MODULE 1

UNIVARIATE SUMMARIES)

Contents

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.

1.1 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 1.1. 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 1.2. Richter scale recordings for 15 major earthquakes.

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

1.1.1 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

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

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 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,

[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.

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]
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.

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

or more succinctly summarized with this equation,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.1.1}$$

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

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{1.1.2}$$

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 $s = \sqrt{0.484} = 0.696$.

Table 1.3. Table showing an efficient calculation of the standard deviation of the earthquake data.

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

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

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

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

The standard deviation can be thought of as "the average difference between the values and the mean." This is, however, not a strict definition because the formula for the standard deviation does not simply add the differences and divide by n as this definition would imply. Notice in Table **??** 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 1.4. 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 1.5. 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.

1.1.2 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 **??**).

Table 1.6. Frequency table of number of countries in two-mine-wide classes.

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

The frequency of individuals in each class is then counted (shown in the second row of Table **??**). 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 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).

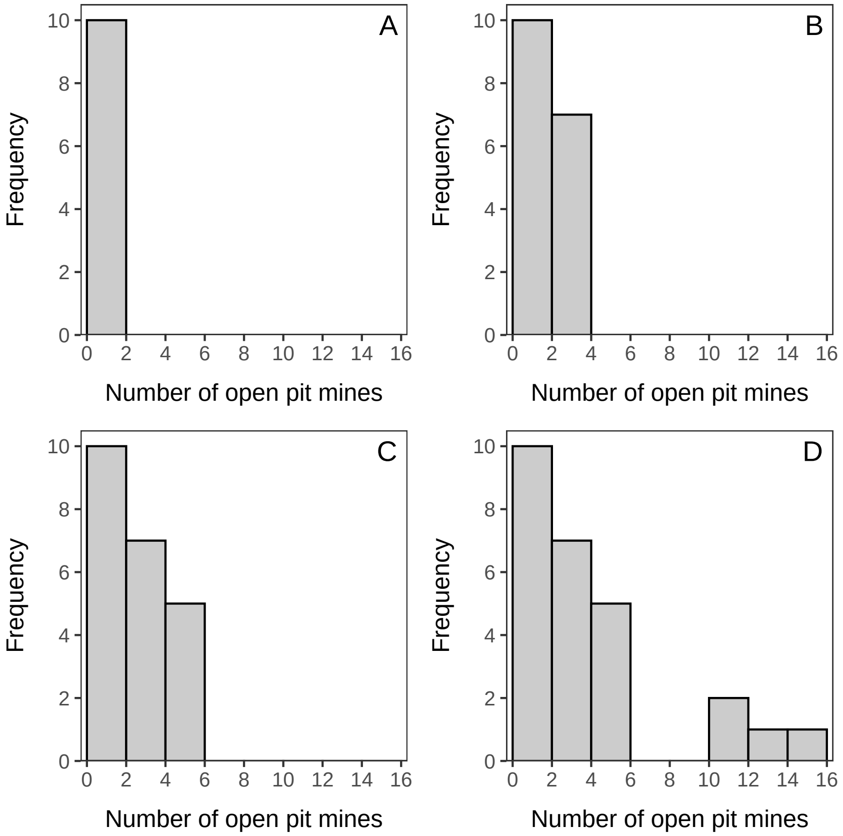

Figure 1.1. Steps (described in text) illustrating the construction of a histogram.

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 **??**).

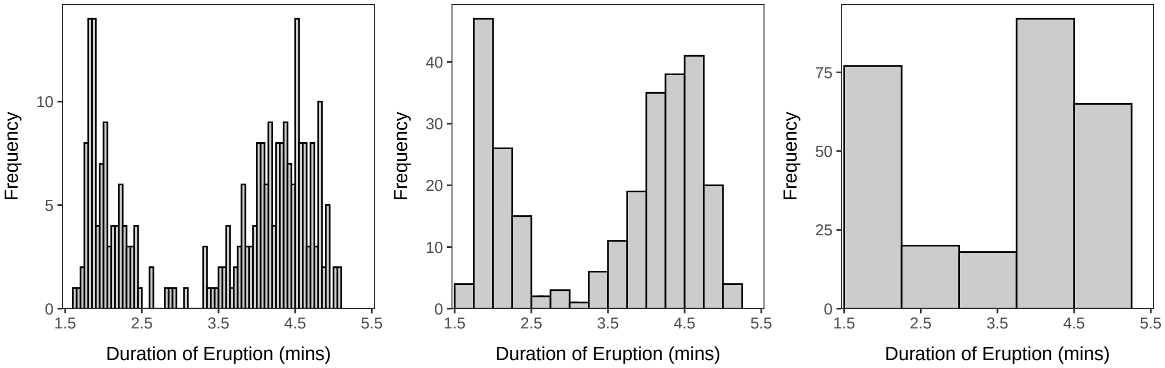

Figure 1.2. Histogram of length of eruptions for Old Faithful geyser with varying number of bins/classes.

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 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.

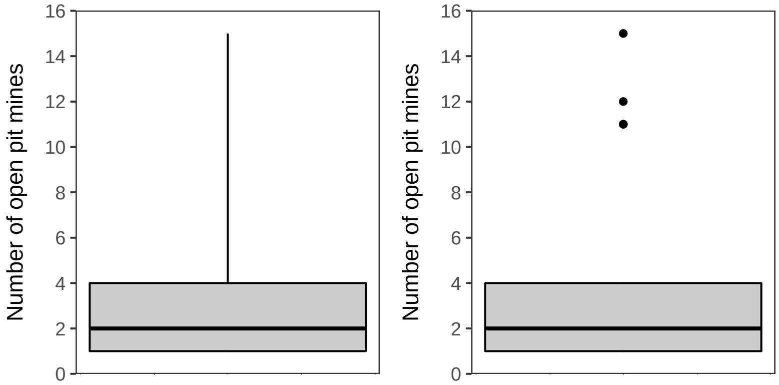

Figure 1.3. Traditional (Left) and modern (Right) boxplots of the open pit mine data.

1.2 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 1.7. 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

1.2.1 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 1.8. 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 1.9. 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 **??**).

1.2.2 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,

Table 1.10. 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

category labels rather than quantitative values are plotted on the x-axis. In addition, to highlight the categorical nature of the data, bars on a bar 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.

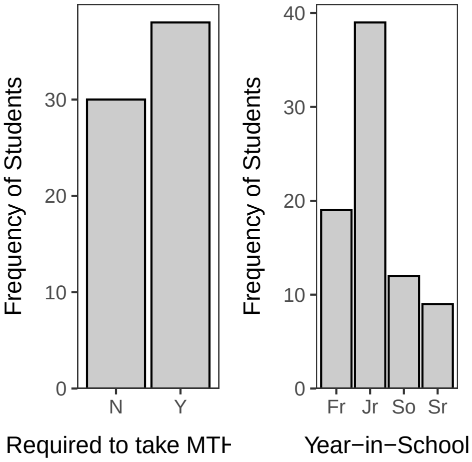

Figure 1.4. Bar charts of the frequency of individuals in MTH107 during Winter 2010 by whether or not they were required to take MTH107 (**Left**) and year-in-school (**Right**).

◇ 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 2

UNIVARIATE EDA

Contents

2.1 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 by four specific attributes:

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

Graphs are used to identify shape and the presence of outliers and to get a general feel for center and dispersion. Numerical summaries, however, are used to specifically describe center and dispersion of the variable. Computing and constructing the required numerical and graphical summaries was described in Module **??**. 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.

2.1.1 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 these two tails is used to identify three different shapes of distributions – symmetric, left-skewed, and right-skewed. If the left- and right-tail of a distribution are approximately equal in shape (length and height), then the distribution is said to be **symmetric** (or more specifically **approximately symmetric**). If the left-tail is stretched out or is longer and flatter than the right-tail, then the distribution is negatively- or **left-skewed**. If the right-tail is stretched out or is longer and flatter than the left-tail, then the distribution is positively- or **right-skewed**. The type of skew is defined by the longer tail; a longer right-tail means the distribution is right-skewed and a longer left-tail means it is left-skewed.

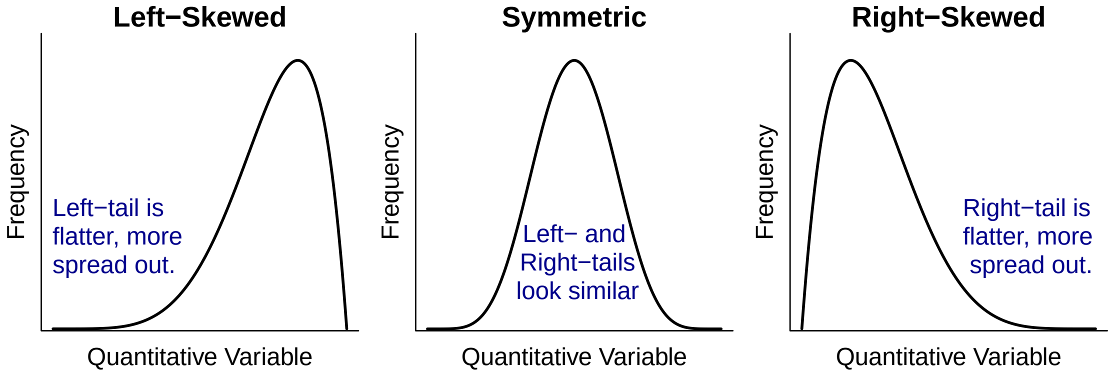

Figure 2.1. Examples of left-skewed (left), symmetric (center), and right-skewed (right) distributions.

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

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

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

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

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

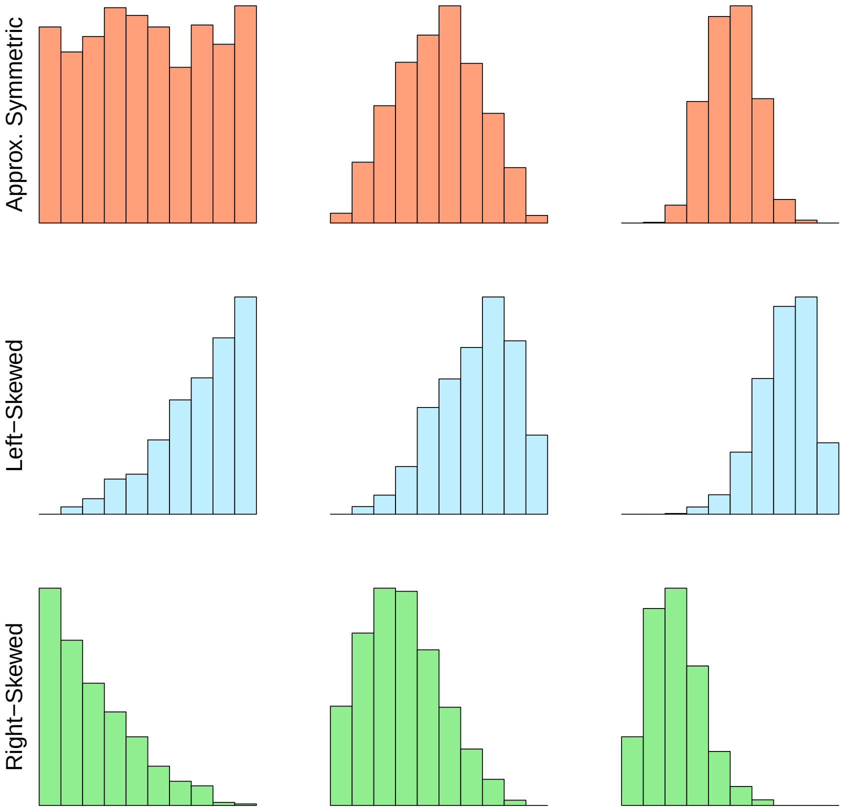

Figure 2.2. Examples of approximately symmetric (top, red), left-skewed (middle, blue), and right-skewed (bottom, green) histograms. Note that the axes labels were removed to focus on the shape of the histograms.

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

2.1.2 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.**

An outlier may be a result of human error in the sampling process. If this is the case, then the value should be corrected or removed. Other times an outlier may be an individual that was not part of the population of interest – e.g., an adult animal that was sampled when only immature animals were being considered. In this case, the individual should be removed from the sample. Still other times, an outlier is part of the population and should generally not be removed from the sample. In fact you may wish to highlight an outlier as an interesting observation! Regardless, it is important that you construct a histogram to determine if outliers are present or not.

Don't let outliers completely influence how you define the shape of a distribution. For example, if the main

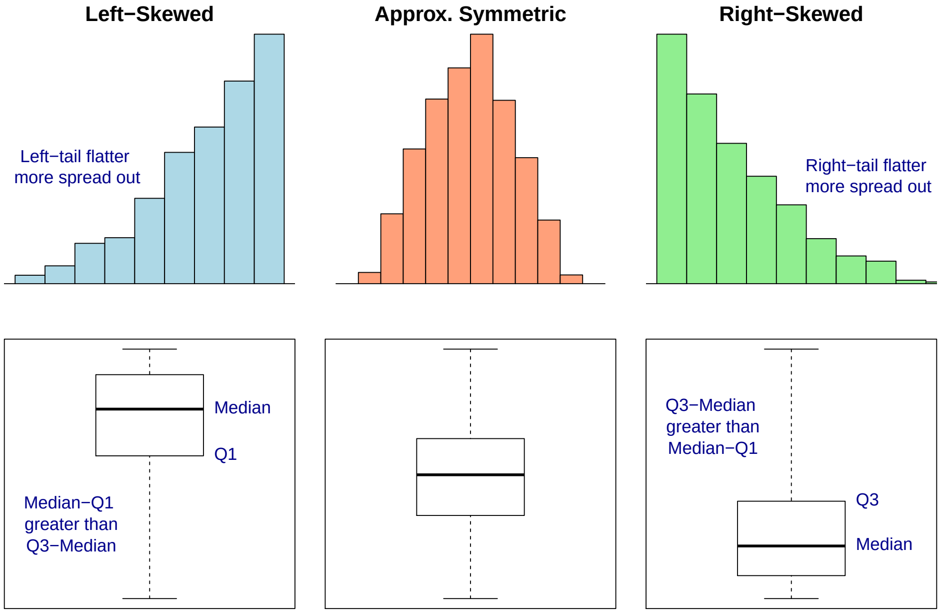

Figure 2.3. Histograms and boxplots for several different shapes of distributions.

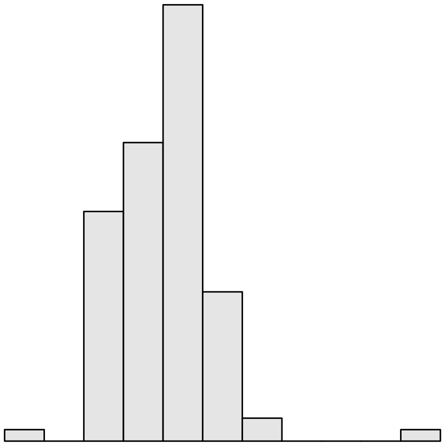

Figure 2.4. Example histogram with an outlier to the right.

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

2.1.3 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 in different ways. The median balances the number of individuals smaller and larger than it. The mean, on the other hand, balances the sum of the distances from it to all points smaller than it and the sum of the distances from it to all points greater than it. Thus, the median is primarily concerned with the **position** of the value rather than the value itself, whereas the mean is very much concerned about the **values** for each individual (i.e., the values are used to find the "distance" from the mean).

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

A plot of the Richter scale data against the corresponding ordered individual number is shown in Figure **??**-Left.[1] The median (blue line) is found by locating the middle position on the individual number axis and then finding the corresponding Richter scale value (move right until the point is intercepted and then move down to the x-axis). The vertical blue line represents the median; i.e., it has the same **number** of individuals (i.e., points) above and below it. In contrast, the mean finds the Richter scale value that has the same total distance to values below it as total distance to values above it. In other words, the mean is the vertical red line placed such that the total **length** of the horizontal dashed red lines is the same to the left as it is to the right. Thus, the median balances the number of individuals above and below the median, whereas the mean balances the total difference in values above and below the mean.

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

◇ **The sum of all differences between individual values and the mean (as properly calculated) equals zero.**

The mean and median differ in their sensitivity to outliers (Figure **??**-Right). For example, suppose that an incredible earthquake with a Richter Scale value of 19.0 was added to the earthquake data set. With this additional individual, the median increases from 7.1 to 7.2, but the mean increases from 7.1 to 7.8. The outlier impacts the value of the mean more than the value of the median because of the way that each statistic measures center. The mean will be pulled towards an outlier because it must "put" many values on the "side" of the mean away from the outlier so that the sum of the differences to the larger values and

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

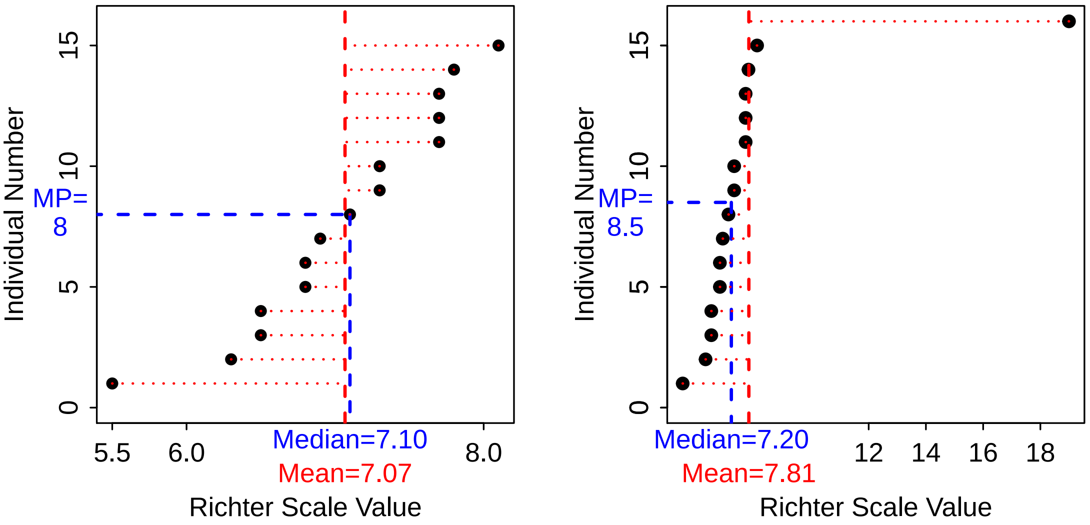

Figure 2.5. Plot of the individual number versus Richter scale values for the original earthquake data (**Left**) and the earthquake data with an extreme outlier (**Right**). The median value is shown as a blue vertical line and the mean value is shown as a red vertical line. Differences between each individual value and the mean value are shown with horizontal red lines.

the sum of the differences to the smaller values will be equal. In this example, the outlier creates a large difference to the right of the mean such that the mean has to "move" to the right to make this difference smaller, move more individuals to the left side of the mean, and increase the differences of individuals to the left of the mean to balance this one large individual. The median on the other hand will simply "put" one more individual on the side opposite of the outlier because it balances the number of individuals on each side of it. Thus, the median has to move very little to the right to accomplish this balance.

◇ **The mean is more sensitive (i.e., changes more) to outliers than the median; it will be "pulled" towards the outlier more than the median.**

The shape of the distribution, even if outliers are not present, also has an impact on the mean and median (Figure **??**). If a distribution is approximately symmetric, then the median and mean (along with the mode) will be nearly identical. If the distribution is left-skewed, then the mean will be less than the median. Finally, if the distribution is right-skewed, then the mean will be greater than the median.

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

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

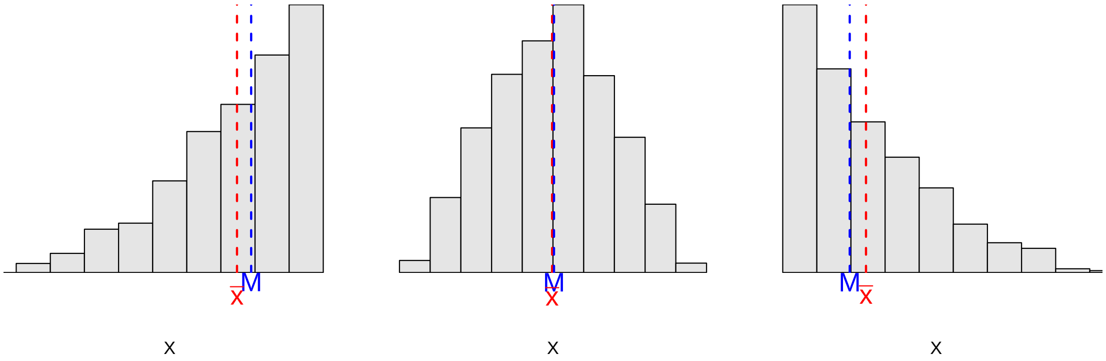

Figure 2.6. Three differently shaped histograms with vertical lines superimposed at the median (M; blue lines) and the mean ($\bar{x}$; red lines).

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

2.1.4 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 the 1) reference to figures and tables, 2) labeling of the figures and tables, 3) that only the mean and standard deviation or the median and IQR are discussed, 4) the range was not used alone as a measure of dispersion, 5) the explanation for why either the median and IQR or the mean and standard deviation were used, and 6) an appendix of R code used was provided.

Number of Open Pit Mines

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

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 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) is 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.

Table 2.1. 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

R Code Appendix:

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

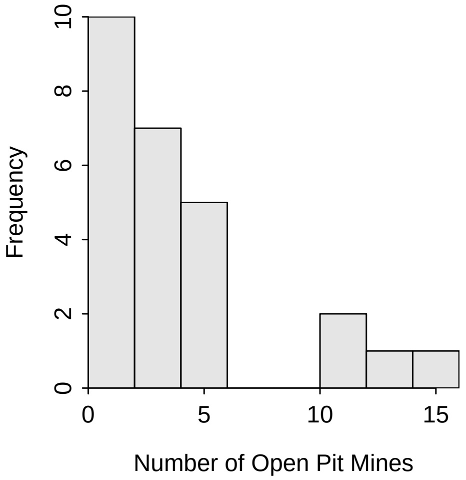

Figure 2.7. Histogram of number of open pit mines in countries with open pit mines.

Lake Superior Ice Cover

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

The shape of number of days of ice cover at gauge 9004 in Lake Superior is approximately symmetric with no obvious outliers (Figure **??**). 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.

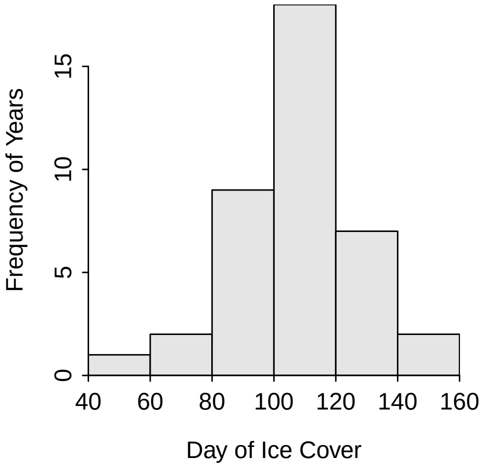

Figure 2.8. Histogram of number of days of ice cover at ice gauge 9004 in Lake Superior.

Table 2.2. Descriptive statistics of number of days of ice cover at ice gauge 9004 in Lake Superior..

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

R Appendix:

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

Crayfish Temperature Selection

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

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

The shape of temperatures selected by the dominant crayfish is slightly left-skewed (Figure **??**) 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.

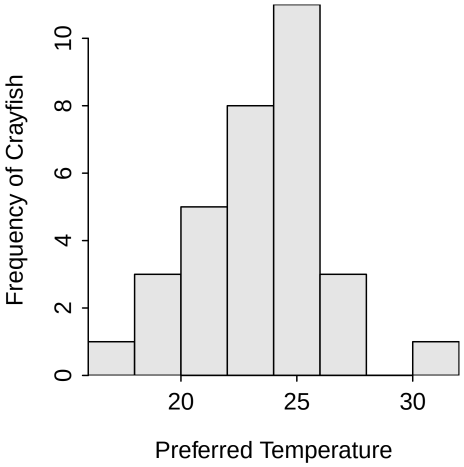

Figure 2.9. Histogram of crayfish temperature preferences.

Table 2.3. Descriptive statistics of crayfish temperature preferences.

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

R Appendix:

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

2.2 Categorical Variable

Interpreting summaries of a single categorical variable is more intuitive and less defined than that for quantitative data. Specifically, one DOES NOT describe shape, center, dispersion, and outliers for categorical data. In this module, methods to construct tables and graphs for categorical data are described and the interpretation of the results demonstrated.

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

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

Table 2.4. Whether (Y) or not (N) MTH107 was required for eight individuals in MTH107 in Winter 2010.

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

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

2.2.1 Example Interpretations

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

Mixture Seed Count

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

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

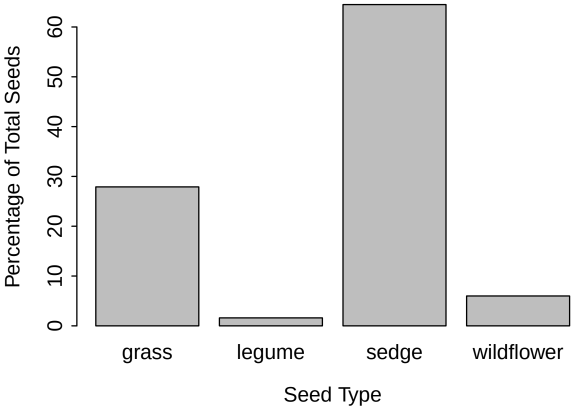

Figure 2.10. Barplot of the percentage of wetland seeds by type.

Table 2.5. Percentage distribution of wetland seeds by type.

grass	legume	sedge	wildflower
27.9	1.6	64.5	6.0

R Appendix:

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