that there is no difference between the rural and urban areas in the proportion of households with a natural tree. Thus, under this null hypothesis, one would expect the proportion of households with a natural tree to be same in both groups. This common proportion is estimated with the proportion of both urban and rural households with a natural tree – i.e., $\frac{153}{421} = 0.363$. Under the null hypothesis, the proportion of rural and the proportion of urban households with a natural tree is 0.363. Because there is a different number of urban and rural households in the study, the actual NUMBER (rather than proportion) of households expected to have a natural tree will differ. The NUMBER of urban households expected to HAVE a natural tree is found by multiplying the number of

urban households by the combined proportion computed above – i.e., $261 * 0.363 = 94.743$. The remaining urban households would be expected to NOT have a natural tree – i.e., $261 - 94.743 = 261(1 - 0.363) = 166.257$. Similar calculations are made for the rural households as follows:

- $160 * 0.363 = 58.08$ rural households to have a natural tree.
- $160 * (1 - 0.363) = 101.92$ rural households to NOT have a natural tree.

These expected frequencies are computed directly and easily from the marginal totals of the original observed frequency table (Table 18.1). For example, when the fractional representation of the decimal proportions are substituted into the calculation for the expected number of urban households with a natural tree that calculation becomes $261 * \frac{153}{421} = \frac{261 * 153}{421}$. A close examination of this formula and the marginal totals in Table 18.1 shows that this value is equal to the product of the corresponding row and column marginal totals in the observed table divided by the total number of individuals. The other expected values are as follows,

- $261 * \frac{268}{421} = \frac{261 * 268}{421} = 166.147$ urban households to NOT have a natural tree.
- $160 * \frac{153}{421} = \frac{160 * 153}{421} = 58.147$ rural households to have a natural tree.
- $160 * \frac{268}{421} = \frac{160 * 268}{421} = 101.853$ rural households to NOT have a natural tree.

All of these expected value calculations follow the same general rule – multiply the row and column totals and divide by the total number of individuals. These expected values are summarized in a two-way table, called the expected frequencies table (Table 18.3).

Table 18.3. Contingency table showing the expected frequency of individuals in urban and rural households that have a natural or an artificial Christmas tree.

| Household | Tree Type | | |
|-----------|------------|------------|------------|
| | Natural | Artificial | |
| Urban | 94.853 | 166.147 | 261 |
| Rural | 58.147 | 101.853 | 160 |
| | 153 | 268 | 421 |

18.2.3 Specifics

The chi-square test is characterized by categorical data of two or more categories recorded for two or more populations. The specifics of the chi-square test are identified in Table 18.4.

In general, confidence intervals are not constructed in relation to a chi-square test because of the complexity of the parameter (i.e., same size as the observed table). Thus, in this book, step 11 will never be computed for a chi-square test.

◊ Step 11 will not be computed for a chi-square test.

18.2.4 Example – Christmas Trees

The 11-steps (Section 15.1) for completing a full hypothesis test for the Christmas tree example presented at the beginning of this chapter are as follows:

Table 18.4. Characteristics of a chi-square test.

- **Null Hypothesis:** “The proportional distribution of individuals into the levels of the response variable is the same for all populations”
- **Alternative Hypothesis:** “The proportional distribution of individuals into the levels of the response variable is NOT the same for all populations.”
- **Statistic:** Observed frequency table.
- **Test Stat:** $\chi^2 = \sum_{cells} \frac{(Observed - Expected)^2}{Expected}$
- **df:** $(r - 1)(c - 1)$ where r = number of rows and c = number of columns
- **Assumptions:** Expected value for each category is ≥ 5 .

1. As stated, α should be set at 0.10.
2. The H_0 : “the proportion of households with a natural tree is the same for urban and rural households” versus H_A : “the proportion of households with a natural tree is NOT the same for urban and rural households.”
3. A chi-square test is required because a categorical response variable with two levels (natural and artificial trees) measured on two populations (urban and rural households) was taken and the distribution of responses is being compared among populations in the null hypothesis.
4. The data appear to be part of an observational study with no clear indication of randomization. The observed frequencies are shown in the following table,

| Household | Tree Type | | |
|-----------|------------|------------|------------|
| | Natural | Artificial | |
| Urban | 89 | 172 | 261 |
| Rural | 64 | 96 | 160 |
| | 153 | 268 | 421 |

5. The expected frequency for each cell is shown in the table below,

| Household | Tree Type | | |
|-----------|------------|------------|------------|
| | Natural | Artificial | |
| Urban | 94.853 | 166.147 | 261 |
| Rural | 58.147 | 101.853 | 160 |
| | 153 | 268 | 421 |

The expected count in each of the four cells of the table is greater than five. Thus, the assumptions are met and the test statistic computed below should reasonably follow a χ^2 distribution.

6. The statistic is the observed frequency table shown in Step 4 above.
7. The test statistic is $\chi^2 = \frac{(89-94.853)^2}{94.853} + \frac{(172-166.147)^2}{166.147} + \frac{(64-58.147)^2}{58.147} + \frac{(96-101.853)^2}{101.853} = 0.3611 + 0.2062 + 0.5891 + 0.3363 = 1.4927$ with 1 df.
8. The p-value is $p = 0.0264$.

```
> ( distrib(4.927,distrib="chisq",df=1,lower.tail=FALSE) )
Error in plot.window(...): need finite 'ylim' values
```

9. The H_0 is not rejected because the p-value is $> \alpha$.
10. There does not appear to be a significant difference between the proportion of rural and the proportion of urban households that have a natural Christmas tree.

Review Exercises

- 18.4** Researchers in Asia (Roberts, 2000) wanted to describe the distribution of the fish genera Cyprinidae in Asian rivers. They collected 228 fish from the Brahmaputra, Irrawaddy, and Salween rivers and recorded whether the fish was a member of the Cyprinidae family or not. Because the rivers were relatively equal in size, they expected the same proportions of Cyprinidae in each of the rivers. Using the data in the table below, test to see if there was a difference in the proportion of Cyprinidae among the rivers at the 5% level.

[Answer](#)

| River | Cyprinidae | |
|-------------|------------|----|
| | Yes | No |
| Brahmaputra | 22 | 51 |
| Irrawaddy | 25 | 53 |
| Salween | 30 | 47 |

- 18.5** The American Nurses Credentialing Center (ANCC) has created guidelines for nursing administration. Some research findings have suggested that ANCC-recognized hospitals also have favorable practice environments for nurses. To study this further and in relation to oncology units, Friese (2005) examined the practice environments and outcomes of nurses working in and out of oncology units in hospitals that adhere and don't adhere to the ANCC guidelines. As part of his study, he determined, through surveys, whether nurses were experiencing high emotional exhaustion (HEE) or not. The results of his study are shown in the table below (note "onc" represent oncology units). Use these results to determine, at the 5% level, if the proportion of nurses experiencing HEE differs among the four categories of hospitals. [Answer](#)

| Clinic Type | HEE | not HEE | total |
|-------------------|-----|---------|-------|
| non-ANCC, non-Onc | 362 | 534 | 896 |
| non-ANCC, Onc | 58 | 92 | 150 |
| ANCC, non-Onc | 197 | 558 | 755 |
| ANCC, Onc | 30 | 125 | 155 |
| total | 647 | 1309 | 1956 |

- 18.6** Fiebach *et al.* (1990) examined the immediate survival of 790 males and 332 females who were hospitalized following a myocardial infarction (i.e., a "heart attack"). During hospitalization, 70 men and 47 women died. Is there a difference, at the 5% level, in mortality rate (proportion of patients that died) between men and women during hospitalization? [Answer](#)

18.7 Eight American undergraduate women were part of a study to determine if whether or not a response is received depends on the size of group addressed (Jones and Foshay 1984). Each student was instructed to say “Hello” to strangers or groups of strangers that they encountered around campus, on the streets in town, in stores, etc. They were told to not make direct eye contact with anyone in the group but to look in the general direction of the group focusing on the shoulders or hair of individuals or the general middle of a group. The students recorded a variety of information for each encounter including how many individuals were in the group and whether at least one person responded to the greeting. The study included 119 people greeted individually, 94 groups of two or three, and 27 groups of four, five or six. They found that 92 of the individuals, 65 of the groups of two or three, and 13 of the groups of four, five, or six responded to the greeting. Determine, at the 5% level, if there is a significant difference in the frequency of responses among the three different sizes of groups (i.e., individuals; two or three; or four, five, or six). [Answer](#)

18.3 Chi-Square Test in R

18.3.1 Data Format

As with the goodness-of-fit test, the data for a chi-square test are entered from summarized data or computed from “raw” data on individuals. The raw individual data must be in the stacked format where one column in the data frame represents the response variable and another column represents the variable that denotes the populations. The raw data must be summarized into a two-way frequency table with `table()` as described in Section ?? and saved into an object. The two-way table must contain frequencies and not proportions or percentages (so don’t use `percTable()`) and must not contain the marginal totals (so don’t use `addMargins()`).

In contrast to the goodness-of-fit test, the summarized data must be entered into a two-dimensional **matrix** rather than a one-dimensional vector. The creation of a matrix first requires that the summarized data be entered into a vector such that the values that will be the first row of the matrix are followed by the values that will form the second row which are followed by the values that will form the third row and so on³. This vector of values will serve as the first argument to `matrix()`. Additionally, `matrix()` should include the number of rows to be in the final matrix in the `nrow=` argument and `byrow=TRUE` to indicate that the values in the vector should be entered into the matrix in a row-wise manner.

- ◊ Observed frequencies are entered into a matrix by first entering the data into a vector such that the values in any row follow the values of the row that preceded it.

The process of entering summarized data into a matrix is better explained by example. Suppose that you are given the observed frequencies shown in this table.

³

The data could be entered such that the values in the first column are followed by the values in the second column and so on, but it is generally easier to enter the values as if you were reading a paragraph – left-to-right, top-to-bottom.

| Location | Species | | | | | | |
|----------|---------|----|----|----|----|----|-----|
| | A | B | C | D | E | F | |
| DI | 34 | 22 | 14 | 13 | 12 | 5 | 100 |
| BP | 62 | 12 | 8 | 7 | 6 | 5 | 100 |
| | 96 | 34 | 22 | 20 | 18 | 10 | 200 |

The observed frequencies, ignoring the marginal sums, are entered into an R object with the following code (notice how the values from the second row follow the values from the first row),

```
> # put frequencies into a vector first
> freq <- c(34,22,14,13,12,5,62,12,8,7,6,5)
[1] 34 22 14 13 12 5 62 12 8 7 6 5
> # allocate those frequencies by row to a matrix with two rows
> obstbl <- matrix(freq,nrow=2,byrow=TRUE)
[,1] [,2] [,3] [,4] [,5] [,6]
[1,] 34   22   14   13   12   5
[2,] 62   12   8    7    6    5
```

It would be better if the rows and columns of this matrix were named. Following the construction of the matrix, the rows and columns of the matrix are named with `rownames()` and `colnames()`, respectively. Each of these functions uses the named matrix object as its only argument and it is assigned a vector that contains the desired names of the rows and columns, respectively. The vector of names must have exactly as many names as there are rows and columns. The rows and columns of the `obstbl` object created above are named with,

```
> rownames(obstbl) <- c("DI","BP")
> colnames(obstbl) <- c("A","B","C","D","E","F")
> obstbl
     A   B   C   D   E   F
DI 34  22  14  13  12  5
BP 62  12   8   7   6   5
```

18.3.2 Chi-Square Test

The chi-square test is performed with `chisq.test()`. The table of summary results, either entered through `matrix()` or generated through `table()`, is the first argument to this function. The only other argument that may be needed is the `correct=` argument for applying the continuity correction as described for the goodness-of-fit test. As per usual, the results of `chisq.test()` should be assigned to an object so that the observed table, expected table, and a visual of the p-value can be easily extracted.

18.3.3 Post-Hoc Analysis

Rejecting the null hypothesis in a chi-square test indicates that there is some difference in the distribution of individuals into the levels of the response variable among some of the populations. However, rejecting the null hypothesis does not indicate which populations are different. In addition, as mentioned previously, confidence intervals are generally not performed with a chi-square test⁴. A post-hoc method for helping determine which

⁴

populations are different following the rejection of a null hypothesis is obtained by observing the so-called Pearson residuals.

A Pearson residual is computed for each cell in the table as,

$$\frac{Observed - Expected}{\sqrt{Expected}}$$

which is basically the appropriately signed square root of the parts in the χ^2 test statistic calculation. Therefore, cells that have Pearson residuals far from zero have contributed substantially to the large χ^2 test statistic that resulted in a small p-value and the ultimate rejection of H_0 . Patterns in where the large Pearson residuals are found may allow one to qualitatively determine which populations differ and, thus, which levels of the response differ the most. This process will be illustrated more fully in the examples and review exercises. The Pearson residuals are obtained from the saved `chisq.test()` object by appending `$residuals` to the object name – e.g., `chi.result$residuals`.

18.3.4 Example - Father Present at Birth

Consider this situation,

Daniel Weiss (in “100% American”) reported the results of a survey of 300 first-time fathers from four different hospitals (labeled as A, B, C, and D). Each father was asked if he was present (or not) in the delivery room when his child was born. The results of the survey are in [FatherPresent.csv](#). Use these data to determine if there is a difference, at the 5% level, in the proportion of fathers present in the delivery room among the four hospitals.

The 11-steps (Section 15.1) for completing a full hypothesis test for this situation are as follows:

1. As stated, α should be set at 0.05.
2. The H_0 : “The proportion of fathers present during the birth of their child is the same for all four hospitals” versus H_A : “The proportion of fathers present during the birth of their child is NOT the same for all four hospitals.”
3. A chi-square test is required because a categorical variable with two levels (present or absent) was measured on four populations (the hospitals) and the distributions into the levels is being compared among populations in the null hypothesis.
4. The data appear to be part of an observational study with no clear indication of randomization (likely a voluntary response survey). The raw data were loaded into R with

```
> fp <- read.csv("data/FatherPresent.csv")
> str(fp)
'data.frame': 300 obs. of  2 variables:
 $ hospital: Factor w/ 4 levels "A","B","C","D": 1 1 1 1 1 1 1 1 1 ...
 $ father   : Factor w/ 2 levels "Absent","Present": 2 2 2 2 2 2 2 2 2 ...
```

It won't be done in this book, but a common exception to this rule is to compute a confidence interval for the difference in proportions when there are only two levels of the response variable and two populations.

5. The observed table was constructed, saved, and submitted to `chisq.test` for, at this stage, observing the expected frequency table.

```
> ( fp.obs <- xtabs(~hospital+father,data=fp) )
      father
hospital Absent Present
  A      9      66
  B     15      60
  C     18      57
  D     19      56
> fp.chi <- chisq.test(fp.obs)
> fp.chi$expected
      father
hospital Absent Present
  A  15.25  59.75
  B  15.25  59.75
  C  15.25  59.75
  D  15.25  59.75
```

The test statistic computed below should reasonably follow a χ^2 distribution, because there are at least five individuals in each cell of the expected table shown above.

6. The statistic is the observed frequency table shown in the output of Step 5 above.
 7. The test statistic is $\chi^2=5$ with 3 df as seen with

```
> fp.chi
Pearson's Chi-squared test with fp.obs
X-squared = 5.0003, df = 3, p-value = 0.1718
```

8. The p-value is $p = 0.1718$ as seen in the results above.
 9. The null hypothesis is not rejected because the p-value is $> \alpha$.
 10. There does not appear to be a significant difference between the proportion of fathers that were present at their child's birth and the hospital where that birth occurred. For comparative purposes, the row-percentage table is seen with

```
> percTable(fp.obs,margin=1,digits=2)
      father
hospital Absent Present Sum
  A 12.00  88.00 100.00
  B 20.00  80.00 100.00
  C 24.00  76.00 100.00
  D 25.33  74.67 100.00
```

18.3.5 Example - Apostle Islands Plants

Consider this situation,

In her Senior Capstone project a Northland College student recorded the dominant (i.e., most abundant) plant species in 100 randomly selected plots on both Devil's Island and the Bayfield Peninsula (i.e., the mainland). There were a total of six "species" (one group was called "other") recorded (labeled as A, B, C, D, E, and F). The results are shown in the table below. Determine, at the 5% level, if the frequency of dominant species differs between the two locations.

| Location | Species | | | | | | |
|----------|-----------|-----------|-----------|-----------|-----------|-----------|------------|
| | A | B | C | D | E | F | |
| DI | 34 | 22 | 14 | 13 | 12 | 5 | 100 |
| BP | 62 | 12 | 8 | 7 | 6 | 5 | 100 |
| | 96 | 34 | 22 | 20 | 18 | 10 | 200 |

The 11-steps (Section 15.1) for completing a full hypothesis test for the Northland College student's Senior Capstone example above is as follows

1. As stated, α should be set at 0.05.
2. The H_0 : "The distribution of dominant plants species is the same between Devil's Island and the Bayfield Peninsula" versus H_A : "The distribution of dominant plants species is NOT the same between Devil's Island and the Bayfield Peninsula."
3. A chi-square test is required because a categorical variable with six levels (plant species) was measured on two populations (Devil's Island and Bayfield Peninsula) and the distributions are being compared in the null hypothesis.
4. The data appear to be part of an observational study where the plots were randomly selected. The observed frequency table given in the background information was entered into R with

```
> freq <- c(34,22,14,13,12,5,62,12,8,7,6,5)
> ai.obs <- matrix(freq,nrow=2,byrow=TRUE)
> rownames(ai.obs) <- c("DI","BP")
> colnames(ai.obs) <- c("A","B","C","D","E","F")
> ai.obs
   A  B  C  D  E  F
DI 34 22 14 13 12 5
BP 62 12  8  7  6 5
```

This observed table was submitted to `chisq.test()` for, at this stage, observing the expected frequency table.

```
> ai.chi <- chisq.test(ai.obs)
> ai.chi$expected
   A  B  C  D  E  F
DI 48 17 11 10 9 5
BP 48 17 11 10 9 5
```

The test statistic computed below should reasonably follow a χ^2 distribution, because there are more than five individuals in each cell of the expected table shown above.

5. The statistic is the observed frequency table shown in the output of Step 4 above.

6. The test statistic is $\chi^2=16.54$ with 5 df as seen with

```
> ai.chi
Pearson's Chi-squared test with ai.obs
X-squared = 16.5442, df = 5, p-value = 0.00545
```

7. The p-value is $p = 0.0055$ as seen in the results above.
 8. The null hypothesis is rejected because the p-value is $< \alpha$.
 9. There does appear to be a significant difference in the distribution of the dominant plants between the two sites. A look at the Pearson residuals,

```
> ai.chi$residuals
      A          B          C          D          E          F
DI -2.020726  1.212678  0.904534  0.9486833  1 0
BP  2.020726 -1.212678 -0.904534 -0.9486833 -1 0
```

and the row-percentage table,

```
> percTable(ai.obs,margin=1,digits=2)
   A  B  C  D  E  F Sum
DI 34 22 14 13 12  5 100
BP 62 12  8  7  6  5 100
```

both suggest that the biggest difference between the two locations is due to “plant A.”⁵

Review Exercises

18.8  Saenz *et al.* (1998) examined the effectiveness of “restrictor plates” (a metal plate designed to reduce “pecking” by pileated woodpeckers (*Dryocopus pileatus*) in reducing damage by pileated woodpeckers) on cavity trees for red-cockaded woodpeckers (*Picoides borealis*) in Eastern Texas. For each red-cockaded woodpecker cavity hole they recorded whether the hole was fit with a restrictor plate or not and, ultimately, whether the cavity hole was damaged or not. The results of their study are recorded in [RestrictorPlates.csv](#). Examine these data to determine, at the 5% level, if restrictor plates reduced the damage done by pileated woodpeckers. [Answer](#)

18.9  On the eastern slopes of the Rocky Mountains in Colorado, Wyoming, and Montana, whitetail deer (*Odocoileus virginianus*), mule deer (*Odocoileus hemionus*), and elk (*Cervus canadensis*) habitats overlap. It has been observed that in these areas where these species interact, diseases common to each species tend to infect more animals than in other areas. To examine this phenomenon, infection information on all three species was observed from individuals killed during the hunting seasons in areas where the habitats overlapped. In particular, it was recorded whether the animal was infected with one of the diseases common to each species or not. These data are recorded in [CervidDisease.csv](#). Test at the 1% significance level if there is a difference in the infection rate among the three species. [Answer](#)

⁵

When “Plant A” is removed from the observed table, the chi-square test performed on the remaining plant species showed no difference in the distribution of the remaining plants between the two locations ($p = 0.9239$). Thus, most of the difference in plant distributions between Devil’s Island and the Bayfield Peninsula appears to be due primarily to “plant A” with more of “plant A” found on the Bayfield Peninsula than on Devil’s Island.

- 18.10** Ashland High School conducted a survey to determine if parents or students favored the idea of uniforms being required apparel for attending school (December 5, 1999, Ashland Daily Press). The surveys were administered to 223 parents at parent-teacher conferences and to 572 students by the Student Council. No other information about the surveys was given in the report. From these surveys it was learned that 70 parents and 101 students FAVORED the wearing of uniforms. Determine, at the 5% level, if there is a difference in the level of support for wearing uniforms between parents and students. [Answer](#)

- 18.11** Five hundred patients participated in a comparison of the effectiveness of three arthritic pain relievers (175 received medication A, 150 received medication B, and 175 received medication C). Each patient used one of the three medications for one month and then was asked if the product was effective. The results showed 115 patients using medication A, 78 patients using medication B, and 140 patients using medication C said their medication was effective. Test, at the 10% level, if there is a difference in effectiveness among the three medications. [Answer](#)

- 18.12** USA Today presented two sets of data on why Americans don't exercise. One set was for 1000 randomly selected men. The other was for 1000 randomly selected women. The results of the surveys are recorded in [Exercise.csv](#). Determine, at the 1% level, if the distribution of men and women differs among the six responses given. [Answer](#)

- 18.13** Fairley *et al.* (1994) gave the results in the table below concerning the age and the number who were positive for human papillomavirus infection among the 290 participants in their study. Test the hypothesis, at the 5% level, that the same proportion for each age-group is HPV-positive. [Answer](#)

| Age | n | HP+ |
|----------|-----|-----|
| under 20 | 27 | 11 |
| 21-25 | 81 | 30 |
| 26-30 | 108 | 34 |
| 31-35 | 74 | 18 |

- 18.14** Passengers aboard the RMS Titanic were classified as to their "class" (first, second, third, or crew) and whether or not they survived the wreck (yes or no). Use the data found in [Titanic.csv](#) to determine if there was a difference, at the 1% level, in the survival rate among the classes of passengers. [Answer](#)

- 18.15** Meliker *et al.* (2004) examined the records of 773 motor-vehicle crashes in southeastern Michigan. Of these, 139 had a driver with a blood alcohol level greater than 0.10% and were defined as alcohol-related crashes. Of these alcohol-related drivers, 79% were male, while 56% of the non-alcohol-related drivers were male. Use this information to determine, at the 1%, if males are more or less likely to be involved in an alcohol-related crash than females. [HINT: I'd construct a 2x2 contingency table (Section ??) of these results with the response variable as columns. Note that the results as presented above are in column-percentage format and the results needed to answer the question are row-percentage format. Also, note that the column totals are given indirectly in the information above but the row totals need to be determined.] [Answer](#)

- 18.16** Shrimp trawlers are required to have turtle exclusion device (TED), that allows most loggerhead sea turtles (*Caretta caretta*) to escape the net, thus reducing turtle mortality due to by-catch. In the Gulf of Mexico, the TEDs were originally required to be 32" x 10" but a new law now requires to TEDs to be 71" x 26" with the thought that turtle mortality would be further reduced by the larger opening. This thought was examined by recording the number of trawl tows that had at least one turtle mortality. In 75 tows with the

original smaller opening there were 16 tows with at least one turtle mortality. In contrast, in 88 tows with the newer larger opening there were 8 tows with at least one turtle mortality. Test at the 10% level if there is a significant difference in the proportion of trawl tows with at least one turtle mortality between trawls with the different sized openings. [Answer](#)

- 18.17**  Researchers observed groups of dolphins off the coast of Iceland near Keflavik in 1998⁶. The researchers recorded the time of the day (“Morning”, “Noon”, “Afternoon”, and “Evenings”) and the main activity of the group, whether travelling quickly (“Travel”), feeding (“Feed”), or socializing (“Social”). The number of dolphin groups observed by each time of day and activity is shown in the table below. Use this information to determine, at the 5% level if the proportion of groups exhibiting each activity differs by time of day.

[Answer](#)

| Time of Day | Activity | | |
|-------------|----------|------|--------|
| | Travel | Feed | Social |
| Morning | 6 | 28 | 38 |
| Noon | 6 | 4 | 5 |
| Afternoon | 14 | 0 | 9 |
| Evening | 13 | 56 | 10 |

- 18.18**  The data in Zoo1.csv contains a list of animals found in several different zoos⁷. In addition, each animal was classified into broad “type” categories (“mammal”, “bird”, and “amph/rep”). The researchers that collected these data wanted to examine if the distribution of broad animal types differed among zoos. Test the researcher’s question at the 5% level [Answer](#)
-

⁶

Data was originally from [here](#).

⁷

These data are stored in a “comma separated values” (CSV) file rather than a “tab delimited text” file. Thus, these data must be loaded into R with `read.csv()` rather than `read.csv()`. The arguments to `read.csv()` are the same as `read.csv()`.

CHAPTER 19

GOODNESS-OF-FIT TEST

Chapter Objectives:

1. Identify when a goodness-of-fit test is appropriate.
2. Perform the 11 steps of a significance test in a goodness-of-fit test situation.

Contents

| | |
|---|-----|
| 19.1 Goodness-of-Fit Test Specifics | 252 |
| 19.2 Goodness-of-Fit Test in R | 259 |

THE CHI-SQUARE DISTRIBUTION and the general chi-square test for a categorical response with two or more populations were introduced in Chapter 18. The goodness-of-fit for a categorical response recorded for one population and compared to a theoretical distribution is focus of this chapter.

19.1 Goodness-of-Fit Test Specifics

It is a common question to determine if the frequency of individuals in the various levels of a categorical variable follow frequencies suggested by a particular distribution. The simplest of these situations occurs when a researcher is making a hypothesis about the percentage or proportion of individuals in one of two categories. The “distribution” of individuals in two categories comes from the proportion in the hypothesis for one group and one minus the proportion in the hypothesis for the other group. In situations with more than two levels, the “distribution” of individuals into the categories likely comes from the hypothesis that a particular theoretical distribution holds true. For example, a researcher may want to determine if frequencies predicted from a certain genetic theory are upheld by the observed frequencies found in a breeding experiment, if the frequency that a certain animal uses various habitats is in proportion to the availability of those habitats, or if the frequency of consumers that show a preference for a certain product (over other comparable products) is non-random.

In each of these cases, the theoretical distribution articulated in the research hypothesis must be converted to statistical hypotheses that will then be used to generate expected frequencies for each level. These expected frequencies will then be statistically compared to the observed frequencies to determine if the theoretical distribution represented in the null hypothesis is supported by the data. The method used for comparing the observed to expected frequencies where the expected frequencies come from a hypothesized theoretical distribution is called a chi-square goodness-of-fit test, or simply a goodness-of-fit test, and is the subject of this section.

19.1.1 The Hypotheses

A goodness-of-fit test is used when a single categorical variable has been recorded and the frequency of individuals in the levels of this variable are to be compared to a theoretical distribution. In its most general form the statistical hypotheses for the goodness-of-fit test will be “wordy.” The language in the statistical hypotheses will relate to whether the “distribution” of individuals into the levels of the variable follows a specific theoretical distribution. The null hypothesis will generically look something like H_0 : “the observed distribution of individuals into the levels follows the ‘theoretical distribution’ ”, where ‘theoretical distribution’ will likely be replaced with more specific language. For example, the research hypothesis that states that “50% of students at Northland are from Wisconsin, 25% are from neighboring states, and 25% are from other states” would be converted to these statistical hypotheses – H_0 : “the proportion of students from Wisconsin, neighboring states, and other states is 0.50, 0.25, and 0.25, respectively” with an H_A : “the proportion of students from Wisconsin, neighboring states, and other states is NOT 0.50, 0.25, and 0.25, respectively.”

- ◊ The statistical hypotheses for a goodness-of-fit test are “wordy” and relate the observed distribution of individuals into levels of the categorical variable to those expected from a theoretical distribution.

The hypotheses are simpler, but you must be more careful, when there are only two levels. For example, the research hypothesis that states that “less than 40% of new-born bear cubs are female” would be converted

to H_0 : “the proportion of bear cubs that are female and male is 0.40 and 0.60, respectively” with an H_A : “the proportion of bear cubs that are female and male is NOT 0.40 and 0.60, respectively.” However, these hypotheses are often simplified to focus on only one level as the other level is implied by subtraction from one. Thus, these statistical hypotheses are more likely to be written as H_0 : “the proportion of bear cubs that are female is 0.40” with an H_A : “the proportion of bear cubs that are female is NOT 0.40.”

- ◊ The statistical hypotheses for a goodness-of-fit test with only two levels of the categorical variable often relate only to the proportion or percentage of individuals in one level.

One may also have expected, from the wording of the research hypothesis about the sex of bear cubs, that the alternative hypothesis would have been H_A : “the proportion of bear cubs that are female is LESS THAN 0.40.” Recall, however, that the chi-square test statistic always represents the two-tailed situation. Thus, the H_A here must reflect that constraint. The researcher will ultimately be able to determine if the proportion is less than 0.40 if the p-value from the goodness-of-fit test indicates a difference and the observed proportion of female bear cubs is less than 0.40.

19.1.2 The Tables

The observed data are usually summarized in a raw frequency table as shown in Section ???. In the context of a goodness-of-fit test this table is called the **observed frequency table**.

In addition to the observed frequency table, a table of expected frequencies must be constructed from the theoretical distribution of proportions in the null hypothesis and the total number of observed individuals (n). Specifically, the expected frequencies are found by multiplying the expected theoretical proportions in the null hypothesis by n . For example, consider this situation,

Bath and Buchanan (1989) surveyed residents of Wyoming by distributing a mailing to a random selection of residents and collecting voluntarily returned surveys. One question asked of the respondents was, “Do you strongly agree, agree, neither agree or disagree, disagree, or strongly disagree with this statement? – ‘Wolves would have a significant impact on big game hunting opportunities near Yellowstone National Park.’” The researchers hypothesized that more than 50% of Wyoming residents would either disagree or strongly disagree with the statement. Of the 371 residents that returned the survey, 153 disagreed and 43 strongly disagreed with the statement.

At first glance it may seem that this variable has five levels – i.e., the levels of agreement offered in the actual survey. However, the researcher’s hypothesis collapsed the results of the survey question into two levels: (1) strongly disagree or disagree combined and (2) all other responses. Thus, the statistical hypotheses for this situation are H_0 : “the proportion of respondents that disagreed or strongly disagreed is 0.50” and H_A : “the proportion of respondents that disagreed or strongly disagreed is NOT 0.50.”

The expected frequencies in each level are derived from the total number of individuals examined and the specific null hypothesis. For example, if the null hypothesis is true, then 50% of the 371 respondents would be expected to disagree or strongly disagree with the statement. In other words, $371 * 0.50 = 185.5$ individuals would be expected to disagree or strongly disagree. Furthermore, the other 50%, or $371 * (1 - 0.50) = 185.5$ would be expected to “not” disagree or strongly disagree. The expectations for the two levels of this variable are summarized as in Table 19.1. The observed frequencies in each category are usually appended to the expected frequency table as another column (Table 19.1).

Table 19.1. Expected and observed frequency of respondents that disagreed or strongly disagreed (i.e., labeled as “Disagree”) with the given statement in the Wyoming survey example.

| Category | Frequency | |
|----------------|-----------|----------|
| | Expected | Observed |
| “Disagree” | 185.5 | 196 |
| not “Disagree” | 185.5 | 175 |
| Total | 371 | 371 |

- ◊ The expected table should maintain at least one decimal in each cell even though the values represent frequencies.

The hypothesis test method developed in the following sections will be used to determine if the differences between the expected and observed frequencies in these categories is “large” enough to suggest that the observed frequencies do not support the distribution represented in the null hypothesis. Before developing this methodology, though, consider the following situation as an illustration where the construction of expected frequencies is bit more complex.

Mendel’s law of independent assortment predicts that the genotypes (i.e., how they look) of the offspring from mating the offspring of a dihybrid cross of homozygous dominant and homozygous recessive parents should follow a 9:3:3:1 ratio. In an experiment to test this, Mendel crossed a pea plant that produces round, yellow seeds (i.e., all dominant alleles, YYWW) with a pea plant that produces green, wrinkled seeds (i.e., all recessive alleles, yyww) such that only round, yellow heterozygous offspring (i.e., YyWw) were produced. Pairs of these offspring were then bred. Mendel’s theory says that $\frac{9}{16}$ of these offspring should be round, yellow; $\frac{3}{16}$ should be round, green; $\frac{3}{16}$ should be wrinkled, yellow; and $\frac{1}{16}$ should be wrinkled, green. Of 566 seeds studied in this experiment, Mendel found that 315 were round, yellow; 108 were round, green; 101 were wrinkled, yellow; and 32 were wrinkled, green. Use these results to determine, at the 5% level, if Mendel’s law of independent assortment is supported by these results.

The statistical hypotheses are as follows,

$$H_0 : \text{“the proportion of RY, RG, WY, and WG individuals will be } \frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \text{ and } \frac{1}{16}, \text{ respectively”}$$

$$H_A : \text{“the proportion of RY, RG, WY, and WG individuals will NOT be } \frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \text{ and } \frac{1}{16}, \text{ respectively”}$$

where RY=“round, yellow”, RG=“round, green”, WY=“wrinkled, yellow”, and WG=“wrinkled, green”. If these proportions are applied to the $n = 566$ observed offspring, then the following frequencies for each genotype would be expected:

- $\frac{9}{16}566 = 318.375$ would be expected to be round, yellow.
- $\frac{3}{16}566 = 106.125$ would be expected to be round, green.
- $\frac{3}{16}566 = 106.125$ would be expected to be wrinkled, yellow.
- $\frac{1}{16}566 = 35.375$ would be expected to be wrinkled, green.

These expected frequencies, along with the observed frequencies, are summarized in Table 19.2.

Table 19.2. Expected and observed frequency of 566 pea seeds in four types.

| Category | Frequency | |
|------------------|-----------|----------|
| | Expected | Observed |
| round, yellow | 318.375 | 314 |
| round, green | 106.125 | 108 |
| wrinkled, yellow | 106.125 | 101 |
| wrinkled, green | 35.375 | 32 |
| Total | 566 | 566 |

19.1.3 Specifics

The goodness-of-fit test is characterized by a single categorical variable with two or more levels. The hypotheses tested usually cannot be converted to mathematical symbols and are thus “word” hypotheses. Specifics of the goodness-of-fit test are shown in Table 19.3.

Table 19.3. Characteristics of a goodness-of-fit test.

- **Hypotheses:** H_0 :“the observed distribution of individuals into the levels follows the ‘theoretical distribution’ ”
 H_A :“the observed distribution of individuals into the levels DOES NOT follow the ‘theoretical distribution’ ”
- **Statistic:** Observed frequency table.
- **Test Stat:** $\chi^2 = \sum_{cells} \frac{(Observed - Expected)^2}{Expected}$
- **df:** Number of levels minus 1.
- **Assumptions:** Expected value in each level is ≥ 5 .

As per our usual steps, it is customary to produce a confidence region following the rejection of a null hypothesis. This is cumbersome in a goodness-of-fit test because there generally is not a single parameter (i.e., there are as many parameters as levels in the variable) for which a single confidence region is computed. However, there is a method for computing these multiple confidence intervals. The method will be discussed here but will only be computed “by hand” in the situation where there are two levels. The method will be implemented when using R (as discussed in a subsequent section) no matter the number of levels.

Let p be the population proportion in a particular level and \hat{p} be the sample proportion in the same interval. The \hat{p} is computed by dividing the frequency of individuals in this level by the total number of individuals in the sample (i.e., n). The \hat{p} is a statistic that is subject to sampling variability measured by $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ for “large” values of n . For “large” values of n the \hat{p} will follow a normal distribution such that a confidence interval for p is computed using the general confidence interval formula found in Section 14.2 and repeated below:

$$\text{“Statistic”} (\pm \text{“scaling factor”}) * SE_{statistic}$$

where the scaling factor is the familiar Z^* . Thus, the confidence interval for p is constructed with

$$\hat{p} \pm Z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Note that one does not need to worry about lower and upper bounds, only confidence intervals will be computed, because of the two-tailed nature of the chi-square test statistic.

In the Wyoming survey example, the proportion of respondents in the sample that either disagreed or strongly disagreed was $\hat{p} = \frac{196}{371} = 0.528$. The standard error for this sample proportion is $\sqrt{\frac{0.528(1-0.528)}{371}} = 0.026$. For a 95% confidence interval the $Z^* = \pm 1.960$ as computed with

```
> ( distrib(0.975,type="q") )
[1] 1.959964
```

Thus, the confidence interval is $0.528 \pm 1.960 * 0.026$ or 0.528 ± 0.051 or $(0.477, 0.579)$. Therefore, one is 95% confident that the true population proportion that either disagreed or strongly disagreed is between 0.477 and 0.579. Because there are only two levels in this example it can also be said with 95% confidence that the population proportion that did not either disagree or strongly disagree is between 0.421 and 0.523.

19.1.4 Example - \$1 Coins

Consider the following situation,

USA Today (June 14, 1995) reported that 77% of the population opposes replacing \$1 bills with \$1 coins. To test if this claim holds true for the residents of Ashland a student selected a sample of 80 Ashland residents and found that 54 were opposed to replacing the bills with coins. Develop a hypothesis test (at the 10% level) to determine if the proportion of Ashland residents that are opposed to replacing bills with coins is different from the proportion opposed for the general population.

The 11-steps (Section 15.1) for completing a full hypothesis test for this example are as follows:

1. As stated, α should be set at 0.10.
2. The H_0 : “The proportion of Ashland residents that oppose replacing the \$1 bill with the \$1 coin is 0.77” and the H_A : “The proportion of Ashland residents that oppose replacing the \$1 bill with the \$1 coin is NOT 0.77.”
3. A goodness-of-fit test is required because a single categorical variable from a single population was recorded and the frequency of responses is being compared to a hypothesized distribution in the null hypothesis.
4. The data appear to be part of an observational study with no clear indication of random selection of individuals.
5. The expected number in the “oppose” level is $80 * 0.77 = 61.6$. The expected number in the “do not oppose” category is $80 * 0.23 = 18.4$. These expectations are shown in the table in the next step. The assumption of more than five individual in all cells of the expected table has been met.

6. The observed table is shown below (along with the expected table).

| Level | Frequency | |
|-----------------|-----------|----------|
| | Expected | Observed |
| “Oppose” | 61.6 | 54 |
| “Do Not Oppose” | 18.4 | 26 |
| Total | 80 | 80 |

7. The test statistic is $\chi^2 = \frac{(61.6-54)^2}{55} + \frac{(18.4-26)^2}{25} = 0.938 + 3.139 = 4.077$ with $2 - 1 = 1$ df.

8. The p-value is $p = 0.0435$ as computed with

```
> ( distrib(4.077,distrib="chisq",df=1,lower.tail=FALSE) )
Error in plot.window(...): need finite 'ylim' values
```

9. The H_0 is rejected because the $p-value < \alpha = 0.10$.
10. The proportion of Ashland residents that oppose replacing the \$1 bill with the \$1 coin does appear to be different from the proportion (0.77) reported for the general population.
11. A 90% confidence interval is warranted using $z^* = 1.645$ as determined with

```
> ( distrib(0.95,type="q") )
[1] 1.644854
```

The sample proportion opposing the \$1 coin is $\frac{54}{80} = 0.68125$ with a standard error of $\sqrt{\frac{0.68125 * 0.31875}{80}} = 0.0521$. Thus, $0.68125 \pm 1.645 * 0.0521$, 0.68125 ± 0.0857 , and $(0.5956, 0.7670)$. Therefore, one is 90% confident that the proportion of all Ashland residents opposed to the \$1 coin is between 0.596 and 0.767.

Review Exercises

- 19.1** In the same study used in the example of this chapter, Bohall-Wood (1987) more closely examined the habitat use of the shrikes observed in the open habitat by looking at four “sub-habitats” within these areas. Of the 1456 shrike observations in this habitat, 149 were in “settled” areas, 944 were in improved pastures, 192 were in overgrown pastures, and 171 were in crop fields. In addition, 20.5% of this habitat was considered to be “settled”, 58.6% was improved pasture, 10.3% was overgrown pasture, and 10.6% was crop fields. Use these results to determine, at the 5% level, if shrikes found in open areas use the sub-habitats in proportion to their availability. [Answer](#)
- 19.2** Between June 11 and 15, 1993, the Times Mirror Center for People and the Press interviewed 1006 adults concerning their views on media treatment of the then newly inaugurated President Clinton. They found 433 of those sampled felt that news organizations were “criticizing Clinton unfairly.” Test the hypothesis (with $\alpha = 0.10$) that more than 45% of all adults feel that Clinton has been criticized unfairly. [Answer](#)
- 19.3** A random selection of consumers present at the Mall of America were allowed to taste three types of cola (Pepsi, Coke, and a generic brand). After tasting each type (which were supplied to each person in a random order) the person was to select which cola they preferred. The results indicated that 57 people preferred Pepsi, 63 preferred Coke, and 34 preferred the generic brand. Is there evidence, at the 5% level, that these customers prefer one brand over the others? [Answer](#)
- 19.4** A particular type of corn is known to have one of four types of kernels: purple-smooth, purple-wrinkled, yellow-smooth, and yellow-wrinkled (see figure below). The purple (P) and smooth (S) alleles are dominant. The cross between heterozygous individuals (i.e., PpSs) should produce a 9:3:3:1 ratio of purple-smooth, purple-wrinkled, yellow-smooth, and yellow-wrinkled individuals. Of the kernels shown in the graphic below (a random picture location but not a random selection of each individual) 32 are purple-smooth, 14 are purple-wrinkled, 8 are yellow-smooth, and 4 are yellow-wrinkled. Use the results to determine, at the 5% level, if the theoretical 9:3:3:1 ratio is upheld with these data. [Answer](#)



19.2 Goodness-of-Fit Test in R

19.2.1 Data Format

A goodness-of-fit test is conducted in R with `chisq.test()` which requires an observed table as the first argument. This observed table is entered from previously summarized data using `c()`. However, raw data consisting of the recorded level for each individual must be summarized to a frequency table, and stored in an object, with `table()` as shown in Section ?? before being submitted to `chisq.test()`.

For example, suppose that the frequencies of shrike observations in the “open”, “mid-successional”, “scattered trees”, “woods”, and “wetland” habitats shown previously are known to be 1456, 43, 112, 44 and 6, respectively. These summarized values are entered directly into a named vector with

```
> ( obs <- c(Open=1456, MidSucc=43, ScatTree=112, Woods=6, Wetland=44) )
  Open  MidSucc  ScatTree    Woods  Wetland
  1456      43       112       6      44
```

However, instead of having summarized frequencies suppose that you have raw data in a variable called `hab.use` in a data frame called `shrike.raw` that looks like this,

```
[1] Open      Open      Open      Open      MidSucc  ScatTree Woods     Woods
Levels: MidSucc Open ScatTree Wetland Woods
```

These raw data must then be summarized into a table like this

```
> ( obs <- xtabs(~hab.use, data=shrike.raw) )
hab.use
  MidSucc      Open ScatTree  Wetland    Woods
  43        1456     112       6      44
```

Note that the two vectors/tables are identical with the exception of the ordering of the levels.

- ◊ If the raw un-summarized data are entered into a vector, then that vector must be summarized with `table()` and assigned to an object before performing the goodness-of-fit test.

19.2.2 Goodness-of-Fit Test

The goodness-of-fit test is computed with `chisq.test()` with a vector or table of observed frequencies in each level of the categorical variable as the first argument. The following arguments may also be used:

- `p=`: a vector of expected proportions for the levels of the theoretical distribution.
- `rescale.p=`: a logical indicating whether the values given in `p=` should be rescaled so that they sum to 1. This rescaling is useful if the proportions entered into `p=` were rounded or are actually the expected frequencies. Using `rescale.p=TRUE` will perform the rescaling.

- **correct=**: a logical indicating whether a so-called “continuity correction” should be used or not. Some authors argue that small chi-square tables with small sample sizes should be corrected for the fact that the chi-square distribution is a continuous distribution. This correction is applied by simply subtracting 0.5 from each observed-expected calculation. The default is to use the correction (=TRUE) which is fine but the results will not match your hand calculations if you do not use the correction. Use `correct=FALSE` to turn off the continuity correction.

◊ A goodness-of-fit test should include the `rescale.p=TRUE` argument in the `chisq.test()` function to correct for any rounding errors in the theoretical proportions.

◊ A chi-square test statistic can be corrected for “continuity” issues with the `correct=TRUE` argument to `chisq.test()`.

The results from `chisq.test()` should be assigned to an object because a variety of useful information can be extracted from this object. For example, suppose that the results of `chisq.test()` were saved into the `chi1` object. With this object, the observed frequencies are extracted with `chi1$observed`, the expected frequencies are extracted with `chi1$expected`, and a visualization of the p-value is obtained with `plot(chi1)`. Finally, confidence intervals for the proportion of individuals in each level are constructed by submitting the saved object to `gofCI` (e.g., `gofCI(chi1)`).

◊ The results from `chisq.test()` should be assigned to an object.

19.2.3 Example - Loggerhead Shrikes

The 11-steps (Section 15.1) for completing a full hypothesis test for the example below is shown below:

*Bohall-Wood (1987) constructed 24 random 16-km transects along roads in counties near Gainesville, FL. Two observers censused each transect once every 2 weeks from 18 October 1981 to 30 October 1982, by driving 32 km/h and scanning both sides of the road for perched and flying shrikes (**Lanius ludovicianus**). The habitat, whether the bird was on the roadside or actually in the habitat, and the perch type were recorded for each shrike observed. Habitats were grouped into five categories. The number of shrikes observed in each habitat was 1456 in open areas, 43 in midsuccessional, 112 in scattered trees, 44 in woods, and 6 in wetlands. Separate analyses were used to construct the proportion of habitat available in each of the five habitat types. These results were as follows: 0.358 open, 0.047 midsuccessional, 0.060 scattered trees, 0.531 woods, and 0.004 wetlands. Use these data to determine, at the 5% level, if shrikes are using the habitat in proportion to its availability.*

1. As stated, α should be set at 0.05.
2. The H_0 : “The distribution of habitat use by shrikes is the same as the proportions of available habitat” versus H_A : “The distribution of habitat use by shrikes is NOT the same as the proportions of available habitat.”
3. A goodness-of-fit test is required because a categorical variable (habitat use) with five levels from a

single population (shrikes in this are) was recorded and will be compared to a theoretical distribution in the null hypothesis.

- The data appear to be part of an observational study where the individuals were not randomly selected but the transects upon which they were observed were. As the summary data is given in the background it was entered into R with

```
> ( obs <- c(Open=1456, MidSucc=43, ScatTree=112, Woods=6, Wetland=44) )
  Open  MidSucc ScatTree    Woods  Wetland
  1456      43     112       6      44
```

The `chisq.test()` function was used at this point, before the assumptions have been assessed, so that R could be used to compute the expected frequencies. The expected proportions of available habitat were first entered into a vector with

```
> ( p.exp <- c(Open=0.358, MidSucc=0.047, ScatTree=0.060, Woods=0.531, Wetland=0.004) )
  Open  MidSucc ScatTree    Woods  Wetland
  0.358     0.047     0.060     0.531     0.004
```

The observed frequencies and expected proportions were then submitted to `chisq.test()` with the results saved to an object with

```
> shrike.chi <- chisq.test(obs, p=p.exp, rescale.p=TRUE)
```

Finally, a “table” of observed and expected frequencies was extracted with

```
> data.frame(obs=shrike.chi$observed, exp=shrike.chi$expected)
   obs      exp
Open 1456 594.638
MidSucc 43 78.067
ScatTree 112 99.660
Woods 6 881.991
Wetland 44 6.644
```

- The test statistic below should follow a χ^2 distribution because there are more than five individuals expected in each habitat level as shown in the output above.
- The appropriate statistic is the observed frequency table shown in the output above.
- The test statistic is $\chi^2=2345$ with 4 df as shown with

```
> shrike.chi
Chi-squared test for given probabilities with obs
X-squared = 2345.071, df = 4, p-value < 2.2e-16
```

- The p-value is $p < 0.00005$ as seen above.
- The H_0 is rejected because the $p - value < \alpha$.
- The shrikes do not appear to use habitats in the same proportions as the availability of the habitat.
- The 95% confidence intervals for the proportion of use in each habitat level is obtained with

```
> gofCI(shrike.chi,digits=3)
   p.obs p.LCI p.UCI p.exp
Open    0.877 0.860 0.892 0.358
MidSucc 0.026 0.019 0.035 0.047
ScatTree 0.067 0.056 0.081 0.060
Woods    0.004 0.002 0.008 0.531
Wetland   0.026 0.020 0.035 0.004
```

From these results it is apparent that the shrikes use the “open” habitat much more often and the “woods” habitat much less often than would be expected if they used all habitats in proportion to their availability.

19.2.4 Example - Modes of Fishing

The 11-steps (Section 15.1) for completing a full hypothesis test for the example below is shown below:

Herriges and King (1999) examined modes of fishing for a large number of recreational saltwater users in southern California. One of the questions asked in their Southern California Sportfishing Survey was what “mode” they used for fishing – “from the beach”, “from a fishing pier”, “on a private boat”, or “on a chartered boat.” The results to this question, along with other data not used here, are found in [FishingModes.csv](#). One hypothesis of interest states that two-thirds of the users will fish from a boat, split evenly between private and charter boats, while the other one-third will fish from land, also split even between those fishing on the beach and those from a pier. Use the data in the mode variable of the data file to determine if this hypothesis is supported at the 10% level.

1. As stated, α should be set at 0.10.
2. The H_0 : “The distribution will follow the proportions of $\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{6}$, and $\frac{1}{6}$ for private boat, charter boat, beach, and pier modes of fishing, respectively” versus H_A : “The distribution will NOT follow the proportions of $\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{6}$, and $\frac{1}{6}$ for private boat, charter boat, beach, and pier modes of fishing, respectively.” [These fractions were found with the following thought process – the two-thirds for “boat” fishing is split in half for one-third each for private and charter boats; the one-third, or two-sixths, for “land” fishing is split in half for one-sixth each for beach and pier fishing.]
3. A goodness-of-fit test is required because a categorical variable (mode) with four levels from a single population (Southern California Sportfishers) was recorded and will be compared to a theoretical distribution in the null hypothesis.
4. The data appear to be part of an observational study where the individuals were not obviously (probably were not) randomly selected. The raw data was read in with

```
> sf <- read.csv("data/FishingModes.csv")
```

The *mode* variable was summarized with

```
> ( obs <- xtabs(~mode,data=sf) )
mode
beach     boat  charter      pier
  134       418     452      178
```

The order of the levels is made note of here so that the expected proportions below can be entered in the same order,

```
> ( p.exp <- c(beach=1/6,boat=1/3,charter=1/3,pier=1/6) )
  beach      boat     charter      pier
0.1666667 0.3333333 0.3333333 0.1666667
```

The `chisq.test()` function is used at this point, before the assumptions have been assessed, so that R could be used to compute the expected frequencies,

```
> sf.chi <- chisq.test(obs,p=p.exp,rescale.p=TRUE)
```

Finally, a “table” of observed and expected frequencies was extracted,

```
> data.frame(obs=sf.chi$observed,exp=sf.chi$expected)
  obs.mode obs.Freq exp
beach      beach      134 197
boat       boat      418 394
charter   charter      452 394
pier       pier      178 197
```

5. The test statistic below should follow a χ^2 distribution because there are more than five individuals expected in each habitat level as shown in the output above.
6. The appropriate statistic is the observed frequency table shown in the output above.
7. The test statistic is $\chi^2=32$ with 3 df as shown with

```
> sf.chi
Chi-squared test for given probabilities with obs
X-squared = 31.9797, df = 3, p-value = 5.285e-07
```

8. The p-value is $p < 0.00005$ as seen above.
9. The H_0 is rejected because the $p - value < \alpha$.
10. The modes of fishing do not appear to match the distribution outlined in the null hypothesis.
11. The 95% confidence intervals for the proportion of use of each mode is obtained with

```
> gofCI(sf.chi,digits=3)
  p.obs p.LCI p.ICI p.exp
beach  0.113 0.097 0.133 0.167
boat   0.354 0.327 0.381 0.333
charter 0.382 0.355 0.410 0.333
pier   0.151 0.131 0.172 0.167
```

From these results it is apparent that the users use the beach slightly less than expected and use the charter boats slightly more than expected. The use of the pier and private boats are not different from what was expected.

19.2.5 Example - Mendelian Genetics II

The 11-steps (Section 15.1) for completing a full hypothesis test for the example below is shown below:

Geneticists hypothesized that three of every four progeny from a cross between two parent fruit-flies known to possess both a dominant and recessive allele would have red eyes. In a carefully controlled experiment, 82 of 151 randomly selected progeny had red-eyes. Test at the 1% level if the percentage of red-eyed progeny in the population of progeny is different than what the researchers hypothesized.

1. As stated, α should be set at 0.01.
2. The H_0 : “The proportion of progeny with red-eyes is 0.75” versus H_A : “The proportion of progeny with red-eyes is NOT 0.75.”
3. A goodness-of-fit test is required because a categorical variable (eye color) with two levels from a single population was recorded and will be compared to a theoretical distribution in the null hypothesis.
4. The data appear to be quasi-experimental in that a specific cross was made but there are very little controls. Selected progeny were randomly selected. As the summary data is given in the background it was entered into R with

```
> ( obs <- c(red=82,nonred=151-82) )
   red nonred
     82      69
```

The `chisq.test()` function was used at this point, before the assumptions have been assessed, so that R could be used to compute the expected frequencies. The expected proportions of available habitat were first entered into a vector with

```
> ( p.exp <- c(red=0.75,nonred=0.25) )
   red nonred
     0.75    0.25
```

The observed frequencies and expected proportions were then submitted to `chisq.test()` with the results saved to an object with

```
> m.chi <- chisq.test(obs,p=p.exp,rescale.p=TRUE)
```

Finally, a “table” of observed and expected frequencies was extracted with

```
> data.frame(obs=m.chi$observed,exp=m.chi$expected)
   obs     exp
red     82 113.25
nonred 69  37.75
```

5. The test statistic below should follow a χ^2 distribution because there are more than five individuals expected in each eye level as shown in the output above.
6. The appropriate statistic is the observed frequency table shown in the output above.
7. The test statistic is $\chi^2=34.49$ with 1 df as shown with

```
> m.chi
Chi-squared test for given probabilities with obs
X-squared = 34.4923, df = 1, p-value = 4.279e-09
```

8. The p-value is $p < 0.00005$ as seen above.
9. The H_0 is rejected because the $p - value < \alpha$.
10. The proportion of red-eyed progeny appears to be different than 0.75. Thus, the Mendelian theory is not supported by these results.
11. The 95% confidence intervals for the proportion of progeny in each eye level is obtained with

```
> gofCI(m.chi,digits=3)
      p.obs p.LCI p.UCI p.exp
red    0.543 0.464 0.620  0.75
nonred 0.457 0.380 0.536  0.25
```

From these results it is apparent that the proportion of progeny with red-eyes was between 0.464 and 0.620 indicating that there were many fewer red-eyed progeny than would be expected from the Mendelian theory.

Review Exercises

19.5  The leader of a local lake association conducted a survey of all members of the association. One question on the survey was, "What is your preferred method of receiving notices from the lake association: by regular mail, by e-mail, by phone, by poster (at the local boat landing), or other?" Of the surveys returned, 47 respondents preferred regular mail, 63 e-mail, 17 phone, 73 by poster, and 8 some other method. OF THE RESPONDENTS WHO DID NOT PREFER SOME OTHER METHOD, is there evidence, at the 5% level, of a difference in the preferred method of contact? [Answer](#)

19.6  Philcox *et al.* (1999) examined patterns in the road-related mortalities of otters (*Lutra lutra*) in Britain from 1971 to 1996. One aspect of their analysis was to examine the sex ratio of road-killed otters. The sex of all otters for which sex could be identified are recorded in *OtterMort.csv*. Use these data to determine if there is a significant (at the 1% level) bias in the sex ratio of road-killed otters. [Answer](#)

19.7  While imprisoned by the Germans during World War II, the English mathematician John Kerrich tossed a coin 10000 times and obtained 5067 heads. Use his results to determine (at the 1% level) whether the coin was fair or not (i.e., equal chance of heads and tails). [Answer](#)

19.8  Fisher claims that the randomization function of its "Studio-Standard" 60-disc CD changer is completely random. To test this assertion, the owner of one of these units randomly filled the CD changer with 20 copies of "The Best of Taj Mahal" and 40 copies of "Beethoven's Greatest." Each CD had 20 songs on it. The owner set out to test the randomness of the CD player by listening to 100 songs chosen by the CD changer. The owner recorded whether a song came from the Taj Mahal (T) CD or the Beethoven (B) CD. The data collected are listed below (organized into rows of 25 for convenience). Test, at the 5% level, the hypothesis that the randomization function on the CD changer is indeed random. [Answer](#)

```
T T B B B B T B T B T B B B T B T B B B B B B B B B B
T T T B B T B T T B T B B T B T B B B T T B T T B
T B B T B B B T B B B B T T B B B B B B B B T T B
B T T B B T B B T T B T B B T B T B B B B T B T B
```

- 19.9** Past data suggest that of the patients that a hospital serves 44% have type O, 45% have type A, 8% have type B, and 3% have type AB blood. In a more recent survey they found that 67 patients had type O, 83 had type A, 29 had type B, and 8 had type AB. Use the more recent results to determine, at the 5% level, if the past results still hold. [Answer](#)

- 19.10** A county district attorney would like to run for the office of state district attorney. She has decided that she will give up her county office and run for state office if more than 65% of her party constituents support her. As her campaign manager, you collected data on 950 randomly selected party members and find that 660 party members support the candidate. Test at the 5% significance level whether she should give up her county office and run for the state office. [Answer](#)

- 19.11** Suppose that you know that a population of deer is at a stable age distribution and stable population size. In addition, it is hypothesized that the survival rate from year-to-year is 50%. Through a random sample of animals from this population you determine that 134 are in the 0-1 age group, 66 are aged 1-2, 30 are aged 2-3, 13 are aged 3-4, 4 are aged 4-5, and 6 are aged 5-6. Use these results to determine, at the 10% level, if the survival rate is indeed 50%. [Hint: Find the expected number of animals in each age category. The expected number in the first age category, X , is found by solving the following equation $X + (0.5^1 + 0.5^2 + 0.5^3 + 0.5^4 + 0.5^5)X = n$ where n is the total number of observed animals. The expected values in the remaining categories are determined from the value of X and the hypothesized survival rate.]

[Answer](#)

- 19.12** Repeat Review Exercise 19.1 using R. [Answer](#)

- 19.13** Repeat Review Exercise 19.2 using R. [Answer](#)

- 19.14** Repeat Review Exercise 19.3 using R. [Answer](#)

- 19.15** Repeat Review Exercise 19.4 using R. [Answer](#)

- 19.16** An Alaskan pollock (*Theragra chalcogramma*) trawling boat will discontinue trawling in an area if the by-catch of king salmon (*Oncorhynchus tshawytscha*) caught in that area exceeds 10% of the catch. In a very large trawl catch the independent observer on the boat randomly sampled 1256 fish and found that 145 were king salmon. Is there evidence, at the 10% level, that the boat should discontinue trawling in that area? [Answer](#)

APPENDICES

REFERENCES

- Allanson, P. 1992. Farm size structure in England and Wales, 1939-1989. *Journal of Agricultural Economics* 43:137–148. [178](#)
- Allen, C. R., S. Demarais, and S. Lutz. 1997. Impact of red imported fire ant population reduction on white-tailed deer fawn recruitment. *Journal of Wildlife Management* 61:911–916. [112, 149](#)
- Andersen, R. and J. D. C. Linnell. 2000. Irriuptive potential in roe deer: Density-dependent effects on body mass and fertility. *Journal of Wildlife Management* 64:698–706. [95](#)
- Bath, A. J. and T. Buchanan. 1989. Attitudes of interest groups in Wyoming toward wolf restoration in Yellowstone National Park. *Wildlife Society Bulletin* 17:519–525. [253](#)
- Bluman, A. G. 2002. *Elementary Statistics: A Step by Step Approach*. 4th edition, McGraw-Hill Companies. [6](#)
- Bohall-Wood, P. 1987. Abundance, habitat use, and perch use of loggerhead shrikes in north-central Florida. *Wilson Bulletin* 99:82–86. [258, 260](#)
- Brylinsky, M. 2001. An evaluation of changes in the yellow perch (*Perca flavescens*) population of Grafton Lake, Kejimkujik National Park, after dam removal. Technical Report Publication No. 59, Acadia Centre for Estuarine Research. [113](#)
- Carroll, K. K. 1975. Experimental evidence of dietary factors and hormone-dependent cancers. *Cancer Research* 35:3374–3383. [149](#)
- Cheshire, W., S. Abashian, and J. Mann. 1994. Botulinum toxin in the treatment of myofascial pain syndrome. *Pain* 59:65–69. [182, 211](#)
- Dudgeon, D. 2000. Large-scale hydrological changes in tropical Asia: Prospects for riverine biodiversity. *BioScience* 50:793–806. [129](#)
- Fairley, C., S. Chen, A. Ugoni, S. Tabrizi, A. Forbes, and S. Garland. 1994. Human papillomavirus infection and its relationship to recent and distant sexual partners. *Obstetrics and Gynecology* 84:755–759. [249](#)
- Fiebach, N., C. Viscoli, and R. Horwitz. 1990. Differences between women and men in survival after myocardial infarction. Biology or methodology? *Journal of the American Medical Association* 263:1092–1096. [242](#)

REFERENCES

REFERENCES

- Friese, C. 2005. Nurse practice environments and outcomes: Implications for oncology nursing. *Oncology Nursing Forum* 32:765–772. [242](#)
- Ginnett, T. F. and E. L. Young. 2000. Stochastic recruitment in white-tailed deer along an environmental gradient. *Journal of Wildlife Management* 64:713–720. [143](#), [150](#)
- Hebblewhite, M. 2000. Wolf and Elk Predator-Prey Dynamics in Banff National Park. Master's thesis, University of Montana. [200](#)
- Herriges, J. and C. King. 1999. Nonlinear income effects in random utility models. *Review of Economics and Statistics* 81:62–72. [262](#)
- Janzen, F. J. and C. L. Morjan. 2002. Egg size, incubation temperature, and posthatching growth in painted turtles (*Chrysemys picta*). *Journal of Herpetology* 36:308–311. [178](#)
- Johnson, R. and P. Kuby. 2000. *Elementary Statistics*. 8th edition, Pacific Grove: Duxbury. [107](#), [108](#)
- Jones, L. M. and N. N. Foshay. 1984. Diffusion of responsibility in a nonemergency situation: Response to a greeting from a stranger. *Journal of Social Psychology* 123:155–159. [243](#)
- Le Boeuf, B. J., D. E. Crocker, D. P. Costa, S. B. Blackwell, P. M. Webb, and D. S. Houser. 2000. Foraging ecology of northern elephant seals. *Ecological Monographs* 70:353–382. [118](#)
- Letty, J., S. Marchandea, J. Clobert, and J. Aubineau. 2000. Improving translocation success: An experimental study of anti-stress treatment and release method for wild rabbits. *Animal Conservation* 3:211–219. [27](#)
- Lock, R. H. 1993. 1993 new car data. *Journal of Statistics Education* 1(1), online journal. [98](#)
- Machowiak, P. A., S. S. Wasserman, and M. M. Levine. 1992. A critical appraisal of 98.6 degrees F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich. *Journal of the American Medical Association* 268:1578–1580. [182](#), [210](#)
- Maniak, P. J., R. D. Lossing, and P. W. Sorensen. 2000. Injured Eurasian ruffe, *Gymnocephalus cernuus*, release an alarm pheromone that could be used to control their dispersal. *Journal of Great Lakes Research* 26:183–195. [13](#)
- Meliker, J. R., R. F. Maiob, M. A. Zimmerman, H. M. Kime, S. C. Smith, and M. L. Wilson. 2004. Spatial analysis of alcohol-related motor vehicle crash injuries in southeastern Michigan. *Accident Analysis and Prevention* 36:1129–1135. [249](#)
- Mladenoff, D. J., R. G. Haight, T. A. Sickley, and A. P. Wydeven. 1997. Causes and implications of species restoration in altered ecosystems: A spatial landscape project of wolf population recovery. *Bioscience* 47:21–31. [113](#)
- Moore, D. S. and G. P. McCabe. 1998. *Introduction to the Practice of Statistics*. 34d edition, W.H.Freeman & Co. [6](#)
- Nicholson, R. and J. Kim. 1997. The relationship of the El-Nino southern oscillation to African rainfall. *Journal of Climatology* 17:117–135. [123](#)
- Owens, R. W. and N. M. Pronin. 2000. Age and growth of pike (*Esox lucius*) in Chivyrkui Bay, Lake Baikal. *Journal of Great Lakes Research* 26:164–173. [13](#), [209](#)
- Peck, S. K. 1985. Effects of aggressive interaction on temperature selection by the crayfish, *Orconectes virilis*. *American Midland Naturalist* 114:159–167. [73](#)

REFERENCES

REFERENCES

- Philcox, C., A. Grogan, and D. MacDonald. 1999. Patterns of otter *Lutra lutra* road mortality in Britain. *Journal of Applied Ecology* 36:748–762. [265](#)
- Ratti, J. R. and E. O. Garton. 1994. Research and experimental design, chapter 1, pp. 1–23. T.A. Bookhout, ed. *Research and management techniques for wildlife and habitats*, The Wildlife Society, Bethesda, MD. [166](#)
- Renner, E. 1970. *Mathematisch-statistische Methoden in der praktischen Anwendung*. Parey: Hamburg, Germany. [178](#)
- Robbins, R. 1990. Signing an organ donor card: Psychological factors. *Death Studies* 14:219–229. [234](#)
- Saenz, D., R. Conner, C. Shackelford, and D. Rudolph. 1998. Pileated woodpecker damage to red-cockaded woodpecker cavity trees in eastern Texas. *Wilson Bulletin* 110:362–367. [248](#)
- Stanford, C. B. 1996. The hunting ecology of wild chimpanzees; implications for the behavioral ecology of Pliocene hominids. *American Anthropologist* 98:96–113. [139](#)
- Suit, P. F. and T. W. Bauer. 1990. Dna quantitation by image cytometry of touch preparations from fresh and frozen tissue. *American Journal of Clinical Pathology* 94:49–53. [150](#)
- Vega Rivera, J. H., W. J. McShea, J. H. Rappole, and J. H. Haas. 1998. Pattern and chronology of prebasic molt for the wood thrush and its relation to reproduction and migration departure. *Wilson Bulletin* 110:384–392. [142](#)
- Waller, D. M. and W. S. Alverson. 1997. The white-tailed deer: A keystone herbivore. *Wildlife Society Bulletin* 25:217–226. [104](#)
- Wang, Y. and D. Finch. 1997. Migration of willow flycatcher along the middle Rio Grande. *Wilson Bulletin* 109:253–268. [148](#)
- Weitz, R. 1979. Barriers to acceptance of genetic counseling among primary care physicians. *Social Biology* 26:189–97. [122](#)

INDEX

- 1-sample Z-Test, *see* Z-Test
1-sample t-Test, *see* t-Test
2-sample t-Test, *see* t-Test
68-95-99.7% Rule, 86
- Accuracy, 165
 α , 185, 186, 188
Alternative Hypothesis, *see* Hypothesis, Alternative
Association
 Definitions, 101
 Measure, 106
- Bar Chart, 78
 Construction, 79
- β , 188
- Boxplot
 Construction, 66
 Interpretation, 66
- Categorical Variable, 16
- Center, 55
- Central Limit Theorem
 Definition, 168
 Effect of n, 169
- Chi-square
 Expected Table, 239
 Goodness-of-Fit Test, 252, 255
 Hypothesis Test, 240
- Coefficient of Determination, *see* r^2
- Column Proportions Table, *see* Table, Proportion
Confidence
 Bounds, 197
 Common Misinterpretations, 195
 Concept, 193, 200
 Effect of C, 202
 Effect of n, 202
- Intervals, 196
 Making narrower, 201
- Continuous Variable, 16
- Convenience Sample, 29
- Correlation
 Characteristics, 108
 Computation, 105, 108
 Estimation, 107
 Interpretation, 106
 Matrix, 108
- Direction
 Definitions, 101
- Discrete Variable, 16
- Dispersion, 61
- Distribution, Distinguishing, 160
- EDA
 Bivariate
 Categorical, 116
 Quantitative, 98
 Univariate
 Categorical, 76
 Quantitative, 48, 72, 81
- Experiment
 Definition, 20
 Multi-Factor, 22
 Principles, 25
 Single-Factor, 20
- Explanatory Variable
 Bivariate EDA, 98
 Regression, 128
- Factor
 Experiment, 20
 Variable in R, 77

- Five Number Summary, 66
- Form
- Definition, 102
- Forward calculation, *see* Normal Distribution, Finding areas
- Frequency Table, *see* Table, Frequency
- Goodness-of-Fit Test, *see* Chi-square
- Histogram
- Construction, 49
 - Interpretation, 52, 54
 - Multiple, 70
- Homoscedasticity, 136
- Hypothesis
- Alternative, 181
 - Null, 181
 - Research, 180
- Hypothesis Testing
- Concept, 182, 186
 - Errors, 188
 - Steps, 207
- Individual, Definition, 9
- Inference
- Definition, 9, 20, 157
- Interaction effect, 23
- Intercept
- Calculation, 135
 - Definition, 130
- IQR
- Calculation, 63, 65
 - When to use, 63, 66
- Level
- Controlling in R, 77
 - Experimental, 21
- Levene's Test, 224, 230
- Line
- Finding best-fit, *see* Regression
 - General Equation, 129
- Margin-of-Error, 197
- Matched-Pairs t-Test, *see* t-Test
- Mean
- Calculation, 57
 - Compared to median, 58
 - How measures center, 58
 - Inference, *see* Z-test and t-Test
 - Population Symbol, 57
 - Population symbol, 84
 - Sample Symbol, 57
- Sensitivity to outliers, 59
- When to use, 60
- Median
- Calculation, 56, 57
 - Compared to mean, 58
 - How measures center, 58
 - Sensitivity to outliers, 59
 - When to use, 60
- Mode, 56
- Natural Variability
- Definition, 5
 - Measure, 160
- Nominal Variable, 16
- Normal Distribution
- 68-95-99.7% Rule, 86
 - Characteristics, 84
 - Finding areas, 94
 - Finding values, 94
 - Symbol, 85
- Null Hypothesis, *see* Hypothesis, Null
- Observational Study, 28
- One-sample t-Test, *see* t-Test
- One-sample Z-Test, *see* Z-Test
- Ordinal Variable, 16
- Outlier, 54, 103
- p-value, 183, 186
- Parameter
- Definition, 10
- Percentage Table, *see* Table
- Population, 10
- Population Distribution
- Definition, 160
 - Normal Distribution, 84
- Power, 188
- Precision, 165, 201
- Predictions
- Regression, 131, 145
- Probability, 173
- Proportions Table, *see* Table, Proportion
- Proportions, Inference, *see* Chi-square
- Quantitative Variable, 15
- Quartile, 62
- Calculation, 65
- r^2 , 136
- Range, 62
- Calculation, 65
- Regression

- Assumptions, 136
- Finding best-fit line, 134, 145
- Predictions, 131
- Purposes, 128
- Rejection Criterion, *see* α
- Replicates, 21, 24
- Residual
 - Computation, 133
 - Definition, 133
- Response Variable
 - Bivariate EDA, 98
 - Experiment, 20
 - Regression, 128
- Reverse calculation, *see* Normal Distribution, Finding values
- Row Proportions Table, *see* Table, Proportion
- RSS, 135
- Sample
 - Definition, 10
- Sample Distribution
 - Definition, 160
 - Histogram, 49
- Sample Size
 - Estimation, 201
- Sampling Distribution
 - Center, 158
 - Definition, 157, 160
 - Dispersion, 157, 159
 - Shape, 158
 - Simulation, 162
- Sampling Variability
 - Definition, 5, 14
 - Measure, 157, 160
- Scatterplot
 - Construction, 98
 - Interpretation, 101
- Scientific Method, 180
- Shape, 52
- Simple Linear Regression, *see* Regression
- Simple Random Sample, 29
- Skewed, 52
- Slope
 - Calculation, 135
 - Definition, 130
- SLR, *see* Regression
- Standard Deviation
 - Calculation, 63, 65
 - Characteristics, 64
 - Interpretation, 63, 64
 - Measure of, 160
- Population symbol, 84
- Sample symbol, 63
- When to use, 66
- Standard Error
 - Definition, 157
 - Effect of n, 169
 - Measure of, 160, 165
- Standard Normal Distribution, *see* Normal Distribution
- Standardization, *see* Normal Distribution, Converting to Z-scale
- Statistic
 - Definition, 11
- Statistics, Field of
 - Definition, 6
- Strength
 - Definition, 103
 - Measure, 105
- Symmetric, 52
- t Distribution
 - Characteristics, 214
- t Test
 - 1-sample, 216
 - 2-sample, 223, 230
- Table
 - Frequency, 76, 116, 117, 123
 - Percentage, 77, 118, 121
 - Proportion, 118, 121
- Table Proportions Table, *see* Table, Proportion
- Treatment, Experimental, 21, 22
- Two-sample t-Test, *see* t-Test
- Two-way Table, *see* Table, Frequency
- Type I Error, *see* Hypothesis Testing, Errors
- Type II Error, *see* Hypothesis Testing, Errors
- Unbiased, 158, 165
- Variability
 - Natural, *see* Natural Variability
 - Sampling, *see* Sampling Variability
- Variable
 - Definition, 10
 - Types, *see* Quantitative, Continuous, Discrete, Categorical, Nominal, or Ordinal
- Variance
 - Calculation, 63
 - Pooled, 223
 - Testing Equality, *see* Levene's Test
- Voluntary Response Sample, 29
- Y-intercept, *see* Intercept

Z-Distribution, *see* Normal Distribution, Standard

Normal

Z-Test, [207](#)