MODULE 1

WHY STATISTICS IS IMPORTANT

Contents

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 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, 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]

The WDNR wants to study elements in the sediment (among other things) in the entire 3000 m^2 area shaded in Figure 1.1. Is it physically possible to examine every square meter of that area? Is it prudent, ecologically and economically, to examine every square meter of this area? The answer, of course, is "no." How then will the WDNR be able to make conclusions about this entire area if they cannot reasonably examine the whole area? The most reasonable solution is to sample a subset of the area and use the results from this sample to make inferences about the entire area.

Methods for properly selecting a sample that fairly represents a larger collection of individuals are an im-

[1] More information at the EPA and the WDNR websites.

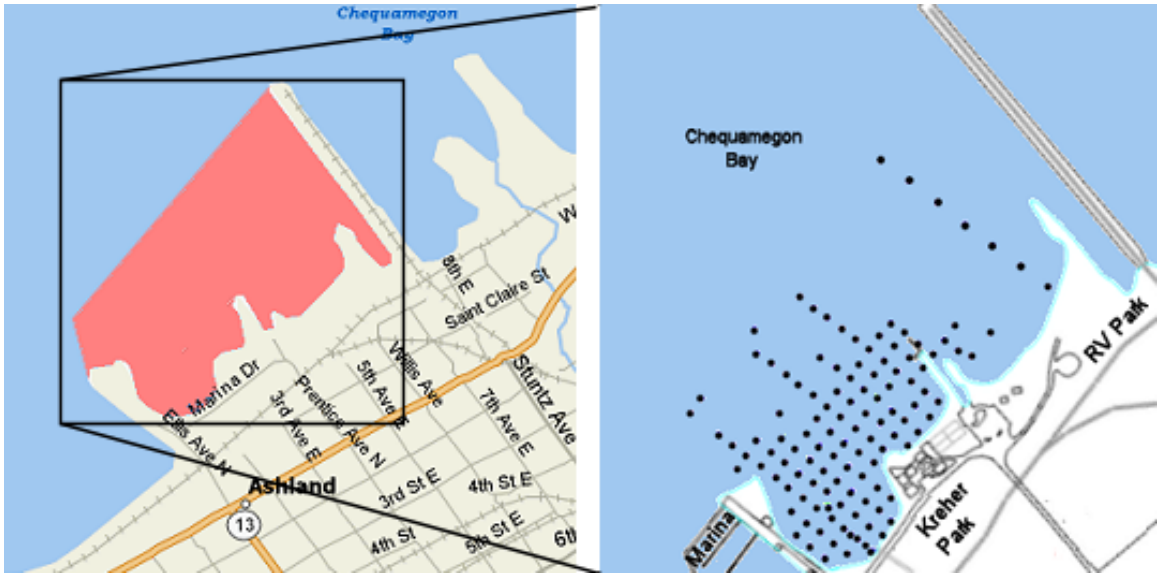


Figure 1.1. Location of the Ashland superfund site (left) with the location of 119 historical sediment sampling sites (right).

portant area of study in statistics. For example, the WDNR would not want to sample areas that are only conveniently near shore because this will likely not be an accurate representation of the entire area. In this example, it appears that the WDNR used a grid to assure a relatively even dispersal of samples throughout the study area (Figure 1.1). Methods for choosing the number of individuals to select and how to select those individuals are discussed in Module 3.

Suppose that the WDNR measured the concentration of lead at each of the 119 locations shown in Figure 1.1. Further suppose that they presented their results at a public meeting by simply showing the list of lead concentration measurements (Table 1.1).[2] Is it easy to make conclusions about what these data mean from this type of presentation?

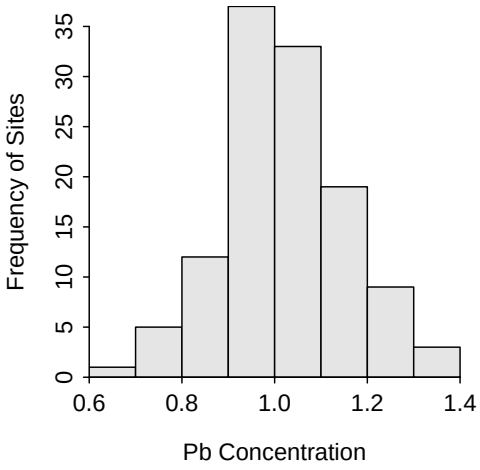
Table 1.1. Lead concentration ($\mu g \cdot m^{-3}$) from 119 sites in Kreher Park superfund site.

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

Instead, suppose that the scientists brought a simple plot of the frequency of observed lead concentrations and brief numerical summaries (Figure 1.2) to the meeting. With these one can easily see that the measurements were fairly symmetric with no obviously "weird" values. The lead concentrations ranged from as low as 0.67 $\mu g \cdot m^{-3}$ to as high as 1.36 $\mu g \cdot m^{-3}$ with the measurements centered on approximately 1.0 $\mu g \cdot m^{-3}$. These summaries are discussed in detail in Module 6. However, at this point, note that summarizing large quantities of data with few graphical or numerical summaries makes it is easier to identify meaning from

[2] These are hypothetical data for this site.

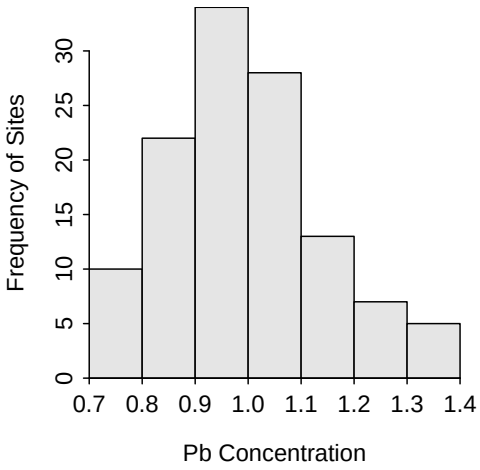
data.



	Pb
Min.	0.67
1st Qu.	0.92
Median	1.01
Mean	1.02
3rd Qu.	1.10
Max.	1.36

Figure 1.2. Histogram and summary statistics of lead concentration measurements ($\mu g \cdot m^{-3}$) 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 the results 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 larger whole will not perfectly represent the large whole. This will lead to some uncertainty in our summaries of the larger whole.



	Pb
Min.	0.71
1st Qu.	0.89
Median	0.98
Mean	0.99
3rd Qu.	1.08
Max.	1.39

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

The results from two different samples do not perfectly agree because each sample contains different individuals (sites in this example), and no two individuals are exactly alike. 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. Natural and sampling variability will be revisited continuously throughout this course.

This may be 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 individuals from the larger whole did not perfectly represent the larger whole. Furthermore, each sample is unique and will likely lead to a (slightly) 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 Major Goals of Statistics

As seen in the Kreher Park example, the field of statistics has two primary purposes. First, statistics provides methods to summarize large quantities of data into concise and informative numerical or graphical summaries. For example, it was 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 was 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).[3]

1.3 Definition of Statistics

Statistics is the science of collecting, organizing, and interpreting numerical information or data (Moore and McCabe 1998). People study statistics for a variety of reasons, including (Bluman 2000):

1. To understand the statistical studies performed in their field (i.e., be knowledgeable about the vocabulary, symbols, concepts, and statistical procedures used in those studies).
2. To conduct research in their field (i.e., be able to design experiments and samples; collect, organize, analyze, and summarize data; make reliable predictions or forecasts for future use; and communicate statistical results).
3. To be better consumers of statistical information.

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

[3] Population and sample are defined more completely in Section 2.1.

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.

1.4 Why Does Statistics (as a tool) Exist?

Besides demonstrating the two major goals of statistics, the Kreher Park example illustrates three "realities" that exist in nature and life that necessitate the need for statistics as tool for understanding. First, in most realistic situations it is not possible or, at least, not reasonable to "see" the entire population. For example, it was not reasonable to sample the sediments throughout the entire contaminated area near Kreher Park. In other examples, is it possible (or reasonable) to examine every Northern Short-Tailed Shew (*Blarina bevicauda*) in Great Lakes states, every person of legal voting age in Wisconsin, or every click on Facebook? Second, as described above, variability exists, both among individuals and results of samples. Third, because we must take samples from populations and those samples are both imperfect representations of the population and sampling variability exists, our conclusions about the population are uncertain. For example, the first sample in the Kreher Park example suggested that the mean lead concentration was 1.02 $\mu g \cdot m^{-3}$, whereas the second sample was 0.98 $\mu g \cdot m^{-3}$. You have also seen this concept when the margin-of-error in poll results are presented. In summary, statistics exist because we must sample instead of observe entire populations, variability is ever present, and the conclusions from samples are uncertain.

MODULE 2

FOUNDATIONAL DEFINITIONS

Contents

STATISTICAL INFERENCE IS THE PROCESS of forming conclusions about a parameter of a population from statistics computed from individuals in a sample.[1] Thus, understanding statistical inference requires understanding the difference between a population and a sample and a parameter and a statistic. And, to properly describe those items, the individual and variable(s) of interest must be identified. Understanding and identifying these six items is the focus of this module.

The following hypothetical example is used throughout this module. Assume that we are interested in the average length of 1015 fish in Square Lake. To illustrate important concepts in this module, assume that all information for all 1015 fish in this lake is known (Figure 2.1). In "real life" this complete information would not be known.

2.1 Definitions

The **individual** in a statistical analysis is one of the "items" examined by the researcher. Sometimes the individual is a person, but it may be an animal, a piece of wood, a location, a particular time, or an event. It is extremely important that you don't always visualize a person when considering an individual in a statistical sense. 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). An individual in the Square Lake example is a fish, because the researcher will collect a set of fish and examine each individual fish.

[1] Formal methods of inference are discussed beginning with Module 13.

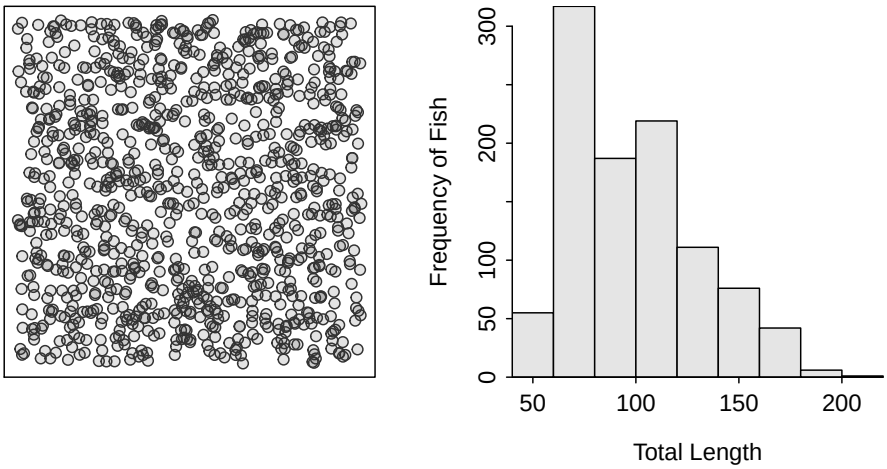


Figure 2.1. Schematic representation of individual fish (i.e., dots; **Left**) and histogram (**Right**) of the total length of the 1015 fish in Square Lake.

The **variable** is the characteristic recorded about each individual. The variable in the Square Lake example is the length of each fish. In most studies, the researcher will record more than one variable. For example, the researcher may also record the fish's weight, sex, age, time of capture, and location of capture. In this module, only one variable is considered. In other modules, two variables will be considered.

A **population** is ALL individuals of interest. In the Square Lake example, the population is all 1015 fish in the lake. The population should be defined as thoroughly as possible including qualifiers, especially those related to time and space, as necessary. This example is simple because Square Lake is so well defined; however, as you will see in the review exercises, the population is often only well-defined by your choice of descriptors.

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." For example, the population average length of all 1015 fish in Square Lake is 98.06 mm and the population standard deviation is 31.49 mm (Table 2.1).[2] 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 usually be "seen."

Table 2.1. 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

The entire population cannot be "seen" in real life. Thus, to learn something about the population, a subset of the population is usually examined. This subset is called a **sample**. The red dots in Figure 2.2 represent a random sample of n=50 fish from Square Lake (note that the sample size is usually denoted by n).
Summaries computed from individuals in a sample are called **statistics**. Specific names of statistics are preceded by "sample." The statistic of interest is always the same as the parameter of interest; i.e., the statistic describes the sample in 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 the sample from Square Lake are shown in Table 2.2 and Figure 2.2. The

[2] We will discuss how to compute and interpret each of these values in later modules.

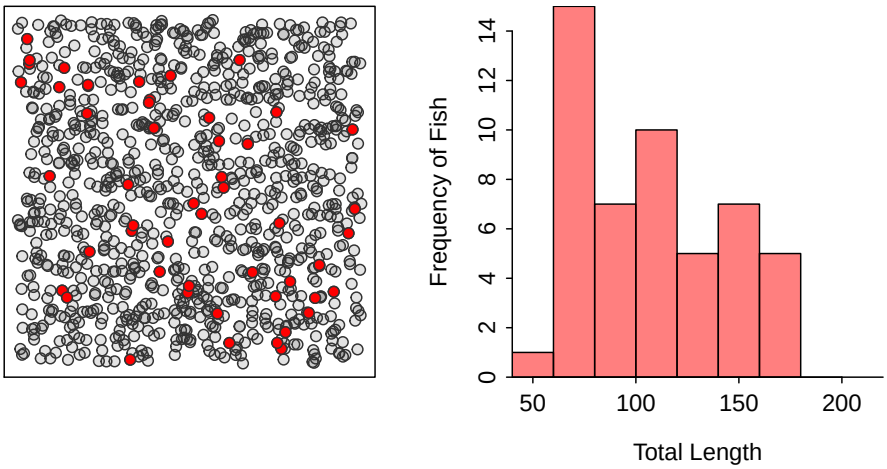


Figure 2.2. Schematic representation (**Left**) of a sample of 50 fish (i.e., red dots) from Square Lake and histogram (**Right**) of the total length of the 50 fish in this sample.

sample mean of 107.5 mm is the best "guess" at the population mean. Not surprisingly from the discussion in Module 1, 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	107.50	34.26	57	77	108	135	171

◇ **An individual is not necessarily a person.**

◇ **Populations and parameters can generally not be "seen."**

2.2 Performing an IVPPSS

In each statistical analysis it is important that you determine the Individual, Variable, Population, Parameter, Sample, and Statistic (**IVPPSS**). First, determine what items you are actually going to look at; those are the individuals. Second, determine what is recorded about each individual; that is the variable. Third, ALL individuals is the population. Fourth, the summary (e.g., mean or proportion) of the variable recorded from ALL individuals in the population is the parameter.[3] Fifth, the population usually cannot be seen, so only a few individuals are examined; those few individuals are the sample. Finally, the summary of the individuals in the sample is the statistic.

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 at from another perspective. A sample is an estimate of the population and a statistic is an estimate of a parameter. Thus, the statistic has to be the same summary (mean or proportion) of the sample as the parameter is of the population.

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

The IVPPSS process is illustrated 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 eastern Maine in 2007. 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.*

- An individual is a rabbit patch in eastern Maine in 2007 (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 eastern Maine in 2007.
- The parameter is the proportion of ALL rabbit patches in eastern Maine in 2007 that showed evidence for New England cottontails.[4]
- The sample is the 70 rabbit patches from eastern Maine in 2007 that were actually examined by the researcher.
- The statistic is the proportion of the 70 rabbit patches from eastern Maine in 2007 actually examined that showed evidence for New England cottontails. [In this case, the statistic would be 53/70 or 0.757.]

In the descriptions above, take note that the individual is very carefully defined (including stating a specific time (2007) and place (eastern Maine)), the population and parameter both use the word "ALL", the sample and statistic both use the specific sample size (70 rabbits), and that the parameter and statistics both use the same summary (i.e., proportion of patches that showed evidence of New England cottontails).

In some situations it may be easier to identify the sample first. From this, and realizing 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 Touristry 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 and 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 at Hawk Ridge.
- The population is ALL years at Hawk Ridge (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 at Hawk Ridge.
- The statistic is the average number of raptors seen in the 1971-2003 sample of years at Hawk Ridge.

Again, note that the individual is very carefully defined (including stating a specific time and place), the population and parameter both use the word "ALL", the sample and statistic both use the specific sample size (32 years), and that the parameter and statistics both use the same summary (i.e., average number of raptors).

[4] Note that this population and parameter cannot actually be calculated but it is what the researcher wants to know.
[5] Information about Hawk Ridge is found here.

◇ **An individual is usually defined by a specific time and place.**

◇ **Descriptions for population and parameter will always include the word "All."**

◇ **Descriptions for sample and statistic will contain the specific sample size.**

◇ **Descriptions for parameter and statistic will contain the same summary (usually average/mean or proportion/percentage). Howeve the summary is for a different set of individuals – the population for the parameter and the sample for the statistic.**

2.2.1 Sampling Variability (Revisited)

It is instructive to once again (see Module 1) consider how statistics differ among samples. Table 2.3 and Figure 2.3 show results from three more samples of n=50 fish from the Square Lake population. The means from all four samples (including the sample in Table 2.2 and Figure 2.2) were quite different from the known population mean of 98.06 mm. Similarly, all four histograms were similar in appearance but slightly different in actual values. These results illustrate that a statistic (or sample) will only approximate the parameter (or population) and that statistics vary among samples. This **sampling variability** is one of the most important concepts in statistics and is discussed in great detail beginning in Module 12.

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	100.48	31.87	45	78	100	120	180
50	99.40	38.28	47	69	90	114	203
50	98.14	32.26	45	71	87	122	174

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

◇ **Parameters are fixed in value, while statistics vary in value.**

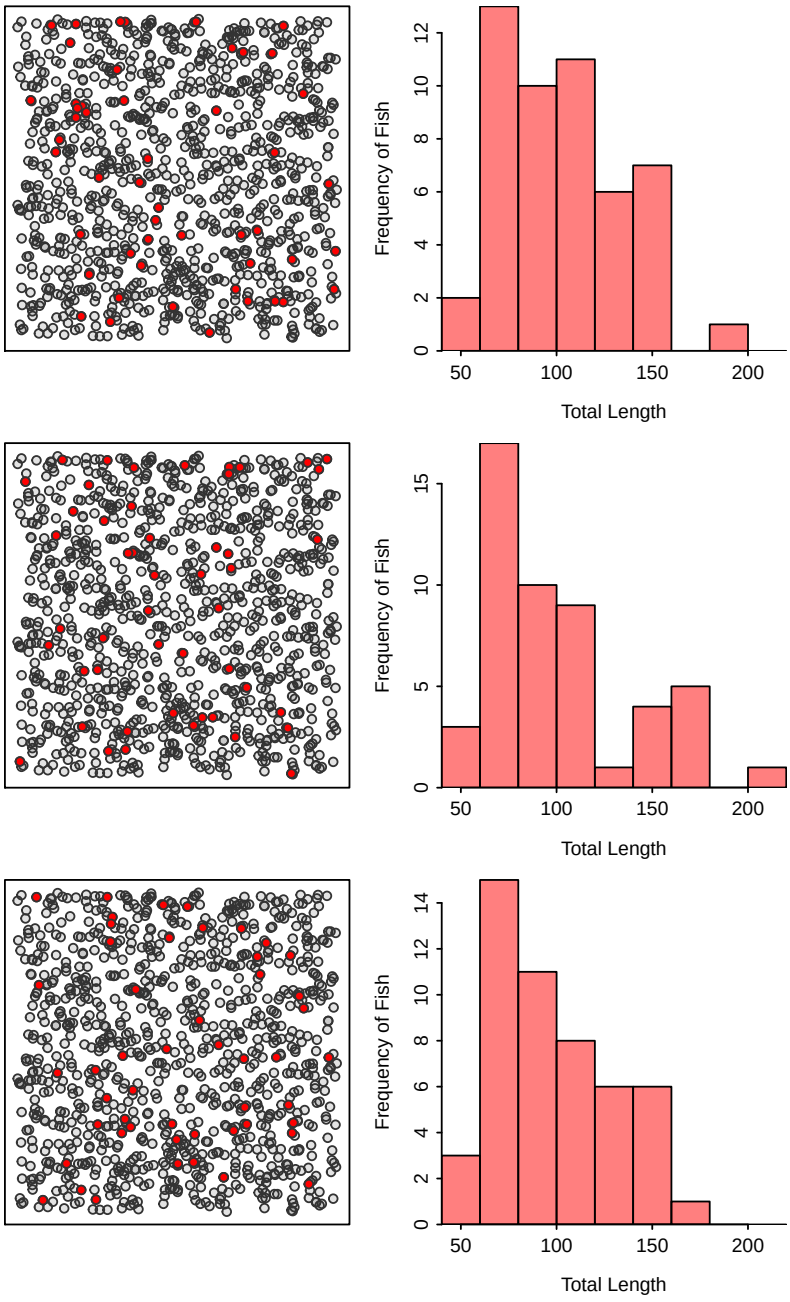


Figure 2.3. Schematic representation (**Left**) of three samples of 50 fish (i.e., red dots) from Square Lake and histograms (**Right**) of the total length of the 50 fish in each sample.

2.3 Variable Types

The type of statistic that can be calculated is dictated by the type of variable recorded. For example, an average can only be calculated for quantitative variables (defined below). Thus, the type of variable should be identified immediately after performing an IVPPSS.

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 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).

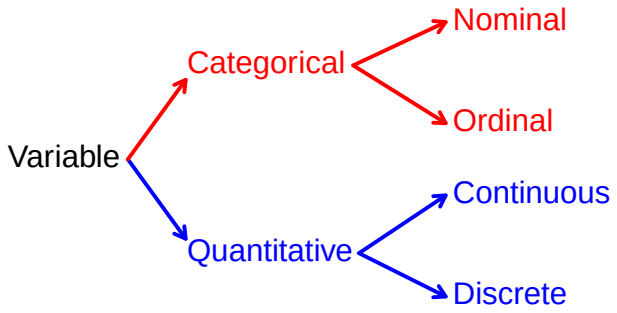


Figure 2.4. Schematic representation of the four types of variables.

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 for 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.

◇ **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 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. For example, 1="Poor", 3="Fair", and 5="Good". 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. It does not matter how precisely quantiative variables are recorded when deciding if the variable is continuous or discrete. For example, the weight of the duck might have been recorded to the nearest kg. However, this was just a choice that was made, the actual values can be continuously finer than kg and, thus, weight is a continuous variable.
6. 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."
7. 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.**

◇ **The precision to which a quantitative variable was recorded does not determine whether it is continuous or discrete. How precisely the variable COULD have been recorded is the important consideration.**

MODULE 3

DATA PRODUCTION

Contents

STATISTICAL INFERENCE IS THE PROCESS of making conclusions about a population from the results of a single sample. To make conclusions about the larger population, the sample must fairly represent the larger population. Thus, the proper collection (or production) of data is critical to statistics (and science in general). In this module, two ways of producing data – (1) Experiments and (2) Observational Studies – are described.

◇ **Inferences cannot be made if data are not properly collected.**

3.1 Experiments

An experiment deliberately imposes a *condition* on individuals to observe the effect on the **response variable**. In a properly designed experiment, all variables that are not of interest are held constant, whereas the variable(s) that is (are) of interest are changed among treatments. As long as the experiment is designed properly (see below), differences among treatments are either due to the variable(s) that were deliberately changed or randomness (chance). Methods to determine if differences were likely due to randomness are developed in later modules. Because we can determine if differences most likely occurred o randomness or changes in the variales, strong *cause-and-effect conclusions* can be made from data collected from carefully designed experiments.

3.1.1 Single-factor Experiments

A **factor** is a variable that is deliberately manipulated to determine its effect on the response variable. A factor is sometimes 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.

For example, suppose that a researcher wants to examine the effect of temperature on the total number of bacterial cells after two weeks. They have inoculated 120 agars[1] with the bacteria and placed them in a chamber where all environmental conditions (e.g., temperature, humidity, light) are controlled exactly. The researchers will use only two temperatures in this simple experiment – 10^oC and 15^oC. All other variables are maintained at constant levels. Thus, temperature is the only factor in this simple experiement because it is the only variable manipulated to different values to determine its impact on the number of bacterial cells.

◇ **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^oC and 15^oC. **Treatments** are the number of unique conditions that individuals in the experiment are exposed to. In a single-factor experiment, 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^oC and 15^oC. Treatments are 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, a replicate is an inoculated agar. The number of replicates is 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 example, the number of replicates is 60 inoculated agars.

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 treatment. 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.[2] Any difference in mean number of bacterial cells will be due to either different temperature treatments or randomness, because all other variables were the same between the two treatments.

◇ **Differences among treatments are either caused by randomness (chance) or the factor.**

The single factor is not restricted to just two levels. For example, more than two temperatures, say 10^oC, 12.5^oC, 15^oC, and 17.5^oC, could have been tested. With this modification, there is still only one factor – temperature – but there are now four levels (and only four treatments).

3.1.2 Multi-factor Experiments – Design and Definitions

More than one factor can be tested in an experiment. In fact, it is more efficient to have a properly designed experiment where more than one factor is varied at a time than it is to use separate experiments in which

[1] An agar, in this case, is a petri dish with a growth medium for the bacteria.
[2] Methods for making this comparison are in Module 19.

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 also examine the effect of relative humidity on the number of bacteria cells. This modified experiment has two factors – temperature (with two levels of 10^oC or 15^oC) and relative humidity (with four levels of 20%, 40%, 60%, and 80%). The number of treatments, or combinations of all factors, in this experiment is found by multiplying the levels of all factors (i.e., $2\times4=8$ in this case). The number of replicates in this experiment is now 15 (i.e., total number of available agars divided by the number of treatments; 120/8).

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

A drawing of the experimental design can be instructive (below). The drawing is a grid where the levels of one factor are the rows and the levels of the other factor are the columns. 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^oC	1	2	3	4
15^oC	5	6	7	8

3.1.3 Multi-factor Experiments – Benefits

The analysis of a multi-factor experimental design is more involved than what will be shown in this course. However, multi-factor experiments have many benefits, which can be illustrated by comparing a multi-factor experiment to separate single-factor experiments. For example, in addition to the two factor experiment in the previous section, consider separate single-factor experiments to determine the effect of each factor separately (further assume that individuals (i.e., agars) can be used in only one of these separate experiments).

To conduct the two separate experiments, randomly split the 120 available agars into two equally-sized groups of 60. The first 60 will be split into two groups of 30 for the first experiment with two temperatures. The second 60 will be split into four groups of 15 for the second experiment with four relative humidities. 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	
10^oC	15^oC
30	30

Relative Humidity			
20%	40%	60%	80%
15	15	15	15

The tabel below was modified from the previous section to show the number of replicates in each treatment of the experiment where both factors were simultaneously manipulated.

	Relative Humidity			
	20%	40%	60%	80%
10^oC	15	15	15	15
15^oC	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 were 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 were 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.[3]

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.[4] The effect of relative humidity is to increase the growth rate for those individuals at 10^oC (black line) but to decrease the growth rate for those individuals at 15^oC (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.

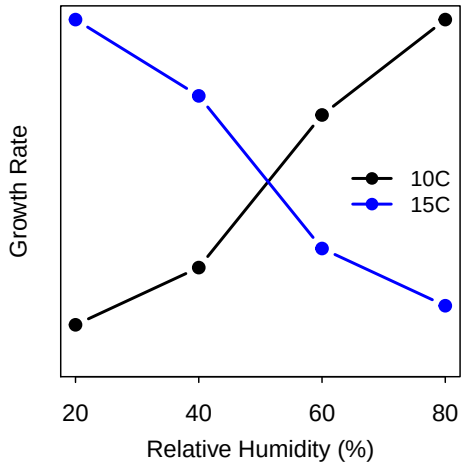


Figure 3.1. Mean growth rates in a two-factor experiment that depict an interaction effect.

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.

3.1.4 Allocating Individuals

Individuals[5] should be randomly allocated (i.e., placed into) to treatments. Randomization will tend to even out differences among groups for variables not considered in the experiment. In other words, randomization

[3] The real importance of this advantage will become apparent when statistical power is introduced in Module 14.
[4] The means of each treatment are plotted and connected with lines in this plot.
[5] When discussing experiments, an "individual" is often referred to as a "replicate" or an "experimental unit."

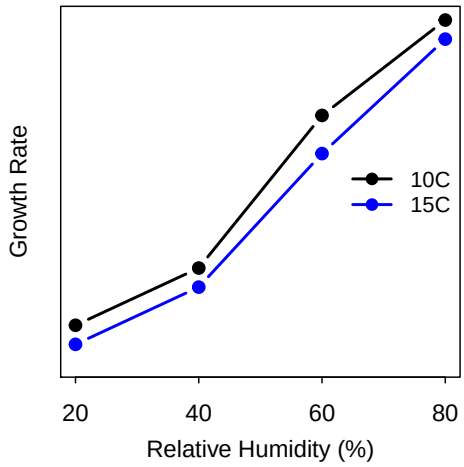


Figure 3.2. Mean growth rates in a two-factor experiment that depict no interaction effect.

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 "10" and 60 marked with "15" could be placed into a hat. One piece of paper would be drawn for each agar and the agar would 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. Agars corresponding to the first 60 numbers drawn from the hat could then be placed into the first treatment. Agars for the next (or remaining) 60 numbers would be placed in the second treatment. This process is essentially the same as randomly ordering 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 1 through 120 is accomplished with

```
> sample(120)
```

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

Thus, the first five (of 60) agars in the 10^oC treatment are 79, 78, 14, 63, and 41. The first five (of 60) agars in the 15^oC treatment are 87, 70, 13, 111, and 89.

In the modified experiment with two factors – temperature and relative humidity – with eight treatments containing 15 agars each, it is more efficient to save the random numbers into an object and then select the numbers in the first 15 positions, then the second 15 positions, etc. Positions are selected from an object by putting the position numbers in square brackets following the object name. Additionally, a colon is used to make a sequence of integers from the number before to the number after the colon.[6]

[6]For example, `1:4` will make an object with the numbers 1, 2, 3, and 4 in it.

```
> ragars2 <- sample(120)
> ragars2[1:15]     # "grab" the first 15 numbers
 [1]  44  34  90  27 100  74  10 102 106   6   9 118  67  59  80
> ragars2[16:30]    # "grab" the second 15 numbers, and so on
 [1]  42  97  46  21  41  53  66  99 107  35  25  56  29  68  79
```

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.[7]

	Relative Humidity			
	20%	40%	60%	80%
10^oC	44,34,$\cdots$	42,97,$\cdots$	33,15,$\cdots$	1,47,$\cdots$
15^oC	39,17,$\cdots$	37,55,$\cdots$	23,24,$\cdots$	77,112,$\cdots$

◇ **Individuals should be randomly allocated to treatments to remove bias.**

3.1.5 Design Principles

There are many other methods of designing experiments and allocating individuals that are beyond the scope of this book.[8] However, all experimental designs contain the following three basic principles.

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

Proper control in an experiment allows for strong cause-and-effect conclusions to be made (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 foreseen or unforeseen variable). Randomly allocating individuals to treatments removes any bias that may be included in the experiment. For example, if we do not 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 as possible at the start of the experiment.

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

[7] Only the first two numbers are shown because of space constraints.
[8] Other common designs include blocked, Latin square, and nested designs.

population. Observational studies cannot be used to make cause-and-effect statements because all variables that may impact 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. For many years scientists thought that the brains of females weighed less than the brains 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 is 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 considered.

◇ Strong cause-and-effect statements CANNOT be made from observational studies.

In observational studies, it is important to understand to which population inferences will refer.[9] To make useful inferences from a sample, the sample must be an unbiased representation of the population. In other words, it must not systematically favor certain individuals or outcomes.

For example, consider that you want to determine the mean length of all fish in a particular lake (e.g., Square Lake from Module 2). 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. 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 of interest. Thus it is important to select a sample from the specified population.

◇ It is important to understand the population before considering how to take a sample.

3.2.1 Types of Sampling Designs

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

A **voluntary response** sample consists of individuals that have chosen 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 consists of individuals who are easiest to reach for the researcher. An example of a convenience sample is when a researcher queries 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 of (i) one major or another, (ii) one or a few "years-in-school" (e.g., Freshman or Sophomores), or (iii) both.

In probability-based sampling, each individual of the population has a known chance of being selected for

[9] Thus, it is very important to first perform an IVPPS as discussed in Module 2.

the sample. The simplest probability-based sample is the **Simple Random Sample** (SRS) where each individual has the same chance of being selected. Proper selection of an SRS requires each individual to be assigned a unique number. The SRS is then formed by choosing random numbers and collecting the individuals that correspond to those numbers.

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 were 1112 transactions during the previous month (i.e., the population). The auditor would then number each transaction from 1 to 1112, 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.

Random numbers are selected in R by including the population size as the first and sample size as the second argument to `sample()`. For example, 30 numbers from between 1 and 1112 is selected with

```
> sample(1112,30)
 [1]  106 1055  578  830 1066  900  525  103  324  795 1057  480  773  649  129  393
[17]  215  982  628 1047  994  200  863  740  574   66  346 1003  941  309
```

Thus, accounts 106, 1055, 578, 830, and 1066 would be the first five (of 30) selected.

There are other more complex types of probability-based samples that are beyond the scope of this course.[10] However, the goal of these more complex types of samples is generally to impart more control into the sampling design.

◇ A proper SRS requires each individual i the population to be assigned a unique number.

If the population is such that a number cannot be assigned to each individual, then the researcher must try to use a method 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 where various aspects of polling are discussed.

- 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?

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

[10] For example, stratified samples, nested, and multistage samples.

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 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.[11] In addition, a "preponderance of evidence" may be arrived at if enough observational studies point to the same conclusion.

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

MODULE 4

GETTING DATA INTO R

Contents

4.1 Setting Up R and Helpers

R is a software environment for performing statistical analyses. RStudio is a helper program that makes it easier to use R. `NCStats` is a set of R functions that make the statistical methods used in this class easier. Herein, I will refer to R but you will interact with R through RStudio.

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. 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 and run again. When writing a report, all tabular and graphical output should be copied from R 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 "comma-separated values" 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.

[1] Specifics for how to format homework assignments is on the course syllabus

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.

Results from an expression are typically saved for further computations by assigning the results 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
```

The result of an expression can be both assigned and printed by surrounding the 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
```

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.

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 modules.

◇ 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.

[2] The spaces between the expression and the parentheses are only needed to increase legibility.

4.3 Working With Data

4.3.1 Data Types

Data in R will be designated as an integer (whole numbers), numeric (non-integer numeric values), character (strings), factor (group membership), or logical (TRUE/FALSE). The type of data largely dictates the type of analysis 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 the group to which an individual belongs. This description allows for certain analyses in later modules.

4.3.2 Entering Data

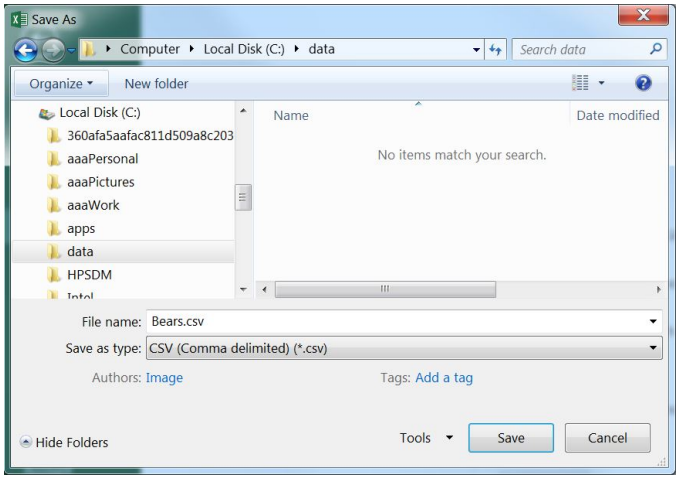
For real data (i.e., several variables from many individuals) it is most efficient to enter data into a comma-separated values (CSV) file and then import that file into R. Creating a CSV file with Microsoft Excel is described below, though there are other ways to create CSV files (see FAQs on class webpage).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

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

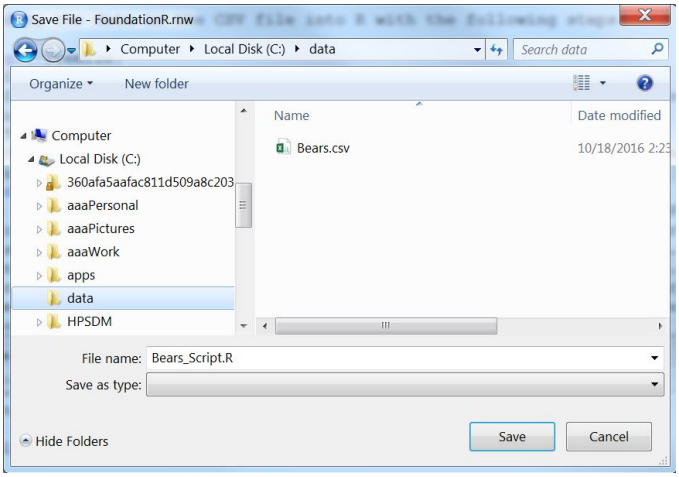
Variable names must NOT contain spaces. For example, don't use *total length* or *length (cm)*. If you feel the need to have longer variable names, then separate the parts with a period (e.g., *length.cm*) or an underscore (e.g., *length_cm*). Furthermore, numerical measurements should NOT include units (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`).

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 are used to load the data in the CSV file into RStudio.

- Open RStudio.
- Open a new script by selecting the `File`, `New File`, `R Script` menu items.
- Type `library(NCStats)` in the new script (i.e., in the upper-left pane).
- 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 (lower-left pane). 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 descried 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.

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 Module 19). For example, the following data are methyl mercury levels recorded in mussels from two locations labeled as "impacted" and "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

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 these notes 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 R 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 (in upper-right pane). Alternatively, the first and last three rows of a data.frame 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, it is useful to examine the structure of the data.frame as 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 `$` notation was used to identify a particular variable (i.e., `loc`) within a data.frame (`bears`). Think of variables as being nested inside data.frames and, thus, to access the variable you must first identify the data.frame in which it exists and then the name of the variable. The `$` 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
```

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 *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.[4] 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"
```

Single variables from a data.frame are vectors.

```
> 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
```

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.

[4] Note that `c` comes from the word "concatenate."

MODULE 5

UNIVARIATE EDA - CATEGORICAL

Contents

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 module, methods to construct tables and graphs for categorical data are described and the interpretation of the results demonstrated.

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

These concepts are illustrated with three data sets. First, data recorded about MTH107 students in the Winter 2010 semester will be used. Specifically, whether or not a student was required to take the courses and the student's year-in-school will be summarized. Whether or not a student was required to take the course for a subset of individuals is shown in Table 5.1.

Table 5.1. Whether (Y) or not (N) MTH107 was required for eight individuals in MTH107 in Winter 2010.

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

Second, the General Sociological Survey (GSS) is a very large survey that has been administered 25 times since 1972. The 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 in GSSEnviroQues.csv.

5.1 Summaries

5.1.1 Frequency and Percentage Tables

A simple method to summarize categorical data is to count the number of individuals in each level of the categorical variable. These counts are called frequencies and the resulting table (Table 5.2) 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 5.2. Frequency table for whether MTH107 was required (Y) or not (N) for eight individuals in MTH107 in Winter 2010.

Required	Freq
Y	5
N	3

The remainder of this module will use results from the entire class rather than the subset used above. For example, frequency tables of individuals by sex and year-in-school for the entire class are in Table 5.3.

Table 5.3. Frequency tables for whether (Y) or not (N) MTH107 was required (Left) and year-in-school (Right) for all individuals in MTH107 in Winter 2010.

Required	Freq
Y	38
N	30

Year	Freq
Fr	19
So	12
Jr	29
Sr	9

Frequency tables are often modified to show the percentage of individuals in each level. **Percentage tables** are constructed from frequency tables by dividing the number of individuals in each level 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 5.4) for students in MTH107 is constructed from Table 5.3 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 5.4).

Table 5.4. Percentage tables for whether (Y) or not (N) MTH107 was required (Left) and year-in-school (Right) for all individuals in MTH107 in Winter 2000.

Required	Perc
Y	55.9
N	44.1

Year	Perc
Fr	27.9
So	17.6
Jr	42.6
Sr	13.2

5.1.2 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, to highlight the categorical nature of the data, bars on a bar plot do not touch. A bar plot for whether or not individuals were required to take MTH107 is in Figure 5.1-Left. This bar plot does not add much to the frequency table because there were only two categories. However, bar plots make it easier to compare the number of individuals in each of several categories as in Figure 5.1-Right.

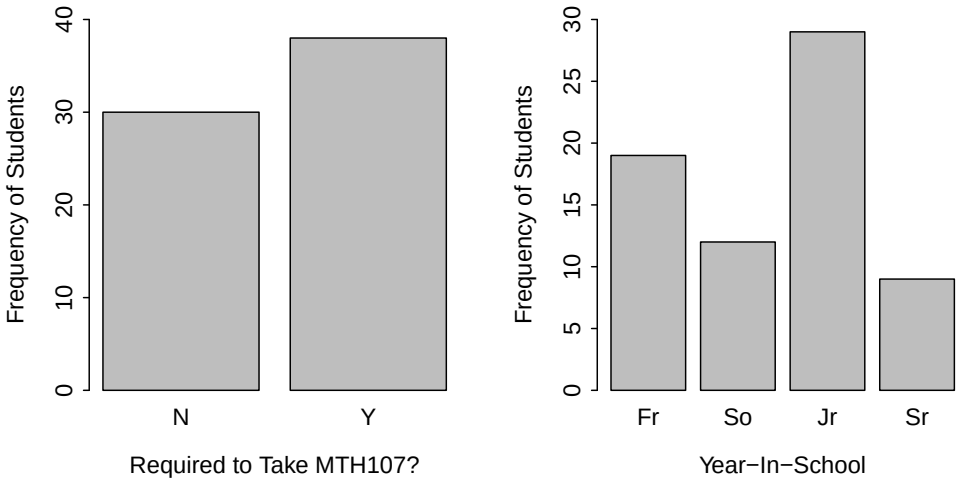


Figure 5.1. Bar charts of the frequency of individuals in MTH107 during Winter 2010 by whether or not they were required to take MTH107 (**Left**) and year-in-school (**Right**).

◇ **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.**

5.1.3 Using in R

The General Sociological Survey (GSS) data are loaded, the structure of the data.frame is examined, and the levels of the *recycle* variable are shown below. These results show the five levels in the *recycle* factor variable, ordered alphabetically as is the default in R. However, the levels should be "Always", "Often", "Sometimes", "Never", and "Not Avail" to follow the natural order of this ordinal variable.

```
> 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 1 ...
 $ tempgen: Factor w/ 5 levels "Extremely","Not",..: 1 1 1 1 1 1 1 1 1 1 ...
> levels(GSS$recycle)
[1] "Always"    "Never"     "Not Avail" "Often"     "Sometimes"
```

The order of a factor variable is controlled by including the ordered level names within a vector given to `levels=` in `factor()`. The names of the levels in this vector must be exactly as they appear in the

original variable and they must be contained within quotes. The levels of *recycle* were reordered below. 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.

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

◇ **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.**

A frequency table of a single categorical variable is computed with `xtabs()`, where the first argument is a one-sided formula of the form `~var` and the corresponding data.frame is in `data=`. The result from `xtabs()` should be assigned to an object for further use. For example, the frequency table is produced, stored in `tabRecycle`, and displayed below. Thus, 1289 respondents answered "Always" to the recycling question.

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

A percentage table is computed by including the saved frequency table as the first argument to `percTable()`.[1] The number of digits of output is controlled with `digits=`. Thus, 36.4% of respondents answered "Always" to the recycling question.

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

A bar plot is produced by giving the saved `xtabs()` object as the first argument to `barplot()`. The x- and y-axes may be explicitly labeled with `xlab=` and `ylab=`, respectively. For example, the bar plot for the recycling data (Figure 5.2) is produced below.

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

5.2 Example Interpretations

For categorical data, an appropriate EDA consists of identifying the major characteristics among the categories. 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. Other examples are below.

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

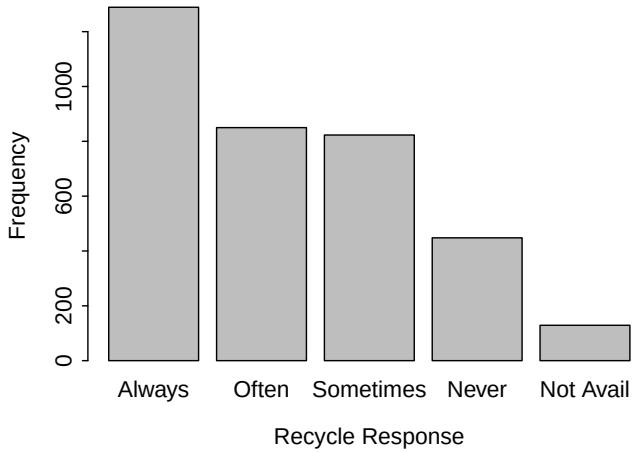


Figure 5.2. Bar chart of the frequency of responses to the recycling question on the GSS.

5.2.1 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 5.5; Figure 5.3). Very few legumes or wildflowers were found in the sample.

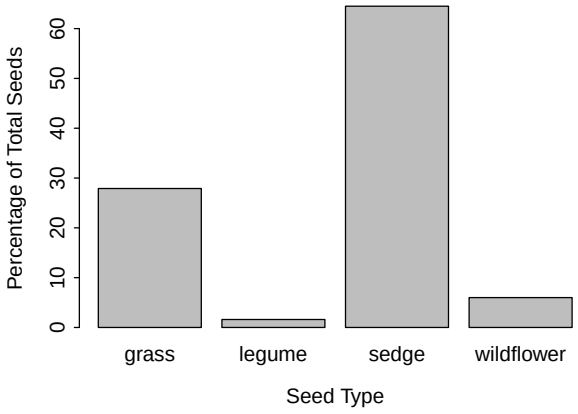


Figure 5.3. Barplot of the percentage of wetland seeds by type.

Table 5.5. Percentage distribution of wetland seeds by type.

grass	legume	sedge	wildflower
27.9	1.6	64.5	6.0

R Appendix:

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

MODULE 6

SUMMARIES FOR ONE QUANTITATIVE VARIABLE)

Contents

SUMMARIZING LARGE QUANTITIES OF DATA WITH few graphical or numerical summaries makes it is easier to identify meaning from data (discussed in Module 1). Numeric and graphical summaries specific to a single quantitative variable are described in this module. Interpretations from these numeric and graphical summaries are described in the next module.

Two data sets will be considered in this module when making calculations "by hand" (i.e., without using R). The first data set consists of the number of open pit mines in countries that have open pit mines (Table 6.1).[1] The second data set is Richter scale recordings for 15 major earthquakes (Table 6.2). A third data set – number of days of ice cover at ice gauge station 9004 in Lake Superior – will be used to demonstrate calculations with R. These data are in LakeSuperiorIce.csv and are loaded into `LSI` below.[2]

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

Table 6.1. Number of open pit mines in countries that have open pit mines.

2.0	11.0	4.0	1.0	15.0	12.0	1.0	1.0	3.0	2.0	2.0	1.0	1.0
1.0	1.0	2.0	4.0	1.0	4.0	2.0	4.0	2.0	1.0	4.0	11.0	1.0

Table 6.2. 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

[1] These data were collected from this page. See Section 4.3.2 for how to enter these data into R.
[2] See Section 4.3.2 for how to access these data. These data are originally from the National Snow and Ice Data Center.

6.1 Numerical Summaries

A "typical" value and the "variability" of a quantitative variable are often described from numerical summaries. Calculation of these summaries is described in this module, whereas their interpretation is described in Module 6. As you will see in Module 6, "typical" values are measures of **center** and "variability" is often described as **dispersion** (or spread). Three measures of center are the median, mean, and mode. Three measures of dispersion are the inter-quartile range, standard deviation, and range.

All measures computed in this module are summary statistics – i.e., they are computed from individuals in a sample. Thus, the name of each measure should be preceded by "sample" – e.g., sample median, sample mean, and sample standard deviation. These measures could be computed from every individual, if the population was known. These values would then be parameters and would be preceded by "population" – e.g., population median, population mean, and population standard deviation.[3]

6.1.1 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,[4]

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 open pit data from Table 6.1 are,

1	1	1	1	1	1	1	1	1	1	2	2	2
2	2	2	3	4	4	4	4	4	11	11	12	15

Because $n = 26$, the $mp = \frac{26+1}{2} = 13.5$. The mp is not an integer so the median is the average of the values in the 13th and 14th ordered positions (i.e., the two positions closest to mp). Thus, the median number of open pit mines in this sample of countries is $\frac{2+2}{2} = 2$.

Consider finding the median of the Richter Scale magnitude recorded for fifteen major earthquakes as another example (ordered data are in Table 6.2). Because $n = 15$, the $mp = \frac{15+1}{2} = 8$. The mp is an integer so the median is the value of the individual in the 8th ordered position, which is 7.1.

◇ Don't forget to order the data when computing the median.

6.1.2 Inter-Quartile Range

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

[3] See Module 2.1 for clarification on the differences between populations and samples and parameters and statistics.

[4] 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.

the median, and then finding the medians of the two halves.[5] 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.[6]
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 open pit mine data (the median was computed in Section 6.1.1). Because $n = 26$ is even, the halves of the data split naturally into two halves each with 13 individuals. Therefore, the $mp = \frac{13+1}{2} = 7$ and the median of each half is the value of the individual in the seventh position. Thus, $Q1 = 1$ and $Q3 = 4$.

1	1	1	1	1	1	1	1	1	1	2	2	2
2	2	2	3	4	4	4	4	4	11	11	12	15

In summary, the first, second, and third quartiles for the open pit mine data are 1, 2, and 4, respectively. These three values separate the ordered individuals into approximately four equally-sized groups – those with values less than (or equal to) 1, with values between (inclusive) 1 and 2, with values between (inclusive) 2 and 4, and with values greater (or equal to) than 4.

As another example, consider finding the quartiles for the earthquake data (Table 6.2). Recall from above (Section 6.1.1) 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, 7.8, 8.1. The middle position for each half is then $mp = \frac{7+1}{2} = 4$. Thus, the median for each half is 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 $Q3 = 7.7$.

The interquartile range (IQR) is the difference between $Q3$ and $Q1$, namely $Q3 - Q1$. However, the IQR (as strictly defined) suffers from a lack of information. For example, what does an IQR of 9 mean? It can have a completely different interpretation if the IQR is from values of 1 to 10 or if it is from values of 1000 to 1009. Thus, the IQR is more useful if presented as both $Q3$ and $Q1$, rather than as the difference. Thus, for example, the IQR for the open pit mine data is from a $Q3$ of 4 to a $Q1$ of 1 and the IQR for the earthquake data is from a $Q3$ of 7 to a $Q1$ of 6.5.

◇ **The IQR can be thought of as the "range of the middle half of the data."**

◇ **When reporting the IQR, explicitly state both $Q3$ and $Q1$ (i.e., do not subtract them).**

6.1.3 Mean

The mean is the arithmetic average of the data. The sample mean is denoted by $\bar{x}$ and the population mean by μ. The mean is simply computed by adding up all of the values and dividing by the number of individuals. If the measurement of the generic variable x on the ith individual is denoted as x_i, then the sample mean is computed with these two steps,

[5] You should review how a median is computed before proceeding with this section.
[6] Some authors put the median into both halves when n is odd. The difference between the two methods is minimal for large n.

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,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{6.1.1}$$

For example, the sample mean of the open pit mine data is computed as follows:

$$\bar{x} = \frac{2+11+4+1+15+...+2+1+4+11+1}{26} = \frac{94}{26} = 3.6$$

Note in this example with a discrete variable that it is possible (and reasonable) to present the mean with a decimal. For example, it is not possible for a country to have 3.6 open pit mines, but it IS possible for the mean of a sample of countries to be 3.6 open pit mines.

◇ **As a general rule-of-thumb, present the mean with one more decimal than the number of decimals it was recorded in.**

6.1.4 Standard Deviation

The sample standard deviation, denoted by s, 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 (i.e., $x_i - \bar{x}$).
3. Square each difference (i.e., $(x_i - \bar{x})^2$).
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

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{6.1.2}$$

The calculation of the standard deviation of the earthquake data (Table 6.2) is facilitated with the calculations shown in Table 6.3. In Table 6.3, note that

- $\bar{x}$ is the sum of the "Value" column divided by $n = 15$ (i.e., $\bar{x} = 7.07$).
- The "Diff" column is each observed value minus $\bar{x}$ (i.e., Step 2).
- The "Diff^2" column is the square of the differences (i.e., Step 3).

- The sum of the "Diff^2" column is Step 4.
- The sample variance (i.e., Step 5) is equal to this sum divided by $n-1=14$ or $\frac{6.773}{14}=0.484$.
- The sample standard deviation is the square root of the sample variance or $s=\sqrt{0.484}=0.696$.

Table 6.3. 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

From this, on average, each earthquake is approximately 0.7 Richter Scale units different than the average earthquake in these data.

◇ In the standard deviation calculations don't forget to take the square root of the variance.

◇ The standard deviation is greater than or equal to zero.

The standard deviation can be thought of as "the average difference between the values and the mean." This is, however, not a strict definition because the formula for the standard deviation does not simply add the differences and divide by n as this definition would imply. Notice in Table 6.3 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 (see Section 7.3 in the next module). Thus, the mean difference will always be zero. This "problem" is 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 differences between the values and the mean. Therefore, "the average difference between the values and the mean" works as a practical definition of the meaning of the standard deviation, but it is not strictly correct.

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

Further note that the mean is the value that minimizes the value of the standard deviation calculation – i.e.,

putting any other value besides the mean into the standard deviation equation will result in a larger value.

Finally, you may be wondering why the sum of the squared differences in the standard deviation calculation is divided by $n-1$, rather than n. Recall (from Section 2.1) 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 Module 12).

6.1.5 Mode

The mode is the value that occurs most often in a data set. For example, one open pit mine is the mode in the open pit mine data (Table 6.4).

Table 6.4. Frequency of countries by each number of open pit mines.

Number of Mines	1	2	3	4	11	12	15
Freq of Countries	10	6	1	5	2	1	1

The mode for a continuous variable is the class or bin with the highest frequency of individuals. For example, if 0.5-unit class widths are used in the Richter scale data, then the modal class is 6.5-6.9 (Table 6.5).

Table 6.5. Frequency of earthquakes by Richter Scale class.

Richter Scale Class	5.5-5.9	6-6.4	6.5-6.9	7-7.4	7.5-7.9	8-8.4
Freq of Earthquakes	1	1	5	3	4	1

Some data sets may have two values or classes with the maximum frequency. In these situations the variable is said to be **bimodal**.

6.1.6 Range

The range is the difference between the maximum and minimum values in the data and measures the ultimate dispersion or spread of the data. The range in the open pit mine data is 15-1 = 14.

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 values of 1 to 10 or if it came from values of 1000 to 1009. Thus, the range is more instructive if presented as both the maximum and minimum value rather than the difference.

6.1.7 Computation of Summaries in R

All summary statistics described above, with the exception of the mode, is calculated in R with `Summarize()`. To summarize a single variable a one-sided formula of the form `~quant` is used, where `quant` generically represents the quantitative variable, along with the `data=` argument. The number of digits after the decimal place is controlled with `digits=`.

```
> 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
```

From this it is seen that the sample median is 114 days, sample mean is 107.8 days, sample IQR is from 97 to 118 days, the sample standard deviation is 21.59 days, and the range is from 48 to 146.

6.2 Graphical Summaries

6.2.1 Histogram

A histogram plots the frequency of individuals (y-axis) in classes of values of the quantitative variable (x-axis). Construction of a histogram begins by creating classes of values for the variable of interest. The easiest way to create a list of classes is to divide the range (i.e., maximum minus minimum value) by a "nice" number near eight to ten, and then round up to make classes that are easy to work with. The "nice" number between eight and ten is chosen to make the division easy and will be the number of classes. For example, the range of values in the open pit mine example is 15-1 = 14. A "nice" value near eight and ten to divide this range by is seven. Thus, the classes should be two units wide (=14/7) and, for ease, will begin at 0 (Table 6.6).

Table 6.6. Frequency table of number of countries in two-mine-wide classes.

Class	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15
Frequency	10	7	5	0	0	2	1	1

The frequency of individuals in each class is then counted (shown in the second row of Table 6.6). The plot is prepared with values of the classes forming the x-axis and frequencies forming the y-axis (Figure 6.1A). The first bar added to this skeleton plot has the bottom-left corner at 0 and the bottom-right corner at 2 on the x-axis, and a height equal to the frequency of individuals in the 0 and 1 class (Figure 6.1B). A second bar is then added with the bottom-left corner at 2 and the bottom-right corner at 4 on the x-axis, and a height equal to the frequency of individuals in the 2 and 3 class (Figure 6.1C). This process is continued with the remaining classes until the full histogram is constructed (Figure 6.1D).

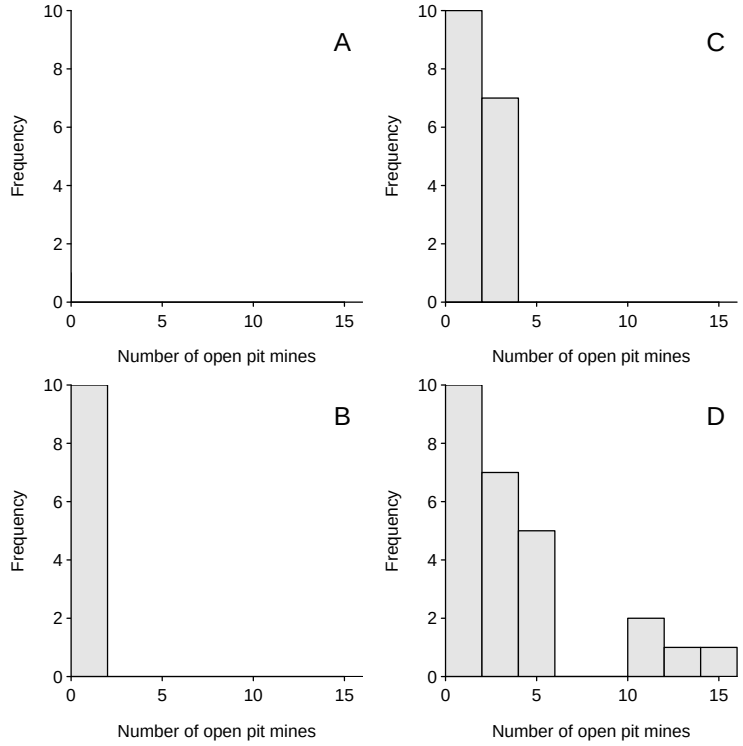


Figure 6.1. Steps (described in text) illustrating the construction of a histogram.

Ideally eight to ten classes are used in 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 6.2).

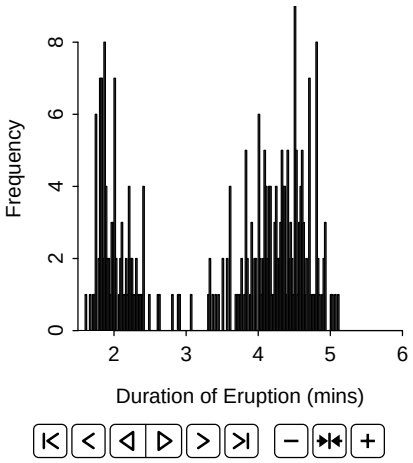


Figure 6.2. Histogram of length of eruptions for Old Faitfhul geyser with varying number of classes.

6.2.2 Boxplot

The **five-number summary** consists of the minimum, Q1, median, Q3, and maximum values (effectively contains the range, IQR, and median). For example, the five-number summary for the open pit mine data is 1, 1, 2, 4, and 15 (all values computed in the previous section). The five-number summary may be displayed as a **boxplot**. A traditional boxplot (Figure 6.3-Left) 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 vertical lines from Q1 to the minimum and from Q3 to the maximum. In modern boxplots (Figure 6.3-Right) the upper line extends from Q3 to the last observed value that is within 1.5 IQRs of Q3 and the lower line 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 traditional and modern boxplots will be the same.

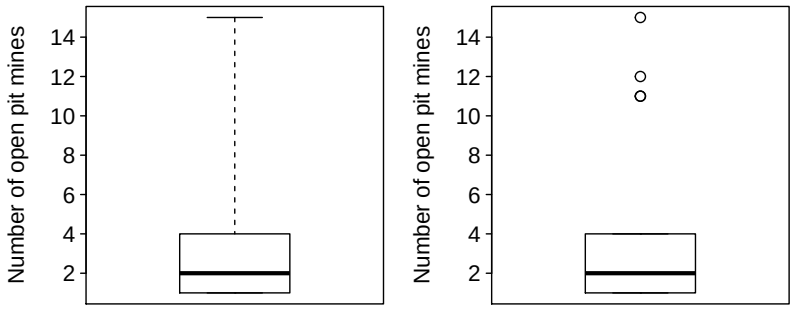


Figure 6.3. Traditional (Left) and modern (Right) boxplots of the open pit mine data.

6.2.3 Construction of Graphs in R

A simple (by default) histogram is constructed in R 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=`.[7] The x-axis label may be improved from the default value by including a label in `xlab=`. The width of the classes may be controlled with a positive integer in `w=`.[8]

```
> hist(~days,data=LSI,xlab="Days of Ice Cover")      # Fig 5.4-Left
> hist(~days,data=LSI,xlab="Days of Ice Cover",w=20) # Fig 5.4-Right
```

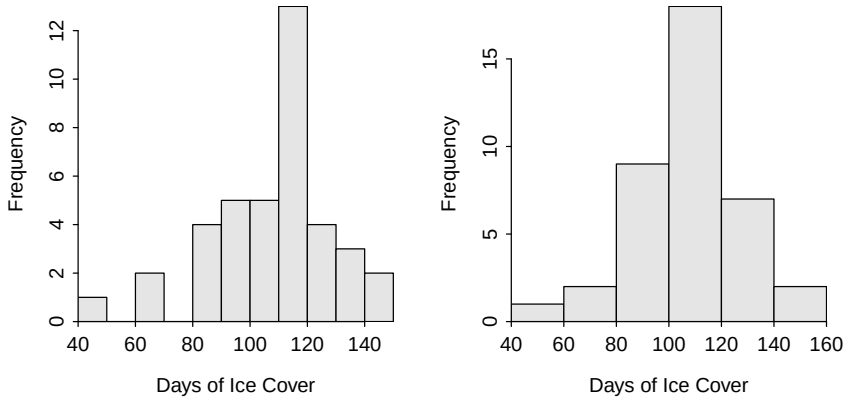


Figure 6.4. Histograms of the duration of ice cover at ice gauge 9004 in Lake Superior using the default class widths (Left) and widths of 20 days (Right).

A modern boxplot of a single variable is constructed in R with `boxplot()`, where the first argument is usually a specific variable in a data.frame. Additionally, the y-axis may be properly labeled with `ylab=`.

```
> boxplot(LSI$days,ylab="Days of Ice Cover")
```

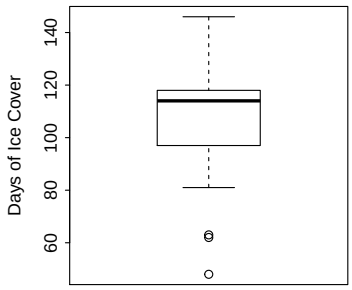


Figure 6.5. Boxplot of the duration of ice cover at ice gauge 9004 in Lake Superior.

[7] Note that this is the same formula used in `Summarize()`.
[8] The endpoints for the classes may also be set by giving a vector of endpoints to `breaks=`.

◇ **The default histogram and boxplot should be modified by properly labeling the axes.**

6.3 Multiple Groups

It is common to need to compute numerical or construct graphical summaries of a quantitative variable separately for 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 `group` is given to `data=`. Summary statistics are separated by group by supplying the same formula and `data=` arguments to `Summarize()`.

As an example, the LSI data.frame contains a `period` variable that indicates whether the ice season was "pre-1975" or "post-1975" (which included 1975). Thus, one may be interested in examining the distribution of annual days of ice for each of these periods period. Histograms (Figure 6.6) and summary statistics separated by period are constructed below.

```
> hist(days~period,data=LSI,ylab="Days of Ice Cover",w=20)
> Summarize(days~period,data=LSI,digits=2)
     period  n nvalid   mean    sd min Q1 median  Q3 max
1 post-1975 22     21 106.76 26.01  48 99  116.0 123 146
2  pre-1975 20     18 109.11 15.59  82 97  110.5 118 137
```

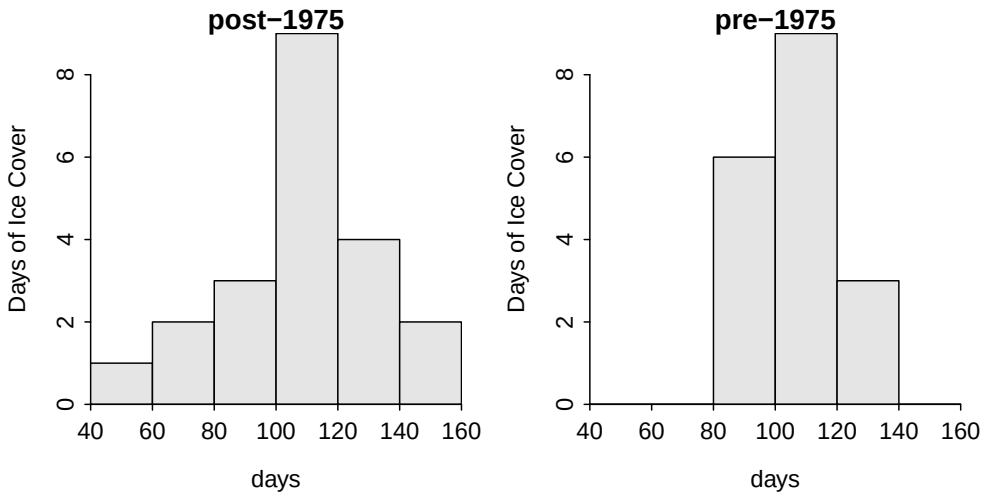


Figure 6.6. Histograms of the duration of ice cover at ice gauge 9004 in Lake Superior by period.

Side-by-side boxplots (Figure 6.7) are an alternative to separated histograms and are constructed by including the same formula and `data=` arguments to `boxplot()`.

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

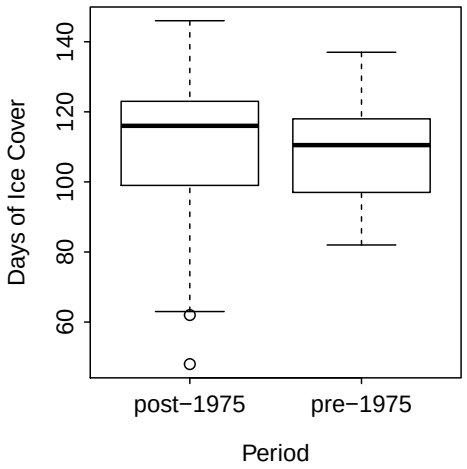


Figure 6.7. Boxplot of the duration of ice cover at ice gauge 9004 in Lake Superior by period.

Note that the formulae above required the grouping variable to be a factor. In some instances, a grouping variable may appear as an integer variable to R. For example, one may want to explore days of ice by decade, but the decade variable is not a factor variable.

```
> 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 ...
 $ period: Factor w/ 2 levels "post-1975","pre-1975": 2 2 2 2 2 2 2 2 2 2 ...
 $ temp  : num  22.9 23 25.7 20 24.8 ...
 $ days  : int  87 137 106 97 105 118 118 136 91 NA ...
```

In these cases, the variale needs to be explicitly converted to a factor variable using `factor()`, as shown below. The use of `factor()` is not needed if R already recognizes the variable as a factor variable.

```
> LSI$decade <- factor(LSI$decade)
> str(LSI)
'data.frame': 42 obs. of  5 variables:
 $ season: int  1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 ...
 $ decade: Factor w/ 5 levels "1950","1960",..: 1 1 1 1 1 2 2 2 2 2 ...
 $ period: Factor w/ 2 levels "post-1975","pre-1975": 2 2 2 2 2 2 2 2 2 2 ...
 $ temp  : num  22.9 23 25.7 20 24.8 ...
 $ days  : int  87 137 106 97 105 118 118 136 91 NA ...
```

MODULE 7

UNIVARIATE EDA - QUANTITATIVE

Contents

A univariate EDA for a quantitative variable is concerned with describing the distribution of 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 used to identify shape and the presence of outliers and to get a general feel for center and dispersion. Numerical summaries, however, are used to specifically describe center and dispersion of the variable. Computing and constructing the required numerical and graphical summaries was described in Module 6. Those summaries are interpreted here to provide an overall description of the distribution of the quantitative variable.

The same three data sets used in Module 6 are used here.

- Number of open pit mines in countries with open pit mines (Table 6.1).
- Richter scale recordings for 15 major earthquakes (Table 6.2).
- The number of days of ice cover at ice gauge station 9004 in Lake Superior.

7.1 Interpreting Shape

A distribution has two tails – a left-tail of smaller or more negative values and a right-tail of larger or more positive values (Figure 7.1). 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 distribution are approximately equal in shape (length and height), then the distribution is said to be **symmetric** (or more specifically **approximately symmetric**). If the left-tail is stretched out or is longer and flatter than the right-tail, then the distribution is negatively- or **left-skewed**. If the right-tail is stretched out or 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.

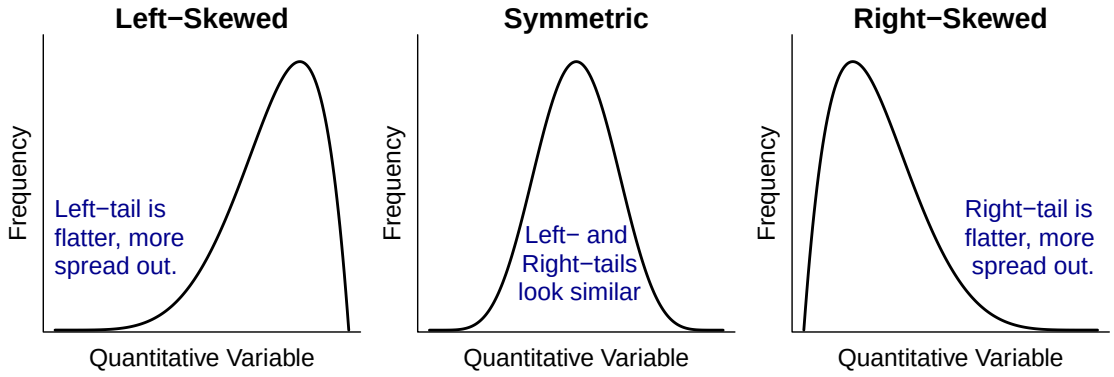


Figure 7.1. Examples of left-skewed (left), symmetric (center), and right-skewed (right) distributions.

◇ **The longer tail defines the type of skew.**

In practice, these labels form a continuum. For example, it may be difficult to discern whether the shape approximately symmetric or one of the skewed distributions. To partially address this issue, "slightly" or "strongly" may be used with "skewed" to distinguish whether the distribution is obviously skewed (i.e., "strongly skewed") or nearly symmetric (i.e., "slightly skewed").

◇ **Symmetric, left-skewed, and right-skewed descriptors are guides; many "real" distributions will not fall neatly into these categories.**

The shape of a distribution is most easily identified from a histogram. Histograms that are examples of each shape are in Figure 7.2. For the sets of skewed distributions, the distributions are less strongly skewed from left-to-right.

The shape of a distribution can also be determined from a boxplot. The relative length from the median to Q1 and the median to Q3 (i.e., the relative position of the median line in the box) indicates the shape of the distribution. If the distribution is left-skewed (i.e., lesser-valued individuals are "spread out"; Figure 7.3-Right), then median-Q1 will be greater than Q3-median. In contrast, if the distribution is right-skewed (i.e., larger-valued individuals are spread out; Figure 7.3-Middle), then Q3-median will be greater than median-Q1. Thus, the median is nearer the top of the box for a left-skewed distribution, nearer the bottom of the box for a right-skewed distribution, and nearer the center of the box for a symmetric distribution (Figure 7.3).

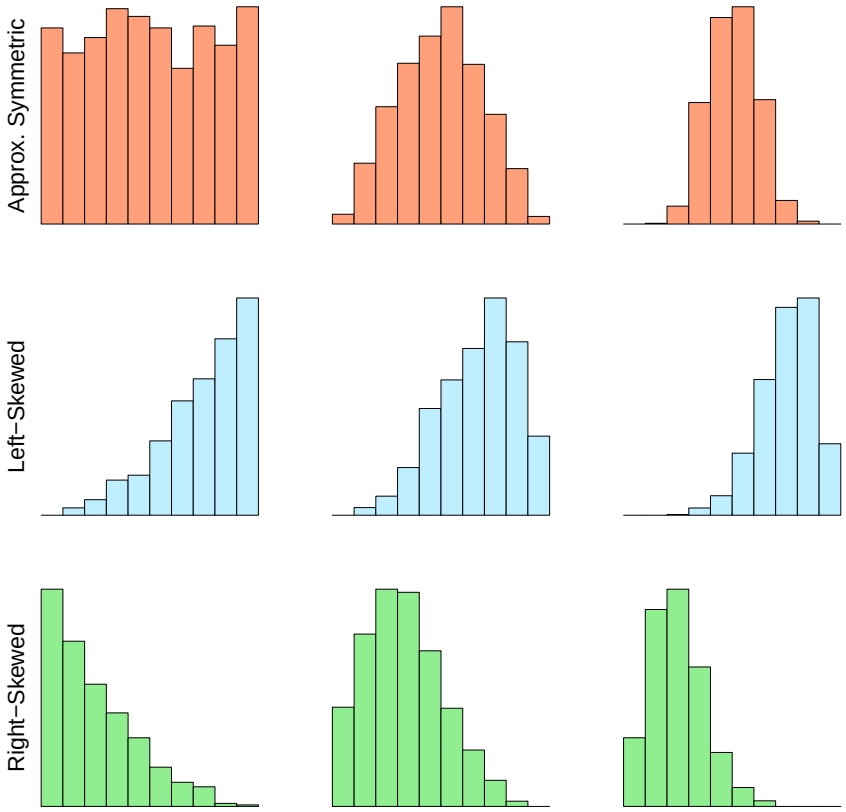


Figure 7.2. Examples of approximately symmetric (top, red), left-skewed (middle, blue), and right-skewed (bottom, green) histograms. Note that the axes labels were removed to focus on the shape of the histograms.

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

7.2 Interpreting Outliers

An outlier is an individual 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 7.4). 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.**
An outlier may be a result of human error in the sampling process. If this is the case, then the value should be corrected or removed. Other times an outlier may be an individual that was 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 should be removed from the sample. Still other times, an outlier is part of the population and should generally not be removed from the sample. In fact you may wish to highlight an outlier as an interesting observation! Regardless, it is important that you construct 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

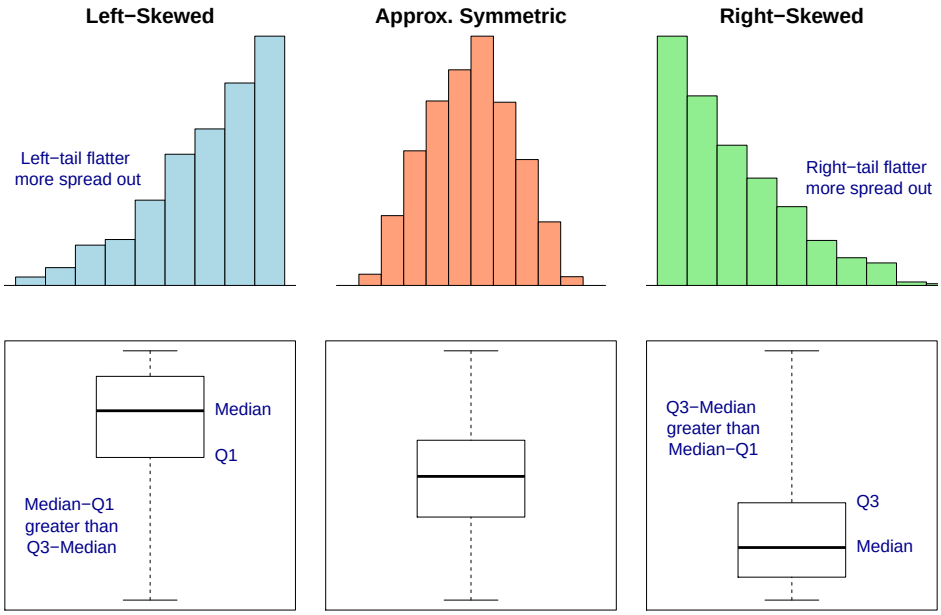


Figure 7.3. Histograms and boxplots for several different shapes of distributions.

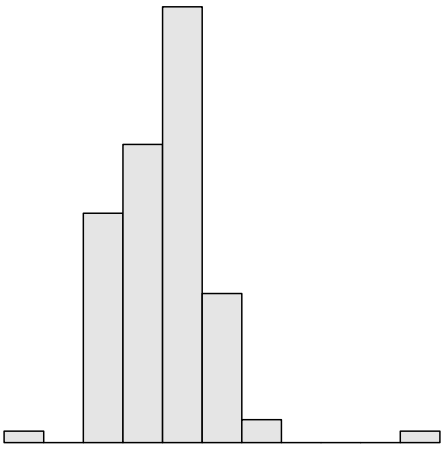


Figure 7.4. Example histogram with an outlier to the right.

cluster of values is approximately symmetric and there is one outlier to the right of the main cluster (as illustrated in Figure 7.4), **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.**

7.3 Comparing the Median and Mean

As mentioned previously, numerical measures will be used to describe the center and dispersion of a distribution. However, which values should be used? Should one use the mean or the median as a measure of center? Should one use the IQR or the standard deviation as a measure of dispersion? Which measures are used depends on how the measures respond to skew and the presence of outliers. Thus, before stating a rule for which measures should be used, a fundamental difference among the measures discussed in Module 6 is explored here.

The following discussion is focused on comparing the mean and the median. However, note that the IQR is fundamentally linked to the median (i.e., to find the IQR, the median must first be found) and the standard deviation is fundamentally linked to the mean (i.e., to find the standard deviation, the mean must first be found). Thus, **the median and IQR will always be used together to measure center and dispersion, as will the mean and standard deviation.**

The mean and median measure center in different ways. The median balances the number of individuals smaller and larger than it. The mean, on the other hand, balances the sum of the distances from it to all points smaller than it and the sum of the distances from it to all points greater than it. Thus, the median is primarily concerned with the **position** of the value rather than the value itself, whereas the mean is very much concerned about the **values** for each individual (i.e., 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 7.5-Left.[1] The median (blue line) is found by locating the middle position on the individual number axis and 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; i.e., it has the same **number** of individuals (i.e., points) above and below 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 placed such 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 total difference in values above and below the mean.

◇ **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 7.5-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 impacts the value of the mean more than 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

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

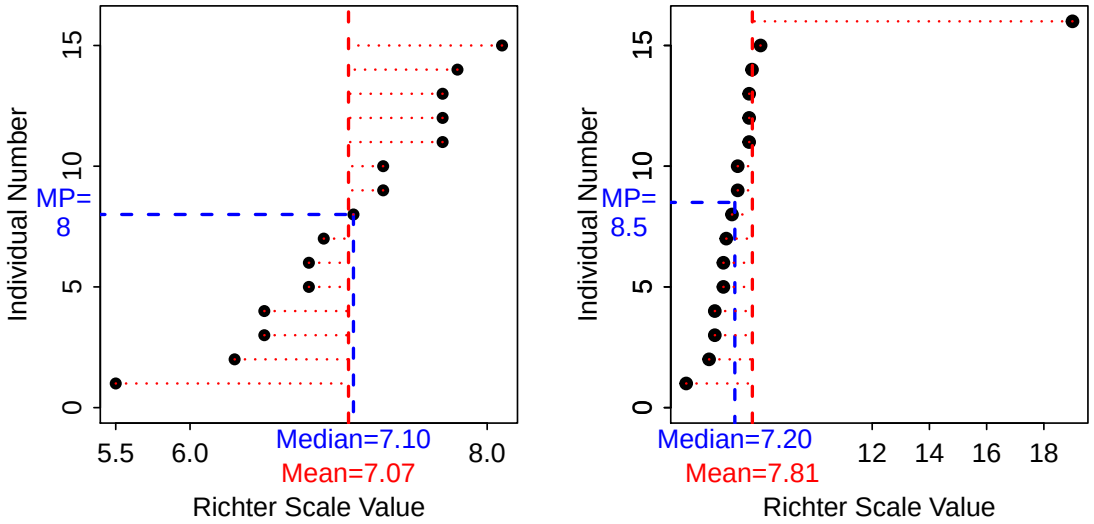


Figure 7.5. Plot of the individual number versus Richter scale values for the original earthquake data (**Left**) and the earthquake data with an extreme outlier (**Right**). The median value is shown as a blue vertical line and the mean value is shown as a red vertical line. Differences between each individual value and the mean value are shown with horizontal red lines.

the sum of the differences to the smaller values will be equal. In this example, the outlier creates a large difference to the right of the mean such that 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 impact on the mean and median (Figure 7.6). If a distribution is approximately symmetric, then the median and mean (along with the mode) will be nearly identical. If the distribution is left-skewed, then the mean will be less than the median. Finally, if the distribution is right-skewed, then the mean will be greater than the median.

◇ **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, the median is used 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.

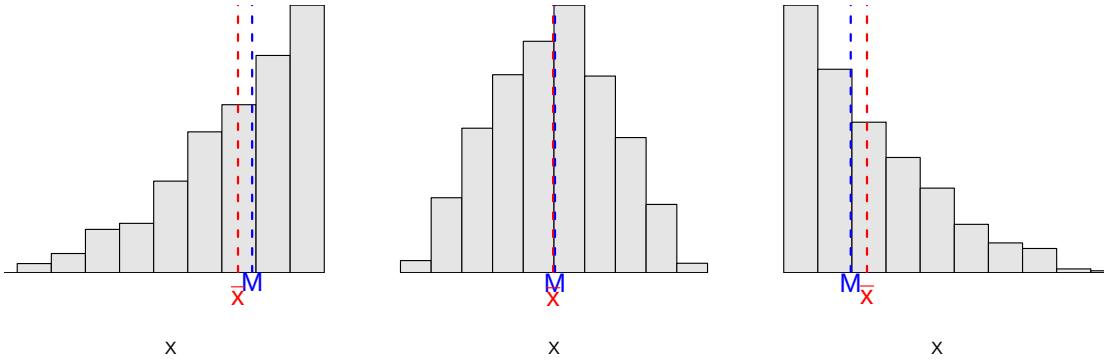


Figure 7.6. Three differently shaped histograms with vertical lines superimposed at the median (M; blue lines) and the mean ($\bar{x}$; red lines).

As note above, the IQR and standard deviation behave similarly to the median and mean, respectively, in the face of outliers and skews. Specifically, the IQR is less sensitive to outliers than the standard deviation.

7.4 Synthetic Interpretations

The graphical and numerical summaries from Module 6 and the rationale described above can be used to construct a synthetic description of the shape, outliers, center, and dispersion of the distribution of a quantitative variable. In the examples below specifically note the 1) reference to figures and tables, 2) labeling of the figures and tables, 3) that only the mean and standard deviation or the median and IQR are discussed, 4) the range was not used alone as a measure of dispersion, 5) the explanation for why either the median and IQR or the mean and standard deviation were used, and 6) an appendix of R code used was provided.

Number of Open Pit Mines

> *Construct a proper EDA for the following situation and data – "The number of open pit mines in countries that have open pit mines (Table 6.1)."*

The number of open pit mines in countries with open pit mines is strongly right-skewed with no outliers present (Figure 7.7). [*I did not call the group of four countries with 10 or more open pit mines outliers because there were more than one or two countries there.*] The center of the distribution is best measured by the median, which is 2 (Table 7.1). The range of open pit mines in the sample is from 1 to 15 while the dispersion as measured by the inter-quartile range (IQR) is from a Q1 of 1.0 to a Q3 of 4.0 (Table 7.1). I chose to use the median and IQR because the distribution was strongly skewed.

Table 7.1. Descriptive statistics of number of open pit mines in countries with open pit mines.

n	mean	sd	min	Q1	median	Q3	max
26.0	3.6	4.0	1.0	1.0	2.0	4.0	15.0

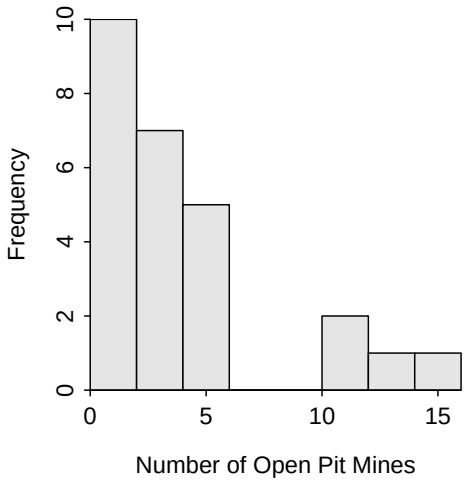


Figure 7.7. Histogram of number of open pit mines in countries with open pit mines.

R Code Appendix:

```
setwd("c:/data/")
mc <- read.csv("MineData.csv")
str(mc)
Summarize(~mines,data=mc,digits=1)
hist(~mines,data=mc,w=2,xlab="Number of open pit mines")
```

Lake Superior Ice Cover

Thoroughly describe the distribution of number of days of ice cover at ice gauge station 9004 in Lake Superior (data are in LakeSuperiorIce.csv).

The shape of number of days of ice cover at gauge 9004 in Lake Superior is approximately symmetric with no obvious outliers (Figure 7.8). The center is at a mean of 107.8 days and the dispersion is a standard deviation of 21.6 days (Table 7.2). The mean and standard deviation were used because the distribution was not strongly skewed and no outlier was present.

Table 7.2. Descriptive statistics of number of days of ice cover at ice gauge 9004 in Lake Superior..

n	nvalid	mean	sd	min	Q1	median	Q3	max
42.0	39.0	107.8	21.6	48.0	97.0	114.0	118.0	146.0

R Appendix:

```
setwd("c:/data/")
LSI <- read.csv("LakeSuperiorIce.csv")
str(LSI)
hist(~days,data=LSI,xlab="Day of Ice Cover",ylab="Frequency of Years",w=20)
Summarize(~days,data=LSI,digits=1)
```

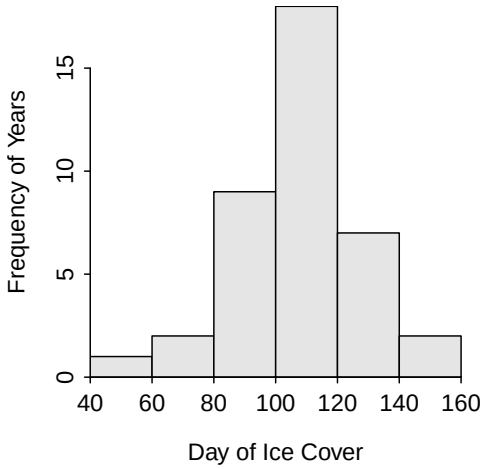


Figure 7.8. Histogram of number of days of ice cover at ice gauge 9004 in Lake Superior.

Crayfish Temperature Selection

Peck (1985) examined the temperature selection of dominant and subdominant crayfish (Orconectes virilis) together in an artificial stream. The temperature (^oC) 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.

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 temperatures selected by the dominant crayfish is slightly left-skewed (Figure 7.9) with a possible weak outlier at the maximum value of 30^oC (Table 7.3). The center is best measured by the median, which is 23^oC (Table 7.3) and the dispersion is best measured by the IQR, which is from 21 to 25^oC (Table 7.3). I used the median and IQR because of the (combined) skewed shape and outlier present.

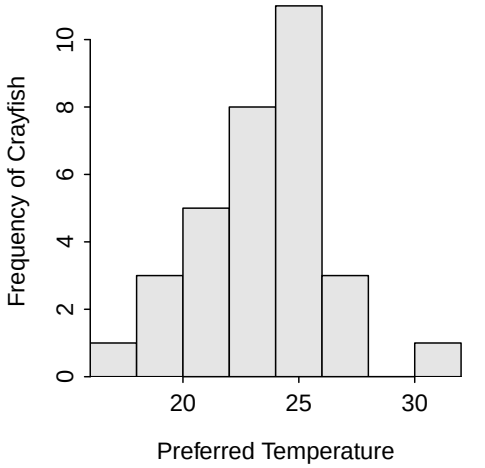


Figure 7.9. Histogram of crayfish temperature preferences.

Table 7.3. 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 Appendix:

```
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)
```

MODULE 8

NORMAL DISTRIBUTION

Contents

A MODEL FOR THE DISTRIBUTION of a single quantitative variable can be visualized by "fitting" a smooth curve to a histogram (Figure 8.1-Left), removing the histogram (Figure 8.1-Center), and using the remaining curve (Figure 8.1-Right) as a model for the distribution of the entire population of individuals. The smooth red curve drawn over the histogram serves as a model for the distribution of the **entire population**. If the smooth curve follows a known distribution, then certain calculations are greatly simplified.

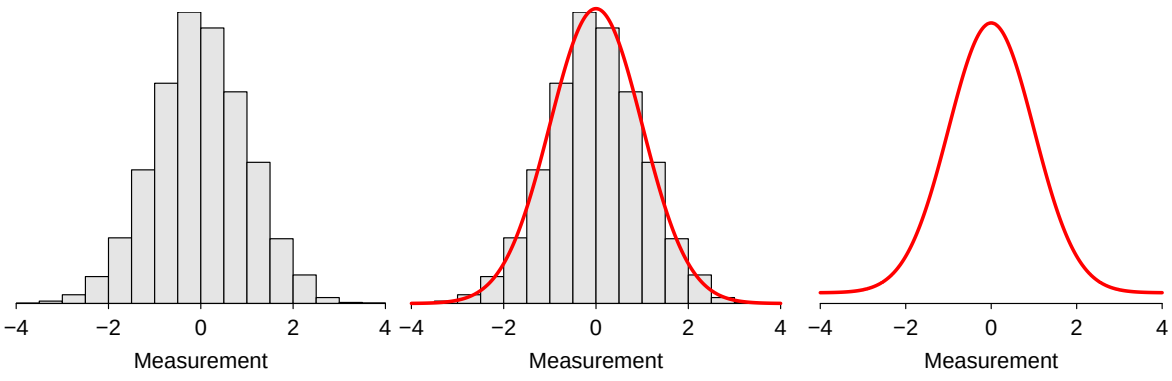


Figure 8.1. Depiction of fitting a smooth curve to a histogram to serve as a model for the distribution.

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 distributions).[1] The use of a normal distribution model to make certain calculations is demonstrated in this module.

8.1 Characteristics of a Normal Distribution

The normal distribution is the familiar bell-shaped curve (Figure 8.1-Right). Normal distributions have two parameters – the population mean, μ, and the population standard deviation, σ – that control the exact shape and position of the distribution. Specifically, the mean μ controls the center and the standard deviation σ controls the dispersion of the distribution (Figure 8.2).

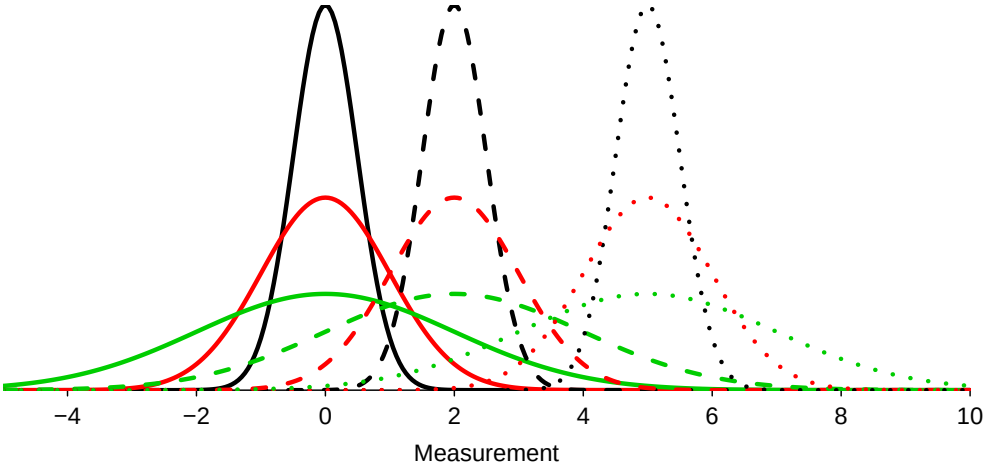


Figure 8.2. Nine normal distributions. Distributions with the same line type have the same value of μ (solid is μ=0, dashed is μ=2, dotted is μ=5). Distributions with the same color have the same value of σ (black is σ=0.5, red is σ=1, and green is σ=2).

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. centered at μ,
3. have inflection points at $\mu \pm \sigma$, and
4. have a total area under the curve equal to 1.

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 μ=0 and a standard deviation of σ=1.

[1] See Module 12.

8.2 Simple Areas Under the Curve

A common problem is to determine the proportion of individuals with a value 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 $\mu g \cdot m^{-3}$, 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 explore calculations for the generic variable X shown in Figure 8.3.

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 8.3-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 8.3-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 are 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.

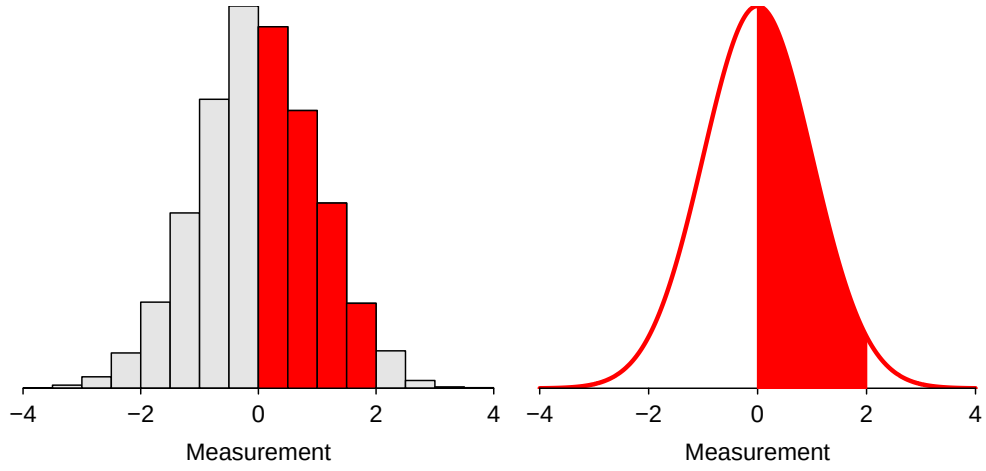


Figure 8.3. Depiction of finding the proportion of individuals between 0 and 2 on a histogram (**Left**) and on a standard normal distribution (**Right**).

◇ **The proportion of individuals between two values of a variable that is normally distributed is the area under the normal distribution between those two values.**

The 68-95-99.7 (or Empirical) Rule states that 68% of individuals that follow a normal distribution have values between $\mu - 1\sigma$ and $\mu + 1\sigma$, 95% have values between $\mu - 2\sigma$ and $\mu + 2\sigma$, and 99.7% have values between $\mu - 3\sigma$ and $\mu + 3\sigma$ (Figure 8.4).

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.

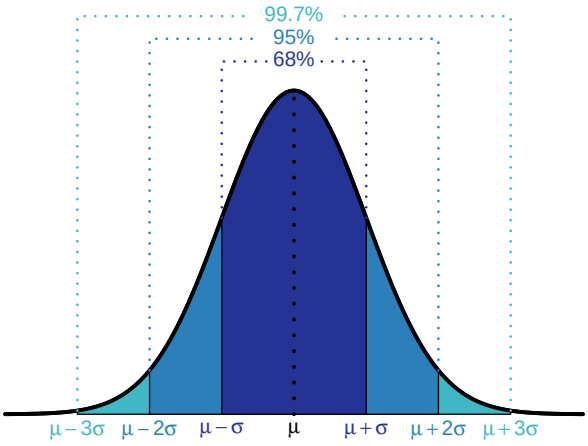


Figure 8.4. Depiction of the 68-95-99.7 (or Empirical) Rule on a normal distribution.

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 a value of A greater than 5 (Figure 8.5) is found by first realizing that 95% of the individuals on this distribution fall between 1 and 5 (i.e., $\pm 2\sigma$ from μ). By subtraction this means that 5% of the individuals must be less than 1 **AND** greater than 5. Finally, because normal distributions are symmetric, the same percentage 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.

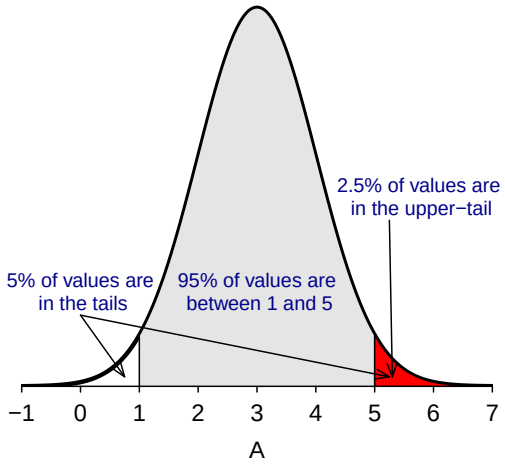


Figure 8.5. The N(3,1) distribution depicting how the 68-95-99.7 Rule is used to compute the percentage of individuals with values greater than 5.

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

8.3 More Complex Areas (Forward Calculations)

Areas under the curve relative to non-integer numbers of standard deviations away from the mean can only be found with the help of special tables or computer software. In this course, we will use 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 and the mean and standard deviation of the normal distribution in the `mean=` and `sd=` arguments, respectively. 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 `lower.tail=FALSE`.

For example, suppose that the heights of a population of students is known to be $H \sim N(66, 3)$. The proportion of students in this population that have a height less than 71 inches is computed below. Thus, approximately 95.2% of students in this population have a height less than 71 inches (Figure 8.6).

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

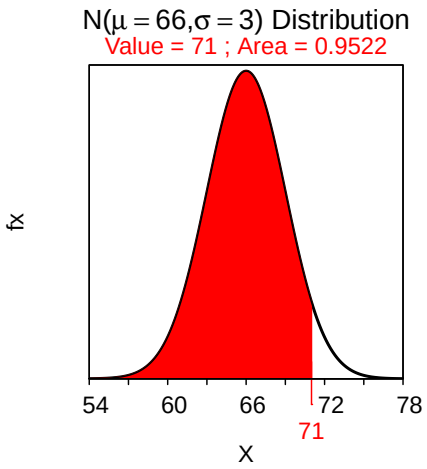


Figure 8.6. Calculation of the proportion of individuals on a $N(66, 3)$ with a value less than 71.

The proportion of students in this population that have a height *greater* than 68 inches is computed below (note use of `lower.tail=FALSE`). Thus, approximately 25.2% of students in this population have a height greater than 68 inches (Figure 8.7).

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

Finding the area between two particular values is a bit more work. To answer "between"-type questions, 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 in Figure 8.8. The area between the two values is B. The area to the left of the larger value corresponds to the area 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).

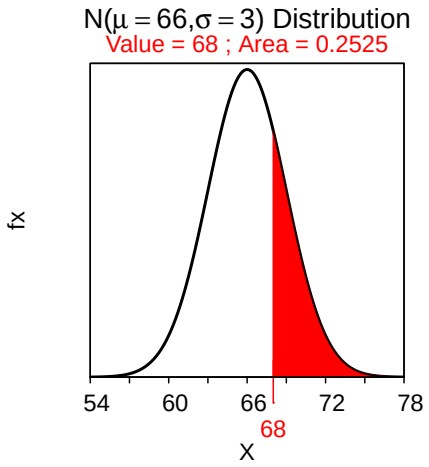


Figure 8.7. Calculation of the proportion of individuals on a $N(66, 3)$ with a value greater than 68.

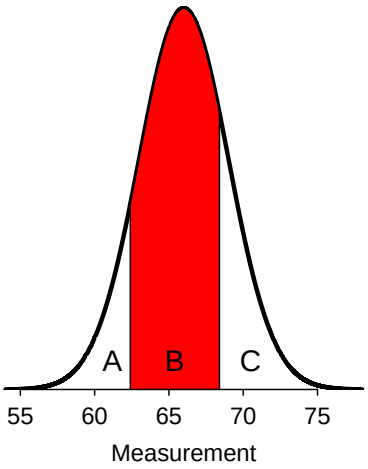


Figure 8.8. Schematic representation of how to find the area between two Z values.

For example, the area between 62 and 70 inches of height is found below. Thus, 81.8% of students in this population have a height between 62 and 70 inches.

```
> ( AB <- distrib(70,mean=66,sd=3) )  # left-of 70
[1] 0.9087888
> ( A <- distrib(62,mean=66,sd=3) )   # left-of 62
[1] 0.09121122
> AB-A                                # between 62 and 70
[1] 0.8175776
```

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

8.4 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 (where the value of X was given and a proportion of individuals was asked for), the calculations in this section give a proportion and ask for a value of X. These types of questions are called **"reverse" normal distribution questions** to contrast them with questions from the previous section.

Reverse questions are also answered with `distrib()`, though the first argument is now the given proportion (or area) of interest. The calculation is treated as a "reverse" question when `type="q"` is given to `distrib()`.[2] For example, the height that has 20% of all students shorter is 63.5 inches, as computed below (Figure 8.9).

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

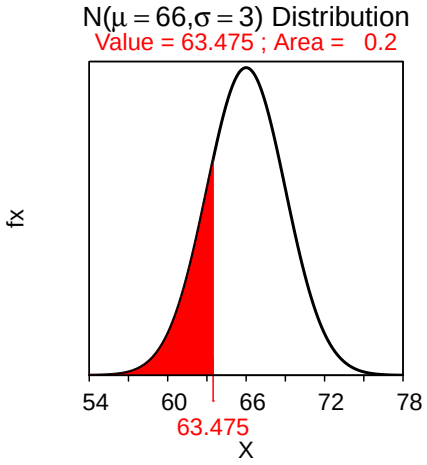


Figure 8.9. Calculation of the height with 20% of all students shorter.

"Greater than" reverse questions are computed by including `lower.tail=FALSE`. For example, 10% of the population of students is taller than 69.8 inches, as computed below (Figure 8.10).

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

"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 8.11 the area D is the symmetric area of interest. If D is 0.80, then C+E must be 0.20.[3] Because D is symmetric about mu, C and E must both equal 0.10. Thus, the

[2] "q" stands for quantile.
[3] Because all three areas must sum to 1.

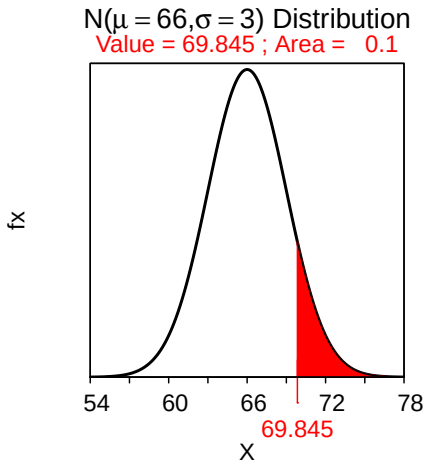


Figure 8.10. Calculation of the height with 10% of all students taller.

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 below.

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

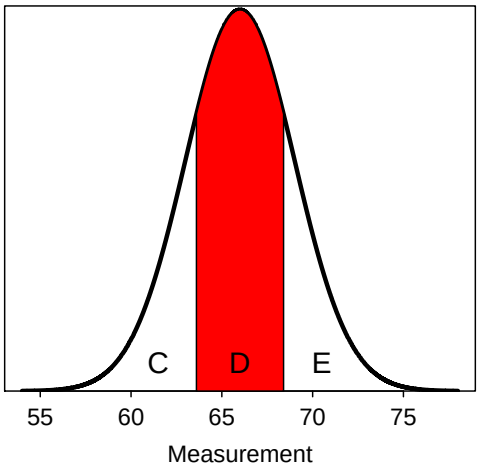


Figure 8.11. Depiction of areas in a reverse between type normal distribution question.

8.5 Distinguish Calculation Types

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

8.6 Standardization and Z-Scores

An individual that is 59 inches tall is 7 inches shorter than average if heights are $N(66, 3)$. Is this a large or a small difference? Alternatively, this same individual is $\frac{-7}{3} = -2.33$ standard deviations below the mean. Thus, a height of 59 inches is relatively rare in this population because few individuals are more than two standard deviations away from the mean.[4] As seen here, the relative magnitude that an individual differs from the mean is better expressed as the number of standard deviations that the individual is 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 mean (i.e., μ). For example, with the height variable above, 69 inches is one standard deviation above the mean, which corresponds to +1 on the standardized scale. Similarly, 60 inches is two standard deviations below the mean, which corresponds to -2 on the standardized scale. Finally, 67.5 inches on the original scale is one half standard deviation above the mean or +0.5 on the standardized scale.

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

$$Z = \frac{\text{"value"} - \text{"center"}}{\text{"dispersion"}} \tag{8.6.1}$$

or, more specifically,

$$Z = \frac{x - \mu}{\sigma} \tag{8.6.2}$$

For example, 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.

Standardized values (Z) follow a $N(0, 1)$. Thus, $N(0, 1)$ 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 on a $N(0, 1)$. This one-to-one relationship is illustrated in Figure 8.12 using the individual with a height of 59 inches and $Z = -2.33$.

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

[4] From the 68-95-99.7% Rule.

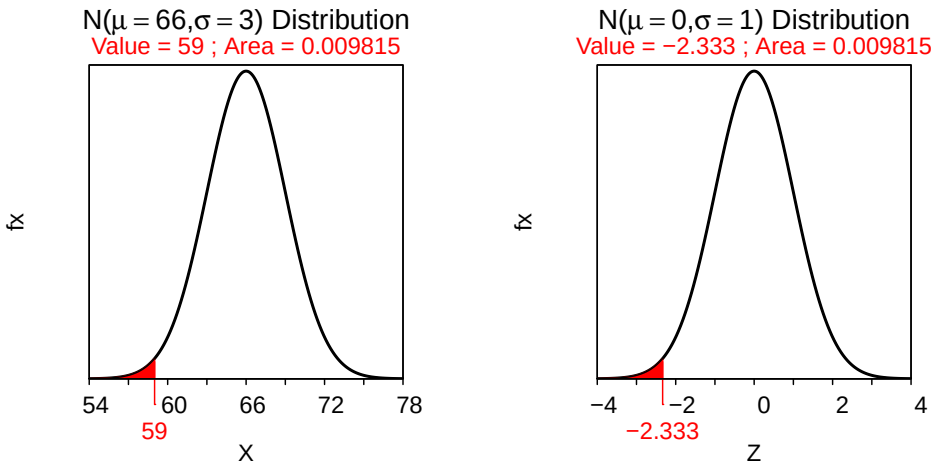


Figure 8.12. Plots depicting the area to the left of 59 on a $N(66, 3)$ (**Left**) and the area to the right of the corresponding Z-score of $Z = -2.33$ on a $N(0, 1)$ (**Right**). Not that the x-axis scales are different.

MODULE 9

BIVARIATE EDA - CATEGORICAL

Contents

TWO-WAY FREQUENCY TABLES summarize two categorical variables recorded on the same individual by displaying levels of the first variable as rows and levels of the second variable as columns. Each cell in this table contains the frequency of individuals that were in the corresponding levels of each variable. These frequency tables are often converted to percentage tables for ease of summarization and comparison among populations. This module explores the construction and interpretation of frequency and percentage tables.

The General Sociological Survey (GSS) is a very large survey that has been administered 25 times since 1972. The purpose of the GSS is to gather data on contemporary American society in order to monitor and explain trends in attitudes, behaviors, and attributes. Data from the following two questions on the GSS are used throughout this module.

- 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 and examined below.

```
> 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 5 ...
 $ grnsol: Factor w/ 5 levels "neither","un",..: 4 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 the levels of factor variables are ordered alphabetically in R (as seen below with `levels()`).

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

The order of levels can be specified using `factor()`. The variable to be reordered is the first argument to `factor()`, as well as the object to the left of the assignment operator. The desired order of the levels is listed in a vector that is given to `levels=`. It is important that the levels in this vector are "spelled" exactly as they appeared originally. Correct orders for *degree* and *grnsol* in the `gss` data.frame are created below.

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

If the natural order of levels is alphabetical or the variable is nominal, then `factor()` is not needed.

◇ **Levels for a factor variable are ordered alphabetically by default in R. If the factor variable is ordinal, then `factor()` with `levels=` may be needed to specify the correct order of levels.**

9.1 Frequency Tables

A common method of summarizing bivariate categorical data is to count individuals that have each combination of levels of the two 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 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). 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

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

If one variables can be considered as the response, then this variable should form the columns of the frequency table. For example, "willingness to cut" could be considered the response variable and it was, appropriately, placed as the column variable in Table 9.1.

Frequency Tables in R

Two-way frequency tables are constructed in R with `xtabs()`, where the first argument is a formula of the form `~rowvar+colvar` and the corresponding data.frame is in `data=`. The result of `xtabs()` should be assigned to an object for further use.

```
> ( tbl1 <- xtabs(~degree+grnsol,data=gss) )
      grnsol
degree vwill will neither  un vun
  ltHS    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 may be added to the margins of a saved table with `addMargins()`. For example, `addMargins()` was used to construct Table 9.1 from `tbl1`.

```
> addMargins(tbl1)
```

9.2 Percentage Tables

Two-way frequency tables may be converted to percentage tables for ease of comparison between levels of the variables and also between populations. 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. However, if the frequencies are converted to percentages, then this comparison is easily made. Three types of percentage tables may be constructed from a frequency table.

9.2.1 Row-Percentage Table

A **row-percentage table** is computed by dividing each cell of the frequency table by the total in the same row of the frequency table and multiplying by 100 (Table 9.2). For example, the value in the "vwill" column and "ltHS" row of the row-percentage table is computed by dividing the value in the "vwill" column and "ltHS" row of the frequency table (i.e., 40; Table 9.1) by the "Sum" of the "ltHS" row of the frequency table (i.e., 646) and multiplying by 100.

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	100.1
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

The value in each cell of a row-percentage table is the percentage OF ALL individuals in that row that 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. This statement must be read carefully. 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.

If the response variable formed the columns, then the row-percentage table allows one to compare percentages in levels of the response (i.e., columns) across groups (i.e., rows). For example, one can see that there is a general decrease in the percentage of respondents that were "very unwilling" to cut their standard of living to protect the environment as the level of education increased (Table 9.2).

Row-Percentage Table in R

Percentage tables are constructed in R by submitting the saved `xtabs()` object to `percTable()`. The number of decimals to display is controlled with `digits=`. A row-percentage table is constructed by including `margin=1`. For example, the code below produced Table 9.2.

```
> percTable(tbl1,margin=1,digits=1)
```

9.2.2 Column-Percentage Table

A **column-percentage table** is computed by dividing each cell of the frequency table by the total in the same column of the frequency table and multiplying by 100 (Table 9.3). For example, the value in the "vwill"

column and "ltHS" row on the column-percentage table is computed by dividing the value in the "vwill" column and "ltHS" row of the frequency table (i.e., 40; Table 9.1) by the "Sum" of the "vwill" column of the 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	99.9	100.0	100.1	100.0	100.1

The value in each cell of a column-percentage table is the percentage OF ALL individuals in that column that 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.

Column-Percentage Table in R

A column-percentage table is constructed by submitting the saved `xtabs()` object to `percTable()` with `margin=2`. For example, the code below produced Table 9.3.

```
> percTable(tbl1,margin=2,digits=1)
```

9.2.3 Total-Percentage Table

Each value in a **total-percentage table** is computed by dividing each cell of the frequency table by the total number of ALL individuals in the frequency table and multiplying by 100. For example, the value in the "vwill" column and "ltHS" row of the table-percentage table (Table 9.4) is computed by dividing the value in the "vwill" column and "ltHS" row of the frequency table (i.e., 40; Table 9.1) by the "Sum" of the entire 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	7.5
Sum	5.3	26.5	24.4	25.6	18.0	99.8

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.

Total-Percentage Table in R

A table-percentage table is constructed by submitting the saved `xtabs()` object to `percTable()` and omitting `margin=`. For example, the code below produced Table 9.4.

```
> percTable(tbl1,digits=1)
```

9.3 Which Table to Use?

Determining which table to use comes from applying one simple rule and practicing with several tables. The rule comes 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 is not made, then the total-percentage table is used.

For example, consider the question – "What percentage of respondents with a bachelor's degree were very unwilling to cut their standard of living to protect the environment?" This question refers to only respondents with bachelor's degrees (i.e., "... of respondents with a bachelor's 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 would be used to answer this question. Therefore, 11.0% of respondents with bachelor's 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 total-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 refers only 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 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, so the total-percentage table would be used. In addition, this question is only concerned with one of the two variables; thus, the answer will come from a marginal distribution. 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.**

MODULE 10

BIVARIATE EDA - QUANTITATIVE

Contents

BIVARIATE DATA OCCURS WHEN TWO variables are 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 and biomass of legumes in fields. This module is focused on describing the bivariate relationship between two quantitative variables. Bivariate relationships between two categorical variables is described in Module 9.

Data on the *weight* (lbs) and highway miles per gallon (*HMPG*) for 93 cars from the 1993 model year are used as an example throughout this module. Ultimately, the relationship between highway MPG and the weight of a car is described. These data are read from 93cars.csv into R and several observations of *HMPG* and *weight* are shown below.[1]

```
> 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
```

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

10.1 Response and Explanatory Variables

The **response variable** is the variable that one is interested in explaining something (i.e., variability) or in making future predictions about. The **explanatory variable** is the variable that may help explain or allow one to predict the response variable. In general, the response variable is thought to depend on the explanatory variable. Thus, the response variable is often called the **dependent variable**, whereas the explanatory variable is often called the **independent variable**.

One may identify the response variable by determining which of the two variables depends on the other. For example, in the car data, highway MPG is the response variable because gas mileage is most likely affected by the weight of the car (e.g., hypothesize that heavier cars get worse gas mileage), rather than vice versa.

In some situations it is not obvious which variable is the response. For example, does the number of mice in the field depend on the number of legumes (lots of feed=lots of mice) or the other way around (lots of mice=not much food left)? Similarly, does area depend on depth or does depth depend on area of the lake? In these situations, the context of the research question is needed to identify the response variable. For example, if the researcher hypothesized that number of mice will be greater if there is more legumes, then number of mice is the response variable. In many cases, the more difficult variable to measure will likely be the response variable. For example, researchers likely wish to predict area of a lake (hard to measure) from depth of the lake (easy to measure).

◇ **Which variable is the response may depend on the context of the research question.**

10.2 Summaries

10.2.1 Scatterplots

A scatterplot is a graph where each point simultaneously represents the values of both the quantitative response and quantitative 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 example, the first individual in the cars data is plotted at x (*Weight*) = 2705 and y (*HMPG*) = 31, whereas the second individual is at x = 3560 and y = 25 (Figure 10.1).

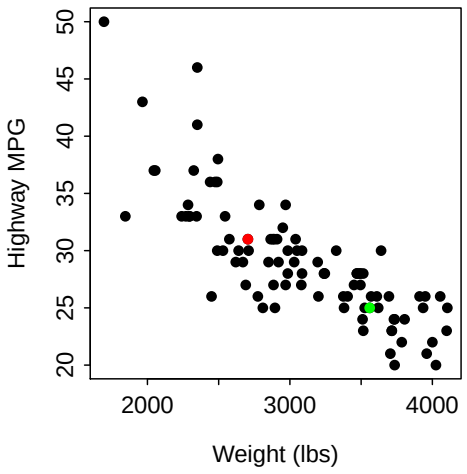


Figure 10.1. Scatterplot between the highway MPG and weight of cars manufactured in 1993. For reference to the main text, the first individual is red and the second individual is green.

Scatterplots are constructed in R with `plot()` with a formula of the form `Y~X`, where `Y` and `X` are variables to be plotted on the y- and x-axes, as the first argument, and the corresponding data.frame in `data=`. The x- and y-axis labels may be modified with `xlab=` and `ylab=`. The character plotted at each point can be changed with `pch=`,[2] which defaults to 1 or an open-circle (Figure 10.2). The scatterplot, excluding the two highlighted points, of highway MPG versus car weight (Figure 10.1) was created with the code below.

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

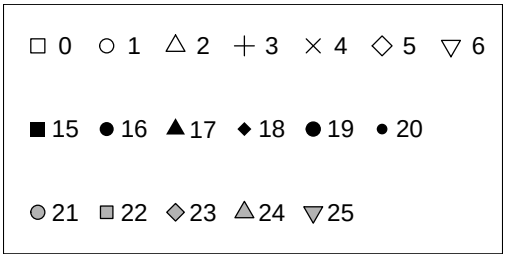


Figure 10.2. Plotting characters available in R and their numerical codes. Note that for values of 21-25 that `bg='gray70'` is used to provide the background color.

10.2.2 Correlation Coefficient

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

$$r = \frac{\sum_{i=1}^{n} \left[\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \right]}{n-1} \tag{10.2.1}$$

where s_x and s_y are the sample standard deviations for the explanatory and response variables, respectively.[3] The formulae in the two sets of parentheses in the numerator are standardized values;[4] thus, the value in each parenthesis is called the standardized x or standardized y, respectively. Using this terminology, Equation (10.2.1) 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 table below illustrates these calculations for the first five individuals in the cars data.[5] 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 are the standardized x values found by dividing the values 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.

[2] This argument is short for "plotting character".
[3] See Section 6.1.4 for a review of standard deviations.
[4] See Section 8.6 for a review of standardized values.
[5] The five cars are treated as if they are the entire sample.

i	HMPG y_i	Weight 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

The meaning and interpretation of r is discussed in more detail in Section 10.3.

The correlation coefficient (r) between two quantitative variables is computed with `corr()` using a formula of the form `Y~X` or `~Y+X`, where `Y` and `X` are the names of quantitative variables, as the first argument and the corresponding data.frame in `data=`. For example, the correlation coefficient between highway MPG and weight for all cars in the car data is -0.81.

```
> corr(HMPG~Weight,data=cars93)
[1] -0.8106581
> corr(~HMPG+Weight,data=cars93)  # alternative form
[1] -0.8106581
```

10.2.3 Pairs of Multiple Variables

Correlation coefficients can be computed or scatterplots can be constructed simultaneously for all pairs of many quantitative variables. A matrix of correlation coefficients is constructed with `corr()` as above using a formula of the form `~X1+X2+X3` (and so on), where the `X1`, `X2`, etc. are all quantitative variables to be used. In some instances, the data.frame may contain missing values (i.e., data that were not recorded). The individuals with missing data are efficiently removed from the correlation matrix with `use="pairwise.complete.obs"` in `corr()`.[6] The number of digits reported in the correlation matrix is controlled with `digits=`. For example, the correlation between highway MPG and size of the fuel tank is -0.786, whereas the correlation between length and weight of the car is 0.806.

```
> corr(~HMPG+FuelTank+Length+Weight,data=cars93,use="pairwise.complete.obs",digits=3)
          HMPG FuelTank Length Weight
HMPG      1.000   -0.786 -0.543 -0.811
FuelTank -0.786    1.000  0.690  0.894
Length   -0.543    0.690  1.000  0.806
Weight   -0.811    0.894  0.806  1.000
```

A matrix of scatterplots is constructed with `pairs()` using the same formula notation as in `corr()`. The plotting character can be changed, as with `plot()`, with `pch=`. Each subplot in the resulting scatterplot matrix (Figure 10.3) 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 10.3 has highway MPG on the y-axis and car weight on the x-axis.

```
> pairs(~HMPG+FuelTank+Length+Weight,data=cars93,pch=21,bg="gray70")
```

[6] Missing data are automatically removed from the scatterplots.

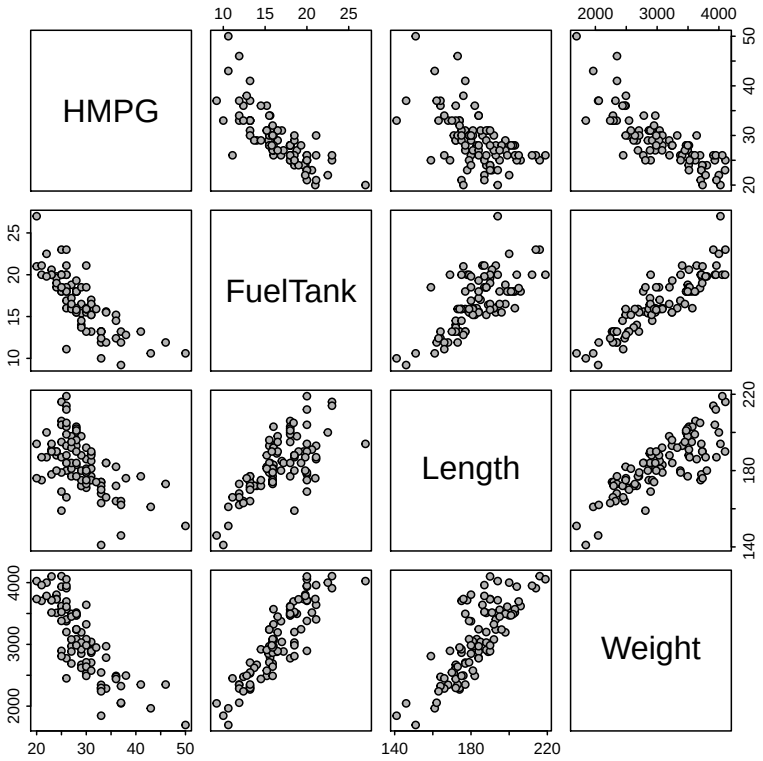


Figure 10.3. Scatterplot matrix of the highway MPG, fuel tank size, length, and weight of cars.

10.3 Items to Describe

Four characteristics should be described for a bivariate EDA with two quantitative variables:

1. **form** of the relationship,
2. presence (or absence) of **outliers**, and
3. **association** or **direction** of the relationship,
4. **strength** of the relationship.

All four of these items can be described from a scatterplot. However, for certain relationships (discussed below), strength is best described from the correlation coefficient.

10.3.1 Form and Outliers

The form of a relationship is determined by whether the "cloud" of points on a scatterplot forms a line or some sort of curve (Figure 10.5). For the purposes of this introductory course, if the "cloud" appears linear then the form will said to be linear, whereas if the "cloud" is curved then the form will be nonlinear. Scatterplots should be considered **linear** unless there is an OBVIOUS curvature in the points.

An outlier is a point that is 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.

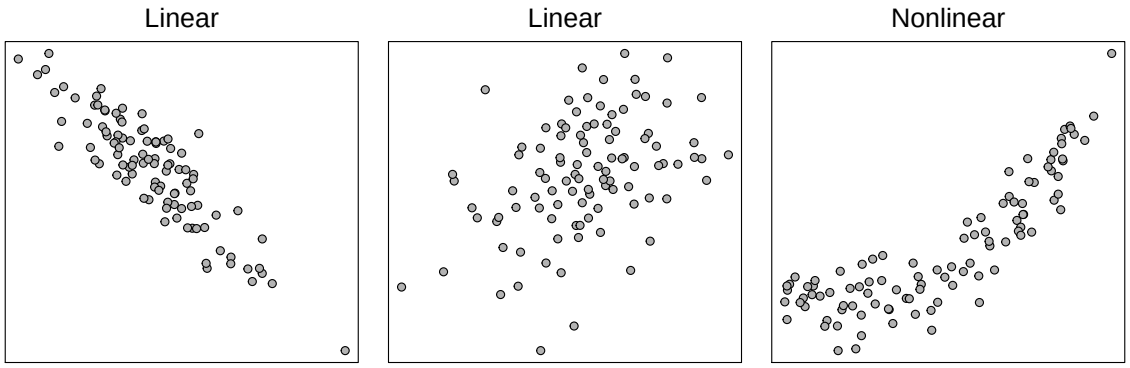


Figure 10.4. Depictions of two linear (Left and Center) and one nonlinear (Right) relationship.

10.3.2 Association or Direction

A positive association is when the scatterplot resembles an increasing function (i.e., increases from lower-left to upper-right; Figure 10.5-Left). For a positive association, most of the individuals are 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 10.5-Right). For a negative association, most of the individuals are above average for one variable and below average for the other variable. No association is when the scatterplot looks like a "shotgun blast" of points (Figure 10.5-Center). For no association, there is no tendency for individuals to be above or below average for one variable and above or below average for the other.

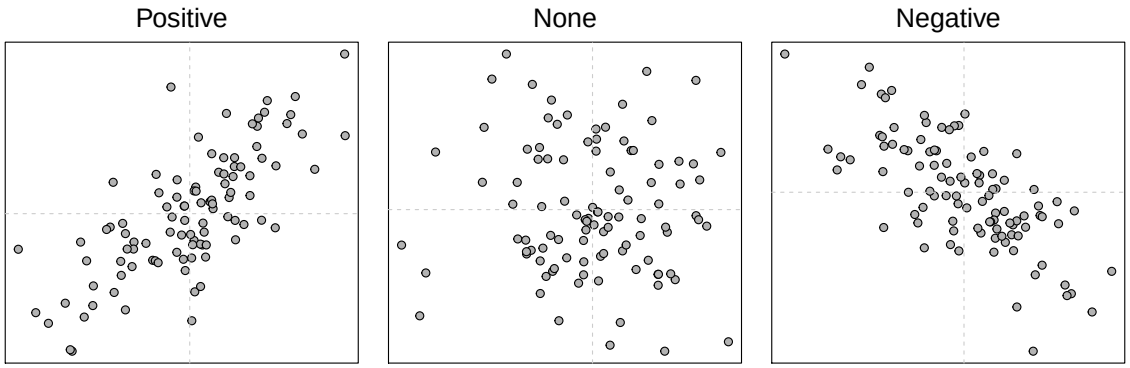


Figure 10.5. Depiction of three types of association present in scatterplots. Dashed vertical lines are at the means of each variable.

10.3.3 Strength (and Association, Again)

Strength is a summary of how closely the points cluster about the general form of the relationship. For example, if a linear form exists, then strength is how closely the points cluster around the line. Strength is difficult to define from a scatterplot because it is a relative term. However, the correlation coefficient (r; Section 10.2.2) is a measure of strength (and association) between two variables, *if the form is linear*.

The sign of r indicates the 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 strength of relationship. Absolute values nearer 1 are stronger relationships.

To better understand how r is a measure of association and strength, reconsider the steps in calculating r from Section 10.2.2. The scatterplots in Figure 10.6 represent a positive (Left) and negative (Right) association. These scatterplots have dashed lines at the mean of both the x- and y-axis variables. Because the mean is subtracted from observed values when standardizing, points that fall above the mean will have positive standardized values and points that fall below the mean will have negative standardized values. The sign for the standardized values are depicted along the axes.

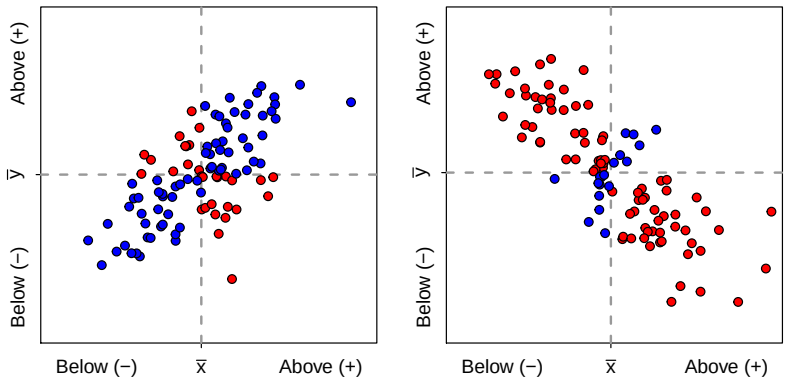


Figure 10.6. Scatterplot with mean lines superimposed and the signs of standardized values for both x and y shown for a positive (**Left**) and negative (**Right**) association. Blue points have a positive product of standardized values, whereas red points have a negative product of standardized values.

Now consider the product of standardized x's and y's in each quadrant of the scatterplots in Figure 10.6. The product of standardized values is positive (blue points) in the quadrant where both standardized values are above average (i.e., both positive signs) and both are below average. The product of standardized values is negative (red points) in the other two quadrants.

Thus, for a positive association (Figure 10.6-Left) the numerator of the correlation coefficient is positive because it is the sum of many positive (blue points) and few negative (red points) products of standardized values. The denominator (recall that it is $n-1$) is always positive. Therefore, r for a positive association is positive. Conversely, for a negative association (Figure 10.6-Right) the numerator of the correlation coefficient is negative because it is the sum of few positive (blue points) and many negative (red points) products of standardized values. Therefore, r for a negative association is negative.

Correlations range from -1 to 1. Absolute values of r equal to 1 indicate a perfect association (i.e., all points 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. How strength and association of the relationship changes along the range of r values is illustrated in Figure 10.7. Categorizations in Table 10.1 can be used as a guideline for describing the strength of relationship between two variables.

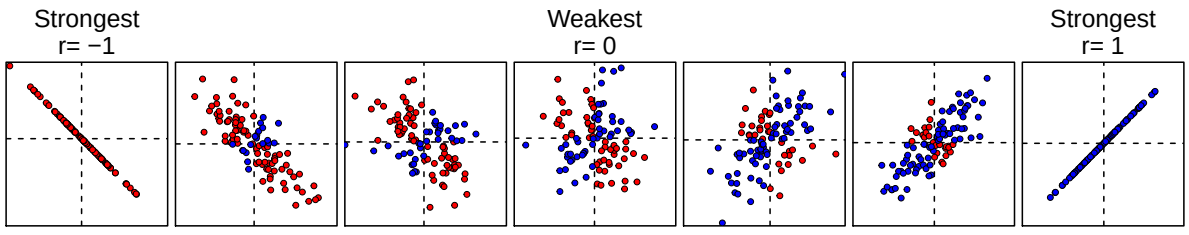


Figure 10.7. Scatterplots along the continuum of r values.

Table 10.1. 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

10.4 Example Interpretations

When performing a bivariate EDA for two quantitative variables, the form, presence (or absence) of outliers, association, and strength should be specifically addressed. In addition, you should state how you assessed strength. Specifically, you should use r to assess strength (see Section 10.3.3) **IF** the relationship is linear without any outliers. However, if the relationship is nonlinear, has outliers, or both, then strength should be subjectively assessed from the scatterplot.

Two other points to consider when performing a bivariate EDA with quantitative variables. First, if outliers are present, do not let them completely influence your conclusions about form, association, and strength. In other words, assess these items ignoring the outlier(s). If you have raw data and the form excluding the outlier is linear, then compute r with the outlier eliminated from the data. Second, the form of weak relationships is difficult to describe because, by definition, there is very little clustering to a form. As a rule-of-thumb, if the scatterplot is not obviously curved, then it is described as linear by default.

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

◇ The form is linear unless there is an OBVIOUS curvature.

Highway MPG and Weight

The following overall bivariate summary for the relationship between highway MPG and weight is made using the calculations from the previous sections.

The relationship between highway MPG and weight of cars (Figure 10.1) appears to be primarily linear (although I see a very slight concavity), negative, and moderately strong with a correlation of -0.81. 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). The correlation coefficient was used to assess strength because I deemed the relationship to be linear without any outliers.

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 linear, with two outliers at total petroleum levels greater than 3000 trillions of BTU, positive, and weak (Figure 10.8-Left). I did not use the correlation coefficient because of the outliers. If the two outliers (Texas and California) are removed then the relationship is linear, with no additional outliers, positive, and weak (r =0.53) (Figure 10.8-Right).

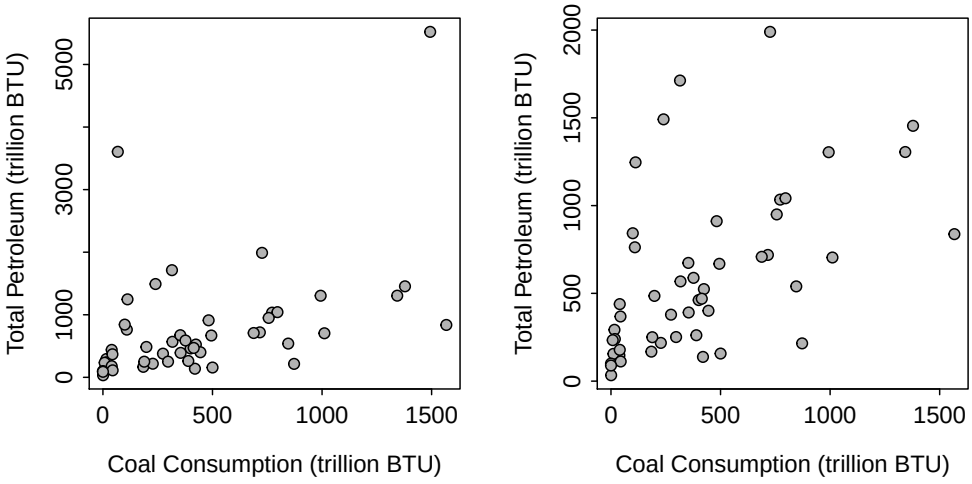


Figure 10.8. Scatterplot of the total consumption of petroleum versus the consumption of coal (in trillions of BTU) by all 50 states and the District of Columbia. The points shown in the left with total petroleum values greater than 3000 trillion BTU are deleted in the right plot.

R Appendix

```
NRG <- read.csv("data/NRG_Consump_2001.csv")
NRG1 <- NRG[-c(5,44),]
plot(TotalPet~Coal,data=NRG,pch=21,bg="gray70",xlab="Coal Consumption (trillion BTU)",
     ylab="Total Petroleum (trillion BTU)")
plot(TotalPet~Coal,data=NRG1,pch=21,bg="gray70",xlab="Coal Consumption (trillion BTU)",
     ylab="Total Petroleum (trillion BTU)")
corr(~Coal+TotalPet,data=NRG1)
```

Hatch Weight and Incubation Time of Geckos

A hobbyist hypothesized that there would be a positive association between length of incubation (days) and hatchling weight (grams) for Crested Geckos (Rhacodactylus ciliatus). To test this hypothesis she collected the incubation time and weight for 21 hatchlings (shown below). Construct a proper EDA for the relationship between incubation time and hatchling weight.

Time	53	54	56	60	60	60	60	60	63	63	77	77	78	81	82	82	83	83	84	90	90
Wt	1.5	1.7	1.4	1.0	1.4	1.5	1.7	1.8	1.4	1.5	1.1	1.6	1.5	1.9	1.4	1.5	1.3	1.7	1.6	1.4	1.8

The relationship between hatchling weight and incubation time for the Crested Geckos is linear, without obvious outliers (*though some may consider the small hatchling at 60 days to be an outlier*), without a definitive association, and weak (r=0.11) (Figure 10.9). I did compute r because no outliers were present and the relationship was linear (or, at least, it was not nonlinear).

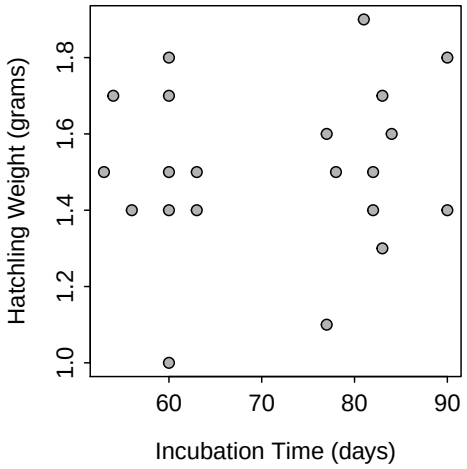


Figure 10.9. Scatterplot of hatchling weight versus incubation time for Crested Geckos.

R Appendix

```
df <- read.csv("data/Gecko.csv")
plot(hatchwt~inctime,data=df,pch=21,bg="gray70",xlab="Incubation Time (days)",
     ylab="Hatchling Weight (grams)")
corr(~inctime+hatchwt,data=df)
```

10.5 Cautions About Correlation

Examining relationships between pairs of quantitative variables is common practice. Using r can be an important part of this analysis, as described above. However, r can be abused through misapplication and misinterpretation. Thus, it is important to remember the following characteristics of correlation coefficients:

- Variables must be quantitative (i.e., if you cannot make a scatterplot, then you cannot calculate r).
- The correlation coefficient only measures strength of **LINEAR** relationships (i.e., if the form of the relationship is not linear, then r is meaningless and should not be calculated).

- The units that the variables are measured in do not matter (i.e., r is the same between heights and weights measured in inches and lbs, inches and kg, m and kg, cm and kg, and cm and inches). This is because the variables are standardized when calculating r.
- The distinction between response and explanatory variables is not needed to compute r. That is, the correlation of GPA and ACT scores is the same as the correlation of ACT scores and GPA.
- Correlation coefficients are between -1 and 1.
- Correlation coefficients are strongly affected by outliers (simply, because both the mean and standard deviation, used in the calculation of r, are strongly affected by outliers).

Additionally, correlation is not causation! In other words, just because a strong correlation is observed it does not mean that the explanatory variable caused the response variable (an exception may be in carefully designed experiments). For example, it was found above that highway gas mileage decreased linearly as the weight of the car increased. One must be careful here to not state that increasing the weight of the car CAUSED a decrease in MPG because these data are part of an observational study and several other important variables were not considered in the analysis. For example, the scatterplot in Figure 10.10, coded for different numbers of cylinders in the car's engine, indicates that the number of cylinders may be inversely related to highway MPG and positively related to weight of the car. So, does the weight of the car, the number of cylinders, or both, explain the decrease in highway MPG?

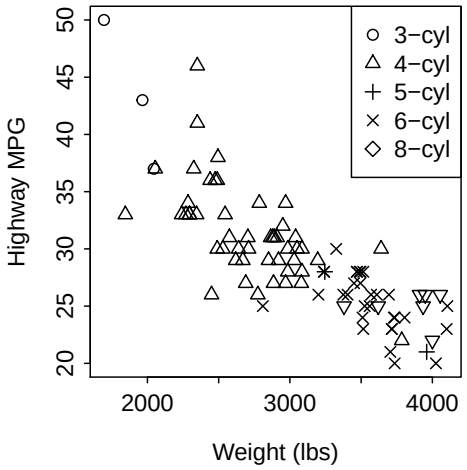


Figure 10.10. Scatterplot between the highway MPG and weight of cars manufactured in 1993 separated by number of cylinders.

More interesting examples (e.g., high correlation between number of people who drowned by falling into a pool and the annual number of films that Nicolas Cage appeared in) that further demonstrate that "correlation is not causation" can be found on the Spurious Correlations website.

Finally, the word "correlation" is often misused in everyday language. "Correlation" should only be used when discussing the actual correlation coefficient (i.e., r). When discussing the association between two variables, one should use "association" or "relationship" rather than "correlation." For example, one might ask "What is the relationship between age and rate of cancer?", but should not ask (unless specifically interested in r) "What is the correlation between age and rate of cancer?".

MODULE 11

LINEAR REGRESSION

Contents

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 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, and a person's score on a German assessment test based on how many years the person studied German.

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 11.1 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 is accomplished by finding the line that best "fits" the points on a scatterplot of the data and using that line to make predictions. Finding and using the "best-fit" line is the topic of this module.

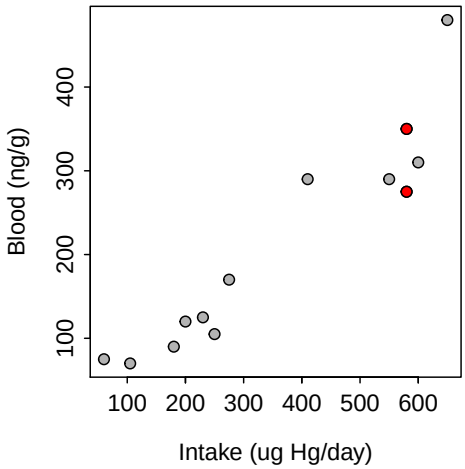


Figure 11.1. Scatterplot of intake of mercury in fish and the mercury in the blood stream. The two individuals mentioned in the main text are shown as red points.

11.1 Response and Explanatory Variables

Recall from Section 10.1 that the response (or dependent) variable is the variable to be predicted or explained and the explanatory (or independent) variable is the variable that will help do the predicting or explaining. In the examples mentioned above, future sales, family expenditures on recreation, the animal's food consumption, and score on the assessment test are response variables and product price, family income, temperature, and years studying German are explanatory variables, respectively. The response variable is on the y-axis and the explanatory variable is on the x-axis of scatterplots.

11.2 Slope and Intercept

The equation of a line is commonly expressed as,

$$y = mx + b$$

where both x and y are variables, m represents the slope of the line, and b represents the y-intercept.[1] 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, in

$$blood = 3.501 + 0.579 * intake$$

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 would be made if the equation had been written as

$$blood = 0.579 * intake + 3.501$$

[1] Hereafter, simply called the "intercept."

◇ 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 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 at 101), then $blood = 3.501 + 0.579 * 101 = 61.98$.[2] The difference between these two values is 61.98-61.40=0.579. Thus, the slope is the change in value of the response variable for a single unit change in the value of the explanatory variable (Figure 11.2). 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 units of mercury in the blood.

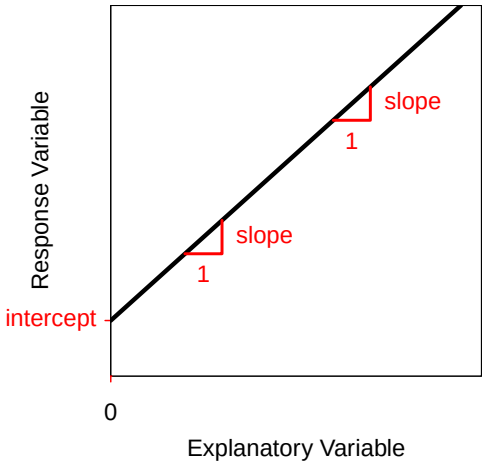


Figure 11.2. Schematic representation of the meaning of the intercept and slope in a linear equation.

To define the intercept, first "plug" $intake = 0$ into the equation discussed above; i.e., $blood = 3.501 + 0.579 * 0 = 3.501$. Thus, the intercept is the value of the response variable when the explanatory variable is equal to zero (Figure 11.2). In this example, the AVERAGE mercury in the blood for an individual with no mercury intake is 3.501. 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*.

[2] For simplicity of exposition, the actual units are not used in this discussion. However, "units" would usually be replaced with the actual units used for the measurements.

11.3 Predictions

Once a best-fit line has been identified (criteria for doing so is discussed in Section 11.5), 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 11.1 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

$$blood = 3.501 + 0.579 * 575 = 336.426$$

A prediction may be visualized by finding the value of the explanatory variable on the x-axis, drawing a vertical line until the best-fit line is reached, and then drawing a horizontal line over to the y-axis where the value of the response variable is read (Figure 11.3).

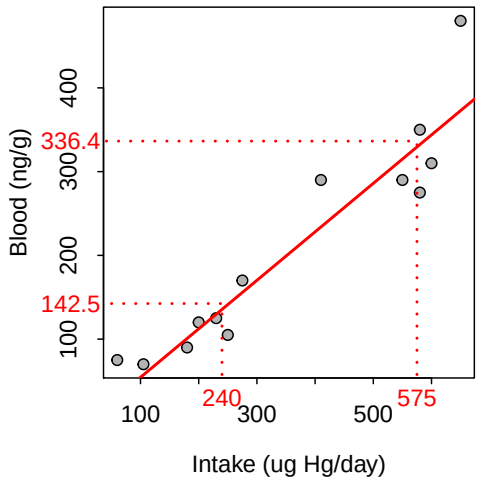


Figure 11.3. Scatterplot between the intake of mercury in fish and the mercury in the blood stream of individuals with superimposed best-fit regression line illustrating predictions for two values of mercury intake.

When predicting values of the response variable, it is important to not extrapolate beyond the range of the data. In other words, predictions with values outside the range of observed values of the explanatory variable should be made cautiously (if at all). An excellent example would be to consider 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. However, using this relationship to predict an individual's height at age 40 would likely result in a ridiculous answer (e.g., over ten feet). The problem here is that the linear relationship only holds for the observed data (i.e., the first ten years of life); 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.

11.4 Residuals

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. Specifically, the residual for an individual is found 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. As shown in the previous section, the predicted level of mercury in the blood for this individual is

$$blood = 3.501 + 0.579 * 650 = 379.851$$

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 the average for individuals with an intake of 650.[3] As a second example, consider an 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

$$blood = 3.501 + 0.579 * 250 = 148.251$$

and the residual is $105 - 148.251 = -43.251$. This negative residual indicates that the observed value is approximately 43 units less than the average for individuals with an intake of 250.

Visually, a residual is the vertical distance between an individual's point and the best-fit line (Figure 11.4).

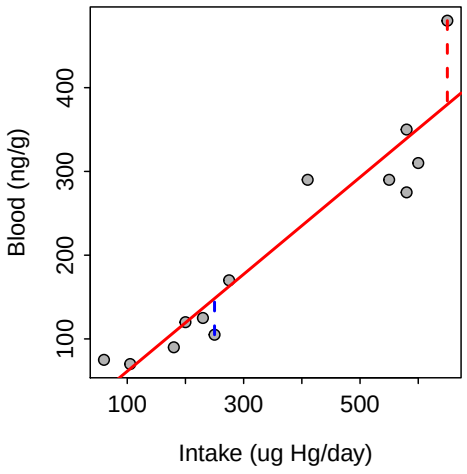


Figure 11.4. Scatterplot between the intake of mercury in fish and the mercury in the blood stream of individuals with superimposed best-fit regression line illustrating the residuals for the two individuals discussed in the main text.

[3] In other words, the observed value is "above" the line.

11.5 Best-fit Criteria

An infinite number of lines can be placed on a graph, but 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, how does one find THE best-fit line from all possible lines. The **least-squares** method described below provides a quantifiable and objective measure of which line best "fits" the data.

Residuals are a measure of how far an individual is from a candidate best-fit line. Residuals computed from all individuals in a data set measure how far all 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 the sum of all squared residuals. The least-squares criterion says that the "best-fit" line is the one line out of all possible lines that has the minimum RSS (Figure 11.5).

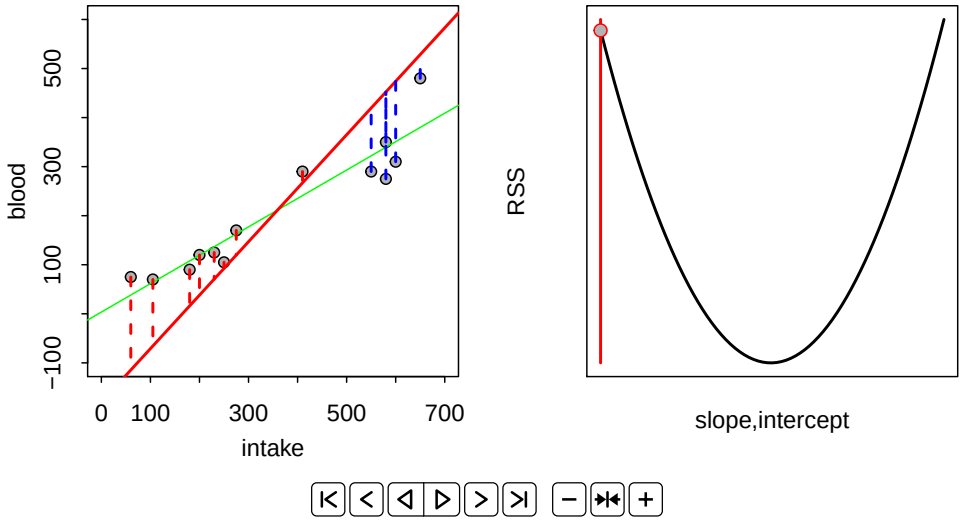


Figure 11.5. An animation illustrating how the residual sum-of-squares (RSS) for a series of candidate lines (red lines) is minimized at the best-fit line (green line).

The discussion thusfar 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

$$slope = r\frac{s_y}{s_x}$$

and an intercept given by

$$intercept = \bar{y} - slope * \bar{x}$$

where $\bar{x}$ and s_x are the sample mean and standard deviation of the explanatory variable, $\bar{y}$ and s_y are the sample mean and standard deviation of the response variable, and r is the sample correlation coefficient between the two variables. Thus, using these formulas finds the slope and intercept for the line, out of all possible lines, that minimizes the RSS.

11.6 Assumptions

The least-squares method for finding the best-fit line only works appropriately if each of the following five assumptions about the data has been met.

1. A line describes the data (i.e., a linear form).
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 data.[4] Description is the primary goal of linear regression used in this course and, thus, only the first two assumptions are considered further.

The linearity assumption appears 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 homoscedasticity assumption states that the variability about the line is the same for all values of the explanatory variable. In other words, the dispersion of the data around the line must be the same along the entire line. Violations of this assumption generally present as a "funnel-shaped" dispersion of points from left-to-right on a scatterplot.

Violations of these assumptions are often evident on a "fitted-line plot", which is a scatterplot with the best-fit line superimposed (Figure 11.6).[5] 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. A violation of one of these assumptions should be obvious on the scatterplot. In other words, there should be a clear curvature or funneling on the plot.

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.**

[4] In contrast to using the model to make inferences about a population model.
[5] Residual plots, not discussed in this text, are another plot that often times is used to better assess assumption violations.

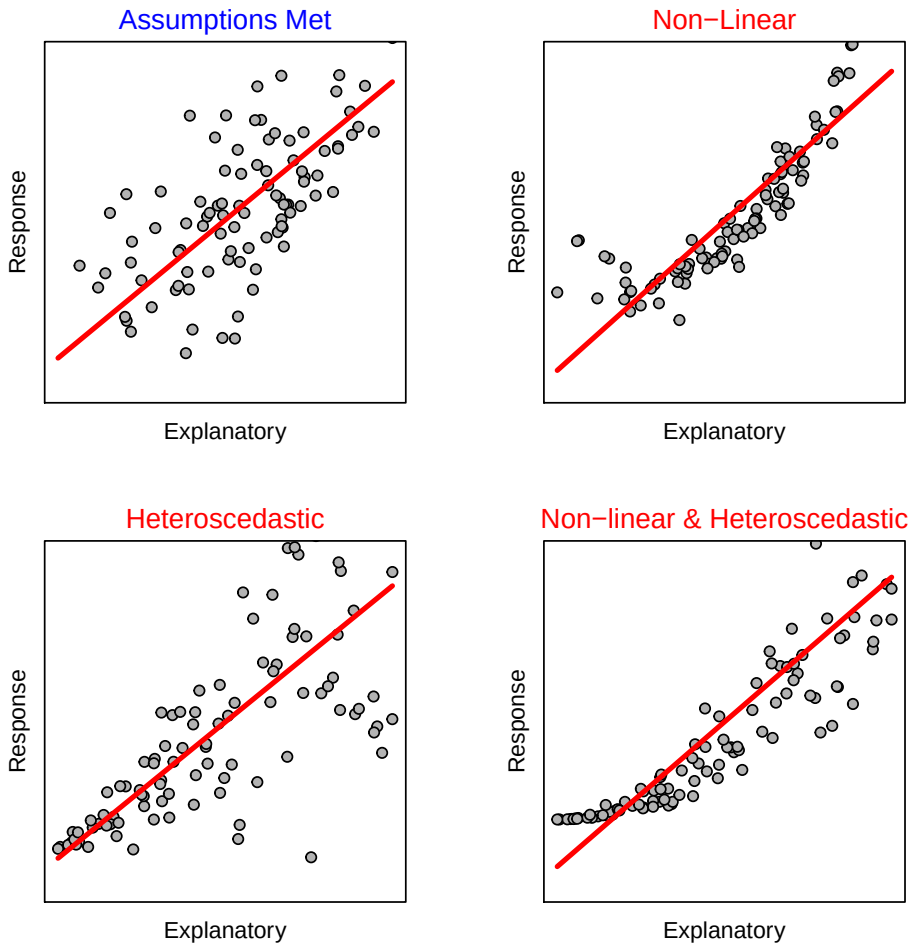


Figure 11.6. Fitted-line plots illustrating when the regression assumptions are met (upper-left) and three common assumption violations.

11.7 Coefficient of Determination

The coefficient of determination (r^2) is the proportion of the total variability in the response variable that is explained away by knowing the value of the explanatory variable and the best-fit model. In simple linear regression, r^2 is literally the square of r, the correlation coefficient.[6] Values of r^2 are between 0 and 1.[7]

The meaning of r^2 can be examined by making predictions of the response variable with and without knowing the value of the explanatory variable. First, consider predicting the value of the response variable without any information about the explanatory variable. In this case, the best prediction is the sample mean of the response variable (represented by the dashed blue horizontal line in Figure 11.7). However, because of natural variability, not all individuals will have this value. Thus, the prediction might be "bracketed" by predicting that the individual will be between the observed minimum and maximum values (solid blue horizontal lines). Loosely speaking, this range is the "total variability in the response variable" (blue box).

[6] Simple linear regression is the fitting of a model with a single explanatory variable and is the only model considered in this module and this course. See Section 10.2.2 for a review of the correlation coefficient.
[7] It is common for r^2 to be presented as a percentage.

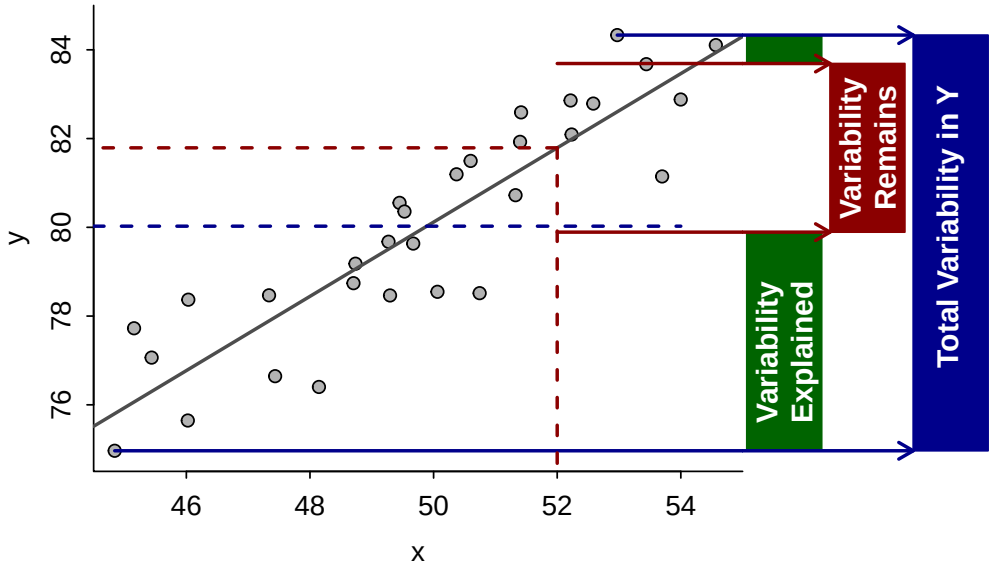


Figure 11.7. Fitted line plot with visual representations of variabilities explained and unexplained. A full explanation is in the text.

Suppose now that the response variable is predicted for an individual with a known value of the explanatory variable (e.g., at the dashed vertical red line in Figure 11.7). 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 same value of the explanatory variable (solid red horizontal lines). Loosely speaking, this range is 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 11.7).

Now, by the definition of r^2, 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., homoscedasticity – see Section 11.6).

If the variability explained away (green box) approaches the total variability in the response variable (blue box), then r^2 approaches 1. This will happen only if the variability around the line approaches zero. In contrast, the variability explained (green box) will approach zero if the slope is zero (i.e., 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.**

11.8 Examples I

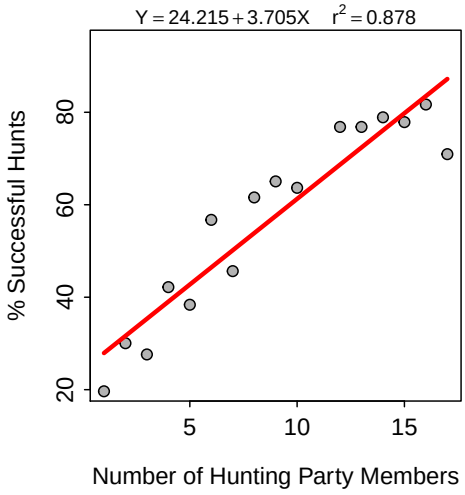
There are twelve questions that are commonly asked about linear regression results. These twelve questions are listed below with some hints about things to remember when answering some of the questions. An example of these questions in context is then provided.

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 intercept, 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. Make sure that this is not an extrapolation.*
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. Make sure that the prediction is not an extrapolation.*
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 below.[8] *Use these results to answer the questions below.*



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. Additionally, 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 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 the percent of successful hunts increases by 3.705, on average, for every increase of one member to the hunting party.

Q: Interpret the value of the intercept in terms of the variables of the problem.

A: The intercept indicates that the percent of successful hunts is 24.215, on average, for hunting parties with no members.

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 number of members observed on the x-axis of the fitted-line plot.

Q: What is the predicted hunt success if the hunting party consists of 12 chimpanzees?

[8] These data are in Chimp.csv.

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.**

11.9 Regression in R

The mercury intake and amount in the blood data is loaded below to be used as an example for finding a regression line with R.

```
> 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 linear regression model is fit to two quantitative variables with `lm()`. The first argument is a formula of the form `response~explanatory`, where `response` contains the response variable and `explanatory` contains the explanatory variable, and the corresponding data.frame is in `data=`. The results of `lm()` should be assigned to an object so that specific results can be extracted.

◇ **The same formula used to make a scatterplot with `plot()` is used in `lm()` to find the best-fit line.**

The regression was fit to the mercury data below. From this it is seen that the intercept is 3.501 and the slope is 0.579.

```
> ( lm1 <- lm(blood~intake,data=merc) )
Coefficients:
(Intercept)       intake
     3.5007       0.5791
```

A fitted-line plot (Figure 11.8) is constructed by submitting the `lm()` object to `fitPlot()`. Aspects of this plot can be adjusted using the same arguments as described for `plot()` in Section 10.2.1.

```
> fitPlot(lm1,pch=21,bg="gray70",xlab="Mercury Intake",ylab="Mercury in the Blood")
```

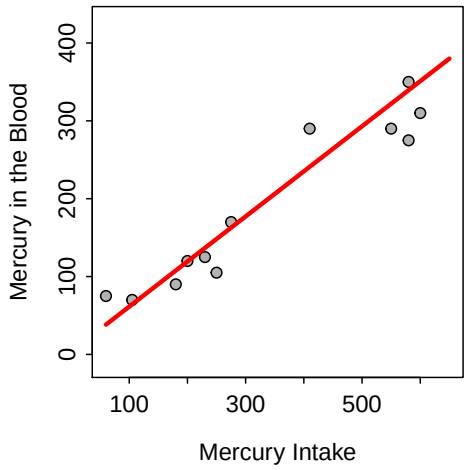


Figure 11.8. Fitted-line plots for the regression of mercury in the blood on mercury intake.

Predicted values from the linear regression are 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 142.5, as obtained below.

```
> predict(lm1,data.frame(intake=240))
       1
142.4895
```

The coefficient of determination is computed by submitting the saved `lm()` object to `rSquared()`. For example, 88.4% of the variability in mercury in the blood is explained by knowing the amount of mercury at intake. [Note the use of `digits=` to control the number of decimals.]

```
> rSquared(lm1,digits=3)
[1] 0.884
```

11.10 Examples II

Car Weight and MPG

In Module 10, 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 twelve typical questions. These data are read into R below and the linear regression model is fit, coefficients extracted, fitted-line plot constructed, and coefficient of determination extracted.

```
> cars93 <- read.csv("data/93cars.csv")
```

```
> ( 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
```

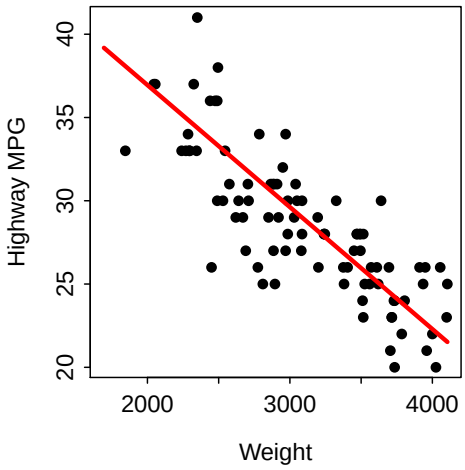


Figure 11.9. Fitted line plot of the regression of highway MPG on weight of 93 cars from 1993.

The simple linear regression model appears to fit the data moderately well as the fitted-line plot (Figure 11.9) shows only a very slight curvature and only very slight heteroscedasticity.[9] The sample slope is -0.0073, the sample intercept is 51.6, 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.

[9] 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.

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 HMPG = 51.6 - 0.0073Weight.

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 -0.0073, on average.

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 average.[10]

Q: What is the predicted highway MPG for a car that weighs 3100 lbs?

A: The predicted highway MPG for a car that weighs 3100 lbs is 51.60137 - 0.00733(3100) = 28.9 MPG. Alternatively, this value is computed 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 11.9).

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.[11]

Q: How much is the highway MPG expected to change if a car is 1000 lbs heavier?

A: If the car was 1000 lbs heavier, you would expect the car's highway MPG to decrease by 7.33 (i.e., 1000 slopes).

[10] This is the correct interpretation of the intercept. However, it is nonsensical because it is an extrapolation; i.e., no car will weigh 0 pounds.

[11] 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.

MODULE 12

SAMPLING DISTRIBUTIONS

Contents

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 difficult becauses statistics depend on the specific individuals in the sample and, thus, vary from sample to sample. For example, recall from Section 2.1 that the mean length of fish differed among four samples "taken" from Square Lake. Thus, to make conclusions about the population from the sample, the distribution (i.e., shape, center, and dispersion) of the statistic computed from all possible samples must be understood.[1] In this module, the distribution of statistics from all possible samples is explored and generalizations are defined that can be used to make inferences. In subsequent modules, this information, along with results from a single sample, will be used to make specific inferences about the population.

◇ **Statistical inference requires considering sampling variability.**

[1] See Module 1 for a review of sampling variability.

12.1 What is a Sampling Distribution?

12.1.1 Definitions and Characteristics

A **Sampling distribution** is the distribution of values of a particular statistic computed from all possible samples of the same size from the same population. The discussion of sampling distributions and all subsequent theories related to statistical inference are based on repeated samples 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. Thus, it is important to note that only one sample is ever actually taken from a population.

The concept of a sampling distribution is illustrated with a population of six students that 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 that the sample mean is computed for each sample (Table 12.1).[2] The sampling distribution of the sample mean from samples of $n = 2$ from this population (Figure 12.1) is the histogram of means from these 15 samples.[3]

Table 12.1. All possible samples of $n = 2$ and corresponding sample mean from the quiz score population.

Scores	Mean	Scores	Mean	Scores	Mean	Scores	Mean	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

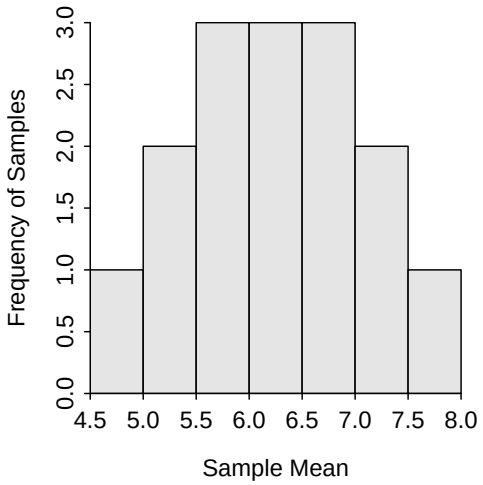


Figure 12.1. Sampling distribution of mean quiz scores from samples of $n = 2$ from the quiz score population.

The mean (=6.000) and standard deviation (=0.845) of the 15 sample means are measures of center and dispersion for the sampling distribution. The standard deviation of statistics (i.e., 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 keep the dispersion of the sampling distribution separate from the dispersion of individuals in the population, which is measured by the standard deviation. Thus, the standard deviation

[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)`.

of all possible sample means is referred to as the standard error of the sample means (SE). The SE in this example is 0.845. The standard deviation is the dispersion of individuals in the population and, in this example, is 1.414.

This example illustrates three major concepts concerning sampling distributions. First, the sampling distribution will more closely resemble a normal distribution than the original population distribution (unless, of course, the population distribution was normal).

Second, the center (i.e., mean) of the sampling distribution 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, the mean of all possible sample means (= 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 estimate of the population mean.

Third, the standard error of the statistic is 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.

◇ **All statistics in this course are unbiased.**

12.1.2 Critical Distinction

Three distributions are considered in statistics. The sampling distribution is the distribution of a statistic computed from all possible samples of the same size from the same population, the population distribution is the distribution of all individuals in a population (see Module 8), and the sample distribution is the distribution of all individuals in a sample (see histograms in Module 6). The sampling distribution is about **statistics**, whereas the population and sample distributions are about **individuals**. For inferential statistics, it is important to distinguish between population and sampling distributions. Keep in mind that one (population) is the distribution of individuals and the other (sampling) is the distribution of statistics.

Just as importantly, 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 dispersion of all individuals in a sample. In contrast, the standard error measures 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.

12.1.3 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 any other sample size is used. For example, the samples and means from each sample of $n = 3$ are 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 12.1.1 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]

[4] One should also look at the results from $n = 4$ in one of the online Review Exercises.

Table 12.2. All possible samples of $n = 3$ and corresponding sample means from the quiz score population.

Scores	Mean	Scores	Mean	Scores	Mean	Scores	Mean	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

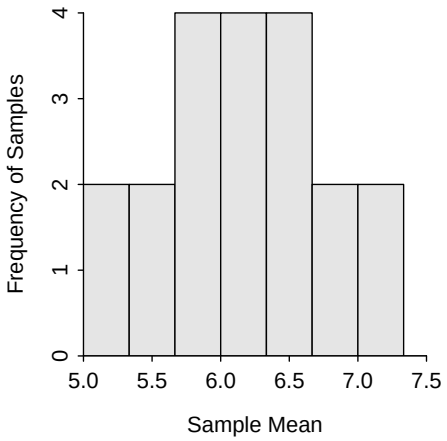


Figure 12.2. Sampling distribution of mean quiz scores from samples of $n = 3$ from the quiz score population.

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 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 corresponding sample medians from the quiz score population.

Scores	Median	Scores	Median	Scores	Median	Scores	Median	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

These examples demonstrate that the naming of a sampling distribution must be specific. For example, the first sampling distribution in this module should be described as the "sampling distribution of sample means from samples of n=2." This last example should be described as the "sampling distribution of sample medians from samples of n=3." Doing this with each distribution reinforces 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.**

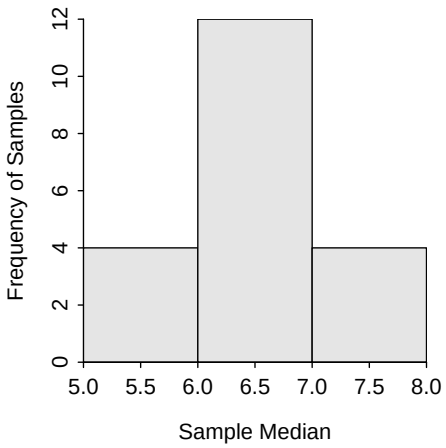


Figure 12.3. Sampling distribution of median quiz scores from $n = 3$ samples from the quiz score population.

12.1.4 Simulating

Exact sampling distributions can only be computed for very small samples taken from a small population. Exact sampling distributions are difficult to show for even moderate 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 and even larger situations?

There are two ways to examine sampling distributions in situations with large sample and population sizes. First, theorems exist that describe the specifics of sampling distributions under certain conditions. One such theorem is described in Section 12.2. Second, 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." We will simulate some sampling distributions here so that the theorem will be easier to understand.

Sampling distributions are simulated by drawing many samples from a population, computing the statistic of interest for each sample, and constructing a histogram of those statistics (Figure 12.4). The computer is helpful with this simulation; however, keep in mind that the computer is basically following the same process as used in Section 12.1.1, with the exception that not every sample is taken.

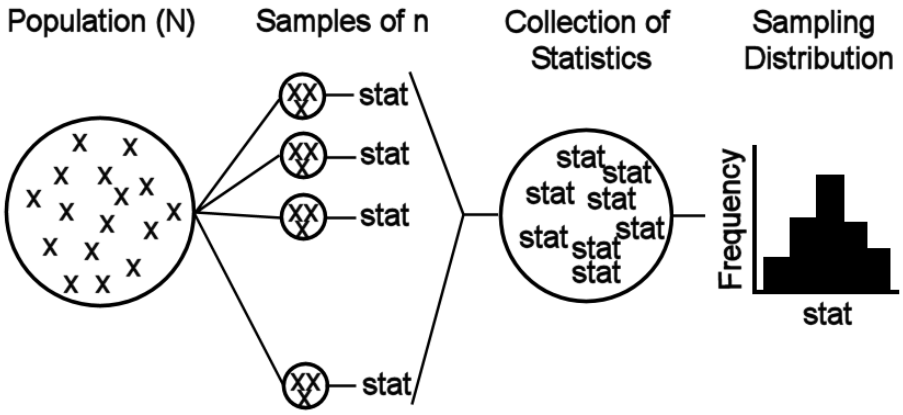
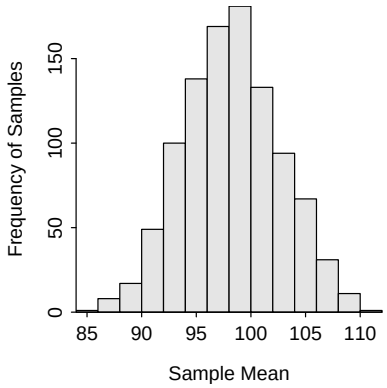


Figure 12.4. Schematic representation of the process for simulating a sampling distribution.

Let's return to the Square Lake fish population from Section 2.1 to illustrate simulating a sampling distribution. 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 a simulated sampling distribution of sample means because it represents the distribution of sample means from 1000, rather than all possible, samples.



	Means
mean	98.25
sd	4.43
min	85.46
Q1	95.14
median	98.26
Q3	101.27
max	110.1

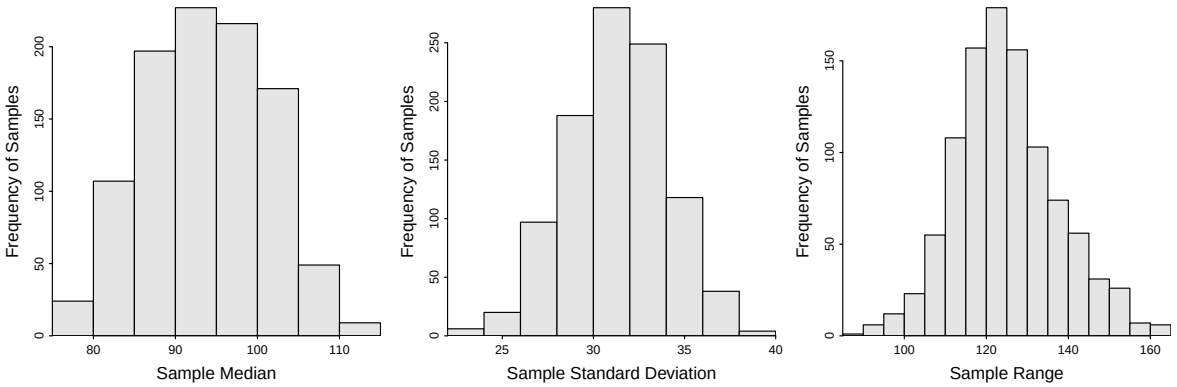
Figure 12.5. Histogram (**Left**) and summary statistics (**Right**) from 1000 sample mean total lengths computed from samples of $n = 50$ from the Square Lake fish population.

As with the actual sampling distributions discussed previously, three characteristics (shape, center, and dispersion) are examined with simulated sampling distributions. First, Figure 12.5 looks at least approximately normally distributed. Second, the mean of the 1000 means (=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" rather than all possible samples. 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).

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 module forward, though, remember that sampling distributions are simulated primarily 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 tool for checking the theory concerning sampling distributions; however, in "real-life" only one sample from the population is needed.**



	Medians
mean	93.45
sd	7.34
min	75.5
max	112.5
Parameter	93

	Std. Devs
mean	31.27
sd	2.76
min	22.98
max	39.32
Parameter	31.49

	Ranges
mean	124.53
sd	12.7
min	88
max	164
Parameter	164

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.

12.2 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 12.1.1) or taking a large number of samples from a large population (Section 12.1.4). 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 to identify 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

- $n \geq 30$,
- $n \geq 15$ and the population distribution is not strongly skewed, **or**
- the population distribution is normally distributed.

Thus, the sampling distribution of $\bar{x}$ should be normally distributed **no matter what the shape of the population distribution is** as long as $n \geq 30$. 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 standard error.

The validity of the CLT can be examined by simulating several (with different n) sampling distributions of $\bar{x}$ from the Square Lake population and from a strongly right-skewed exponential distribution (Figure 12.7). Several observations about the CLT can be made from Figure 12.7. First, the sampling distribution is approximately normal for $n \geq 30$ for both scenarios and is approximately normal for smaller n for the Square Lake example because that population is only slightly skewed. Second, the means of all sampling

distributions in both examples are approximately equal to μ, regardless of n. Third, the dispersion of the sampling distributions (i.e., the SE of the means) becomes smaller with increasing n. Furthermore, the SE from the simulated results closely match the SE expected from the CLT.

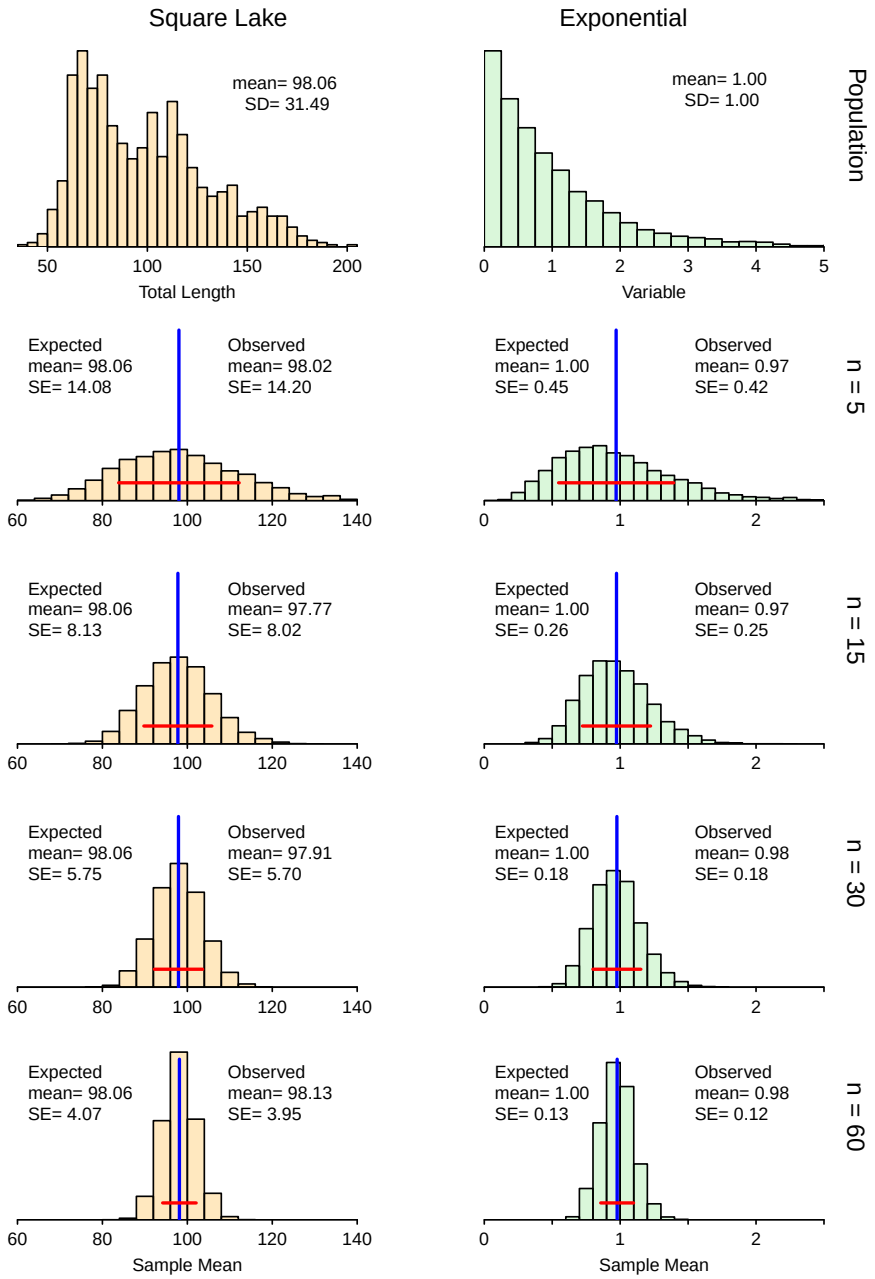


Figure 12.7. Sampling distribution of the sample mean simulated from 5000 samples of four different sample sizes extracted from the Square Lake fish population (Left) and an exponential population (Right). The shapes of the populations are shown in the top histogram. On each simulated sampling distribution, the vertical blue line is the mean of the 5000 means and the horizontal red line represents ±1SE from the mean.

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 similar statistics. The standard error is a measure of precision; i.e., a high SE means low precision and a low SE means high precision.

The targets in Figure 12.8 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 (i.e., 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. Targets in which the red dots are closely clustered are considered precise. Similarly, sampling distributions that exhibit little variability (low dispersion) are considered precise.

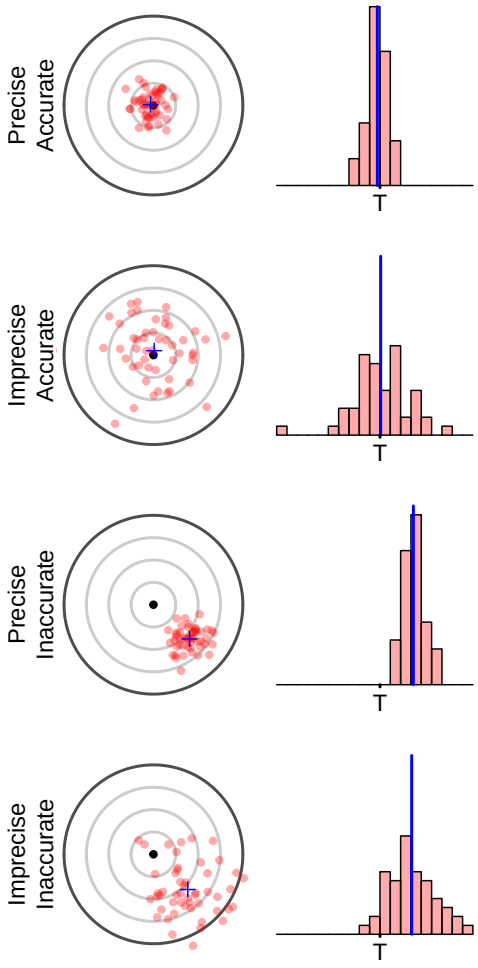


Figure 12.8. The center of each target (i.e., the bullseye) and the point marked with a "T" (for "truth") represent the parameter of interest. Each dot on the target represents a statistic computed from a single sample and, thus, the many red dots on each target represent repeated samplings from the 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.

MODULE 13

PROBABILITY INTRODUCTION

PROBABILITY is the "language" used to describe the proportion of times that a random event will occur. The language of probability is at the center of statistical inference (see Modules 14 and 16). Only a minimal understanding of probability is required to understand most basic inferential methods, including all of those in this course. Thus, only a short, example-based, introduction to probability is provided here.[1]

13.1 Probability of Individuals

The most basic forms of probability assume that items are selected 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 individuals were "thoroughly mixed" and more realistic examples will require randomization.[2]

If every item 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"); Figure 13.1-Left). Similarly, the probability of randomly selecting a woman from a room with 20 women and 30 men is 0.4 ($= \frac{20}{50}$; Figure 13.1-Right). In both examples, the calculation can be considered a probability because (i) individuals were randomly selected and (ii) a proportion of a total was computed.

[1] A deeper understanding of probability is required to understand more complex inferential methods beyond those in this course.
[2] See Module 3 for methods to randomly select or allocate individuals.

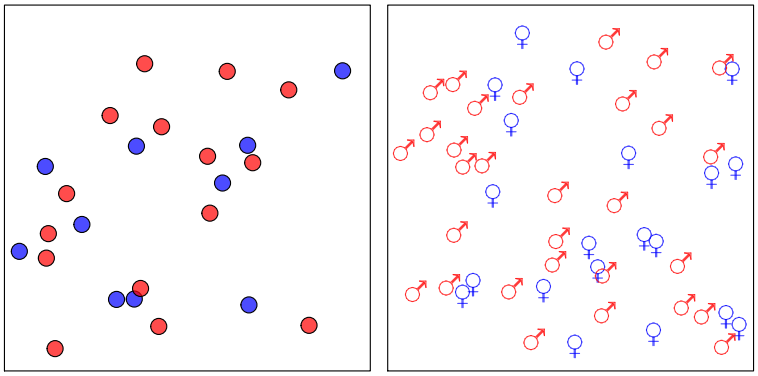


Figure 13.1. Depictions of a 'box' with 15 red balls and 10 blue balls (Left) and a 'room' with 30 men and 20 women (Right).

The two previous examples are simple because the selection is from a small, discrete set of items. Probabilities may be computed for a continuous variable if the distribution of that variable is known for the entire population. 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 (or reasonably approximated). For example, as shown in Module 8, if it can be assumed that heights is $N(66, 3)$, then the proportion of individuals in the population taller than 71 inches tall is 0.0478 (Figure 13.2).[3] This result is a probability because (i) the individual was randomly selected and (ii) the proportion of all individuals of interest in the entire population was found.

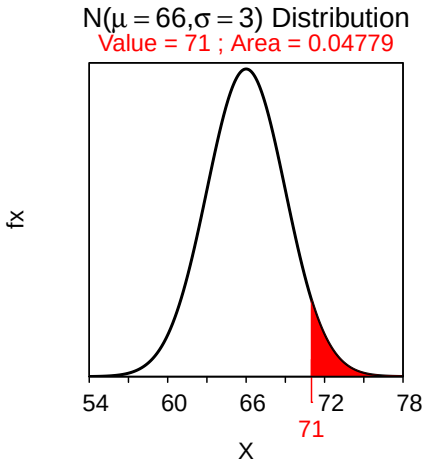


Figure 13.2. Calculation of the probability that a randomly selected individual from a $N(66, 3)$ population will have a height greater than 71 inches.

13.2 Probability of Statistics

The probability of a statistic computed from a random sample can also be found because the Central Limit Theorem (CLT) explains the distribution of statistics from all possible samples from a population Module

[3] As computed with `distrib(71,mean=66,sd=3,lower.tail=FALSE)`.

12. Probability calculations from sampling distributions will be the basis for making statistical inferences in Modules 14 and 16. These calculations are introduced here.

If the sample size is large enough, then the CLT states that the sampling distribution of sample means is approximately normal and the methods from Module 8 may then be used to compute probabilities. Therefore questions such as "what is the probability of observing a sample mean of less than 95 mm from a sample of $n = 50$ from Square Lake?" can be answered. This question is answered by first recalling that, for the length of all fish in Square Lake, μ =98.06 and σ =31.49. Because $n = 50$ is greater than 30, the CLT says that the distribution of the sample means from these samples is $\bar{x} \sim N(98.06, \frac{34.19}{\sqrt{50}})$ or $\bar{x} \sim N(98.06, 4.835)$. Thus, the proportion of samples of $n = 50$ from Square Lake with an $\bar{x} < 95$ mm is 0.2634, which comes from computing the area less than 95 on a $N(98.06, 4.835)$ distribution (Figure 13.3-Left).[4]

```
> ( distrib(95,mean=98.06,sd=34.19/sqrt(50)) )
[1] 0.2634127
```

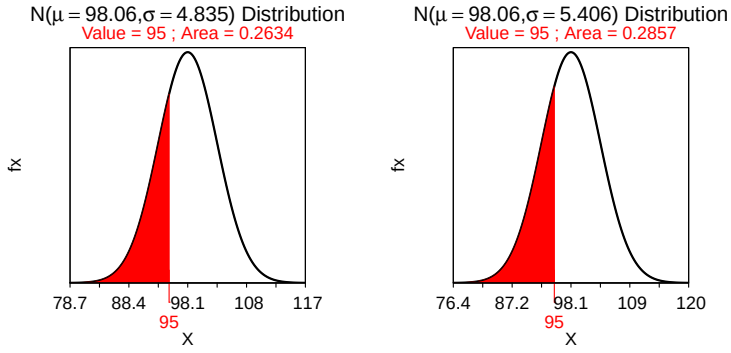


Figure 13.3. Proportion of sample means less than 95 mm on a $N(98.06, 4.84)$ (Left) and $N(98.06, 5.406)$ (Right) distribution.

Consider another question – "what is the probability of observing a sample mean of more than 95 mm in a sample of $n = 40$ from Square Lake?" 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 $\bar{x} \sim N(98.06, 5.406)$ (Note the different value of the SE). Thus, the answer to this question is the area to the right of 95 on a $N(98.06, 5.406)$ or 0.7143 (Figure 13.3-Right).

```
> ( 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.**

[4] 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 12.1.1). 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=`.

Consider two more Square Lake example questions. First, "what is the probability of observing a sample mean of more than 95 mm in a sample of n =10 from Square Lake?" This question is again about a statistic, but because $n < 15$ and the population is not known to be normal it is not known that the sampling distribution will be normal. Thus, this question cannot be answered. Second, "What is the probability that a fish will have a length less than 85 mm?" This question is about an individual, not a statistic as in the previous questions. Thus, the population distribution, NOT the sampling distribution, is appropriate here. However, this question also cannot be answered because the population distribution is not known to be normally distributed.

Two points are illustrated with the last two questions. First, population distributions are used for questions about individuals and sampling distributions are used for questions about statistics. Second, if the distribution is not known to be normal, no matter which distribution is used, then the probability cannot be computed.[5]

One issue you may have noticed is that these calculations require knowing the mean, standard deviation, and shape (if $n < 30$) of the population. However, the population usually cannot be "seen" (recall Module 1) and, thus, it is uncomfortable to assume 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. These questions, while not yet realistic, will help you to better understand sampling distibutions for when they are needed to make inferences in later modules.

13.3 A Process for Handling Probability Questions

As seen in the previous two sections, probability questions may use either the population distribution or the sampling distribution. To properly answer these questions it is important to determine (i) which of these two distributions to use, (ii) whether that distribution is normal or not, and (iii) the specific characteristics (i.e., mean and dispersion) of that distribution.

The type of distribution to use is dictated by whether the question is about an individual or about a statistic. Questions about individuals require using the population distribution, whereas questions about statistics require using the sampling distribution. Information about the population distribution, such as whether it is normally distributed or not and what the mean and standard deviation are, will be provided in the background information provided. In contrast, specifics about the sampling distribution must be identified from applying the rules of the Central Limit Theorem to information provided in the background. For both distributions, the probability question cannot be answered if the distribution is not normal. Both distributions are centered on μ, but the population distribution uses the standard DEVIATION as a measure of dispersion, whereas the sampling distribution uses the standard ERROR.

[5] At least with the techniques in this course.

MODULE 14

HYPOTHESIS TESTS

Contents

A STATISTIC IS AN IMPERFECT ESTIMATE of a parameter because of sampling variability. There are two calculations using the results of a single sample that recognize this imperfection and allow conclusions to be made about a parameter. First, a researcher may form an *a priori* hypothesis about a parameter and then use the information in the sample to make a judgment about the "correctness" of that hypothesis. Second, a researcher may form, from the information in the sample, a range of values that is likely to contain the parameter. The first method is called *hypothesis testing* and is the subject of this module. The second method consists of constructing a *confidence region*, which is introduced in Module 16. Specific applications of these two techniques are described in Modules 17-21.

14.1 Hypothesis Testing & The Scientific Method

In its simplest form, the scientific method has four steps:

1. Observe and describe a natural phenomenon.
2. Formulate a hypothesis to explain the phenomenon.
3. Use the hypothesis to predict new observations.
4. Experimentally test the predictions.

If the results of the experiment do not match the predictions, then the hypothesis is rejected and an alternative hypothesis is proposed. If the results of the experiment closely match the predictions, then belief in the hypothesis is gained, though the hypothesis will likely be subjected to further experimentation.

Statistical hypothesis testing is key to using the scientific method in many fields of study and, in fact, closely follows the scientific method in concept. Statistical hypothesis testing begins by formulating two competing statistical hypotheses from a research hypothesis. One of these hypotheses (the null) is used to predict the parameter of interest. Data is then collected and statistical methods are used to determine whether the observed statistic closely matches the prediction made from the null hypothesis or not. Probability (Module 13) is used to measure the degree of matching with sampling variability taken into account. This process and the theory underlying statistical hypothesis testing is explained in detail in this module.

14.2 Statistical Hypotheses

Hypotheses are classified into two types: (1) research hypothesis and (2) statistical hypotheses. A research hypothesis is a "wordy" statement about the question or phenomenon that the researcher is testing. Four example research hypotheses are:

1. A medical researcher is concerned that a new medicine may change patients' mean pulse rate (from the "known" mean pulse rate of 82 bpm for individuals in the study population not using the new medicine).
2. A chemist has invented an additive to car batteries that she thinks will extend the current 36 month average life of a battery.
3. An engineer wants to determine if a new type of insulation will reduce the average heating costs of a typical house (which are currently \$145 per month).
4. A researcher is concerned whether, on average, Alzheimer's caregivers at a particular facility are clinically depressed (as suggested by a mean Beck Depression Inventory (BDI) score greater than 25)

Research hypotheses are converted to statistical hypotheses that are mathematical and more easily subjected to statistical methods. There are two types of statistical hypotheses: (1) the null hypothesis and (2) the alternative hypothesis. The **null hypothesis**, abbreviated as H_0, is a specific statement of no difference between a parameter and a specific value or between two parameters. The H_0 ALWAYS contains an equals sign because it always represents "no difference." The **alternative hypothesis**, abbreviated as H_A, always states that there is some sort of difference between a parameter and a specific value or between two parameters. The type of difference comes from the research hypothesis and will require use of a less than ($<$), greater than ($>$), or not equals ($\neq$) sign. Null and alternative hypotheses that correspond to the four research hypotheses above are:

1. $H_A : \mu \neq 82$ and $H_0 : \mu = 82$ (where μ represents the mean pulse rate for individuals in the study population that take the new medicine; thus, the alternative hypothesis represents a change from the "normal" pulse rate).
2. $H_A : \mu > 36$ and $H_0 : \mu = 36$ (where μ represents the mean life of batteries with the new additive; thus, this alternative hypothesis represents an extension of the current battery life).
3. $H_A : \mu < 145$ and $H_0 : \mu = 145$ (where μ represents the mean monthly heating bill for houses that receive the new type of insulation; thus, this alternative hypothesis represents a decline in heating bills from the previous "normal" amount).
4. $H_A : \mu > 25$ and $H_0 : \mu = 25$ (where μ represents the mean BDI score; thus, this alternative hypothesis represents a mean score that indicates clinical depression).

The sign used in the alternative hypothesis comes directly from the wording of the research hypothesis (Table 14.1). 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. 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 14.1. 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	

14.2.1 Hypothesis Testing Concept

Statistical hypothesis testing begins by using the null hypothesis to predict what value one should expect for the mean in a sample. So, for the Square Lake example (from Module 1), if $H_0 : \mu = 105$ and $H_A : \mu < 105$, then one would expect, if the null hypothesis is true, that the observed sample mean would be 105. If the observed sample mean was NOT equal to 105 and sampling variability did not exist, then the prediction based on the null hypothesis would not be supported and one would conclude that the null hypothesis was incorrect. In other words, one would conclude that the population mean was not equal to 105.

Of course, sampling variability does exist and it 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, 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 process to determine 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 $\bar{x}$ for samples of $n = 50$ in the Square Lake example is $\frac{\sigma}{\sqrt{n}} = \frac{31.5}{\sqrt{50}} = 4.45$. With this sampling variability, an observed sample mean of 103 would be considered reasonably close to 105 and one would have more belief in $H_0 : \mu = 105$ (Figure 14.1). However, an observed sample mean of 90 is further away from 105 than one would expect based on sampling variability alone and belief in $H_0 : \mu = 105$ would lessen (Figure 14.1).

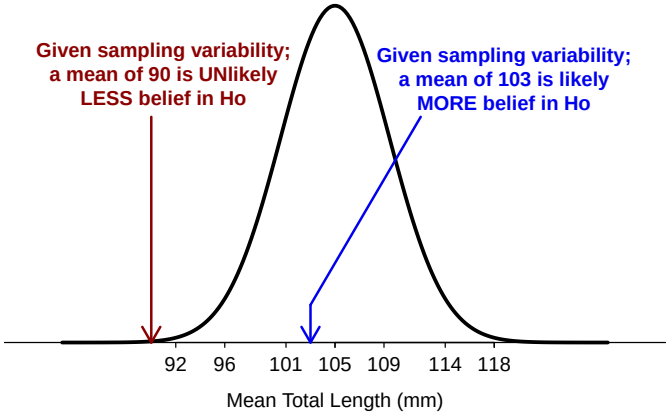


Figure 14.1. Sampling distribution of samples means of n=50 from the Square Lake population ASSUMING that μ=105.

While the above procedure is intuitively appealing, the conclusions are not as clear when the examples chosen (i.e., samples means of 103 and 90) are not as extremely close or distant from the null hypothesized value. For example, what would one concludeif the observed sample mean was 97?. A first step in creating a more objective decision criteria is to compute the "p-value." A p-value is the probability of the observed statistic or a value of the statistic more extreme assuming that the null hypothesis is true. The p-value is described in more detail below given its centrality to making conclusions about statistical hypotheses.

The meaning of the phrase "or more extreme" in the p-value definition is derived from the sign in H_A (Figure 14.2). 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 is clear from Figure 14.2 why "less than" and "greater than" are one-tailed alternatives and "not equals" is a two-tailed alternative.

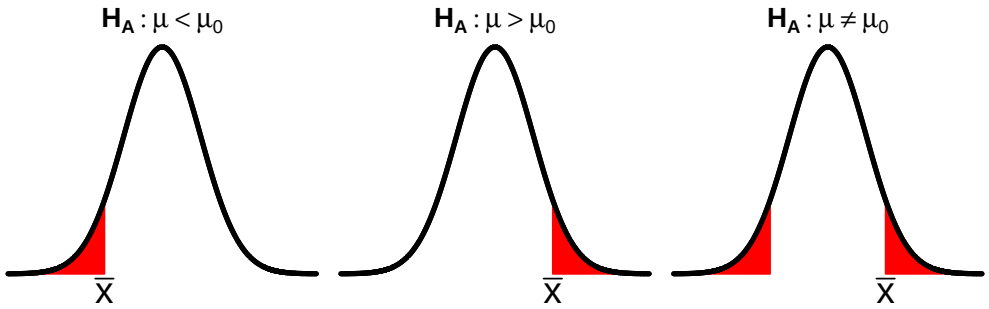


Figure 14.2. Depiction of "or more extreme" (red areas) in p-values for the three possible alternative hypotheses.

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 from 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 means.[1] The null distribution in the Square Lake example is thus $\bar{x} \sim N(105, 4.45)$ because $n = 50 > 30$ (so the Central Limit Theorem holds), $H_0 : \mu = 105$, 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 with $n = 50$ from Square Lake (as it was in Table 2.2). The p-value in this case would be "the probability of observing $\bar{x} = 100$ or a smaller value assuming that $\mu = 105$." This probability is computed by finding the area to the left of 100 on a $N(105, 4.45)$ null distribution and is the exact same type of calculation as that made in Section 13.2. Thus, this p-value of $p = 0.1308$ is computed as below and shown in Figure 14.3.

```
> ( distrib(100,mean=105,sd=31.49/sqrt(50)) )
[1] 0.1307722
```

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 null hypothesized value. 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.

[1] Of course, unless the null hypothesis happens to be perfectly true.

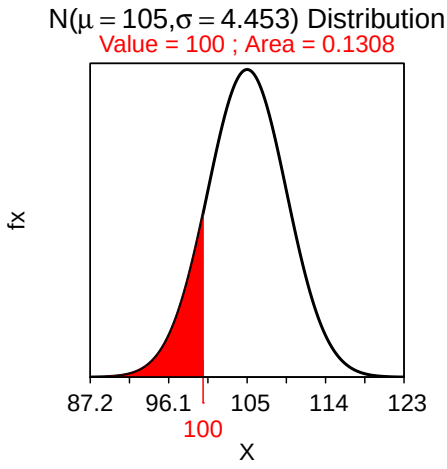


Figure 14.3. Depiction of the p-value for the Square Lake example where $\bar{x} = 100$ and $H_A : \mu < 105$.

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 and symbolized with α, such that p-values less than α are considered small and would result in rejecting 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.

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 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.

The null hypothesis in the Square Lake example is not rejected because the p-value (i.e., 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.

14.3 Test Statistics and Effect Sizes

Instead of reporting the observed statistic and the resulting p-value, it may be of interest to know how "far" the observed statistic was from the hypothesized value of the parameter. This 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 difference 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 3 but does not seem "different" if the standard error is 15 (Figure 14.4).

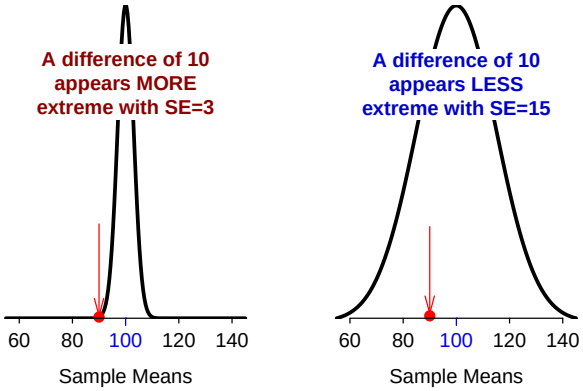


Figure 14.4. Sampling distribution of samples means with SE=3 (Left) and SE=15 (Right). A single observed sample mean of 90 (a difference of 10 from the hypothesized mean of 100) is shown by the red dot and arrow.

The difference between the observed statistic and the hypothesized parameter is standardized to a common scale by dividing by the standard error of the statistic. The result is called a *test statistic* and is generalized with

$$\text{Test Statistic} = \frac{\text{Observed Statistic} - \text{Hypothesized Parameter}}{SE_{\text{Statistic}}} \tag{14.3.1}$$

Thus, the test statistic (14.3.1) measures how many standard errors the observed statistic is away from the hypothesized parameter. A relatively large value is indicative of a difference that is likely not due to randomness (i.e., sampling variability) and suggests that the null hypothesis should be rejected.

The test statistic in the Square Lake Example is $\frac{100-105}{\frac{31.49}{\sqrt{50}}} = -1.12$. Thus, the observed mean total length of 100 mm is 1.12 standard errors below the null hypothesized mean of 105 mm. From our experience, a little over one SE from the mean is not "extreme" and, thus, it is not surprising that the null hypothesis was not rejected.

There are other forms for calculating test statistics, 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.

14.4 Hypothesis Testing Concept Summary

In summary, hypotheses are statistically examined with the following procedure.

1. Construct null and alternative 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.
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.

Statisticians say "do not reject H_0" rather than "accept H_0 as true" when the p-value $> \alpha$ for two reasons. First, there are several other possible values, besides the specific value in 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 rejected.[2] So, we don't 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 mm"). 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.

[2] In fact, for example, the values in a 95% confidence interval – see Module 16 – represent all possible hypothesized values that would not be rejected with a two-tailed H_A using $\alpha = 0.05$.

MODULE 15

HYPOTHESIS TEST ERRORS

Contents

DECISIONS ABOUT HYPOTHESES BASED ON STATISTICS may, at times, be incorrect. In this module, two types of errors that can be made are described and the probability of making these errors is described. The concepts in Module 14 should be understood before proceeding here.

15.1 Errors

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 15.1). 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 15.1. 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

The decision in the Square Lake example of Module 14 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 but unknown μ. Decisions can be made, however, that affect the magnitude of β (discussed below with power).

15.2 Statistical Power

A concept that is very closely related to decision-making errors is the idea of statistical power, or just **power** for short. 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. 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.

Power is equal to $1 - \beta$ and, thus, like β it cannot be computed directly because the actual mean (μ_A) is 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"). This reverse calculation on the null distribution gives $\bar{x}$=99.2928.

```
> ( rejreg <- distrib(0.10,mean=105,sd=31.49/sqrt(50),type="q") )
[1] 99.29279
```

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 the null distribution (i.e., a "forward" calculation on the actual distribution).
   - This computation is to find the area to the left of $\bar{x}$=99.2928 on $N(98.06, \frac{31.49}{\sqrt{50}})$. The area to the left of this Z is 0.6090.

```
> ( distrib(rejreg,mean=98.06,sd=31.49/sqrt(50)) )
[1] 0.6090419
```

Thus, the power to detect a $\mu_A = 98.06$ was 0.6090. This means that in only about 61% of the samples will the false $H_0 : \mu = 105$ be correctly rejected. Thus, it is not too surprising that H_0 was not rejected in this example.

Even though power can usually not be calculated, a researcher can make decisions that will positively affect power (Figure 15.1). For example, a researcher can increase power by increasing α or n. Increasing n is more beneficial because it does not result in an increase in Type I errors as would occur with increasing α.

In addition, power decreases as the difference between the hypothesized mean (μ_0) and the actual mean (μ_A) decreases (Figure 15.1). 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; Figure 15.1). 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. For example, if n could be doubled in the Square Lake example to 100, then the power to correctly reject $H_0 : \mu = 105$ would increase to approximately 0.82 (Figure 15.1).

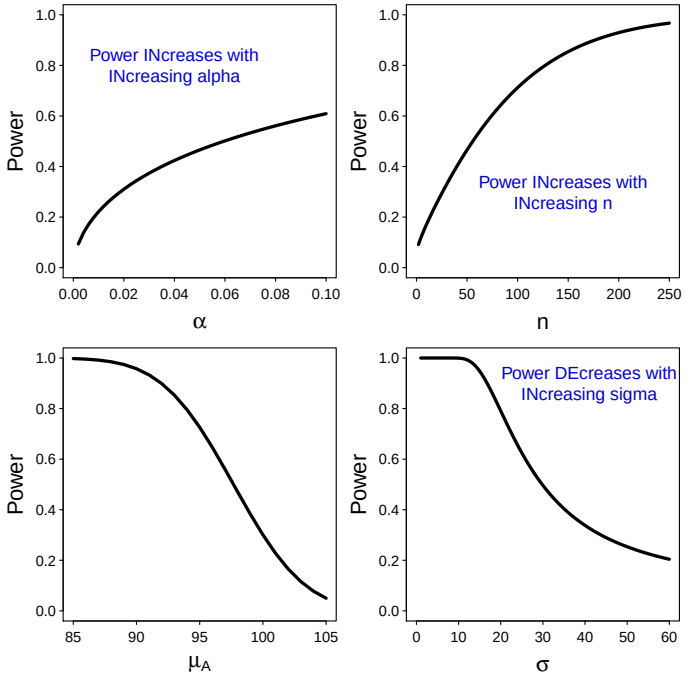


Figure 15.1. The relationship between one-tailed (lower) power and α, n, actual mean (μ_A), and σ. In all situations where the variable does not vary, $\mu_0 = 105$, $\mu_A = 98.06$, $\sigma = 31.49$, $n = 50$, and $\alpha = 0.05$.

MODULE 16

CONFIDENCE REGIONS

Contents

THE FINAL RESULT FROM A HYPOTHESIS TEST (Module 14) can feel uneventful – i.e., either conclude that the parameter may be equal to or different from the hypothesized value.[1] If the parameter is thought to be different from the hypothesized value we might then say that our best guess at the parameter is the observed statistic. However, as shown in Module 12, a statistic is an imperfect estimate of the unknown parameter because of sampling variability. This imperfectness can be recognized by computing a range of values that is likely to contain the 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." The range in the last phrase acknowledges sampling variability and is called a confidence region. In this module, the concept, calculation, and interpretation of confidence regions is explored.

16.1 Confidence Concept

An understanding of what it means to be "95% confident" requires examination of multiple samples from a population, as was done in Module 12 when considering sampling variability. In this initial discussion, only 95% confidence intervals (CI), where a range (i.e., bounded on both ends) is computed, are considered. These simplifying restrictions and unrealistically knowing population values are used here only so that the **concept** of confidence intervals can be more easily explained. General methods for constructing other types of confidence regions with other levels of confidence are in Section 16.2.

[1] Depending on the H_A it may be known if the parameter is more or less than the hypothesized value.

In the Square Lake example (introduced in Module 1), it was known that μ=98.06 and σ=31.49 (Table 2.1). Additionally, $\bar{x} = 100.04$ was obtained from the first sample of n=50 (Table 2.2). A 95% CI for μ is defined as $\bar{x} \pm 2SE_{\bar{x}}$. The 95% CI for the mean total length for the Square Lake population, computed from this one sample, is $100.04 \pm 2\frac{31.49}{\sqrt{50}}$, 100.04 ± 8.91, or (91.13,108.95). This interval contains μ (i.e., 98.06 is between 91.13 and 108.95). In other words, this particular CI accomplished 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 of 100 95% confidence intervals shown in Figure 16.1 did not contain μ. Thus, the researcher would have concluded that μ was in an incorrect interval four times in these 100 samples. The concept of "confidence" in confidence regions is related to determining how often the intervals correctly contain the parameter.

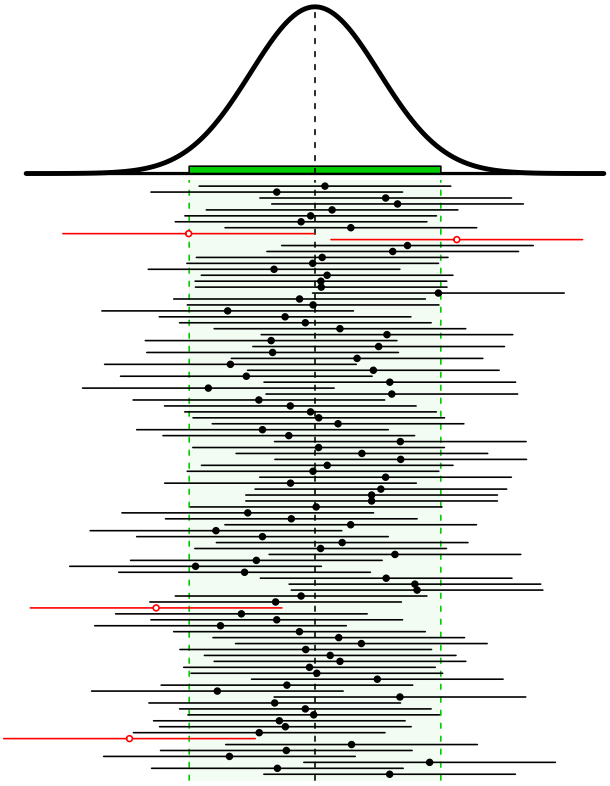


Figure 16.1. Sampling distribution of the sample mean (top) and 100 95% confidence intervals (horizontal lines) from samples of n=50 from the Square Lake population. Confidence intervals that do NOT contain μ=98.06 are shown in red and with an open circle. The green shaded area represents 95% of the sample means. See text for more explanation.

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 16.1. Note that any sample that produced a mean (dot on the CI line) inside the expected range of sample means (light green area) also produced a 95% CI that contained μ (i.e., black CI line with a solid circle). 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% CIs.

The specifics for constructing confidence regions with different levels of confidence is 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 16.2-Left) and 100 80% CIs (Figure 16.2-Right) for the Square Lake fish length data.

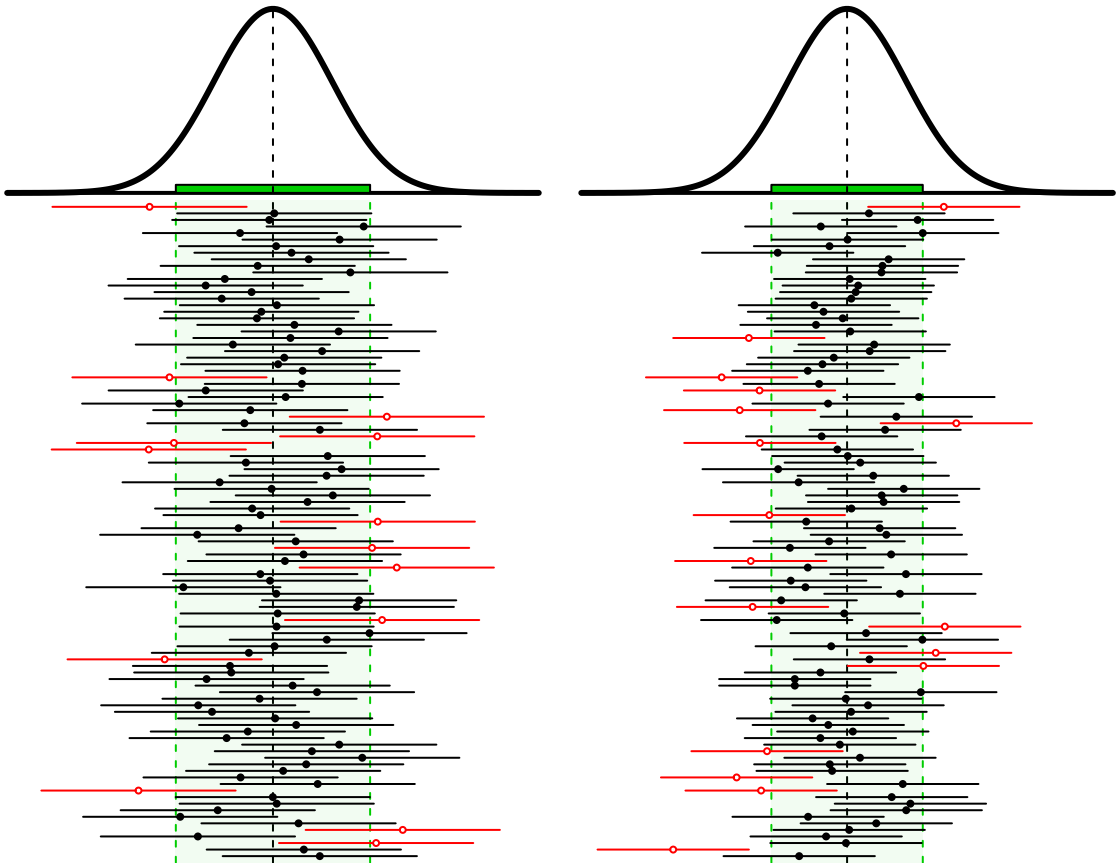


Figure 16.2. Sampling distribution of the sample mean (**tops**) and 100 random 90% (**Left**) and 80% (**Right**) confidence intervals (horizontal lines) from samples of n=50 from the Square Lake population. Confidence intervals that do NOT contain μ are shown in red.

The concept of confidence regions can be difficult to grasp at first. Thus, one should consider the following subtleties about the concept of a confidence region:

- A CI is a random variable like any other statistic. That is, each sample results in a different 95% CI (see CI lines in Figures 16.1 and 16.2) just like it results in a different $\bar{x}$ (see dots on CI lines in Figures 16.1 and 16.2).
- Any CI either contains the parameter (e.g., μ) or not. However, on average, 95% of 95% CIs will contain the parameter and 5% will not. That is, 95% of all possible 95% CIs wil contain the parameter.
- A 95% CI is a technique that "works correctly" 95% of the time. In other words, 95% of all 95% CI "capture" the unknown parameter.

Because of these subtleties, confidence regions are often misinterpreted. Common misinterpretations are listed below with an explanation for 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 contained in the confidence interval." [*This is incorrect because the population mean is constant (not random), it either is or is not in a particular CI, and it will never change whether it is or is not in that CI. The CI, not the parameter, is random.*]
2. "There is a 95% probability that the sample mean is contained in the confidence interval." [*This is incorrect for the simple fact that CI 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 CI (see next section).*]
3. "95% of all 95% confidence intervals are contained in the confidence interval." [*First, this is physically impossible because each CI is the same width (if n and the level of confidence stay constant). Second, it is not important how many CI are contained in a CI; interest is in whether the parameter is in the interval or not.*]

◇ **Confidence intervals are constructed for parameters, not statistics.**

◇ **Care and specificity must be used when interpreting and describing confidence intervals.**

16.2 Constructing Confidence Regions

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, as intervals or bounds, and for nearly all **parameters**.

The level of confidence (C) used will be determined by the α chosen for the hypothesis test; specifically, $C = 100(1-\alpha)\%$. For example, if α is set at 0.05, then the level of confidence will be $100(1-0.05)\%$ (Table 16.1). Thus, 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 hypothesis testing is linked to the proportion of errors made with confidence regions.

Table 16.1. Several common confidence levels (C) and the corresponding probability of a Type I error (α).

α	C
0.01	99%
0.05	95%
0.10	90%

The type of confidence region depends on the type of alternative hypothesis (Table 16.2). If the alternative hypothesis is two-tailed (i.e., $\neq$), then the confidence region will be an interval (i.e., a range will be computed, as in Section 16.1). However, if the alternative hypothesis is one-tailed, then a 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 a range of possible values). 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.

Table 16.2. Confidence regions and their interpretation in relation to alternative hypotheses (H_A) types.

H_A	Confidence Region	Interpretation
$\neq$	Interval	Parameter in interval
$<$	Upper Bound	Parameter less than upper bound
$>$	Lower Bound	Parameter greater than lower bound

Fortunately, most confidence regions follow the same basic form of

$$\text{“Statistic”} + \text{“scaling factor”} * SE_{statistic}$$

where “Statistic” represents the statistic used to estimate the parameter, $SE_{statistic}$ is the standard error of that statistic, and “scaling factor” $* SE_{statistic}$ is called the margin-of-error. The scaling factor is computed from a known distribution. When σ is known, the scaling factor is computed from a $N(0,1)$ and is called Z^*. Thus, in the case when a confidence interval is being constructed for μ and σ is known, the specific formula for the confidence region is

$$\bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

The “scaling factor” serves to control the width and type of confidence region. The magnitude of the scaling factor controls the relative width of the region such that the parameter is contained in the region at a rate according to the level of confidence. For example, the scaling factor for a 99% confidence region will be set such that 99% of the confidence regions will contain the parameter.

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 Z^* is the two values such that an area equal to the level of confidence is contained between them (Figure 16.3-Left). The two values that delineate these boundaries will be the same value but with different signs because the $N(0,1)$ is symmetric about zero. Thus, a confidence interval is computed with a scaling factor of $\pm Z^*$.

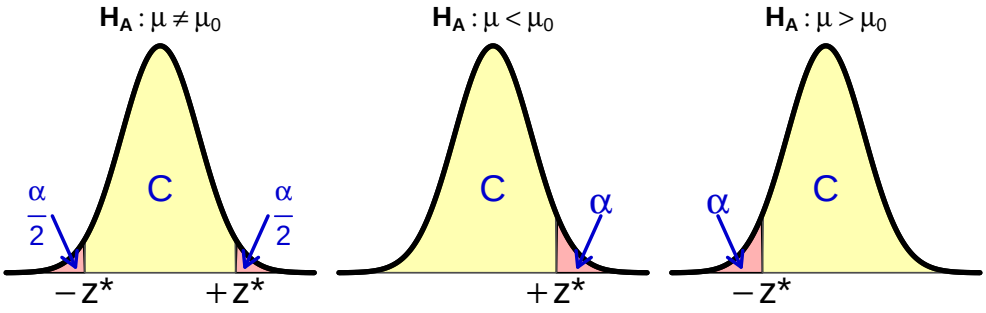


Figure 16.3. Areas (yellow) that define Z^* for confidence regions of a parameter in a hypothesis test.

In contrast, if the alternative hypothesis is a “less than”, then an upper confidence bound is desired and Z^* has an area equal to the level of confidence LESS THAN it (Figure 16.3-Middle). As the level of confidence will always be greater than 50%, this definition will produce a positive 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 Z^* has an area equal to the level of confidence GREATER THAN it (Figure 16.3-Right). This definition produces a negative Z^* so that the scaling factor will be $-Z^*$.

◇ **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.**

Constructing a proper confidence region should follow the five steps below. These steps are illustrated in three examples further below.

1. Identify the level of confidence (i.e., $C = 100(1-\alpha)\%$; Table 16.1).
2. Identify the type of confidence regions – interval, lower bound, or upperbound (Table 16.2).
3. Determine the scaling factor.
4. Compute the actual confidence region.
5. Interpret the confidence region.

Consider the Square Lake example where $H_A : \mu < 105$, $\alpha = 0.05$, and $\bar{x}$=100.04 from $n = 50$ (Table 2.2).

1. $C = 95\%$ $(100(1-0.05))$.
2. Upper confidence bound because H_A is a "less than."
3. $Z^* = +1.645$ as found with [Note that `mean=0` and `sd=1` are the default settings for `distrib()` and can, thus, be omitted when finding a Z^*.]

```
> ( distrib(0.95,type="q") )
[1] 1.644854
```

4. $100.04 + 1.645\frac{31.49}{\sqrt{50}}$, $100.04 + 7.33$, or 107.37.
5. One is 95% confident that the 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 μ.

Second, suppose that the Lake Superior ice cover data from Table 7.2 (note that $\bar{x}$=107.8 and n=42) was tested with $H_A : \mu \neq 100$, $\sigma = 22$, and $\alpha = 0.01$.

1. $C = 99\%$ $(100(1-0.01))$.
2. Confidence interval because H_A is a "not equals."
3. $Z^* = \pm 2.576$ as found with

```
> ( distrib(0.995,type="q") )
[1] 2.575829
```

4. $107.8 \pm 2.576\frac{22}{\sqrt{42}}$, 107.8 ± 8.74, or $(99.06, 116.54)$.
5. One is 95% confident that the mean annual number of days of ice cover on Lake Superior is between 99.1 and 116.5 days. By confident, it is meant that 95% of all 95% confidence regions will contain μ.

Finally, suppose that the second example hypothesis test in Module 14 about battery life (i.e., $H_A : \mu > 36$ vs $H_0 : \mu = 36$) is being tested with $\alpha = 0.10$. Further suppose that σ =7 and that $\bar{x}$ =45 from n =40.

1. $C = 90\%$ $(100(1-0.10))$.
2. Lower confidence bound because H_A is a "greater than."
3. Z^* =-1.282 as found with

```
> ( distrib(0.90,type="q",lower.tail=FALSE) )
[1] -1.281552
```

4. $45 - 1.282\frac{7}{\sqrt{40}}$, $45 - 1.42$, or 43.58
5. One is 90% confident that the mean life for ALL batteries with the additive is more than 43.58 months. By confident, it is meant that 95% of all 95% confidence regions will contain μ.

16.3 Hypothesis Tests and Confidence Region Relationship

An alternative conceptualization of confidence intervals can show how confidence regions and hypothesis tests are related. This conceptualization rests on considering the sample means that would be "reasonable to see" from populations with various values of μ. A graphic is contructed below using the Square Lake population as an example and assuming that σ is known (=31.49), $n = 50$, and 95% CIs are used.

First, compute the most common 95% of sample means assuming that $\mu = 70$; i.e., $70 \pm 1.960\frac{31.49}{\sqrt{50}}$ or $(61.27, 78.73)$. This range is plotted as a vertical rectangle centered on $\mu = 70$ (left-most rectangle) in Figure 16.4-Left. Next, compute and plot the same range for a slightly larger μ (e.g., with $\mu = 71$, plot $(62.27, 78.73)$). Then repeat these steps for sequentially larger values of μ until a plot similar to Figure 16.4-Left is constructed.

Consider very carefully what Figure 16.4-Left represents. The vertical rectangles represent the ranges 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 essence, each vertical line represents the sample means that are likely to be observed from a population with a given population mean (x-axis).

Now suppose that $\bar{x}$=100.04 (Table 2.2). Draw a horizontal line across Figure 16.4 at this value and then draw vertical lines down from where the horizontal line first enters and last leaves the band of possible sample means (Figure 16.4-Right). The x-axis values that these vertical lines intercept are an approximate 95% CI for μ. The approximation is only as close as the intervals used to construct the rectangles (i.e., 1.0 mm were used here). However, the results from this graphical approach (i.e., $(92, 108)$) compare favorably to the previous results from using the CI formula (i.e., $(91.27, 108.73)$).

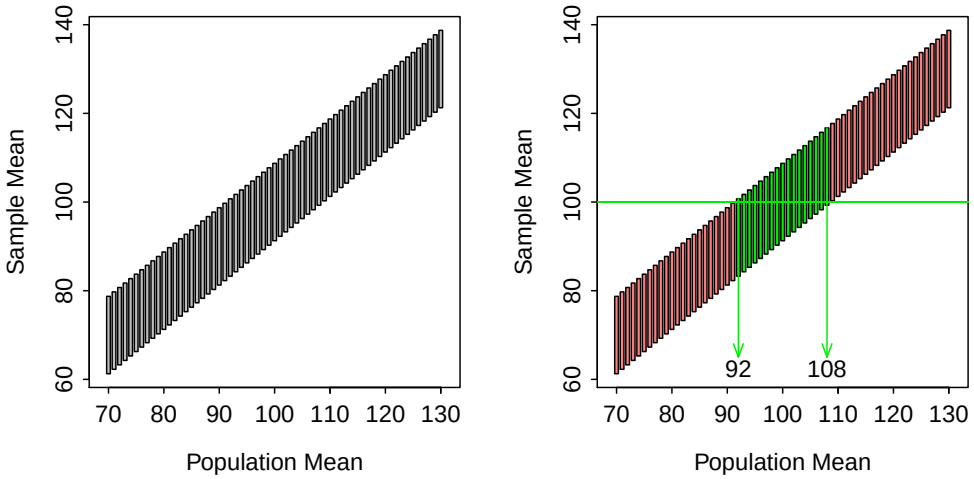


Figure 16.4. Range (95%) of sample means that would be produced by particular population means in the Square Lake fish length example (**Left**) and the ranges intercepted by $\bar{x} = 100.04$ mm (**Right**).

Surely, the CI formula (Section 16.2) is a more efficient and precise way to construct confidence intervals. However, this conceptualization illustrates that a confidence interval (or region, more generally) consists of population means that are likely to produce the observed sample mean. Thus, a confidence region represents possible null hypothesized population means that WOULD NOT BE rejected during hypothesis testing.

◇ **A confidence region represents null hypothesized values that would NOT be rejected.**

16.4 Precision and Sample Size

The width of a confidence interval explains how precisely the parameter is estimated. For example, narrow intervals represent precise estimates of the parameter. The width of a confidence interval is directly related to the margin-of-error which 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 confidence interval gets smaller.

A small standard error means that sampling variability is low and the parameter is precisely estimated by the statistic. Smaller standard errors are obtained only by increasing the sample size. A smaller standard deviation would also result in a smaller SE, but the standard deviation cannot be made smaller (i.e., it is an inherent characteristic of the population).

A smaller scaling factor is obtained by reducing the level of confidence. For example, a 90% confidence interval uses a $Z^* = \pm 1.645$ whereas a 95% confidence interval uses a $Z^* = \pm 1.960$. Thus, decreasing the confidence level narrows the CI. However, reducing the level of confidence will also increase the number of confidence intervals that do not contain the parameter. Thus, reducing the level of confidence may not be the best choice for narrowing the confidence interval.

The margin-of-error formula can be solved for n.

$$\begin{aligned} m.e. &= Z^* \frac{\sigma}{\sqrt{n}} \\ \sqrt{n} &= \frac{Z^* \sigma}{m.e.} \\ n &= \left(\frac{Z^* \sigma}{m.e} \right)^2 \end{aligned}$$

This formula can be used to find the n required to estimate μ within $\pm m.e.$ units with C% confidence assuming that σ is known. For example, suppose that one wants to determine n 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. From this, $m.e.$=5, σ=34.91, and Z^*=1.645 (found previously for 90% confidence).[2] 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 to meet these constraints. Note that sample size calculations are always rounded up to the next integer because rounding down would produce a sample size that does not meet the desired criteria.

◇ Always round sample size calculations up to the next integer.

The margin-of-error and confidence level in these calculations need to come from the researcher's beliefs about 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 best guesses are often used instead. In practice, a researcher will often prepare a graph with varying values of σ to make an informed decision of what sample size to choose.

[2] Strictly, $Z^* \pm$=1.645, but the sign is inconsequential due to squaring in the sample size formula.

MODULE 17

1-SAMPLE Z-TEST

Contents

A FOUNDATION FOR MAKING STATISTICAL INFERENCES was provided in Modules 12-16. Most of the material in Modules 14 and 16 is related to a 1-Sample Z-test, which is formalized in this module. Other specific hypothesis tests are in Modules 18-21.

17.1 11-Steps of Hypothesis Testing

Hypothesis testing is a rigorous and formal procedure for making inferences about a parameter from a statistic. The 11 steps listed below will help make sure that all aspects important to hypothesis testing are completed. These steps should be used for all hypothesis tests in this and ensuing modules.

1. State the rejection criterion (α).
2. State the null and alternative hypotheses to be tested and define the parameter(s).
3. Identify (and explain why!) the hypothesis test to use (e.g., 1-Sample t, 2-sample t, etc.).
4. Collect the data (address study type 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. State your rejection decision about H_0.
10. Summarize your findings in terms of the problem.
11. Compute and interpret an appropriate confidence region for the parameter.

The order of some of these steps is arbitrary. However Steps 1-3 **MUST** be completed before collecting data (Step 4). Further note that Step 11 is completed to provide a more definitive statement about the value of the parameter when H_0 was rejected (i.e., if the parameter differs from the hypothesized value, then provide a range for which the actual parameter may exist).

17.2 1-Sample Z-Test Specifics

A 1-Sample Z-Test tests $H_0 : \mu = \mu_0$, where μ_0 represents a specific value of μ, when σ is known. Other specifics of this test were discussed in previous modules and are summarized in Table 17.1.

Table 17.1. 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:** $\bar{x} + Z^* \frac{\sigma}{\sqrt{n}}$
- **Assumptions:**
  1. σ is known
  2. $n \geq 30$, $n \geq 15$ and the **population** is not strongly skewed, OR the **population** is normally distributed.
- **Use with:** Quantitative response, one group (or population), σ known.

The only test that can possibly be confused with a 1-Sample Z-Test is a 1-Sample t-Test (Module 18), which tests the same null hypothesis but when σ is unknown.

17.2.1 Example - Intra-class Travel

Below are the 11-steps (Section 17.1) for completing a full hypothesis test for the following situation:

> *A dean wants to determine if it takes more than 10 minutes, on average, to go between classes. To test this hypothesis, she collected a random sample of 100 intra-class travel times and found a mean of 10.12 mins. Assume from previous studies that the distribution of intra-class times is symmetric with a standard deviation of 1.60 mins. Test the dean's hypothesis with $\alpha = 0.10$.*

1. α=0.10.
2. $H_0 : \mu = 10$ mins vs. $H_A : \mu > 10$ mins, where μ is the mean time for ALL intra-class travel events.
3. A 1-Sample Z-Test is required because (i) a quantitative variable (intra-class travel time) was measured, (ii) individuals from one group (or population) is considered (students at the Dean's school), and (iii) σ is thought to be known (=1.60 mins).
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. (i) $n = 100 \geq 30$ and (ii) σ is thought to be known (=1.60 mins).
6. $\bar{x}$=10.12.
7. $Z = \frac{10.12-10}{\frac{1.60}{\sqrt{100}}} = \frac{0.12}{0.16}$=0.75.
8. p-value=0.2266.
9. H_0 is not rejected because the p-value > α=0.10.
10. It appears that the mean time for **all** intra-class travel events is not more than 10 minutes.
11. A 90% lower confidence bound will use $Z*$=-1.282. The lower confidence bound is thus $10.12 - 1.282 * \frac{1.60}{\sqrt{100}}$=9.91. Thus, I am 90% confident that the mean intra-class travel time is more than 9.91 minutes; further evidence that the mean intra-class travel time is not greater than 10 minutes.

R Appendix:

```
( z <- distrib(10.12,mean=10,sd=1.60/sqrt(100),lower.tail=FALSE) )
( zstar <- distrib(0.90,lower.tail=FALSE,type="q") )
```

17.3 1-Sample Z-Test in R

If raw data exist, the calculations for a 1-Sample Z-test can be efficiently computed with `z.test()`. This function requires the vector of quantitative data as the first argument, the hypothesized value for μ in `mu=`, and the known σ in `sd=`. Additionally, the type of alternative hypothesis may be declared in `alt=`, where `alt="two.sided"` (the default), `alt="less"`, and `alt="greater"` correspond to the "not equals", "less than", and "greater than" hypotheses, respectively. Finally, the level of confidence may be given as a proportion (between 0 and 1) in `conf.level=` (which defaults to 0.95).

17.3.1 Body Temperature

Below are the 11-steps (Section 17.1) for completing a full hypothesis test for the following situation:

> *Machowiak et al. (1992) critically examined the belief that the mean body temperature is 98.6^{o}F by measuring body temperatures in a sample of healthy humans. Use their data in BodyTemp.csv, with $\sigma = 0.63^o$F and $\alpha = 0.01$ to determine if the mean body temperature differs from 98.6^{o}F.*

1. α=0.01.
2. $H_0 : \mu = 98.6^o$F vs. $H_A : \mu \neq 98.6^o$F, where μ is the mean body temperature for ALL healthy humans. [*Note that not equals was used because the researchers want to determine if the temperature is* ***different from*** $98.6^o F$.]
3. A 1-Sample Z-Test is required because (i) a quantitative variable (i.e., body temperature) was measured, (ii) individuals from one group (or population) is considered (i.e., healthy humans), and (iii) σ is thought to be known ($= 0.63^o$F).
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.
5. (i) $n = 130 \geq 30$ and (ii) σ is thought to be known ($= 0.63^o$F).
6. $\bar{x} = 98.25^o$F (Table 17.2).
7. $Z = -6.35$ (Table 17.2).
8. p-value< 0.00005 (Table 17.2).
9. Reject H_0 because p-value$< \alpha = 0.01$.
10. It appears that the mean body temperature of ALL healthy humans is less than 98.6^{o}F. [*Note that the test was for a difference but because $\bar{x} < 98.6$ this more specific conclusion can be made.*]
11. I am 99% confident that the mean body temperature (μ) for ALL healthy humans is between 98.1 and 98.4^{o}F (Table 17.2).

Table 17.2. Results from 1-Sample Z-Test for mean body temperature.

```
z = -6.3482, n = 130.000, Std. Dev. = 0.630, Std. Dev. of the sample mean =
0.055, p-value = 2.178e-10
99 percent confidence interval:
 98.10690 98.39156
sample estimates:
mean of bt$temp
       98.24923
```

R Appendix:

```
bt <- read.csv("data/BodyTemp.csv")
( bt.z <- z.test(bt$temp,mu=98.6,sd=0.63,conf.level=0.99) )
```

MODULE 18

1-SAMPLE T-TEST

Contents

PRIOR TO THIS MODULE, hypothesis testing methods required knowing σ, which is a parameter that is seldom known. When σ is replaced by its estimator, s, the test statistic follows a Student's t rather than a standard normal (Z) distribution. In this module, the t-distribution is described and a 1-Sample t-Test for testing that the mean from one population equals a specific value is discussed.

18.1 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 18.1). 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 s rather than σ in the standard error calculation.[1] The degrees-of-freedom are related to n and generally come from the denominator in the standard deviation calculation. As the degrees-of-freedom increase, the t-distribution becomes narrower, taller, and approaches the standard normal distribution (Figure 18.1).

[1] Recall that the sample standard deviation is a statistic and is thus subject to sampling variability.

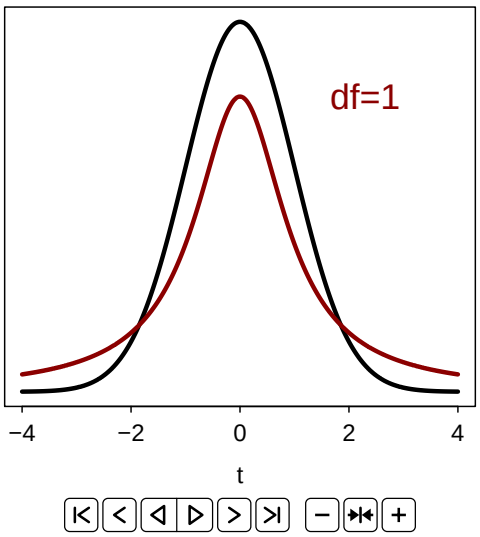


Figure 18.1. Standard normal (black) and t-distributions (red) with varying degrees-of-freedom.

Proportional areas on a t-distribution are computed using `distrib()` similar to what was described for a normal distribution in Modules 8 and 12. The major exceptions for using `distrib()` with a t-distribution is that `distrib="t"` must be used and the degrees-of-freedom must be given in `df=` (how to find df is discussed in subsequent sections). For example, the area right of $t = -1.456$ on a t-distribution with 9 df is 0.9103 (Figure 18.2).

```
> ( distrib(-1.456,distrib="t",df=9,lower.tail=FALSE) )
[1] 0.9103137
```

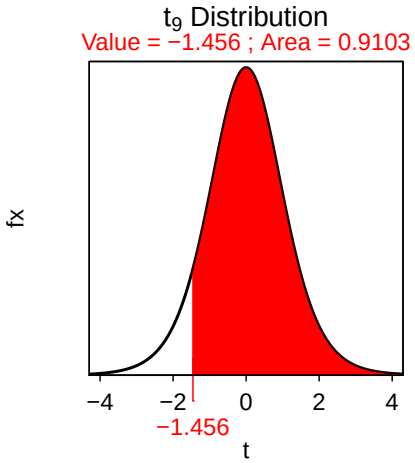


Figure 18.2. Depiction of the area to the right of $t = -1.456$ on a t-distribution with 9 df.

Similarly, the t with an upper-tail area of 0.95 on a t-distribution with 19 df is -1.729 (Figure 18.3).[2]

```
> ( distrib(0.95,distrib="t",type="q",df=19,lower.tail=FALSE) )
[1] -1.729133
```

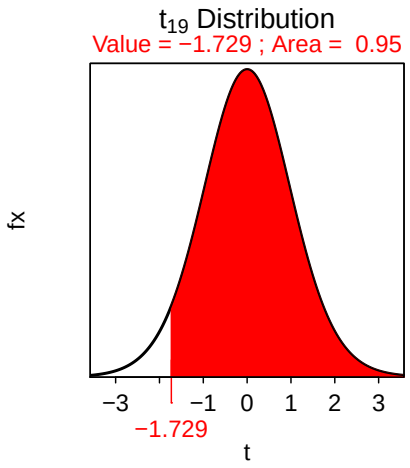


Figure 18.3. Depiction of the value of t with an area to the right of 0.95 on a t-distribution with 19 df.

18.2 1-Sample t-Test Specifics

A 1-Sample t-Test is similar to a 1-Sample Z-test in that both test the same H_0. The difference, as discussed above, is that when σ is replaced by s, the test statistic becomes t and the scaling factor for confidence regions becomes a t^*. Other aspects are similar between the two tests as shown in Table 18.1.[3]

Table 18.1. 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:** $\bar{x} + t^* \frac{s}{\sqrt{n}}$
- **df**: $n - 1$
- **Assumptions:**
  1. σ is UNknown
  2. $n \geq 40$, $n \geq 15$ and the **sample** (i.e., histogram) is not strongly skewed, OR the **sample** is normally distributed.
- **Use with:** Quantitative response, one group (or population), σ UNknown.

[2] This "reverse" calculation would be t^* for a 95% lower confidence bound.
[3] Compare Table 18.1 to Table 17.1.

18.2.1 Example - Purchase Catch of Salmon?

Below are the 11-steps (Section 17.1) for completing a full hypothesis test for the following situation:

> *A prospective buyer will buy a catch of several thousand salmon if the mean weight of all salmon in the catch 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?*

1. α=0.05.
2. $H_0 : \mu = 19.9$ lbs vs. $H_A : \mu > 19.9$ lbs where μ is the mean weight of ALL salmon in the catch.
3. A 1-Sample t-Test is required because (1) a quantitative variable (weight) was measured, (ii) individuals from one group (or population) were considered (this catch of salmon), and (iii) σ is **UN**known.[4]
4. The data appear to be part of an observational study with random selection.
5. (i) n=50 $\geq$ 40 and (ii) σ is unknown.
6. $\bar{x} = 20.1$ lbs (and $s = 0.76$ lbs).
7. $t = \frac{20.1-19.9}{\frac{0.76}{\sqrt{50}}} = \frac{0.2}{0.107} = 1.87$ with df = 50-1 = 49.
8. p-value = 0.0337.
9. H_0 is rejected because the p-value $< \alpha$.
10. The average weight of ALL salmon in this catch appears to be greater than 19.9 lbs; thus, the buyer should accept this catch of salmon.
11. I am 95% confident that the mean weight of ALL salmon in the catch is greater than 19.92 lbs (i.e., $20.1 - 1.677\frac{0.76}{\sqrt{50}} = 20.1 - 0.18 = 19.92$).

R Appendix:

```
( pval <- distrib(1.87,distrib="t",df=49,lower.tail=FALSE) )
( tstar <- distrib(0.95,distrib="t",type="q",df=49,lower.tail=FALSE) )
```

18.3 1-Sample t-Test in R

If raw data exist, the calculations for a 1-Sample t-test can be efficiently computed with `t.test()`. The arguments to `t.test()` are the same as those for `z.test()`, with the exception that `sd=` is not used with `t.test()`. Thus, `t.test()` requires the vector of quantitative data as the first argument, the null hypothesized value for μ in `mu=`, the type of alternative hypothesis in `alt=` (again, can be `alt="two.sided"` (the default), `alt="less"`, or `alt="greater"`), and the level of confidence as a proportion in `conf.level=` (defaults to 0.95). The use of `t.test()` is illustrated in the following example.

18.3.1 Example - Crab Body Temperature

Below are the 11-steps (Section 17.1) for completing a full hypothesis test for the following situation:

> *A marine biologist wants to determine if the body temperature of crabs exposed to ambient air temperature is different than the ambient air temperature. The biologist exposed a sample of 25 crabs to an air temperature of 24.3°C for several minutes and then measured the body temperature of each crab (shown below). Test the biologist's question at the 5% level.*
>
> `22.9,22.9,23.3,23.5,23.9,23.9,24.0,24.3,24.5,24.6,24.6,24.8,24.8,`
> `25.1,25.4,25.4,25.5,25.5,25.8,26.1,26.2,26.3,27.0,27.3,28.1`

[4] If σ 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."

1. $\alpha = 0.05$.
2. $H_0 : \mu = 24.3^o$C vs. $H_A : \mu \neq 24.3^o$C, where μ is the mean body temperature of ALL crabs.
3. A 1-Sample t-Test is required because (1) a quantitative variable (temperature) was measured, (ii) individuals from one group (or population) were considered (an ill-defined population of crabs), and (iii) σ is **UN**known.
4. The data appear to be part of an experimental study (the temperature was controlled) with no suggestion of random selection of individuals.
5. (i) n = 25 $\geq$ 15 and the sample distribution of crab temperatures appears to be only slightly right-skewed (Figure 18.4) and (ii) σ is **UN**known.
6. $\bar{x} = 25.0^o$C (Table 18.2).
7. $t = 2.713$ with 24 df (Table 18.2).
8. p-value = 0.0121 (Table 18.2).
9. H_0 is rejected because the p-value $< \alpha$.
10. It appears that the average body temperature of ALL crabs is greater than the ambient temperature of 24.3^oC.
11. I am 95% confident that the mean body temperature of ALL crabs is between 24.5^oC and 25.6^oC (Table 18.2).

Table 18.2. Results from 1-Sample t-Test for body temperature of crabs.

```
t = 2.7128, df = 24, p-value = 0.01215
95 percent confidence interval:
 24.47413 25.58187
sample estimates:
mean of x
   25.028
```

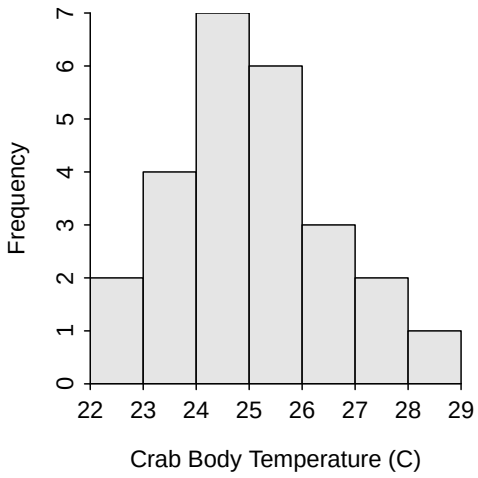


Figure 18.4. Histogram of the body temperatures of crabs exposed to an ambient temperature of 24.3^oC.

R Appendix:

```
df <- read.csv("data/CrabTemps.csv")
hist(~ct,data=df,xlab="Crab Body Temp (C)")
( ct.t <- t.test(df$ct,mu=24.3,conf.level=0.95) )
```

MODULE 19

2-SAMPLE T-TEST

Contents

While it is often useful to test whether a population mean differs from a specific value (i.e., with the 1-Sample t-Test of Module 18), there are many instances where interest is in whether means from two groups (or populations) differ. For example, is there a difference in mean income between males and females, in mean test scores between students from high- and low-income families, in mean percent body fat between raccoons from southern and northern Wisconsin, or in mean amount of milk produced from cows provided with a hormone or a placebo. In all of these situations, interest is identifying if a difference in population means exists between two groups (males and females, students from high- and low-income families, raccoons from southern and northern Wisconsin, cows given a hormone or a placebo). A **2-Sample t-Test** is used in these situations and is the subject of this module.

19.1 2-Sample t-Test Specifics

In a 2-Sample t-Test, $H_0 : \mu_1 = \mu_2$ states that the two population means are equal. This 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 H_0, the "parameter" is $\mu_1 - \mu_2$ and the corresponding statistic is $\bar{x}_1 - \bar{x}_2$. Thus, a 2-Sample t-Test is focused on the difference in population means.

When looking at the "general" test statistic formula (i.e., Equation (14.3.1)) of

$$\text{Test Statistic} = \frac{\text{Observed Statistic} - \text{Hypothesized Parameter}}{SE_{\text{Statistic}}}$$

it is apparent that the SE of $\bar{x}_1 - \bar{x}_2$ (i.e., the 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 Section 19.2), the standard error of $\bar{x}_1 - \bar{x}_2$ is

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where n_1 and n_2 are the sample sizes for the two groups and s_p^2 is the "pooled sample variance" computed as a weighted average of the two sample variances (s_1^2 and s_2^2), or

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The degrees-of-freedom for the 2-Sample t-Test with equal variances come from the denominator of the pooled variance calculation; i.e., $df = n_1 + n_2 - 2$. The specifics of the 2-Sample t-Test are in Table 19.1.

Table 19.1. 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:** $(\bar{x}_1 - \bar{x}_2) + t^* \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$
- **df:** $n_1 + n_2 - 2$
- **Assumptions:** $n_1 + n_2 \geq 40$, $n_1 + n_2 - 2 \geq 15$ and **each sample** (i.e., histogram) is not strongly skewed, OR **each sample** is normally distributed.
- **Use with:** Quantitative response, two groups (or populations), individuals are independent between groups.

◇ **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.**

A 2-Sample t-Test is often used to test an alternative hypothesis of simply finding a difference between the two groups. However, if the null hypothesis is rejected in these instances (thus, identifying a significant difference between the two groups), then care should be taken to specifically describe how the two groups differ. If the statistic is negative, then the mean of the first group is lower than the mean of the second group and, if the statistic is positive, then the mean of the first group is larger than the mean of the second group. The values of the confidence region should be used to identify how much larger or smaller the mean from one group is compared to the mean of the other group.

19.2 Testing for Equal Variances

As noted above, the methods of a 2-Sample t-Test differ depending on whether the two population variances are equal or not. This should present a problem to you because the population variances are parameters and are typically not known.[1] The question of whether these parameters are equal or not is answered with a hypothesis test, as has been done with all other questions about parameters.

A Levene's Test is used to determine whether two population variances are equal. The specifics of the Levene's test are not examined in detail here, rather you only need to know that $H_0 : \sigma_1^2 = \sigma_2^2$ is tested against $H_A : \sigma_1^2 \neq \sigma_2^2$. We will use computer software to compute the p-value for this test (without further detail). If the Levene's Test p-value $< \alpha$, then H_0 is rejected and the population variances are considered unequal. If the p-value $> \alpha$, then H_0 is not rejected and the population variances are considered equal.

19.2.1 Example - Corn and Fertilizers

Below are the 11-steps (Section 17.1) for completing a full hypothesis test for 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 19.2 to test the researcher's hypothesis at the 5% level.*

Table 19.2. Summary statistics 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

1. $\alpha = 0.05$.
2. $H_0 : \mu_s - \mu_o = 0$ vs $H_A : \mu_s - \mu_o > 0$, where μ is the mean plant height, s represents synthetic fertilizer, and o represents organic fertilizer. [*Note that positive differences represent larger values for synthetic fertilizer; thus, H_A represents synthetic fertilizer producing taller plants.*]
3. A 2-Sample t-Test is required because (i) a quantitative variable (height) was measured, (ii) two groups are being compared (synthetic and organic fertilizers), and (iii) plants in the two groups were **IN**dependent as the plants were not paired, plants were not tested over time, etc.
4. The data appear to be part of an experiment (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. (i) $n_s + n_o = 50 > 40$, (ii) individuals in the two groups are independent as discussed above, and (iii) the population variances appear to be equal because the Levene's Test p-value (0.1341) is $> \alpha$.
6. $\bar{x}_s - \bar{x}_0 = 51.46 - 47.49 = 3.97$. Additionally,

$$s_p^2 = \frac{(25-1)5.975^2 + (25-1)6.721^2}{25+25-2} = 40.44$$

and

$$SE_{\bar{x}_s - \bar{x}_o} = \sqrt{40.44\left(\frac{1}{25} + \frac{1}{25}\right)} = 1.799$$

[1] Actually, the population variances don't have to be known, it just needs to be known whether they are equal or not.

7. $t = \frac{3.97-0}{1.799} = \frac{3.97}{1.799} = 2.207$ with 25+25-2 = 48 df.
8. p-value = 0.0161.
9. The H_0 is rejected because the p-value $< \alpha$.
10. The average height of the corn plants appears to be greater for plants grown with synthetic fertilizer than for plants grown with organic fertilizer.
11. I am 95% confident that plants grown with synthetic fertilizer are more than 0.95 cm taller, on average, than plants grown with the organic fertilizer. [Note $3.97 - 1.677 * 1.799 = 3.97 - 3.02 = 0.95$.]

R Appendix:

```
( pval <- distrib(2.207,distrib="t",df=48,lower.tail=FALSE) )
( tstar <- distrib(0.95,distrib="t",df=48,type="q",lower.tail=FALSE) )
```

19.2.2 Example - Music and Anxiety

Below are the 11-steps (Section 17.1) for completing a full hypothesis test for the following situation:

> *An oral surgeon conducted an experiment to determine if background music decreased the anxiety level of patients during tooth extraction. Over a one-month period, 32 patients had a tooth removed while listening to music and 36 had a tooth removed without listening to music. 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 more 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 at the 5% level.*

1. $\alpha = 0.05$.
2. $H_0 : \mu_w - \mu_{wo} = 0$ vs $H_A : \mu_w - \mu_{wo} < 0$, where μ is the mean anxiety level, w represents patients "with", and wo represents "without" music. [*Note that negative numbers represent lower anxiety values in patients in the "with music" treatment. Thus, H_A suggests lower anxiety in paients with music.*]
3. A 2-Sample t-Test is required because (i) a quantitative variable (anxiety level) was measured, (ii) two groups are being compared (music or no music), and (iii) individuals in the two groups are independent (i.e., they were not paired, were not otherwise related, etc.).
4. The data appear to be an experiment as the music treatment was imparted by the surgeon, but there is no obvious random selection or allocation in this study.
5. (i) $n_w + n_{wo} = 68 > 40$, (ii) individuals in the two groups are independent as described above, and (iii)the two population variances appear to be equal because the Levene's Test p-value of 0.089 is greater than α.
6. $\bar{x}_w - \bar{x}_{wo} = 4.2 - 5.9 = -1.7$. Additionally,

$$s_p^2 = \frac{(32-1)1.2^2 + (36-1)1.9^2}{32+36-2} = 2.59$$

and

$$SE_{\bar{x}_w - \bar{x}_{wo}} = \sqrt{2.59\left(\frac{1}{32} + \frac{1}{36}\right)} = 0.391$$

7. $t = \frac{-1.7-0}{0.391} = -4.348$ with 32+36-2 = 66 df.

8. p-value < 0.00005.
9. H_0 is rejected because the p-value $< \alpha$.
10. The mean anxiety level appears to be lower when music was played for the patients.
11. I am 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. [Note -1.7+1.668*0.391 = -1.7+0.65 = -1.05.

R Appendix:

```
( pval <- distrib(-4.348,distrib="t",df=66) )
( tstar <- distrib(0.95,distrib="t",df=66,type="q") )
```

19.3 2-Sample t-Tests in R

19.3.1 Data Format

Data must be in stacked format (as described in Section 4.3.2) for a 2-Sample t-Test. Stacked data has measurements in one column and group labels for the measurement in another column. Thus, each row corresponds to a measurement and the group for a single individual. As an example, BOD measurements from either the inlet or outlet to an aquaculture facility are shown below. These data are stacked because each row corresponds to one individual (a water sample) with one column of (BOD) measurements and another column for which group the individual belongs.

```
   BOD    src
 6.782  inlet
 5.809  inlet
 8.063 outlet
 8.001 outlet
```

19.3.2 Levene's Test

Before conducting a 2-Sample t-Test, the assumption of equal population variances must be tested with Levene's test. The Levene's test is computed with `levenesTest()`, where the first argument is a model formula of the form `response~group`, where `response` represents the quantitative measurements and `group` represents the group factor variable.[2] The data.frame containing `response` and `group` is given in `data=`.

19.3.3 2-Sample t-Test

A 2-Sample t-Test is computed with `t.test()`, where the first argument is the same formula as in `levenesTest()` (and, thus, same `data=`). Additionally, the following arguments may need to be specified.

- `mu=`: The specific value in H_0. For a 2-Sample t-Test this is usually 0, which is the default.
- `alt=`: A string that indicates the type of H_A (i.e., `"two.sided"` (default), `"greater"`, or `"less"`).
- `conf.level=`: The level of confidence (default is `0.95`) used for the confidence region of $\mu_1 - \mu_2$.
- `var.equal=`: A logical value that indicates whether the two population variances should be considered equal or not. If `TRUE`, then the pooled sample variance is calculated and used in the standard error. The default `FALSE`, to assume UNequal variances.

◇ `var.equal=TRUE` must be in `t.test()` to assume equal variances. This is NOT the default.

[2] This is the same model formula introduced in Section 6.3 for summarizing multiple groups of data.

R computes the difference among groups as the alphabetically "first" level minus the alphabetically "second" level. For example, if the two levels are *inlet* and *outlet*, then R will compute $\bar{x}_{outlet} - \bar{x}_{inlet}$. If this is not the order you want, then you need to change the order of the levels by using `levels=` in `factor()` (as described in Modules 5 and 9). For example, the order of the levels of *src* in the *aqua* data.frame is changed below.

```
> aqua$src <- factor(aqua$src,levels=c("outlet","inlet"))
> levels(aqua$src)
[1] "outlet" "inlet"
```

19.3.4 Example - BOD in Aquaculture Water

Below are the 11-steps (Section 17.1) for completing a full hypothesis test for the following situation:

> *An aquaculture farm takes water from a stream and returns it to the stream after it has circulated through the fish tanks. The owner has taken steps to reduce the level of organic matter in the water released back into the stream. However, he is still concerned that water returned to the stream may contain heightened levels of organic matter. To determine if this is true, he took samples of water at the intake and, at other times, downstream from the outlet and recorded the biological oxygen demand (BOD) as a measure of the organics in the effluent (a higher BOD at the outlet would imply heightened levels of organics are being released to the stream). The owner's data are recorded in BOD.csv.* Test for any evidence (i.e., at the 10% level) to support the owner's concern.

1. $\alpha = 0.10$.
2. $H_0 : \mu_{outlet} - \mu_{inlet} = 0$ vs $H_A : \mu_{outlet} - \mu_{inlet} > 0$, where μ is the mean BOD, *outlet* represents the outlet source, and *inlet* represents the inlet source. [*Positive differences represent larger values at the outlet, which implies that BOD is higher in the water released from the facility. Thus, H_A represents the owner's concern. Further note that the order of subtraction could have been reversed such that the owner's concern would require a "less than" H_A. This is simply a matter of choice. However, note that the order of the levels has to be changed in R to use my choice of hypotheses.*]
3. A 2-Sample t-Test is required because (i) a quantitative variable (BOD level) was measured, (ii) two groups are being compared (outlet and inlet), and (ii) the individuals in the groups were **IN**dependent (note that it said that the outlet samples came from different times then the inlet samples).
4. The data appear to be part of an observational study with no obvious randomization.
5. (i) n = 20 >15 and the histograms (Figure 19.1) are inconclusive about the shape because of the small sample size in each group (it appears that the *inlet* data is not strongly skewed, whereas the *outlet* data is skewed, which may invalidate the results of this hypothesis test; however, I continued to make a complete example), (ii) individuals in the two groups are independent as discussed above, and (iii) the variances appear to be equal because the Levene's test p-value (=0.5913) is greater than α.
6. $\bar{x}_{outlet} - \bar{x}_{inlet} = 8.69 - 6.65 = 2.03$ (Table 19.3).
7. $t = 8.994$ with 18 df (Table 19.3).
8. p-value < 0.00005 (Table 19.3).
9. 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, the aquaculture facility appears to add to the biological oxygen demand of the water and the farmer's concern is warranted.
11. I am 90% confident that the mean BOD measurement at the outlet is AT LEAST 1.73 GREATER than the mean BOD measurement at the inlet (Table 19.3).

Table 19.3. Results from the 2-Sample t-Test for differences in BOD between the inlet and outlet of an aquaculture facility.

```
t = 8.994, df = 18, p-value = 2.224e-08
90 percent confidence interval:
 1.732704      Inf
sample estimates:
mean in group outlet  mean in group inlet
              8.6873                6.6538
```

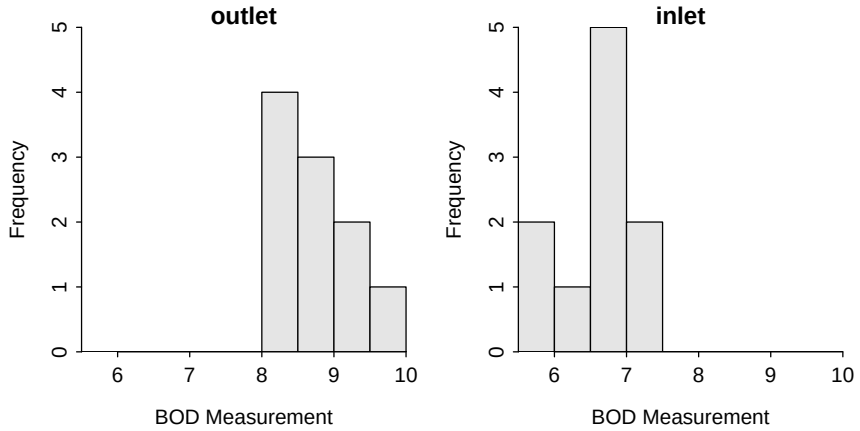


Figure 19.1. Histogram of the BOD measurements at the outlet and inlet of the aquaculture facility.

R Appendix:

```
aqua <- read.csv("data/BOD.csv")
aqua$src <- factor(aqua$src,levels=c("outlet","inlet"))
hist(BOD~src,data=aqua,xlab="BOD Measurement")
levenesTest(BOD~src,data=aqua)
( aqua.t <- t.test(BOD~src,data=aqua,var.equal=TRUE,alt="greater",conf.level=0.90) )
```

MODULE 20

CHI-SQUARE TEST

Contents

Situations where a categorical response variable is recorded would be summarized with a frequency or percentage table (see Modules 5 and 9). The appropriate test statistic in these situations is a chi-square rather than a t. The Chi-Square Test test statistic follows a chi-square distribution, which is introduced below. The rest of this module is dedicated to the general Chi-Square Test where the distribution of a categorical response variable is compared between two or more groups (or populations). The related goodness-of-fit test for a categorical response recorded for only one group (or population) is introduced in Module 21.

20.1 Chi-Square Distribution

A chi-square (χ^2) distribution is generally right-skewed (Figure 20.1), with the exact shape dictated by the degrees-of-freedom (df; as df increase, the sharpness of the skew decreases; Figure 20.1). In its simplest form, the χ^2 distribution arises as a sampling distribution for the χ^2 test statistic,

$$\chi^2 = \sum_{cells} \frac{(Observed - Expected)^2}{Expected}$$

where "Observed" and "Expected" represent the observed and expected individuals in the cells of frequency tables (see Module 5 and Module 9) and "cells" generically represents the number of cells in one of these tables. Thus, the χ^2 distribution arises from comparing frequencies in two tables.[1]

[1] Subsequent sections demonstrate how this test statistic is used to compare observed frequencies (i.e., from a sample) to a table of expected frequencies (i.e., from a null hypothesis).

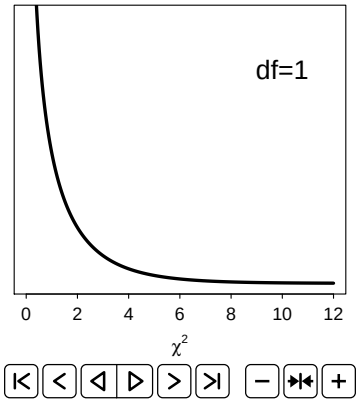


Figure 20.1. χ^2 distributions with varying degrees-of-freedom.

Unlike the normal and t distributions, the χ^2 distribution always represents 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 a "negative tail" being "folded over" onto what is the "positive tail." Thus, all probability (i.e., area) calculations on a χ^2 distribution represent the two-tailed alternative hypotheses.

Proportional areas on a χ^2 distribution are computed with `disrib()`, similarly to what was described for normal and t distributions in Modules 8, 12, and 18. The major difference for using `distrib()` with a χ^2 distribution is that `distrib="chisq"` must be used and the degrees-of-freedom must be given to `df=` (how to find the df will be discussed in subsequent sections). In addition, if calculating a p-value, then `lower.tail=FALSE` is always used because the upper-tail probability represents the two-tailed alternative hypothesis inherent to all Chi-Square Tests. For example, the area right of $\chi^2 = 6.456$ on a χ^2 distribution with 2 df is 0.0396 (Figure 20.2).

```
> ( distrib(6.456,distrib="chisq",df=2,lower.tail=FALSE) )
[1] 0.03963669
```

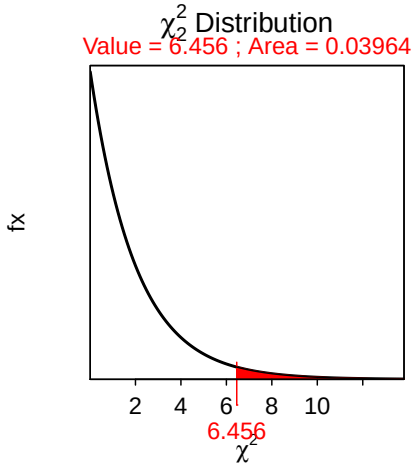


Figure 20.2. Depiction of the area to the right of $\chi^2 = 6.456$ on a χ^2 distribution with 2 df.

20.2 Chi-Square Test Specifics

Researchers commonly want to compare the distribution of individuals into the levels of a categorical variable among two or more groups (or populations). For example, researchers may want to determine if the disribution of failing students differs between males and females, if the distribution 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 have a categorical response variable (fail or not, play sport or not, plant species, answer to five-choice question) compared among two or more groups (gender, income category, two locations, neighboring counties). The Chi-Square Test, the subject of this module, can be used for each of these situations.[2]

20.2.1 Hypotheses

The statistical hypotheses for a Chi-Square Test are "wordy." To explore this, let's first assume that a two-way frequency table (see Module 9) will summarize the data where the rows correspond to separate groups and the columns correspond to levels of the response variable. In this organization, the Chi-Square Test null hypothesis is that the row percentages are equal – i.e., "the percentage distribution of individuals into the levels of the response variable is the same for all groups." The alternative hypothesis states that there is some difference among the row percentages – i.e., "the percentage distribution of individuals into the levels of the response variable is NOT the same for all groups."

As one example (more are shown below), consider the following:

> *An association of Christmas tree growers in Indiana sponsored a survey of Indiana households to help improve the marketing of Christmas trees. Of the 261 rural households, 64 had a natural tree (as compared to an artificial tree). Of the 160 urban households, 89 had a natural tree. Use these results to determine, at the 10% level, if the distribution of households with a natural tree differed between rural and urban households.*

The hypotheses for this situation are,

H_0 : "the distrubution of households into the tree types is the same for urban and rural households"
H_A : "the distrubution of households into the tree types is NOT the same for urban and rural households"

20.2.2 Tables

As noted above, all two-way frequency tables used for a Chi-Square Test will be organized such that the response variable forms the columns and the groups to be compared form the rows. With this organization, the row-percentage table becomes the table of primary interest 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-percentage table is equal, given sampling variability.

The observed raw data must be organized into a two-way frequency table as described in Module 9. For example, the Christmas tree data is summarized as in Table 20.1. 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 to determine if each row of the row-percentage table is statistically equivalent or not. Thus, the row-percentage table computed from the frequency table is useful when interpreting the results of a Chi-Square Test (Table 20.2).

[2] The Chi-Square Test is quite flexible and can be derived from different types of hypotheses than those described here.

Table 20.1. Frequency of individuals in urban and rural households that have a natural or an artificial Christmas tree.

	Tree Type		
Household	Natural	Artificial	
Urban	89	172	**261**
Rural	64	96	**160**
	153	**268**	**421**

Table 20.2. Percentage of individuals within urban and rural households that have a natural or an artificial Christmas tree.

	Tree Type		
Household	Natural	Artificial	
Urban	34.1	65.9	**100.0**
Rural	40.0	60.0	**100.0**
	36.3	**63.7**	**100.0**

The Chi-Square Test requires constructing a table of expected values that are derived from the null hypothesis. Specifically, the "expected" table contains the expected frequency of individuals in each level of the response variable for each group assuming that the distribution of responses does not differ among groups. These expected table are computed from the margins of the observed table, but are best explained with an illustrative example.

In the Christmas tree example, the null hypothesis states that there is no difference in the distribution of households with a natural tree between the rural and urban areas. Thus, under this null hypothesis, one would expect the proportion (or percentage) of households with a natural tree to be the same in both groups. The proportion of households with a natural tree, regardless of location, is $\frac{153}{421}$=0.363. Thus, 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 common 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 (i.e., $160 * 0.363 = 58.080$ expected to have a natural tree and $160 * (1 - 0.363) = 101.920$ expected to NOT have a natural tree.

These expected frequencies are computed directly and easily from the marginal totals of the observed frequency table (Table 20.1). For example, substituting the fractional representation of the decimal proportions into the calculation of the expected number of urban households with a natural tree gives $261 * \frac{153}{421}$=$\frac{261*153}{421}$=94.853[3]. A close examination of this formula and the marginal totals in Table 20.1 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 follow a similar pattern 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.

Thus, all expected values in a Chi-Square Test are calculated by multiplying the row and column totals of the frequency table and dividing by the total number of individuals. These expected values are summarized in a two-way table, called the expected frequencies table (Table 20.3).

[3]Note a slight difference here because 0.363 was rounded to three decimals, whereas the fraction is not rounded.

Table 20.3. The expected frequency of individuals in urban and rural households that have a natural or an artificial Christmas tree.

	Tree Type		
Household	Natural	Artificial	
Urban	94.853	166.147	**261**
Rural	58.147	101.853	**160**
	153	**268**	**421**

20.2.3 Specifics

The Chi-Square Test is characterized by a categorical response variable recorded for two or more groups (or populations). The specifics of the Chi-Square Test are in Table 20.4.

Table 20.4. Characteristics of a Chi-Square Test.

- **Null Hypothesis:** "The distribution of individuals into the levels of the response variable is the same for all groups"
- **Alternative Hypothesis:** "The distribution of individuals into the levels of the response variable is NOT the same for all groups."
- **Statistic:** Observed frequency table.
- **Test Statistic:** $\chi^2 = \sum_{cells} \frac{(Observed - Expected)^2}{Expected}$
- **df:** $(r-1)(c-1)$ where r = number of rows and c = number of columns
- **Assumptions:** Expected value for each category is ≥ 5.
- **Use with:** Categorical response, two or more groups (or populations).

In general, a confidence region is not constructed for a Chi-Square Test because of the complexity of the statistics and parameter. Thus, in this course, Step 11 for a hypothesis test will not be computed for a Chi-Square Test.

20.2.4 Example – Christmas Trees

Below are the 11-steps (Section 17.1) for a full hypothesis test for the Christmas tree example.

1. α=0.10.
2. H_0: "distribution of households by type of tree is the same for urban and rural households" vs. H_A: "distribution of households by type of tree is NOT the same for urban and rural households."
3. A Chi-Square Test is required because (i) a categorical response variable was recorded (type of tree) and (ii) two groups are being compared (urban and rural households).
4. The data appear to be part of an observational study with no clear indication of randomization.
5. The expected frequency in each of the four cells is greater than five (Table 20.3).
6. The observed frequency table is in Table 20.1.
7. $\chi^2 = \frac{(89-94.853)^2}{94.853} + \frac{(172-166.147)^2}{166.147} + \frac{(64-58.147)^2}{58.147} + \frac{(96-101.853)^2}{101.853} = 0.3611 + 0.2062 + 0.5891 + 0.3363 = 1.4927$ with 1 df.
8. p-value=0.2218.
9. H_0 is not rejected because the p-value is $> \alpha$.
10. There does not appear to be a significant difference in the distribution of Christmas tree types among rural and urban households.
11. Not performed for Chi-Square Test.

R Appendix:

```
( distrib(1.4927,distrib="chisq",df=1,lower.tail=FALSE) )
```

20.3 Chi-Square test in R (Raw Data)

The data for a Chi-Square Test may be computed from raw data on individuals (this section) or entered from summarized data (see Section 20.4). Raw data must be in stacked format where one column in the data.frame represents the response variable and another column represents the groups (see Sections 4.3.2 and 19.3). Raw data must be summarized into a two-way frequency table with `xtabs()` as described in Module 9. The two-way table must contain frequencies, not proportions or percentages (don't use `percTable()`), without marginal totals (don't use `addMargins()`).

The Chi-Square Test is performed with `chisq.test()`, which takes an observed frequency table either entered through `matrix()` or summarized with `xtabs()` as the first argument. The only other argument needed is `correct=FALSE` so that the continuity correction is not used.[4] The results of `chisq.test()` should be assigned to an object. The Chi-Square test statistic and p-value are extracted by simply printing the saved object. The expected frequency table is returned by appending `$expected` to the saved object.

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 groups. However, rejecting the null hypothesis does not indicate which groups are different. In addition, as mentioned previously, confidence intervals are generally not performed with a Chi-Square Test. A post-hoc method for helping determine which groups differ is obtained by observing the Pearson residuals.

A Pearson residual is computed for each cell in the table as,

$$\frac{Observed - Expected}{\sqrt{Expected}}$$

which is the appropriately signed square root of the parts in the χ^2 test statistic calculation. Therefore, cells that have Pearson residuals far from zero 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 groups 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`.

20.3.1 Example - Father Present at Birth

Below are the 11-steps (Section 17.1) for completing a full hypothesis test for the following 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.*

[4] The continuity correction is not used here simply so that the results using R will match hand-calculations. The continuity correction should usually be used.

1. α=0.05.
2. H_0:"distribution of fathers presence (or not) during the birth of their child is the same for all four hospitals" vs. H_A : "the distribution of fathers presence during the birth of their child is NOT the same for all four hospitals."
3. A Chi-Square Test is required because (i) a categorical variable (present or absent) was recorded and (ii) four groups are being compared (the hospitals).
4. The data appear to be part of an observational study with no clear indication of randomization (likely a voluntary response survey).
5. There are at least five individuals in each cell of the expected table (Table 20.5).
6. The statistic is the observed frequency table (Table 20.6).
7. χ^2=5.000 with 3 df (Table 20.7).
8. p-value=0.1718 (Table 20.7).
9. H_0 is not rejected because the p-value is $> \alpha$.
10. The distribution of father's presence (or not) at their child's birth does not seem to differ significantly among hospitals where that birth occurred. For comparative purposes, the row-percentage table is in Table 20.8.

R Appendix:

```
setwd("c:/data/")
fp <- read.csv("FatherPresent.csv")
( fp.obs <- xtabs(~hospital+father,data=fp) )
( fp.chi <- chisq.test(fp.obs,correct=FALSE) )
fp.chi$expected
percTable(fp.obs,margin=1,digits=1)
```

Table 20.5. Expected frequency table for father's presence (or absence) during child birth among four hospitals.

hospital	Absent	Present
A	15.25	59.75
B	15.25	59.75
C	15.25	59.75
D	15.25	59.75

Table 20.6. Observed frequency table for father's presence (or absence) during child birth among four hospitals.

hospital	Absent	Present
A	9	66
B	15	60
C	18	57
D	19	56

Table 20.7. Results from the Chi-Square Test for differences in father's presence during child birth among four hospitals.

```
X-squared = 5.0003, df = 3, p-value = 0.1718
```

Table 20.8. Percentage of father's presence (or absence) during child birth among four hospitals.

```
        father
hospital Absent Present   Sum
       A   12.0    88.0 100.0
       B   20.0    80.0 100.0
       C   24.0    76.0 100.0
       D   25.3    74.7 100.0
```

20.4 Chi-Square test in R (Summarized Data)

Two-way frequency datat that has already been summarized (outside of R) must be entered into a two-dimensional matrix. The frequencies must first be entered into a vector with the first row of values followed by the second row and so on. This vector is then the first argument to `matrix()`, which will also include the number of rows in the frequency table in `nrow=` and `byrow=TRUE` (which causes the values in the vector to be entered into the matrix in a row-wise manner). The process of entering summarized data into a matrix is better explained by example.

Suppose that you are given this observed frequency table.

	Species						
Location	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 first entered into a vector called `freq` below, which is then transformed into a two-row matrix called `obstbl`.

```
> ( 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
> ( 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
```

The matrix is more informative if the rows and columns are named with `rownames()` and `colnames()` as shown below.

```
> 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
```

Once this observed table is constructed, the chi-square tests is performed exactly as described in the previous sections (i.e., starting with `chisq.test()`).

20.4.1 Example - Apostle Islands Plants

Below are the 11-steps (Section 17.1) for completing a full hypothesis test for the following 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.*

	Species						
Location	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**

1. α=0.05.
2. H_0: "the distribution of dominant plants species is the same between Devil's Island and the Bayfield Peninsula" vs. 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 (i) a categorical variable with six levels (plant species) was recorded and (ii) two grouops are being compared (Devil's Island and Bayfield Peninsula).
4. The data appear to be part of an observational study where the plots were randomly selected.
5. There are more than five individuals in each cell of the expected table (Table 20.9).
6. The statistic is the observed frequency table given in the background.
7. χ^2=16.54 with 5 df (Table 20.10).
8. p-value=0.0055 (Table 20.10).
9. H_0 is rejected because the p-value is $< \alpha$.
10. 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 (Table 20.11) and the row-percentage table (Table 20.12) both suggest that the biggest difference between the two locations is due to "plant A."[5]

R Appendix:

```
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.chi <- chisq.test(ai.obs) )
ai.chi$expected
ai.chi$residuals
percTable(ai.obs,margin=1,digits=1)
ai.obs1 <- ai.obs[,-1]
( ai.chi1 <- chisq.test(ai.obs1) )
```

[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.

Table 20.9. Expected frequency table for dominant plant species on Devil's Island and the Bayfield Peninsula.

```
    A  B  C  D E F
DI 48 17 11 10 9 5
BP 48 17 11 10 9 5
```

Table 20.10. Results from the Chi-Square Test for differences in the distribution of dominant plant species between Devil's Island and the Bayfield Peninsula.

```
X-squared = 16.5442, df = 5, p-value = 0.00545
```

Table 20.11. Pearson residuals from the Chi-Square Test for differences in the distribution of dominant plant species between Devil's Island and Bayfield Peninsula.

```
           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
```

Table 20.12. Percentage of dominant plant species within each location (Devil's Island and Bayfield Peninsula).

```
    A  B  C  D  E F Sum
DI 34 22 14 13 12 5 100
BP 62 12  8  7  6 5 100
```

MODULE 21

GOODNESS-OF-FIT TEST

Contents

It is common to determine if the frequency of individuals in the levels of a categorical response variable follow frequencies suggested by a particular theory or 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 category and one minus the proportion in the hypothesis for the other category. 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 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 a Goodness-of-Fit Test, the subject of this module.

21.1 Goodness-of-Fit Test Specifics

21.1.1 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," relating whether the "distribution" of individuals into the levels of the response variable follows a specific theoretical distribution or not. The null hypothesis will generally be like H_0 : "the 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 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 hypotheses are simpler, but you must be more careful, when there are only two levels of the response variable. For example, a research hypothesis of "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 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."

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 from Section 20.1, however, that the chi-square test statistic always represents a two-tailed situation. Thus, the H_A here reflects 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.

21.1.2 The Tables

For a Goodness-of-Fit Test, the data are summarized in an observed frequency table as in Module 5. Additionally, a table of expected frequencies must be constructed from the theoretical distribution in the null hypothesis and the total number of observed individuals (n). Specifically, the expected frequencies are found by multiplying the expected proportions from the theoretical distribution in the null hypothesis by n. For example, consider this situation:

> Bath and Buchanan (1989) surveyed residents of Wyoming by distributing a mailing to random 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. These expectations are summarized in Table 21.1.

Table 21.1. 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.

	Frequency	
Category	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.**

Consider the following situation where 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 : "the proportion of RY, RG, WY, and WG individuals will be $\frac{9}{16}$, $\frac{3}{16}$, $\frac{3}{16}$, and $\frac{1}{16}$, respectively"

H_A : "the proportion of RY, RG, WY, and WG individuals will NOT be $\frac{9}{16}$, $\frac{3}{16}$, $\frac{3}{16}$, and $\frac{1}{16}$, 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:

- $\frac{9}{16} \cdot 566 = 318.375$ would be expected to be round, yellow.
- $\frac{3}{16} \cdot 566 = 106.125$ would be expected to be round, green.
- $\frac{3}{16} \cdot 566 = 106.125$ would be expected to be wrinkled, yellow.
- $\frac{1}{16} \cdot 566 = 35.375$ would be expected to be wrinkled, green.

These expected frequencies are summarized in Table 21.2.

Table 21.2. Expected and observed frequency of 566 pea seeds in four types.

	Frequency	
Category	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

The hypothesis test method developed in the following sections will be used to determine if the differences between the expected and observed frequencies is "large" enough to suggest that the observed frequencies do not support the distribution represented in the null hypothesis.

21.1.3 Specifics

The Goodness-of-Fit Test is characterized by a single categorical response variable. The hypotheses tested usually cannot be converted to mathematical symbols and are thus "wordy." Specifics of the Goodness-of-Fit Test are in Table 21.3.

Table 21.3. 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 Statistic:** $\chi^2 = \sum_{cells} \frac{(Observed - Expected)^2}{Expected}$
- **df:** Number of levels minus 1.
- **Assumptions:** Expected value in each level is ≥ 5.
- **Confidence Interval (for one level):** $\hat{p} \pm Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- **Use with:** Categorical response, one group (or population), comparing to a theoretical distribution.

It is cumbersome to produce a confidence interval 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 response variable). Confidence intervals can be calculated for the proportion in each level as shown below. However, confidence intervals will only be "hand"-calculated when there are two levels. When using R (as discussed in a subsequent section), confidence intervals will be computed for all levels, 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 variable with that 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 16.2 and repeated below:

$$\text{“Statistic”} + \text{“scaling factor”} * SE_{statistic}$$

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, $Z^*=\pm 1.960$.[1] 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 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.

21.1.4 Example - $1 Coins

Below are the 11-steps (Section 17.1) for completing a full hypothesis test for 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.*

1. α=0.10.
2. H_0: “the proportion of Ashland residents that oppose replacing the $1 bill with the $1 coin is 0.77” vs. 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) a single categorical variable was recorded (opinion about $1 coin), (ii) a single group (or population) was considered (Ashland residents), and (iii) 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).

[1] This Z^* is computed with `distrib(0.975,type="q")`

Level	Frequency Expected	Observed
"Oppose"	61.6	54
"Do Not Oppose"	18.4	26
Total	80	80

7. $\chi^2=\frac{(61.6-54)^2}{55}+\frac{(18.4-26)^2}{25}=0.938+3.139=4.077$ with $2-1=1$ df.
8. p-value=0.0435.
9. 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. I am 90% confident that the proportion of all Ashland residents opposed to the \$1 coin is between 0.596 and 0.767. $[\frac{54}{80}\pm 1.645*\sqrt{\frac{0.68125*0.31875}{80}}=0.68125\pm 1.645*0.0521=0.68125\pm 0.0857=(0.5956, 0.7670).]$

R Appendix:

```
( distrib(4.077,distrib="chisq",df=1,lower.tail=FALSE) )
( distrib(0.95,type="q") )
```

21.2 Goodness-of-Fit Test in R

21.2.1 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 summarized data using `c()` or raw data is summarized to a frequency table with `xtabs()` as in Module 5.

For example, suppose that the frequency of shrike observations in the "mid-successional", "open", "scattered trees", "woods", and "wetland" habitats is known to be 43, 1456, 112, 44 and 6, respectively. These summarized values are entered directly into a named vector below.

```
> ( obs <- c(MidSucc=43,Open=1456,ScatTree=112,Woods=6,Wetland=44) )
 MidSucc     Open ScatTree    Woods  Wetland
      43     1456      112        6       44
```

However, instead of having summarized frequencies, suppose that the individual habitat observations were stored in a variable called `hab.use` in the `df` data.frame. These raw data must be summarized into a frequency table.

```
> ( obs <- xtabs(~hab.use,data=shrike.raw) )
hab.use
 MidSucc     Open ScatTree  Wetland    Woods
      43     1456      112        6       44
```

21.2.2 Goodness-of-Fit Test

The Goodness-of-Fit Test is computed with `chisq.test()` with a observed frequencies as the first argument and the following arguments:

- `p=`: a vector of expected proportions for the levels of the theoretical distribution.
- `rescale.p=TRUE`: rescales the values in `p=` to sum to 1. Rescaling is useful if the proportions in `p=` were rounded or are expected frequencies.
- `correct=FALSE`: indicates to not use a "continuity correction."[2]

The results from `chisq.test()` should be assigned to an object so that useful information can be extracted. The chi-square test statistics and p-value are extracted by typing the name of the saved object, the expected values are extracted by appending `$expected` to the object, and a visual of the p-value is obtained by submitting the object to `plot()`. In addition, confidence intervals for the proportions of individuals in each level are constructed by submitting the saved object to `gofCI()`.

21.2.3 Example - Loggerhead Shrikes

Below are the 11-steps (Section 17.1) for completing a full hypothesis test for the following situation:

> *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. α=0.05.
2. H_0:"distribution of habitat use by shrikes is the same as the proportions of available habitat" vs. H_A:"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 (i) a categorical variable was recorded (habitat use), (ii) a single group (or population) was considered (shrikes in this area), and (iii) the observed distribution is compared to a theoretical distribution.
4. 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.
5. There are more than five individuals expected in each habitat level (Table 21.4).
6. The statistic is the observed frequency table in Table 21.4.
7. χ^2=2345.1 with 4 df (Table 21.5).
8. p-value< 0.00005 (Table 21.5).
9. H_0 is rejected because the p-value< α.
10. The shrikes do not appear to use habitats in the same proportions as the availability of the habitat.

[2]Some statisticians 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. We will not use the continuity correction in this course so that R calculations will match hand calculations.

11. The 95% confidence intervals for the proportion of use in each habitat level are in Table 21.6. From these results it appears that the shrikes use the "open" habitat much more often and the "woods" habitat mush less often than would be expected if they used all habitats in proportion to their availability.

R Appendix:

```
( obs <- c(Open=1456,MidSucc=43,ScatTree=112,Woods=6,Wetland=44) )
( p.exp <- c(Open=0.358,MidSucc=0.047,ScatTree=0.060,Woods=0.531,Wetland=0.004) )
( shrike.chi <- chisq.test(obs,p=p.exp,rescale.p=TRUE) )
data.frame(obs=shrike.chi$observed,exp=shrike.chi$expected)
gofCI(shrike.chi,digits=3)
```

Table 21.4. Observed and expected frequencies for the Goodness-of-Fit Test for shrike habitat use.

	obs	exp
Open	1456	594.638
MidSucc	43	78.067
ScatTree	112	99.660
Woods	6	881.991
Wetland	44	6.644

Table 21.5. Results from the Goodness-of-Fit Test for shrike habitat use.

```
X-squared = 2345.071, df = 4, p-value < 2.2e-16
```

Table 21.6. Observed proportions, 95% condidence intervals for the proprtions, and expected proportions for shrike habitat use.

	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

21.2.4 Example - Modes of Fishing

The 11-steps (Section 17.1) for a hypothesis test for this example is 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. α=0.10.
2. 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" vs. 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." [*Thought process – the two-thirds for "boat" fishing is split to one-third each for private and charter boats; the one-third, or two-sixths, for "land" fishing is split to one-sixth each for beach and pier fishing.*]
3. A Goodness-of-Fit Test is required because (i) a categorical variable was recorded (mode), (ii) a single group (or population) was considered (Southern California Sportfishers), and (iii) the obseved distribution is compared to a theoretical distribution.
4. The data appear to be part of an observational study where the individuals were not obviously (probably were not) randomly selected.
5. There are more than five individuals expected in each mode (Table 21.7).
6. The statistic is the observed frequency table in Table 21.7.
7. χ^2=31.980 with 3 df (Table 21.8).
8. p-value< 0.00005 (Table 21.8).
9. H_0 is rejected because the p-value< α.
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 in Table 21.9. From these results it is apparent that the users use the beach slightly less than expected and use charter boats slightly more than expected. The use of the pier and private boats are not different from expected.

R Appendix:

```
setwd("c:/data/")
sf <- read.csv("FishingModes.csv")
obs <- xtabs(~mode,data=sf)
p.exp <- c(beach=1/6,boat=1/3,charter=1/3,pier=1/6)
( sf.chi <- chisq.test(obs,p=p.exp,rescale.p=TRUE) )
data.frame(obs=sf.chi$observed,exp=sf.chi$expected)
gofCI(sf.chi,digits=3)
```

Table 21.7. Observed and expected frequencies for the Goodness-of-Fit Test for modes of fishing.

	obs.mode	obs.Freq	exp
beach	beach	134	197
boat	boat	418	394
charter	charter	452	394
pier	pier	178	197

Table 21.8. Results from the Goodness-of-Fit Test for modes of fishing.

```
X-squared = 31.9797, df = 3, p-value = 5.285e-07
```

Table 21.9. Observed proportions, 95% condidence intervals for the proprtions, and expected proportions for modes of fishing.

	p.obs	p.LCI	p.UCI	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

21.2.5 Example - Mendelian Genetics II

Below are the 11-steps (Section 17.1) for completing a full hypothesis test for the following situation:

> *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 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 was hypothesized.*

1. α=0.01.
2. H_0:"The proportion of progeny with red eyes is 0.75" vs. H_A:"The proportion of progeny with red eyes is NOT 0.75."
3. A Goodness-of-Fit Test is required because (i) a categorical variable was recorded (red eye color or not), (ii) a single group (or population) was considered in the experiment, and (iii) the observed distribution is compared to a theoretical distribution.
4. The data appear to be experimental in that a specific cross was made and the environment in which they were raised was controlled. Progeny were randomly selected.
5. There are more than five individuals expected in each eye level (Table 21.10).
6. The appropriate statistic is the observed frequency table in Table 21.10.
7. χ^2=34.49 with 1 df (Table 21.11)
8. p-value< 0.00005 (Table 21.11).
9. 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. From the 95% confidence intervals in Table 21.12 it appears that the proportion of progeny with red eyes was between 0.464 and 0.620, which indicates that there were many fewer red-eyed progeny than would be expected from the Mendelian theory.

R Appendix:

```
obs <- c(red=82,nonred=151-82)
p.exp <- c(red=0.75,nonred=0.25)
( m.chi <- chisq.test(obs,p=p.exp,rescale.p=TRUE) )
data.frame(obs=m.chi$observed,exp=m.chi$expected)
gofCI(m.chi,digits=3)
```

Table 21.10. Observed and expected frequencies for the Goodness-of-Fit Test for the genetic cross experiment.

```
       obs    exp
red     82 113.25
nonred  69  37.75
```

Table 21.11. Results from the Goodness-of-Fit Test for the genetic cross experiment.

```
X-squared = 34.4923, df = 1, p-value = 4.279e-09
```

Table 21.12. Observed proportions, 95% condidence intervals for the proprtions, and expected proportions for eye colors in the genetic cross experiment.

```
       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
```

MODULE 22

FILTERING DATA IN R

Contents

In the Module 4, you learned how to retrieve data from the class webpage, enter your own data into a CSV file, load that data into R, and how to view that data in R. In this module, we will learn how to create subsets (i.e., filter) a data.frame into smaller data.frames. For example, you may want to create a data.frame that contains just male bears from a data.frame with both male and female bears, or a data.frame that only contains sales during summer months from a data.frame that contains all sales. Less often you may wish to eliminate a particular individual from the data.frame, perhaps if it is considered to be erroneous.

22.1 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. 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 22.1). The results from `filterD()` should be assigned to an object, which is then the name of the new data.frame.

The following are examples of new data.frames created from *bears* (which was created in the previous module). The name of the new data.frame (i.e., object 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.

Table 22.1. Condition operators used in `filterD()` and their results. Note that *var* 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 (see examples in main text).

Condition Operator	Individuals Returned from Original Data Frame
var == *value*	all individual that are **equal** to the given value
var != *value*	all individuals that are **NOT equal** to the given value
var > *value*	all individuals that are **greater than** the given value
var >= *value*	all individuals that are **greater than or equal** to the given value
var < *value*	all individuals that are **less than** the given value
var <= *value*	all individuals that are **less than or equal** to the given value
condition , *condition*	all individuals that **meet both conditions**
condition | *condition*	all individuals that **meet one or both conditions**[1]

- 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 weighed at least 150 kg.

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

22.2 Selecting Individuals

In some instances, you may need to select or exclude an individual from a data.frame. Positions within an object are identified within square brackets. As data.frames are two-dimensional objects they are indexed by a row and a column, in that order. For example, the item in the third row and second column of `bears` is selected below.

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

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 with `bears$weight.kg`. As a better example, the entire third row is selected below (note that the column designation was omitted).

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

Multiple rows are selected by combining row indices together with `c()`. For example, the third, fifth, and eighth rows are selected below (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
```