Chapter Two: Descriptive Methods
We previously said that descriptive statistics is made up of various techniques used to summarize the information contained in a set of data. The following descriptive techniques will comprise the major topics covered in this chapter.

- distributional
- graphical
- computational

Before broaching these topics, however, we will discuss the prerequisite topics of scales of measurement and summation notation.
The four scales of measurement as promulgated by S.S. Stevens are as follows.

1. nominal
2. ordinal
3. (equal) interval
4. ratio
The following are some characteristics of the nominal scale.

- Least sophisticated (informative) of the four scales.
- Classifies persons or things on the basis of the characteristic being assessed.
- No information regarding quantity or amount is imparted by this scale.
- Characterizes persons or things as being similar or dissimilar insofar as the characteristic being assessed is concerned.
- Has no ability to make “greater than” or “less than distinctions.
- Examples include blood typing, gender and area codes.
The following are some characteristics of the ordinal scale.

- Classifications based on this scale incorporate the attributes of “greater than” and “less than”.
- Cannot provide information as to *how much* less or more one category represents than another. That is, the difference between 10 and 20 may be quite different from the difference between 30 and 40.
- Examples include most ranking procedures as well as assessments on scales with categories such as none, mild, moderate and severe.
The following are some characteristics of the interval scale.

- Provides information as to *how much* less or more one category represents than another.

- Scale points with equal differences represent equal differences in the characteristics being assessed. For example, insofar as the characteristic being measured is concerned, the difference between 10 and 20 on this scale is the same as the difference between 30 and 40.

- Has an arbitrary zero point.

- Examples include the Fahrenheit thermometer and the Julian calendar.
The following are some characteristics of the ratio scale.

- Identical to the interval scale except that the ratio scale has a true zero point.
- Examples include commonly used measures of weight and height.
A simpler view of data holds that they are either continuous or discrete.

- A **continuous** variable is one that, at least theoretically, can take on any value in some specified range. Examples include weight, blood pressure and temperature.

- A **discrete** variable is one that is not continuous. Examples include most count data such as the numbers of persons living in households.

- Discrete variables that can take only one of two values, e.g., male or female, dead or alive, are termed **dichotomous** variables.
**Summation notation** is the notation used to show exactly how data are to be summed.
Suppose five numbers are written down in arbitrary order. Call the first number $x_1$, the second $x_2$ and so forth. If we wanted to indicate that these numbers are to be summed we could write the instruction

$$x_1 + x_2 + x_3 + x_4 + x_5.$$
A shorter form of this notation can be written as

$$\sum_{i=1}^{5} x_i.$$ 

The notation $\sum x$ indicates that the $x$ values are to be summed while the $i$ subscript on the $x$ acts as a place holder for the numbers 1 through 5. The notation $i = 1$ shows that the summation is to begin with $x_1$ while the 5 indicates that the summation is to end with $x_5$. In other words, all the numbers in the set are to be summed.
When all of the numbers in a set of arbitrary size \((n)\) are to be summed we use the notation

\[
\sum_{i=1}^{n} x_i
\]

Thus, if we wish to indicate that the squares of a set of numbers of arbitrary size are to be summed \(w\) would use the notation

\[
\sum_{i=1}^{n} x_i^2.
\]
Some Rules of Summation

1. \( \sum_{i=1}^{n} c = nc \)

2. \( \sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i \)

3. \( \sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \)

4. \( \sum_{i=1}^{n} (x_i - y_i) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \)
An Example Application of Some Summation Rules

\[ \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} \]

rule 4

\[ = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} \]

rule 1

\[ = \sum_{i=1}^{n} x_i - \frac{n \bar{x}}{} \]

\[ = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i \]

\[ = 0 \]
A frequency distribution shows the number of responses of each type.

A relative frequency distribution shows the proportion of responses of each type.

A cumulative frequency distribution shows the number of responses that are less than or equal to specified values.

A cumulative relative frequency distribution shows the proportion of responses that are less than or equal to specified values.
# Distributions of Perceived Pain Measures

<table>
<thead>
<tr>
<th>Pain Category</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Cumulative Frequency</th>
<th>Cumulative Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe</td>
<td>4</td>
<td>.07</td>
<td>60</td>
<td>1.00</td>
</tr>
<tr>
<td>Moderate</td>
<td>8</td>
<td>.13</td>
<td>56</td>
<td>.93</td>
</tr>
<tr>
<td>Mild</td>
<td>17</td>
<td>.28</td>
<td>48</td>
<td>.80</td>
</tr>
<tr>
<td>None</td>
<td>31</td>
<td>.52</td>
<td>31</td>
<td>.52</td>
</tr>
</tbody>
</table>
It is sometimes more informative to arrange data into groups or intervals of values rather than dealing with individual values. In constructing tables of this sort two related questions must be answered.

- Into how many intervals should the values be grouped?
- How long should the intervals be?

There are no hard and fast answers to these questions but a rule of thumb suggests that there be no fewer than six and no more than 15 intervals. Another helpful suggestion is that, when plausible, class interval widths of 5 units, 10 units or some multiple of 10 should be used in order to make the summarization more comprehensible.
### Example Grouped Distribution

Table: Grouped distributions of systolic blood pressures using eight intervals.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Cumulative Frequency</th>
<th>Cumulative Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>142-149</td>
<td>2</td>
<td>.014</td>
<td>144</td>
<td>1.000</td>
</tr>
<tr>
<td>134-141</td>
<td>40</td>
<td>.278</td>
<td>142</td>
<td>.986</td>
</tr>
<tr>
<td>126-133</td>
<td>36</td>
<td>.250</td>
<td>102</td>
<td>.708</td>
</tr>
<tr>
<td>118-125</td>
<td>21</td>
<td>.146</td>
<td>66</td>
<td>.458</td>
</tr>
<tr>
<td>110-117</td>
<td>18</td>
<td>.125</td>
<td>45</td>
<td>.313</td>
</tr>
<tr>
<td>102-109</td>
<td>8</td>
<td>.056</td>
<td>27</td>
<td>.188</td>
</tr>
<tr>
<td>94-101</td>
<td>14</td>
<td>.097</td>
<td>19</td>
<td>.132</td>
</tr>
<tr>
<td>86-93</td>
<td>5</td>
<td>.035</td>
<td>5</td>
<td>.035</td>
</tr>
</tbody>
</table>
It is often more informative to present distributions as graphs rather than in the tabular form shown previously. Many graphical forms are available. Here you will learn about the bar graph, histogram, polygon, and stem-and-leaf display.
Bar graphs can be constructed for frequency, relative frequency, cumulative frequency, and cumulative relative frequency distributions.

Response categories are depicted along the horizontal (x) axis.

Frequencies, relative frequencies, cumulative frequencies or cumulative relative frequencies are noted along the vertical (y) axis.

The frequency, relative frequency, cumulative frequency, or cumulative relative frequency is read as the height, measured against the y axis, of a bar placed above the specified category.

Normally used with discrete rather than continuous data.
Figure: Relative frequency bar graph for pain scores.
Characteristics of Histograms

- Same general characteristics as bar graphs but are used with continuous data while bar graphs are used with discrete data.
- Bars used in histograms are contiguous while those used in bar graphs are not.
- Because histograms are used with continuous data, it is usually the upper and lower real limits of the intervals that are labeled on the x axis rather than the midpoint.
Figure: Relative frequency histogram of blood pressure measures in 12 categories.
Characteristics of Polygons

- Polygons can be constructed for frequency, relative frequency, cumulative frequency, and cumulative relative frequency distributions.
- Polygons are constructed in a manner similar to histograms except that instead of placing a bar over each interval, a dot is placed at a height appropriate to the y axis.
- In the case of frequency and relative frequency polygons the dot is placed at the midpoint of the interval while for cumulative distributions the dot is placed at the upper real limit of the interval.
These dots are then connected with straight lines that are joined to the $x$ axis at the lower and, in the cases of frequency and relative frequency polygons, upper ends of the distribution. For frequency and relative frequency polygons the points at which the line is brought down to the $x$ axis are at what would be the midpoints of an additional interval at each end of the distribution.

Cumulative frequency and cumulative relative frequency polygons are not connected to the baseline at the upper end of the distribution and are connected at the lower end at the upper real limit of an additional interval added to the lower end of the distribution.
Polygons are particularly convenient when one wishes to compare two or more distributions as when male and female blood pressures are to be plotted in the same graph.
Figure: Relative frequency polygon of blood pressure measures in 12 categories.
Figure: Relative frequency polygon imposed on histogram of blood pressure measures in 12 categories.
Figure: Cumulative relative frequency polygon of blood pressure measures in 12 categories.
Characteristics of Stem-and-leaf Displays

- Similar to frequency histogram.
- Primary advantage over the histogram is that it preserves the values of the displayed variable.
Figure: Stem-and-leaf display of blood pressure measures.
Construction of Stem-and-leaf Displays

1. Divide each observation into a “stem” and “leaf” component as described below.

2. List the stem components from lowest to highest values as would be done on the x axis of a histogram.

3. Place the leaf components associated with each stem over the stem in ascending order.

The **stem** of a number is defined as all of the digits in the number except for the right most. The **leaf** is then the right-most digit. Thus, the stem for the blood pressure value of 86 is 8 and the leaf is 6.
It is often desirable to describe some particular characteristic of a data set numerically. Perhaps the most familiar measure of this sort is what is commonly referred to as the “average” or more precisely, the *arithmetic mean* of a set of data. We now examine four distinct categories of such measures. They are, measures of central tendency, measures of variability, measures of relative standing, and measures associated with distribution shape.
Measures of central tendency provide information about typical or average values of a data set. There are many such measures but we will consider only the mean, median, and mode as these are most commonly used.
The arithmetic mean is the best known of the measures of central tendency and is what most people refer to as the “average.” It is calculated by summing all the observations in the set of data and dividing this sum by the number of observations. The modifier “arithmetic” is sometimes used to distinguish it from other, less familiar, means.
The Sample Mean

\[ \bar{x} = \frac{\sum x}{n} \]

Where \( x \) represents the values whose mean is to be calculated and \( n \) represents the number of observations in the sample.
The Population Mean

\[ \mu = \frac{\sum x}{N} \]

Where \( x \) represents the values whose mean is to be calculated and \( N \) represents the number of observations in the population.
Among the many properties of the mean are the following.

- It is unambiguously defined in that its method of calculation is generally recognized.
- It is unique in that a data set has one and only one mean.
- Its value is influenced by all observations in the data set.
Example

Find the mean of the numbers 3, 5, 4, 8, 7.

Solution

\[ \bar{x} = \frac{3 + 5 + 4 + 8 + 7}{5} = 5.4 \]
There are a number of different ways to define and calculate the median. By far the most common definition maintains that the median is the value that divides a dataset into two equal parts so that the number of values that are greater than or equal to the median is equal to the number of values that are less than or equal to the median. A less known definition of the median states that it is a point on the scale of measurement located such that half the observations are above and below the point. Thus, one definition is value based while the other is scale based.
The most common way of computing the median when the number of values is odd is to order the observations in terms of magnitude, then choose the middle value as the median. Thus, to find the median of the numbers 3, 5, 4, 8, 7, 0, 12 we arrange the values as

\[
\begin{array}{c}
\overbrace{0}^{3 \text{ values}} & \overbrace{3}^{\text{median}} & \overbrace{4}^{3 \text{ values}} & \overbrace{5} & \overbrace{7} & \overbrace{8} & \overbrace{12}
\end{array}
\]

with the middle number, 5, is chosen as the median.
A more formal way of expressing this is given by

\[
\text{Median (} n \text{ odd)} = x_{\frac{n+1}{2}}
\]

where \( n \) is the number of observations and \( \frac{n+1}{2} \) is the subscript of \( x \).
When the number of observations in the data set is even, there is no middle value to choose as the median. In this case, the median is computed as the mean of the two middle values. Thus, to find the median of the numbers 14, 8, 3, −1, 0, 12, 12, and 11 we arrange the values as

\[
\begin{align*}
-1 & \ 0 \ 3 \ 8 \ 11 \ 12 \ 12 \ 14 \\
\end{align*}
\]

and obtain \( \frac{8+11}{2} = 9.5 \) as the median.
A more formal way of expressing this is given by

\[
\text{Median (n even)} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}
\]

where \( \frac{n}{2} \) and \( \frac{n}{2} + 1 \) are the subscripts identifying the two middle values.
A scale based method of calculating the median is given by

\[
\text{Median} = LRL + (w) \left[ \frac{(0.5)(n) - cf}{f} \right]
\]

where \( LRL \) is the lower real limit of the median interval, \( w \) is the width of the median interval calculated as the difference between the upper and lower real limits of that interval, \( n \) is the total number of observations, \( cf \) is the cumulative frequency \textit{up to} the median interval and \( f \) is the frequency of the median interval.

Two special cases occur when (a) the median falls on the upper (lower) real limit of an interval and, (b) when the median interval spans more than one scale interval.
Example

Applying the scale based method of finding the median to the values 0, 1, 1, 2, 2, 2, 2, 3 yields

\[ 1.5 + (1.0) \frac{(0.5)(8) - 3}{4} = 1.75 \]
When computing the scale based median, three different scenarios are possible.

- When a median interval is identified, Formula 2.5 is applied.
- When half the observations fall below and half above a real limit with the interval above the limit having nonzero frequency, the real limit is taken as the median.
- When half the observations fall below and half above a real limit with the interval above the limit having zero frequency, the midpoint of the zero frequency interval(s) is taken as the median.
Among the properties of the median are the following.

- It may be defined and calculated in a number of different ways.
- Given a specific definition and manner of calculation, it is unique in that a data set has one and only one median.
- It is insensitive to extreme observations.
The mode of a data set is the score or scores in the set that occur most frequently. If all scores in the set occur with equal frequency there is no mode. If two or more scores occur with equal frequency and that frequency is greater than that of the other scores in the set, then there will be more than one mode.

In the case of nominal or ordinal level data, the modal category can be found. The modal category is the one that has the greatest frequency. If two or more categories have equal frequencies and that frequency is greater than that of all other categories, then there will be more than one modal category.
The modes of the following data set are 9 and 7.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
</tr>
</tbody>
</table>