MODULE 1

WHY STATISTICS IS IMPORTANT

1.1 Realities

The city of Ashland performed an investigation in the area of Kreher Park (Figure **??**) 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 **??**. 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 important 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 **??**). Methods for choosing the number of individuals to select and how to select those individuals are discussed in Module **??**.

Suppose that the WDNR measured the concentration of lead at each of the 119 locations shown in Figure **??**. Further suppose that they presented their results at a public meeting by simply showing the list of lead

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

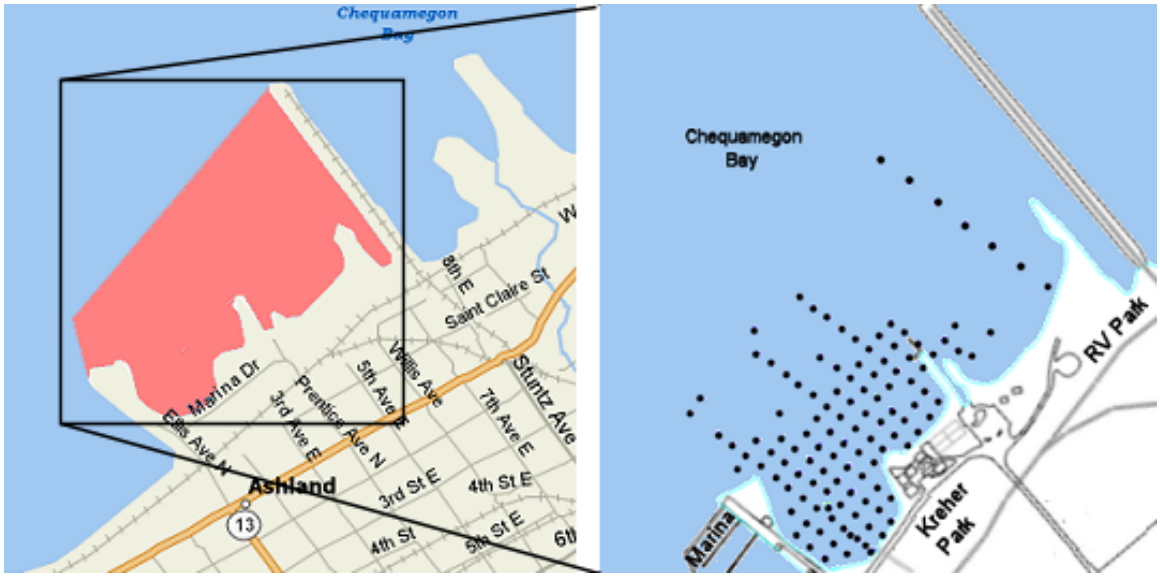


Figure 1.1. Location of the Ashland superfund site (left) with the location of 119 historical sediment sampling sites (right).

concentration measurements (Table **??**).[2] Is it easy to make conclusions about what these data mean from this type of presentation?

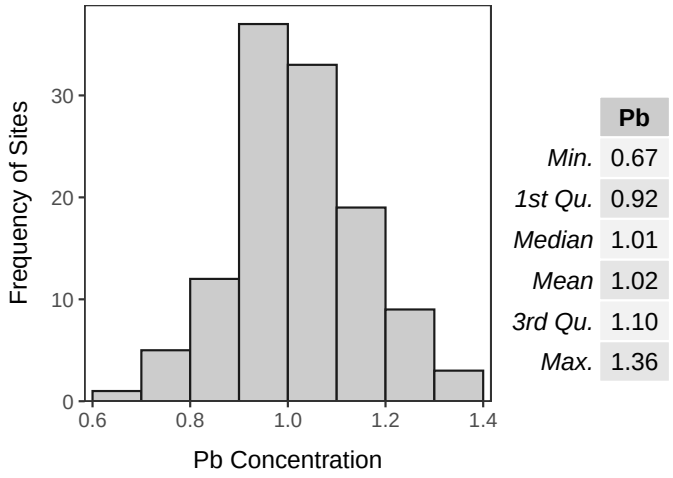
Table 1.1. Lead concentration ($\mu g \cdot m^{-3}$) from 119 sites in Kreher Park superfund site.

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 **??**) 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 **??**. 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 data.

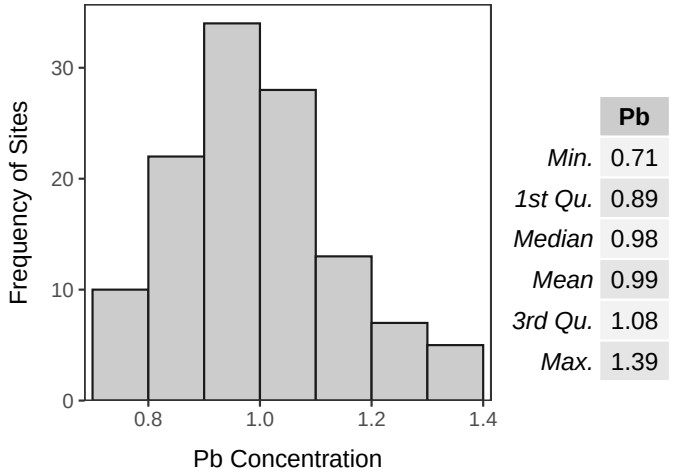
A critical question at this point is whether or not the results from the one sample of 119 sites perfectly represents the results for the entire area. One way to consider this question is to examine the results obtained from another sample of 119 sites. The results from this second sample (Figure **??**) 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.

[2] These are hypothetical data for this site.



	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.



	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 **??**) 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 **??** and **??**, than it was from the full list of lead measurements in Table **??**. 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 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.

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

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

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 **??**). 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.

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.

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

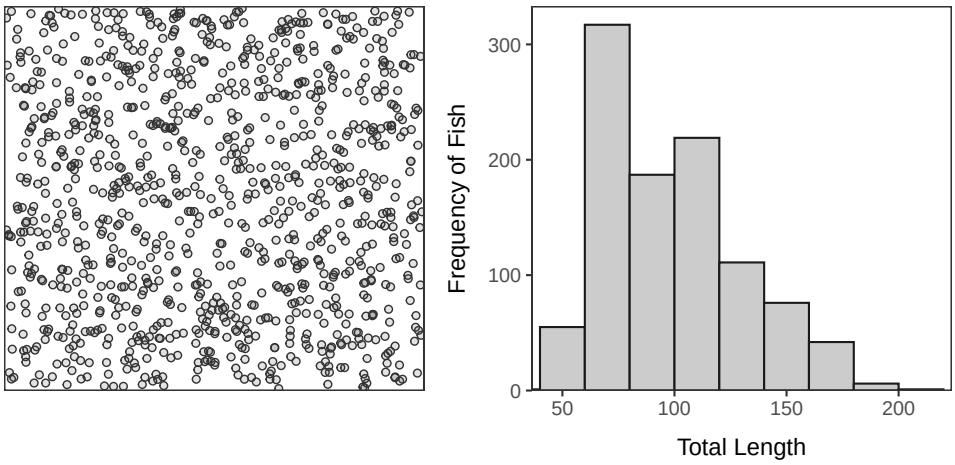


Figure 2.1. Schematic representation of individual fish (i.e., dots; **Left**) and histogram (**Right**) of the total length of the 1015 fish in Square Lake.

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] 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 **??** 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 **??** and Figure **??**. The sample mean of 107.5 mm is the best "guess" at the population mean. Not surprisingly from the discussion in Module **??**, the sample mean does not perfectly equal the population mean.

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

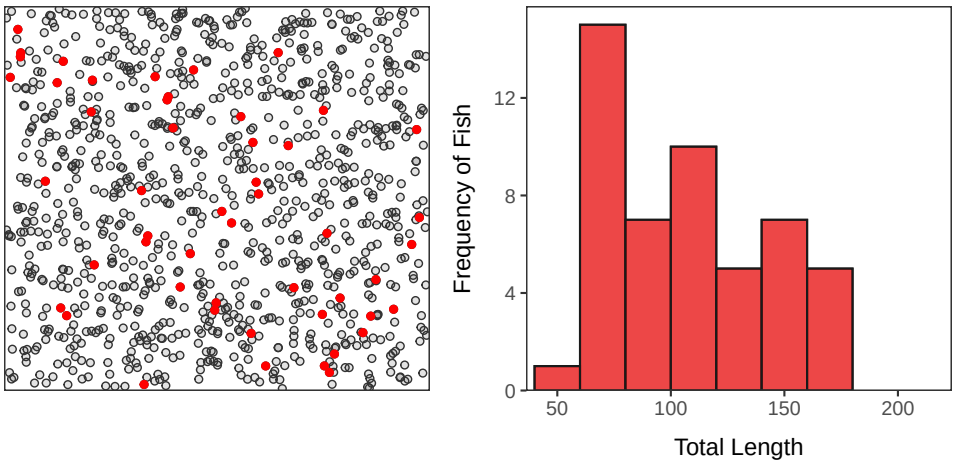


Figure 2.2. Schematic representation (**Left**) of a sample of 50 fish (i.e., red dots) from Square Lake and histogram (**Right**) of the total length of the 50 fish in this sample.

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.

The IVPPSS process is illustrated for the following situation:

A University of New Hampshire graduate student (and Northland College alum) investigated

[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.

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).

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

[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.

◇ **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 **??**) consider how statistics differ among samples. Table **??** and Figure **??** 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 **??** and Figure **??**) 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 **??**.

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.**

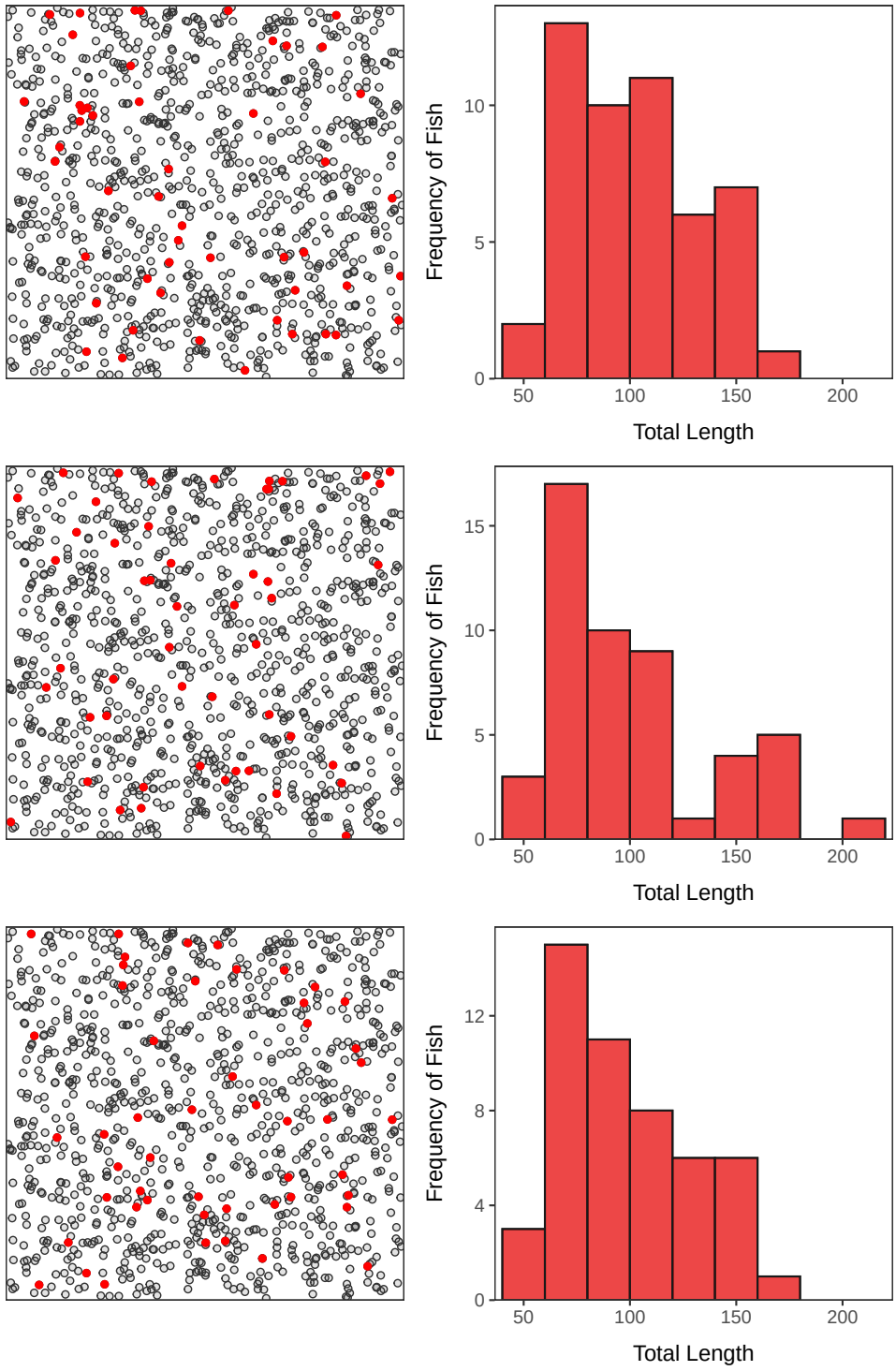


Figure 2.3. Schematic representation (**Left**) of three samples of 50 fish (i.e., red dots) from Square Lake and histograms (**Right**) of the total length of the 50 fish in each sample.

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 **??**). **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 **??**).

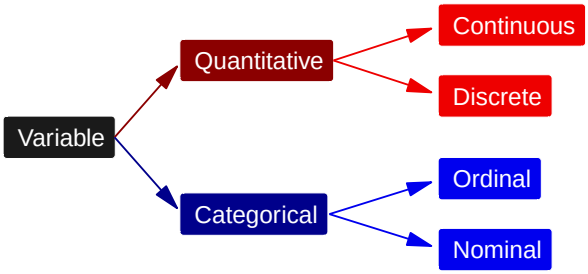


Figure 2.4. Schematic representation of the four types of variables.

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

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 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

[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 **??**.

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 **??**-Left.[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.

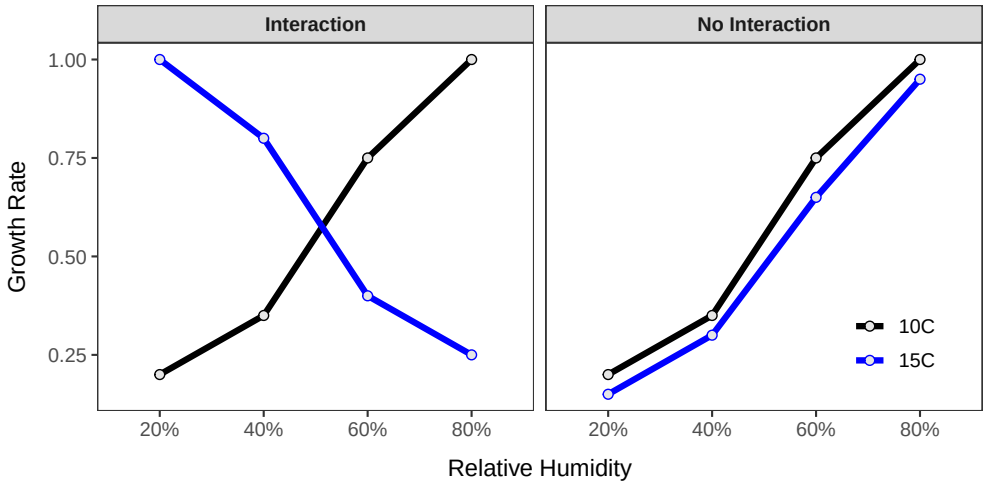


Figure 3.1. Mean growth rates in a two-factor experiment that depict an interaction effect (left) and no interaction effect (right).

Multi-factor experiments are used to detect the presence or absence of interaction, not just the presence of it. The hypothetical results in Figure **??**-Right 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 **??**.
[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."

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, the random numbers would be divided into 8 groups each with 15 numbers.

```
> sample(120)
```

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

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 (only the first two are shown because of space constraints).

	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,$\cdots$

$\diamond$ 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.[6] 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 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.[7] 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.

[6] Other common designs include blocked, Latin square, and nested designs.
[7] Thus, it is very important to first perform an IVPPS as discussed in Module **??**.

For example, consider that you want to determine the mean length of all fish in a particular lake (e.g., Square Lake from Module **??**). 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 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.[8] 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 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

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

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

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

MODULE 4

UNIVARIATE SUMMARIES)

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

4.1 Quantitative Variable

Two data sets will be considered in this section about summarizing quantitative variables. The first data set consists of the number of open pit mines in countries with open pit mines (Table **??**).[1] The second data set is Richter scale recordings for 15 major earthquakes (Table **??**).

Table 4.1. 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 4.2. 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

4.1.1 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 **??**. As you will see in Module **??**, "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

[1] These data were collected from this page. See Section **??** for how to enter these data into R.

mean, and sample standard deviation. These measures could be computed from every individual, if the population was known. The values would then be *parameters* and would be preceded by "population" – e.g., population median, population mean, and population standard deviation.[2]

Median

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

The process for finding the median consists of three steps,[3]

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 **??** are (data are wrapped for convenience),

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 **??**). 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.

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 the median, and then finding the medians of the two halves.[4] 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.[5]

[2] See Module **??** for clarification on the differences between populations and samples and parameters and statistics.
[3] 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.
[4] You should review how a median is computed before proceeding with this section.
[5] Some authors put the median into both halves when n is odd. The difference between the two methods is minimal for large n.

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 **??**). 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 **??**). Recall from above (Section **??**) 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 $Q1$ of 1 to a $Q3$ of 4 and the IQR for the earthquake data is from a $Q1$ of 6.5 to a $Q3$ of 7.

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

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

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,

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,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.1.1}$$

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.**

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

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{4.1.2}$$

The calculation of the standard deviation of the earthquake data (Table **??**) is facilitated with the calculations shown in Table **??**. In Table **??**, 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 $s = \sqrt{0.484} = 0.696$.

From this, on average, each earthquake is approximately 0.70 Richter Scale units different than the average earthquake in these data.

Table 4.3. 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

◇ 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 **??** 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 **??** 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 **??**) 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 **??**).

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 **??**).

Table 4.4. 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 **??**).

Table 4.5. 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**.

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.

4.1.2 Graphical Summaries

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 **??**).

The frequency of individuals in each class is then counted (shown in the second row of Table **??**). The plot is prepared with values of the classes forming the x-axis and frequencies forming the y-axis (Figure **??**A). The first bar added to this skeleton plot has the bottom-left corner at 0 and the bottom-right corner at 2 on the

Table 4.6. 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

x-axis, and a height equal to the frequency of individuals in the 0-1 class (Figure **??**B). 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 2-3 class (Figure **??**C). This process is continued with the remaining classes until the full histogram is constructed (Figure **??**D).

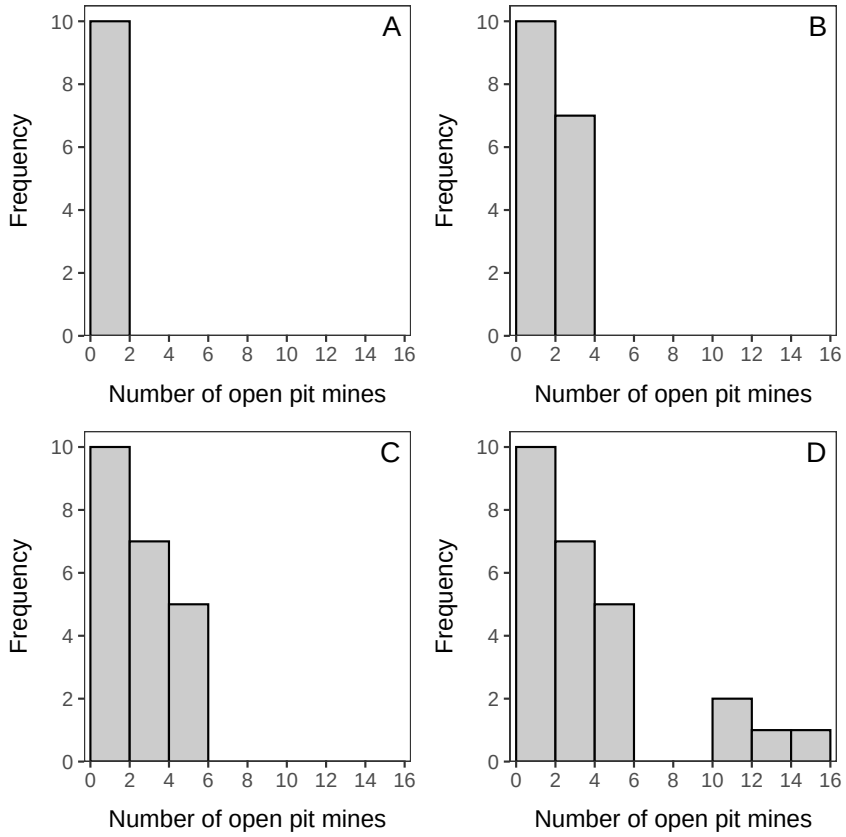


Figure 4.1. Steps (described in text) illustrating the construction of a histogram.

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 **??**).

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 sections). The five-number summary may be displayed

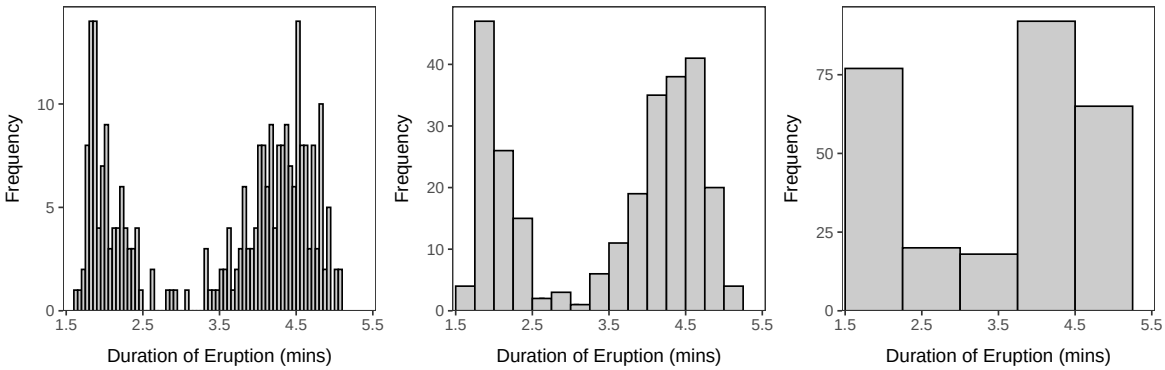


Figure 4.2. Histogram of length of eruptions for Old Faithful geyser with varying number of bins/classes.

as a **boxplot**. A traditional boxplot (Figure **??**-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 **??**-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 traditional and modern boxplots will be the same.

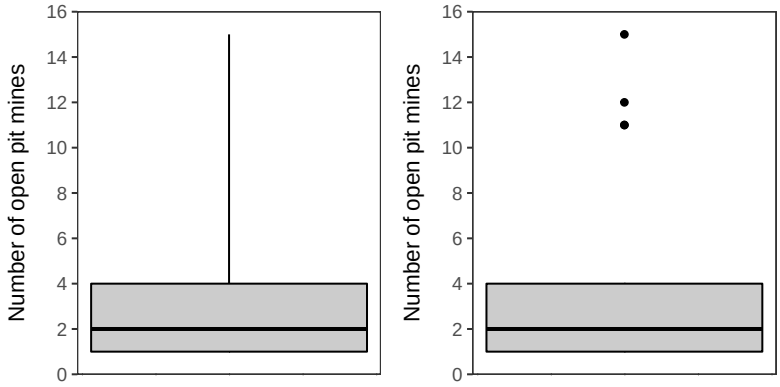


Figure 4.3. Traditional (Left) and modern (Right) boxplots of the open pit mine data.

4.2 Categorical Variable

In this section, methods to construct tables and graphs for categorical data are described. Interpretation of the results is demonstrated in the next module. The concepts are illustrated with data about MTH107 students from the Winter 2020 semester. 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 **??**.

Table 4.7. Whether (Y) or not (N) MTH107 was required for eight individuals in MTH107 in Winter 2020.

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

4.2.1 Numerical Summaries

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 **??**) 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 4.8. Frequency table for whether MTH107 was required (Y) or not (N) for eight individuals in MTH107 in Winter 2020.

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 **??**.

Table 4.9. Frequency tables for whether (Y) or not (N) MTH107 was required (Left) and year-in-school (Right) for all individuals in MTH107 in Winter 2020.

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 **??**) for students in MTH107 is constructed from Table **??** 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 **??**).

4.2.2 Graphical Summaries

Bar Charts

Bar charts are used to display the frequency or percentage of individuals in each level of a categorical variable. Bar charts 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

Table 4.10. Percentage tables for whether (Y) or not (N) MTH107 was required (Left) and year-in-school (Right) for all individuals in MTH107 in Winter 2020.

Required	Perc
Y	55.9
N	44.1

Year	Perc
Fr	27.9
So	17.6
Jr	42.6
Sr	13.2

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

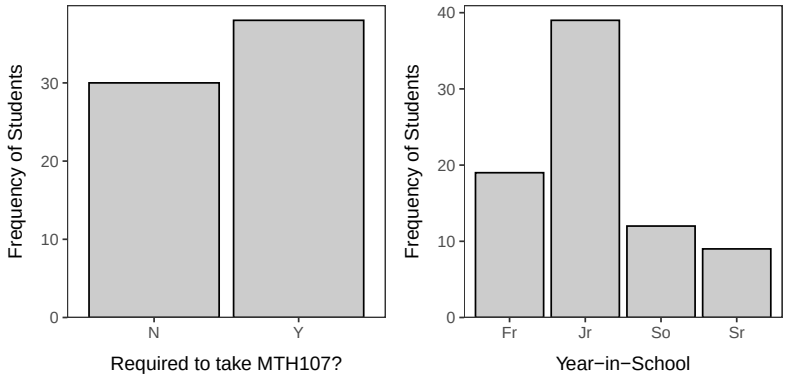


Figure 4.4. Bar charts of the frequency of individuals in MTH107 during Winter 2010 by whether or not they were required to take MTH107 (**Left**) and year-in-school (**Right**).

◇ **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.**

MODULE 5

UNIVARIATE EDA

5.1 Quantitative Variable

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 with these four 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 **??**. 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 **??** are used here.

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

5.1.1 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 **??**). The relative appearance of the tails is used to identify the shape of a distribution.

If the left- and right-tail are approximately equal in length and height, then the distribution is **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**.

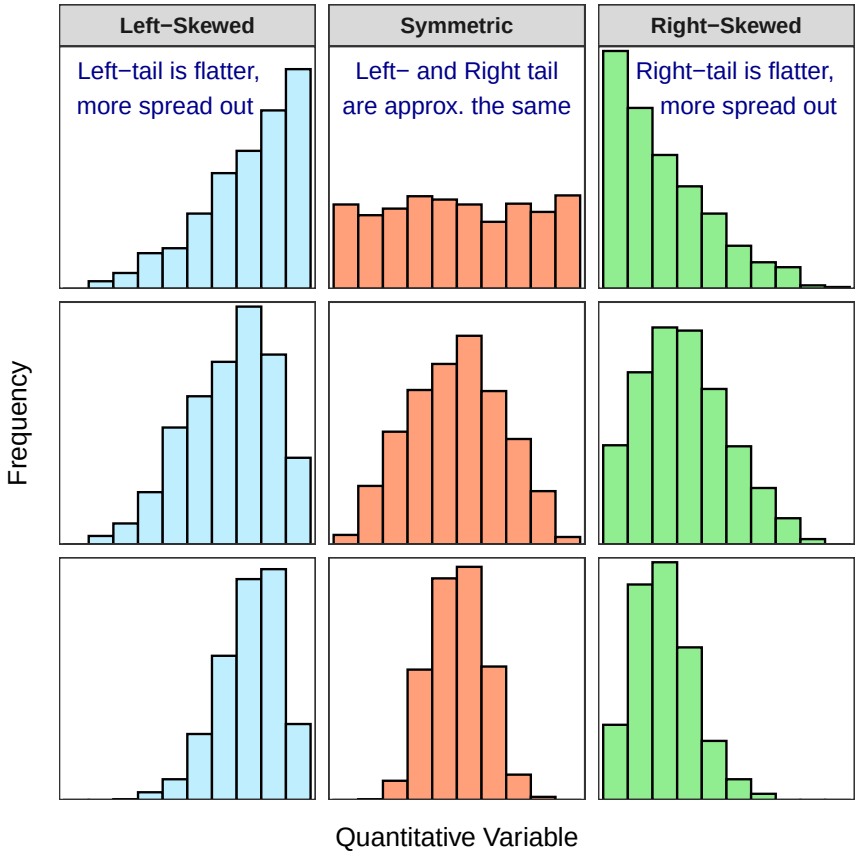


Figure 5.1. Examples of left-skewed, approximately symmetric, and right-skewed histograms. The skewed distributions are more skewed in the top row and less skewed in the bottom row.

◇ **The longer tail defines the type of skew; a longer right-tail means the distribution is right-skewed and a longer left-tail means it is left-skewed.**

In practice, these labels form a continuum (Figure **??**). For example, it may be difficult to discern whether the shape is approximately symmetric or skew. 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").

◇ **Shape terms may be modified with "approximately", "slightly", or "strongly."**

A distribution is **bimodal** if there are two distinct peaks (Figure **??**). The shape may be "bimodal left-skewed" if the left peak is shorter, "bimodal symmetric" if the two peaks are the same height, or "bimodal right-skewed" if the right peak is shorter.

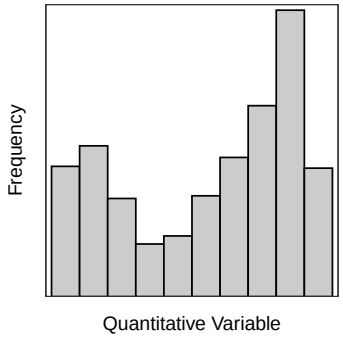


Figure 5.2. Example of a bimodal left-skewed histograms.

Shape may be identified from a histogram or a boxplot (Figure **??**). Shape is most easily determined from a histogram, as you can focus simply on the "longest" tail. With boxplots, one must examine the relative length from the median to Q1 and the median to Q3 (i.e., the position of the median line in the box). If the distribution is left-skewed (i.e., lesser-valued individuals are "spread out"), then the median-Q1 will be greater than Q3-median. In contrast, if the distribution is right-skewed (i.e., larger-valued individuals are spread out), then the 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 **??**).

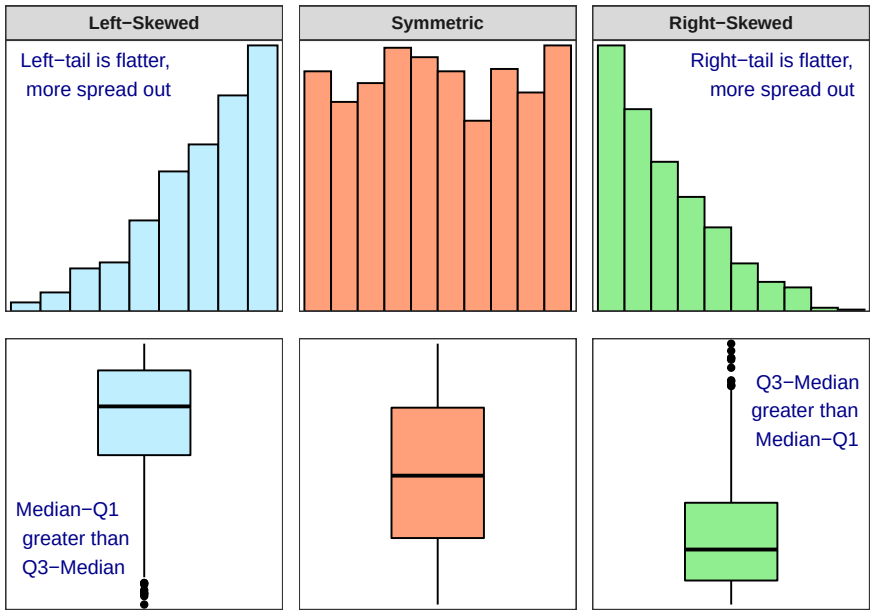


Figure 5.3. Histograms and boxplots for several different shapes of distributions.

◇ **Shape is easier to describe from a histogram than a boxplot.**

5.1.2 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 **??**). In general, outliers must be **on the margins of the histogram, should be separated by one or two missing bars, and should only be one or two individuals.**

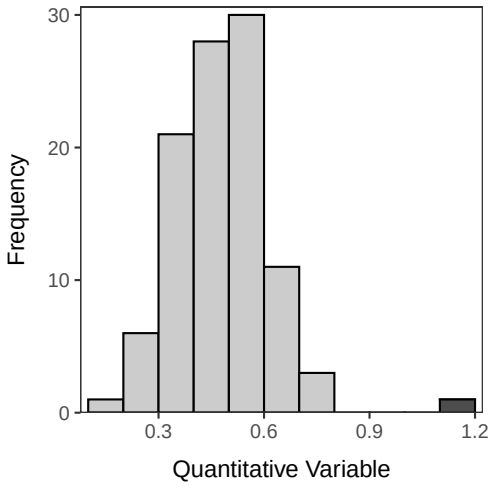


Figure 5.4. Example histogram with an outlier to the right (dark gray).

An outlier may be a result of human error in the sampling process and, thus, it 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 – and, thus, it 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 to 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 cluster of values is approximately symmetric and there is one outlier to the right of the main cluster (as illustrated in Figure **??**), **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.**

5.1.3 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 **??** 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 differently. The median balances the number of individuals smaller and larger than it. The mean, on the other hand, balances the sum of the distances to all points smaller than it and the sum of the distances 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 concerned with 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 data) are not considered when calculating the median; whereas the actual values are considered when calculating the mean.**

A plot of the Richter scale data against the corresponding ordered individual numbers is shown in Figure **??**-Top. The median (blue line) is the Richter scale value that corresponds to the middle individual number (i.e., move right from the individual number until the point is intercepted and then move down to the x-axis). Thus, the median (blue line) 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 (vertical red line) is placed such that the total **length** of the horizontal dashed red lines (distances from mean to point) is the same to the left as to the right.

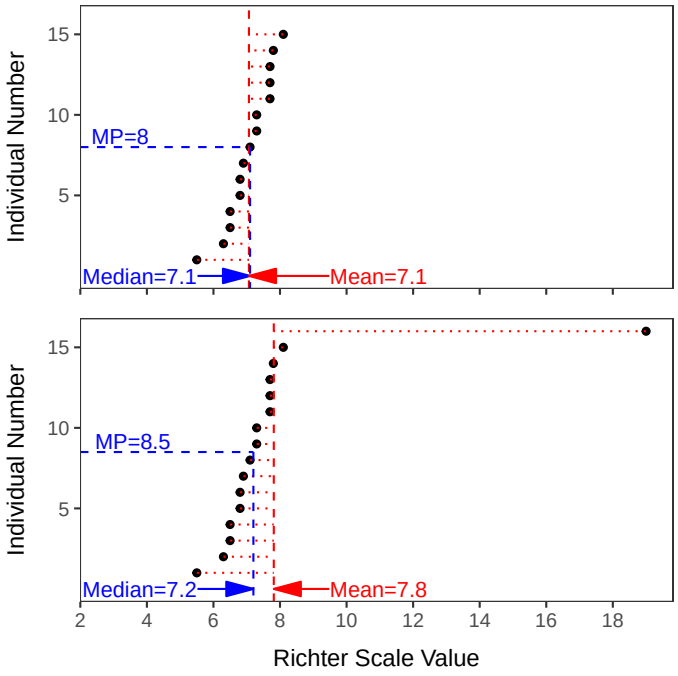


Figure 5.5. Plot of the individual number versus Richter scale values for the original earthquake data (**Top**) and the earthquake data with an extreme outlier (**Bottom**). The median value is shown as a blue vertical line and the mean value is shown as a red vertical line. Differences between each individual value and the mean value are shown with horizontal red lines.

◇ **The mean balances the distance to individuals above and below the mean. The median balances the number of individuals above and below the median.**

◇ **The sum of all differences between individual values and the mean equals zero.**

The mean and median differ in their sensitivity to outliers (Figure **??**). 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 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 **??**). 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.

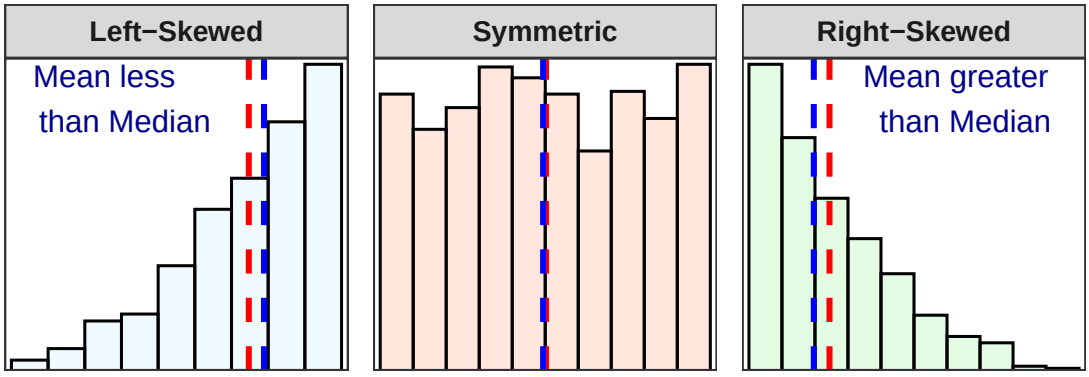


Figure 5.6. Three histograms with vertical dashed lines marking the median (blue) and the mean (red).

◇ **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 differently. 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.

As noted 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.

◇ **In this course, center and dispersion will be measured by the median and IQR if outliers are present or the distribution is more than slightly skewed, and the mean and standar deviation will be used if no outliers are present and the distribution is symmetric or only slightly skewed.**

5.1.4 Synthetic Interpretations

The graphical and numerical summaries from Module **??** 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 that 1) shape and outliers are described from the histogram, 2) center and dispersion are described ONLY from the mean and standard deviation OR the median and IQR are discussed, 3) the specific position of outliers (if present) is explained, 4) an explanation is given for why either the median and IQR or the mean and standard deviation were used, and 5) the range was not used alone as a measure of dispersion.

Number of Open Pit Mines

Construct a proper EDA for the number of open pit mines in countries that have open pit mines as summarized in Table **??** *and Figure* **??**.

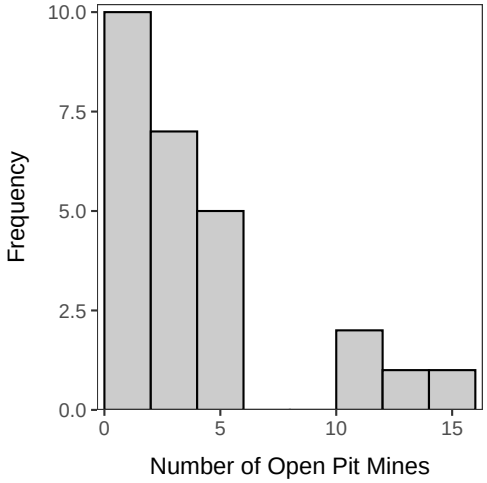


Figure 5.7. Histogram of number of open pit mines in countries with open pit mines.

The number of open pit mines in countries with open pit mines is strongly right-skewed with no outliers present (Figure **??**). [*I did not call the group of four countries with 10 or more open pit mines outliers*

Table 5.1. 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

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 **??**). 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) from a Q1 of 1.0 to a Q3 of 4.0 (Table **??**). I chose to use the median and IQR because the distribution was strongly skewed.

Lake Superior Ice Cover

Thoroughly describe the distribution of number of days of ice cover at ice gauge station 9004 in Lake Superior from Figure **??** *and Table* **??**.

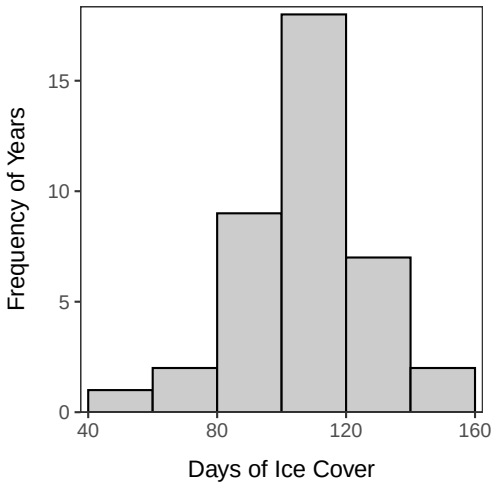


Figure 5.8. Histogram of number of days of ice cover at ice gauge 9004 in Lake Superior.

Table 5.2. 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

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

Crayfish Temperature Selection

Peck (1985) examined the temperature selection of dominant and subdominant crayfish (Orconectes virilis) together in an artificial stream. The temperature (°C) selection by the dominant crayfish in the presence of subdominant crayfish in these experiments was recorded below. Thoroughly describe all aspects of the distribution of selected temperatures from Figure **??** *and Table* **??**.

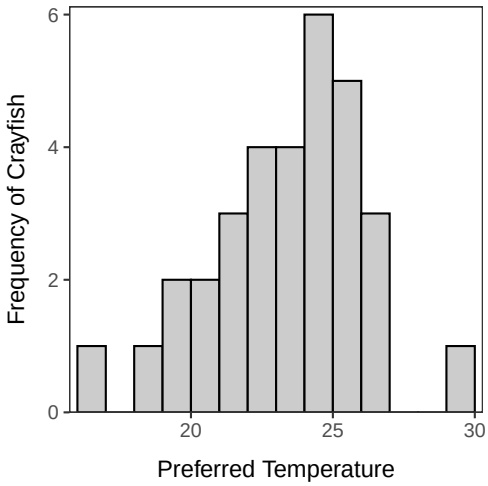


Figure 5.9. Histogram of crayfish temperature preferences.

Table 5.3. 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

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

5.2 Categorical Variable

An appropriate EDA for a categorical variable 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; the numerical values in the graph or table are not simply repeated.

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

5.2.1 Example Interpretations

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 displayed the results in Figure* **??**. *Use these results 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 (Figure **??**). Very few legumes or wildflowers were found in the sample.

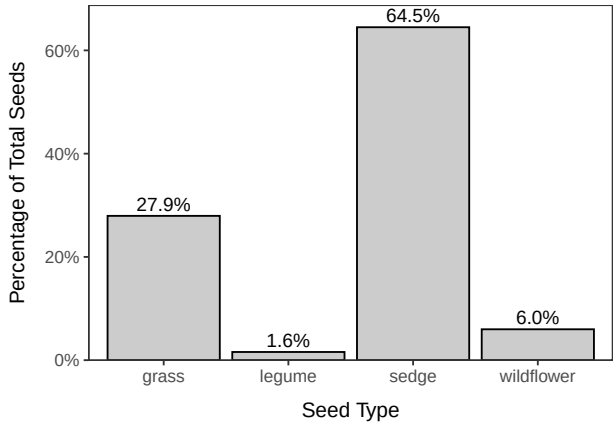


Figure 5.10. Bar chart of the percentage of wetland seeds by type.

GSS Recycling

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?" The results are displayed in Figure **??** *and Table* **??***. Use these results to describe the distribution of answers to the question.*

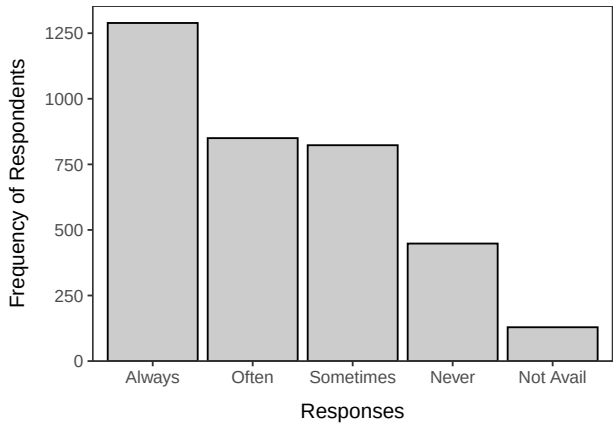


Figure 5.11. Barplot of the percentage of wetland seeds by type.

Table 5.4. Frequency of respondents by response to the question about recycling.

Always	Often	Sometimes	Never	Not Avail
1289	850	823	448	129

More than twice as many respondents always recycled compared to never recycled, with approximately equal numbers in between that often or sometimes recycled.

MODULE 6

NORMAL DISTRIBUTION

A MODEL FOR THE DISTRIBUTION of a single quantitative variable can be visualized by "fitting" a smooth curve to a histogram (Figure **??**-Left), removing the histogram (Figure **??**-Center), and using the remaining curve (Figure **??**-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.

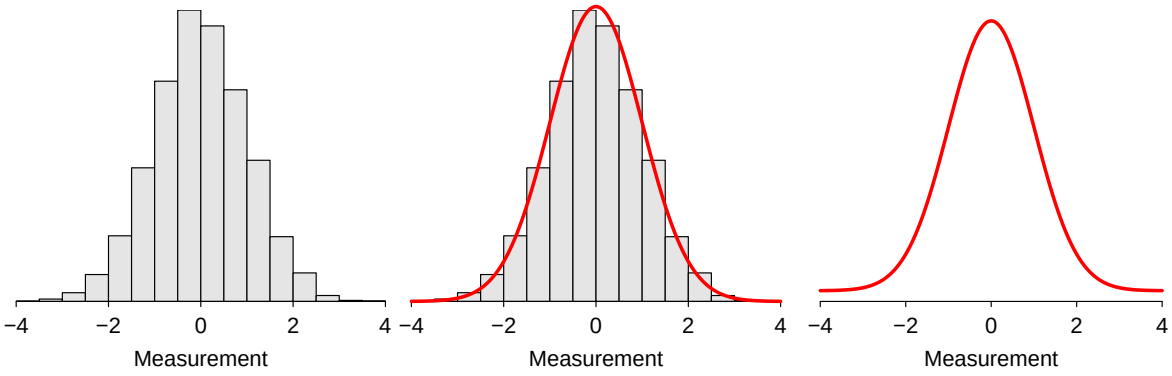


Figure 6.1. Depiction of fitting a smooth curve to a histogram to serve as a model for the distribution.

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.

[1] See Module **??**.

6.1 Characteristics of a Normal Distribution

The normal distribution is the familiar bell-shaped curve (Figure **??**-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 **??**).

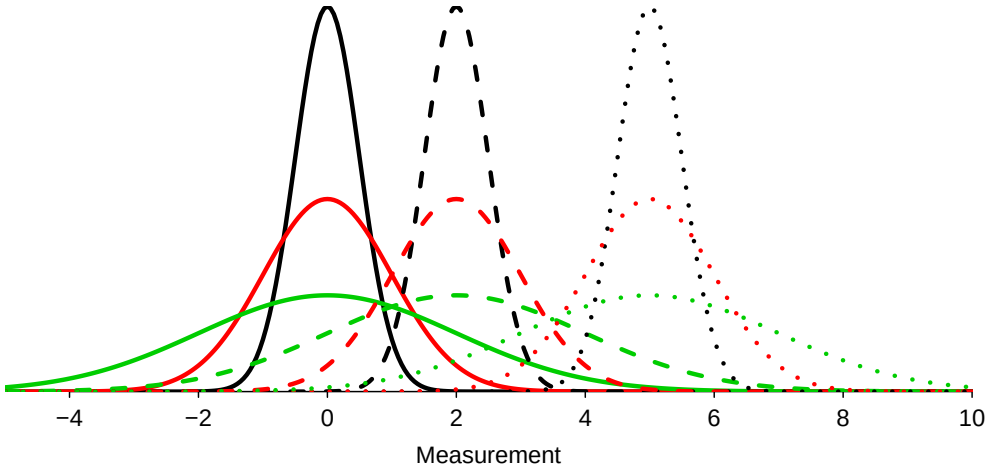


Figure 6.2. Nine normal distributions. Distributions with the same line type have the same value of μ (solid is μ=0, dashed is μ=2, dotted is μ=5). Distributions with the same color have the same value of σ (black is σ=0.5, red is σ=1, and green is σ=2).

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.

6.2 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 **??**.

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 **??**-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 **??**-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.

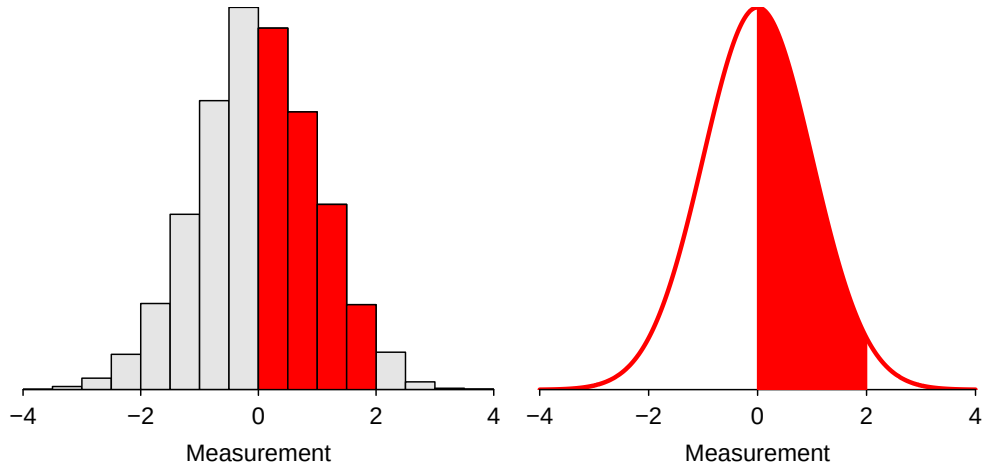


Figure 6.3. Depiction of finding the proportion of individuals between 0 and 2 on a histogram (**Left**) and on a standard normal distribution (**Right**).

◇ **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 **??**).

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.

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 **??**) 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.

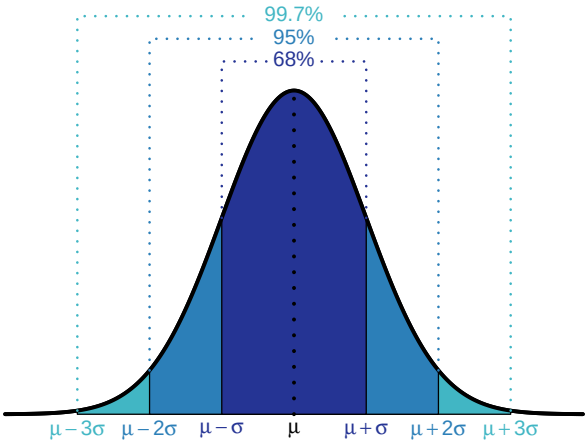


Figure 6.4. Depiction of the 68-95-99.7 (or Empirical) Rule on a normal distribution.

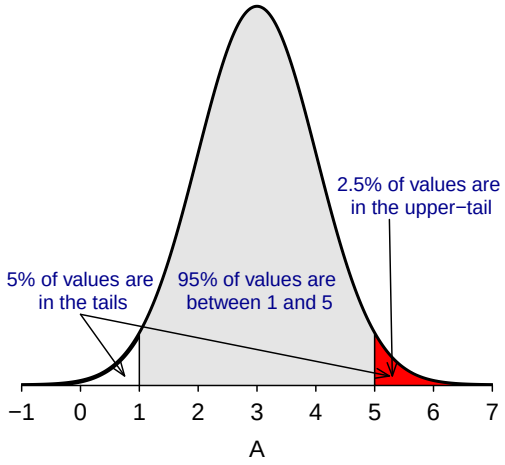


Figure 6.5. The N(3,1) distribution depicting how the 68-95-99.7 Rule is used to compute the percentage of individuals with values greater than 5.

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

6.3 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 **??**).

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

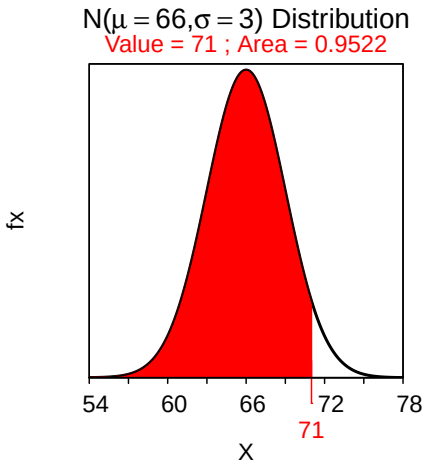


Figure 6.6. Calculation of the proportion of individuals on a $N(66, 3)$ with a value less than 71.

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 **??**).

```
> ( 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 **??**. 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).

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
```

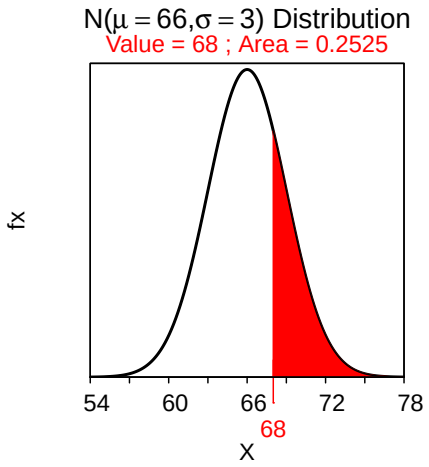


Figure 6.7. Calculation of the proportion of individuals on a $N(66, 3)$ with a value greater than 68.

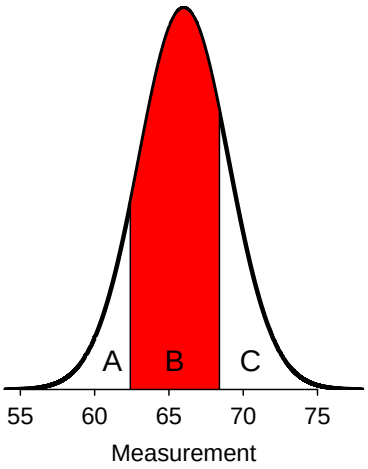


Figure 6.8. Schematic representation of how to find the area between two Z values.

```
[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.**

6.4 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 **??**).

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

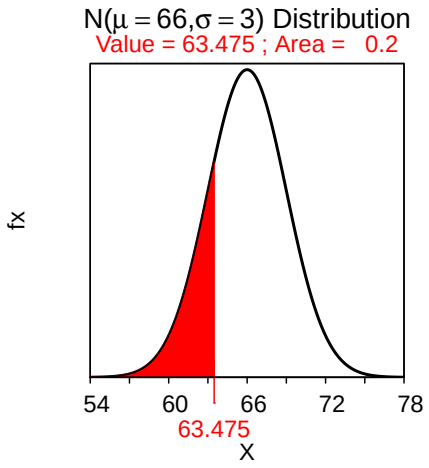


Figure 6.9. Calculation of the height with 20% of all students shorter.

"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 **??**).

```
> ( 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 **??** 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 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

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

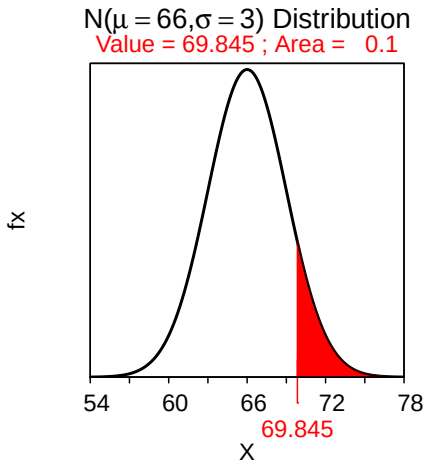


Figure 6.10. Calculation of the height with 10% of all students taller.

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
```

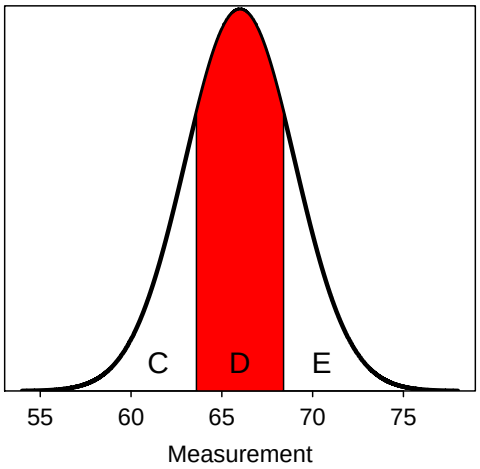


Figure 6.11. Depiction of areas in a reverse between type normal distribution question.

6.5 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.

6.6 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

$$Z = \frac{\text{"value"} - \text{"center"}}{\text{"dispersion"}} \tag{6.6.1}$$

or, more specifically,

$$Z = \frac{x - \mu}{\sigma} \tag{6.6.2}$$

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 **??** 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.

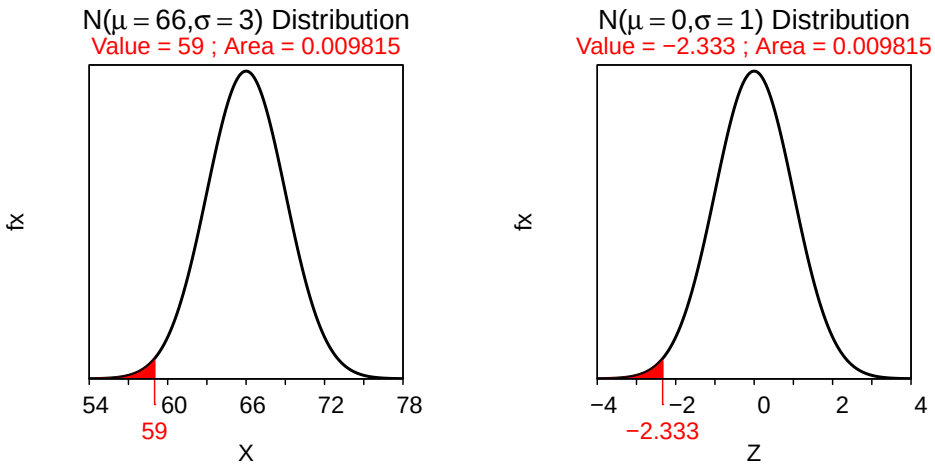


Figure 6.12. Plots depicting the area to the left of 59 on a $N(66, 3)$ (**Left**) and the area to the right of the corresponding Z-score of $Z = -2.33$ on a $N(0, 1)$ (**Right**). Not that the x-axis scales are different.

MODULE 7

BIVARIATE EDA - QUANTITATIVE

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 **??**.

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
```

7.1 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

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

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.**

7.2 Summaries

7.2.1 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 **??**).

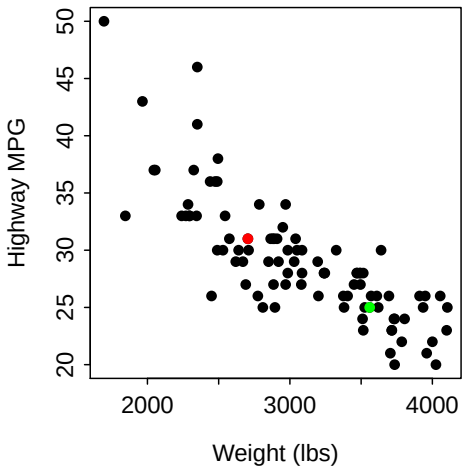


Figure 7.1. Scatterplot between the highway MPG and weight of cars manufactured in 1993. For reference to the main text, the first individual is red and the second individual is green.

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 **??**). The scatterplot, excluding the two highlighted points, of highway MPG versus car weight (Figure **??**) was created with the code below.

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

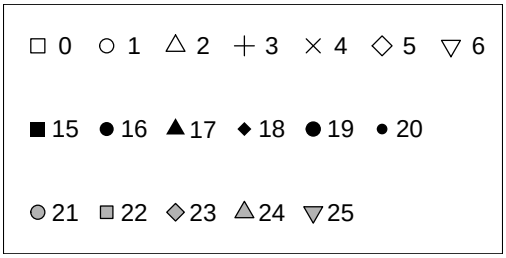


Figure 7.2. Plotting characters available in R and their numerical codes. Note that for values of 21-25 that `bg='gray70'` is used to provide the background color.

7.2.2 Correlation Coefficient

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

$$r = \frac{\sum_{i=1}^{n} \left[\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \right]}{n-1} \tag{7.2.1}$$

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 (**??**) 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 **??** for a review of standard deviations.
[4] See Section **??** 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 **??**.

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
```

7.2.3 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 **??**) 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 **??** 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.

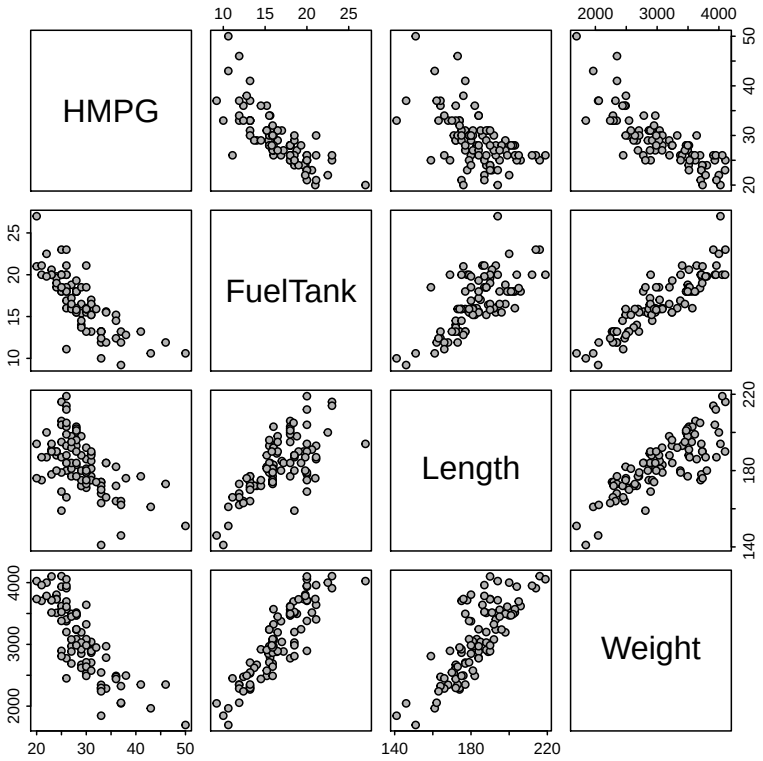


Figure 7.3. Scatterplot matrix of the highway MPG, fuel tank size, length, and weight of cars.

7.3 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.

7.3.1 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 **??**). 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.

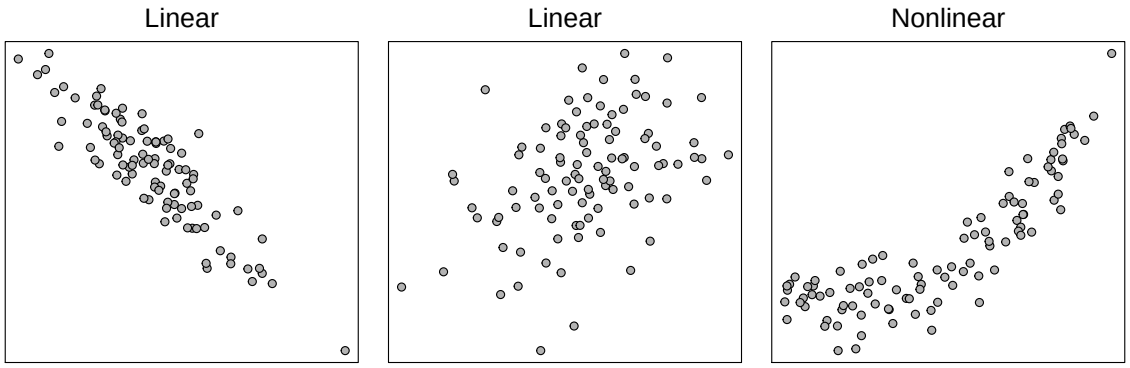


Figure 7.4. Depictions of two linear (Left and Center) and one nonlinear (Right) relationship.

7.3.2 Association or Direction

A positive association is when the scatterplot resembles an increasing function (i.e., increases from lower-left to upper-right; Figure **??**-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 **??**-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 **??**-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.

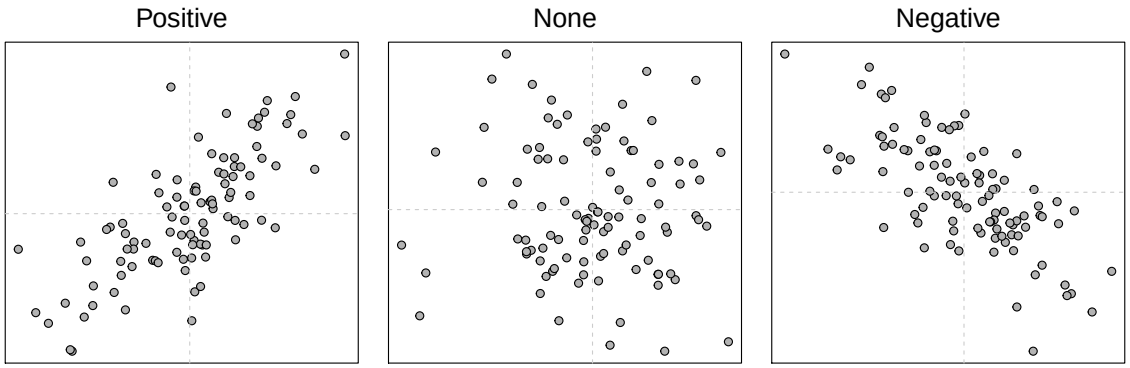


Figure 7.5. Depiction of three types of association present in scatterplots. Dashed vertical lines are at the means of each variable.

7.3.3 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 **??**) 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 **??**. The scatterplots in Figure **??** 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.

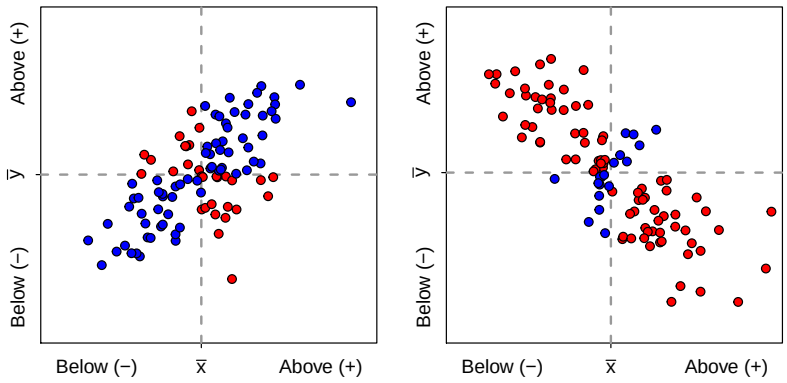


Figure 7.6. Scatterplot with mean lines superimposed and the signs of standardized values for both x and y shown for a positive (**Left**) and negative (**Right**) association. Blue points have a positive product of standardized values, whereas red points have a negative product of standardized values.

Now consider the product of standardized x's and y's in each quadrant of the scatterplots in Figure **??**. 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 **??**-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 **??**-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 **??**. Categorizations in Table **??** can be used as a guideline for describing the strength of relationship between two variables.

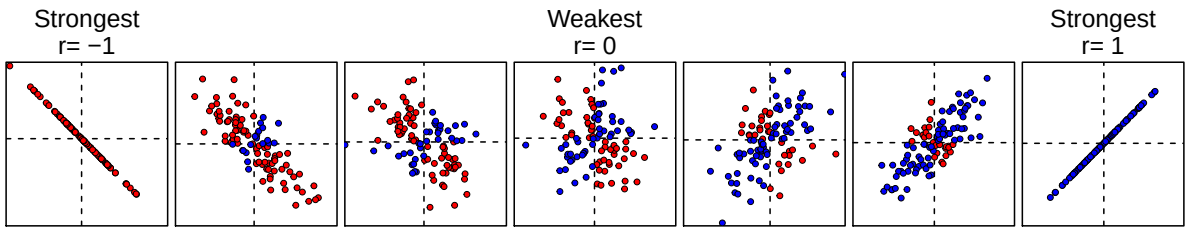


Figure 7.7. Scatterplots along the continuum of r values.

Table 7.1. 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

7.4 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 **??**) **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 **??**) 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 **??**-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 **??**-Right).

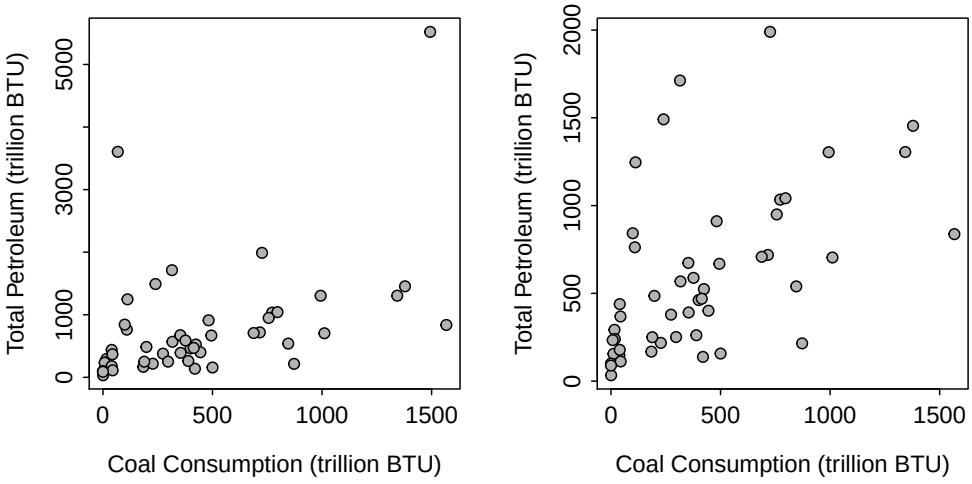


Figure 7.8. Scatterplot of the total consumption of petroleum versus the consumption of coal (in trillions of BTU) by all 50 states and the District of Columbia. The points shown in the left with total petroleum values greater than 3000 trillion BTU are deleted in the right plot.

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 **??**). I did compute r because no outliers were present and the relationship was linear (or, at least, it was not nonlinear).

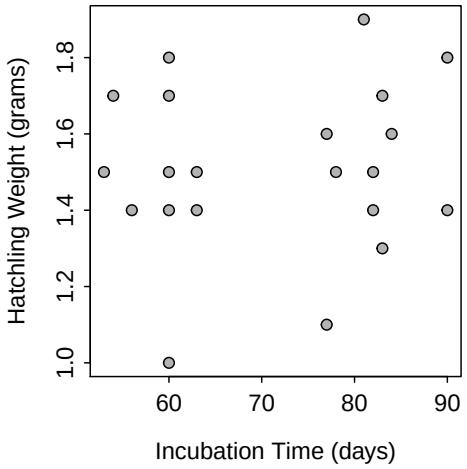


Figure 7.9. Scatterplot of hatchling weight versus incubation time for Crested Geckos.

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)
```

7.5 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 **??**, 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?

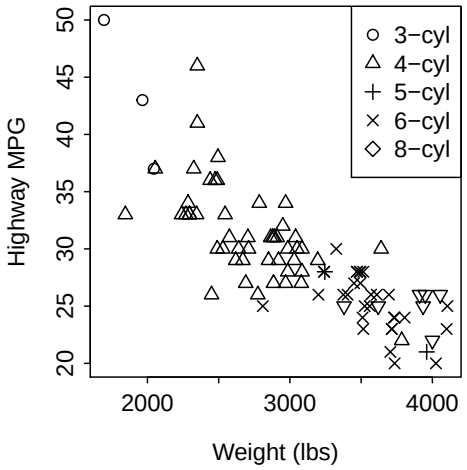


Figure 7.10. Scatterplot between the highway MPG and weight of cars manufactured in 1993 separated by number of cylinders.

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?".