Chapter 2
Summarizing and Graphing Data

-- Organizing and Displaying Data
  Frequency Tables

-- Graphical Displays

-- Numerical Measures
  Measures of Central Tendency
  Measures of Variation
  Measures of Position
Organizing and Displaying Data
Example 2.1: NFL Data

Table 2.1 shows an example of raw data, which gives the 40 National Football League (NFL) players’ weights (in pounds).


Table 2.1: NFL Data

Very Little information can be obtained from looking at the above raw data.
The most common method of organizing raw data is to construct a frequency table.

**Frequency Table** is a table including different classes (or groups) and their respective frequencies (or counts within the classes) is known as a frequency table. A frequency table can be constructed using the following steps:

Step 1: Determine number of classes or groups depending on the sample size $n$. A rule of thumb is $c = \ln(n)$, where $c$ is the number of classes, ‘$\ln$’ stands for the natural logarithm.

Step 2: Determine the range of the data as $Range = Maximum - Minimum$.

Step 3: Determine the width or length of the class intervals as $\frac{Range}{c}$.

Step 4: Determine the classes by introducing an extra decimal position so that each data value belongs to exactly one class.

Step 5: Use tally marks to put all data values in different classes and count the tallies to find the frequencies for each class.
Example 2.1 Frequency table for the NFL data in table 2.1 can be constructed as follows:

Step 1: \( c = \ln(n) = \ln(40) \approx 4 \).

Step 2: \( \text{Range} = \text{Maximum} - \text{Minimum} = 331 - 152 = 179 \).

Step 3: \( \text{Class Width} = \text{Range} / c = 179 / 4 \approx 45 \).

Step 4: The classes are 152-196, 197-241, 242-286, and 287-331.

Step 5: The class boundaries are 151.5-196.5, 196.5-241.5, 241.5-286.5, and 286.5-331.5.

Table 2.2: Frequency table from the NFL data in table 2.1

<table>
<thead>
<tr>
<th>Class Limits</th>
<th>Class Boundaries</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>152-196</td>
<td>151.5-196.5</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>197-241</td>
<td>196.5-241.5</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>242-286</td>
<td>241.5-286.5</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>287-331</td>
<td>286.5-331.5</td>
<td></td>
<td>7</td>
</tr>
</tbody>
</table>
Some of the terms commonly used to describe a frequency table are given below.

**Lower class limits** are the smallest numbers that can belong to the different classes. In table 2.2 the values 152, 197, 242 and 287 are the lower class limits.

**Upper class limits** are the largest numbers that can belong to the different classes. In table 2.2 the values 196, 241, 286 and 331 are the upper class limits.

**Class boundaries** are the numbers used to separate the classes but without the gaps. To obtain the class boundaries, find the gaps created by class limits. Subtract the half of the gap to each of the lower limit and add the half of the gap to each of the upper limit. In table 2.2 the gap between class limits is 1 so we subtract 0.5 to each of the lower limits and add 0.5 to each of the upper limit to get 151.5, 196.5, 241.5, 286.5, 331.5 as the class boundaries.
Class midpoints are the mid points of each class. Class midpoints are found by adding the lower class boundary to upper class boundary and dividing by 2. 174, 219, 264, 309 are the class midpoints in table 2.2.

Class width is the difference between two consecutive lower class limits or two consecutive lower class boundaries. In table 2.2 class width is 45.

Additional guidelines for constructing a frequency table
1. Usually there are between 5 to 20 classes, that is, to avoid having too few or too many classes. The researcher decides how many classes to use.
2. The classes are mutually exclusive (or non-overlapping).
3. The classes are continuous.
4. The classes are exhaustive (accommodate all the data).
5. The classes are equal in width.
Cumulative and Relative Frequency Tables

Frequency of the class is the number of data values contained in a specific class. Sometimes it is helpful to construct cumulative or relative frequency tables. The cumulative frequency for a class is the sum of the frequencies for that class and all previous classes. The cumulative frequency table is basically the list of classes and their respective cumulative frequencies.

Relative frequencies (Rel. Freq.) are obtained by dividing the class frequency by the total number of observations, which is the sum of all frequency.

\[
\text{Relative Frequency} = \frac{\text{Frequency}}{\text{Total Frequency}}
\]
An example of cumulative frequency table can be seen in table 2.3 for the data in table 2.2.

Table 2.3 shows the cumulative frequency (Cum. Freq.) and relative frequency (Rel. Freq.) table of the NFL data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>152-196</td>
<td>151.5-196.5</td>
<td></td>
<td>12</td>
<td>12</td>
<td>12/40</td>
</tr>
<tr>
<td>197-241</td>
<td>196.5-241.5</td>
<td></td>
<td>7</td>
<td>19</td>
<td>7/40</td>
</tr>
<tr>
<td>242-286</td>
<td>241.5-286.5</td>
<td></td>
<td>14</td>
<td>33</td>
<td>14/40</td>
</tr>
<tr>
<td>287-331</td>
<td>286.5-331.5</td>
<td></td>
<td>7</td>
<td>40</td>
<td>7/40</td>
</tr>
</tbody>
</table>
Graphical Displays
After the data have been organized into a frequency table, they can be presented in a graphical form. The purpose is to describe the data using tabular or pictorial forms. There are many graphical displays, such as, Stem-Leaf Display, Bar Graph, Pie Chart, Histogram, Relative Frequency Histogram, Box Plot, etc.
Histogram

We use a visual tool called a histogram to analyze the shape of the distribution of the data. It is a graph consisting of bars of equal width drawn adjacent to each other (without gaps). The horizontal scale represents the classes of quantitative data values and the vertical scale represents the frequencies. The heights of the bars correspond to the frequency values.
Histogram

Basically a graphic version of a frequency distribution.

Histogram

Classes

151.5-196.5
196.5-241.5
241.5-286.5
286.5-331.5
Histogram

The bars on the horizontal scale are labeled with one of the following:

(1) Class boundaries

(2) Class midpoints

(3) Lower class limits (introduces a small error)

Horizontal Scale for Histogram: Use class boundaries or class midpoints.

Vertical Scale for Histogram: Use the class frequencies.
Relative Frequency Histogram

Has the same shape and horizontal scale as a histogram, but the vertical scale is marked with relative frequencies instead of actual frequencies.
Distribution Shapes

Uniform & symmetrical

Skewed right

Symmetrical

Skewed left

Bimodal & symmetrical

Bimodal & skewed right
Stem-Leaf Display

Stem-Leaf display is the simplest method of summarizing a numerical data set when the number of observation in the data is not too large. From a quick glance of the data, a grouping is established, the common leading digits are known as stems and the remaining digits are known as leaves. The Stem-Leaf display formed for the NFL data is shown on the next page with stem bring 1, 2, and 3 and leaves are the two trailing digits.
Stem-Leaf display shows a quick distributional pattern of the data. As it can be seen in the above example, the measurements within the leaves are not necessarily ordered. Stems can be repeated for lower and higher values whenever necessary.
Note that in the previous Stem-Leaf display there are three groups (or stems). If there is a need to have more groups, each group could be divided as low and high as follows:

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-low</td>
<td></td>
</tr>
<tr>
<td>1-high</td>
<td>56 78 92 87 52 75 73 81 74 59 84 92</td>
</tr>
<tr>
<td>2-low</td>
<td>34 48 47 49 21 42 36 09 18 13 41</td>
</tr>
<tr>
<td>2-high</td>
<td>58 76 67 96 64 53 67 74 54 73 53</td>
</tr>
<tr>
<td>3-low</td>
<td>20 12 04 25 27 31</td>
</tr>
<tr>
<td>3-high</td>
<td></td>
</tr>
</tbody>
</table>
Bar Graph

Uses bars of equal width to show frequencies of categories of qualitative data. Vertical scale represents frequencies or relative frequencies. Horizontal scale identifies the different categories of qualitative data.

A *multiple bar graph* has two or more sets of bars, and is used to compare two or more data sets.
Multiple Bar Graph

Figure 2.4 Bar Graph for the Cookie Shop Data
Multiple Bar Graph

Figure 2.5 Bar Graph for Snack Choices Data

![Preferred snack choices of students at Hillary's high school](image)
Pie Chart

Figure 2.6 Pie Chart for Federal Expenditure Data
Skewness

A distribution is skewed if one of its tails is longer than the other. The first distribution shown below has a positive skew. This means that it has a longer tail in the positive direction. The distribution in the middle has a negative skew since it has a longer tail in the negative direction. Finally, the third distribution is symmetric and has no skew. Distributions with positive skew are sometimes called "skewed to the right" whereas distributions with negative skew are called "skewed to the left". And the symmetric distributions are often called bell shaped.

Figure 2.7 Skewness
Numerical Measures
Measures of Central Tendency
Arithmetic Mean (Mean)
the measure of center obtained by adding the values and dividing the total by the number of values

What most people call an average.
Notation

\( \Sigma \) denotes the sum of a set of values.

\( x \) is the variable usually used to represent the individual data values.

\( n \) represents the number of data values in a sample.

\( N \) represents the number of data values in a population.
Notation

\( \bar{x} \) is pronounced ‘x-bar’ and denotes the mean of a set of sample values

\[
\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}
\]

\( \mu \) is pronounced ‘mu’ and denotes the mean of all values in a population

\[
\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum x}{N}
\]

The sample mean is an unbiased estimator of the population mean. And it is true for any arbitrary population.
Example 2.5: The data represent scores on an exam in a statistics class of 10 students:

\[ 88 \quad 70 \quad 75 \quad 82 \quad 89 \quad 82 \quad 90 \quad 91 \quad 65 \quad 77 \]

Find the mean score.

\[
\bar{x} = \frac{\sum x}{n} = \frac{(88 + 70 + 75 + 82 + 89 + 82 + 90 + 91 + 65 + 77)}{10} = 80.9
\]
Example 2.6: Let us consider a population of four values 1, 2, 3, and 4.

Then the population mean, \( \mu = \frac{1 + 2 + 3 + 4}{4} = 2.5 \).

There are sixteen possible samples of size 2 with replacement

\{1,1\}, \{1,2\}, \{1,3\}, \{1,4\}, \{2,1\}, \{2,2\}, \{2,3\}, \{2,4\}, \{3,1\}, \{3,2\}, \{3,3\}, \{3,4\}, \{4,1\}, \{4,2\}, \{4,3\}

and \{4,4\}

Corresponding sample means are 1, 1.5, 2, 2.5, 1.5, 2, 2.5, 3, 2, 2.5, 3, 3.5, 2.5, 3, 3.5 and 4.

The mean of all sixteen means is:

\[ \frac{1 + 1.5 + 2 + 2.5 + 1.5 + 2 + 2.5 + 3 + 2 + 2.5 + 3 + 3.5 + 2.5 + 3 + 3.5 + 4}{16} = 2.5. \]
Mean of Grouped Frequency Data

The mean in case of grouped data is found using the following formula:

$$\bar{x} = \frac{\sum f \cdot x_m}{n}$$

where, $x_m$ is the class midpoint.
Example 2.7: Find the sample mean using the data in table 2.4.

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency f</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-10</td>
<td>1</td>
</tr>
<tr>
<td>11-15</td>
<td>2</td>
</tr>
<tr>
<td>16-20</td>
<td>3</td>
</tr>
<tr>
<td>21-25</td>
<td>5</td>
</tr>
<tr>
<td>26-30</td>
<td>4</td>
</tr>
<tr>
<td>31-35</td>
<td>3</td>
</tr>
<tr>
<td>36-40</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2.4: A Grouped Data

First we need to find the midpoints of each class where \( x_m = \frac{Lower \ Limit + Upper \ Limit}{2} \).
Form a column \((f \cdot x_m)\) the calculation is shown table 2.5:

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency ((f))</th>
<th>Midpoint ((x_m))</th>
<th>(f \cdot x_m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-10</td>
<td>1</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>11-15</td>
<td>2</td>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td>16-20</td>
<td>3</td>
<td>18</td>
<td>54</td>
</tr>
<tr>
<td>21-25</td>
<td>5</td>
<td>23</td>
<td>115</td>
</tr>
<tr>
<td>26-30</td>
<td>4</td>
<td>28</td>
<td>112</td>
</tr>
<tr>
<td>31-35</td>
<td>3</td>
<td>33</td>
<td>99</td>
</tr>
<tr>
<td>36-40</td>
<td>2</td>
<td>38</td>
<td>76</td>
</tr>
<tr>
<td>(\sum f = n = 20)</td>
<td></td>
<td>(\sum (f \cdot x_m) = 490)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.5: Grouped Mean Calculation

\[
\bar{x} = \frac{\sum f \cdot x_m}{n} = \frac{490}{20} = 24.5
\]
Median

Median can be for a population or for a sample. Due to similarity in calculation, we will discuss only for samples. Median $M$ is the midpoint of the distribution meaning; half of the observations are smaller than the median and the other half are larger than the median.

It is not affected by an extreme value - is a resistant measure of the center
Finding the Median

1. Arrange the observations in order, from the smallest to the largest.

2. Find \( \frac{n+1}{2} \) which is the position of the median where \( n \) is the number of observation in the data. Then locate the \( \frac{n+1}{2}th \) positioned measurement in the ordered list which is the median.

3. If the number of observations \( n \) is odd, the median \( M \) is the central observation in the ordered list. If the number of the observations \( n \) is even, the median \( M \) is the mean of the two central observations in the ordered list.
Example 2.8: A random sample of National Basketball Association (NBA) players’ heights (in feet) is: 6.52, 6.39, 6.78, 7.12, 6.23, 6.68, 6.94.

To find the median height count the number of observations in the sample n=7.

Ordered sample is: 6.23, 6.39, 6.52, 6.68, 6.78, 6.94, 7.12.

The location of the median $M$ is $\frac{n+1}{2} = \frac{7+1}{2} = 4$.

The median of this data set is 6.68, the fourth value of the ordered data set from the smallest to the largest.
Example 2.9: Let a random sample of 8 NBA players’ heights (in feet) be 6.52, 6.39, 6.78, 7.12, 6.23, 6.68, 6.88, 6.94.

Ordered sample is: 6.23, 6.39, 6.52, 6.68, 6.78, 6.88, 6.94, 7.12.

The location of the median $M$ is $\frac{n+1}{2} = \frac{8+1}{2} = 4.5$.

The median of this data set is 6.73, the average of the fourth value 6.68 and the fifth value 6.78 of the ordered data set from the smallest to the largest.
Mode

Mode is another measure of center. Mode(s) is (are) the most frequent value(s) in the data set. Mode is also called the most typical value in the data set. Data set can have one, more than one, or no mode.

Mode is the only measure of central tendency that can be used with nominal data.
Bimodal  two data values occur with the same greatest frequency.

Multimodal  more than two data values occur with the same greatest frequency.

No Mode  no data value is repeated.
Mode - Examples

a. 5.40 1.10 0.42 0.73 0.48 1.10
   - Mode is 1.10

b. 27 27 27 55 55 55 88 88 99
   - Bimodal - 27 & 55

c. 1 2 3 6 7 8 9 10
   - No Mode
Definition

**Midrange** is a rough estimate of the center of the data. It is computed by adding the lowest and the highest values and dividing it by 2,

\[ \text{Midrange} = \frac{\text{lowest value} + \text{highest value}}{2} \]
Example 2.13: According to the consumer reports, the prices per ounce in cents of the barbecue-flavored potato chips in a sample of 6 brands are

\[
\begin{align*}
&19 & 19 & 27 & 28 & 18 & 35 \\
\text{Midrange} &= \left( \frac{18 + 35}{2} \right) = 26.5 \text{ cents.}
\end{align*}
\]
Relation between Mean, Median and mode

The mean and median cannot always be used to identify the shape of the distribution.

When the distribution is negatively skewed, \( \text{Mean} < \text{Median} < \text{Mode} \).

When the distribution is positively skewed, \( \text{Mode} < \text{Median} < \text{Mean} \).

When the distribution is symmetric or bell shaped, \( \text{Mean} = \text{Median} = \text{Mode} \).

Figure 2.8 Examples of Symmetric and Skewed Distributions
Measures of Variation
Measures of central tendency give us measures of where the center of a set of data is, but this is not sufficient to characterize the data.

Example 2.14: Consider the following two data sets:

Data set A: 50, 60, 70, 80, 90

Data set B: 69, 69, 70, 71, 71

Both these data sets have a mean of 70 yet the first data set is more widely dispersed than the second data set. So a measure of variation is needed to characterize data sets.

Selected measures of variation, such as, range, mean absolute deviation, variance, and standard deviation are discussed below.
Range is the simplest measure of variation. The range (R) is the difference between the maximum value and the minimum value of the measurements.

\[ R = \text{Maximum} - \text{Minimum}. \]

Range for data set \( A \) is 40.

Range for data set \( B \) is 2.

Data set A has more variability. Range is a simple measure of variability, but it uses only 2 values from the data set and ignores all other values.
Mean Absolute Deviation (MAD)

Mean of absolute deviations from the mean.

Population mean absolute deviation is computed as \( \frac{\sum |x - \mu|}{N} \).

Sample mean absolute deviation is computed as \( \frac{\sum |x - \bar{x}|}{n} \).
Example 2.15: A random sample of 7 National Basketball Association (NBA) players’ heights are 6.52, 6.39, 6.78, 7.12, 6.23, 6.68, and 6.94 feet.

Sample mean, $\bar{x} = \frac{\sum x}{n} = \frac{6.52 + 6.39 + 6.78 + 7.12 + 6.23 + 6.68 + 6.94}{7} = \frac{46.66}{7} = 6.67$.

Mean absolute deviation, $MAD = \frac{\sum |x - \bar{x}|}{n} = \frac{1}{7} \left( |6.52 - 6.67| + |6.39 - 6.67| + |6.78 - 6.67| + |7.12 - 6.67| + |6.23 - 6.67| + |6.68 - 6.67| + |6.94 - 6.67| \right) = \frac{0.15 + 0.28 + 0.11 + 0.45 + 0.44 + 0.01 + 0.27}{7} = \frac{1.71}{7} = 0.2443$. 
Variance

The variance is the measure of the spread of data around the mean. Formulas for variance differ slightly depending on whether we are using a sample or the entire population.

The population variance is denoted by $\sigma^2$ (lower case sigma square),

$$
\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N} = \frac{\sum (x - \mu)^2}{N}
$$

A sample variance is denoted by $s^2$ and computed as:

$$
s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1} = \frac{\sum (x - \bar{x})^2}{n - 1}
$$

Computational formula for sample variance $s^2$ is

$$
s^2 = \frac{n \sum x^2 - (\sum x)^2}{n(n - 1)}
$$
Standard Deviation

Standard deviation is the square root of the variance.

The population standard deviation, \( \sigma = \sqrt{\sigma^2} \).

The sample standard deviation:

\[
s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{n\sum x^2 - (\sum x)^2}{n(n-1)}}.
\]

The variance is easier to comprehend but the standard deviation has more direct use through \textit{Z-score, Empirical rule, Chebyshev’s inequality}, and so on to name a few.
Example 2.16: Amount of total snow fall in Minnesota in 2011 is estimated from a sample of 10 locations as given below.

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>INCHES</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDOVER</td>
<td>19.0</td>
</tr>
<tr>
<td>MINNETONKA</td>
<td>18.0</td>
</tr>
<tr>
<td>CHASKA</td>
<td>17.5</td>
</tr>
<tr>
<td>HUTCHINSON</td>
<td>16.5</td>
</tr>
<tr>
<td>PLYMOUTH</td>
<td>17.0</td>
</tr>
<tr>
<td>DEEPHAVEN</td>
<td>16.0</td>
</tr>
<tr>
<td>HILLTOP</td>
<td>14.5</td>
</tr>
<tr>
<td>CHANHASSEN</td>
<td>14.0</td>
</tr>
<tr>
<td>WACONIA</td>
<td>14.0</td>
</tr>
<tr>
<td>BUFFALO</td>
<td>13.5</td>
</tr>
</tbody>
</table>

Find the sample variance and the sample standard deviation using both main and shortcut formulas.
Using the main formula, \[ s^2 = \frac{\sum(X - \bar{X})^2}{n - 1} \]

First find the sample mean \[ \bar{X} = \frac{\sum X}{n} = \frac{19.0 + 18.0 + 17.5 + \cdots + 13.5}{10} = \frac{160.0}{10} = 16.0 \]

The table below shows the calculation of the sample variance

<table>
<thead>
<tr>
<th>X</th>
<th>X - \bar{X}</th>
<th>(X - \bar{X})^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.0</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>18.0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>17.5</td>
<td>1.5</td>
<td>2.25</td>
</tr>
<tr>
<td>16.5</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>17.0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>16.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14.5</td>
<td>-1.5</td>
<td>2.25</td>
</tr>
<tr>
<td>14.0</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>14.0</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>13.5</td>
<td>-2.5</td>
<td>6.25</td>
</tr>
</tbody>
</table>

\[ \sum X = 160 \quad \sum (X - \bar{X}) = 0 \quad \sum (X - \bar{X})^2 = 33 \]

Table 2.6: Computations for Variance
Variance, \( s^2 = \frac{33}{9} = 3.667 \)

Standard deviation, \( s = \sqrt{3.667} = 1.915 \)

Note that the sum of deviation from the mean is always zero. i.e. \( \sum (x - \bar{x}) = 0 \)

Let's calculate the sample variance using shortcut formula from the above

<table>
<thead>
<tr>
<th>( x )</th>
<th>( x^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.0</td>
<td>361</td>
</tr>
<tr>
<td>18.0</td>
<td>324</td>
</tr>
<tr>
<td>17.5</td>
<td>306.25</td>
</tr>
<tr>
<td>16.5</td>
<td>272.25</td>
</tr>
<tr>
<td>17.0</td>
<td>289</td>
</tr>
<tr>
<td>16.0</td>
<td>256</td>
</tr>
<tr>
<td>14.5</td>
<td>210.25</td>
</tr>
<tr>
<td>14.0</td>
<td>196</td>
</tr>
<tr>
<td>14.0</td>
<td>196</td>
</tr>
<tr>
<td>13.5</td>
<td>182.25</td>
</tr>
</tbody>
</table>

| \( \sum x = 160 \) | \( \sum x^2 = 2593 \) |

Table 2.7: Variance Using Shortcut Formula
Sample variance using the shortcut formula,

\[
s^2 = \frac{n\sum x^2 - (\sum x)^2}{n(n-1)} = \frac{10*2593 - (160)^2}{10*9} = 3.667,
\]

and the sample standard deviation is 1.915. The sample variance and the sample standard deviation both are same as calculated by the original formula but the shortcut formula uses less computational steps.
Coefficient of Variation

In order to compare the variation in two or more different distributions when the unit of measurement is not the same, we use coefficient of variation (CV). Coefficient of variation expresses the standard deviation as a percentage of the sample mean.

\[
CV = \frac{\text{standard deviation}}{\text{mean}} \times 100\%.
\]
Example 2.17: The trading post on Grand Mesa is a small, family-run store in Colorado. Grand Mesa region contains many good fishing lakes, so the Trading post sells spinners (a type of fishing lure). The store has eight different types of spinners. The price (in dollars) are:

2.10   1.95   2.60   2.00   1.85   2.25   2.15   2.25

We can calculate mean $\bar{x} = 2.14$ and standard deviation $s = 0.22$.

The coefficient of variation $CV = \frac{s}{\bar{x}} \times 100\% = \frac{0.22}{2.14} \times 100\% = 10.28\%$.

This means the standard deviation of the spinner prices is only 10.28% of the mean.

For a population data use the formula $CV = \frac{\sigma}{\mu} \times 100\%$. 
Chebyshev’s Theorem

Chebyshev’s Inequality or Theorem

For a population, Chebyshev’s inequality applies to any distribution, which states that for \( k > 1 \), at least \( \left(1 - \frac{1}{k^2}\right) \times 100\% \) measurements will fall between \( \mu - k\sigma \) and \( \mu + k\sigma \) or at least \( \left(1 - \frac{1}{k^2}\right) \times 100\% \) of the measurements will fall within the \( k \) standard deviation of the mean.

For a sample, Chebyshev’s inequality applies to any data, which states that for \( k > 1 \), at least \( \left(1 - \frac{1}{k^2}\right) \times 100\% \) measurements will fall between \( \bar{X} - ks \) and \( \bar{X} + ks \) or at least \( \left(1 - \frac{1}{k^2}\right) \times 100\% \) of the measurements will fall within the \( k \) standard deviation of the mean.

Letting \( k = 2 \), we get at least three fourth or 75% of the data values will fall within 2 standard deviation of the mean of the data set. Note that \( k \) is not necessarily an integer.
Example 2.18: Let the mean price of houses in a certain neighborhood be $50,000, and the standard deviation be $10,000. Find the price range for which at least 75% of the houses will sell.

Using the Chebyshev’s theorem, 75% of the data values will fall within 2 standard deviation of the mean, that is, in the interval $(\bar{X} - 2s, \bar{X} + 2s)$.

Here, $\bar{X} - 2s = 50000 - 2(10000) = 30000$ and $\bar{X} + 2s = 50000 + 2(10000) = 70000$.

Hence, at least 75% of the houses will be sold in the area will have a price range from $30,000 to $70,000.
The Empirical Rule

Empirical Rule

When a distribution is symmetric and bell shaped, approximately 68% of the data values will fall within 1 standard deviation of the mean or within \((\bar{x} - s, \bar{x} + s)\), approximately 95% of the data values will fall within 2 standard deviation of the mean or within \((\bar{x} - 2s, \bar{x} + 2s)\) and approximately 99.7% (or almost all) of the data values will fall within 3 standard deviation of the mean or within \((\bar{x} - 3s, \bar{x} + 3s)\).

Note that for a population measurements the corresponding intervals are \((\mu - \sigma, \mu + \sigma)\), \((\mu - 2\sigma, \mu + 2\sigma)\), and \((\mu - 3\sigma, \mu + 3\sigma)\), respectively.
Figure 2.9 Empirical Rule

- $\mu - 3\sigma$: 99.74%
- $\mu - 2\sigma$: 95.44%
- $\mu - \sigma$: 68.26%
- $\mu + \sigma$: 68.26%
- $\mu + 2\sigma$: 95.44%
- $\mu + 3\sigma$: 99.74%
Example 2.19: A random sample of 7 National Basketball Association (NBA) players’ heights (in feet) are 6.52, 6.39, 6.78, 7.12, 6.23, 6.68, 6.94.

Sample mean, $\bar{x} = \frac{\sum x}{n} = \frac{6.52 + 6.39 + 6.78 + 7.12 + 6.23 + 6.68 + 6.94}{7} = \frac{46.66}{7} = 6.67$.

Sample variance, $s^2 = \frac{n\sum x^2 - (\sum x)^2}{n(n-1)}$, where $\sum X^2 = (6.52)^2 + (6.39)^2 + \cdots + (6.94)^2$

$= 311.6042$. Then $s^2 = \frac{7(311.6042) - (46.66)^2}{7(7-1)} = 0.0970$.

Sample standard deviation, $s = \sqrt{s^2} = \sqrt{0.0970} = 0.3114$.

In the application of the empirical rule, assuming that the distribution is bell shaped, approximately 68% of the measurements for the respective population are between $6.67 - 0.3114 = 6.3586$ and $6.67 + 0.3114 = 6.9814$ i.e. between (6.3586, 6.9814) which is the one standard deviation of the mean interval, approximately 95% of the measurements are between $6.67 - 2(0.3114) = 6.0472$ and $6.67 + 2(0.3114) = 7.2928$, i.e. between (6.0472, 7.2928) which is the two standard deviation of the mean interval and almost all measurements or 99.7% are between $6.67 - 3(0.3114) = 5.7358$ and $6.67 + 3(0.3114) = 7.6042$, i.e. between (5.7358, 7.6042) which is the three standard deviation of the mean interval.
Example 2.20: IQ scores have a bell-shaped distribution with a mean of 100 and a standard deviation of 15. What percentage of scores are

(a) between 70 and 130?

\[ \mu - 2\sigma = 100 - 2 \times 15 = 70 \quad \text{and} \quad \mu + 2\sigma = 100 + 2 \times 15 = 130. \]
The interval (70, 130) is the two standard deviation from the mean interval. So that approximately, 95% of the scores falls between (70, 130).

(b) Greater than 115?

Since \( \mu + \sigma = 100 + 15 = 115 \), we know that within one standard deviation there are 68% of the scores, hence there are 32% scores are outside, and 16% of the scores are greater than 115.
Z - Score

A standardized score is also called a *z-score* for an observation is obtained by subtracting the mean from that observation and then dividing the result by the standard deviation.

The population *z-score*, \( z = \frac{x - \mu}{\sigma} \).

The sample *z-score* is computed as, \( z = \frac{x - \bar{x}}{s} \).

Note:
1. *z-score* measures the number of standard deviations that a data value falls above or below the mean.
2. If the *z-score* is positive, the actual value is greater than the mean.
3. If the *z-score* is negative, the actual value is lower than the mean.
4. If the *z-score* is zero, the actual value is same as the mean.
Example 2.21: Human body temperature has a mean of $98.2^\circ F$ and a standard deviation of $0.62^\circ F$. Convert $102.5^\circ F$ to $z$-score.

$$z - \text{score}, \quad Z = \frac{X - \overline{X}}{s} = \frac{102.5 - 98.2}{0.62} = 6.935.$$ 

Example 2.22: A student scored 85 on a test where the standard deviation was 3. The student’s $z$-score was 2.33. Find the test average.

If $Z = \frac{X - \overline{X}}{s}$, then $\overline{X} = X - Z * s = 85 - 3 * 2.33 = 78.01.$
Measures of Position
The measures are due to the position of the measurement relative to the overall distribution.

**Quantiles**

In general, the measures of positions are known as quantiles. Here, we will discuss several quantiles, such as, quartiles, deciles, and percentiles.

**Quartiles**

Quartiles are the three partitioning measurements when the overall distribution is divided into four equal proportions. The three quartiles are denoted as $Q_1$, $Q_2$, and $Q_3$.

$Q_1$ (the first quartile) is the value for which 25% measurements are lower or equal.

$Q_2$ (the second quartile or the median) is the value for which 50% measurements are lower or equal.

$Q_3$ (the third quartile) is the value for which 75% measurements are lower or equal.
In order to calculate the quartiles we can first compute the median of the whole data which is the second quartile (Q₂).

Q₁ is the median of the data set below Q₂ (median of the lower half data values) and Q₃ is the median of the data set above Q₂ (median of the upper half data values).

Example 2.23: Find Q₁, Q₂ and Q₃ of the data: 15, 13, 6, 5, 12, 50, 22, 18

Step 1. Ordered data: 5, 6, 12, 13, 15, 18, 22, 50

Step 2. Q₂ = Median of the whole data = \( \frac{(n+1)}{2} \)th value = 4.5th value = \( \frac{13+15}{2} \) = 14

Step 3. Q₁ = Median of the lower half data = 9

Step 4. Q₃ = Median of the upper half data = 20
Deciles

Deciles are the nine partitioning measurements when the overall distribution is divided into ten equal proportions. The deciles are denoted as $D_1, D_2, \ldots, D_9$. For example, $D_2$ (the second decile) is the value for which 20% measurements are lower or equal.

$D_5$ is the same as $Q_2$ (median) of the distribution.

Percentiles

The partitioning measurements when the overall distribution is divided into 100 equal proportions. For example, $P_{17}$ (the seventeenth percentile) is the value for which 17% measurements are lower or equal.

The computation of the quantiles can be generalized as follows:

Step 1: Order the data from the smallest to the largest.

Step 2: Compute the locator, $L = (n + 1)p$, where $n$ is the sample size and $p$ is the proportion corresponding to the percentage of interest.

Step 3: Locate the $L^{th}$ measurement in the ordered data, interpolate whenever $L$ is not an integer.
Note that slight variations in methods are used by different authors but the method introduced here will produce the most suitable measures of positions.

Example 2.24: A random sample of 7 National Basketball Association (NBA) players’ heights (in feet) are 6.52, 6.39, 6.78, 7.12, 6.23, 6.68, 6.94. Find the 25th percentile.

Step 1: Ordered data: 6.23, 6.39, 6.52, 6.68, 6.78, 6.94, 7.12.

Step 2: \( L = (n + 1)p = (7 + 1)0.25 = 2 \)nd value.

Step 3: \( P_{25} = Q_1 = 6.39 \).

Example 2.25: Find the second decile (D2) from the NBA data given in example 2.24.

Step 1: Ordered data: 6.23, 6.39, 6.52, 6.68, 6.78, 6.94, 7.12.

Step 2: \( L = (n + 1)p = (7 + 1)0.2 = 1.6 \).

Step 3: \( D_2 = 6.23 + (6.39 - 6.23)0.6 = 6.326 \).
Box plot uses measures of positions to determine the shape of the distribution and to identify the outlier(s).

A box is plotted either horizontally or vertically.

Box plots are based on the five number summary: minimum (the lowest value of the data set), \( Q_1 \), Median, \( Q_3 \) and maximum (the highest value of the data set).

The box spans the quartiles and the lines are extended from the box to the extremes and show the full spread of the data. The lower (or left) end of the box corresponds to the first quartile, the upper (or right) end of the box corresponds to the third quartile, and the median lies in between which is marked by a line.
An outlier is an observation that lies outside the overall pattern of a distribution. Outliers should be investigated carefully. Often they contain valuable information about the process under investigation or the data gathering and recording process.

All the values outside the interval \((Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)\) are considered outliers. Where \(IQR = \text{Interquartile Range} = Q_3 - Q_1\).

Often, \(Q_1 - 1.5 \times IQR\) is known as Lower Fence (LF) and \(Q_3 + 1.5 \times IQR\) as Upper Fence (UF).

That means all the values less than \((Q_1 - 1.5 \cdot IQR)\) and greater than \((Q_3 + 1.5 \cdot IQR)\) are considered outliers.

Lines extended from the box to the smallest and largest values that are not suspected outliers.
Example 2.27: To construct the box-plot for the NBA players’ heights data, Median = 6.68, \( Q_1 = 6.39 \), in computing \( Q_3 \), \( L = (n+1)p = (7+1)0.75 = 6 \), and \( Q_3 = 6.94 \). IQR = Inter-quartile Range = \( Q_3 - Q_1 = 6.94 - 6.39 = 0.55 \). \( LF = Q_1 - 1.5 \times IQR = 6.39 - 1.5(0.55) = 5.565 \) and \( UF = Q_3 + 1.5 \times IQR = 6.94 + 1.5(0.55) = 7.765 \). There is no data value below 5.565 and above 7.765. So, there is no outlier in the data.

![Figure 2.10 Boxplot for Example 2.27](image-url)
Example 2.28: Check for outliers in the data set: 5, 13, 6, 12, 18, 15, 50, 22

Step 1. Ordered data: 5, 6, 12, 13, 15, 18, 22, 50

Step 2. First we find $Q_1$ and $Q_3$. In this case, $Q_1=9$ and $Q_3=20$.

Step 3. Find $IQR = Q_3 - Q_1$. In this case, $IQR = 11$.

\[(Q_1 - 1.5 \cdot IQR) = -7.5\]

\[(Q_3 + 1.5 \cdot IQR) = 36.5\]

Hence only one value outside $(-7.5, 36.5)$ is 50.

Figure 2.11 Boxplot for Example 2.28