UBC SAUDER SCHOOL OF BUSINESS

LESSON 1

Statistical Foundations for Real Estate Analysis

Note: Selected readings can be found under "Online Readings" on your Course Resources webpage

Assigned Reading

1. UBC Real Estate Division. 2014. *BUSI 344 Course Workbook*. Vancouver: UBC Real Estate Division.
   Lesson 1: Statistical Foundations for Real Estate Analysis; Glossary (end of workbook)

Recommended Reading

1. UBC Real Estate Division. 2009. *Advanced Computer-Assisted Mass Appraisal*. Vancouver: UBC Real Estate Division.
   Chapter 1: Univariate Statistical Analysis (in particular, students should review the sections on inferential statistics and hypothesis testing for samples)
   Chapter 2: Multivariate Analysis

2. UBC Real Estate Division. 2009. *Foundations of Real Estate Mathematics*. Vancouver: UBC Real Estate Division.
   Chapter 11: Introduction to Statistics and Simple Data Description
   Chapter 12: Univariate Data Descriptive Measures
   Chapter 13: Introduction to Multivariate Data Analysis
   Chapter 16: An Introduction to Statistics used in Real Property

3. Huff, D. 1954. *How to Lie with Statistics*. New York: W.W. Norton.

4. Rumsey, D. 2007. *Intermediate Statistics for Dummies*. Wiley Publishing Inc.

Learning Objectives

After completing this lesson, the student should be able to:

1. define a "variable" and its properties;

2. do simple mathematical summations;

3. calculate and understand the difference between absolute and percentage changes in variables;

4. arrange data in groups and differentiate the features of grouped and ungrouped data;

5. determine the absolute and relative frequencies of data;

6. draw and interpret line graphs and histograms;

7. calculate and understand the differences between various measures of central tendency: the mean, the median, and the mode;

8. calculate and interpret the various measures of dispersion: the standard deviation, the variance, the coefficient of variation, coefficient of dispersion, and the range;

9. explain the term "correlation" and be able to describe a linear correlation coefficient; and

10. explain how to use a multivariate regression equation to obtain the predicted value of a dependent variable.

Instructor's Comments

This lesson introduces statistical analysis, setting the foundations for practical use of statistics in real estate applications. This is a quick overview of a subject many students find challenging. After this introductory lesson, we will proceed quickly to applications of statistical techniques. If you do not feel confident in your understanding of basic statistics upon completing this lesson, then you are advised to seek out further statistics help on your own. For example, you may wish to review the suggested readings above or find your own statistics help beyond the course. If you are unsure of the basic statistics presented here, then you will have difficulty effectively applying them later, so you are advised to review this material in-depth until you are comfortable with it.

Introduction to Statistics and Simple Data Description

What are Statistics?

The word *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 measures (this will be defined below).

Statistics are used in one form or another by most people every day. Newspapers quote figures about rising house prices and inflation rates, commercials cite statistics about one group having fewer cavities, sports commentators talk about batting averages and weather reporters 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 real estate. 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. Statistics also provide the foundation for forecasting.

In this lesson, the emphasis is on some of the uses of statistics as exemplified above; however, the discussion will not be exhaustive of all aspects of statistics. In particular, the goals of this lesson are two-fold:

1. to enable the reader to read and understand material employing simple statistics; and
2. to enable the reader to use simple statistical measures and techniques to describe and analyze data.

To accomplish this, the lesson will focus on descriptive statistics – that is, the process of describing and summarizing data. Interested readers may consult any of the statistics texts contained in the references to study in more detail such topics as probability and inferential statistics.[1]

[1] Inferential statistics refers to that branch of statistics focusing on the drawing of reasoned conclusions from data.

It is often stated that statistics can be employed to prove any point because numbers and data are so easily manipulated. This statement may be true if the reader has no familiarity with statistics and/or data analysis and is therefore unable to fully understand what is being presented. But this is the fault of either the reader or the user rather than the statistics. For example, consider a claim that the crime rate increased by 100% in a particular neighbourhood over some period of time. Note that the situation where the rate increased from 1,000 crimes to 2,000 crimes is very different from the situation where the rate increased from one to two, although the percentage increase is 100% in both cases. The book, *How to Lie with Statistics* (Huff, 1954), provides many other excellent examples of how statistics may be misused. What is important in such cases is that the reader have some background so that the best questions may be asked and the appropriate conclusions reached. The material discussed here should aid the reader in this task.

It's all Greek to me!

Many mathematical symbols come from the Greek alphabet. In fact, the word alphabet comes from the first two letters of the Greek alphabet…alpha (α) and beta (ß). Some of the most common Greek letters used in mathematics in BUSI 344 are listed below:

Upper case sigma	Σ	represents the sum of a set of numbers (e.g., $X_1 + X_2 + X_3$)
Lower case mu	μ	the mean or the arithmetical average
Lower case sigma	σ	the standard deviation – a measure of how spread out the data is
	σ^2	the variance – the standard deviation squared
Lower case chi	χ^2	"chi-squared" is used in a number of statistical tests

Definitions and Simple Mathematics

Before proceeding, some definitions and mathematical expressions need to be introduced. While the presentation in this lesson 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 "GENDER" 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 cannot 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 "GENDER" 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 any 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. Generally, discrete variables do 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 in the course.

Variables are frequently denoted with *subscripts*; for example, X_i where the subscript "i" takes on the values of, say, 1, 2, 3 and 4". 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_1 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

$$\mathbf{TOTAL = X_1 + X_2 + X_3 + X_4} \qquad \textbf{(Equation 1.1)}$$

A shorthand version would use the *summation sign* (the capital Greek letter sigma Σ), as follows:

$$\sum_{i=1}^{n} \mathbf{X_i = X_1 + X_2 + X_3 + \ldots + X_n} \qquad \textbf{(Equation 1.2)}$$

which says that the value of X_i for each value of i ranging from 1 to n should be added together, where n is a positive integer[2] greater than or equal to 1. In many cases, the subscripts are dropped if no confusion arises as to which index is being used. For example,

$$\sum_{i=1}^{m} X_i \quad \text{is often written simply as } \Sigma X$$

Throughout this course, 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 course will employ housing unit data which might be obtained from real estate boards, Multiple Listing Service® systems, 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 in which the numerator is the change in the value from the base to the new value, and the denominator is the base value. The formula to be used is as follows:

$$\text{Percentage change} = 100 \times \frac{\text{Final Value} - \text{Base Value}}{\text{Base Value}} \qquad \textbf{(Equation 1.3)}$$

It is important to note for absolute and percentage changes that 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 large. Without information on the size of the base, the percentage change cannot be fully understood – recall the crime rate example presented earlier. Thus, in the event that the percentage measure would be misunderstood, both the base and the percentage change should be reported. In the above two examples, the $15,000 change would represent a 75% increase over $20,000 and a 5% increase over $300,000.

[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.

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 1.1

For a given month in 2006, 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 1.1: Hypothetical Housing Prices

$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

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 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 2006 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, especially 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 1.1, the following frequency distribution may be constructed:

Table 1.2: Absolute Frequency Distribution of Housing Prices

Data Value	Absolute Frequency
$34,000	1
50,000	2
65,000	1
66,000	4
69,000	1
71,000	3
79,000	1
80,000	2
81,000	1
85,000	2
95,000	2
99,000	1
100,000	1
110,000	2
156,000	1

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.[3] With the 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 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 should 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.

For the data from Illustration 1.1, one might choose groups 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 1.3: Frequency Distribution for Grouped Housing Prices

Group	Group Frequency	Midpoint
$30,000.00 - 39,999.99	1	35,000
$40,000.00 - 49,999.99	0	45,000
$50,000.00 - 59,999.99	2	55,000
$60,000.00 - 69,999.99	6	65,000
$70,000.00 - 79,999.99	4	75,000
$80,000.00 - 89,999.99	5	85,000
$90,000.00 - 99,999.99	3	95,000
$100,000.00 - 109,999.99	1	105,000
$110,000.00 - 119,999.99	2	115,000
$120,000.00 - 129,999.99	0	125,000
$130,000.00 - 139,999.99	0	135,000
$140,000.00 - 149,999.99	0	145,000
$150,000.00 - 159,999.99	1	155,000

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 1.4.

Table 1.4: Frequency Distribution for Grouped Housing Prices

Group	Group Frequency	Midpoint
$30,000.00 - 49,999.99	1	40,000
$50,000.00 - 69,999.99	8	60,000
$70,000.00 - 89,999.99	9	80,000
$90,000.00 - 109,999.99	4	100,000
$110,000.00 - 129,999.99	2	120,000
$130,000.00 - 149,999.99	0	140,000
$150,000.00 - 169,999.99	1	160,000

[3] The term "range" is also used for width in this context.

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 data items. Groups with a frequency of zero (such as $130,000 to $149,999.99) can be deleted.

Using the data from Illustration 1.1, 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 1.1, 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, $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 will always sum to one.

Illustration 1.2

Using the data from Illustration 1.1 and the groupings from Table 1.4, relative frequencies may be calculated.

Table 1.5: Absolute and Relative Frequency for Grouped Housing Prices

Group	Absolute Group Frequency	Relative Group Frequency
$ 30,000.00 - 49,999.99	1	1/25 = .04
$ 50,000.00 - 69,999.99	8	8/25 = .32
$ 70,000.00 - 89,999.99	9	9/25 = .36
$ 90,000.00 - 109,999.99	4	4/25 = .16
$110,000.00 - 129,999.99	2	2/25 = .08
$130,000.00 - 149,999.99	0	0/25 = .00
$150,000.00 - 169,999.99	1	1/25 = .04
Total Observations	25	1.00

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.

Lists, groupings, and frequency distributions can all be used to help you find patterns in the data during the initial stages of data analysis. Computer software tools such as Microsoft Excel, SPSS, and NCSS have many options for manipulating data and producing these types of reports.

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 1.3

Using the data from Illustration 1.1, two frequency distributions may be derived. Figure 1.1 shows a line graph with absolute frequencies. A line graph with relative frequencies is presented in Figure 1.2.

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.

Figure 1.1: Line Graph for Absolute Frequencies of Ungrouped Data on Hypothetical Sale Prices

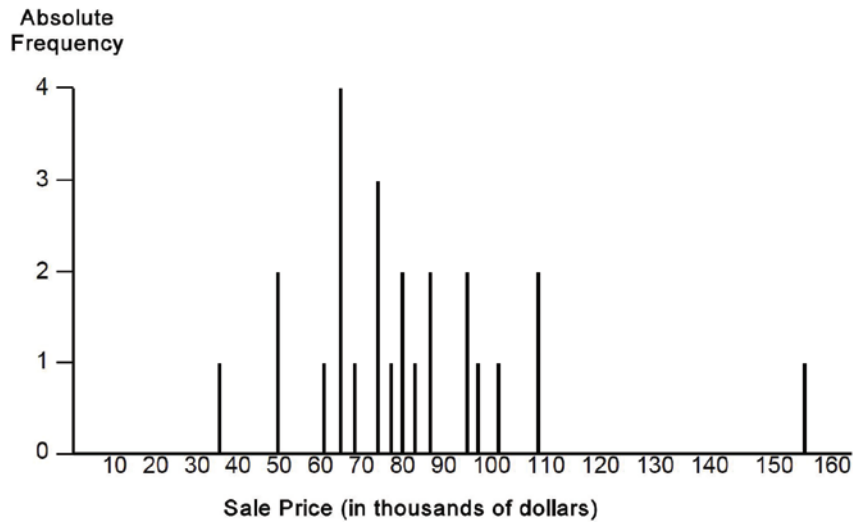


Figure 1.2: Line Graph for Relative Frequencies of Ungrouped Data on Hypothetical Sale Prices

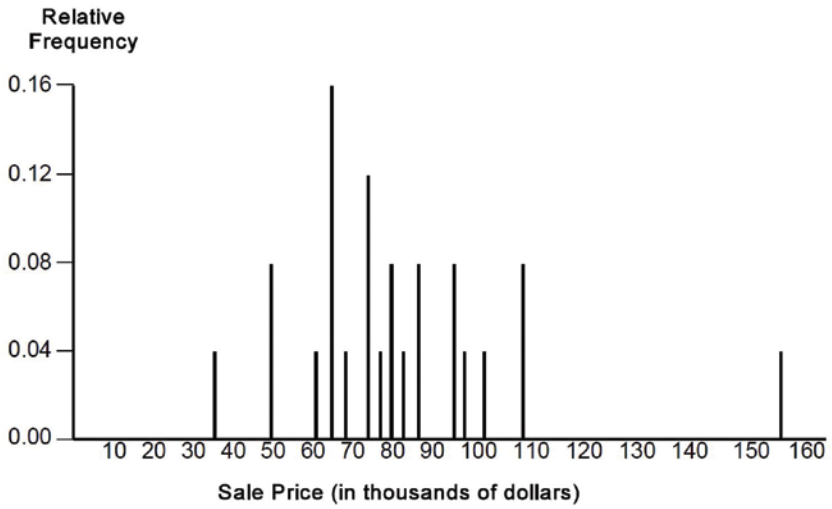


Illustration 1.4

Using the frequency distributions from Illustration 1.2, 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 1.3 presents the histogram for absolute frequencies, while Figure 1.4 presents the histogram for relative frequencies. Figure 1.5 presents a histogram showing both absolute and relative frequencies.

Figure 1.3: Histogram of Absolute Frequencies for Grouped Data on Hypothetical Sale Prices

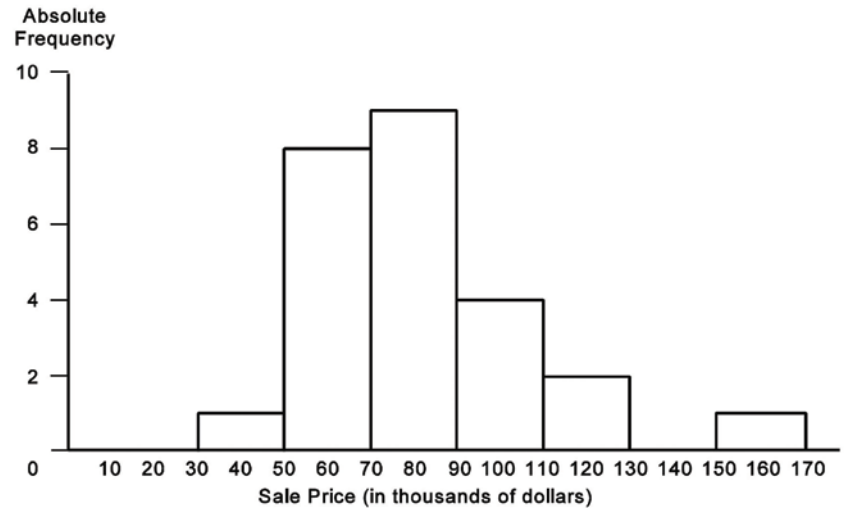


Figure 1.4: Histogram of Relative Frequencies for Grouped Data on Hypothetical Sale Prices

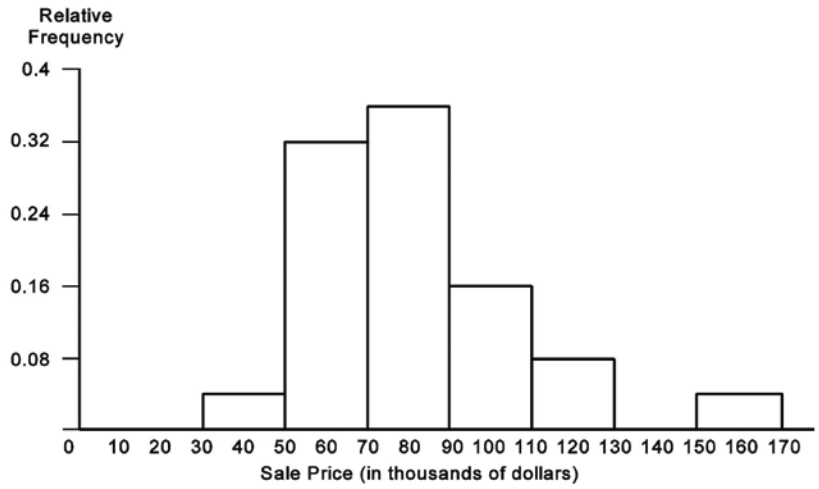
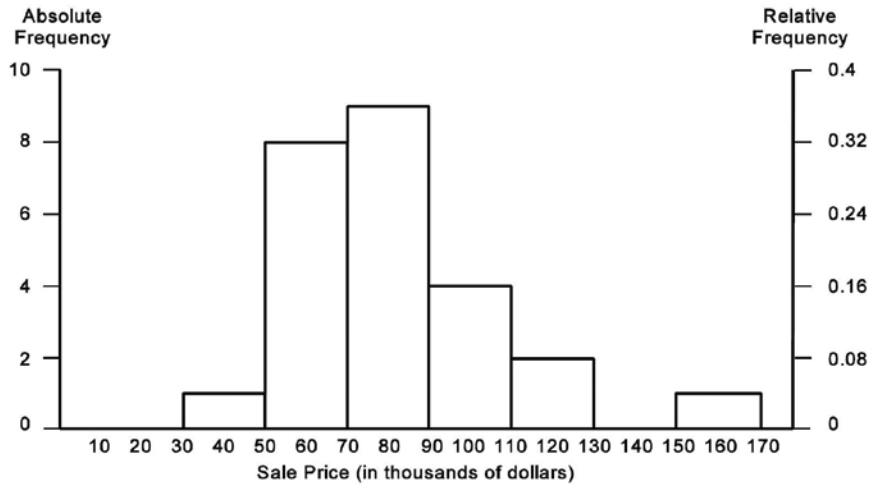


Figure 1.5: Histogram of Absolute and Relative Frequencies for Grouped Data on Hypothetical Sale Prices



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 1.5

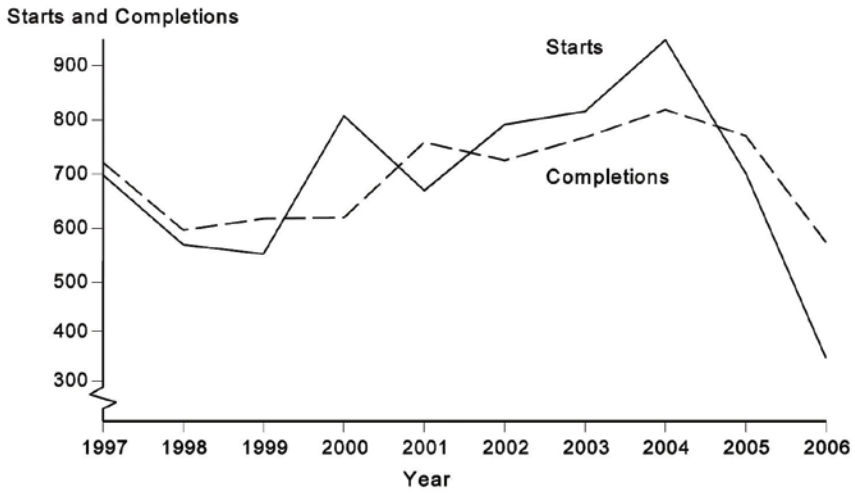
The following is hypothetical data on single family housing starts and completions in a large city:

Table 1.6: Housing Starts and Completions, 1997-2006

Year	Starts	Completions
1997	699	722
1998	570	597
1999	553	618
2000	808	620
2001	670	759
2002	792	726
2003	816	768
2004	948	819
2005	702	771
2006	360	574

These data may be displayed graphically to increase the ability of the reader to understand them and make comparisons.

Figure 1.6: Housing Starts and Completions, 1997-2006



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, straight 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 Illustration 1.5, 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 1.6, where the section of the vertical axis from zero to 300 has been removed.

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 are not available should be distinguished, for example by using n/a for not available.
- Units of measurement should be clearly indicated.

Summary: Simple Data Description

This section presented some basic definitions and notations that are fundamental to the study of statistics. 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 are left ungrouped.

Univariate Data Descriptive Measures

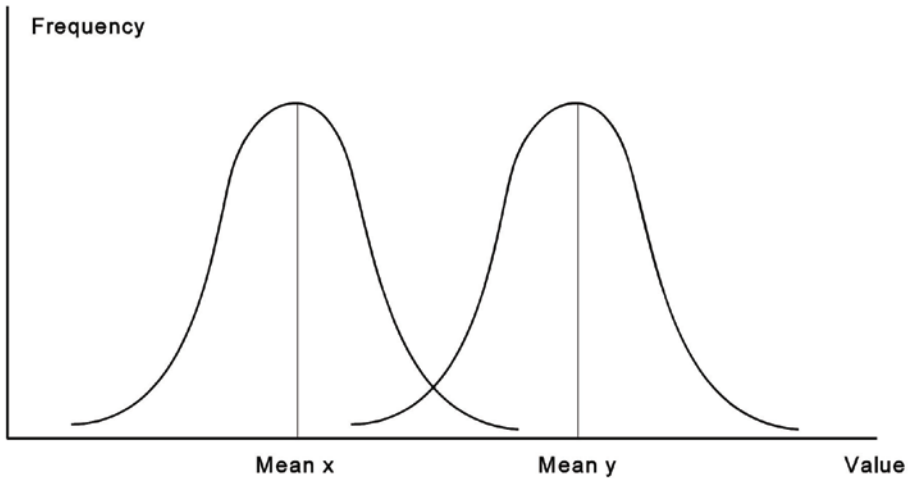
In the previous section, several techniques used to describe data were discussed. In particular, the importance of frequency distributions and histograms was discussed. These techniques provide information by listing every data value or by combining values into groups. These techniques are potentially 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.

Measures of Central Tendency

There are several measures of central tendency that are commonly used, and each of them has its own attributes. However, these measures are often incorrectly called "averages". The term "average" is imprecise and should be avoided as its use can be misleading. In place of this term, the precise names of the various measures of central tendency should be used.

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. The mean of the X distribution is less than the mean of the Y distribution which helps to determine the horizontal placement of the distributions.

Figure 1.7: Mean of X and Mean of Y



Arithmetic 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. The mean of the three numbers 2, 8, and 11 is $(2 + 8 + 11) \div 3 = 21 \div 3 = 7$.

Illustration 1.6

Using the data on housing prices below find the arithmetic mean:

$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.

$$\text{Mean} = \frac{\$2,000,000}{25} = \$80,000$$

Some notation can be used to simplify the calculations. The symbol μ (the small Greek letter mu) is generally used for the arithmetic mean, thus using Equation 1.2 we can use the following notation for the mean:

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \textbf{(Equation 1.4)}$$

This is a shorthand method for denoting the calculations involved in computing the arithmetic mean. This notation will be used repeatedly. Software tools such as MS-Excel, SPSS, and NCSS have built in functionality to calculate the arithmetic mean.

Illustration 1.7

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

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

$$n = 5$$

$$\mu = \frac{\sum_{i=1}^{5} x_1}{n} = \frac{\$95}{5} = \$19.00$$

Illustration 1.8

Using the data in the frequency distribution below, the arithmetic mean may be calculated using Equation 1.4.

Data Value	Frequency
$34,000	1
50,000	2
65,000	1
66,000	4
69,000	1
71,000	3
79,000	1
80,000	2
81,000	1
85,000	2
95,000	2
99,000	1
100,000	1
110,000	2
156,000	1
	25

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

$$= \$2,000,000$$

$$\mu = \frac{\$2,000,000}{25} = \$80,000$$

Weighted Mean

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:

$$\mu = \frac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i} \qquad \textbf{(Equation 1.5)}$$

where W_i is the weighting factor for the value X_i. The following illustration demonstrates the use of the formula.

Illustration 1.9

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 the quiz, and the mid-term examination is worth twice that of the quiz. It would make no sense to simply calculate the mean of the three marks (85 + 80 + 71 divided by 3) since they are weighted differently. Thus, the following table should be constructed:

Mark	Weight	(Weight)(Mark)
85	3	255
80	2	160
71	1	71
	Total 6	Total 486

The weights used in this illustration are derived from the statements in the problem about the relative importance of the exam types. Note that the weights are generally not of the form 1, 2 and 3, but rather are derived from the relative importance or frequencies.

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 1.5 from above:

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

Another illustration of a weighted mean can be demonstrated by how the grade for BUSI 344 is calculated. The multiple choice assignments are given a weight of 10% towards your final grade, Project 1 is given a weight of 15% towards your final grade, Project 2, 25% and the Final Exam accounts for 50% of your final grade. So, if you get 75% on each of your multiple choice assignments, 78% on Project 1, 84% on Project 2, and 82% on the exam, what is your final grade? This can be done reasonably quickly using an Excel table (below), where the column Weighted Grade is the Weight multiplied by the Grade and the TOTAL is simply the SUM of the Weighted Grades.

Course Component	Weight (w)	Grade (x)	Weighted Grade
Multiple Choice	10%	75	7.5
Project 1	15%	78	11.7
Project 2	25%	84	21.0
Final Exam	50%	82	41.0
TOTAL	**100%**		**81.2**

So, your final grade is 81.2%. Well done!

In terms of the arithmetic notation shown in Equation 1.5 this would be written as:

$$\mu = \frac{\sum_{i=1}^{4} w_i x_i}{\sum_{i=1}^{4} w_i} = \frac{(0.10)(75) + (0.15)(78) + (0.25)(84) + (0.50)(82)}{0.10 + 0.15 + 0.25 + 0.50} = \frac{81.2}{1} = 81.2$$

Note: recall from Page 1.3 that the i=1 at the bottom of the upper case sigma (the summation sign) means that the sum starts with the first value in the series and the 4 at the top of the summation sign means that the sum ends with the fourth value in the series. Since our two series (w and X) have only four values, that means we are using all of them. In the numerator (top term of the fraction) we take the product of each w_i times X_i and then sum them all up, the denominator (bottom term of the fraction) simply sums the four weights.

As practice examples for the weighted mean, try using two sets of grades for the Multiple Choice, Project 1, Project 2, and Final Exam marks respectively:

Set 1: 62%, 85%, 85%, and 90%
Set 2 100%, 80%, 80%, and 62%

Here are two tables for you to fill in

Set 1

Course Component	Weight (w)	Grade (X)	Weighted Grade
Multiple Choice	10%	62	
Project 1	15%	85	
Project 2	25%	85	
Final Exam	50%	90	
TOTAL	**100%**		

Set 2

Course Component	Weight (w)	Grade (X)	Weighted Grade
Multiple Choice	10%	100	
Project 1	15%	80	
Project 2	25%	80	
Final Exam	50%	62	
TOTAL	**100%**		

This will show that when the weight is low (as in the case of the Multiple Choice grades which are worth only 10% of the total grade), a low grade does not impact the final mark very much but when the weight is high (as in the case of the Final Exam being worth 50% of the total grade) a low grade can bring the overall score down quite a bit. The weighted means for the two sets of marks are 85.2 and 73.0 respectively.

The mean (arithmetic or weighted) 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 and it is the most universally understood. Typically, when the word "average" is used in day-to-day language it is the arithmetic mean that is being quoted.

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 are greater than the median and half the values are less than the median.

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

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

Illustration 1.10

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 1.11

Calculate the median for the housing prices given 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

Practice Example

Calculate the median for the following data items: 15, 25, 16, 19, 13, 17

Median = $\frac{16+17}{2}$ = 16.5

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 1.10, 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 be 5 even if the larger value (35) is 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. Finally, for many statistical procedures (which are beyond the scope of this course), the median is more complicated to use than the arithmetic mean.

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 1.16). 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 1.12

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 1.13

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

Illustration 1.14

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 1.15

Suppose you are asked to find the mode of the distribution of housing prices given in Illustration 1.11 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.

$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

Illustration 1.16

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 the 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.

Practice Examples

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

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

Mode = 15 rooms.

Find the mode for the following distribution of car colours in a given neighbourhood:

red, black, blue, green, red, black, blue, black, green, white, brown, black

The mode is black.

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

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

Modes = 53, 56, and 59.

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

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 1.16), because it may not exist (as in Illustration 1.13), and because there may be more than one mode (as in Illustration 1.14). It has the advantage that it can be calculated even if the data values are not numeric as in the car colour practice example above. In that example it is impossible to calculate 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 the individual extreme values although it may itself be an extreme value.

Illustration 1.17

For the data presented in the housing value practice example above (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 1.8 is a symmetric distribution, while the distribution in Figure 1.9 is said to be skewed to the right. The distribution in Figure 1.10 is said to be skewed to the left.

Figure 1.8: Symmetric Distribution

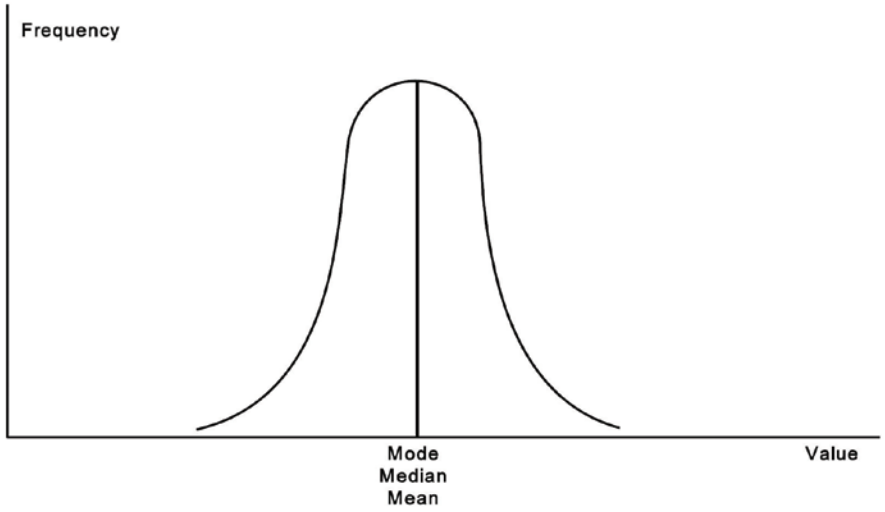


Figure 1.9: Non-Symmetric Distribution: Skewed to the Right

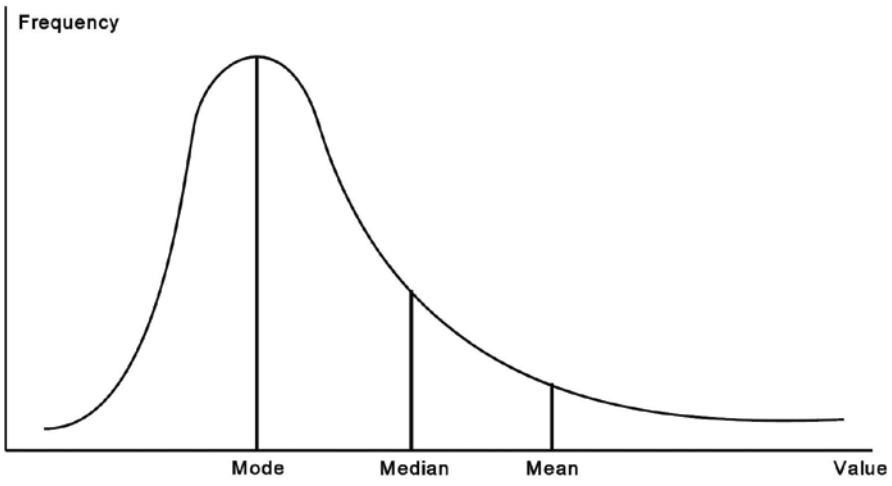
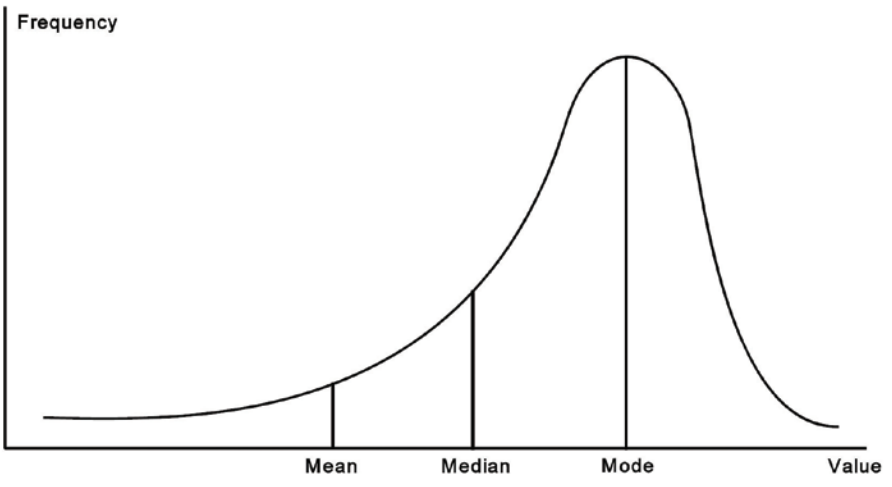


Figure 1.10: Non-Symmetric Distribution: Skewed to the Left



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 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.

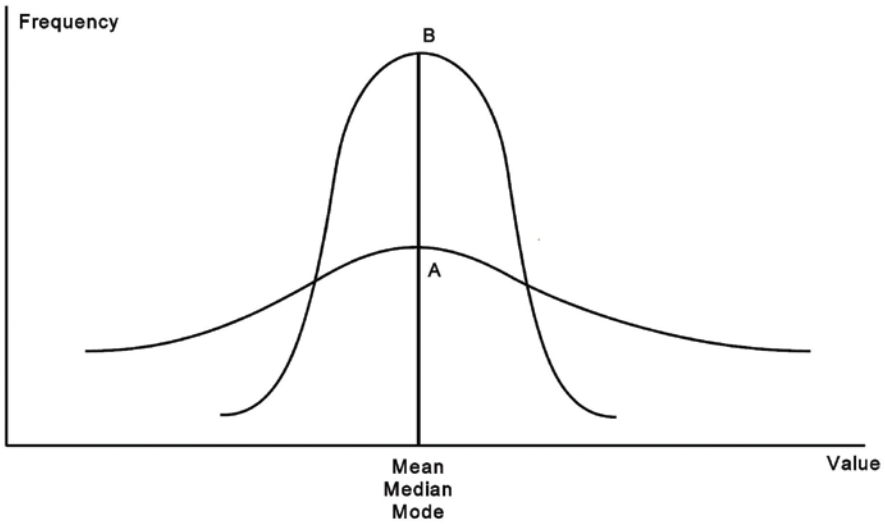
The choice between the mean, median, and mode should be made independently of whether the variable under consideration is discrete or continuous. For example, if the number of rooms in a set of houses (which is clearly discrete) is being analyzed, then the mode might seem appealing because it is a whole number – this would not in itself be a good reason to choose this measure.

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.

Measures of Dispersion

While the measures of central tendency place a distribution on the graph, they give no idea 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 Distribution B.

Figure 1.11: Distributions with Varying Dispersion



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.

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 greatest value in the distribution while the minimum is the least 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 1.18

For the data provided in Illustration 1.11 (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 1.19

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 1.18, but the distributions are quite different.

Practice Example

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

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

Standard Deviation

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

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}} \qquad \textbf{(Equation 1.6)}$$

where σ = standard deviation of the distribution
μ = arithmetic mean of the distribution
X_i = the values for which the standard deviation is being calculated
n = the number of data items
$\sqrt{}$ = symbol for the square root

To simplify this equation a little we will break it down into its components:

- since we are measuring dispersion the first thing we need to calculate is the difference between each data item and the mean;
- then all of these differences are squared (this simply turns any negative differences positive);
- then the squared differences are all summed and then divided by the number of data items – this gives us the mean of all the squared differences; and
- finally, since all the differences were squared initially, we need to take the square root to get the standard deviation.

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 1.11, the standard deviation of Distribution B is smaller than that of Distribution A (this particular comparison is valid only because the means are equal; more will be said about this below).

Illustration 1.20

Calculate the standard deviation for the following data values:

6, 10, 15, 9

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

X_i	μ	$X_i - \mu$	$(X_i - \mu)^2$
6	10	-4	16
10	10	0	0
15	10	5	25
9	10	-1	1

$$\sum_{i=1}^{4}(x_i - \mu)^2 = 16 + 0 + 25 + 1 = 42$$

$$\sigma = \sqrt{\frac{42}{4}} = \sqrt{10.5} = 3.24$$

Practice Example

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

$900	$950	$875	$975	$1,050
$1,000	$1,025	$1,100	$925	$850

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

Mean = $965
Standard Deviation = $75.99

The standard deviation is small relative to the mean, suggesting that the distribution is compact around the mean.

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 1.8), then approximately 68% of all 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.

The normal curve that is mentioned in the previous paragraph has some very special properties and has been the subject of much statistical research since the early 1900s. Suppose you examine all the final grades for BUSI 344 and suppose further that when you plot them on a histogram, the data forms a normal distribution, i.e., the data is spread out in a nice bell shaped curve as shown in Figure 1.8. If that normal distribution has a mean of 75 and a standard deviation of 6 then statistical research tells us that roughly 68% of all students' grades will fall between 69 and 81 (75 ± 6; one standard deviation away from the mean) and roughly 95% of all students' grades will fall between 63 and 87 (75 ± 12; two standard deviations away from the mean). This is discussed further in Lesson 2.

Standard Deviation (Sample)

The formula given in Equation 1.6 is for the Standard Deviation of a population. In general, most software products will calculate a Standard Deviation of a sample. The only difference is the divisor in the formula; for a population it is **n**, for a sample it is **n-1**. Thus the Standard Deviation for a sample will be slightly greater than the Standard Deviation for a matching set of data that represents the entire population.

$$\sigma_s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} \qquad \textbf{(Equation 1.7)}$$

In Illustration 1.20, if the four numbers represented a sample of a larger population, then the standard deviation would be:

$$\sigma_s = \sqrt{\frac{42}{3}} = \sqrt{14} = 3.74$$ [4]

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 instead of the standard deviation.

The variance, which is the square of the standard deviation, is measured in squared units. For example, if the 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. It is important to note the difference in units here – you cannot compare two numbers if their units are different. You can compare two variances if both of them are measured in dollars squared, but you cannot compare a variance measured in dollars squared (apartment rents) and a variance measured in feet squared (ceiling height).

[4] In this course, for the most part you will be dealing with samples although in some multiple choice questions the data for the population is given. If your answer does not match one of the options provided, then you should try the alternate divisor in the calculation. Note that SPSS produces the sample Standard Deviation as its default. In Excel, the function STDEV produces the sample Standard Deviation, the function STDEVP produces the Standard Deviation for the population.

Illustration 1.21

Using the data from Illustration 1.20, calculate the variance. This is simply the standard deviation squared.

$$\sigma^2 = (\sigma)^2 = (3.24)^2 = 10.50 \text{ dollars squared}$$

As mentioned previously, many software tools have built in functionality to calculate the standard deviation and variance.

The mathematical equation for the variance is simply Equation 1.6 or Equation 1.7 without the square root sign:

$$\boldsymbol{\sigma^2 s} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1} \qquad \textbf{(Equation 1.8)}$$

This equation may look more complicated than it is. The variance is calculated by finding the difference between each observation and the mean, squaring this difference, adding them together, and then dividing the sum by the number of observations.

Coefficient of Variation

Another measure of dispersion is the *coefficient of variation* (COV). Such a measure is necessary if the standard deviations for two variables are to be compared. For example, suppose that the standard deviation of prices of houses sold in Vancouver is $40,000 and 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 it is defined as follows:

$$\text{Coefficient of variation} = \frac{\sigma}{\mu}(100) \qquad \textbf{(Equation 1.9)}$$

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

Illustration 1.22

For a distribution of house prices, assume that the mean is $100,000 and the standard deviation is $40,000; for the distribution of incomes of the owners, assume that the mean is $20,000 and the standard deviation is $20,000. While the standard deviation is larger for the house 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:

$$\text{COV prices} = \frac{\$40,000}{\$100,000} \times 100 = 40\%$$

while for incomes, the coefficient of variation is:

$$\text{COV income} = \frac{\$20,000}{\$20,000} \times 100 = 100\%$$

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 1.23

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.

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

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

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; if you prefer risk, or want to maximize the expected gain, you would probably choose alternative A.

Practice Example

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

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%

The coefficient of variation provides a “relative” measure of dispersion. In essence, it shows how “big” or “small” the standard deviation is compared to the mean. This means that the coefficient of variation measures the size of the standard deviation “relative” to the size of the mean, effectively showing how tightly distributed a set of data is, or alternatively how widely dispersed it may be. As an illustration say you have two sets of data, one has a standard deviation of 50,000 and the other has a standard deviation of 5. Which distribution is more tightly spread around its mean? You cannot tell this from the standard deviation alone. But, if we measure the standard deviation relative to the mean we can determine which set of data has the “tightest” distribution. So, if the mean

of the first distribution is 125,000, the coefficient of variation is 50,000 divided by 125,000 or 40%, showing that the standard deviation is 40% of the mean. In the second distribution the mean is 8, so the coefficient of variation is 5 divided by 8 or 62.5%, showing that the standard deviation is 62.5% of the mean. This shows that the first distribution is less dispersed, with data more tightly packed around the mean of 125,000. The second distribution is more widely dispersed, with more "spread" between the data points and the mean of 8.

Coefficient of Dispersion

The last measure of dispersion to be discussed is the *coefficient of dispersion*. This is a measure of dispersion around the *median*; the three previous measures all dealt with dispersion around the mean. The coefficient of dispersion (COD) is to the median what the standard deviation is to the mean. Mathematically, the COD is the mean of the absolute deviations from the median all divided by the median. The formula for the COD is as follows:

$$COD = \left(\left[\sum_{i=1}^{n} ABS(X_i - X_{med})\right] \div n\right) \div X_{med} \qquad \textbf{(Equation 1.10)}$$

where COD = coefficient of dispersion
X_{med} = median of the distribution
X_i = the values for which the coefficient of dispersion is being calculated
n = the number of data items
ABS = the absolute value (difference from zero, ignore negative signs)

A quick comparison of the standard deviation formula in Equation 1.6 with the coefficient of dispersion formula above illuminates subtle differences. The standard deviation uses the differences between each data item and the mean and then squares those differences; the coefficient of dispersion uses the median and the absolute value of the differences to achieve its measure of dispersion.

As the dispersion or spread increases, the COD increases. Conversely, as the COD decreases, the dispersion decreases and the distribution is more compact. The COD is usually expressed as a percentage, representing a percentage of the median.

Illustration 1.24

Calculate the coefficient of dispersion for the following data values:

6, 10, 15, 9, 7

First, the median n must be determined: it is 9. Then, the coefficient of dispersion may be calculated by filling out the following table.

X_i	X_{med}	$X_i - X_{med}$	$ABS(X_i - X_{med})$
6	9	-3	3
10	9	1	1
15	9	6	6
9	9	0	0
7	9	-2	2

$$\sum_{i=1}^{5} ABS(x_i - x_{med}) = 3 + 1 + 6 + 0 + 2 = 12$$

$$COD = (12 \div 5) \div 9 = 0.267 \text{ or } 26.7\% \text{ (or rounded up to } 27\%)$$

Summary: Univariate Data Descriptive Measures

This section discussed several measures used to summarize a set of data values. 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, coefficient of variation, and coefficient of dispersion. 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; if the preferred measure of central tendency is the median, then the coefficient of dispersion should be used.

NOTE FROM THE TUTOR

Most of the statistics discussed on the previous pages will be used throughout the course, so it is important that you understand what they are measuring and how to interpret them. For the multiple choice assignments and projects, you will use computer software to calculate these statistics, but it will be up to you to use them in the correct context.

Multivariate Data Analysis Defined

In the preceding section, the discussion has been limited to the analysis of only one variable at a time. For example, the discussion focussed on the distributions of such variables as house prices, incomes and rents. In this section, the discussion moves to an examination of more than one variable at one time. In particular, various techniques will be examined which describe the relationship between two or more variables in what is called *multivariate data analysis*. For example, there is often an interest in the relationship between the price and size of housing units. This section will introduce techniques to examine such a relationship.

In the present context, the concern is not with *causal relationships* but rather with *statistical relationships*. A causal relationship is one in which there is an explicit cause and effect relationship between the two variables (e.g., generally the taller you are the more you will weigh). A statistical relationship simply indicates that there is some sort of consistent relationship between two variables without any presumed causal relationship.

Some of the techniques used to describe single variables will be generalized to handle two or more variables at one time. Techniques will also be presented which may be used to predict the value of one variable given the value of another variable.

Graphical Analysis of the Relationship Between Two Variables

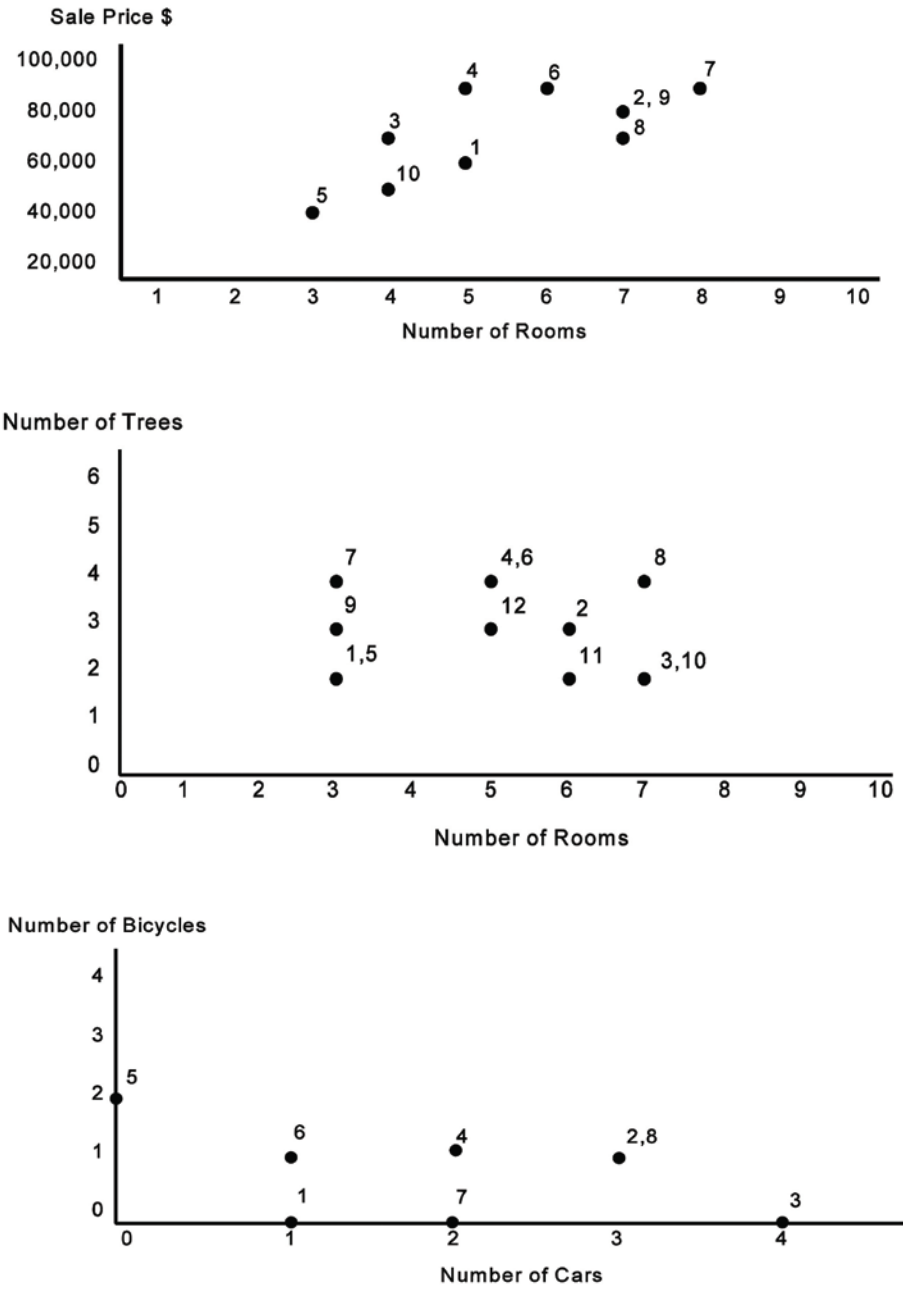
Given data on two variables, a set of points can be plotted on the graph. The points can then be examined to determine whether any type of relationship exists between the two variables. This type of graph is called a *scatter diagram* or *scatterplot*.

Consider Figure 1.12, which provides three graphs:

- sale price versus number of rooms;
- number of trees versus number of rooms; and
- number of cars versus number of bicycles.

The data points have been labelled in the graphs for the purpose of illustration, although this is not generally done, particularly when the number of data points is large.

Figure 1.12: Scatter Diagrams



The results:

- For sale price versus number of rooms, while there is no precise relationship between the two variables, the distribution of data points suggests that as the number of rooms increases, so does the sales price. This appears to be a sensible result.
- There appears to be no obvious relationship between the number of rooms and the number of trees. This seems intuitively obvious and the scatter diagram confirms this expectation.
- Finally, it appears that the number of bicycles decreases as the number of cars increases. This too seems intuitively obvious, and is confirmed by the scatter diagram.

The primary reason for graphing two variables is to aid in identifying and describing the statistical relationship between them. While such relationships may be discernible from the data itself, they may be more easily identified by examining a graph. However, as the number of observations increases, it becomes more time consuming to graph the data. Thus, other methods to identify and describe a relationship may be more appropriate, one of which is regression analysis.

Correlation Coefficient

The correlation coefficient, abbreviated r, is a measure that is designed to indicate the strength of the relationship between two variables. The linear correlation coefficient r, is a measure of the linear relationship between X and Y. It is a summary measure just as mean and standard deviation are summary measures. The correlation coefficient may be positive, negative, or zero.

- If the correlation coefficient is positive, as one variable increases (decreases), the other variable will increase (decrease) – that is, they move in the same direction.
- If the correlation coefficient is negative, as one variable increases (decreases), the other variable will decrease (increase) – that is, they move in opposite directions.
- If the correlation coefficient is zero (or close to zero), there is no relationship – that is, is no tendency for one variable to increase or decrease as the other changes.

The terms most commonly used to describe these situations are *positively correlated*, *negatively correlated*, and *uncorrelated*, respectively.

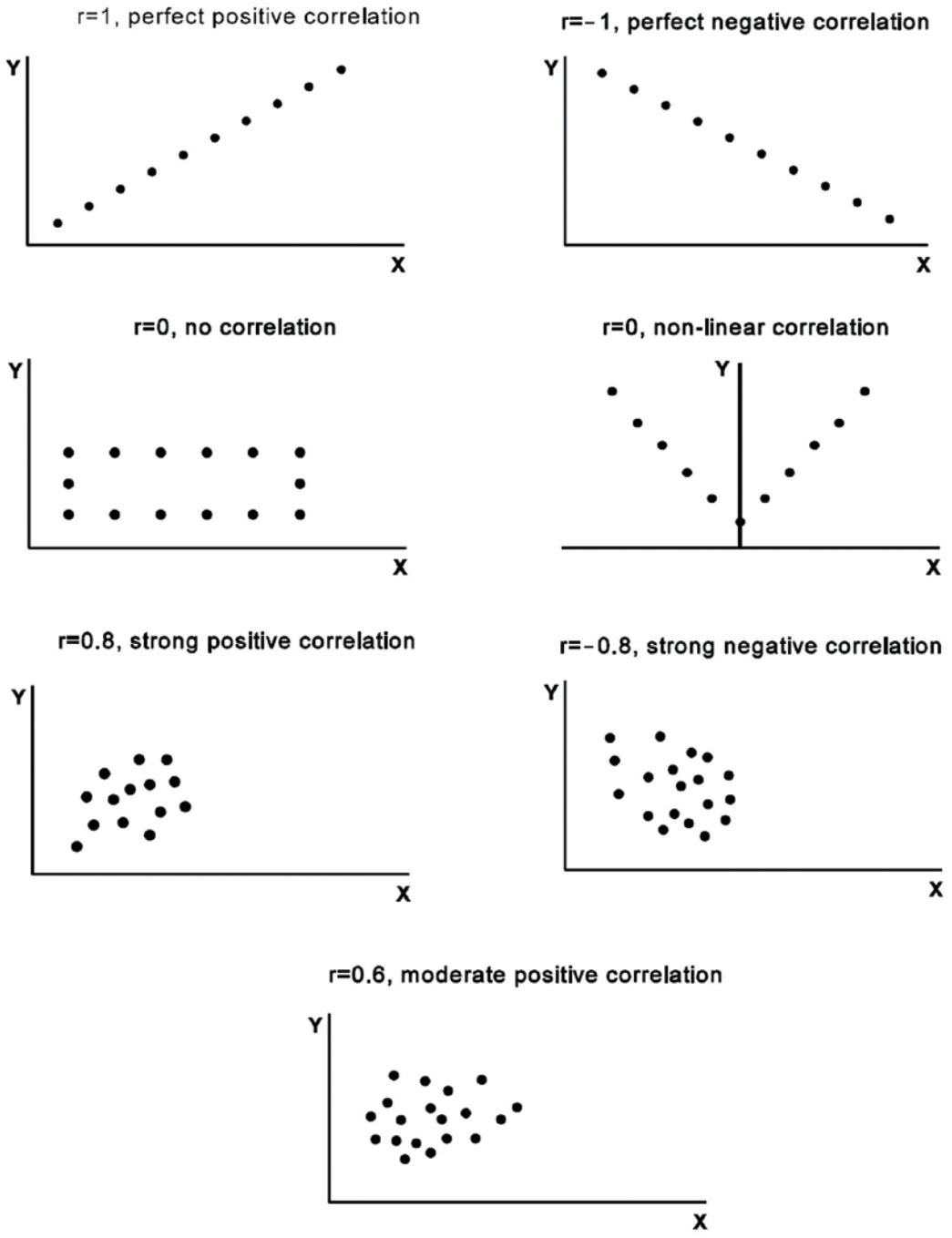
The mathematical definition of the correlation coefficient means that it takes on values only from -1.0 to +1.0. If the correlation coefficient equals +1.0, then the two variables are *perfectly positively correlated*. In this case, when the data pairs are graphed, they will all lie on straight line sloping upwards and to the right. Similarly, if the correlation coefficient is -1.0, then the two variables are said to be *perfectly negatively correlated*. In this case, when the data pairs are graphed, they will all lie on a straight line sloping downwards and to the right. As the correlation coefficient gets close to +1.0 or -1.0, the points become more compactly distributed around a straight line. If the correlation coefficient is 0, then the two variables are said to be *uncorrelated*. There are numerous configurations which would yield a correlation coefficient of 0.

NOTE FROM THE TUTOR

The mathematical formula to calculate a correlation coefficient is quite complex and will not be reproduced here. Students in this course will always use computer software to calculate the correlation coefficients and will not be expected to calculate this statistic on the exam. If you are interested in more background on the correlation coefficient and its calculation, type correlation coefficient into your favourite internet search engine for a wide variety of information on this statistic.

Figure 1.13 illustrates a variety of correlation relationships within data, ranging from r = -1 (perfect negative correlation), to r = 0 (no correlation or non-linear correlation), and to r = +1 (perfect positive correlation). In examining these figures, note that higher correlation coefficients are related to data points more compactly distributed around a straight line.

Figure 1.13: Correlation Relationships



Keep in mind that a strong correlation does not mean there is a causal relationship between the variables. Consider the following example: studies have shown there is a large and positive correlation between the number of bars and the number of churches in certain towns. However, it is unlikely there is any causal relationship between the two variables. If the number of churches increases, the number of bars will not necessarily increase and vice versa. There is probably a high correlation between the number of bars and churches because they are both related to the population of the city. As the population increases, the number of bars increase and the number of churches increase.

Furthermore, $r = .5$ does not mean the strength of the relationship between X and Y is "halfway" between perfect correlation and no correlation. In general, $r > .8$ indicates a strong relationship, $.4 < r < .8$ indicates a moderate relationship, and $r < .4$ indicates a weak relationship.

Theory and Method of Regression Analysis

The objective of regression analysis in appraisal is to model the relationship between property characteristics and sale price, so that the sale price can be estimated from the property characteristics. For example, the relationship between housing size and sale price can be investigated from data on square feet of living area and sale price. Table 1.7 shows such data for twenty-five recently sold single-family residences, and Figure 1.14 illustrates the resulting scatter diagram. For ease of illustration, the horizontal line has been truncated at 1,000 square feet, because all twenty-five properties sold were larger than 1,000 square feet.

The data appear to have an upward sloping relationship. This makes intuitive sense – as square feet increases, so does price. We will add a "line of best fit", by "eyeballing" what appears to be the line that best indicates the relationship in this data.

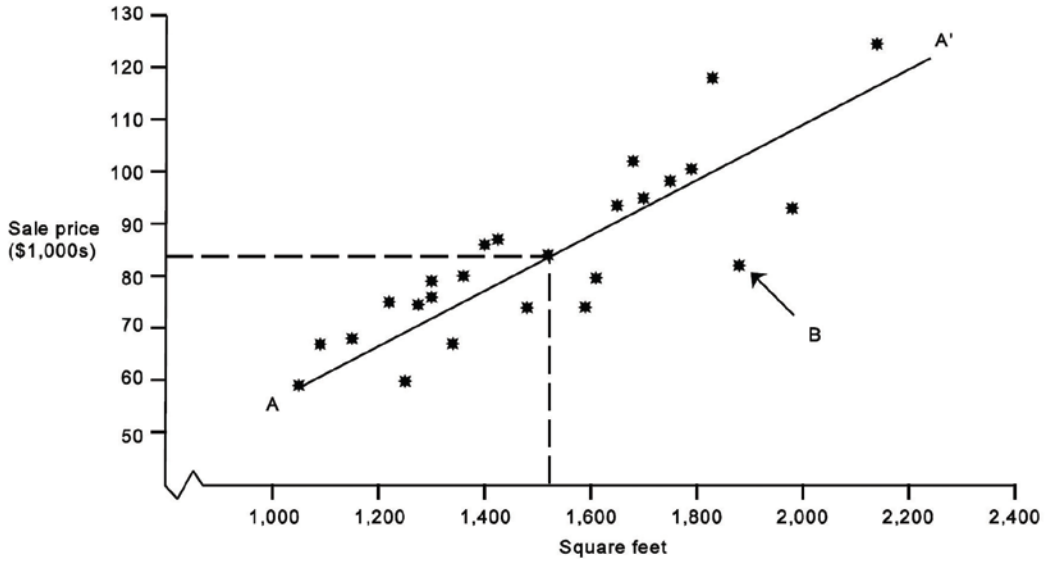
For the data in Table 1.7, the correlation coefficient between square feet and sale price has been calculated and is equal to 0.839 (the CORREL function in Microsoft Excel will perform this calculation). This indicates a strong positive linear relationship between square feet and sale price for this data.

We can use this line to estimate the sale price of an unsold property, by noting its square footage and reading the corresponding estimated sale price from the line. For example, to estimate the sale price of an unsold house with 1,500 square feet, draw a vertical line upward from 1,500 square feet to line A'. Then draw a second line horizontally from line A' to the vertical axis. This process is illustrated by the dashed lines in Figure 1.14. The estimated sale price of the house is approximately $84,000.

Table 1.7: Regression Data

Sale number	Square feet	Sale price
1	1,050	$59,000
2	1,090	66,900
3	1,150	68,000
4	1,220	75,000
5	1,250	59,800
6	1,275	74,500
7	1,300	79,000
8	1,300	75,900
9	1,340	67,000
10	1,360	80,000
11	1,400	86,000
12	1,425	87,000
13	1,480	73,900
14	1,520	84,000
15	1,590	74,000
16	1,610	79,600
17	1,650	93,500
18	1,680	102,000
19	1,700	94,900
20	1,750	98,200
21	1,790	100,500
22	1,830	118,000
23	1,880	82,000
24	1,980	93,000
25	2,140	124,500

Figure 1.14: Plot of Regression Data



Regression analysis is a more scientific, objective, and efficient method of fitting this line. It uses the principle that a straight line can be determined by one point and the slope. In fact, the regression equation for one independent variable:

$$s = b_0 + b_1 X_1 \qquad \textbf{(Equation 1.11)}$$

is simply the equation of a straight line, where b_1 is the slope and b_0 the point at which the line intersects the vertical axis. The slope of line A' in Figure 1.14 thus corresponds to b_1. The major difference is that the slope of A' was "eyeballed", whereas b_1 is calculated.

Consider point B in Figure 1.14, which corresponds to Sale 23 in Table 1.7. The property has 1,880 square feet and sold for $82,000. Based on line A', the estimated sale price is approximately $100,000. In statistical terms, this difference is called the amount of *error* (e_i) in the estimate, although calling this an error is misleading in that nothing has been done wrong! Regression analysis calculates b_0 and b_1 in a manner that minimizes the sum of squared differences between actual and predicted prices; that is, multiple regression analysis (MRA) minimizes:

$$\sum e_i^2 = \sum (S_i - \hat{S}_i)^2 \qquad \textbf{(Equation 1.12)}$$

where S_i is the actual sale price of property i and $\hat{S}_i$ is the estimated price of property i.

In the present example, the regression of sales prices on square feet (using the data in Table 1.7) produces the equation:

$$S_i = \$11{,}493 + \$47{,}90X_i \qquad \textbf{(Equation 1.13)}$$

where X_i is square feet of living area. That is, on average, sales prices increase at a rate of $47.90 per square foot of living area. Lesson 6 will explain in detail how to use computer software to produce regression lines such as this.

The regression coefficient, $b_1 = \$47.90$, is the slope of the regression line. The "constant", $b_0 = \$11{,}493$, is the value at which the regression line intersects the vertical axis when $X = 0$ (not shown in Figure 1.14). This equation minimizes the sum of the squared differences between actual and predicted prices (S_i and $\hat{S}_i$); any other equation would produce a larger sum of squared differences. In the example of an unsold house of 1,500 square feet, the regression estimated sale price is:

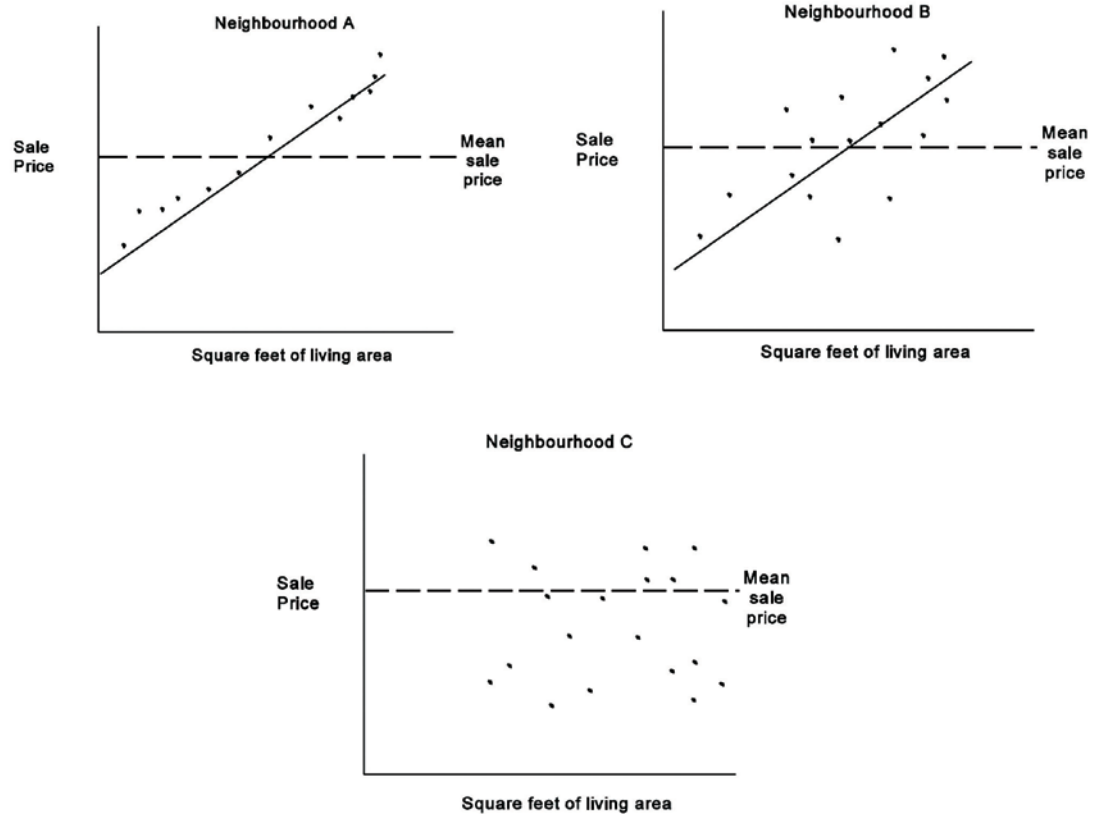
$$S_i = \$11{,}493 + \$47.90(1{,}500) = \$83{,}343$$

which agrees closely with the previous "eyeball" estimate.

How well does this equation estimate sale prices? Consider Figure 1.15, which illustrates a regression of sales prices on square feet of living area for three different neighbourhoods (in each case, models are developed from sales in only that neighbourhood). In plots for Neighbourhoods A and B, the regression line has the same slope and intersects the vertical axis at the same point. However, this equation does not estimate property sale prices with equal precision in all three.

- In Neighbourhood A, actual sales prices lie very close to the sale prices predicted by the regression line. In other words, the sum of squared differences, $\sum e_i^2$, is small, and it appears that we can be confident of the regression-estimated sale prices.
- In Neighbourhood B, actual sales prices are loosely fitted by the regression line.
- In Neighbourhood C, there is virtually no relationship between living area and sale price, making it impossible to draw the line of best fit. $\sum e_i^2$ is very large and the average sale price would produce almost equally good sale price estimates.

Figure 1.15: Comparison of Regression Results



We conclude that regression analysis is a useful predictor of property sale prices when $\sum e_i^2$ is small, but not when it is large.

One means of minimizing $\sum e_i^2$ is to use additional variables. In Figure 1.14, many points probably lie below A' because they represent properties with some negative features, such as poor physical condition. Other points probably lie above A' because they represent properties with generally positive features, such as good physical condition.

The model might be respecified, then, as:

$$S = b_0 + b_1X_1 + b_2X_2 \qquad \textbf{(Equation 1.14)}$$

where X_2 is a variable that represents physical condition relative to some norm, and may be either positive (better than normal condition) or negative (poorer than normal condition). Again, MRA would calculate the regression coefficients b_0, b_1, and b_2 in such a way as to minimize $\sum e_i^2$, where, in this case, the predicted values are a function of both living area and physical condition. Note that the importance of any one variable in the regression equation is directly related to its contribution in reducing $\sum e_i^2$.

Thus, MRA is a statistical technique for estimating unknown data on the basis of known and available data. In appraisal applications, the unknown data are market values. The known and available data are sales prices and property characteristics of both the sold and unsold properties.

Our example so far has been simplistic, using only one variable to predict sale price. A more realistic example would contain additional independent variables.

As a slightly more in-depth illustration, consider the equation:

$$S = \$7{,}800 + \$32.10X_1 - \$746X_2$$

where

X_1 is square feet of living area; and
X_2 is effective age.

In this case, b_0 is \$7,800, b_1 is \$32.10, and b_2 is -\$746. For a house with 2,000 square feet and an effective age of 5 years, the predicted value is:

$$\begin{aligned} \mathbf{S} &= \mathbf{\$7{,}800 + (\$32.10 \times 2{,}000) - (\$746 \times 5)} \\ &= \$7{,}800 + \$64{,}200 - \$3{,}730 \\ &= \$68{,}270 \end{aligned}$$

The regression coefficients for the variables are derived from analysis of sale prices and state each variable's contribution (or influence on) price.

Summary: Regression Analysis

This section introduced multivariate data analysis, providing a variety of techniques for examining the relationships between two or more variables. Our focus is on multiple regression analysis (MRA). We will continue our examination of MRA in Lessons 6 to 8, when illustrating mass appraisal model building.

Conclusion

This lesson covered an overview of the basic statistical skills required for real estate applications, including the terminology, techniques, and statistical measures used in carrying out an analysis. Statistical analysis has a variety of uses in real estate and can help describe data, measure the quality of work performed, and can also help to make inferences about expected sales prices of other properties. A single statistical measure alone may not provide adequate information about data. Therefore, in order to properly analyze data in applications in this course, an understanding of the significance of each statistical measure is required. By comparing several statistical measures against each other, and against known values, the strength of a statistic can be measured and more information about a data set can be determined. It should also be noted that the information provided by statistical analysis may only illustrate statistical relationships, and is not concerned with causality. A basic understanding of these tools and techniques is necessary for the use of more complex statistical software applications covered later in this course.

GLOSSARY

The terminology raised in this lesson and throughout the course can be reviewed in the glossary at the end of the workbook. As well, we have provided an index for easy future reference.

COURSE BULLETINS

Remember to check the Course Resources webpage for course bulletins and note any changes in your workbook and manual.

Review and Discussion Questions

Note: You can find suggested answers for Review and Discussion Questions under "Online Readings" on the Course Resources webpage.

1. In what situation would absolute frequencies be more useful than relative frequencies?

2. If your boss asked you to take a pay cut this year of 20%, promising an increase of 25% next year, would you take this deal? Discuss this situation with the other students on the BUSI 344 discussion forum webpage.

3. Imagine that you are reading a crime report that says that the murder rate in Small Town, BC increased by 100% compared to the previous year, while there was a 50% increase in Toronto, ON. Would you conclude that Small Town is becoming more dangerous than Toronto? Discuss this with other students and explain what additional information you would need to draw your conclusions.

4. An agent selling a property in Toronto states to a prospective purchaser that the average income in the area is $111,000, in order to entice the buyer to the high-class neighbourhood. The same agent then writes a letter to the municipal tax office in order to protest a proposed property tax increase for Toronto. In the letter, she states that the average income in the area is only $70,000. Discuss with other students how it is possible for the agent to state both figures without lying to one of the parties.

5. A condo developer commissions a study about average family size. The study reports that the average family size is 2.5 people so she builds 300 units with two bedrooms and a den. However, the units don't sell well because most of the buyers complain that the condos are either too small or too large for their needs. What went wrong?

6. Discuss with other students on the BUSI 344 forum what type of study you would commission if you were the developer in Question 5 above.

7. If the range for Distribution 1 is greater than the range for Distribution 2, what can you conclude about their respective standard deviation values? Can you conclude the same thing about their coefficient of variation?

8. Real estate investments are considered a good hedge against inflation. What does this tell you about the correlation between house prices and inflation?

9. If you read that college educated people are less likely to be married, should you forego a college education in the hopes of getting married? Discuss with other students on the BUSI 344 discussion forum the relationship between causation and correlation.

10. Some skeptics claim that Canada adopted the metric system to hide price inflation. How does a $0.20/litre increase in gas prices relate to a $0.20/gallon increase in percentage terms (assume there are 4 litres per gallon)? Similarly, why is smoked salmon often sold on the basis of price per 100 grams instead of price per kilogram?

ASSIGNMENT 1

LESSON 1: Statistical Foundations for Real Estate Analysis

The following questions should be submitted using the Real Estate Division's website *www.realestate.ubc.ca*. See "How to Submit Multiple Choice Assignments" in the Real Estate Division Student Handbook for more information.

TIP FROM THE TUTOR

The multiple choice "quizzes" are provided mainly to ensure you read the material carefully and to help ensure you keep on schedule. As such, they do not account for a large amount of the overall grade: approximately 1% per assignment or 0.05% per question. You should go through the questions carefully, but you should not be obsessing on the details. Assuming your time is limited, the projects are a better place for you to budget your time.

Marks: 1 mark per question.

THE NEXT TWO (2) QUESTIONS ARE BASED ON THE FOLLOWING INFORMATION:

Johnny, a real estate developer, has recently developed a large tract of land near Scotch Creek, BC. Sales have been slow thus far. Although there are 102 beautiful lots available for purchase, only ten have sold. The following are observations of the prices for the sales completed thus far:

Lot 1 (X_1)	=	$ 104,555	Lot 40 (X_6)	=	$ 210,445
Lot 5 (X_2)	=	$ 117,000	Lot 41 (X_7)	=	$ 100,000
Lot 21 (X_3)	=	$ 203,500	Lot 42 (X_8)	=	$ 105,100
Lot 22 (X_4)	=	$ 150,000	Lot 98 (X_9)	=	$ 112,500
Lot 35 (X_5)	=	$ 127,300	Lot 99 (X_{10})	=	$ 125,000

Based on this information, calculate the following sums:

1. $\sum_{i=1}^{6} X_i - \sum_{i=7}^{10} X_i$

(1) $ 575,055
(2) $ 470,200
(3) $ 445,055
(4) $ 917,655

2. $\frac{\sum_{i=1}^{5} X_i}{5}$

(1) $ 140,471
(2) $ 104,555
(3) $ 123,455
(4) $ 125,000

Assignment 1 continues on next page

3. The following are hypothetical data for residential real estate sales activity:

Year	Sales
1	13,850
2	14,458
3	14,008
4	15,839
5	16,445

Which one of the following statements is TRUE?

(1) The percentage change between year 1 and 2 is larger than between year 4 and 5, and the absolute change between year 1 and 2 is smaller than between year 4 and 5.
(2) The percentage change between year 1 and 2 is smaller than between year 4 and 5, and the absolute change between year 1 and 2 is larger than between year 4 and 5.
(3) The percentage change between year 1 and 2 is smaller than between year 4 and 5, and the absolute change between year 1 and 2 is smaller than between year 4 and 5.
(4) The percentage change between year 1 and 2 is larger than between year 4 and 5, and the absolute change between year 1 and 2 is larger than between year 4 and 5.

THE NEXT TWO (2) QUESTIONS ARE BASED ON THE FOLLOWING INFORMATION:

The following represents hypothetical data on the number of days between the listing date and the sale date for 50 single family units:

37	32	42	26	42	49	25	34	35	32
49	37	22	38	41	40	29	36	29	33
36	40	34	39	38	31	37	38	32	38
28	44	38	42	31	31	38	39	42	31
39	28	31	25	42	37	31	42	47	31

4. What is the relative frequency of 42 as the number of days?

(1) 0.1
(2) 5
(3) 6
(4) 0.12

5. What is the relative frequency of the group 38-43 days?

(1) 0.24
(2) 0.34
(3) 0.36
(4) 0.40

Assignment 1 continues on next page

6. The relative frequency of a dataset group is determined to be .15 while the absolute frequency is 18 (for the same group). What is the total number of observations in the dataset?

(1) 100
(2) 120
(3) 90
(4) None of the above

7. Which of the above variables are discrete?

A. Year the house was built
B. Price of the house
C. Length of front yard
D. Number of Fireplaces

(1) A and D
(2) B and C
(3) D and B
(4) None of the above

8. When writing the following news headlines, the writers used statistics to make an assertion. In which of these cases was the variable measured a continuous variable?

A. "Real estate markets booming: the number of sales triples"
B. "Americans gaining weight every year: obesity has doubled among adults"
C. "Corporate profits going down"
D. "Mad cow disease reduces the number of animals in continental Europe"
E. "Salaries of real estate salespeople greater than ever"

(1) A, B, C, and E
(2) C and E
(3) A and D
(4) Only E

9. The following is a distribution of selling prices for 12 homes:

$ 100,000	$ 95,000	$ 100,000	$ 540,000
$ 155,000	$ 96,000	$ 155,000	$ 133,000
$ 120,000	$ 135,000	$ 80,000	$ 175,000

What is the standard deviation for the distribution of sale prices?

(1) $157,000
(2) $141,202
(3) $14,120,166,667
(4) $118,828

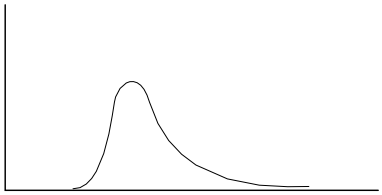
Assignment 1 continues on next page

10. For a given distribution and a given mean, an increase in the value of the standard deviation(s) will:

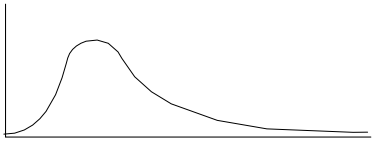
(1) decrease the representativeness of the arithmetic mean.
(2) have no impact on the representativeness of the arithmetic mean.
(3) improve the representativeness of the arithmetic mean.
(4) have an ambiguous impact on the representativeness of the arithmetic mean.

11. Which of the above distributions would have a median value that is less than the mean value?

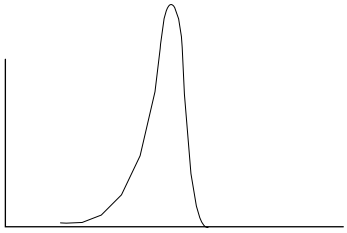
A.



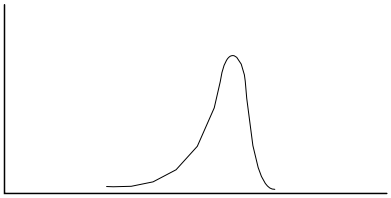
B.



C.



D.



(1) A and B
(2) B and D
(3) C only
(4) C and D

12. The standard deviation is:

(1) the difference between the maximum and minimum values in a distribution.
(2) the square of variance.
(3) a measure of dispersion of data around the mean of the distribution, measured in the same units as the original variable.
(4) a measure of dispersion of data around the mean of the distribution, measured in percentage terms.

Assignment 1 continues on next page

THE NEXT TWO (2) QUESTIONS RELATE TO THE FOLLOWING EQUATION FOR THE PRICE OF A HOUSE:

$$\text{PRICE} = 35{,}000 + 25{,}000 \times \text{Number of Rooms} + 7{,}000 \times \text{Number of Bathrooms}$$

13. A house with 5 rooms and 3 bathrooms would sell for:

(1) $149,000 more than a house with 2 rooms and 1 bathroom.
(2) $32,000 more than a house with 4 rooms and 2 bathrooms.
(3) the same as a house with 4 rooms and 4 bathrooms.
(4) $19,000 less than a house with 6 rooms and 1 bathroom.

14. An increase in the number of bathrooms from 2 to 6 would increase the sale price by:

(1) $ 7,000
(2) $35,000
(3) $28,000
(4) $42,000

THE NEXT TWO (2) QUESTIONS ARE BASED ON THE FOLLOWING INFORMATION:

Last year a small realty firm paid its four clerks $35,000 each, five junior agents $55,000 each, five senior agents $90,000 each, and the firm's managing partner $340,000.

15. What are the mean and median salaries paid at the firm?

(1) Mean: $80,333.33; Median $90,000.
(2) Mean: $55,000; Median $80,333.33.
(3) Mean: $80,333.33; Median $55,000.
(4) Mean: $90,000; Median $55,000.

16. Suppose the firm hires another managing partner. His salary is $340,000 per year. What effect will this have on the mean and the median salaries of the group, respectively?

(1) No change, no change
(2) Increase, no change
(3) No change, increase
(4) Increase, increase

17. The mean is:

(1) the calculated arithmetic average of the data values in a distribution.
(2) the middle data value in a distribution.
(3) a measure of dispersion in a dataset.
(4) the data value that occurs most frequently in a distribution.

Assignment 1 continues on next page

18. The following table provides information on the price and number of stories of eight apartment buildings:

X_i Stories	Y_i Sale Price
1	$1,000,000
3	$2,000,000
4	$4,000,000
6	$4,000,000
8	$5,000,000
9	$7,000,000
11	$8,000,000
14	$9,000,000

What is the linear correlation coefficient between X and Y?

(1) Close to 0
(2) Close to -1
(3) Close to 1
(4) Close to 0.5

PLEASE REFER TO THE INSTRUCTIONS FOR PROJECT 1 ON PAGES PROJECTS.1 THROUGH PROJECTS.10 TO ANSWER THE FOLLOWING TWO QUESTIONS.

19. Which of the following best describes the goal of Project 1?

(1) Apply mass appraisal techniques to specify an additive linear regression model that can be used to set condominium property tax assessments.
(2) Review a database of condominium sales, identify those that fit the market of University Gate, determine the best unit of comparison, and estimate sales prices for the University Gate project.
(3) Illustrate how SPSS is a mandatory tool for all statistical analysis in real property appraisal.
(4) Analyze model residuals for the University Gate condominium project and apply these towards an AVM for condominiums.

20. Project 1 requires you to analyze a database of comparable sales called “condosales”, which you will use in applying the direct comparison approach to value. An initial task is to explore the data. Which of the following is NOT a required step in the data exploration for this project?

(1) Apply graphical analysis to identify relationships between possible dependent variables and various independent variables.
(2) Identify variables that require transformation or recoding to be a more useful format.
(3) Review the data for any missing or incomplete information.
(4) Explore how basic multivariate regression may reveal linear or non-linear patterns in the data.

20 Total Marks

Assignment 1 continues on next page

Viewing Assignment Answer Guides

As soon as your assignment has been received and processed by the Real Estate Division, you can immediately download the answer guide. See your Student Handbook or visit your Course Resources webpage for more information on how to download assignment answer guides.

Planning Ahead

Project 1 requires you to carry out data analysis based on Lessons 1 to 5 with the objective of producing a table of prices for a 32-unit condominium complex.. You should read ahead to Project 1 so that you have a better idea of what is expected on this assignment. You may want to carry out some preliminary data investigation experimenting with the information presented in Lesson 1 now, while the materials are fresh in your mind, rather than waiting until Project 1 is due.

End of Assignment 1