Chapter 3

**Key Ideas**

Measures of Center: Mean, Median, Mode
Skewness
Measures of Variation: Range, Standard Deviation, Variance
Chebyshev's Theorem and The Empirical Rule (a.k.a. 68-95-99.7 Rule)
Measures of Relative Standing: Z-Score, Quartile, Percentile, Interquartile Range (IQR)
Exploratory Data Analysis: Outlier, 5-Number Summary, Boxplot, Modified Boxplot

## Section 3-1: Overview

In Chapter 2, we explored different ways to display data sets visually. However, we now need to come up with a way to describe data sets that will allow us to compare various characteristics of different sets. This can be accomplished by computing values (statistics) from the observations that represent some aspect of the data set.

Before discussing measures of center and the formulas for their calculation, here is a quick review of summation notation.
Sigma (Summation) Notation is a way to concisely represent the summation of many numbers at a time. Here is what all the parts stand for:

Ending Value

Some function of the index

Indicates that we are summing

$$\sum_{i=1}^{k} f(i) = f(1) + f(2) + f(3) + ... + f(k)$$

*Index* letter

Starting Value

Examples

$$\sum_{i=1}^{3} i^2 = 1^2 + 2^2 + 3^2 = 1 + 4 + 9 = 14$$

$$\sum_{i=0}^{4} \frac{i+1}{2} = \frac{0+1}{2} + \frac{1+1}{2} + \frac{2+1}{2} + \frac{3+1}{2} + \frac{4+1}{2} = \frac{1}{2} + \frac{2}{2} + \frac{3}{2} + \frac{4}{2} + \frac{5}{2} = \frac{15}{2}$$

$$\sum_{i=1}^{3} Observation\ i = Observation\ 1 + Observation\ 2 + Observation\ 3$$

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$$

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + ... + x_n$$

## Section 3-2: Measures of Center

A *measure of center* is a value representing the center, or middle, of a data set. There are several different measures that may be used depending on the nature of the observations.

- The **mean** of the data is the arithmetic mean (or average) of the observations. Simply add up all of the values and divide by the total number of observations. The symbol for the mean is $\bar{x}$. The formula for the mean is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where the *n* observations are given by $x_1, x_2, x_3, ..., x_n$.
- The **median** of the data is the middle value of the data set if the observations are arranged in order. Note that for an *odd* number (*n*) of observations, there is an exact middle value in the $\frac{n+1}{2}$ position. However, for an *even* number (*n*) of observations, there are two middle values in the $\frac{n}{2}$ and $\frac{n}{2}+1$ positions. The convention is to average those two values to get the median. Sometimes, the median is denoted $\tilde{x}$.
- The **mode** of the data is the most common value of the observations. If 2 values tie for the largest frequency, the data is called *bimodal.* If more than 2 values tie, it is called *multimodal.* Finally, if no values repeat (i.e. all values are tied for largest), then the data set is said to have *no mode*.

- The **midrange** is the average of the minimum and maximum values in the dataset. We won’t use this value at all.

Examples

Data set #1:<br>
1, 1, 3, 5, 6

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5) = \frac{1}{5}(1+1+3+5+6) = \frac{16}{5} = 3.2$$

$\tilde{x} = 3$ (Note: the 3 is found in the $\frac{n+1}{2} = \frac{5+1}{2} = 3^{\text{rd}}$ position)

Mode = 1

Data set #2:<br>
2, 3, 4, 5, 9, 12

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{6}\sum_{i=1}^{6} X_i = \frac{1}{6}(X_1 + X_2 + X_3 + X_4 + X_5 + X_6) = \frac{1}{6}(2+3+4+5+9+12) = \frac{35}{6} = 5.83$$

$\tilde{x} = \frac{4+5}{2} = \frac{9}{2} = 4.5$ (Note: we average the numbers 4 and 5, located in the $\frac{n}{2} = \frac{6}{2} = 3^{\text{rd}}$ and $\frac{n}{2} + 1 = 4^{\text{th}}$ positions)

No Mode

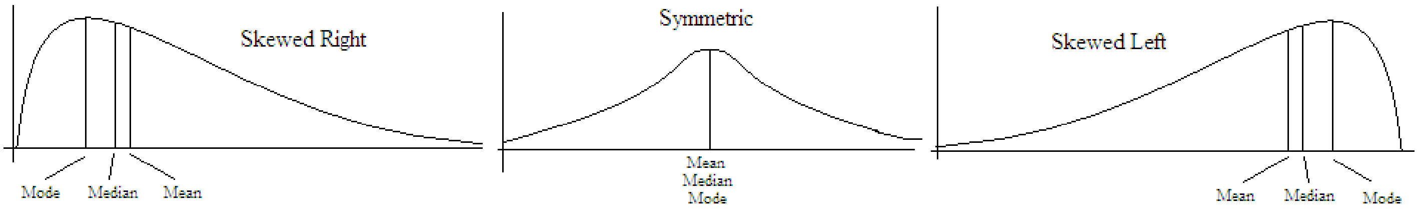
Note: Suppose a data set is represented as a frequency distribution using intervals as the groups, but you want to calculate the mean. Since you do not know the exact observation values, it is customary to treat all observations in an interval as being the midpoint of the interval. For example, if the interval 20-30 has a frequency of 10, you should pretend that there are 10 observations with a value of 25 for the purposes of computing the mean.

Since there are 3 different measures of centers, it seems reasonable to ask which is best to use. There are advantages and disadvantages to each of them, depending on the nature of the data set. These are listed below.

| Measure | Advantages | Disadvantages |
|---|---|---|
| Mean | Easy to Compute<br>Sample Means tend to Vary Less<br>Good properties as sample size increases (more to come on that later) | Sensitive to extreme values (outliers) |
| Median | Resistant to outlying values<br>Good for skewed data (see below) | Harder to calculate<br>Less useful than the mean for inference (more to come on that later) |
| Mode | Easy to compute<br>Good for qualitative (categorical) data | Not very useful for quantitative data |

Skewness

Using the mean, median, and mode together can help to describe the *skewness* of a data set. A data set is considered skewed if the values extend more to one side of the distribution than the other.



## Section 3-3: Measures of Variation

Now that we can measure the center of a data set, it may be useful to be able to measure how much the values differ from each other. In other words, we would like to distinguish between the following data sets, all of which have $\bar{x} = \tilde{x} = \text{mode} = 3$:

3 3 3 3 3      0 3 3 3 6      -58 2 3 3 65      0 3 3 4 5

While the centers of the data sets are the same, the variation is clearly different. The first data set always gives the same values (no variation), whereas the third data set has observations that vary wildly in positive and negative directions. To distinguish between these sets, we use *measures of variation*.

- **Range** is the simplest measure of variation. It is the difference between the maximum and minimum values. The formula is: *Range* = *Maximum* – *Minimum*. There is one shortcoming of range, though. We still cannot distinguish between data sets like 0 5 5 5 5 5 5 5 5 5 10 and 0 1 2 3 4 5 6 7 8 9 10 (clearly, the data set on the right seems to vary more). Both have a range of 10 however.
- The **Standard Deviation** is a better way to measure variation. First, we look at the difference between each data value and the mean: $x_i - \bar{x}$. Then, to make sure all the distances are positive, we square that difference: $(x_i - \bar{x})^2$. Next, we add up these differences for all of the observations: $\sum_{i=1}^{n}(x_i - \bar{x})^2$. Finally, we divide by ($n$ –1) and take the square root to in some way undo the squaring from before. This gives the formula: $Std.Dev. = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Standard Deviation is denoted *s*, and it is in the same units as the observations in the data set.
- The **Variance** is the square of the standard deviation, $s^2$. Since its units are in squared units of the original observations, its value is harder to interpret than standard deviation. As a result, it is not used as much.

Example

Consider the data set 1 2 3 4 5. We see that $\bar{x} = \tilde{x} = 3$. Let's find the range, standard deviation, and variance.

*Range* = *Maximum* – *Minimum* = 5 – 1 = 4.

To find standard deviation, we can use a table to calculate each part separately.

| | $x_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|
| | 1 | 1 – 3 = -2 | 4 |
| | 2 | 2 – 3 = -1 | 1 |
| | 3 | 3 – 3 = 0 | 0 |
| | 4 | 4 – 3 = 1 | 1 |
| | 5 | 5 – 3 = 2 | 4 |
| Sum | --- | --- | 10 |

We see that $\sum_{i=1}^{5}(x_i - \bar{x})^2 = 10$, so:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{5}(x_i - \bar{x})^2} = \sqrt{\frac{1}{4}(10)} = \sqrt{2.5} = 1.58$$

Finally, the variance is $s^2 = 2.5$.

Now, consider the data set 1 1 3 5 5. We see that $\bar{x} = \tilde{x} = 3$ again. Let's find the range, standard deviation, and variance.

*Range* = *Maximum* – *Minimum* = 5 – 1 = 4.

To find standard deviation, we can use a table to calculate each part separately.

| | $x_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|
| | 1 | 1 – 3 = -2 | 4 |
| | 2 | 1 – 3 = -2 | 4 |
| | 3 | 3 – 3 = 0 | 0 |
| | 4 | 5 – 3 = 2 | 4 |
| | 5 | 5 – 3 = 2 | 4 |
| Sum | --- | --- | 16 |

We see that $\sum_{i=1}^{5}(x_i - \bar{x})^2 = 16$, so:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{5}(x_i - \bar{x})^2} = \sqrt{\frac{1}{4}(16)} = \sqrt{4} = 2$$

In this data set, the variance is $s^2 = 4$. So we see that the second data set has more variation than the first one, which makes sense. The observations are spread further away from the mean than in the first set of data.

Some Notation

In a *sample*, the standard deviation is denoted $s$ and the variance $s^2$.
In a *population*, we call the standard deviation $\sigma$ (sigma) and the variance $\sigma^2$ (sigma squared).

A quick note about the standard deviation formula: Often, people wonder why we divide by $n - 1$ in the formula instead of $n$, which is the number of observations. The reason is because of something called *degrees of freedom*. We already need to know the mean $\bar{x}$ for the standard deviation formula. Therefore, if we know only $n - 1$ of the observations, we could figure out the last one, since the mean tells you the sum of the observations. As a result, it turns out that dividing by $n - 1$ will make the standard deviation an *unbiased estimator*, meaning that as the sample size increases, $s$ will not consistently overestimate or underestimate the true population standard deviation. (see page 103 in the book for more discussion).

Also, for discussion on why we square the differences instead of taking absolute value, see page 102.

Comparing Variation Among Data Sets

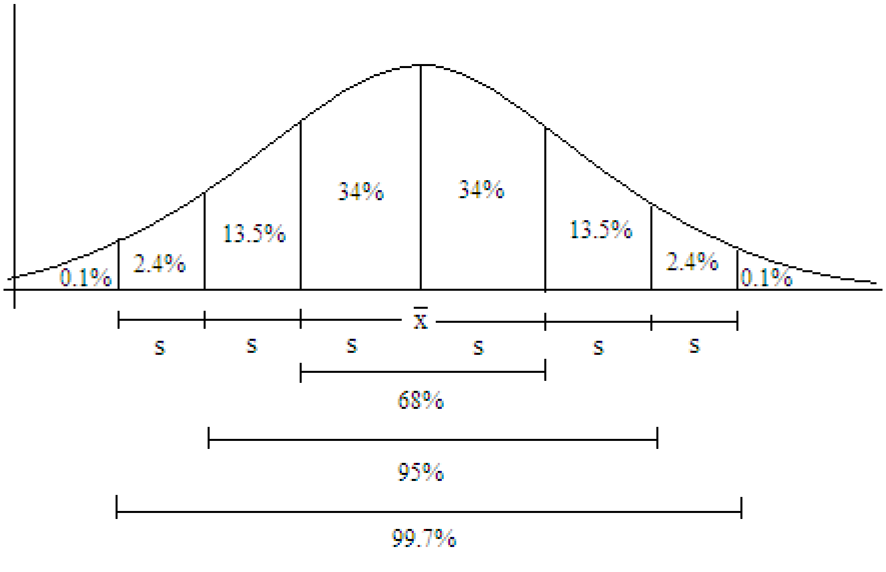
Sometimes, we may want to compare the variation in two data sets, but the units are not the same. For instance, consider the ACT and the SAT college entrance exams. Since the ACT scores range 0-36, but SAT scores range 0-1600, we will always have a larger mean and standard deviation for SAT scores because the numbers are larger. However, we may want to compare the variation in scores on the exams *relative to their observation sizes*. To do this, we use the **coefficient of variation**, which is denoted $CV = \frac{s}{\bar{x}} \cdot 100\%$. This calculates the percent of variation relative to the mean, and essentially puts the variation in any data set on the same scale.

Applications of Standard Deviation

There are a couple of useful theorems relating to standard deviation that result in some "rules of thumb" that are especially useful for people who want a good idea of what is going on in a data set without rigorous calculation.

**The Empirical Rule** (also called the 68-95-99.7 Rule): If a data set has a bell-shaped (normal) distribution, then:

(i) About 68% of the observations fall within one standard deviation of the mean (between $\bar{x} - s$ and $\bar{x} + s$)
(ii) About 95% of the observations fall within two standard deviations of the mean (between $\bar{x} - 2s$ and $\bar{x} + 2s$)
(iii) About 99.7% of the observations fall within three standard deviations of the mean (between $\bar{x} - 3s$ and $\bar{x} + 3s$)
(iv) Furthermore, since a normal distribution is symmetric, we get the other percentages shown below (all approximate)



A note: This rule is an approximation based on percentiles of the normal distribution, which we will discuss later in the course.

**Chebyshev's Theorem**: In <u>any</u> data set (even skewed ones), the proportion of values lying within $k$ standard deviations of the mean is always greater than or equal to $1 - \frac{1}{k^2}$, where $k > 1$. In particular, this means:

(i) More than 75% of the observations fall within two standard deviations of the mean (between $\bar{x} - 2s$ and $\bar{x} + 2s$)
(ii) More than 89% of the observations fall within three standard deviations of the mean (between $\bar{x} - 3s$ and $\bar{x} + 3s$)

## Section 3-4: Measures of Relative Standing

It is certainly useful having measures of center and variation for data sets, but what if you wanted to compare two data sets with different units to each other? For instance, what if you wanted to compare test grades for two students from two different classes, where one class took the SAT and the other took the ACT? The mean and standard deviation for the SAT class would definitely be higher, but that is only because the SAT is 1600 points total as opposed to 36 for the ACT. In order to compare these students' scores, we need a measure of *relative standing*.

### Z-Scores

To compare two groups, then, what can be done? First of all, we could subtract the group mean from each observation in each class. This would put the center of each data set at 0. Next, we can divide each observation by the group standard deviation, which scales everything down to a standard deviation of 1. Since both data sets now have a mean 0 and standard deviation 1, observations can be compared. Subtracting the mean and dividing by the standard deviation for an observation is called finding the *z-score*.

**Z-Score** (for a sample): $z = \dfrac{x - \bar{x}}{s}$

Here, $x$ is the observation, $\bar{x}$ is the sample mean, and $s$ is the sample standard deviation

**Z-Score** (for a population): $z = \dfrac{x - \mu}{\sigma}$

Here, $x$ is the observation, $\mu$ is the population mean, and $\sigma$ is the population standard deviation.

A z-score tells you how many standard deviations above (or below) the mean an observation is located.
For example, $z = -1.45$ means the observation is located 1.45 standard deviations *below* the mean.
*To compare two observations*, simply compare the z-scores for each observation. Whichever observation has a larger z-score has a higher *relative standing* (i.e. it is higher compared to other observations in its data set).

Example: Carol and Meredith, two sisters, are arguing about who has to do more homework compared to their classmates. Carol is in college, and does about 7 hours of homework a week. Meredith is in high school and does 4.5 hours of homework a week. For the purposes of this example, let's suppose college students' study hours have a mean of 5 hours and a standard deviation of 1, while high school students have a mean of 3 hours and a standard deviation of 0.5. Who studies more relative to their peers?

Carol's z-score is: $z = \dfrac{x - \mu}{\sigma} = \dfrac{7 - 5}{1} = 2$, so Carol's hours are 2 standard deviations above the mean for college students.

Meredith's z-score is: $z = \dfrac{x - \mu}{\sigma} = \dfrac{4.5 - 3}{0.5} = 3$, so Meredith's hours are 3 standard deviations above the mean for high school students.

So we see that while Meredith studies fewer hours a week, she still studies more *relative* to her classmates.

### Interpreting Z-Scores

From Chebyshev's Theorem, we know that more than 75% of observations will always fall within 2 standard deviations of the mean.
Furthermore, for bell-shaped distributions, 95% of observations fall within 2 standard deviations of the mean.
For this reason, we say that observations with z-scores between –2 and 2 (within 2 s.d. of the mean) are **Ordinary Values**.
Observations with $z < -2$ or $z > 2$ (further than 2 s.d. from the mean) are called **Unusual Values**.

### Quartiles

One of our measures of center was the median, which was the middle observation of the data set. This is the unique point where 50% of the data set lies above and below that value. It addition to the median, there are other special values called **quartiles**, which mark each 25% of the data set. They are marked with a Q and a number denoting which quarter of the data they represent.

**$Q_1$:** First Quartile – 25% of the observations are below this point, and 75% above.
**$Q_2$:** Second Quartile – 50% of the observations are below this point, and 50% above (this is the median).
**$Q_3$:** Third Quartile – 75% of the observations are below this point, and 25% above.

### Percentiles

A more general form of quartiles splits the data set at *any* percent, instead of just 25%, 50%, or 75%. For example, the 90th percentile is the observation where 90% of the data lies below that value, and 10% above. What follows is a general method for finding percentiles (including quartiles).

1. Decide what percent of the data you want below the value. Call this $k$ (e.g. for $Q_1$, you would let $k = 0.25$, for the 25th percentile).
2. Sort the data.
3. Multiply $k$ by the total number of observations $n$. Let $L = kn$.
   (note: in the book, they let $k$ be the percent number – i.e. 25 instead of 0.25 – and the method is a little different)
4. If $L$ is a whole number, average the observations in the $L^{\text{th}}$ and $(L+1)^{\text{st}}$ positions in the data set to get the percentile.
5. Otherwise, round $L$ up to the next whole number, and use the observation in the $L^{\text{th}}$ position as the percentile.
6. Often, the percentile is denoted $P_k$, where the number in the subscript is either the percent or the decimal version of the percent.

Example: Consider the following dataset – 4, 5, 2, 6, 8, 10, 2, 4, 0, 34, 11, 3. There are $n = 12$ observations.
Let's find all 3 quartiles and the 80th percentile. First, we sort the data: 0, 2, 2, 3, 4, 4, 5, 6, 8, 10, 11, 34
Finding $Q_1$: $k = 0.25$, and $L = kn = 0.25(12) = 3$. Since $L$ is a whole number, we average the 3rd and 4th observations to get $Q_1 = 2.5$.
Finding $Q_2$: $k = 0.5$, and $L = kn = 0.5(12) = 6$. Since $L$ is a whole number, we average the 6th and 7th observations to get $Q_2 = 4.5$.
Finding $Q_3$: $k = 0.75$, and $L = kn = 0.75(12) = 9$. Since $L$ is a whole number, we average the 9th and 10th observations to get $Q_2 = 9$.
Finding $P_{80}$: $k = 0.80$, and $L = kn = 0.80(12) = 9.6$. Now we round $L$ up to get 10, and use the 10th observation. So $P_{80} = 10$.

Another Measure of Variation
As we learned before, the mean and standard deviation are heavily influenced by extremely high or low values, and in those situations the median should be used as a measure of center. However, what should we use to measure variation? One measure that is often used with the median is called the **interquartile range (IQR)**.
It is defined by: **IQR = $Q_3 – Q_1$**.

## Section 3-5: Exploratory Data Analysis (EDA)

*Exploratory Data Analysis* is the way that many scientists discover trends in data sets that are not immediately obvious. This is done through the use of graphs and basic statistics (i.e. mean, median, mode, standard deviation, quartiles, IQR, etc.). Other than the graphs and statistics we have already discussed, there are a few more ways scientists can do this.

Outliers
One important feature of a data set is something called an **outlier**. This is an observation that is much higher or lower than most of the other observations in the data set. Often, outliers represent some sort of error (i.e. the measurement device malfunctioned, someone wrote the number down wrong, etc.). However, other outliers represent some sort of anomaly that may provide insight into the process being examined in the data set. These values also have a big effect on the mean and standard deviation, and are therefore important to identify.

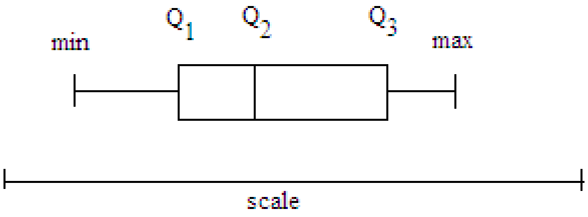
*The Outlier Rule*: If an observation is 1.5 times the IQR above $Q_3$ or below $Q_1$, then it is an outlier. This means if an observation falls outside the interval ($Q_1 – 1.5 \cdot \text{IQR}$, $Q_3 + 1.5 \cdot \text{IQR}$), then it is an outlier.

If an outlier is determined to be valid (not a malfunction or mistake), statisticians will often examine graphs and statistics for data sets excluding the outlier in order to see what effect it has on the interpretation of the data set.

Boxplots
A **boxplot** is a very useful plot that divides a data set into its quartiles. While not as useful as a histogram for a single data set, it is incredibly useful in comparing data sets to each other. Before we discuss how to create a boxplot, though, we need to first define a *5-Number Summary*.

A **5-Number Summary** is a collection of five numbers in a data set: the minimum value, $Q_1$, $Q_2$, $Q_3$, and the maximum value. Usually, it is written in the following form: (min, $Q_1$, $Q_2$, $Q_3$, max).

A boxplot (also called a box-and-whisker diagram) is formed by a box from $Q_1$ to $Q_3$, two lines running from min to $Q_1$ and $Q_3$ to max, and vertical lines at each of the 5 numbers in the 5-number summary (see below).



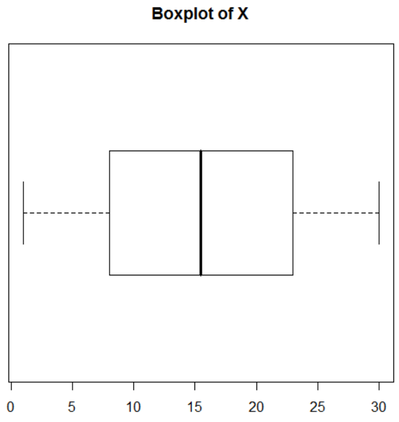
The thing to remember with boxplots is that 25% of the observations fall between each pair of vertical lines (25% in the two "whiskers", 25% in each half of the box). Thus, a smaller distance between vertical lines means observations are densely packed in that area (this would translate into higher bars on a histogram). A wider area between lines means the data values are more spread out.

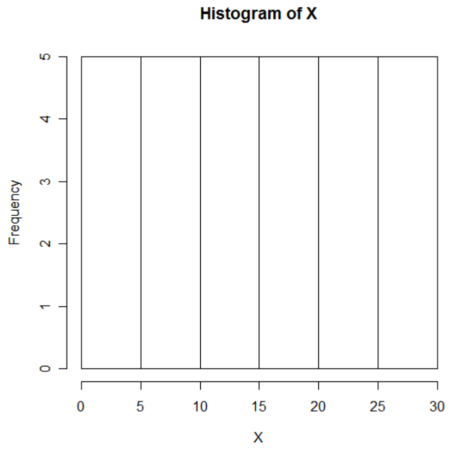
A **modified boxplot** uses the same design, except all outliers are displayed by a symbol (usually a *), and the whiskers only extend to the highest and lowest observations that are *not* outliers (here, outliers are determined by the outlier rule above).

Examples

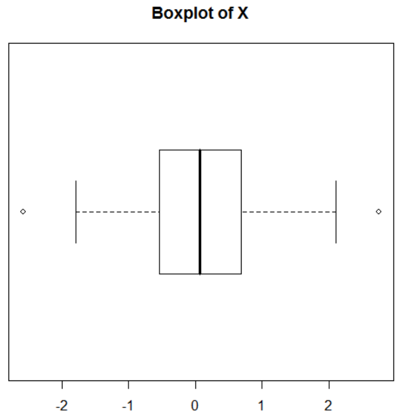
Here are some side-by-side histograms and boxplots for various data sets:

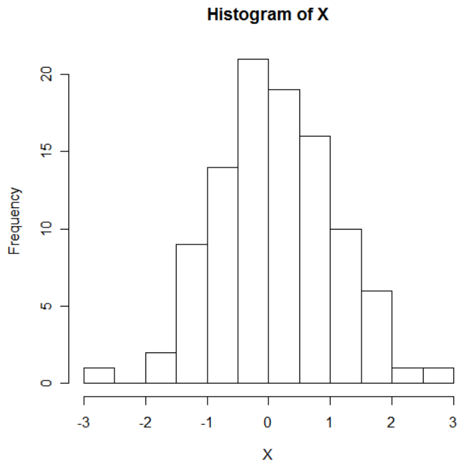
*Data Set #1*

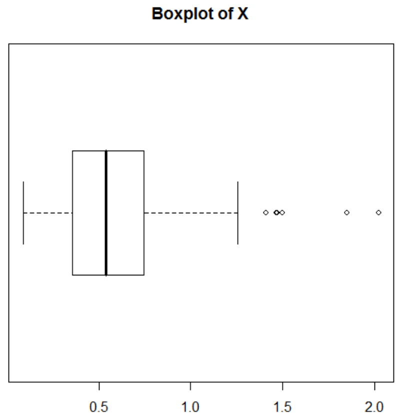


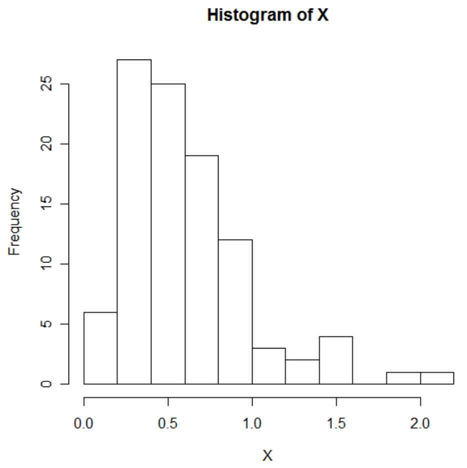


*Data Set #2*



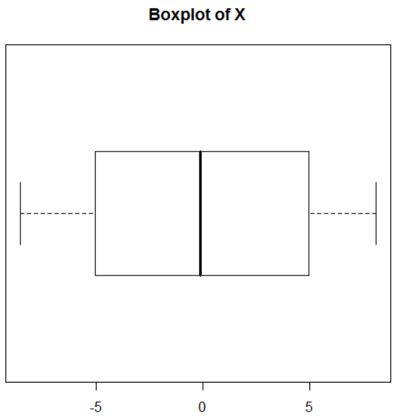


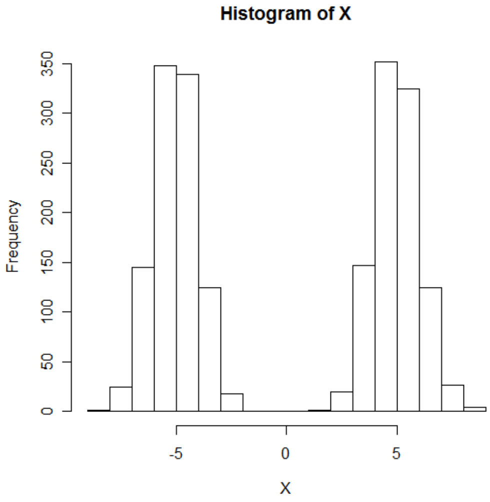




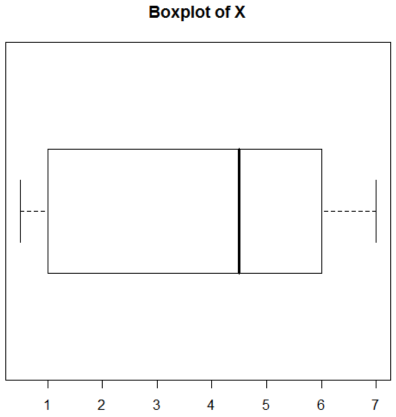
*Data Set #3*

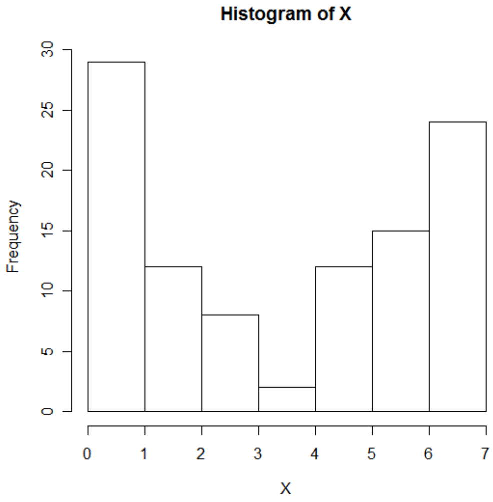
*Data Set #4*





*Data Set #5*


Boxplot of X


Histogram of X
Frequency
X