

CHAPTER ONE

INTRODUCTION

Welcome to the study of statistics! It has been our experience that many students face the prospect of taking a course in statistics with a great deal of anxiety, apprehension, and even dread. They fear not having the extensive mathematical background that they assume is required, and they fear that the contents of such a course will be irrelevant to their work in their fields of concentration.

Although it is true that an extensive mathematical background is required at more advanced levels of statistical study, it is not required for the introductory level of statistics presented in this book. Greater reliance is placed on the use of the computer for calculation and graphical display so that we may focus on issues of conceptual understanding and interpretation. While hand computation is de-emphasized, we believe, nonetheless, that a basic mathematical background – including the understanding of fractions, decimals, percentages, signed (positive and negative) numbers, exponents, linear equations and graphs – is essential for an enhanced conceptual understanding of the material.

As for the issue of relevance, we have found that students better comprehend the power and purpose of statistics when it is presented in the context of a substantive problem with real data. In this information age, data are available on a myriad of topics. Whether our interests are in health, education, psychology, business, the environment, and so on, numerical data may be accessed readily to address our questions of interest. The purpose of statistics is to allow us to analyze these data to extract the information that they contain in a meaningful way and to write a story about the numbers that is both compelling and accurate.

Throughout this book we make use of a series of real data sets that are available online through our companion website: www.cambridge.org/Stats-Stata2e. We will pose relevant questions and learn the appropriate methods of analysis for answering such questions. Students will learn that more than one statistic or method of analysis typically is needed to address a question fully. Students also will learn that a detailed description of the data, including possible anomalies, and an ample characterization of results, are critical components of any data analytic plan. Through this process we hope that this book will help students come to view statistics as an integrated set of data analytic tools that when used together in a meaningful way will serve to uncover the story contained in the numbers.

The Role of Statistical Software in Data Analysis

From our own experience, we have found that the use of a statistics software package to carry out computations and create graphs not only enables a greater emphasis on conceptual understanding and interpretation, but also allows students to study statistics

in a way that reflects statistical practice. We have selected the latest version of Stata available to us at the time of writing, version 16, for use with this text. We have selected Stata because it is a well-established comprehensive package with a robust technical support infrastructure that is widely used by behavioral and social scientists. In addition, not only does Stata include a menu-driven “point and click” interface, making it accessible to the new user, but it also includes a command line or program syntax interface, allowing students to be guided from the comfortable “point and click” environment to the beginnings of statistical programming. Like MINITAB, JMP, Data Desk, Systat, SPSS, and SPlus, Stata is powerful enough to handle the analysis of large data sets quickly. By the end of the course, students will have obtained a conceptual understanding of statistics as well as an applied, practical skill in how to carry out statistical operations.

Statistics: Descriptive and Inferential

The subject of statistics may be divided into two general branches: descriptive and inferential. *Descriptive statistics* are used when the purpose of an investigation is to *describe* the data that have been (or will be) collected. Suppose, for example, that a third-grade elementary school teacher is interested in determining the proportion of children who are firstborn in her class of 30 children. In this case, the focus of the teacher’s question is her own class of 30 children and she will be able to collect data on all of the students about whom she would like to draw a conclusion. The data collection operation will involve noting whether each child in the classroom is firstborn or not; the statistical operations will involve counting the number who are, and dividing that number by 30, the total number of students in the class, to obtain the proportion sought. Because the teacher is using statistical methods merely to describe the data she collected, this is an example of descriptive statistics.

Suppose, on the other hand, that the teacher is interested in determining the proportion of children who are firstborn in *all* third-grade classes in the city where she teaches. It is highly unlikely that she will be able to (or even want to) collect the relevant data on all individuals about whom she would like to draw a conclusion. She will probably have to limit the data collection to some randomly selected smaller group and use *inferential statistics* to generalize to the larger group the conclusions obtained from the smaller group. *Inferential statistics* are used when the purpose of the research is not to describe the data that have been collected, but to generalize or make inferences based on it. The smaller group on which she collects data is called the *sample*, whereas the larger group to whom conclusions are generalized (or inferred), is called the *population*. In general, two major factors influence the teacher’s confidence that what holds true for the sample also holds true for the population at large. These two factors are the method of sample selection and the size of the sample. Only when data are collected on *all* individuals about whom a conclusion is to be drawn (when the sample *is* the population and we are therefore in the realm of descriptive statistics), can the conclusion be drawn with 100 percent certainty. Thus, one of the major goals of inferential statistics is to assess the degree of certainty of inferences when such inferences are drawn from sample data. Although this text is divided roughly into two parts, the first on descriptive statistics and the second on inferential statistics, the second part draws heavily on the first.

Variables and Constants

In the previous section, we discussed a teacher's interest in determining the proportion of students who are firstborn in the third grade of the city where she teaches. What made this question worth asking was the fact that she did not expect everyone in the third grade to be firstborn. Rather, she quite naturally expected that in the population under study, birth order would vary, or differ, from individual to individual and that only in certain individuals would it be first.

Characteristics of persons or objects that vary from person to person or object to object are called *variables*, whereas characteristics that remain constant from person to person or object to object are called *constants*. Whether a characteristic is designated as a variable or as a constant depends, of course, on the study in question. In the study of birth order, birth order is a variable; it can be expected to vary from person to person in the given population. In that same study, grade level is a constant; all persons in the population under study are in the third grade.

.....
EXAMPLE 1.1 Identify some of the variables and constants in a study comparing the math achievement of tenth-grade boys and girls in the southern United States.

Solution *Constants:* Grade level; Region of the United States
Variables: Math achievement; Sex

.....
EXAMPLE 1.2 Identify some of the variables and constants in a study of math achievement of secondary-school boys in the southern United States.

Solution *Constants:* Sex; Region of the United States
Variables: Math achievement; Grade level

Note that grade level is a constant in Example 1.1 and a variable in Example 1.2. Because constants are characteristics that do not vary in a given population, the study of constants is neither interesting nor informative. The major focus of any statistical study is therefore on the variables rather than the constants. Before variables can be the subject of statistical study, however, they need to be numerically valued. The next section describes the process of measuring variables so as to achieve that goal.

The Measurement of Variables

Measurement involves the observation of characteristics on persons or objects, and the assignment of numbers to those persons or objects so that the numbers represent the amounts of the characteristics possessed. As introduced by S. S. (Stanley Smith) Stevens (1946) in a paper, *On the Theory of Scales of Measurement*, and later described by him in a chapter, "Mathematics, measurement, and psychophysics," in the *Handbook of Experimental Psychology*, edited by Stevens (1951), we describe four levels of measurement in this text. Each of the four levels is defined by the nature of the observation and the way in which the numbers assigned correspond to the amount of the underlying characteristic that has been observed. The level of measurement of a variable determines which

numerical operations (e.g., addition, subtraction, multiplication, or division) are permissible on that variable. If other than the permissible numerical operations are used on a variable given its level of measurement, one can expect the statistical conclusions drawn with respect to that variable to be questionable.

Nominal Level

The nominal level of measurement is based on the simplest form of observation – whether two objects are similar or dissimilar; for example, whether they are short versus non-short, male versus female, or college student versus non-college student. Objects observed to be similar on some characteristic (e.g., college student) are assigned to the same class or category, while objects observed to be dissimilar on that characteristic are assigned to different classes or categories. In the nominal level of measurement, classes or categories are *not* compared as say, taller or shorter, better or worse, or more educated or less educated. Emphasis is strictly on observing whether the objects are similar or dissimilar. As befitting its label, classes or categories are merely named, but not compared, in the *nominal* level of measurement.

Given the nature of observation for this level of measurement, numbers are assigned to objects using the following simple rule: if objects are dissimilar, they are assigned different numbers; if objects are similar, they are assigned the same number. For example, all persons who are college students would be assigned the same number (say, 1); all persons who are non-college students also would be assigned the same number different from 1 (say, 2) to distinguish them from college students. Because the focus is on distinction and not comparison, in this level of measurement, the fact that the number 2 is larger than the number 1 is irrelevant in terms of the underlying characteristic being measured (whether or not the person is a college student). Accordingly, the number 1 could have been assigned, instead, to all persons who are non-college students and the number 2 to all persons who are college students. Any numbers other than 1 and 2 also could have been used as well.

While the examples in this section (e.g., college student versus non-college student) may be called *dichotomous* in that they have only two categories, nominal variables also may have more than two categories (e.g., car manufacturers – Toyota, Honda, General Motors, Ford, Chrysler, etc.).

Ordinal Level

The ordinal level of measurement is not only based on observing objects as similar or dissimilar, but also on ordering those observations in terms of an underlying characteristic. Suppose, for example, we were not interested simply in whether a person was a college student or not, but rather in ordering college students in terms of the degree of their success in college (e.g., whether the college student was below average, average, or above average). We would, therefore, need to observe such ordered differences among these college students in terms of their success in college and we would choose numbers to assign to the categories that corresponded to that ordering. For this example, we might assign the number 1 to the below average category, the number 2 to the average category, and the number 3 to the above average category. Unlike in the nominal level of

measurement, in the ordinal level of measurement, it is relevant that 3 is greater than 2, which, in turn, is greater than 1, as this ordering conveys in a meaningful way the ordered nature of the categories relative to the underlying characteristic of interest. That is, comparisons among the numbers correspond to comparisons among the categories in terms of the underlying characteristic of success in college. In summary, the ordinal level of measurement applies two rules for assigning numbers to categories: (1) different numbers are assigned to persons or objects that possess different amounts of the underlying characteristic, and (2) the higher the number assigned to a person or object, the less (or more) of the underlying characteristic that person or object is observed to possess. From these two rules it does *not* follow, however, that equal numerical differences along the number scale correspond to equal increments in the underlying characteristic being measured in the ordinal level of measurement. While the differences between 3 and 2 and between 2 and 1 in our college student success example are both equal to 1, we cannot infer from this that the difference in success between above average college students and average college students equals the difference in success between average college students and below average college students.

We consider another example that may convey more clearly this idea. Suppose we line 10 people up according to their size place and assign a number from 1 to 10, respectively to each person so that each number corresponds to the person's size place in line. We could assign the number 1 to the shortest person, the number 2 to the next shortest person, and so forth, ending by assigning the number 10 to the tallest person. While, according to this method, the numbers assigned to each pair of adjacent people in line will differ from each other by the same value (i.e., 1), clearly, the heights of each pair of adjacent people will not necessarily also differ by the same value. Some adjacent pairs will differ in height by only a fraction of an inch, while other adjacent pairs will differ in height by several inches. Accordingly, only some of the features of this size place ranking are reflected or modeled by the numerical scale. In particular, while, in this case, the numerical scale can be used to judge the relative order of one person's height compared to another's, differences between numbers on the numerical scale cannot be used to judge how much taller one person is than another. As a result, statistical conclusions about variables measured on the ordinal level that are based on other than an ordering or ranking of the numbers (including taking sums or differences) cannot be expected to be meaningful.

Interval Level

An ordinal level of measurement can be developed into a higher level of measurement if it is possible to assess how near to each other the persons or objects are in the underlying characteristic being observed. If numbers can be assigned in such a way that equal numerical differences along the scale correspond to equal increments in the underlying characteristic, we have what is called an *interval level of measurement*. As an example of an interval level of measurement, consider the assignment of yearly dates, the chronological scale. Because one year is defined as the amount of time necessary for the Earth to revolve once around the Sun, we may think of the yearly date as a measure of the number of revolutions of the Earth around the Sun up to and including that year. Hence, this assignment of numbers to the property *number of revolutions of the Earth around the*

Sun is on an interval level of measurement. Specifically, this means that equal numerical differences for intervals (such as 1800 CE to 1850 CE and 1925 CE to 1975 CE) represent equal differences in the number of revolutions of the Earth around the Sun (in this case, 50). In the interval level of measurement, therefore, we may make meaningful statements about the amount of *difference* between any two points along the scale. As such, the numerical operations of addition and subtraction (but not multiplication and division) lead to meaningful conclusions at the interval level and are therefore permissible at that level. For conclusions based on the numerical operations of multiplication and division to be meaningful, we require the ratio level of measurement.

Ratio Level

An interval level of measurement can be developed into a higher level of measurement if the number zero on the numeric scale corresponds to zero or “not any” of the underlying characteristic being observed. With the addition of this property (called an *absolute zero*), ratio comparisons are meaningful, and we have what is called the *ratio level of measurement*. Consider once again the chronological scale and, in particular, the years labeled 2000 CE and 1000 CE. Even though 2000 is numerically twice as large as 1000, it does not follow that the number of revolutions represented by the year 2000 is twice the number of revolutions represented by the year 1000. This is because on the chronological scale, the number 0 (0 CE) does not correspond to zero revolutions of the Earth around the Sun (i.e., the Earth had made revolutions around the Sun many times prior to the year 0 CE). In order for us to make meaningful multiplicative or ratio comparisons of this type between points on our number scale, the number 0 on the numeric scale must correspond to 0 (none) of the underlying characteristic being observed.

In measuring height, not by size place, but with a standard ruler, for example, we would typically assign a value of 0 on the number scale to “not any” height and assign the other numbers according to the rules of the interval scale. The scale that would be produced in this case would be a ratio scale of measurement, and ratio or multiplicative statements (such as “John, who is 5 feet tall, is twice as tall as Jimmy, who is 2.5 feet tall”) would be meaningful. It should be pointed out that for variables to be considered to be measured on a ratio level, “not any” of the underlying characteristic only needs to be meaningful theoretically. Clearly, no one has zero height, yet using zero as an anchor value for this scale to connote “not any” height is theoretically meaningful.

Choosing a Scale of Measurement

Why is it important to categorize the scales of measurement as nominal, ordinal, interval, or ratio? If we consider college students and assign a 1 to those who are male college students, a 2 to those who are female college students, and a 3 to those who are not college students at all, it would not be meaningful to add these numbers nor even to compare their sizes. For example, two male college students together do not suddenly become a female college student, even though their numbers add up to the number of a female college student ($1 + 1 = 2$). And a female college student who is attending school only half-time is not suddenly a male college student, even though half of her number is the number of a male college student ($2/2 = 1$). On the other hand, if we were dealing with a ratio-leveled

THE MEASUREMENT OF VARIABLES

7

height scale, it would be possible to add, subtract, multiply, or divide the numbers on the scale and obtain results that are meaningful in terms of the underlying trait, height. In general, and as noted earlier, the scale of measurement determines which numerical operations, when applied to the numbers of the scale, can be expected to yield results that are meaningful in terms of the underlying trait being measured.

TABLE 1.1 Hierarchy of scales of measurement

- | |
|---|
| <ol style="list-style-type: none"> 1. Ratio 2. Interval 3. Ordinal 4. Nominal |
|---|

Said differently, any numerical operation can be performed on any set of numbers; whether the resulting numbers are meaningful, however, depends on the particular level of measurement being used.

Note that the four scales of measurement exhibit a natural hierarchy, or ordering, in the sense that each level exhibits all the properties of those below it (see Table 1.1). Any characteristic that can be measured on one scale listed in Table 1.1 can also be measured on any scale below it in that list. Given a precise measuring instrument such as a perfect ruler, we can measure a person's height, for example, as a ratio-scaled variable, in which case we could say that a person whose height is 5 feet has twice the height of a person whose height is 2.5 feet. Suppose, however, that no measuring instrument were available. In this situation, we could, as we have done before, "measure" a person's height according to size place or by categorizing a person as tall, average, or short. By assigning numbers to these three categories (such as 5, 3, and 1, respectively), we would create an ordinal level of measurement for height.

In general, it may be possible to measure a variable on more than one level. The level that is ultimately used to measure a variable should be the highest level possible, given the precision of the measuring instrument used. A perfect ruler allowed us to measure heights on a ratio level, while the eye of the observer allowed us to measure height only on an ordinal level. If we are able to use a higher level of measurement but decide to use a lower level instead, we would lose some of the information that would have been available to us on the higher level. We would also be restricting ourselves to a lower level of permissible numerical operations.

.....
EXAMPLE 1.3 Identify the level of measurement (nominal, ordinal, interval, or ratio) most likely to be used to measure the following variables:

1. Ice cream flavors
2. The speed of five runners in a one-mile race, as measured by the runners' order of finish, first, second, third, and so on.
3. Temperature measured in degrees Celsius.
4. The annual salary of individuals.

Solution

1. The variable ice cream flavors is most likely measured at the nominal level of measurement because the flavors themselves may be categorized simply as being the same or different and there is nothing inherent to them that would lend themselves to a ranking. Any ranking

would have to depend on some extraneous property such as, say, taste preference. If numbers were assigned to the flavors as follows,

Flavor	Number
Vanilla	0
Chocolate	1
Strawberry	2
etc.	etc.


meaningful numerical comparisons would be restricted to whether the numbers assigned to two ice cream flavors are the same or different. The fact that one number on this scale may be larger than another is irrelevant.

- This variable is measured at the ordinal level because it is the order of finish (first, second, third, and so forth) that is being observed and not the specific time to finish. In this example, the smaller the number the greater the speed of the runner. As in the case of measuring height via a size place ranking, it is not necessarily true that the difference in speed between the runners who finished first and second is the same as the difference in speed between the runners who finished third and fourth. Hence, this variable is not measured at the interval level. Had time to finish been used to measure the speed of the runners, the level of measurement would have been ratio for the same reasons that height, measured by a ruler, is ratio-leveled.
- Temperature measured in degrees Celsius (centigrade) is at the interval level of measurement because each degree increment, no matter whether from 3 to 4°C or from 22 to 23°C, has the same physical meaning in terms of the underlying characteristic, heat. In particular, it takes 1 calorie to raise the temperature of 1 gram of water by 1 degree Celsius, no matter what the initial temperature reading on the Celsius scale. Thus, equal differences along the Celsius scale correspond to equal increments in heat, making this scale interval-leveled. The reason this scale is not ratio-scaled is because 0 degrees Celsius does not correspond to “not any” heat. The 0 degree point on the Celsius scale is the point at which water freezes, but even frozen water contains plenty of heat. The point of “not any” heat is at -273 degrees Celsius. Accordingly, we cannot make meaningful ratio comparisons with respect to amounts of heat on the Celsius scale and say, for example, that at 20 degrees Celsius there is twice the heat than at 10 degrees Celsius.
- The most likely level of measurement for annual salary is the ratio level because each additional unit increase in annual salary along the numerical scale corresponds to an equal additional one dollar earned no matter where on the scale one starts, whether it be, for example, at \$10,000 or at \$100,000; and, furthermore, because the numerical value of 0 on the scale corresponds to “not any” annual salary, giving the scale a true or absolute zero. Consequently, it is appropriate to make multiplicative comparisons on this scale, such as “Sally’s annual salary of \$100,000 is twice Jim’s annual salary of \$50,000.”

Discrete and Continuous Variables

As we saw in the last section, any variable that is not intrinsically numerically valued, such as the ice cream flavors in Example 1.3, may be converted to a numerically valued variable. Once a variable is numerically valued, it may be classified as either discrete or continuous.

Although there is really no exact statistical definition of a discrete or a continuous variable, the following usage generally applies. A numerically valued variable is said to be *discrete* (or *categorical* or *qualitative*) if the values it takes on are integers or can be thought of in some unit of measurement in which they are integers. A numerically valued variable is said to be *continuous* if, in any unit of measurement, whenever it can take on the values a and b , it can also theoretically take on all the values between a and b . The limitations of the measuring instrument are not considered when discriminating between discrete and continuous variables. Instead, it is the nature of the underlying variable that distinguishes between the two types.

 **Remark.** As we have said, there is really no hard and fast definition of discrete and continuous variables for use in practice. The words discrete and continuous do have precise mathematical meanings, however, and in more advanced statistical work, where more mathematics and mathematical theory are employed, the words are used in their strict mathematical sense. In this text, where our emphasis is on statistical practice, the usage of the terms discrete and continuous will not usually be helpful in guiding our selection of appropriate statistical methods or graphical displays. Rather, we will generally use the particular level of measurement of the variable, whether it is nominal, ordinal, interval, or ratio.

.....

EXAMPLE 1.4 Let our population consist of all eighth-grade students in the United States, and let X represent the region of the country in which the student lives. X is a variable, because there will be different regions for different students. X is not naturally numerically valued, but because X represents a finite number of distinct categories, we can assign numbers to these categories in the following simple way: 1 = Northeast, 2 = North Central, 3 = South, and 4 = West. X is a discrete variable, because it can take on only four values. Furthermore, because X is a nominal-leveled variable, the assignment of 1, 2, 3, and 4 to Northeast, North Central, South, and West, respectively, is arbitrary. Any other assignment rule would have been just as meaningful in differentiating one region from another.

.....

EXAMPLE 1.5 Consider that we repeatedly toss a coin 100 times and let X represent the number of heads obtained for each set of 100 tosses. X is naturally numerically valued and may be considered discrete because the only values it can take on are the integer values 0, 1, 2, 3, and so forth. We may note that X is ratio-leveled in this example because 0 on the numerical scale represents “not any” heads.


.....


EXAMPLE 1.6 Consider a certain hospital with 100 beds. Let X represent the percentage of occupied beds for different days of the year. X is naturally numerically valued as a proportion of the number of occupied beds. Although X takes on fractional values, it is considered discrete because the proportions are based on a count of the number of beds occupied, which is an integer value.

.....

EXAMPLE 1.7 Let our population consist of all college freshmen in the United States, and let X represent their heights, measured in inches. X is numerically valued and is continuous because all possible values of height are theoretically possible. Between any two heights

exists another theoretically possible height. For example, between 70 and 71 inches in height, exists a height of 70.5 inches and between 70 and 70.5 exists a height of 70.25 inches, and so on.

 **Remark.** Even if height in Example 1.7 were reported to the nearest inch (as an integer), it would still be considered a continuous variable because all possible values of height are theoretically possible. Reporting values of continuous variables to the nearest integer is usually due to the lack of precision of our measuring instruments. We would need a perfect ruler to measure the exact values of height. Such a measuring instrument does not, and probably cannot, exist. When height is reported to the nearest inch, a height of 68 inches is considered to represent all heights between 67.5 and 68.5 inches. While the precision of the measuring instrument determines the accuracy with which a variable is measured, it does not determine whether the variable is discrete or continuous. For that we need only to consider the theoretically possible values that a variable can assume.

 **Remark.** In addition to the problem of not being able to measure variables precisely, another problem that often confronts the behavioral scientist is the measurement of traits, such as intelligence, that are not directly observable. Instead of measuring intelligence directly, tests have been developed that measure it indirectly, such as the IQ test. While such tests report IQ, for example, as an integer value, IQ is considered to be a continuous trait or variable, and an IQ score of 109, for example, is thought of as theoretically representing all IQ scores between 108.5 and 109.5.

Another issue, albeit a more controversial one, related to the measurement of traits that are not directly observable, is the level of measurement employed. While some scientists would argue that IQ scores are only ordinal (given a good test of intelligence, a person whose IQ score is higher than another's on that test is said to have greater intelligence), others would argue that they are interval. Even though equal intervals along the IQ scale (say, between 105 and 110 and between 110 and 115) may not necessarily imply equal amounts of change in intelligence, a person who has an IQ score of 105 is likely to be closer in intelligence to a person who has an IQ score of 100 rather than to a person who has an IQ score of 115. By considering an IQ scale and other such psychosocial scales as ordinal only, one would lose such information that is contained in the data and the ability to make use of statistical operations based on sums and differences, rather than merely on rankings.

Another type of scale, widely used in attitude measurement, is the Likert scale, which consists of a small number of values, usually five or seven, ordered along a continuum representing agreement. The values themselves are labeled typically from strongly disagree to strongly agree. A respondent selects that score on the scale that corresponds to his or her level of agreement with a statement associated with that scale. For the same reasons noted earlier with respect to the IQ scale, for example, the Likert scale is considered by many to be interval rather than strictly ordinal.

.....

EXAMPLE 1.8 For each variable listed, describe whether the variable is discrete or continuous. Also, describe the level of measurement for each variable. Use the following data set excerpted from data analyzed by Tomasi and Weinberg (1999). The original study was carried