MODULE 5

UNIVARIATE EDA - QUANTITATIVE

Objectives:

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

Contents

Once data have been collected (Module 3), it is important to develop a “feel” for the data, to identify what types of values each variable takes, and to determine if there are any “issues” in the data. This first step in a statistical analysis is called Exploratory Data Analysis (EDA). We will begin by examining the distribution of each variable by itself, called a univariate EDA, and then examine pairs of variables, called a bivariate EDA (see Modules 8 and 9). Additionally, the methods employed differ for quantitative and categorical variables. Quantitative variables are the focus of this module, whereas categorical variables are the focus of Module 6.

5.1 Items to Describe

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

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

Graphs are used to identify shape and the presence of outliers and to get a general feel for center and dispersion. However, numerical summaries are used to specifically describe center and dispersion of the data.

◇ Shape, center, dispersion, and outliers are described for each quantitative variable.

◇ Shape and outliers are described from graphs; center and dispersion are describe with numerical summaries.

Three primary data sets will be explored throughout this module.

- Measurements of water consumption in one hour by mice (Table 5.1).[1]
- Richter scale recordings for 15 major earthquakes (Table 5.2).
- The number of days of ice cover at ice gauge station 9004 in Lake Superior (data in LakeSuperiorIce.csv).[2] The *days* variable is the total number of days of ice cover at this site for nearly every ice season from 1955-56 to 1996-97 (three years were missing). These data are loaded into `LSI` below.

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

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

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

Table 5.2. Richter scale recordings for 15 major earthquakes.

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

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

[2] See Section 4.3.2 for a description of how to access these data. These datat are originally from the National Snow and Ice Data Center.

5.2 Histograms

5.2.1 General Construction

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

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

These steps are illustrated with the mouse water consumption data. The easiest way to create a list of classes is to divide the difference between the maximum and minimum values in the data 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. In this example, the range of values is 18.4-9.4 = 9.0. A "nice" value between eight and ten to divide this range by is nine. Thus, the classes of data should be one unit wide and, for ease, will begin at 9 mm (Table 5.3).

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

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

The number of individuals with a value of the variable in each class is called a frequency and are shown in the second column of Table 5.3. The plot is prepared with values of the classes forming the x-axis and frequencies forming the y-axis (Figure 5.1-Left). The first bar added to this skeleton plot has the bottom-left corner at 9 and the bottom-right corner at 10 on the x-axis, and a height equal to the frequency of individuals in the 9 to 9.9 class (Figure 5.1-Center). A second bar is then added with the bottom-left corner at 10 and the bottom-right corner at 11 on the x-axis, and a height equal to the frequency of individuals in the 10 to 10.9 class (Figure 5.1-Right). This process is continued with the remaining classes until the full histogram is constructed (Figure 5.2).

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

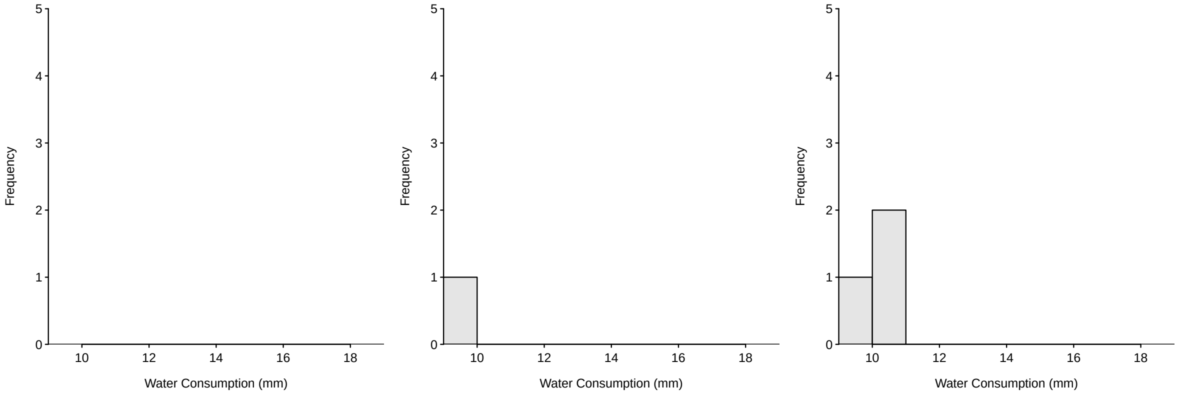


Figure 5.1. Steps illustrating the development of a histogram.

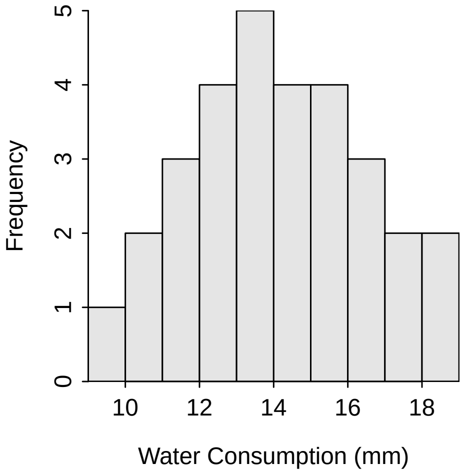


Figure 5.2. Histogram of water consumption (mm) by mice.

5.2.2 Histograms in R

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

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

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

[3] `xlab=` is for the "x-axis label."

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

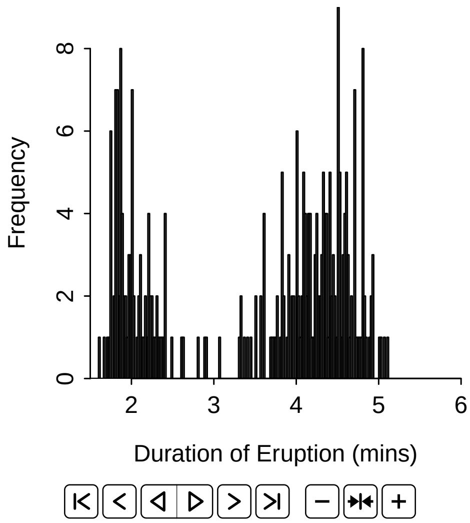


Figure 5.3. Histogram of length (minutes) of eruptions for Old Faithful geyser with varying number of classes.

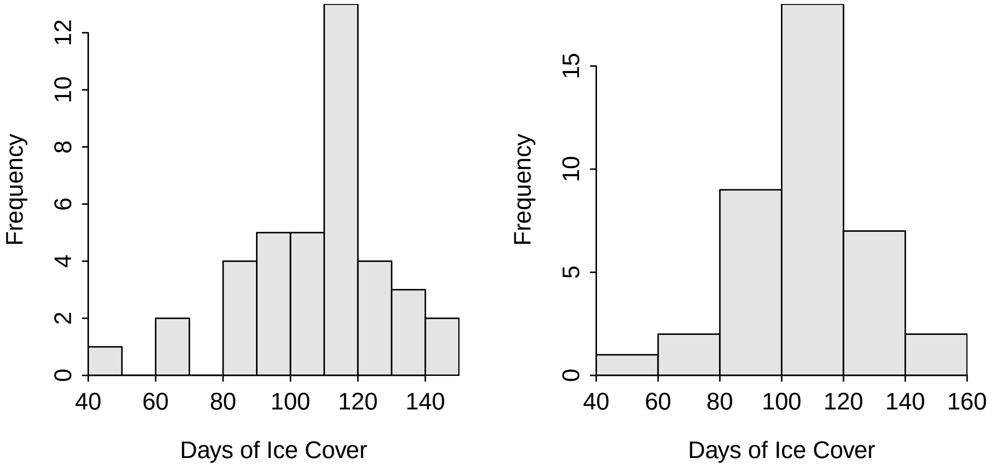


Figure 5.4. Histograms of the duration of ice cover at ice gauge 9004 in Lake Superior using the default class widths (Left) and widths of 20 days (Right).

Review Exercises

5.1 Histograms are constructed from what type of variables? *Answer*

5.2 What type of values are plotted on the x-axis of a histogram? *Answer*

5.3 What type of values are plotted on the y-axis of a histogram? *Answer*

5.4 What is the ideal number of bars on a histogram? *Answer*

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

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

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

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

5.3 Interpreting Shape

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

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

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

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

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

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

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

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

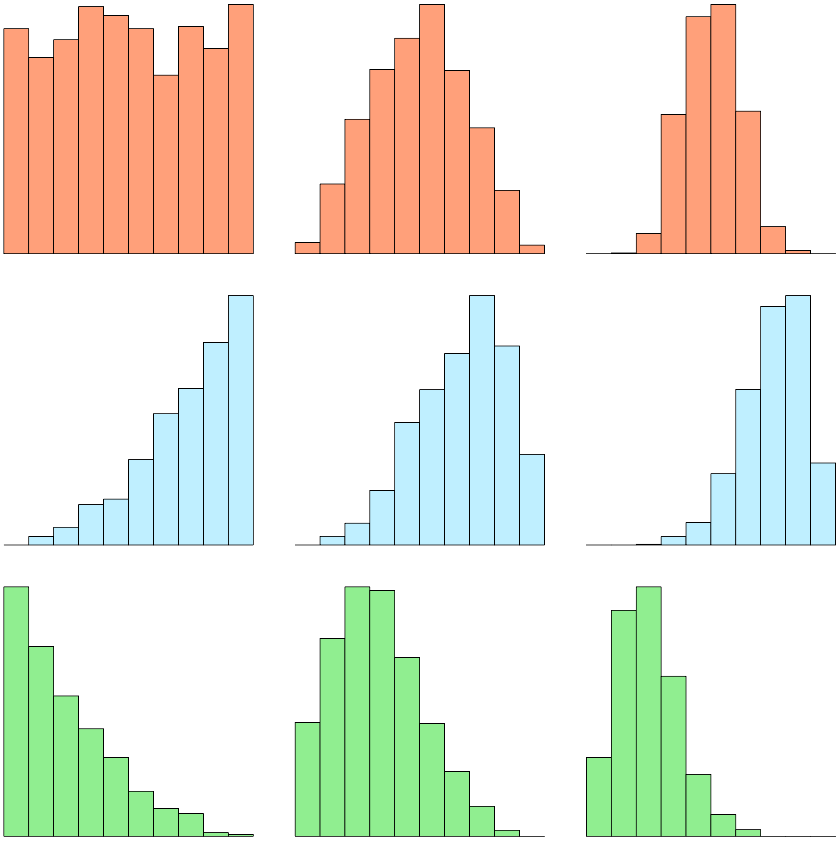


Figure 5.5. Examples of approximately symmetric (top, red), left-skewed (middle, blue), and right-skewed (bottom, green) histograms. Note that the axes labels were removed to focus attention on the shape of the histograms. Each histogram was constructed from n=1000 individuals and the x-axis range is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

5.4 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 5.6). In general, outliers must be on the margins of the histogram, should be separated by one or two missing bars, and should only be one or two individuals.

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

An outlier may occur as 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's value 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 cluster of values is approximately symmetric and there is one outlier to the right of the main cluster (as

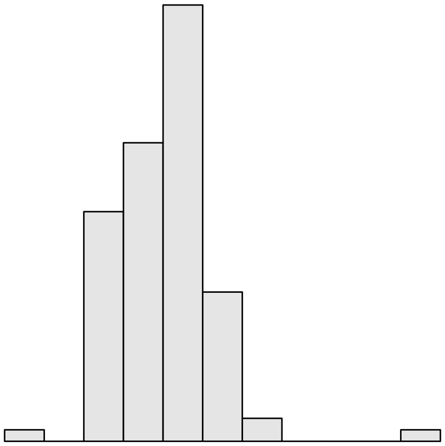


Figure 5.6. Example histogram with an outlier to the right.

illustrated in Figure 5.6), **DON'T** call the distribution right-skewed. You should describe this distribution as approximately symmetric with an outlier to the right.

◇ Not all outliers warrant removal from your sample.

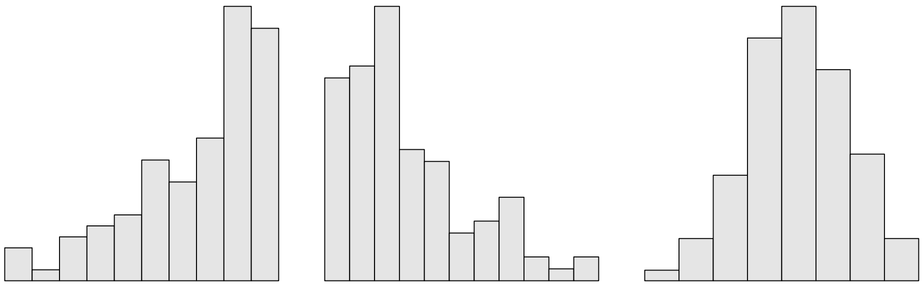
◇ Don't let outliers completely influence how you define the shape of a distribution.

Review Exercises

5.7 What is a distribution with a long left-tail called? *Answer*

5.8 What is a distribution with a long right-tail called? *Answer*

5.9 What is the shape of the distribution on the left below? *Answer*



5.10 What is the shape of the distribution in the center above? Answer

5.11 What is the shape of the distribution on the right above? Answer

5.12 Comment on the shape and presence of outliers in Figure 1.2. Answer

5.5 Measures of Center

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

The modes, medians, and means computed in this section are summary statistics – i.e., they are computations from individuals in a sample. Thus, they should specifically be called the sample mode, sample median, and sample mean. The mode, median, and mean can also be computed from every individual in the population, if it is known. The computed values would then be parameters and would be called the population mode, population median, and population mean. See Section 2.1 for clarification on the differences between populations and samples and parameters and statistics.

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

◇ **Measures of center computed from individuals in a sample are preceded by "sample"; those computed from all individuals in a population are preceded by "population."**

5.5.1 Mode

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

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

∆ **Bimodal**: The shape of a distribution with two peaks or "humps."

5.5.2 Median

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

The process for finding the median consists of three steps,[6]

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

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

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

Because $n = 30$, the $mp = \frac{30+1}{2} = 15.5$. The mp is not an integer so the median is the average of the values in the 15th and 16th ordered positions (i.e., the two positions closest to mp). Thus, the median water consumption in this sample of mice is $\frac{13.7+14.0}{2} = 13.85$ mm.

As another example, consider finding the median of the Richter Scale magnitude recorded for fifteen major earthquakes (ordered data in Table 5.2). Because $n = 15$, the $mp = \frac{15+1}{2} = 8$. The mp is an integer so the median is the value of the individual in the 8th ordered position, which is 7.1.

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

5.5.3 Mean

The mean is the arithmetic average of the data. The sample mean is denoted by $\bar{x}$ and the population mean by μ. 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,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5.5.1}$$

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

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

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

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

5.5.4 Measures of Center in R

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

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

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

Review Exercises

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

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

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

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

5.15 Compute the mean and median of the creatine phosphate data in Exercise 5.5. Answer

5.16 Compute the mean and median of the carbon monoxide data in Exercise 5.6. Answer

[7] Data collected from USGS.

[8] Data collected from U.S. census.

5.5.5 Comparing the Median and Mean

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

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

A plot of the Richter scale data against the corresponding ordered individual number is shown in Figure 5.7-Left.[9] The median (blue line) is found by locating the middle position on the individual number axis and then finding the corresponding Richter scale value (move right until the point is intercepted and then move down to the x-axis). The vertical blue line represents the median, and it can be seen that it has the same **number** of individuals (i.e., points) below it as above it. In contrast, the mean finds the Richter scale value that has the same total distance to values below it as total distance to values above it. In other words, the mean is the vertical red line so that the total **length** of the horizontal dashed red lines is the same to the left as it is to the right. Thus, the median balances the number of individuals above and below the median, whereas the mean balances the difference in values above and below the mean.

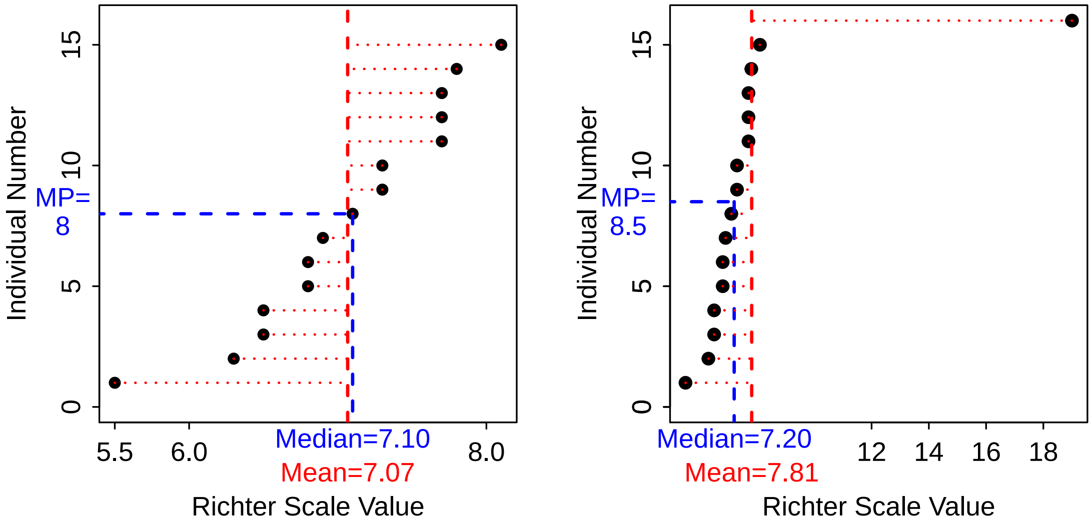


Figure 5.7. Plot of the individual number versus Richter scale values for the original earthquake data (**Left**) and the earthquake data with an extreme outlier (**Right**). The median value is shown as a blue vertical line and the mean value is shown as a red vertical line. Differences between each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

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

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

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

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

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

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

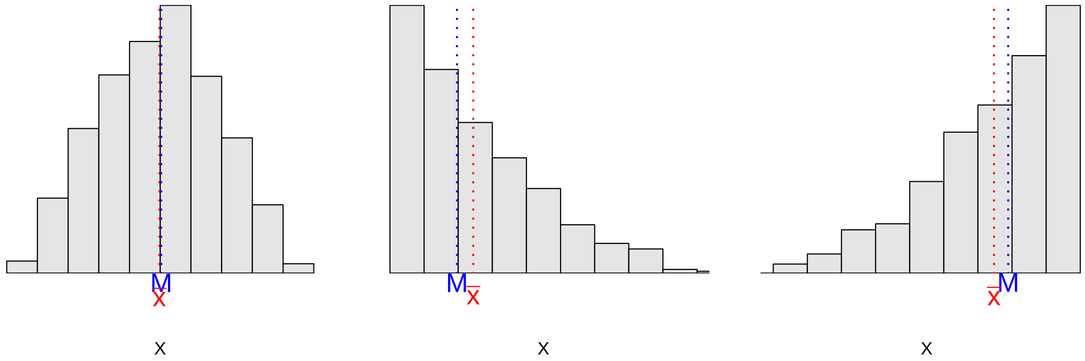


Figure 5.8. Three differently shaped histograms with vertical lines superimposed at the median (M; blue lines) and the mean ($\bar{x}$; red lines).

◇ **The mean and median are equal for symmetric distributions.**

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

As shown above, the mean and median measure center in different ways. The question now becomes "which measure of center is better?" The median is a "better" measure of center when outliers are present. In addition, the median gives a better measure of a typical individual when the data are skewed. Thus, in this course, 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.

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

Review Exercises

5.17 Is the mean less than, approximately equal to, or greater than the median for the distribution shown in Exercise 5.9? *Answer*

5.18 Is the mean less than, approximately equal to, or greater than the median for the distribution shown in Exercise 5.10? *Answer*

5.19 Is the mean less than, approximately equal to, or greater than the median for the distribution shown in Exercise 5.11? *Answer*

5.20 Is the mean divided by the median less than 1, equal to 1, or greater than 1 for a symmetric distribution? *Answer*

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

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

5.6 Measures of Dispersion

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

The range, IQR, and standard deviation computed in this section are summary statistics – i.e., they are computations from individuals in a sample. Thus, they should all be preceded with "sample." See Section 2.1 for clarification on the differences between populations and samples and parameters and statistics.

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

◇ **Measures of dispersion computed from individuals in a sample are preceded by "sample"; those computed from all individuals in a population are preceded by "population."**

5.6.1 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 mouse consumption data (Table 5.1) is 18.4-9.4 = 9.0.

The range should never be used by itself as a measure of dispersion. The range is extremely sensitive to outliers and is best used only to show all possible values present in the data. The range (as strictly defined) also suffers from a lack of information. For example, what does a range of 9 mean? It can have a completely different interpretation if it came from 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.

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

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

5.6.2 IQR

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

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

These calculations are illustrated with the earthquake data (Table 5.2). Recall from above (Section 5.5.2) that the median (=7.1) is in the eighth position of the ordered data. The value in the eighth position will not be included in either half. Thus, the two halves of the data are 5.5 6.3 6.5 6.5 6.8 6.8 6.9 and 7.3 7.3 7.7 7.7 7.7 7.8 8.1. Each half contains seven individuals, so the middle position for each half is $mp = \frac{7+1}{2} = 4$. Thus, the median for each half 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 $Q3 = 7.7$.

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

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

As another example, consider the quartiles of the mouse consumption data (the median was computed in Section 5.5.2). Because $n = 30$ is even, the halves of the data split naturally with 15 individuals in each half. Therefore, the $mp = \frac{15+1}{2} = 8$ and the median of each half is the value of the individual in the eighth position. Thus, $Q1 = 12.5$ and $Q3 = 15.8$. In summary, the first, second, and third quartiles for the mouse water consumption data are 12.5, 13.85, and 15.8, respectively. These three values separate the ordered individuals into approximately four equally-sized groups – those with values less than 12.5, with values between 12.5 and 13.85, with values between 13.85 and 15.8, and with values greater than 15.8.

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

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

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

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

5.6.3 Standard Deviation

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

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

These steps are neatly summarized with

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

The calculation of the standard deviation of the earthquake data (Table 5.2) is facilitated with the calculations shown in Table 5.4. In Table 5.4, note that $\bar{x}$ is equal to the sum of the "Value" column divided by $n = 15$

[12] This statement is not strictly correct as will become obvious. However, this is an acceptable general interpretation of s.

(i.e., $\bar{x} = 7.07$). The "Diff" column which contains each observed value minus the calculated $\bar{x}$ (i.e., Step 2). The "Diff^2" column contains the square of the previously calculated 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 $\frac{6.773}{14} = 0.484$. Finally, the sample standard deviation is the square root of the sample variance or $s = \sqrt{0.484} = 0.696$. Thus, on average, each earthquake is approximately 0.7 Richter Scale units different than the average earthquake in these data.

Table 5.4. 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

Δ **Standard Deviation**: "Essentially" the average deviation or difference of individuals from the mean.

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

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

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

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

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

At the beginning of this section, the standard deviation was defined as "essentially the average difference between the values and the mean." **Essentially** was emphasized because the formula for the standard deviation does not simply add together the differences and divide by n as this definition would imply. Notice in Table 5.4 that the sum of the differences from the mean is 0. This will be the case for all standard deviation calculations using the correct mean, because the mean balances the distance to individuals below the mean with the distance of individuals above the mean (review Section 5.5.5). 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 difference between the values and the mean. Therefore, the original definition of the standard deviation is strictly incorrect; however, it works well as a practical definition of the meaning of the standard deviation.

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

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, why is the sum of the squared differences divided by $n-1$, rather than n, in the standard deviation calculation? Recall (from Section 2.1) that statistics are meant to estimate parameters. The sample standard deviation is supposed to estimate the population standard deviation (σ). Theorists have shown that if we divide by n, s will consistently underestimate σ. Thus, s calculated in this way would be a biased estimator of σ. Theorists have found, though, that dividing by $n-1$ will cause s to be an unbiased estimator of σ. Being unbiased is generally good – it means that on average our statistic estimates our parameter (this concept is discussed in more detail in Module 12).

5.6.4 Measures of Dispersion in R

The minimum, maximum, Q1, Q3, and standard deviation are calculated with `Summarize()` as described previously for the mean and median. Thus, s =21.59, the IQR is from $Q1$ =97.00 to $Q3$ =118.00, and the range is from 48.00 to 146.00.

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

Review Exercises

5.23 Compute the range, IQR, and standard deviation for the maximum gauge heights of the Bois Brule River in Brule, WI from Exercise 5.13 both "by hand" and with R. *Answer*

5.24 Compute the range, IQR, and standard deviation for the population density of Wisconsin counties from Exercise 5.14 both "by hand" and with R. *Answer*

5.25 Compute the range, IQR, and standard deviation of the creatine phosphate data in Exercise 5.5. *Answer*

5.26 Compute the range, IQR, and standard deviation of the CO data in Exercise 5.6. *Answer*

5.7 Overall Summaries

Overall numerical summaries come from considering the relationship between measures of center and dispersion. From the previous section it was seen that the standard deviation and mean are conceptually linked, as are the median and IQR. Indeed, the linked measure of center must be computed first in both dispersion calculations. Thus, if the mean is used to measure center, then the standard deviation must be used to measure dispersion. Similarly, if the median is used to measure center, then the IQR must be used to measure dispersion.[13]

5.7.1 Boxplots

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

The five-number summary may be displayed as a **boxplot**. A traditional boxplot (Figure 5.9) 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 value and from Q3 to the maximum value. The vertical lines have been modified on modern boxplots to allow easier detection of outliers. Specifically, the upper line extends from Q3 to the last observed value that is within 1.5 IQRs of Q3 and the lower line extends from Q1 to the last observed value that is within 1.5 IQRs of Q1. Observed values outside of the whiskers are termed "outliers" by this algorithm and are typically plotted with circles or asterisks. If no individuals are deemed "outliers" by this algorithm, then the two traditional and modern boxplots will be the same.

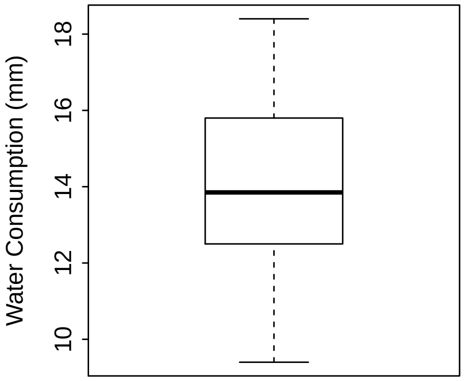


Figure 5.9. Boxplot of the mouse consumption data.

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

The relative length from the median to Q1 and the median to Q3 (i.e., the position of the median line in the box) indicates the shape of the distribution. If the distribution is left-skewed (i.e., lesser-valued individuals

[13] Recall that the range will never be used by itself.

are "spread out"; Figure 5.10-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 5.10-Middle), then Q3-median will be greater than median-Q1. Thus, if the distribution is right-skewed then the median will be closer to Q1 than to Q3, if the distribution is left-skewed then the median will be closer to Q3 than to Q1, and if the distribution is approximately symmetric (Figure 5.10-Left) then the median will be in the middle of the box.

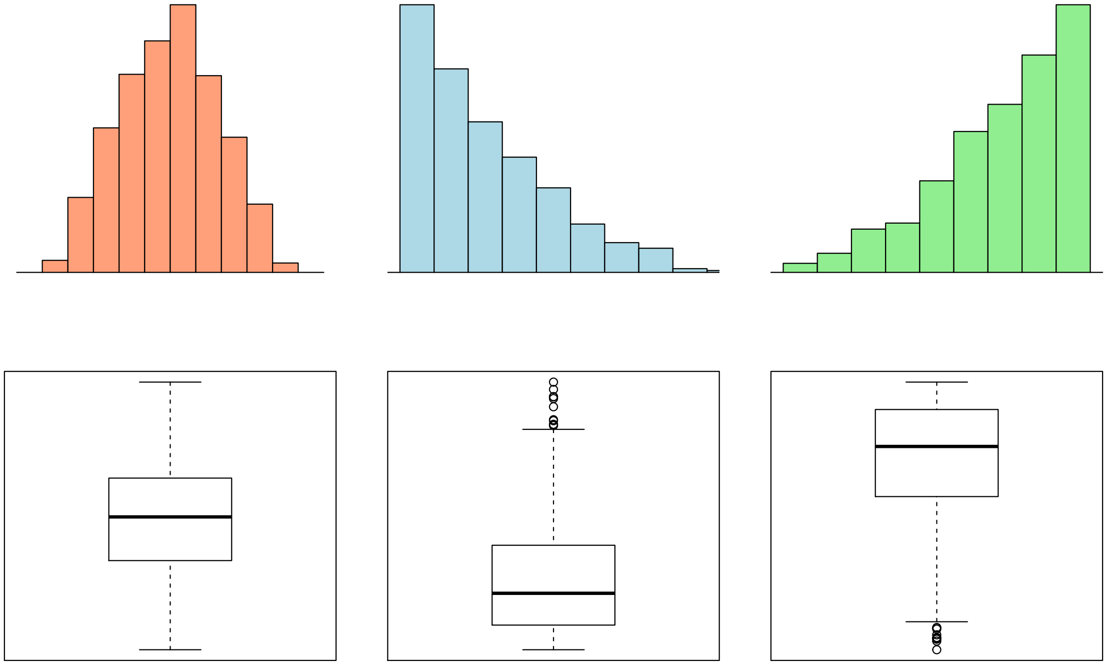


Figure 5.10. Histograms and boxplots for several different shapes of distributions.

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

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

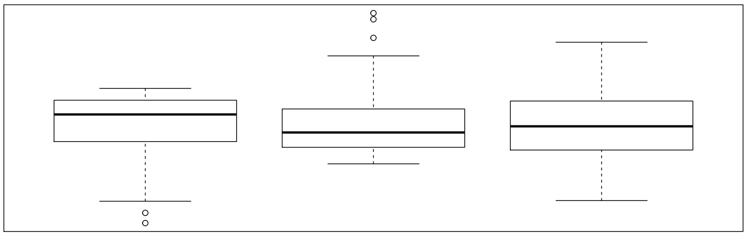
A boxplot is constructed in R with `boxplot()`. This function requires only the name of the quantitative variable as the first argument although the x- and y-axes are labeled with `xlab=` and `ylab=`, respectively.

Review Exercises

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

5.28 R Construct a boxplot for the population density of Wisconsin counties from Exercise 5.14. *Answer*

5.29 What is the shape of the left boxplot below? *Answer*



5.30 What is the shape of the middle boxplot above? *Answer*

5.31 What is the shape of the right boxplot above? *Answer*

5.32 If the distribution is skewed left, which measure should you generally use to measure center? *Answer*

5.33 Which measure of center should you generally use for a right-skewed distribution? *Answer*

5.34 Which measure of center should you generally use for a symmetric distribution? *Answer*

5.35 Which measure of dispersion should you generally use for a symmetric distribution? *Answer*

5.36 Which measure of dispersion should you generally use for a left-skewed distribution? *Answer*

5.37 Which measure of dispersion should you generally use for a right-skewed distribution? *Answer*

5.38 Is Q3-Q2 less than, approximately equal to, or greater than Q2-Q1 if the data are left-skewed? *Answer*

5.39 What is the shape of the distribution if Q3-Q2 is greater than Q2-Q1? *Answer*

5.8 Multiple Groups

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

As an example, suppose that you want to examine the average annual days of ice for each decade (using the `LSI` data). One might expect to use the `days~decade` formula except that the `decade` variable is not a factor.[14] This can be connverted to a factor by including the variable to the left of the assignment operator and in `factor()`. The desired grouping variable may already be a factor in many data.frames and, thus, will not require modification with `factor()`.

```
> LSI$decade <- factor(LSI$decade)
> str(LSI)
'data.frame': 42 obs. of  4 variables:
 $ season: int  1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 ...
 $ decade: Factor w/ 5 levels "1950","1960",..: 1 1 1 1 1 2 2 2 2 2 ...
 $ temp  : num  22.9 23 25.7 20 24.8 ...
 $ days  : int  87 137 106 97 105 118 118 136 91 NA ...
```

Histograms (Figure 5.11) and summary statistics separated by decade are then constructed as below.

```
> hist(days~decade,data=LSI,ylab="Days of Ice Cover",w=20)
> Summarize(days~decade,data=LSI,digits=2)
  decade  n nvalid   mean    sd min     Q1 median    Q3 max
1   1950  5      5 106.40 18.73  87  97.00  105.0 106.0 137
2   1960 10      8 113.12 14.80  91 104.20  116.0 119.8 136
3   1970 10     10 115.50 19.19  82 105.80  115.0 124.0 146
4   1980 10     10 103.80 24.88  48  90.25  116.0 118.0 123
5   1990  7      6  96.00 28.53  62  72.00  100.5 114.0 132
```

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

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

[14] It was not a factor because the data in `decade` looks numeric to R.

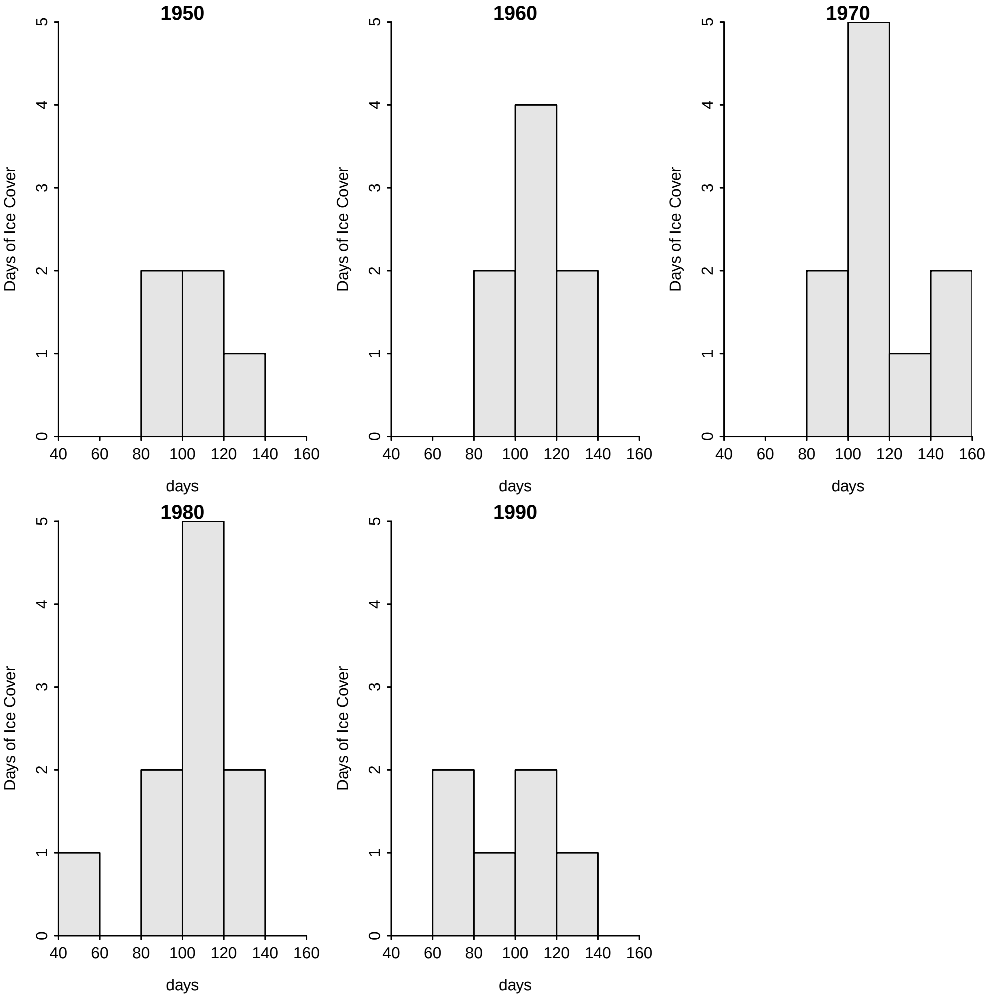


Figure 5.11. Histograms of the duration of ice cover at ice gauge 9004 in Lake Superior by each decade.

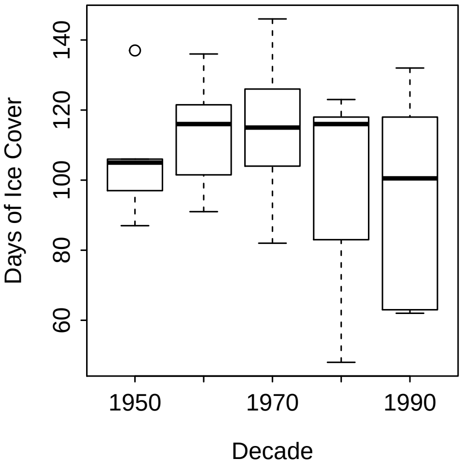


Figure 5.12. Boxplot of the duration of ice cover at ice gauge 9004 in Lake Superior by each decade.

Review Exercises

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

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

5.9 Example Interpretations

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

Mouse Consumption Example

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

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

Table 5.5. Descriptive statistics of mouse water consumption.

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

R Appendix:

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

Crayfish Temperature Selection

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

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

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

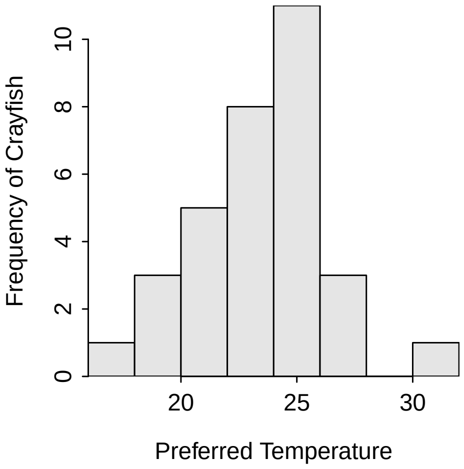


Figure 5.13. Histogram of crayfish temperature preferences.

Table 5.6. Descriptive statistics of crayfish temperature preferences.

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

R 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)
```

Review Exercises

5.41 R Construct a proper EDA for the creatine phosphokinase data presented in Exercise 5.5. Make sure to defend your choice of numerical summaries. *Answer*

5.42 R The Dow Jones Travel Index tracks the cost of hotel and car-rental rates in 20 major cities. For its May 7, 1996, survey the following rates were given for the 20 cities: 152, 180, 167, 119, 115, 113, 119, 135, 140, 126, 114, 133, 205, 104, 149, 124, 127, 161, 106, and 179. Thoroughly describe the distribution of these data. [*Note: You can use fewer than the ideal number of bars on your histogram because the sample size is so small in this situation.*] *Answer*

5.43 R The data in Zoo2.csv contains the physical size (in acres) of a sample of zoos from around the United States. Perform a univariate EDA on the *size* variable. *Answer*