CHAPTER 8
INTRODUCTION TO STATISTICAL ANALYSIS

LEARNING OBJECTIVES:

After studying this chapter, a student should understand:

- notation used in statistics;
- how to represent variables in a mathematical form for statistical purposes;
- how to construct frequency distributions, histograms, and bar graphs;
- measures of central tendency including calculation of mean, median and mode;
- measures of dispersion including calculation of maximum, minimum, range, standard deviation, variance and coefficient of variation.

I. INTRODUCTION TO STATISTICS AND SIMPLE DATA DESCRIPTION

A. What Is Statistics?

Statistics refers to a variety of processes that deal with data. In particular, statistics includes the collection, assembly, classification, summarization, presentation, and analysis of data. Analysis includes reaching conclusions about data and making decisions based on the data. Statistics may also refer to a single piece of data, or summary measure (this will be defined below).

Statistics are used in one form or another by most people every day. The newspapers quote figures about rising house prices and inflation rates, commercials cite statistics about one group having fewer cavities, sports reporters talk about batting averages and weathermen talk about the probability of rain. Each of these examples involves the use of statistics, although each is different in its own way.

Statistics have a variety of uses in appraisal. One of the more important uses is to measure the quality of work performed. Statistical analysis of sales can also be used to make inferences about expected sales prices of other properties. Statistics can help identify the particular attributes, and their relative importance, that significantly contribute to observed sale prices.

In this chapter the emphasis is on some of the uses of statistics exemplified above; however, the discussion will not be exhaustive of all aspects of statistics. In particular, the goals of this chapter are to enable the reader to:

- read and understand material employing simple statistics; and
- use simple statistical measures and techniques to describe and analyze data.

To accomplish this, the chapter will focus on descriptive statistics, that is, the process of describing and summarizing data.

---

1 Data will be defined in detail below. Briefly, data refers generally to numerical information.
B. Definitions and Simple Mathematics

Before proceeding, some definitions and mathematical expressions need to be introduced. While the presentation in this chapter is not highly mathematical in nature, some basic terms are relevant.

A variable is a symbol or name which can take on any number of a predetermined set of values. The variable "SEX" can have the values "male" and "female". The variable "N" which may be used to represent the number of children in a family can take on any value from among 0, 1, 2, 3 and so on. "N", in this case, may not take the value 3.7 because a family may not have 3.7 children. A similar statement can be made about the number of rooms in a house (with the exception of 0 because a house must have at least one room); however, if N represents the number of bathrooms in a house, it could assume the values 0, .5, 1, 1.5, and so on, where .5 is a half-bathroom. From one example to another, the variable name (such as "SEX" and "N") can be used to represent different variables; however, within any given problem or example, each variable name should refer to only one variable and be used consistently.

Variables may be either continuous or discrete. Continuous variables are those variables which can theoretically take on any value between two other given values. Examples of continuous variables include height, weight, and number of square feet in a house. For example, height can be 5', or 5'2", or 5' 2.35" and so on. This type of variable often represents a measurement. Discrete variables are all other variables, and can take on only a limited number of values. Discrete variables may not take on fractional values and they often represent counts. For example, if N represents the number of students in a statistics class, N would be a discrete variable. N would also be discrete if it represented the number of rooms in a house. The importance of this distinction will be discussed later.

Variables are frequently denoted with subscripts, for example, $X_i$ where the subscript "i" takes on the values 1, 2, 3, and 4, etc. In this example, "i" is the subscript. In this case, $X_i$ would be a shorthand version for $X_1$, $X_2$, $X_3$, and $X_4$, each of which would assume some value. For example, $X_i$ might equal 7, $X_2$ might equal 39, $X_3$ might equal 3, and $X_4$ might equal 6. If X is a variable which represents the number of rooms in a series of houses, then the first house might have 7 rooms ($X_1 = 7$), the second might have 39 rooms ($X_2 = 39$) and so on. If the sum of the values of X is required, it is possible to write

$$\text{TOTAL} = X_1 + X_2 + X_3 + X_4$$  \hspace{1cm} \text{Equation 1}

A shorthand version would use the summation sign (the capital Greek letter sigma $\Sigma$), as follows:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n$$  \hspace{1cm} \text{Equation 2}

which says that the values of $X_i$ for each value of $i$ ranging from 1 to $n$ should be added together, where $n$ is a positive integer\(^2\) with $n$ greater than or equal to 1. This concept is clarified in the following illustration.

---

\(^2\) An integer is a counting, or whole number. Positive integers are those integers greater than zero. For example, 1, 2, 3 and so on.
Illustration 1

By using Equation 2, it follows that:

\[ \sum_{i=1}^{1} X_i = X_1 \]
\[ \sum_{i=1}^{2} X_i = X_1 + X_2 \]
\[ \sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4 \]
\[ \sum_{i=3}^{4} X_i = X_3 + X_4 \]

Now, if \( X_1 = 7 \), \( X_2 = 39 \), \( X_3 = 3 \), and \( X_4 = 6 \), then:

\[ \sum_{i=1}^{1} X_i = X_1 = 7 \]
\[ \sum_{i=1}^{2} X_i = X_1 + X_2 = 7 + 39 = 46 \]
\[ \sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4 = 7 + 39 + 3 + 6 = 55 \]
\[ \sum_{i=3}^{4} X_i = X_3 + X_4 = 3 + 6 = 9 \]

Example 1

If \( X_1 = 7 \), \( X_2 = 39 \), \( X_3 = 3 \), and \( X_4 = 6 \), then find the indicated sums.

(a) \( \sum_{i=2}^{4} X_i \)  
(b) \( \sum_{i=1}^{4} X_i \)  
(c) \( \sum_{i=1}^{3} X_i \)

Solution:

(a) 48  
(b) 55  
(c) 49
Example 2

If \( X_1 = 10, X_2 = 3, X_3 = 7, X_4 = 9, X_5 = 2 \), then calculate the following sums

\[
\begin{align*}
(a) & \quad \sum_{i=1}^{5} X_i \\
(b) & \quad \sum_{i=1}^{3} X_i \\
(c) & \quad \sum_{i=3}^{5} X_i
\end{align*}
\]

Solution:

(a) 31  
(b) 20  
(c) 18

Again, the summation sign (\( \sum \)) is simply a shorthand method for indicating that a series of numbers should be added together. In many cases, the subscripts are dropped if no confusion arises as to which index is being used. For example,

\[
\sum_{i=1}^{m} X_i \quad \text{is often written as} \quad \sum X
\]

Throughout this chapter, reference will be made to the use of and the process of analyzing data. Data refers to information on one or more variables which has been collected. Data may be numerical (for example, housing values) or non-numerical (for example, types of construction material). Data might be obtained by asking questions, counting, or looking in census books, among many other methods. Many of the examples in this chapter will employ housing unit data which might be obtained from real estate boards, multiple listing services, or land registry offices.

It is common to examine data which are expressed in absolute terms such as house values, population, and number of houses sold. However, when changes in data values are considered, problems in comparisons may occur. For example, if it is known that the value of a house increased by \$15,000\ during the year, it makes a difference whether the change is from \$20,000\ to \$35,000\ or from \$300,000\ to \$315,000. Thus, it is common to convert the absolute changes to percentage changes. To calculate a percentage, the base value, or initial value, must first be determined. The percentage change is 100 times a fraction. The fraction's numerator is the change in the value from the base to the new value and the fraction's denominator is the base value. The formula to be used is as follows:

\[
\text{Percentage change} = 100 \times \frac{\text{Final Value} - \text{Base Value}}{\text{Base Value}}
\]

Equation 3

Several illustrations will help explain the use of this formula.
**Illustration 2**

Assume Mr. Smith buys a house for $20,000 and later sells it for $35,000. The percentage change may be calculated as follows:

| Base value | $20,000 |
| Final value | $35,000 |
| Change | $15,000 |

Percentage change = \( \frac{15,000}{20,000} \times 100 \times .75 = 75\% \)

**Illustration 3**

Suppose an apartment building is purchased for $300,000 and sold one month later for $315,000. The percentage change may be calculated as follows:

| Base value | $300,000 |
| Final value | $315,000 |
| Change | $15,000 |

Percentage change = \( \frac{15,000}{300,000} \times 100 \times .05 = 5\% \)

**Illustration 4**

Suppose a suite is purchased for $50,000 and then sold for $20,000 because the roof caves in. The percentage change in price may be calculated as follows:

| Base value | $50,000 |
| Final value | $20,000 |
| Change | $30,000 |

Percentage change = \( \frac{-30,000}{50,000} \times 100 \times -.60 = -60\% \)

**Example 3**

Calculate the percentage change from $500 to $600 where these represent monthly rents.

**Solution:**

Percentage change = 20%

**Example 4**

Calculate the percentage change from 600 to 500 where these represent the number of housing units completed in an area in two consecutive years.

**Solution:**
Percentage change = -16.67%

**Example 5**

Calculate the percentage change from $500 to $500 where these represent the monthly rent of an apartment from one month to the next.

**Solution:**
Percentage change = 0.00%

**Example 6**

The following table presents data on the number of units sold in a residential area of a major city:

<table>
<thead>
<tr>
<th>YEAR</th>
<th>UNITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989</td>
<td>1,050</td>
</tr>
<tr>
<td>1990</td>
<td>1,400</td>
</tr>
<tr>
<td>1991</td>
<td>1,200</td>
</tr>
</tbody>
</table>


**Solution:**
1989 to 1990
absolute change = 350
percentage change = 33.33%

1990 to 1991
absolute change = -200
percentage change = -14.29%

1989 to 1991
absolute change = 150
percentage change = 14.29%

As noted at the outset, changes in values over time may be expressed in absolute or percentage terms. Each, when considered alone, may be misleading. A large percentage change due to a small base value should not be compared to a large percentage change when the base is itself large. This is because without information on the size of the base, the percentage cannot be fully understood. Thus, in the event that the percentage measure would be misunderstood, both the base and the percentage change should be reported.

**C. Simple Data Description**

In this section, some methods of data description will be outlined. To discuss the various techniques, the following hypothetical data on housing values will be used.
Illustration 5

For a given month in 1991, assume that 25 homes are sold in a particular neighbourhood for the following prices. These data are arranged in ascending order for convenience, although this need not be done.

Table 1
Hypothetical Housing Prices

<table>
<thead>
<tr>
<th>Data Value</th>
<th>Absolute Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$34,000</td>
<td>66,000</td>
</tr>
<tr>
<td>50,000</td>
<td>66,000</td>
</tr>
<tr>
<td>50,000</td>
<td>66,000</td>
</tr>
<tr>
<td>65,000</td>
<td>69,000</td>
</tr>
<tr>
<td>66,000</td>
<td>71,000</td>
</tr>
<tr>
<td>71,000</td>
<td>71,000</td>
</tr>
<tr>
<td>79,000</td>
<td>80,000</td>
</tr>
<tr>
<td>80,000</td>
<td>80,000</td>
</tr>
<tr>
<td>81,000</td>
<td>85,000</td>
</tr>
<tr>
<td>85,000</td>
<td>85,000</td>
</tr>
<tr>
<td>95,000</td>
<td>95,000</td>
</tr>
<tr>
<td>99,000</td>
<td>100,000</td>
</tr>
<tr>
<td>100,000</td>
<td>110,000</td>
</tr>
<tr>
<td>110,000</td>
<td>156,000</td>
</tr>
</tbody>
</table>

One method of describing these 25 sales prices is simply to list them as has been done above. This method provides a maximum amount of information as each and every sale is detailed and there is no need for any guess work or assumptions. Notice that when the data are arranged in ascending order, it is very easy to determine the highest and lowest values. However, this method becomes cumbersome quite quickly as the number of data items increases. One can imagine how much space would be required to list out the price for every house that was sold in 1991 in a large city with an active real estate market. Because of this, several methods have been devised so that the data may be organized and summarized in a more compact way.

An absolute frequency distribution is a more compact method of description, particularly when a particular value appears more than once in the distribution. This technique lists each value and the number of times or frequency that the value occurs in the list of data items. For the convenience of the reader, the values in a frequency distribution should be listed in ascending order.

Using the data from Illustration 5, the following frequency distribution may be constructed:

Table 2
Absolute Frequency Distribution of Housing Prices

<table>
<thead>
<tr>
<th>Data Value</th>
<th>Absolute Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$34,000</td>
<td>1</td>
</tr>
<tr>
<td>50,000</td>
<td>2</td>
</tr>
<tr>
<td>65,000</td>
<td>1</td>
</tr>
<tr>
<td>66,000</td>
<td>4</td>
</tr>
<tr>
<td>69,000</td>
<td>1</td>
</tr>
<tr>
<td>71,000</td>
<td>3</td>
</tr>
<tr>
<td>79,000</td>
<td>1</td>
</tr>
<tr>
<td>80,000</td>
<td>2</td>
</tr>
<tr>
<td>81,000</td>
<td>1</td>
</tr>
<tr>
<td>85,000</td>
<td>2</td>
</tr>
<tr>
<td>95,000</td>
<td>2</td>
</tr>
<tr>
<td>99,000</td>
<td>1</td>
</tr>
<tr>
<td>100,000</td>
<td>1</td>
</tr>
<tr>
<td>110,000</td>
<td>2</td>
</tr>
<tr>
<td>156,000</td>
<td>1</td>
</tr>
</tbody>
</table>
In this Illustration, the absolute frequency distribution is a more compact method because it involves listing only 15 values (and their associated frequencies) rather than all 25 values. However, this method could still be cumbersome if there were a large number of distinct values. Thus, it is sometimes preferable to group data before constructing a frequency distribution.

In grouping data, the practice is to place into the same group (or cell) data values which are close together. For ease of analysis, each group should have the same width. In the example data at hand, each width might be $5,000 or $10,000, depending on the needs of the analysis. The widths could be any amount, but generally they are chosen to be convenient sizes so that the midpoint is also a convenient number with which to work (the reasons for this will be discussed later). It is essential that the groups be designed so that every value will fit into some group (this property says that the groups totally exhaust the set of possible values). Further, the groups should be designed so that they do not overlap (this property says that the groups should be mutually exclusive). Thus, the groups should be designed so that every possible value will fit into one and only one group.

It is important to choose the group bounds with the data values in mind. The key point is that the bounds must be chosen so as to allow the placement of each value into a cell. Because all of the data values in Illustration 5 are whole dollar amounts, the bounds could also be whole dollar amounts. If the data values included cents, then it would be necessary to include cents amounts in the bounds.

For the data from Illustration 5, one might choose group widths which have a width of $10,000 (actually $9,999.99 so that all possible values may be placed into a group unambiguously). Thus the groups would be:

<table>
<thead>
<tr>
<th>GROUP</th>
<th>GROUP FREQUENCY</th>
<th>MIDPOINT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$30,000.00 - 39,999.99</td>
<td>1</td>
<td>35,000</td>
</tr>
<tr>
<td>$40,000.00 - 49,999.99</td>
<td>0</td>
<td>45,000</td>
</tr>
<tr>
<td>$50,000.00 - 59,999.99</td>
<td>2</td>
<td>55,000</td>
</tr>
<tr>
<td>$60,000.00 - 69,999.99</td>
<td>6</td>
<td>65,000</td>
</tr>
<tr>
<td>$70,000.00 - 69,999.99</td>
<td>4</td>
<td>75,000</td>
</tr>
<tr>
<td>$80,000.00 - 89,999.99</td>
<td>5</td>
<td>85,000</td>
</tr>
<tr>
<td>$90,000.00 - 99,999.99</td>
<td>3</td>
<td>95,000</td>
</tr>
<tr>
<td>$100,000.00 - 109,999.99</td>
<td>1</td>
<td>105,000</td>
</tr>
<tr>
<td>$110,000.00 - 119,999.99</td>
<td>2</td>
<td>115,000</td>
</tr>
<tr>
<td>$120,000.00 - 129,999.99</td>
<td>0</td>
<td>125,000</td>
</tr>
<tr>
<td>$130,000.00 - 139,999.99</td>
<td>0</td>
<td>135,000</td>
</tr>
<tr>
<td>$140,000.00 - 149,999.99</td>
<td>0</td>
<td>145,000</td>
</tr>
<tr>
<td>$150,000.00 - 159,999.99</td>
<td>1</td>
<td>155,000</td>
</tr>
</tbody>
</table>

It would have been incorrect to have, for example, one group with bounds of $30,000 and $35,000, and the next group with bounds of $40,000 and $45,000, as there would be no group for a value such as $36,000. Using the earlier terminology, these groups would not be exhaustive because the $36,000 figure would be left out. Similarly, it would have been incorrect to have one group with bounds of $30,000 and $35,000, and the next

---

3 The term "range" is also used for width in this context.
group with bounds of $35,000 and $40,000, as the groups would overlap. The value $35,000 could thus be placed in two groups rather than one. Using the earlier terminology, these groups would not be mutually exclusive. Groups with a frequency of zero (such as 130,000.00 to 139,999.99) may be deleted.

Alternatively, the group width could have been chosen to be $20,000 (actually $19,999.99 so that all possible values may be placed into a group). In this case, the frequency distribution would be as shown in Table 4.

<table>
<thead>
<tr>
<th>GROUP FREQUENCY</th>
<th>MIDPOINT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ 30,000.00 - 49,999.99</td>
<td>40,000</td>
</tr>
<tr>
<td>$ 50,000.00 - 69,999.99</td>
<td>60,000</td>
</tr>
<tr>
<td>$ 70,000.00 - 89,999.99</td>
<td>80,000</td>
</tr>
<tr>
<td>$ 90,000.00 - 109,999.99</td>
<td>100,000</td>
</tr>
<tr>
<td>$110,000.00 - 129,999.99</td>
<td>120,000</td>
</tr>
<tr>
<td>$130,000.00 - 149,999.99</td>
<td>140,000</td>
</tr>
<tr>
<td>$150,000.00 - 169,999.99</td>
<td>160,000</td>
</tr>
</tbody>
</table>

In this case, there are only seven groups as opposed to the 13 groups obtained when the group width was $10,000 (9,999.99). Both may be compared to the first absolute frequency distribution which has 15 values and the original data list which had 25 items. Groups with a frequency of zero (such as $130,000 to 149,999.99) can be deleted. Using the data from Illustration 5, two group widths have been demonstrated, each of which allows a different number of groups. There are no absolute rules regarding the number of groups that should be used for any given situation. In fact, the data should dictate how many groups should be used, and it is quite likely that more than one number of groups could be used in any example. What is important is that there is a trade-off between ease of presentation and detail of information provided by the data as the number of groups changes. Increasing the number of groups provides more information but decreases the ease of presentation. As the number of groups decreases, ease of presentation increases while the amount of information provided decreases. For example, given only the frequency distribution with seven groups above, it is unknown whether the nine sales in the $70,000.00 to $89,999.99 group are all clustered near $70,000, clustered near $89,999.99, or spread evenly throughout the interval. Each of these situations is quite different.

If there are a few extreme values which are considerably larger or smaller than most of the other values, then the final and/or initial group might be open-ended. An open-ended group is one in which one of the two bounds is unspecified, for example "$25,000+", or "over $25,000". This may be contrasted with closed-ended groups which specify both bounds, such as $25,000.00 to $29,999.99. In Illustration 5, if there were another observation with a value of $225,000, the final group could be "over $170,000" which is open-ended, rather than having several intermediate groups with no entries. The problem with open-ended groups is that the reader does not know if the extreme value is $175,000, $225,000, $350,000 or $1,000,000 or any other large value. Using open-ended intervals poses an additional problem in that the group midpoint cannot be calculated, and as will be seen later, the midpoint of a group is an important measurement.

Relative frequencies are often presented in addition to the absolute frequencies. The relative frequency for any single value or cell is the frequency for that value or cell divided by the total number of data points. For any particular problem, the relative frequencies always sum to one, because the sum of the numerators in the relative frequencies is equal to the total number of data points which is equal to the denominator.
Illustration 6

Using the data from Illustration 5, relative frequencies may be calculated.

<table>
<thead>
<tr>
<th>GROUP</th>
<th>ABSOLUTE GROUP FREQUENCY</th>
<th>RELATIVE GROUP FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>$30,000.00 - 49,999.99</td>
<td>1</td>
<td>1/25 = .04</td>
</tr>
<tr>
<td>$50,000.00 - 69,999.99</td>
<td>8</td>
<td>8/25 = .32</td>
</tr>
<tr>
<td>$70,000.00 - 89,999.99</td>
<td>9</td>
<td>9/25 = .36</td>
</tr>
<tr>
<td>$90,000.00 - 109,999.99</td>
<td>4</td>
<td>4/25 = .16</td>
</tr>
<tr>
<td>$110,000.00 - 129,999.99</td>
<td>2</td>
<td>2/25 = .08</td>
</tr>
<tr>
<td>$130,000.00 - 149,999.99</td>
<td>0</td>
<td>0/25 = .00</td>
</tr>
<tr>
<td>$150,000.00 - 169,999.99</td>
<td>1</td>
<td>1/25 = .04</td>
</tr>
<tr>
<td>Total Observations</td>
<td>25</td>
<td>1.00</td>
</tr>
</tbody>
</table>

For the group $30,000.00 to $49,999.99, there is one observation. Thus, the relative frequency for this group is 1 divided by 25 (the total number of observations). For the group $50,000.00 to $69,999.99, there are eight observations. The relative frequency for this group is thus 8 divided by 25 or .32. As can be seen, the relative frequencies sum to one.

D. Pictorial Data Description

The distribution of data values is sometimes more easily seen in a pictorial representation of the data. With ungrouped data, a line graph may be drawn which shows the relationship between the data value and either the absolute or relative frequency of that data value. The data values would be shown on the horizontal axis, while the frequencies would be shown on the vertical axis.

Illustration 7

Using the data from Illustration 5, two frequency distributions may be derived. Figure 1 shows a line graph with absolute frequencies. A line graph with relative frequencies is presented in Figure 2. These two line graphs are combined in Figure 3 with absolute frequencies shown on the left vertical scale and relative frequencies shown on the right vertical scale.
Introduction to Statistical Analysis

Figure 1
Line Graph for Absolute Frequencies of Ungrouped Data on Hypothetical Sale Prices

Figure 2
Line Graph for Relative Frequencies of Ungrouped Data on Hypothetical Sale Prices
For grouped data, the pictorial relationship between frequencies and data values is referred to as a bar graph, or histogram. Because each group has width, bars are used rather than lines to indicate the frequency. The width of the bar is the same as the width of the group, and in the histogram, the bar is centred around the group midpoint.

**Illustration 8**

Using the frequency distributions from Illustration 6, a histogram can be drawn for sales price data. The group width is $20,000 so that will be the width of each bar.

Figure 4 presents the histogram for absolute frequencies, while Figure 5 presents the histogram for relative frequencies. Figure 6 presents a histogram showing both absolute and relative frequencies.
Figure 4
Histogram of Absolute Frequencies for Grouped Data on Hypothetical Sale Prices

Figure 5
Histogram of Relative Frequencies for Grouped Data on Hypothetical Sale Prices
Example 7

For the following data, which represent numbers of square feet in fifteen living rooms, construct a frequency distribution and display the frequency distribution graphically.

\[
\begin{align*}
138 & \quad 164 & \quad 158 \\
146 & \quad 158 & \quad 140 \\
168 & \quad 126 & \quad 138 \\
146 & \quad 173 & \quad 140 \\
164 & \quad 145 & \quad 138 \\
\end{align*}
\]

Solution:

\[
\begin{array}{ccc}
\text{Value} & \text{Absolute Frequency} & \text{Relative Frequencies} \\
126 & 1 & 1/15 = .07 \\
138 & 3 & 3/15 = .20 \\
140 & 2 & 2/15 = .13 \\
145 & 1 & 1/15 = .07 \\
146 & 2 & 2/15 = .13 \\
158 & 2 & 2/15 = .13 \\
164 & 2 & 2/15 = .13 \\
168 & 1 & 1/15 = .07 \\
173 & 1 & 1/15 = .07 \\
15 & 1 & 1.00 \\
\end{array}
\]
Example 8

Using the data from Example 7 above, first group it and then construct a frequency distribution and histogram.

Solution:
Choosing a cell width of 10 would yield the following frequency distribution:

<table>
<thead>
<tr>
<th>Cell</th>
<th>Absolute Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>125-134</td>
<td>1</td>
<td>1/15 = .07</td>
</tr>
<tr>
<td>135-144</td>
<td>5</td>
<td>5/15 = .33</td>
</tr>
<tr>
<td>145-154</td>
<td>3</td>
<td>3/15 = .20</td>
</tr>
<tr>
<td>155-164</td>
<td>4</td>
<td>4/15 = .27</td>
</tr>
<tr>
<td>165-174</td>
<td>2</td>
<td>2/15 = .13</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.00</td>
</tr>
</tbody>
</table>
The line graphs and histograms described above are two demonstrations of graphs which may be used to describe data according to their frequencies. Graphs may also be used to present other types of data visually, as the following illustration shows.

**Illustration 9**

The following is hypothetical data on single family housing starts and completions in a large city:

<table>
<thead>
<tr>
<th>YEAR</th>
<th>STARTS</th>
<th>COMPLETIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>699</td>
<td>722</td>
</tr>
<tr>
<td>1983</td>
<td>570</td>
<td>597</td>
</tr>
<tr>
<td>1984</td>
<td>553</td>
<td>618</td>
</tr>
<tr>
<td>1985</td>
<td>808</td>
<td>620</td>
</tr>
<tr>
<td>1986</td>
<td>670</td>
<td>759</td>
</tr>
<tr>
<td>1987</td>
<td>792</td>
<td>726</td>
</tr>
<tr>
<td>1988</td>
<td>816</td>
<td>768</td>
</tr>
<tr>
<td>1989</td>
<td>948</td>
<td>819</td>
</tr>
<tr>
<td>1990</td>
<td>702</td>
<td>771</td>
</tr>
<tr>
<td>1991</td>
<td>360</td>
<td>574</td>
</tr>
</tbody>
</table>

These data may be displayed graphically to increase the ability of the reader to understand it and make comparisons.
Graphs and tables can aid in interpreting and analyzing data; however, they can also be misleading. To avoid any difficulties in interpretation, readers should keep the following points in mind when constructing graphs:

- If enough data points are available, the lines in the graphs and charts may be drawn as continuous curves. However, with only a small number of data points, straights lines joining each point should be used.

- Graphs should be clearly labelled with a descriptive title. All lines or entries in the graph itself should also be clearly labelled. Where necessary, different types of graphics (straight lines, dashes, dots and dashes) should be used to clearly delineate each relationship.

- In some cases, there may be no data within a given range. For example, in Figure 7, none of the data points on starts or completions are less than 300. In these cases, the scale may be truncated, that is, the range for which there are no values may be deleted from the scale. This fact should be noted clearly in the graph by a break. Immediately above the break in the scale, a data value should be given as a reference point. This has been done in Figure 7.

When tables are used, the following guidelines should be observed:

- A complete descriptive title should appear with the table to clearly indicate what data are being presented.

- The source(s) of the data should appear below the table.

- Entries which are zero and those for which data is not available (use n.a.) should be distinguished.
• Units of measurement should be clearly indicated.

E. Summary

In this section, some basic definitions and notations that are fundamental to the study of statistics have been presented. The section concentrated on methods of describing data values. It is noted that the most complete information is provided to the reader when every individual data value is listed, or when a frequency distribution on ungrouped data is presented. However, either of these methods may prove to be cumbersome in particular cases. In an effort to overcome this problem, data values may be placed together into groups with equal widths. Frequency distributions may again be derived, and they may be depicted in histograms or bar graphs. The grouping technique, while perhaps more convenient, does not provide the reader with as much detailed information as is provided when data is left ungrouped.

II. UNIVARIATE DATA DESCRIPTIVE MEASURES

A. Introduction

In the previous section, several techniques used to describe data were discussed. In particular, the importance of frequency distributions and histograms was discussed. These techniques provide information by listing every data value or by grouping them into cells. These techniques are potentially awkward if there are many distinct data values. Thus, it is often useful to employ single numbers, or summary measures, to describe and typify the data values. This section focuses on measures of central tendency, which refer to simple figures that are some sort of middle value; and measures of dispersion, which refer to how spread out the distribution of data values is. There are other summary measures; however, they are rarely used.

B. Measures of Central Tendency

Although there are several measures of central tendency commonly used, each of which has its own attributes, each is often called an "average". Use of the term "average" can be misleading. In place of this term, the precise names of the various measures of central tendency should be used: mean, median and mode.

The measures of central tendency help to place the distribution on the horizontal scale of a graph. The following graph of two frequency distributions illustrates this point. In this graph, the distributions of two variables (X and Y) and their respective means (to be defined below) are shown. That the mean of the X distribution is less than the mean of the Y distribution helps to determine the horizontal placement of the distributions.
1. **Mean**

The most commonly used measure of central tendency is the arithmetic mean, or simply the mean. The mean is also referred to as the expected value. To compute the mean for a set of data, it is first necessary to calculate the sum (total) of all the numbers in the distribution, and then divide the sum by the number of data points in the set.

**Illustration 10**

To find the mean (more precisely, the arithmetic mean) of the numbers 6, 15, and 21, first find the sum (which is 42). Then divide the sum (42) by the number of observations (3) to get the mean. In this case, the mean is 14.

**Illustration 11**

Using the data on housing prices from Illustration 5 (repeated below), find the arithmetic mean. The data values are:

| $34,000 | 66,000 | 71,000 | 81,000 | 99,000 |
| 50,000 | 66,000 | 71,000 | 85,000 | 100,000 |
| 50,000 | 66,000 | 79,000 | 85,000 | 110,000 |
| 65,000 | 69,000 | 80,000 | 95,000 | 110,000 |
| 66,000 | 71,000 | 80,000 | 95,000 | 156,000 |

First, the sum must be calculated. In this example, the sum is $2,000,000. Next, the sum must be divided by the number of data values, which is 25. The arithmetic mean is $80,000.

\[
\text{Mean} = \frac{\$2,000,000}{25} = \$80,000
\]
Example 9

The following numbers represent the number of days ten houses have been listed with a multiple listing service:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>47</td>
<td>15</td>
</tr>
<tr>
<td>53</td>
<td>23</td>
<td>4</td>
</tr>
<tr>
<td>90</td>
<td>18</td>
<td>93</td>
</tr>
<tr>
<td>117</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Find the mean number of days a house has been listed.

Solution:
mean number of days = 51

Some notation can be used to simplify the calculations. If $X_i$ for $i=1$ to $n$ are the values for which the mean is required, then the sum of the numbers is denoted by:

$$\sum_{i=1}^{n} X_i$$  

Equation 4

The symbol $\mu$ is generally used for the arithmetic mean. Thus,

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$  

Equation 5

This is a shorthand method for denoting the calculations involved in computing the arithmetic mean. This notation will be used repeatedly.

Illustration 12

Calculate the arithmetic mean for the following per square foot costs of construction for five single family units.

- $17.20
- $22.30
- $15.30
- $19.10
- $21.10

$$\sum_{i=1}^{5} X_i = 95.00$$

$n = 5$
If a frequency distribution has been constructed, the method of calculating the arithmetic mean changes slightly. In Illustration 12, each data value appeared only once so that the formula for the mean may be written as

$$\mu = \frac{\sum_{i=1}^{5} X_i}{n} = \frac{\$95}{5} = \$19.00$$

This formulation incorporates the frequency with which each value occurs. In this case, each frequency is one, but this need not be the case. Hence, it is useful to generalize the calculation. This may be generalized as follows:

$$\mu = \frac{\sum_{i=1}^{m} f_i X_i}{n}$$

where $f_i$ is the frequency with which the value $X_i$ occurs and,

$$\sum_{i=1}^{m} f_i X_i = f_1 X_1 + f_2 X_2 + f_3 X_3 + \ldots + f_m X_m$$

This method does not require adding each value as many times as it occurs; rather, it lists each value just once (multiplied by its frequency). Note that the sum of the frequencies equals $n$, that is

$$\sum_{i=1}^{m} f_i = n$$

where $m$ is the number of different data values.
Illustration 13

Using the data in the frequency distribution of Illustration 5, the arithmetic mean may be calculated using Equation 6.

<table>
<thead>
<tr>
<th>Data Value</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$34,000</td>
<td>1</td>
</tr>
<tr>
<td>$50,000</td>
<td>2</td>
</tr>
<tr>
<td>$65,000</td>
<td>1</td>
</tr>
<tr>
<td>$66,000</td>
<td>4</td>
</tr>
<tr>
<td>$69,000</td>
<td>1</td>
</tr>
<tr>
<td>$71,000</td>
<td>3</td>
</tr>
<tr>
<td>$79,000</td>
<td>1</td>
</tr>
<tr>
<td>$80,000</td>
<td>2</td>
</tr>
<tr>
<td>$81,000</td>
<td>1</td>
</tr>
<tr>
<td>$85,000</td>
<td>2</td>
</tr>
<tr>
<td>$95,000</td>
<td>2</td>
</tr>
<tr>
<td>$99,000</td>
<td>1</td>
</tr>
<tr>
<td>$100,000</td>
<td>1</td>
</tr>
<tr>
<td>$110,000</td>
<td>2</td>
</tr>
<tr>
<td>$156,000</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \mu = \frac{\sum_{i=1}^{m} f_i X_i}{n} \]

\[ \sum_{i=1}^{15} f_i X_i = (1)(34,000) + (2)(50,000) + (1)(65,000) + \\
(4)(66,000) + (1)(69,000) + (3)(71,000) + \\
(1)(79,000) + (2)(80,000) + (1)(81,000) + \\
(2)(85,000) + (2)(95,000) + (1)(99,000) + \\
(1)(100,000) + (2)(110,000) + (1)(156,000) = 2,000,000 \]

\[ n = \sum_{i=1}^{m} f_i = 25 \]

\[ \mu = \frac{\sum_{i=1}^{m} f_i X_i}{n} = \frac{2,000,000}{25} = 80,000 \]

Note that the number of terms to be summed is 15 (the number of distinct values of \( X \)), rather than 25 (the total number of observations of \( X \) values).

The formula for the arithmetic mean can also be expressed in terms of relative rather than absolute frequencies. The formula may be rewritten as follows:

\[ \mu = \sum_{i=1}^{m} \frac{f_i}{n} (X_i) \]  

Equation 9
The only difference between this and the previous formula is that \( f_i \) is divided by \( n \), rather than \( fX_i \) being divided by \( n \). These two formulations yield identical arithmetic means. Notice that \( f/n \) is the formula for a relative frequency. Thus, it should be apparent that the summation of the relative frequencies must equal one.

**Illustration 14**

Using the data from Illustration 5, absolute and relative frequencies may be calculated. The arithmetic mean can then be calculated using the relative frequency formulation (Equation 9).

<table>
<thead>
<tr>
<th>Data Value</th>
<th>Absolute Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$34,000</td>
<td>1</td>
<td>.04</td>
</tr>
<tr>
<td>50,000</td>
<td>2</td>
<td>.08</td>
</tr>
<tr>
<td>65,000</td>
<td>1</td>
<td>.04</td>
</tr>
<tr>
<td>66,000</td>
<td>4</td>
<td>.16</td>
</tr>
<tr>
<td>69,000</td>
<td>1</td>
<td>.04</td>
</tr>
<tr>
<td>71,000</td>
<td>3</td>
<td>.12</td>
</tr>
<tr>
<td>79,000</td>
<td>1</td>
<td>.04</td>
</tr>
<tr>
<td>80,000</td>
<td>2</td>
<td>.08</td>
</tr>
<tr>
<td>81,000</td>
<td>1</td>
<td>.04</td>
</tr>
<tr>
<td>85,000</td>
<td>2</td>
<td>.08</td>
</tr>
<tr>
<td>95,000</td>
<td>2</td>
<td>.08</td>
</tr>
<tr>
<td>99,000</td>
<td>1</td>
<td>.04</td>
</tr>
<tr>
<td>100,000</td>
<td>1</td>
<td>.04</td>
</tr>
<tr>
<td>110,000</td>
<td>2</td>
<td>.08</td>
</tr>
<tr>
<td>156,000</td>
<td>1</td>
<td>.04</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

\[
\mu = \frac{1}{n} \sum_{i=1}^{n} f_i (X_i) = (.04)(34,000) + (.08)(50,000) + (.04)(65,000) \\
+ (.16)(66,000) + (.04)(69,000) + (.12)(71,000) \\
+ (.04)(79,000) + (.08)(80,000) + (.04)(81,000) \\
+ (.08)(85,000) + (.08)(95,000) + (.04)(99,000) \\
+ (.04)(100,000) + (.08)(110,000) + (.04)(156,000) \\
= $80,000
\]

The method used to calculate the arithmetic mean for grouped data can also be used to calculate a weighted arithmetic mean. This type of mean is employed when the data values have different frequencies or levels of importance. The formula to use when a weighted arithmetic mean is calculated is as follows:

\[
\mu = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}
\]

where \( w_i \) is the weighting factor for the value \( X_i \). The following illustration demonstrates the use of the formula:
Illustration 15

You are asked to calculate the mean grade for a real estate statistics course where the final examination mark is 85, the mid-term examination mark is 80 and the quiz mark is 71. The final examination mark is weighted three times as much as a quiz, and one and one-half times as much as the midterm examination. It would make no sense to simply calculate the mean of the three marks since they are to be weighted differently. Thus, the following table should be constructed:

<table>
<thead>
<tr>
<th>Mark</th>
<th>Weight</th>
<th>(Weight)(Mark)</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>3</td>
<td>255</td>
</tr>
<tr>
<td>80</td>
<td>2</td>
<td>160</td>
</tr>
<tr>
<td>71</td>
<td>1</td>
<td>71</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>Total 486</td>
</tr>
</tbody>
</table>

This table is identical to the frequency distribution table that would be constructed if there were six examination marks (three 85s, two 80s and one 71), each of which counted the same. Applying Equation 10 above:

\[
\mu = \frac{\sum_{i=1}^{3} w_i X_i}{\sum_{i=1}^{3} w_i} = \frac{(3)(85) + (2)(80) + (1)(71)}{3 + 2 + 1} = \frac{486}{6} = 81
\]

If \( w_i \) is replaced by \( f_i \) in Equation 10, the formula is identical to Equation 6 because:

\[
\sum_{i=1}^{m} w_i = n
\]

The mean may also be calculated for grouped data. The mean is given by:

\[
\mu = \frac{\sum_{i=1}^{m} f_i M_i}{n} \quad \text{Equation 11}
\]

where \( f_i = \) the frequency for group \( i \)
\( M_i = \) the midpoint for group \( i \)
\( n = \) total number of data values = \( \sum_{i=1}^{m} f_i \)

Actually, the value of the mean given by Equation 11 is only an estimate of the mean because it will change as the group bounds change, and as the midpoint of the group changes also. Thus, while it is easier to calculate the mean with grouped data than with individual data, some accuracy is sacrificed.
Illustration 16

Using the data from Illustration 5 as it was originally grouped with widths of $10,000 in section one, an estimate of the mean can be obtained as follows:

$$
\mu = \frac{\sum_{i=1}^{13} f_i M_i}{25} = \frac{(1)(35,000) + (0)(45,000) + (2)(55,000) + (6)(65,000) + (4)(75,000) + (5)(85,000) + (3)(95,000) + (1)(105,000) + (2)(115,000) + (0)(125,000) + (0)(135,000) + (0)(145,000) + (1)(155,000))}{25}
$$

$$
= \frac{2,035,000}{25}
$$

$$
= $81,400
$$

This is larger than the true value of the mean as calculated in Illustrations 11, 13, and 14. As noted above, this can occur generally because this calculation provides only an estimate of the mean. It occurs in this particular case because the group midpoint tends to be larger than the values in the group, and thus the sum obtained by using group midpoints is larger than the sum obtained by using the actual data. An alternative estimate of the mean can also be obtained using the seven groups given by cell widths of $20,000 as follows:

$$
\mu = \frac{\sum_{i=1}^{7} f_i M_i}{25} = \frac{(1)(40,000) + (8)(60,000) + (9)(80,000) + (4)(100,000) + (2)(120,000) + (0)(140,000) + (1)(160,000))}{25}
$$

$$
= \frac{2,040,000}{25}
$$

$$
= $81,600
$$

This grouping gives an over-estimate of the true mean for the same reasons as before.

The mean is affected by each data item under consideration. This is advantageous as it increases the reliability of the mean. However, this implies that it is also affected by extreme values in the group. For example, the mean of the numbers 3, 5, and 100 is 36, but this is not particularly representative of any of the three values as the mean has been affected by the one extreme value of 100. Another problem that arises because every value is considered in calculating the mean is that the necessary computations can be lengthy since all values must be summed. One further problem is that a mean cannot be calculated for grouped data when one of the groups is open-ended because that group has no midpoint. Despite these problems, the arithmetic mean remains the most commonly used measure of central tendency because its properties are preferable to those of the other measures of central tendency.
2. **Median**

A second measure of central tendency is called the median. The median is defined as the middle data value in a distribution in which the data values are arranged in ascending (or descending) order. As such, it is the most central value because half of the values lie above it, and half lie below it.

To calculate the median for any distribution, these steps should be followed:

- arrange the data in ascending (or descending) order;
- select the middle value as the median if there is an odd number of data values in the distribution;
- compute the arithmetic mean of the two middle values if there is an even number of data values in the distribution. This median is the mean of the two middle data items.

**Illustration 17**

To calculate the median for the data items 5,2,4,35,9, first arrange the data in ascending order as follows:

```
2,4,5,9,35
```

Since there is an odd number of data items, the middle value, 5, is the median value.

You should verify that the median is the same if the data is arranged in descending order. You should also note that the arithmetic mean of this distribution is 11.

**Illustration 18**

Calculate the median for the data items 3,9,6,12.

First arrange the data in ascending order as follows:

```
3,6,9,12.
```

Since there is an even number of data items in this distribution, the median is the arithmetic mean of the two middle values (6 and 9), which is 7.5.

**Illustration 19**

Calculate the median for the housing prices given in Illustration 11 (repeated below). Because there is an odd number of prices (25), the median is the middle price, which is the 13th price when they are arranged in ascending order. The median is $79,000.

```
$34,000 66,000 71,000 81,000 99,000  
50,000 66,000 71,000 85,000 100,000  
50,000 66,000 79,000 85,000 110,000  
65,000 69,000 80,000 95,000 110,000  
66,000 71,000 80,000 95,000 156,000  
```
Example 10

Calculate the median for the following data items:

15, 25, 16, 17, 19, 13

Solution:

Median = \(\frac{16 + 17}{2} = \frac{33}{2} = 16.5\)

Example 11

Calculate the median for the following data items:

13, 23, 19, 100, 1,000, 18, 10,000

Solution:

Median = 23

The median as a measure of central tendency does not explicitly consider the value of every data item in the distribution, and as such, it is unaffected by extreme values. This may be contrasted with the mean which is affected by extreme values. In Illustration 17, 35 is an extreme value because it is much larger than the other data items in the distribution. The median is unaffected by this value, and in fact, the median would have been 5 even if the largest value (35) had been replaced by 10, 40, or 5,000 because 5 would still be the middle value in ascending or descending order. Further, it is generally possible to calculate a median for grouped data when there is an open-ended group, although the procedure is complicated. A disadvantage of the median is that it requires that the data be arranged in ascending or descending order.

3. Mode

A third measure of central tendency is the mode. The mode of a distribution is that data value which occurs most frequently in the distribution. The mode is the least used measure of central tendency because it may not be a central value - it could be an extreme value which occurs most frequently (see Illustration 24). A series of data values may have more than one mode if more than one value occurs most frequently. If every value occurs with equal frequency, there is no mode.

Illustration 20

The mode for the series 5,7,8,3,5 is 5 as it appears most often in the distribution. This series is called unimodal because it has only one mode.

Illustration 21

There is no mode for the set of numbers 5,7,8,3 because every value occurs with the same frequency.
Illustration 22

For the set of numbers 5, 7, 8, 3, 5, 8, there are two modes because both 5 and 8 occur with the highest frequency (twice). This set of numbers is called bimodal because there are two modes.

Illustration 23

Suppose you are asked to find the mode of the distribution of housing prices given in Illustration 11 and repeated below. In this case, the mode is $66,000 as this value occurs most frequently (four times) in this distribution. Note that this value is different from the mean of the distribution ($80,000) and the median of the distribution ($79,000) calculated earlier. This occurs because each measure of central tendency is affected by different factors.

Illustration 24

Suppose seven houses were listed with a multiple listing service. Below are the number of days each one was listed before it sold.

1, 2, 3, 5, 9, 100, 100

The mode for this distribution is 100 days as this value appears most frequently. However, 100 days is not a central value; rather, it is an extreme value and thus is not particularly representative of most of the values in the distribution.

Example 12

Find the mode for the following distribution where the numbers represent the number of rooms in each of eight houses:

13, 15, 13, 17, 16, 15, 19, 15

Solution:

Mode = 15 rooms

Example 13

Find the mode for the following distribution of living room sizes in square feet for seven houses:

100, 292, 370, 150, 293, 294, 101

Solution:

There is no mode.
Example 14

The following numbers represent grades on an examination given in a real estate statistics course. Find the mode for this distribution.

57, 53, 59, 56, 53, 58, 59, 55, 56

Solution:

Modes = 53, 56, and 59

Example 15

Suppose you are asked to find the average housing value for the fifteen houses in a neighbourhood. Because you know there are three averages (measures of central tendency), you calculate all three. Find the mean, median, and mode.

$138,000 164,000 158,000
146,000 158,000 140,000
168,000 126,000 138,000
146,000 173,000 140,000
164,000 145,000 138,000

Solution:

Mean = $149,466.67
Median = $146,000.00
Mode = $138,000.00

The mode suffers as a measure of central tendency because it may not be representative (as in Illustration 24), because it may not exist (as in Illustration 21), and because there may be more than one mode (as in Illustration 22). It has the advantage that it can be calculated even if the data values are not numeric. For example, in analyzing the responses to the question, "What is your favourite colour?", it is impossible to calculate a mean or median, but it is possible to determine the mode, which would be the colour most frequently mentioned. Further, the mode is unaffected by individual extreme values, although it may itself be an extreme value.

Illustration 25

For the data presented in Example 15 (repeated here), which measure of central tendency is most appropriate?

$138,000 164,000 158,000
146,000 158,000 140,000
168,000 126,000 138,000
146,000 173,000 140,000
164,000 145,000 138,000

In this example, the three measures of central tendency are relatively close together; that is, the mean is $149,470, the median is $146,000, and the mode is $138,000. Further, there are no extreme values in this distribution. As such, any of the three measures could be used. In situations like this, the mean would probably be best to use because it takes all of the values into consideration, and is not adversely affected by extreme values. Further, the mean is the best known "average".
The following three frequency distributions show three possible relationships between the mean, median, and mode. It is assumed that there are enough data values to connect all of the points and draw a smooth curve.
Figure 9 is a symmetric distribution, while the distribution in Figure 10 is said to be skewed to the right. The distribution in Figure 11 is said to be skewed to the left.

These are only three possible distributions, and they illustrate the possibility that one or more of the measures of central tendency may not be very representative.

The strengths and weaknesses of the mean, median and mode as measures of central tendency can now be reviewed. The mean is affected by every value (even extreme values), while the median and mode are not. Most sophisticated statistical tests use the mean, a few use the median, and virtually none use the mode. The mean requires that all of the values be summed, the median requires that they be arranged in descending/ascending order, and the mode requires that a frequency distribution be constructed. The mean and median always exist for numerical data and will be unique. The mode may not exist and it may not be unique. The mean and median require numerical data values, while the mode does not.

The choice between the mean, median, and mode should be made independently of whether the variable under consideration is discrete or continuous. For example, if the number of rooms in a set of houses -which is clearly discrete - is being analyzed, then the mode should not be chosen as the most desirable measure simply because it is a whole number.

It is impossible to provide a set of rules which indicates for each and every example which measure of central tendency is most appropriate. The circumstances of each example must dictate which measure is most appropriate. If there is any doubt, more than one such measure should be presented. This discussion highlights the problems associated with trying to select a single measure to summarize all of the data values in a set, rather than presenting the entire frequency distribution.
C. **Measures of Dispersion**

While the measures of central tendency place a distribution on the graph, they give no idea of how spread out or dispersed a distribution is. For example, the two distributions shown below have the same mean, median, and mode; however, the distributions are quite different. Distribution A is much more spread out than is distribution B.

Because of the likelihood that distributions may be quite different (i.e., more or less spread out), it is necessary to develop measures to indicate the dispersion of a distribution.

1. **Range**

The maximum and minimum values in a distribution are two measures that must be determined so that the range of the distribution can be calculated. The maximum is the largest value in the distribution while the minimum is the smallest value in the distribution. The range is equal to the difference between the maximum and the minimum. The range is obviously determined by extreme values but it ignores the dispersion among the values other than the minimum and maximum. Note that the range can be the same for two distributions despite the fact that the values other than the maximum and minimum are quite different. Because of this, the range is not a very useful measure of dispersion.
Illustration 26

For the data provided in Illustration 23 (repeated below), the range may be computed. The minimum value in the distribution is $34,000, and the maximum is $156,000. Thus,

$34,000  66,000  71,000  81,000  99,000  
50,000  66,000  71,000  85,000  100,000  
50,000  66,000  79,000  85,000  110,000  
65,000  69,000  80,000  95,000  110,000  
66,000  71,000  80,000  95,000  156,000

Range = maximum value - minimum value = $156,000 - $34,000 = $122,000

Illustration 27

Calculate the range for the following sales prices of twelve apartment buildings:

$1,056,000  1,105,000  1,098,000  1,178,000  
1,061,000  1,099,000  1,130,000  1,160,000  
1,073,000  1,057,000  1,120,000  1,149,000

Range = maximum value - minimum value = $1,178,000 - $1,056,000 = $122,000

Notice the range is the same as in Illustration 26, but the distributions are quite different.

Example 16

Suppose there are eight apartment buildings of different sizes. The numbers given below indicate the number of suites in each building. Use the maximum, minimum, and range to determine how dispersed the sizes are.

15, 23, 19, 47, 92, 16, 19, 93

Solution:

Maximum number of suites = 93
Minimum number of suites = 15
Range = 78

2. Standard Deviation

The most commonly employed measure of dispersion is the standard deviation for which the symbol is \( \sigma \) (the small Greek letter sigma). The formula for the standard deviation is as follows:\(^4\)

\[
\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}}
\]

Equation 12

---

\(^4\) See Appendix 1 for a discussion of the differences in calculating the standard deviation for populations and samples.
where
\[ \sigma = \text{standard deviation of the distribution;} \]
\[ \mu = \text{arithmetic mean of the distribution;} \]
\[ X_i = \text{the values for which the standard deviations is being calculated;} \]
\[ n = \text{the number of data items; and} \]
\[ \sqrt{ } = \text{symbol for the square root}. \]

The standard deviation measures the dispersion of the raw data around the mean of the distribution. As the dispersion or spread increases, the standard deviation increases. Conversely, the dispersion decreases and the distribution is more compact as the standard deviation decreases. Put another way, as the standard deviation increases, the arithmetic mean becomes less representative of the other values in the distribution. For the two distributions in Figure 12, the standard deviation for distribution B is smaller than for distribution A (this particular comparison is valid only because the means are equal; more will be said about this below).

**Illustration 28**

Calculate the standard deviation for the following data values

\[ 6, 10, 15, 9 \]

First, the mean must be calculated, which equals 10. Then, the standard deviation may be calculated by filling out the table below.

<table>
<thead>
<tr>
<th>( X_i )</th>
<th>( \mu )</th>
<th>( X_i - \mu )</th>
<th>((X_i - \mu)^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>10</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \sigma = \sum_{i=1}^{4} (X_i - \mu)^2 = 16 + 0 + 25 + 1 = 42 \]

\[ \sigma = \sqrt{\frac{42}{4}} = \sqrt{10.5} = 3.24 \]

This formula is time consuming, particularly as the number of data items increases. It can be shown algebraically that the formula given above is equivalent to:

\[ \sigma = \sqrt{\frac{1}{n} \left( \sum_{i=1}^{n} X_i^2 - n\mu^2 \right)} \]  

**Equation 13**

Using the formula, it is not necessary to compute \((X_i - \mu)\) for each item in the distribution, a fact which greatly simplifies the calculations. Consider the following illustration.
Illustration 29

Using the data from Illustration 28, calculate the standard deviation using Equation 13. The values used in that Illustration are 6, 10, 15 and 9.

\[
\mu = \frac{\sum_{i=1}^{4} X_i}{4} = \frac{6 + 10 + 15 + 9}{4} = 10
\]

\[
\sum_{i=1}^{4} X_i^2 = 6^2 + 10^2 + 15^2 + 9^2 = 36 + 100 + 225 + 81 = 442
\]

\[
\sigma = \sqrt{\frac{1}{4} \left[ 442 - (4 \times 10^2) \right]} = \sqrt{\frac{1}{4} (442 - 400)}
\]

\[
\sigma = \sqrt{\frac{1}{4} (42)} = \sqrt{10.5} = 3.24
\]

Illustration 30

A luxury apartment building has ten similar suites with the following distribution of monthly rents:

| $900 | $950   |
| $875 | $975   |
| $1050| $1000  |
| $1025| $1100  |
| $925 | $850   |

Calculate the mean and standard deviation for this distribution of monthly rents.

The mean monthly rent is $965. For the standard deviation, equation 13 may be used if no calculator is readily available. The equation requires that

\[
\sum_{i=1}^{10} X_i^2 = 9,370,000
\]

The calculation of the standard deviation is then straightforward using Equation 13:

\[
\sigma = \sqrt{\frac{1}{10} \left( \sum_{i=1}^{10} X_i^2 - n \mu^2 \right)} = \sqrt{\frac{1}{10} \left[ 9,370,000 - (10)(965)^2 \right]}
\]

\[
= \sqrt{\frac{1}{10} (9,370,000 - 9,312,250)} = \sqrt{\frac{1}{10} (57,750)}
\]

\[
= \sqrt{5,775} = 75.99
\]
This is small relative to the mean suggesting that the distribution is compact around the mean. This may be confirmed using the calculator and the procedure described in Illustration 11.

Under certain assumptions, the standard deviation may be used to determine how the values of a variable are distributed relative to the mean. If the distribution is bell-shaped, or normal (the distribution is symmetric about the mean, mode, and median which are all equal as in Figure 9), then approximately 68\% of all of the values in the distribution will be located within one standard deviation of the mean. Further, under the same conditions, approximately 95\% of all of the values will be located within two standard deviations of the mean.

3. Variance

The standard deviation is measured in the same units as the original variable. For example, if the distribution of sales prices of single family houses is listed in dollar amounts, then the standard deviation is also measured in dollars. Occasionally, the variance is used as a measure of dispersion rather than the standard deviation.

The variance, which is the square of the standard deviation, is measured in squared units. For example, if sales are listed in dollars, the variance is measured in dollars squared. The variance provides the same information about a distribution as the standard deviation, although it is not necessary to compute a square root. The variance is given by the following equation:

\[
\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2
\]

Equation 14

For computational purposes, the formula

\[
\sigma^2 = \frac{1}{n} \left( \sum_{i=1}^{n} X_i^2 - n\mu^2 \right)
\]

Equation 15

should be used which is equivalent to Equation 14 and similar to equation 13.

Illustration 31

Using the data from Illustration 28, calculate the variance. From Equation 14 or 15, it is seen immediately that:

\[
\sigma^2 = (\sigma)^2 = (3.24)^2 = 10.50
\]

4. Coefficient of Variation

The last measure of dispersion to be discussed is the coefficient of variation. Such a measure is necessary if the standard deviations for two variables are to be compared. For example, suppose that the standard deviation of house prices sold in Vancouver is $40,000, while the standard deviation of the incomes of the occupants of the houses is $20,000. It might be misleading to compare these two absolute deviations because one may have much less variation than the other relative to its mean. If the mean price of the houses is $100,000 while the mean income of the owners is $20,000, the relative variations are quite different. To make these comparisons, a measure called the coefficient of variation is used, and is defined as follows:
Coefficient of variation \[ \frac{\sigma}{\mu} \times 100 \] \hspace{1cm} \text{Equation 16}

This measure expresses variation relative to the mean in percentage terms.

**Illustration 32**

For a distribution of house prices, assume the mean is $100,000 and the standard deviation is $40,000, while for the distribution of incomes of the owners, the mean is $20,000 and the standard deviation is $20,000. While the standard deviation is larger for prices, this comparison is inappropriate. To compare these in percentage terms, the coefficients of variation should be calculated. For prices, the coefficient of variation is:

\[
CV \text{ prices } = \frac{\$40,000}{\$100,000} \times 100 = 40\%
\]

while for incomes, the coefficient of variation is:

\[
CV \text{ incomes } = \frac{\$20,000}{\$20,000} \times 100 = 100\%
\]

This shows that, despite the fact that the standard deviation of incomes is smaller for incomes than for prices, the distribution of incomes is more dispersed about its mean ($20,000) than is the distribution of house prices about its mean ($100,000).

**Illustration 33**

Suppose you must choose between two alternative real estate investments. The expected gain (mean) from investment A is $150,000 with a standard deviation of $30,000. Investment B has an expected gain (mean) of $100,000 and a standard deviation of $5,000. Which alternative would you choose and why?

The mean can be taken as a measure of the expected gain from the investment, while the standard deviation can be taken (and often is) as a measure of the risk associated with the investment. Investment B has a smaller standard deviation or measure of risk than Investment A. However, this comparison is not sound because the expected values (means) are different. Thus, the coefficient of variation should be calculated to make an accurate comparison.

Investment A: Coefficient of Variation = \[
\frac{\$30,000}{\$150,000} \times 100 = 20\%
\]

Investment B: Coefficient of Variation = \[
\frac{\$5,000}{\$100,000} \times 100 = 5\%
\]

Thus, relative to their means, there is less variability (or risk) associated with investment B than with investment A. The choice to be made depends on a subjective evaluation of the expected gains and risks. If you are risk averse, you would probably choose alternative B, while if you prefer risk, or want to maximize the expected gain, you would probably choose alternative A.
Example 17

Calculate the maximum, minimum, range, standard deviation, variance, and coefficient of variation for the following fifteen housing prices:

| $138,000 | 164,000 | 158,000 |
| 146,000 | 158,000 | 140,000 |
| 168,000 | 126,000 | 138,000 |
| 146,000 | 173,000 | 140,000 |
| 164,000 | 145,000 | 138,000 |

Solution:
- Maximum Value = $173,000
- Minimum value = $126,000
- Range = $47,000
- Standard deviation = $13,250.74
- Variance = 175,582,110.55 squared dollars
- Coefficient of variation = 8.87%

D. Summary

In this section, several measures which may be used to summarize a set of data values have been discussed. These measures do not require the listing of every data value as do frequency distributions; however, they do not provide as much information as do frequency distributions. Two types of summary measures were discussed - measures of central tendency and measures of dispersion. The former group includes the arithmetic mean, the median, and the mode. These help to locate the centre of the distribution. Each of these has its strengths and weaknesses, and each may be appropriate or inappropriate for any given problem. It is important that as many of these measures as necessary be provided so that adequate information about the distribution is provided.

The measures of dispersion include the maximum, minimum, range, standard deviation, variance, and coefficient of variation. These measures indicate how spread out or dispersed the distribution is. The most commonly used measure is the standard deviation which provides information on how dispersed the values are around the mean of the distribution. To compare the dispersion for two sets of data values, the coefficient of variation should be used.

III. REVIEW OF DEFINITIONS AND TERMS

The reader should be able to define and use the following terms as presented in this chapter:

**Absolute frequency**
- number of times each data value occurs within a distribution

**Arithmetic mean**
- a measure of central tendency, given by the sum of all the numbers under consideration, divided by the number of numbers. The mean is denoted by \( \mu \) and is also referred to as the expected value

**Average deviation**
- the arithmetic mean of the absolute deviations of a set of numbers from a measure of central tendency such as the median
**Bar graph**
a graph of the relationship between the data values for ungrouped data and their frequencies for grouped data. Same as a histogram

**Bimodal**
refers to a distribution with two modes

**Close-ended group**
a group for which both bounds are specified

**Coefficient of variation**
a measure designed to allow comparisons between the standard deviations of two data sets

**Continuous variable**
variable which can theoretically assume any value between two other given values

**Discrete variable**
variable which can assume only a limited number of values

**Distribution**
shorthand term for frequency distribution

**Expected Value**
see Arithmetic mean

**Frequency distribution**
a listing of data values accompanied by the number of times each value occurs

**Group midpoint**
the midpoint of a group when data has been classified in groups

**Grouped data**
refers to data which has been classified into groups

**Histogram**
see bar graph

**Line graph**
a graph of the relationship between the data values and their frequencies for ungrouped data

**Maximum**
a measure of dispersion given by the largest value in the distribution

**Measure of central tendency**
a measure that can be used as typical value or middle value

**Measure of dispersion**
a measure which indicates how spread out a distribution is

**Median**
measure of central tendency given by the middle value or the mean of the two middle values when the data are arranged in ascending (or descending) order

**Minimum**
a measure of dispersion given by the smallest value in the distribution

**Mode**
a measure of central tendency given by the most frequent value. If all values occur with equal frequency, there is no mode

**Open-ended group**
a data group for which either the upper or lower bound is unspecified

**Percentage**
a statistic used to facilitate data comparisons

**Range**
a measure of dispersion given by the maximum minus the minimum
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative frequency</td>
<td>number of times each data value occurs within a distribution, divided by the total number of data values</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>measure of dispersion of a distribution around its mean</td>
</tr>
<tr>
<td>Statistics</td>
<td>collection, assembly, classification, summarization, presentation, and analysis of data. Also refers to summary measures.</td>
</tr>
<tr>
<td>Summary measure</td>
<td>a single number or measure which summarizes a distribution without listing every value</td>
</tr>
<tr>
<td>Summation sign</td>
<td>mathematical symbol to denote the sum of several numbers: ( \sum )</td>
</tr>
<tr>
<td>Unimodal</td>
<td>refers to a distribution with only one mode</td>
</tr>
<tr>
<td>Variable</td>
<td>symbol which can take on any of a predetermined set of values</td>
</tr>
<tr>
<td>Variable name</td>
<td>name used to refer to a variable</td>
</tr>
<tr>
<td>Variance</td>
<td>a measure of dispersion of a distribution around its mean; the square of the standard deviation</td>
</tr>
<tr>
<td>Weighted arithmetic mean</td>
<td>an arithmetic mean where each of the values is assigned a weight indicating its importance or frequency</td>
</tr>
</tbody>
</table>
APPENDIX 1
Populations, Samples, and Standard Deviation

Statistical analysis may be performed on either population or sample data. A population is the entire universe or entire set of observations, while a sample represents a subset of the population. The population may change from one problem to the next. For example, the population may be all of the houses in Vancouver, or the population may be all of the houses in a particular neighbourhood in Vancouver, or the population may be all of the houses which sold in that neighbourhood in Vancouver in a given year. A sample of five houses might be drawn from any of these three populations.

The statistical measures in this chapter have focussed on population data. The equivalent measures for samples, called sample statistics, are related to population statistics but have some small differences. For example, in the standard deviation equation (Equation 12) in this chapter, it has been implicitly assumed that the data constituted a population rather than a sample. This equation would change slightly when calculating the standard deviation for a sample in that the denominator would be \( n - 1 \) rather than \( n \). This would make the denominator smaller, which ultimately results in a larger standard deviation measure, reflecting the increased potential for error when using a sample as compared to an entire population. Note that this difference will be greater in smaller samples and that in larger samples the magnitude of this adjustment will be negligible. In terms of notation, the standard deviation of the population is usually shown by the \( \sigma \) symbol (the small Greek letter sigma) while the standard deviation of a sample is shown by \( s \).

When the observations in a population are being analyzed, full information is available because all of the relevant observations are present. However, there are many cases when only a sample from a population is available for analysis. This may occur for several reasons.

1. It may be difficult and costly to define or find every observation in the population. For example, if the population consists of every housing unit in a large city, it could be difficult to find and list each item in the population.

2. Once the observations are defined, it may be expensive to obtain the necessary data for every observation in the population. For example, it would be expensive to determine the number of square feet in every housing unit in a large city.

3. It may not be feasible or reasonable to obtain the necessary data for every observation. For example, if the problem is to find out how long light bulbs last, it would not be reasonable to let every one burn until it burned out since no useful light bulbs would exist after all the data had been collected.

For these reasons, it is often necessary to deal with samples rather than populations.
SELECTED REFERENCES


The above are but three of a very large number of general statistic texts. Each covers the topics discussed in this book. The book below contains many examples for the layperson on the uses and abuses of statistics.