CHAPTER 1: IMPORTANT INITIAL CONCEPTS

Introduction to Statistics

"What is statistics?" The answer is rather simple yet very complex. Statistics is a science that involves collecting, organizing, analyzing and interpreting information called data. One major objective of working with this data is to equip individuals to make meaningful decisions and evaluations when presented with information presented in a quantitative form. If raw data are not organized and analyzed, they are simply a mass of numbers or symbols which have no meaning. Learning to work with statistics provides an individual with numerous possibilities for testing hypotheses and solving practical problems using the scientific method. Statistics provide an extremely valuable research tool. While the use of statistics can be highly informative, each person must remember that they are only a means to an end. That end is providing the opportunity to draw reliable research conclusions based on careful analyses. Researchers must be careful to ensure that the use of various statistical tools is driven by the desire to produce analysis that will enable reliable research conclusions, not by the tools themselves.

Statistical tools can be used by those seeking to answer research questions such as, “Why do certain individuals identify with a certain political party?” or “Do rural backgrounds of students affect academic success?” Statistics do not provide direct answers to these and other questions, but they represent a vital part of the research process that can produce such answers.
Important Statistical Terms and Concepts

The process of learning how to use statistics effectively begins with an introduction of several terms and concepts that are central to the use of statistics. **Descriptive statistics** are the product of the processes of collecting and organizing data. These represent the first two processes involved in the science of statistics. These data can be collected from wide variety of sources. Researchers may collect data by making observations and recording them. Other researchers may obtain data from government sources or other existing databases such as the *Statistical Abstracts of the United States*. Still others may conduct a survey as a means of collecting data. After data collection, these data can be organized in ways which make trends, proportions, and summaries clear and apparent. Simply organizing data in some pattern may be very important and can be far more meaningful than a mass of unorganized values. Compare the scores reported in the two figures below. Which set of data is more meaningful? What basic conclusions ben be drawn from the data presented in Figure 1:2?

<table>
<thead>
<tr>
<th>Figure 1:1</th>
<th>Figure 1:2</th>
</tr>
</thead>
<tbody>
<tr>
<td>98, 97, 71, 95, 83, 94, 90, 75, 72, 85, 78, 67, 73, 51, 68, 65, 82, 59, 86, 56, 80, 74, 82, 75, 42,</td>
<td>Score</td>
</tr>
<tr>
<td>Score</td>
<td>Frequency</td>
</tr>
<tr>
<td>90–100</td>
<td>6</td>
</tr>
<tr>
<td>80-89</td>
<td>7</td>
</tr>
<tr>
<td>70-79</td>
<td>11</td>
</tr>
<tr>
<td>60-69</td>
<td>3</td>
</tr>
<tr>
<td>Under 60</td>
<td>4</td>
</tr>
<tr>
<td>Total=</td>
<td>31</td>
</tr>
</tbody>
</table>
Simple data analysis can also be an activity that falls within the scope of descriptive statistics. Preliminary analysis of data is usually accomplished by using two types of statistical measures. These two measures are called **measures of central tendency** and **measures of variation**. The most common measures of central tendency are the **mean**, **median**, and **mode**. Most students are already familiar with the arithmetic mean which is frequently called the **average**. Measures of central tendency represent an attempt to produce a statistic that is representative of a larger group of numbers. As the term central tendency indicates, the value of these statistics tends toward the middle of the group as a whole. The most common measures of variation are **range**, **variance**, and **standard deviation**. These measures describe how widely the data are dispersed from a central point (described by the measures of central tendency). Later chapters will provide a more complete explanation of each of these concepts and instruction on computing each statistic.

The second major area of study in the field of statistics is called **statistical inference**. Statistical inferences are far more complex than descriptive statistics. Statistical inference is the process of drawing conclusions or making predictions about a **population** (N) on the basis of **samples** (n) drawn from that population. A **population** is the complete set of individuals who share at least one characteristic. All individuals who earn under $10,000 per year, all students who attend a particular university, all students in a statistics class, and all registered voters in the United States are examples of populations. A **sample** (n) is a sub-set of the population that has been selected to represent the larger group. A polling organization which selects 2000 registered voters selected from all those registered in the United States has drawn a sample for its research. **Representative samples** are those whose characteristics are the same (or nearly the same) as the
larger population from which they have been drawn. Statistical inference allows researchers to use information drawn from representative samples to make inferences about populations as a whole. Later chapters provide a more complete discussion of sampling and statistical inference.

Researchers seeking to employ statistics as a tool for organizing, analyzing, and evaluating data must begin by collecting **quantified** data. Data is quantified by measuring and expressing it as a numerical value of some kind. The level of measurement employed in the process of quantifying the data determines the type of data that is produced. The three types of quantitative data are **nominal, ordinal, and interval**. Whether individuals employ descriptive statistics or statistical inference, at the very beginning it is critical to understand how and why a researcher quantifies data. Depending on the level of **measurement**, numbers are used to:

1. Categorize or label (nominal data)
2. Rank or order (ordinal data)
3. Score (interval data)

These three designations are called the **nominal, ordinal, and interval** levels of measurement. Each of these levels of measurement is important. There are statistics which have been created for use with **only one** specific level of measurement. The first question which **must always** be asked before conducting statistical analyses is: "What type or level of measurement is the data?"

The answer to this question becomes extremely important later in the text when students are asked to decide what statistic must be used with what type data. Failure to accurately determine the level of measurement will result in selecting the wrong statistic for the analysis which in turn will result in false research conclusions. These false conclusions then negate the entire research project.
The lowest and least precise level of measurement in statistics is the **nominal** level. The nominal level of measurement involves assigning data to mutual exclusive categories. Other frequently used terms associated with nominal data are **labeling** and **naming**. Examples of nominal data are plentiful. A limited number of these are as follows:

<table>
<thead>
<tr>
<th>Grades</th>
<th>Productivity</th>
<th>Gender</th>
<th>400 Meter Dash</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Fail</td>
<td>2. Nonproductive</td>
<td>2. Female</td>
<td>2. Did not finish</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race</th>
<th>Geographical</th>
<th>Time</th>
<th>Political Parties</th>
</tr>
</thead>
</table>

Every cases must be assigned to one and only one category when employing the nominal level of measurement. As in the examples above, researchers often assign numbers to the categories or classifications of data when using nominal measurement as a means of simplifying the process of collecting and coding the data. The use of numerical values to represent these categories does not indicate that one group is higher or lower than the others or carry any meaningful information other than the identification of the group that a particular case falls into.

The reason the nominal level of measurement is considered the lowest level of measurement is that it is not as informative as ordinal and interval levels of measurement. It allows one to say only that things are different. The nominal measurement generally leaves the researcher with many unanswered questions about these data. As a result, statistics which are
designed for analyzing nominal data are useful but less precise than statistics designed to analyze other types of data. Nominal data are generally organized in a data matrix similar to the 2 x 2 matrix for pass and failure rates of men and women shown in Figure 2:3.

<table>
<thead>
<tr>
<th></th>
<th>Pass</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Women</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

For a higher level of measurement, the researcher would move up the hierarchy of complexity to the ordinal level. At the ordinal level of measurement, data placed in a ranked order based on the degree to which they demonstrate a specific characteristic. At this level, it is possible to conclude that a subject has more or less of that characteristic than another. Some examples of ordinal data are as follows:

<table>
<thead>
<tr>
<th>Grades</th>
<th>Miss America Contest</th>
<th>400 Meter Dash</th>
<th>Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. B</td>
<td>2. 1st Runner Up</td>
<td>2. 2nd</td>
<td>2. Good</td>
</tr>
<tr>
<td>3. C</td>
<td>3. 2nd Runner Up</td>
<td>3. 3rd</td>
<td>3. Fair</td>
</tr>
<tr>
<td>4. D</td>
<td>4. 3rd Runner Up</td>
<td>4. 4th</td>
<td>4. Poor</td>
</tr>
<tr>
<td>5. F</td>
<td>5. 4th Runner Up</td>
<td>5. 5th</td>
<td>5. Very Poor</td>
</tr>
</tbody>
</table>

Ordinal data yields more information about relationships among categories than nominal, yet the exact distance between classes and scales is still unknown. Being superior is higher than poor and 4th is lower than 1st. Statistics used for analyzing ordinal data are more reliable than those associated with nominal data; however, for greater accuracy, one must move to the interval level of measurement.
Unlike nominal and ordinal, the **interval level of measurement** provides the exact distance or interval between categories. Interval data are constant units of measurement which appear at equal **intervals** between points on a scale. For most statistical studies or research projects, interval level data are the most useful and descriptive. Examples of interval level data are:

<table>
<thead>
<tr>
<th>Grades</th>
<th>400 Meter Dash</th>
<th>Fahrenheit</th>
<th>Weight</th>
<th>Incomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>45 seconds</td>
<td>32 degrees</td>
<td>98 lbs.</td>
<td>$10,000</td>
</tr>
<tr>
<td>80%</td>
<td>50 seconds</td>
<td>60 degrees</td>
<td>110 lbs.</td>
<td>$20,000</td>
</tr>
<tr>
<td>70%</td>
<td>55 seconds</td>
<td>90 degrees</td>
<td>210 lbs.</td>
<td>$50,000</td>
</tr>
<tr>
<td>60%</td>
<td>60 seconds</td>
<td>212 degrees</td>
<td>250 lbs.</td>
<td>$100,000</td>
</tr>
<tr>
<td>50%</td>
<td>65 seconds</td>
<td>3000 degrees</td>
<td>300 lbs.</td>
<td>$500,000</td>
</tr>
</tbody>
</table>

Interval data yield exact information which may be important to the outcome of a research project. Since interval level data are more precise than nominal and ordinal, the researcher can utilize statistical processes which are more powerful and accurate. Unlike nominal and ordinal measurements, averages and other statistics which require exact distances can be calculated.

The examples given demonstrate that it is often possible to produce a quantitative measurement in three different ways. For example, an individual running the 400 meter dash may finish (nominal), finish 3rd (ordinal), or run the dash in 55 seconds (interval). If one indicates that an individual finished (nominal) in a race, one still has no idea how well the individual actually performed in the race. The runner may have finished last. By knowing that a runner finished 3rd (ordinal), more knowledge about the outcome is apparent, but one still does not know if the runner is a world class sprinter. If a runner finished the race in 55 seconds
taking a test may pass (nominal), earn a C (ordinal), or score a 75% (interval).

Interval data provides sufficient precision to allow the researcher to convert it into ordinal or
class and can compare the score to the minimum passing level to determine if the
student has passed or failed. On the other hand, if the only information available is that the
student has passed, there is no way to determine how that student ranks in the class or the number
or items answered correctly.

Interval data are the most precise, ordinal data less precise, and nominal data the least
precise. Meaningful ordinary arithmetic operations such as obtaining a mean, can only be
performed with interval data. Numerical values are usually assigned to nominal and ordinal data,
but the results of any arithmetic operations would probably be meaningless.

While considering levels of measurement or scales, it is also important to know that any
type of data, nominal, ordinal or interval, may be classified as continuous or discrete. Discrete
variable values can only differ by fixed amounts. There can be no fractional units between the
values. For example, family size, number of siblings, number of arrests, and number of crimes
committed in a large city in one month are all discrete variables. One cannot have two and one-

1 World record held by Michael Johnson at 43.18 seconds.
half arrests, five and one-half siblings, or twenty and two-thirds crimes committed. These variables demand the use of whole numbers, and it is impossible for a third observation to appear between two observations. Continuous variable values may take on any value on a continuum. For example, height, weight, temperature, and age are continuous values. Age difference between two people may be a year, month, day, hour, minute or second. In short, observations that can take on the value of any real number in some continuous interval are called continuous. Data measured on the interval level may be either discrete or continuous. The nominal and ordinal levels of measurement are discrete.

The objective of this chapter has been to acquaint the student with some basic statistics vocabulary and the notion of quantifying data. These terms and concepts are extremely important and become the first building block which must be comprehended before the student is ready to proceed to the next slightly more complex statistical step. A review of these statistical steps appears at the end of the chapter.

---

2 Very frequently, aggregate data such as national average are reported in this manner, but it is well understood that in reality such a division is impossible.
Step 1  What is the study of statistics? The science of collecting, organizing, analyzing, and interpreting information called data.

Step 2  What are descriptive statistics? They are statistics which describe distributions of data.

Step 3  What is statistical inference? Statistical inference is interpreting data by drawing inferences for a population from knowledge of a sample.

Step 4  What is the nominal level of measurement?

Step 5  What is the ordinal level of measurement? At the ordinal level of measurement, the data are ranked.

Step 6  What is the interval level of measurement? At the interval level, the data are at equal intervals on a scale.
EXERCISES - CHAPTER 1

I. Short answer questions:

(1) Define the term "statistics".

(2) What is a population? What is a sample?

(3) Other than those given in the text, give three examples of populations and three examples of samples.

(4) The following are some won-lost records for the teams in the Western Division of the American Baseball League. Organize these data in a meaningful way.

Milwaukee 58-60, Cleveland 61-38, Chicago 62-44, Minnesota 50-61, Kansas City 30-77.

(5) The data in question 5 are what level of measurement as presented in the daily newspaper?

II. For each of the following, indicate (N) for nominal, (O) or ordinal, and (I) for interval:

(6) ______ Book bestseller list, 1st-10th
(7) ______ Marital Status (Married, Single, Divorced)
(8) ______ Interest Rates
(9) ______ City Populations
(10) ______ State of Residency
(11) ______ Open or Closed
(12) ______ Temperature
(13) ______ County Fair Ribbons: Purple, Blue, Red, White
(14) ______ Networks: ABC, NBC, CBS
(15) ______ Top Ten Songs
(16) ______ Height (in inches)
(17) ______ Weight (lbs)
(18) ______ Olympic Medals: Gold, Silver, Bronze
(19) ______ Veteran / Non-Veteran
(20) ______ Orchestra: 1st chair, 2nd chair, 3rd chair
(21) ______ Win or Lose
(22) ______ Hotels: 1, 2, 3, 4, or 5 star
(23) ______ Texture: Raised or Flat
(24) ______ Military Ranks
(25) ______ Domestic or Foreign
(26) ______ Speed (mph)
(27) ______ Income Taxes
(28) _____ On or Off
(29) _____ Batting Averages
(30) _____ IQ Scores
(31) _____ Boy Scout Ranks
(33) _____ Grades of Meat
(34) _____ Classes of Mail: 1st, 2nd, 3rd,
(35) _____ College Faculty: Professor, Associate Professor, Assistant Professor, Instructor
(36) _____ Calories
(37) _____ Golf Scores
(38) _____ Home or Away Game
(39) _____ Top Stock Performers
(40) _____ Gallons
(41) _____ Decibels
(42) _____ LSAT Scores
(43) _____ US News Rankings of Colleges and Universities
(44) _____ Guilty or Not Guilty Verdict

(45) Can a mean be obtained for nominal or ordinal data? Why or why not?

III. For a math review, solve the following (52-65). Let \( x = 5 \) and \( y = 10 \).

(46) \( 9 + 5x = \) \hspace{1cm} (53) \( 2y^2 + 5x = \)

(47) \( .96y + x = \) \hspace{1cm} (54) \( x^2 + y^2 + x^3 = \)

(48) \( x + 6 = \) \hspace{1cm} (55) \( .61x + 50 - 36 = \)

(49) \( xy + xy = \) \hspace{1cm} (56) \( 2x + 7y = \)

(50) \( x^2 - y^2 = \) \hspace{1cm} (57) \( 1.06y - 99 + 10 = \)

(51) \( 3xy - 4xy + x = \) \hspace{1cm} (58) \( x^2 - 2 + y^3 \)

(52) \( 3x^2 = \) \hspace{1cm} (59) \( .75x + 10 - 22 = \)