Unit 1

Introduction to Statistics

Unit Objectives

Upon completion of this unit, you should be able to:

1. recognize basic statistical terminology;

2. effectively group and categorize data;

3. graphically display data and identify the shape of the graph; and

4. compute measures of center and variation.

Instructor's Notes

Introduction

Almost daily, we are bombarded by newspaper and television reports with the results of the latest surveys, polls, and scientific studies. For example, on December 10, 2001, the *Today* show on NBC informed viewers that researchers in Denmark published the results of a study finding there is no reliable evidence that regular mammograms reduce the risk of women dying from breast cancer. Are we to assume that this latest study suggests that mammograms are not a useful method of detecting and preventing breast cancer? If that is the case, what are women to do? How will they effectively protect themselves? Does this mean women no longer need mammograms?

The average American with no background in statistics must always rely on researchers, the government, and the media to honestly report and interpret the results of a given study. Most often, the reports are delivered in such a way as to suggest that the results of a given study are meaningful enough to cause a change in lifestyles, attitudes, or beliefs.

Ultimately, the primary goal of studying elementary statistics should be to gain a reasonable understanding of statistical concepts so that they may be applied to everyday life. By the end of this course, you will be able to read surveys and studies with a critical eye, rely on your own interpretation of the results, make your own judgments, and, finally, decide if the results are significant enough to warrant some type of change in lifestyle or beliefs.

The Nature of Statistics

In an attempt to get students to become an active participant in a statistics class, it is useful to ask the students to gather some data. This data will be used for various purposes throughout the term. The assignment for the first night might look like this:

> Conduct the following survey: collect data from 20 males and 20 females (18 years old or over). For each subject you will need to record the following: gender, age, height (in inches), arm span (in inches), length of the humerus bone (in inches), and handedness. The 'handedness' category is used to identify the subject as either left or right-handed. (Explicit directions are provided regarding the measurement of the humerus bone from the shoulder to the elbow.) Please organize your data using a table.

After the students gather this information, the individual results are compiled into a larger table. We can then proceed to discuss the type of information provided by the table. At first, we can see that the table provides only descriptive information about the subjects in our study. We can see how many males were surveyed, how many females are left handed, etc. Can we use this information to make conclusions about the subjects in the study?

Students quickly recognize that the data may be used to determine the average height, arm span, or age of males (or females) in the study. Likewise, we may wish to determine the proportion of males (or females) in the study who are left-handed. Answering questions about average values and proportions of subjects in a study fall under the label of **descriptive statistics**. In general, descriptive statistics refers to the methods of organizing and summarizing information from a set of data.

Once we recognize that we **can** make conclusions about the subjects in the given study it is reasonable to ask if we can extend these observations to the general population. In other words, suppose we find that the average height for women in our study is 5'5", can we then say that the average height for all women in America is 5'5"? This type of question falls under the label of **inferential statistics**. In general, we may think of inferential statistics as referring to the methods involved in using sample data to make generalizations about an entire population.

What other questions might we like to answer?

- If we can develop a formula to predict the arm span of a male given his height, then may we use the same formula for females?
- Does gender have any effect on handedness? i.e., are more men left-handed?
- Is age related to height?
- Can the length of the humerus bone predict height?
- Is there a relationship between arm span and height?

Many of these questions will be addressed throughout the rest of the study guide.

At this point it is important to call attention to the difference between a **sample** and a **population**. The term **population** refers to the collection of all individuals or measurements of observation or interest. While the term **sample** will be used in reference to a portion of the population from which information is collected.

There are many ways in which a population may be sampled. Our study will focus primarily on methods of collecting **random** or **simple random** samples – samples determined completely by chance. It is important to keep in mind that while a census will provide excellent information regarding the entire population, it is too costly and time consuming. Therefore we seek to gather a representative sample so that we might extend the results of research beyond the participants in a study. In general, the question of how large is "large enough" when determining sample size is closely related to the diversity of measurements within a group.

When referring to a population or sample, it is often important to give the **quantity** being measured or the **quality** being observed. For instance, suppose that the Department of Agriculture is doing a study of the weights of the potatoes in an experimental field in Aroostook County, Maine. In this example, the population is the weights of all of the potatoes in the field. It is **not** sufficient to say that the population consists of all potatoes in the field. We must also identify the quantity to be measured. Otherwise, we do not know if we are interested in values such as the diameter, length, weight, time to mature, or any other quantitative measurement of the potato. On the other hand, we might be interested in a **quality** of the potato such as color or taste. Therefore, it is often useful to identify the characteristic of interest.

Written Assignment

Reminder: these written assignments are for your benefit and are **NOT** to be turned in for a grade.

Do problems 1.2, 1.15, 1.23, 1.27, 1.29

Organizing Data

In a nutshell, statistics is the study of how to collect, organize, analyze, and interpret information from data. Our goal is to learn useful methods for organizing data once it has been collected.

Like mathematics, statistics can be thought of as a language with its own vocabulary, grammar, and syntax. In order to be proficient with organizing, displaying, and summarizing data it will be necessary to become familiar with many new vocabulary terms. Knowledge of such terms will aid us in the classification of the data, which is important when selecting the appropriate statistical methods for analyzing data.
In statistics, the word **variable** is used to identify a characteristic that varies from subject to subject. Every variable may be further classified as either **qualitative** or **quantitative**. Quantitative variables are numerical in nature (e.g., height, hours spent commuting to

work, etc.), while qualitative variables are non-numerical (e.g., eye color, gender, political affiliation, etc.). Each quantitative variable may be further classified as either **discrete** or **continuous**. Discrete variables are typically associated with counting (e.g., number of leaves on a tree), while continuous variables typically involve some sort of measurement (e.g., body temperature in °F). Formally, **data** can be thought of as the information collected when observing the values of a variable.

While qualitative variables are non-numerical, it is sometimes useful to "code" qualitative data using numbers. This is particularly useful if we have a large number of observations and would like to use a computer to tabulate, organize, or analyze the data. For example, suppose we were interested in determining if body temperature for females is different than that of males. Our study could include 100,000 subjects chosen at random where both body temperature and gender were recorded for each subject. Ideally, we would like to quickly determine the proportion of females and males in our study. To do so we can code the data by assigning a number to each category: 0 for "male" and 1 for "female." To determine the proportion of females in the study we merely ask a computer to quickly count the number of 1's. This example is very simplistic but the idea may be extended to more complicated techniques of analyzing data.

Now that we feel comfortable with the idea of the classification of data we may move on to the basics of organizing data.

Example: Suppose the results of the hypothetical first assignment include the following heights (in inches) for 20 women over the age of 18:

72	69	62	63	64	62	65	64	68	66
62	68	69	70	64	71	64	67	66	65

At first glance this list of numbers does not tell us much about the heights of the women in our study. We might recognize that it would be useful to list the numbers in either descending or ascending order so that we might get a sense of how high or how low the numbers go. Listing the numbers in ascending order we get the following:

62	62	62	63	64	64	64	64	65	65
66	66	67	68	68	69	69	70	71	72

Now we can easily notice several things:

- The lowest recorded height is 62 inches.
- The highest recorded height is 72 inches.
- The difference in heights between the tallest woman and smallest woman in this survey is 10 inches.
- The height that occurs most frequently is 64 inches.

The difference in height between the tallest woman and the shortest woman gives us an idea of the spread of our observations. However, this does not give us any sense of whether the heights are clumped around some central value or if they are evenly distributed between 62 inches and 72 inches.

If we were to combine the results of all surveys for every student in the class, we could easily have a list of heights for 500 women. As the number of data observations increases we would find that listing numbers in ascending order is quite tedious and time consuming. Furthermore, hundreds of observations will make it more difficult to determine the degree to which the data is spread out and to identify a number that occurs most frequently. It is clear that merely listing the numbers in increasing order can be somewhat limiting.

In this situation, it is useful to group our data. We might decide that we want to group our data in intervals, i.e., up to 63 inches, 64 to 66 inches, 67 to 69 inches, 70 to 72 inches, 73 inches and above. In this situation we can look at our data and see that there are four observations that fall into the category of "up to 63 inches," and so on. This grouping process will allow us to determine if there are any large clusters of values thereby getting a better sense of the spread of the data.

We may better see the results of grouping the data by making a graph or visual representation of the data. An **histogram** is a graph of the data displaying both of the groups (or classes) and the frequency with which the observations fall into the various groups. The method of sorting data into groups requires some forethought. Generally, when choosing the number of classes it is important to pick enough classes so that the grouping is worthwhile. Choosing too many classes can be as bad as choosing too few. Typically, anywhere from 5 to 15 classes are used. Most importantly, when constructing classes there should be no gaps or overlap. Think of the problems that could arise if a particular observation could not be classified because it could be considered a member of more than one group.

Reconsider the women's height example with the previously defined classes: up to 63 inches, 64 to 66 inches, 67 to 69 inches, 70 to 72 inches, 73 inches and above. If the original directions instruct the student to round each height to the nearest whole inch then this represents an acceptable division of classes. However, suppose instead that the instructions on the assignment ask the student to round each height to the nearest half-inch. Then the previously defined classes are no longer acceptable because there are gaps between the classes. What happens if a woman's height is measured to be 65.5 inches? To which class is this person assigned? In this case we would need to redefine the classes. Since we only have 20 observations then it seems reasonable to use 6 classes. The classes should have the same width – with the possible exception of the first and last class being defined similarly to the previous example when we want to include extreme values. An acceptable choice for the classes could be: $61 \leq 63$, $63 \leq 65$, $65 \leq 67$, $67 \leq 69$, $69 \leq 71$, and $70 \leq 73$. Where "$61 \leq 63$ inches" includes all values starting at 61 up to but not including 63. A similar symbol is used in the text. Using this method we can see that a measurement such as "63.5" would fall into the $63 \leq 65$ class.

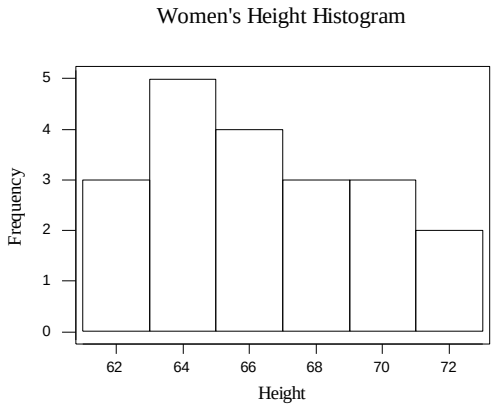
A frequency table is given below for our example. **Relative frequencies** are computed by dividing the frequency by the total number of observations. Relative frequencies should add up to 1. This may be off a little due to round-off error, but should never be off by more than one percent. An histogram may be constructed for both a frequency table and a relative frequency table. Both histograms will look virtually the same, the only difference being the scaling on the vertical axis.

Height	Frequency	Rel. Freq.
61≤ 63	3	0.15
63≤ 65	5	0.25
65≤ 67	4	0.20
67≤ 69	3	0.15
69≤ 71	3	0.15
71≤ 73	2	0.10
Totals	20	1.00

The **cumulative frequency** for a class is the sum of the frequencies for that class and all previous classes. Notice how the left column for height has changed.

Height	Cumulative Freq.
up to 63	3
up to 65	8
up to 67	12
up to 69	15
up to 71	18
up to 73	20

The frequency histogram associated with the frequency table above is shown below.



Data may be sorted by non-numerical classes as well. For example, if we had a list of numbers representing the scores on a test we could group them according to letter grade and construct the associated histogram.
Throughout this section we are mainly concerned with organizing our data and using histograms. There are many other types of graphs and visual displays. Although no

discussion will take place in this study guide, you should be familiar with stem-and-leaf plots, dot-plots, and pie charts.

We may use a table, graph, or formula to provide the values of the observations and the frequency in which they occur. Formally, this is called the **distribution** of the data set. When viewed graphically the shape of the distribution may take on one of many common forms. Names and pictures of these common distributions are given in your text. Be familiar with these names and shapes.

Written Assignment

Do problems 2.1, 2.7, 2.14, 2.17, 2.18, 2.27, 2.42, 2.43, 2.53, 2.58, 2.60, 2.90, 2.96, 2.97, 2.102

Descriptive Measures

We cannot predict **exactly** what a single member of a population will look like, how tall they will be, or how many times their heart will beat in sixty seconds. However, there are some **average** characteristics that an individual is likely to exhibit.

Consider the following statements:

- The average salary of a NBA player is $4.5 million.
- The manual transmission Toyota Corolla averages 41 miles per gallon on the highway.
- A survey of 560 adult women revealed average height to be 5'5".

In each of the previous statements, one number is used to describe a sample or population. In particular, each number is used to represent a central or typical value of the associated population or sample. Such a number is called a **measure of center**.

The most common measures of center are the mean, median, and mode. The **mode** is the value or observation that occurs most frequently. Note: it is possible for a data set to have multiple modes or no mode at all. The **median** is the central value or observation of an ordered list. In other words, the median will separate the bottom 50% of a data set from the top 50%. The **mean** is the average of all values in a data set. To construct the mean we merely add up all observations and divide by the total number of observations in the data set.

When computing the median it is important to first order the data set from smallest to largest. The technique of finding the middle depends upon the number of observations in the data set. If there are an odd number of observations then the median is precisely the middle number. If there are an even number of observations then the median is the mean (or average) of the two middle numbers.

Example: Reconsider the women's height survey. The ordered list is given below:

62	62	62	63	64	64	64	64	65	65
66	66	67	68	68	69	69	70	71	72

The mode is 64 because it occurs most frequently. The median is 65.5, the average of 65 and 66 since these are the middle two values. The mean is 66.05, the sum of all values divided by 20.

Suppose the survey of a second student in the class yielded the following results:

62	62	62	63	64	64	64	64	65	65
66	66	67	68	68	69	69	70	84	87

The mode and median remain the same at 64 and 65.5, respectively. However, the new mean is 67.45.

We can see that for these two data sets all but the last two values are the same. In the second set, there are two large values of 84 and 87 – representing two women that are exceptionally tall at 7 feet and 7 feet 3 inches, respectively. From this example we can see that the mode and the median are not affected by such extreme values. We call a measure with this property **resistant**. The mean is **sensitive** to extremely high or low values. Extreme values such as the 84 and 87 in the above data set are called **outliers**.

How do we know which measure of center to use?

If we are dealing with qualitative data then the only appropriate measure of center is the mode. If we are working with quantitative data then any of the three measures are appropriate, however some are more useful in certain situations. For example, if L.L. Bean is interested in determining which fly-rod is purchased most frequently, then they would compute the mode. This way they could decide which rod to carry in larger numbers. On the other hand, consider the year-end report of a Fortune 500 company. When reporting an "average" salary it would be most meaningful to use the median because the salaries of the top-level executives would pull up the mean, which is sensitive to outliers. However, if we are interested in reporting an "average" that is computed using all values of a data set then the mean is the only appropriate choice.

For our purposes in this class we will use the mean most frequently because it makes use of all data values and can be analyzed more conveniently using statistical methods.

A measure of center is a summary of a data set rolled up into one number. However, sometimes using one number to describe a set of data is not always meaningful.

Example: Consider the following two data sets:

Data Set #1: 100, 100, 100, 100, 100
Data Set #2: 90, 90, 100, 110, 110

Each data set contains 5 values and each data set has a mean of 100. However, when looking at the two sets we can see that they are very different. In the first set there is no variability whatsoever – all values are the same. In the second set, on average, the values are about 10 units away from the mean. It is clear that we can have two very different data sets with the same mean. Given this circumstance it seems reasonable to want to report a second number to describe the variability (or spread) of the data.

The easiest way to numerically represent the spread of the data is to compute the **range**, which is the difference between the highest and lowest value. Since this value pertains only to the highest and lowest values, it tells us nothing about the values in between.

Ideally, a measure of variation should describe how the values vary among themselves. In order to be consistent, we would like to see how the numbers vary in relation to a particular value, say the "center." In other words, we will try to determine how far each value is from the center (or the mean) of our distribution. If the majority of the data set lies close to the center, then there will be little variation. Therefore, the greater the spread of the data the greater the measure of variation. The **standard deviation** is a measure of variation that is computed by determining how far, on average, the observations are from the mean.

The method for computing standard deviation is not easy. However, most calculators and computers can compute it rather quickly. Your text carefully develops the formula and provides several examples on how to compute standard deviation.

Example: Returning briefly to our two data sets:

Data Set #1: 100, 100, 100, 100, 100 has a mean of 100 and a standard deviation of 0.
Data Set #2: 90, 90, 100, 110, 110 has a mean of 100 and a standard deviation of 10.

This new information about standard deviation tells us that there is no variability in the first set and that the values in the second set, on average, differ from the mean by about 10 units.

To compute both the mean and the standard deviation we need to know if the data set is from a population or a sample. Throughout this text it will be important to distinguish between a population and a sample and the associated notation.

Notation: the sample mean is denoted by $\bar{x}$, while the population mean is denoted by μ. Similarly, the sample standard deviation is denoted by s, while the population standard deviation is denoted by σ. Be sure to pay close attention to the difference in formulas for computing these values. In particular,

sample mean $\bar{x} = \dfrac{\sum x}{n}$ population mean $\mu = \dfrac{\sum x}{N}$

sample std. deviation $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ population std. deviation $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$

where n = size of sample and N = size of population.

Section 3.4 details a method in which five numbers (minimum, 1st quartile, median, 3rd quartile, and maximum) combined with a graphical display (a box-plot) may be used to describe the center and the measure of variation of a data set. The modified box-plot is useful because it identifies potential outliers by plotting them individually. Box-plots are useful when comparing two or more data sets – provided the same scale is used. It is useful to be familiar with this graphical display; however, we will not make much use of it throughout the course.

As previously stated, it is often important to differentiate between sample data and population data. Generally, we will use sample data but it is necessary that we become acquainted with the notation and terminology associated with population data. The sample mean is a descriptive measure of a sample. Likewise, the population mean is a descriptive measure of a population. A descriptive measure of a sample is called a **statistic** and a descriptive measure for a population is called a **parameter**. Thus, $\bar{x}$ and s are statistics, whereas μ and σ are parameters. Given the fact that we rarely know everything about a population we will often use statistics to estimate parameters. For example, we will use the sample mean $\bar{x}$ to estimate the population mean μ.

It is often common to compare observed values to each other by determining their relative standing. In other words, we can compare to values by determining how far each is from the mean. To do this we compute a **standardized** score called a *z*-score for bell-shaped distributions. The *z*-score is a value that represents the number of standard deviations a value is from the mean.

For a given value x, its *z*-score is computed using the formula $z = \frac{x - \mu}{\sigma}$.

Negative *z*-scores tell us that a value is less than the mean. Positive *z*-scores tell us that a value is greater than the mean.

Written Assignment

As a reminder, these assignments are designed as homework to allow you to apply the material learned from the readings. These are for your benefit and will **not** be turned in for a grade.

Do problems 3.1, 3.3, 3.5, 3.9, 3.26, 3.42, 3.45, 3.46, 3.53, 3.55, 3.85, 3.121-3.124

You are now ready to take the Unit 1 test. Please send in the test, with the attached assignment cover, as your assignment for Unit 1.

Unit 1 Test

Use only hand calculators on this exam; Minitab or any other computer software is **not** allowed. The textbook (but no other notes) may be used during the exam. Be sure to submit all your work in order to receive partial credit.

1. Identify each number as either continuous or discrete.
   a. The average speed of cars passing a speed trap on the Maine Turnpike between 3 PM and 6 PM on a given Monday.
   b. A census taker wants to know the number of Maine families with preteens.
   c. The temperature of the ocean at various depths.

2. For each of the following variables, indicate whether it is qualitative or quantitative.
   a. The admitting diagnosis of patients admitted to a mental health clinic.
   b. The weights of babies born in a hospital during a year.
   c. The class rank of high school seniors at the local high school.
   d. The religion of the sample subjects.

3. Identify each study as either experimental or observational.
   a. A sample of fish is taken from the Androscoggin River, in Maine, to measure the level of mercury in the fish.
   b. A research scientist gives a weight loss drug to a group of 500 patients and a placebo to another 500 patients to determine if the weight loss drug has an effect on the patients.

4. In response to a poll on a *Dateline* NBC program about wildlife conservation, 1276 of 1450 callers said they would be willing to spend more money on imported fossil fuels in order to eliminate the possibility of oil drilling in national parks set aside as wildlife preserves. NBC followed with the announcement that 88% of Americans are willing to spend money to protect wildlife preserves. Do you think that the group of people who responded is likely to be representative of all Americans? Explain your answer.

5. Construct a frequency table with 4 classes for the following data on the charge for monthly long distance phone bills for the last year: $17.06, $20.96, $25.97, $26.41, $22.02, $27.34, $18.67, $24.88, $24.07, $25.35, $23.39, $20.60.

6. Find the original data from the stem-and-leaf plot.

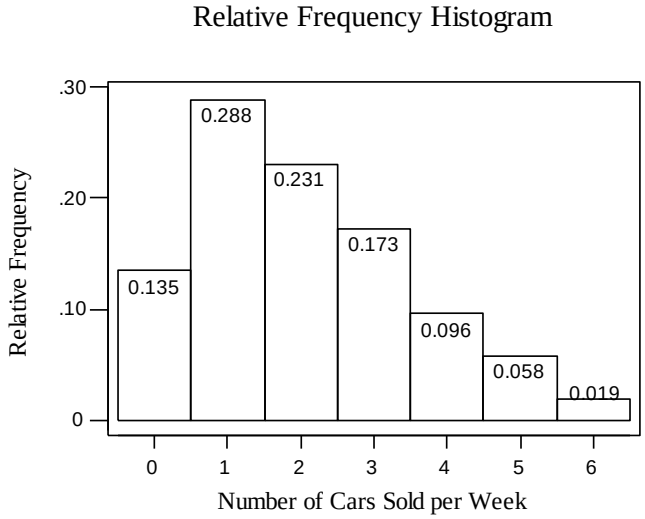
Stem	Leaves
6	4 7
7	1 1 8
8	4
9	0 3 8 6

7. The winners of the NCAA wrestling championships for the years 1968-1997 are given in the table below.

Champion	Frequency	Relative Frequency
Oklahoma State	5	
Iowa State	6	
Oklahoma	1	
Iowa	17	
Arizona State	1	
Total:		

   a. Compute relative frequencies for each class and fill in the appropriate column of the table. Round all relative frequencies to 3 decimal places.
   b. Draw the relative frequency histogram for the above table.

8. Consider the following relative frequency histogram displaying the number of cars sold per week last year for a given sales rep, Ronnie, at Emerson Toyota.



   Given that there are 52 weeks in a year, approximately how many times did Ronnie sell 2 cars per week? Round your answer to the nearest whole week.

9. The average retail price for bananas in 1994 was 46.0 cents per pound, as reported by the U.S. Department of Agriculture. A recent random sample of 8 supermarkets gave the following prices for bananas in cents per pound.

42	45	49	50	51	51	52	53

   a. Find the mean price for bananas per pound. Round to 3 decimal places.
   b. Find the median price for bananas per pound.
   c. Find the mode price for bananas per pound.

10. The number of absences for five children in a local kindergarten class is as follows.

Kid	# Absences x	x^2
Sophie	3	
Bert	7	49
Billy	6	
Joey	8	
Julie	2	
	$\sum x =$	$\sum x^2 =$

a. Complete the table and use the computing formula $s = \sqrt{\dfrac{n\sum x^2 - \left(\sum x\right)^2}{n(n-1)}}$ to find the standard deviation for the number of absences. Round to 3 decimal places. Note: if you would prefer to use the other formula for standard deviation then adjust the table accordingly.

b. What is the range for the number of absences?

c. Which appears to be a better measure of variation for this data set: range or standard deviation?

d. What is the variance?

11. True or False. For the data set {2,3,4,3,6,75}, the median is a better measure of center than the mean. Explain.

12. Suppose that the mean score on this test is 75.6 with a standard deviation of 7.8. Suppose also that the scores follow a bell-shaped distribution.

a. Convert a score of 80 to a z-score. Round your answer to 2 decimal places.

b. How many standard deviations is a score of 80 from the mean?