PART ONE

UNIVARIATE ANALYSIS

Describing Variable Distributions

CHAPTER TWO

LEVELS OF MEASUREMENT AND AGGREGATION

Learning Objectives

1. Summarize the role of variables in research.
2. Identify the four levels of measurement variables can have.
3. Describe the difference between variables that identify qualities compared with variables that identify quantities.
4. Explain the differences among raw frequencies, proportions, percentages, and rates.
5. Define the units of analysis in any particular data set.

Science cannot progress without reliable and accurate measurement of what it is you are trying to study. The key is measurement, simple as that.

—Robert D. Hare

When you can measure what you are speaking about, and express it in numbers, you know something about it.

—The Lord Kelvin

INTRODUCTION

In Chapter 1, we examined various sampling techniques that can be used for selecting a sample from a given population. Once we have selected our sample, we can begin the process of collecting information. The information we gather is usually referred to as "data" and in its entirety is called a "data set." In this chapter, we will take a closer look at the types of variables that can make up a data set.

This may be the first time you have been formally exposed to statistics, but we are sure each of you has some idea what a variable is even though you may not call it that. A **variable** is any element to which different values can be attributed. Although society is beginning to acknowledge the fluidity of gender, it is still a variable that is typically measured using two values, male and female. Race/ethnicity is a variable with many values, such as American Indian, African American, Asian, Latinx, Caucasian, and mixed. Age is another variable that can take on different values, such as 2, 16, 55 years, and so on.

Variable: Characteristic or property that can vary or take on different values or attributes.

As we noted in the last chapter, in explanatory research, we are interested in explaining a dependent variable by using one or more independent variables. In research, the dependent variable is expected to vary or change depending on variation or change in the independent variable. In this causal type of explanation, the independent variable is the cause and the dependent variable the effect or outcome. The entire set of values a variable takes on is called a **frequency distribution** or an **empirical distribution**. In a given data set, a frequency distribution is a distribution (a list) of outcomes or values for a variable. It is referred to as an empirical distribution because it is a distribution of empirical (real and observed) data, and it is called a frequency distribution because it tells us how frequent each value or outcome is in the entire data set. For example, suppose we conducted a survey from a sample of 100 persons in your class at your university. In one question we asked for respondent's age. Suppose this "age" variable ranged from 18 to 42. There might be 15 people who were 18 years of age, 30 people who were 19 years of age, 17 people who were 20 years of age, only 1 person who was 42 years of age, and so on. An empirical, or frequency, distribution would tell you not only what the different ages were but also how many people in the sample were represented by each age in the entire distribution.

Frequency or empirical distribution: Distribution of values that make up a variable distribution.

In contrast, a characteristic of your sample that does not vary in a data set is called a **constant**. Unlike a variable, whose values vary or are different, a constant has only one value. For example, if you have a sample of inmates from a male correctional institution, the value for "gender" would be considered a constant—"male." Since all elements of the sample would be male, respondent's gender would not vary in that data set. Similarly, if you selected a sample of 20-year-olds from the sophomore class at a state university, age would be a constant rather than a variable in that sample because all members of the sample would be the same age (20 years).

Constant: Characteristic or property that does not vary but takes on only one value.

Notice that a given characteristic, such as respondent's gender or age, is not always a variable or a constant. Under different conditions, it may be one or the other. For example, in a sample of male prisoners, gender is a constant, but age is a variable because the male inmates are likely to be different ages. In the sample of 20-year-old sophomore students from a university, age is a constant, and respondent's gender is a variable because some persons in the sample would be male and some would be female.

We can classify variables in many different ways and make several distinctions among them. First, there are differing levels of measurement that can be associated with variables. The next section of the chapter examines these measurement differences, beginning with the classification of variables as either continuous or categorical variables. We then examine the four measurement classifications within these broad categories: nominal, ordinal, interval, and ratio measurement. The second section of the chapter addresses the difference between independent and dependent variables and the different ways of reporting the features of variables. In the final section, you will learn how to identify the units of analysis in a research design so that you can state conclusions about the relationships between your variables in the appropriate units. As you will soon see, understanding this information is extremely important as every statistical application we use typically depends on a variable's levels of measurement. And when we interpret statistical results, we can only generalize that result to the units of analysis that were observed in the sample.

LEVELS OF MEASUREMENT

Recall that data generally come from one of three places: They are gathered by us personally, gathered by another researcher, or gathered by a government agency. Doing research on a previously collected data set is often referred to as "secondary data analysis" because the data already existed and had been analyzed before. No matter how they were collected, however, data sets are by definition simply a collection of many variables. For illustrative purposes, imagine that we were interested in the relationship between levels of student drinking and drug use and student demographic characteristics such as gender, age, religion, and year in college (freshman, sophomore, junior, senior). Table 2.1 displays the small data set we might have obtained had we investigated this issue by collecting surveys from 20 college students (a random sample, of course).

Table 2.1 Example of the Format of a Data Set From a Survey of 20 College Students

ID Number	Gender	Age	College Year	*GPA*	Average Month		Religion
					# Drinks	# Times Used Marijuana	
1	Female	19	Sophomore	2.3	45	22	Catholic
2	Male	22	Senior	3.1	30	10	Other
3	Female	22	Senior	3.8	0	0	Protestant
4	Female	18	Freshman	2.9	35	5	Jewish
5	Male	20	Junior	2.5	20	20	Catholic
6	Female	23	Senior	3.0	10	0	Catholic
7	Male	18	Freshman	1.9	45	25	Not religious
8	Female	19	Sophomore	2.8	28	3	Protestant
9	Male	28	Junior	3.3	9	0	Protestant
10	Female	21	Junior	2.7	0	0	Muslim
11	Female	18	Freshman	3.1	19	2	Jewish
12	Male	19	Sophomore	2.5	25	20	Catholic
13	Female	21	Senior	3.5	2	0	Other
14	Male	21	Junior	1.8	19	33	Protestant
15	Female	42	Sophomore	3.9	10	0	Protestant
16	Female	19	Sophomore	2.3	45	0	Catholic
17	Male	21	Junior	2.8	29	10	Not religious
18	Male	25	Sophomore	3.1	14	0	Other
19	Female	21	Junior	3.5	5	0	Catholic
20	Female	17	Freshman	3.5	28	0	Jewish

Qualitative or categorical variables: Values that refer to qualities or categories. They tell us what kind, what group, or what type a value is referring to.

To measure the extent to which each student used alcohol and marijuana, let's say we asked them these questions: "How many drinks do you consume in an average month? By 'drinks' we mean a beer, a mixed drink, or a glass of wine." "How many times during an average month do you use marijuana?" Each of the other variables in the table relates to other information about each student in the sample. Everything listed in this table, including the respondent's identification number, is a variable. All of these variables combined represent our data set. The first thing you may notice about these variables is that some are represented by categories and some are represented by actual numbers. Gender, for example, is divided into two categories, female and male. This type of variable is often referred to as a **qualitative** or **categorical variable**, implying that the values represent qualities or categories only. The values of this variable have no numeric or quantitative meaning. Other examples in the data set of qualitative variables include college year and religion.

The rest of the variables in our data set, however, have values that do represent numeric values that can be quantified—hence the name **quantitative or continuous variables**. The values of quantitative variables can be compared in a numerically meaningful way. Respondent's identification number, age, grade point average, number of drinks, and number of times drugs were used are all quantitative variables. We can compare the values of these variables in a numerically meaningful way. For example, from Table 2.1, we can see that respondent 1 has a lower grade point average than respondent 19. We can also see that respondents 7 and 16 have the highest levels of alcohol consumption in the sample.

Quantitative or continuous variables: Values that refer to quantities or different measurements. They tell us how much or how many.

In Table 2.1, it is relatively easy to identify which variables are qualitative and which are quantitative simply because the qualitative variables are represented by **alphanumeric data** (by letters rather than by numbers). Data that are represented by numbers are called **numeric data**. A good way to remember the distinction between these two types of data is to note that alphanumeric data consist of letters of the alphabet, whereas numeric data consist of numbers.

Alphanumeric data: Values of a variable that are represented by letters rather than by numbers.

Numeric data: Values of a variable that represent numerical qualities.

It is certainly possible to include alphanumeric data in a data set, as we have done in Table 2.1, but when stored in a computer, as most data are, alphanumeric data take up a great deal of space, and alphanumeric data are difficult to statistically analyze. For this reason, these data are usually converted to or represented by numeric values. For example, females may arbitrarily be identified with the number 1, rather than with the word "female," and males with the number 2. Assigning numbers to the categorical values of qualitative variables is called "coding" the data. Of course, which numbers get assigned to qualitative variables (for example, 1 for females and 2 for males) is arbitrary because the numeric code (number) assigned has no real quantitative meaning. Males could be given either a 1 or a 2, or a 0, with females coded either a 2 or a 1; it makes no difference. The numbers would still only be representing qualities or categories.

Table 2.2 redisplays the data in Table 2.1 numerically as they would normally be stored in a computer data set. Because values of each variable are represented by numbers, it is a little more difficult to distinguish the qualitative variables from the quantitative variables. You have to ask yourself what each of the values really means. For example, for the variable gender, what does the "1" really represent? It represents the code for a female student and is therefore not numerically meaningful. Similarly, the number "1" coded for the religion variable represents those students who said they were Catholic, and the code "3" represents those students who said they were Jewish. There is nothing inherently meaningful about the numbers 1 and 3. They simply represent categories for the religion variable and we changed the letters of the alphabet to numbers. For the variable age, what does the number 19 represent? This is actually a meaningful value—it tells us that this respondent was 19 years of age, and it is therefore a quantitative variable.

Table 2.2 Example of the Data Presented in Table 2.1 as They Would Be Stored in a Computer Data File

ID Number	Gender	Age	College Year	*GPA*	Average Month		Religion
					# Drinks	# Times Drugs Used	
1	1	19	2	2.3	45	22	1
2	2	22	4	3.1	30	10	6
3	1	22	4	3.8	0	0	2
4	1	18	1	2.9	35	5	3
5	2	20	3	2.5	20	20	1

(Continued)

Table 2.2 (Continued)

ID Number	Gender	Age	College Year	*GPA*	Average Month		Religion
					# Drinks	# Times Drugs Used	
6	1	23	4	3.0	10	0	1
7	2	18	1	1.9	45	25	5
8	1	19	2	2.8	28	3	2
9	2	28	3	3.3	9	0	2
10	1	21	3	2.7	0	0	4
11	1	18	1	3.1	19	2	3
12	2	19	2	2.5	25	20	1
13	1	21	4	3.5	2	0	6
14	2	21	3	1.8	19	33	2
15	1	42	2	3.9	10	0	2
16	1	19	2	2.3	45	0	1
17	2	21	3	2.8	29	10	5
18	2	25	2	3.1	14	0	6
19	1	21	3	3.5	5	0	1
20	1	17	1	3.5	28	0	3

Level of measurement: Mathematical nature of the values for a variable.

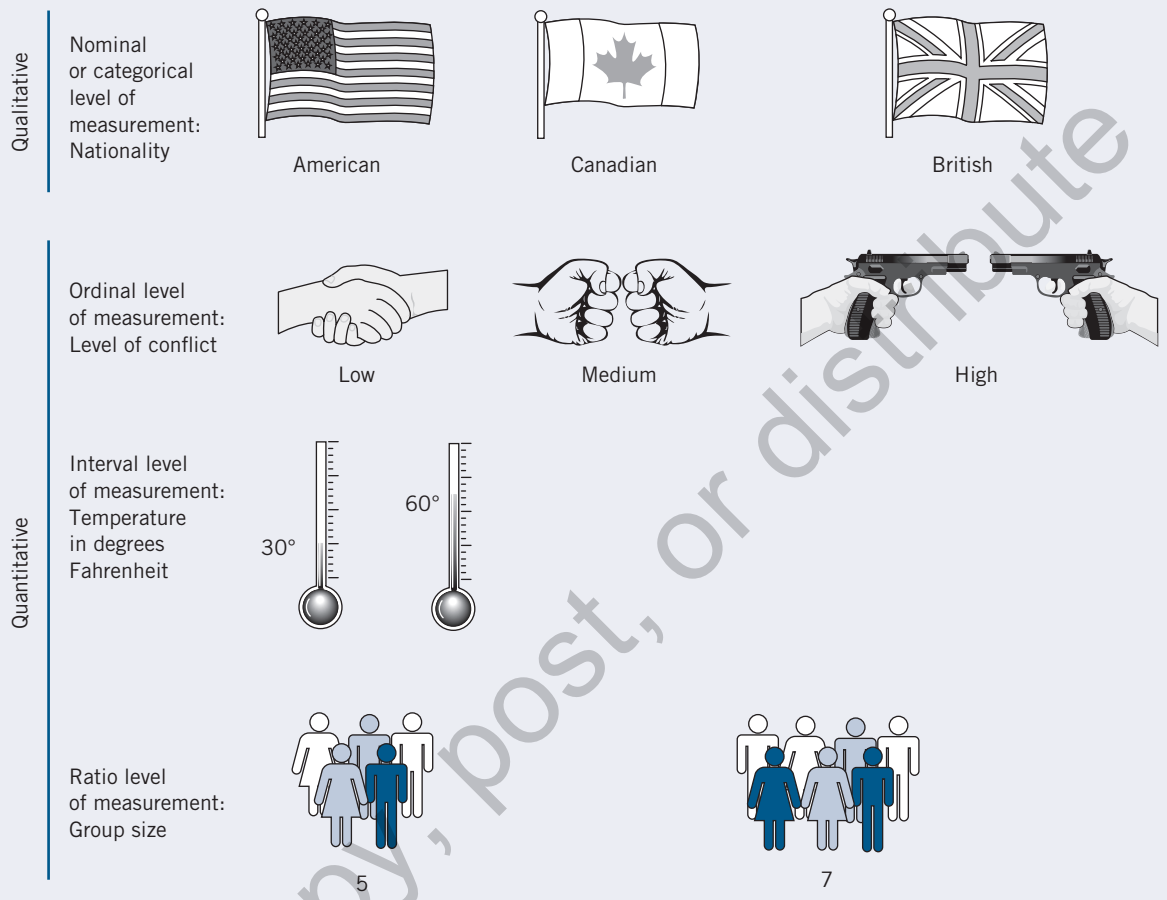
In addition to distinguishing between qualitative and quantitative, we can differentiate among variables in terms of what is called their **level of measurement**. The four levels of measurement are (1) nominal, (2) ordinal, (3) interval, and (4) ratio. Figure 2.1 depicts the difference among these four levels of measurement.

Nominal Level of Measurement

Nominal-level variables: Values that represent categories or qualities of a case only.

Variables measured at the nominal level are exclusively qualitative in nature. The values of **nominal-level variables** convey classification or categorization information *only*. Therefore, the only thing we can say about two or more nominal-level values of a variable is that they are different. We cannot say that one value reflects more or less of the variable than the other. The most common types of nominal-level variables are gender (male and female), religion (Protestant, Catholic, Jewish, Muslim, etc.), and political party (Democrat, Republican, Independent, etc.). The values of these variables are distinct from one another and can give us only descriptive information about the type or label attached to a value. Notice we can say that males are different from females but not that they have more "gender." We can say that Protestants have a different religion than Catholics or Jews, but again, not that they have more "religion." The only distinction we can make with nominal-level variables is that their values are different.

Figure 2.1 Levels of Measurement



Because they represent distinctions only of kind (one is merely different from the other), the categories of a nominal-level variable are not related to one another in any meaningful numeric way. This is true even if the alphanumeric values are converted or coded into numbers. For example, in Table 2.2, the values assigned to the variables gender and religion are given numeric values. Remember, however, that these numbers were simply assigned for convenience and have no numeric meaning. The fact that Catholics are assigned the code of 1 and Protestants are assigned the code of 2 does not mean that Protestants have twice as much religion as Catholics or that the Protestant religion is "more than" the Catholic religion. The only thing that the codes of 1 and 2 mean is that they refer to different religions. Because we cannot make distinctions of "less than" or "more than" with them, then, nominal-level variables do not allow us to rank-order the values of a given variable. In other words, nominal-level measurement does not have the property of order. It merely reflects the fact that some values are different from others. Consequently, mathematical operations cannot be performed with nominal-level data. With our religion variable, for example, we cannot subtract a 2 (Protestant) from a 3 (Jewish) to get a 1 (Catholic). Do you see how meaningless mathematical operations are with variables measured at the nominal level?

Ordinal Level of Measurement

Ordinal-level variables: Values that not only represent categories but also have a logical order.

The values of **ordinal-level variables** not only are categorical in nature, but the categories also have some type of relationship to each other. This relationship is one of order or transitivity. That is, categories on an ordinal variable can be rank-ordered from high (more of the variable) to low (less of the variable) even though they still cannot be exactly quantified. As a result, although we can know whether a value is more or less than another value, we do not know exactly *how much* more or less. The properties of ordinal-level measurement are clearer with an example.

Let's say that on a survey, we have measured income in such a way that respondents simply checked the income category that best reflected their annual income. The categories the survey provided are as follows:

1. Less than $20,000
2. $20,001 to $40,000
3. $40,001 to $60,000
4. $60,001 to $80,000
5. more than $80,001

Now suppose that one of our respondents (respondent 1) checked the first category and that another respondent (respondent 2) checked the third category. We don't know the exact annual income of each respondent, but we do know that the second respondent makes more than the first. Thus, in addition to knowing that our respondents have different annual incomes (nominal level), we also know that one income is more than the other. In reality, respondent 1 may make anywhere between no money and $20,000, but we can never know. Had we measured income in terms of actual dollars earned per year, we would be able to make more precise mathematical distinctions between respondents' annual incomes. Suppose we had a third person (respondent 3) who checked the response of more than $80,000. The property of transitivity says that if respondent 1 makes less than respondent 2, and if respondent 2 makes less than respondent 3, then respondent 1 also makes less than respondent 3. The rank order is thus:

1. Less than $20,000 respondent 1
2. $20,001 to $40,000
3. $40,001 to $60,000 respondent 2
4. $60,001 to $80,000
5. more than $80,001 respondent 3

Other examples of ordinal-level variables include the "Likert-type" response questions found on surveys that solicit an individual's attitudes or perceptions. You are probably familiar with this type of survey question. A typical one follows: "Please respond to the following statement by circling the appropriate number: '1' Strongly Agree, '2' Agree, '3' Disagree, '4' Strongly Disagree." The answers to these questions represent the ordinal level of measurement. Often these categories are displayed like this:

1	2	3	4
Strongly Agree	Agree	Disagree	Strongly Disagree

Response categories that rank-order attitudes in this way are often called *Likert* responses after Rensis Likert, who is believed to have developed them back in the 1930s. There are other ways to measure judgements using a Likert-type response. For example, the aggression questionnaires (AQs) in the literature are designed to measure an individual's propensity to feel anger and hostility (Buss & Warren, 2000). It consists of 34 items, such as "Given enough provocation, I may hit another person," "When people annoy me, I may tell them what I think of them," and "I have trouble controlling my temper." Individuals taking the AQ are asked to respond to the statements using a five-point Likert-type scale from "not at all like me," which is coded 0, to "completely like me," coded 5. Tracey Skilling and Geoff Sorge (2014) used the AQ to assess the validity of two other scales, one intended to measure criminal attitudes and another intended to measure antisocial attitudes. Using a sample of delinquent offenders in Canada, they found that all three measures were significantly related to each other, indicating that they were each measuring antisocial attitudes.

Interval Level of Measurement

In addition to enabling us to rank-order values, **interval-level variables** allow us to quantify the numeric relationship among them. To be classified as an interval-level variable, the difference between values along the measurement scale must be the same at every two points. For example, the difference in temperature on the Fahrenheit scale between 40 degrees and 41 degrees is the same 1 degree difference as the difference between 89 degrees and 90 degrees. This 1 degree difference is the same difference that exists between all the other values on the Fahrenheit scale.

Interval-level variable: In addition to an inherent rank order, a value's numeric relationship to other values is known. There is an equal and constant distance between adjacent values. Therefore, the values can be added and subtracted.

Another characteristic of interval-level measurement is that the zero point is arbitrary. An arbitrary zero means that, although a value of zero is possible, zero does not mean the absence of the phenomenon. A meaningless zero is an arbitrary zero. For example, a temperature on the Fahrenheit scale of 0 degrees does not mean that there is no temperature outside; it simply means that it is cold! Zero degrees on the Fahrenheit scale is arbitrary. These characteristics allow scores on an interval scale to be added and subtracted, but meaningful multiplication and division cannot be performed. This level of measurement is represented in Figure 2.1 by the difference between two Fahrenheit temperatures. Although 60 degrees is 30 degrees hotter than 30 degrees, 60 in this case is not twice as hot as 30. Why not? Because heat does not begin at 0 degrees on the Fahrenheit scale.

Social scientists often treat indices (see the AQ earlier) that were created by combining responses to a series of variables measured at the ordinal level as interval-level measures. Another example of an index like this could be created with responses to the Core Institute's (2020) questions about friends' disapproval of substance use (see Table 2.3). The survey has 13 questions on the topic, each of which has the same three response choices. If Do Not Disapprove is valued at 1, Disapprove is valued at 2, and Strongly Disapprove is valued at 3, then the summed index of disapproval would range from 12 to 36. The average could then be treated as a fixed unit of measurement. So a score of 20 could be treated as if it were 4 more units than a score of 16 and so on.

Ratio Level of Measurement

Ratio-level variables have all the qualities of interval-level variables, and the numeric difference between values is based on a natural, or true-zero, point. A true-zero point means that a score of zero indicates that the phenomenon is absent. For example, if people were asked how many hours they worked last month and they replied "zero hours," it would mean that there was a complete absence of work—they were unemployed that month. Ratio measurement allows meaningful use of multiplication and division, as well as addition and subtraction. We can therefore divide one number by another to form a ratio—hence the name of this level of measurement. Suppose we were conducting a survey of the victimization experiences of residents in

Ratio-level variables: Variables that we assume can be added and subtracted as well as multiplied and divided and that have true-zero points.

Table 2.3 Ordinal-Level Variables Can Be Added to Create an Index With Interval-Level Properties: Core Alcohol and Drug Survey

How Do You Think Your Close Friends Feel (or Would Feel) About You . . . (Mark One for Each Line)	Do Not Disapprove	Disapprove	Strongly Disapprove
a. Trying marijuana once or twice			
b. Smoking marijuana occasionally			
c. Smoking marijuana regularly			
d. Trying cocaine once or twice			
e. Taking cocaine regularly			
f. Trying LSD once or twice			
g. Taking LSD regularly			
h. Trying amphetamines once or twice			
i. Taking amphetamines regularly			
j. Taking one or two drinks of an alcoholic beverage (beer, wine, liquor) nearly every day			
k. Taking four or five drinks nearly every day			
l. Having five or more drinks in one sitting			
m. Taking steroids for bodybuilding or improved athletic performance			

Source: Adapted from *Core Alcohol and Drug Survey: Long Form* 2020 from the Core Institute.

rural areas and asked them to provide their annual income in dollars. This variable would be an example of the ratio-level of measurement because it has both a true-zero point and equal and known distances between adjacent values. For example, a value of no income, "zero dollars," has inherent meaning to all of us, and the difference between $10 and $11 is the same as that between $55,200 and $55,201.

There are a few variables in Table 2.2 that are measured at the ratio level. One is the number of drinks respondents had in an average month. Notice that there were a few respondents who had "0" drinks—this is an absolute zero! And a college student who drinks an average of 20 drinks a month has 10 more drinks than someone who has 10 drinks a month and 10 fewer drinks than someone who has an average of 30 drinks a month. We have not shown you how to calculate the mean yet, but imagine we calculate the average number of drinks a senior in college has from this table and find that it is 10.5 drinks. We then calculate the average number of drinks a first-year student has as 31.75 drinks. Because this is a ratio-level variable with an absolute zero, we could now take the ratio of drinks consumed by a first-year student compared with a senior to be (31.75 / 10.5 = 3.02) and say that first-year students consume about 3 times as much alcohol as seniors! Does this seem accurate to you? Because we can do this, the level of measurement is called "ratio."

For most statistical analyses in social science research, the interval and ratio levels of measurement can be treated as equivalent. In addition to having numerical values, both the interval and ratio levels also involve **continuous measures**. That is, the numbers indicating the values of variables are points on a continuum, not discrete categories. Because of this, researchers often treat variables measured at the interval and ratio levels as comparable. They then refer to this as the **interval–ratio level of measurement**. In this text, we generally rely on this distinction.

Continuous measure: Measure with numbers indicating the values of variables as points on a continuum.

Interval–ratio level of measurement: Variables that we assume can be added and subtracted as well as multiplied and divided regardless of whether they have true-zero points.

FOUR TYPES OF MEASUREMENT

Nominal: Values represent categories or qualities of a case only.

Ordinal: Values not only represent categories but also have a logical order.

Interval: In addition to an inherent rank order, a value's relationship to other values is known. There is an equal and constant distance between adjacent values.

Ratio: Not only can distances be determined between values, but these distances are based on a true-zero point.

The Case of Dichotomies

Dichotomies are variables having only two values and are a special case from the standpoint of levels of measurement. Although variables with only two categories are generally thought of as nominally measured, we can also think of a dichotomy as indicating the presence or absence of an attribute. Suppose, for example, we were interested in differences between individuals who had never used illegal drugs in the last year and those who had used at least one illegal drug in the last year. We could create a variable that indicated this dichotomous distinction by coding those individuals who said they did not use any illegal substance as 0 and those who had used at least one illegal substance as 1. Viewed in this way, there is an inherent order to the two values: In one group, the attribute of consuming illegal substances is absent (those coded 0), and in another, it is present (those coded 1). When we code variables like this as 0 or 1, they are often called **binary variables**.

Dichotomy: Variable having only two values.

Binary variable: Dichotomous variable that has been coded 0 or 1.

Comparing Levels of Measurement

Table 2.4 summarizes the types of comparisons that can be made with different levels of measurement, as well as the mathematical operations that are legitimate. All four levels

Table 2.4 Properties of Measurement Levels

Examples of Comparison Statements	Appropriate Math Operations	Relevant Level of Measurement			
		Nominal	Ordinal	Interval	Ratio
A is equal to (not equal to) B	= (≠)	√	√	√	√
A is greater than (less than) B	> (<)		√	√	√
A is three more than (less than) B	+ (–)			√	√
A is twice (half) as large as B	× (÷)				√

of measurement allow researchers to assign different values to different cases. All three quantitative measures allow researchers to rank cases in order.

WAYS OF PRESENTING VARIABLES

In this section, we examine some of the most commonly used pieces of information you will confront: counts, rates, ratios, proportions, and percentages. These are simply different ways in which to present, describe, and compare variables.

Counts and Rates

Count or frequency: Number of units in the sample that has a particular value in a variable distribution.

The most elementary way of presenting information is to present the counts or frequencies of a variable. A **count** or **frequency** is simply the number of times that a value occurs in your data. The numbers of violent victimizations recorded by the National Crime Victimization Survey (NCVS), which includes rapes, robberies, and assaults, by age for 2018 are presented here:

Age Group	Number of Victims (*f*)
12–17	422,460
18–24	478,740
24–34	650,560
35–49	703,770
50–64	579,770
65 and older	113,030

These numbers tell us exactly how many victims of violent crime there were in the United States in 2018 in each of the six age groups. Violent crimes include rapes, robberies, and assaults. We can see from these data that there were fewer victims in the age 65 and older group than in any other age group. The highest number of victims of violent crime appeared in the 35 to 49 age group (703,770 victims). Based on these counts, who has the greatest vulnerability to becoming a victim? Do those who are aged 35 to 49 have a greater risk of becoming the victim of violence compared with those aged 18 to 24, or those aged 12 to 17? The short answer is no. The long answer is that if we want to make comparisons across different categories, whether they be age categories, gender, race/ethnicity, city, year, or any other aggregation, it is not possible to produce conclusions of relative risk. Why? Because simple counts and frequencies do not take into consideration the size of the total at-risk population within each category. Although we may sometimes come to the same conclusion, using simple frequencies to make these comparisons most often leads to misleading conclusions.

Rate: Number of a phenomenon divided by the total possible, which is then multiplied by a constant such as 1,000, 10,000, or 100,000.

To make comparisons accurately across units with different population sizes, it is important to control for the size of the populations you are comparing. To do this, it is necessary to calculate the **rate** of an occurrence.

Proportions: Number of some value in a variable distribution that is divided by total possible scores.

Percentages: Number of some value in a variable distribution that is divided by total possible scores and then is multiplied by 100.

Proportions and Percentages

Two other common techniques used to present information about variables are **proportions** and **percentages**. These measures are really special kinds of ratios obtained by dividing

CASE STUDY

The Importance of Rates for Victimization Data

Let's assume we want to assess how much risk a person across each age group has of experiencing a violent victimization. Table 2.5 presents the same violent victimization data along with the population counts for each age group. Rates are derived by dividing the observed number of occurrences or phenomena by the total number that could theoretically have been observed within the population of interest. In addition, rates are usually standardized according to some population base, such as a rate per 1,000, 10,000, or per 100,000 people:

$$\text{Rate} = \frac{\text{Number in subset}}{\text{Total number}} \times \text{Constant(e.g.,1,000)} \qquad (2\text{–}1)$$

As formula 2-1 shows, to derive the victimization rate of violence within age categories, we must first divide the number of victims of violent crimes observed within an age group by the total number of potential victims within this age group. This latter number would be the entire population for this age group because, theoretically, everyone in the age group could have become a victim of a violent crime. We then multiply that ratio by some population standard to get a rate per 1,000, or a rate per 10,000 population.

Let's calculate the rate of violent crime for those aged 18 to 24 using the population standard of 1,000:

$$\left(\frac{478{,}740}{27{,}143{,}454}\right) = .0176 \times 1{,}000 = 17.6$$

We obtain a rate of violent crime for those aged 18 to 24 of 17.6 per 1,000 of those aged 18 to 24. When we calculate the victimization rates for each age category displayed in Table 2.5, a very different picture of vulnerability emerges. After standardizing for the size of the at-risk population, we see that those between the ages of 18 and 24 have the highest risk of violent victimization compared with all other age categories. This rate is just a bit larger than that for the aged 12 to 17 age group. Except for those aged 65 years and older, those in the 35 to 49 year age group actually have the lowest rate of victimization.

Let's look at another dramatic example of how a frequency count can mislead you because of differences in population size, whereas a rate will not. Let's assume that in 2018 there were 99 murders in Kansas City, Missouri. In that same year, there were 49 murders in Baton Rouge, Louisiana. From the numbers, Kansas City is more dangerous to live in compared with Baton Rouge; in fact, there were almost twice as many murders there. But before you pack your bags and move to Baton Rouge, stop and think about it. Can you compare these raw frequency counts? No. You can't compare relative risk by using simple frequency counts in this case! In fact, the population of Baton Rouge at the time was only 230,212, whereas the population in Kansas City was nearly a half-million (465,514). Now let's calculate the rate of homicide per 100,000 people in each city:

$$\text{Rate for Kansas City} = \left(\frac{99}{465{,}514}\right) = .000212 \times 100{,}000 = 21.2$$

$$\text{Rate for Baton Rouge} = \left(\frac{49}{230{,}212}\right) = .000212 \times 100{,}000 = 21.2$$

Table 2.5 Violent Crime Victims, Total Population, and Violent Crime Rates per 1,000 by Age Group, 2018

Age Group	Number of Victims	Total Population	Rate per 1,000
12–17	422,460	24,633,684	17.1
18–24	478,740	27,143,454	17.6
24–34	650,560	39,891,724	16.3
35–49	703,770	65,240,931	10.7
50–64	579,770	41,860,232	13.8
65 and older	113,030	34,991,753	3.2

Source: Adapted from *Criminal Victimization, 2018* by Morgan and Oudekerk, 2019, from the Bureau of Justice Statistics, U.S. Department of Justice.

Amazing! The relative risk for becoming a murder victim in both Kansas City and Baton Rouge was the same in 2018. A final analogy that is often used to underscore the notion of relative risk will help cement this point. If you are like us, every time we are about to take off in an airplane, we get a bit nervous. In fact, when Ronet and Ray's son was very young, they occasionally took separate flights to ensure that if the airplane crashed, one of them would be alive to take care of him. Unfortunately, they weren't thinking very critically because they all took the same taxi to the airport even when they were on different flights. The problem in this scenario is that they had a greater risk of being in an accident in the taxi on the way to and from the airport than they did on the flight. On average, flying kills about 200 people a year in the United States, whereas driving kills an average of 32,300 people (National Safety Council, 2020). Let's assume a 2019 U.S. population of 308,745,538 and plug this number into a rate per 100,000 as we have done here:

$$\text{Rate of Death for Flying} = \left(\frac{200}{308,745,538}\right) = .00000065 \times 100,000 = .065$$

$$\text{Rate of Death for Driving} = \left(\frac{32,300}{308,745,538}\right) = .0001046 \times 100,000 = 10.46$$

Ratio: Expresses the relationship between two numbers and indicates their relative size.

As you can see, making risk-based calculation without good data can lead to very flawed assumptions. Remember that a **ratio** is a number that expresses the relationship between two numbers and indicates their relative size. As you saw earlier, the ratio of x to y is determined by dividing x by y. A ratio for the relative risk of dying while driving compared with flying is 10.46 /.065 = 161. Wow! This tells us that the risk of dying while driving is 161 times greater compared with flying. Think about that the next time you get behind the wheel of your car! Buckle up!

---

the number of observations from a subset of your sample by the total number in your sample. In other words, a proportion is obtained by dividing the number of counts for a given event (f) by the total number of events (n). More specifically, proportions are obtained using the following formula:

$$\text{Proportion} = \frac{\text{Number in subset of sample}}{\text{Total number in sample}} = \frac{f}{n} \qquad (2\text{-}2)$$

A proportion may also be called a **relative frequency** because it expresses the number of cases in a given subset (f) relative to the total number of cases (n). In this text, we use the terms "proportion" and "relative frequency" interchangeably.

Relative frequency: See Proportions.

Percentages are obtained simply by multiplying a proportion by 100. This standardizes the numbers to a base of 100, which is generally easier for an audience to interpret:

$$\text{Percent} = \frac{f}{n} \times 100 = \text{Proportion} \times 100 \tag{2-3}$$

Let's go through an example. Using violent victimization data from the NCVS for 2018, Table 2.6 presents the total number of each type of victimization, the total number of each that was reported to police, the proportion reported, and the percent reported to police. If we were attempting to understand the differences in reporting behavior across different types of crimes, comparing the number of crimes reported would not tell us anything about which crime was most likely to be reported. However, examining either the proportion or the percentage columns tells us a great deal. We can easily see that rape and sexual assaults (.25) are the least likely violent crimes to be reported to police. The violent crime most likely to be reported to police is robbery. It is also interesting to note that an equal percentage (45%) of victimizations perpetrated by strangers and intimate partners were reported to police.

Table 2.6 Total Number, Number Reported, Proportion, and Percentage of Crimes Reported to Police by Type of Crime (NCVS 2018)

Type of Crime	Total number (n)	Number Reported (f)	Proportion (f/n)	Percent (f/n) × 100
Violent Crime	6,385,520	2,720,232	0.43	43
Rape/Sexual Assault	734,630	182,923	0.25	25
Robbery	573,100	358,761	0.63	63
Assault	5,077,790	2,183,450	0.43	43
Aggravated	1,058,040	640,114	0.61	61
Simple	4,019,750	1,543,584	0.38	38
Selected Characteristics of Violent Crime				
Domestic Violence	1,333,050	626,534	0.47	47
Intimate Partner	847,230	381,254	0.45	45
Stranger Violence	2,493,750	1,109,719	0.45	45
Violence involving Injury	1,449,530	787,095	0.54	54
Violence involving a weapon	1,329,700	801,809	0.60	60

Source: Adapted from Morgan and Oudeker, 2019, *Criminal Victimization, 2018,* Bureau of Justice Statistics, U.S. Department of Justice.

UNITS OF ANALYSIS

Units of analysis: Particular units or aggregations (e.g., people and cities) that constitute an observation in a data set.

The final issue we discuss in this chapter is often referred to as the unit of analysis. The **units of analysis** are the particular units or objects we have gathered our data about and to which we apply our statistical methods. Stated differently, our units of analysis are whatever constitutes the observations in our data set. For example, are our observations or data points made up of persons? Prisons? Court cases? States? Nations? In research, we employ many different levels of aggregation for research. Sometimes we use questionnaires or interviews to obtain data from individuals. The NCVS, for example, interviews individuals in households from around the United States and asks them about their experiences with criminal victimization. In this particular research, the unit of analysis is the individual or person because the data are obtained from individual respondents, but these data can also be aggregated to the household level.

In other instances, the unit of analysis is a group or collectivity. Often, these data originally were collected from individuals and then combined, or aggregated, to form a collectivity. For example, the Federal Bureau of Investigation (FBI) collects information about the number of crimes reported by individuals to local police departments. However, the FBI aggregates this information, identifying what state the report came from and, in some cases, what city and/or county. Depending on what data you use, then, the unit of analysis may be states, counties, or cities.

As an example of data at the state level of analysis, Table 2.7 presents the percentage change from 2017 to 2018 in the total crime rate by state. Even though the information for this variable originally came from individual law enforcement agencies in each state, the units of analysis in this case are the states because that is how the data are aggregated here.

Table 2.7 Percentage Change in the Total Crime Rate in the U.S. 2017–2018, by State

State	Percentage Change	State	Percentage Change
Alabama	–3.88	Montana	–3.21
Alaska	–4.85	Nebraska	–9.01
Arizona	–7.68	Nevada	–6.76
Arkansas	–6.83	New Hampshire	–9.45
California	–4.42	New Jersey	–10.78
Colorado	–0.34	New Mexico	–8.81
Connecticut	–6.05	New York	–3.64
Delaware	–5.42	North Carolina	–2.83
District of Columbia	1.76	North Dakota	–6.71
Florida	–8.7	Ohio	–9.07

State	Percentage Change	State	Percentage Change
Georgia	–9.79	Oklahoma	–0.16
Hawaii	1.03	Oregon	–1.72
Idaho	–11.16	Pennsylvania	–8.24
Illinois	–3.88	Rhode Island	–5.68
Indiana	–7.9	South Carolina	–5.24
Iowa	–18.28	South Dakota	–8.41
Kansas	–1.76	Tennessee	–4.37
Kentucky	–8.8	Texas	–7.12
Louisiana	–2.8	United States (total)	–6.44
Maine	–9.78	Utah	–13.56
Maryland	–8.66	Vermont	–12.8
Massachusetts	–10.13	Virginia	–7.27
Michigan	–6.65	Washington	–6.13
Minnesota	–9.2	West Virginia	–17.54
Mississippi	–12.8	Wisconsin	–13.02
Missouri	–6.54	Wyoming	–2.72

Source: Adapted from FBI. 2020. 2018 Crime in the United States. Table 4. Crime in the U.S. by Region, Geographic Division, and State, 2017-2018. https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018/topic-pages/tables/table-4

A positive percentage change score indicates that the crime rate increased between 2017 and 2018 and a negative score indicates that it decreased for the given state. As you can see, the total crime rate decreased in virtually all states, which has been the trend in state crime rates for the past several years.

Again, understanding the units of analysis is important when making statistical interpretations from data as you will see in the next chapter. We can only make generalizations about the units of analysis for which our data represent. For example, if we have state-level data and we find that states that have higher rates of poverty also tend to have higher rates of murder, we can only generalize this finding to the states not to counties or cities. Nor can we say that individuals who live under conditions of poverty are more likely to experience a homicide if we only have state-level data. Only if we were analyzing individual victims and offenders could we make statements about individual factors related to lethal violence. We will remind you of this throughout the book!

SUMMARY

We hope that you now have a better understanding of the differences among the types of variables and the many levels of measurement used in research. It is essential that you be fully familiar with these concepts so that you can understand their statistical applications. We have classified the two most general measurement levels of variables as being qualitative and quantitative. Qualitative variables tell us "what kind" or "what category" a variable's value denotes, and the values of quantitative variables give us numeric information regarding "how much" or the "quantity" a value contains. Within these two categories, we have also specified the conditions under which a variable can be defined as measured at the nominal, ordinal, interval, or ratio level. These levels are hierarchical in nature and can be thought of as a sort of quantitative hierarchy. Values of nominal-level variables only differ in kind or quality and have no numerical distinction. Ordinal variables have values with a rank order. In addition to an inherent rank order, the distance between categories of an interval-level variable has a known and constant value. And finally, not only can distances be determined between values of a ratio-level variable, but also these distances are based on a true-zero point. The remainder of the chapter examined the differences between simple counts of a phenomenon (referred to as frequencies) and rates, ratios, proportions, and percentages. The final section discussed the units of analysis used in research.

KEY TERMS ➤ REVIEW KEY TERMS WITH eFLASHCARDS.

alphanumeric data 27
binary variable 33
categorical variable 26
constant 25
continuous measure 33
continuous variable 27
count 34
dichotomy 33
distribution 000
empirical distribution 25
frequency 34
frequency distribution 25
interval 000
interval-level variable 31
interval–ratio level of measurement 33
level of measurement 28
nominal 000
nominal-level variable 28
numeric data 27
ordinal 000
ordinal-level variable 30
percentages 34
proportions 34
qualitative variable 26
quantitative variable 27
rate 34
ratio 36
ratio-level variable 31
relative frequency 37
units of analysis 38
variable 24

KEY FORMULAS

Rate (equation 2-1):

$$\text{Rate} = \frac{\text{Number in subset}}{\text{Total number}} \times \text{Constant (e.g., 1,000)}$$

Proportions (equation 2-2):

$$\text{Proportion} = \frac{\text{Number in subset of sample}}{\text{Total number in sample}} = \frac{f}{n}$$

Percentages (equation 2-3):

$$\text{Percent} = \frac{f}{n} \times 100 = \text{Proportion} \times 100$$

PRACTICE PROBLEMS

1. For each of the following variables, define the level of measurement as either qualitative or quantitative and, further, as one of the four more distinct levels: nominal, ordinal, interval, or ratio:
   a. A convicted felon's age in years
   b. A driver's score on the Breathalyzer exam
   c. The fine for a parking ticket

d. The specific offense code of a felony

e. A defendant's gender

f. Fines levied on industrial companies convicted of violating the Clean Air Act

2. What distinguishes a variable measured at the ordinal level from a variable measured at the interval level of measurement? What more does the ratio level of measurement add to this?

3. In a study examining the effects of arrest on convicted drunk drivers' future drunk-driving behavior, which is the independent variable and which is the dependent variable?

4. If we are interested in determining the extent to which males and females are more or less afraid to walk outside alone at night, which variable would we designate as our independent variable and which as our dependent variable?

5. To compute a rate of violent crime victimizations against people 14–18 years old, what would we use as the numerator and what as the denominator?

6. What are the advantages of rates over frequency counts? Give an example.

7. From the following table, compute the proportions and percentages of the household crime victimizations that were reported to the police by the loss value of the victimization:

	f	Proportion	%
Less than $10	16		
$10–$49	39		
$50–$99	48		
$100–$249	86		
$250–$999	102		
$1,000 or more	251		
	$n = 542$		

8. Javier Ramos and Marin Wenger (2019) compared re-offending rates between formerly incarcerated native- and foreign-born individuals. They found that immigrants were much less likely to reoffend as compared to their native-born counterparts. In this study, what were the units of analysis? What were the independent and dependent variables?

9. To test the existence of a relationship between unemployment and crime, we use data from 50 states of the United States. What are the units of analysis? What would you select to be the independent variable? What would you deem to be the dependent variable?

10. Suppose we are interested in the amount of time police departments took to respond to reports of crime. We track response times for several police departments within large metropolitan areas to see whether there are any differences based on the location of the jurisdiction. In this study, what are the units of analysis?

SPSS EXERCISES

Data for Exercise	
Data Set	**Description**
2012 states data.sav	This data set compiles official statistics from various official sources, such as the census, health department records, and police departments. It includes basic demographic data, crime rates, and incidence rates for various illnesses and infant mortality for entire states.

Variables for Exercise	
Variable Name	**Description**
State	The state from which data were collected.
Murdercat	An indicator for the rate at which murders occur.
perindpoverty	The proportion of individuals below the poverty line.
BurglaryRt	An indicator for the rate at which burglaries occur.
MedianIncome	The median income for a state.

1. First, take a look at the description and variable information in "variable view" in the 2010 states data set. What is the unit of analysis for this data?
2. In variable view, find the following variables and look at the variable and value labels. For each variable, identify if it is a (1) quantitative or qualitative measure, (2) its level of measurement, and (3) if appropriate, the unit that it is measured in:
   a. State
   b. Murdercat
   c. perindpoverty
   d. BurglaryRt
   e. MedianIncome
3. In data view, find Alabama and Alaska (rows 1 and 2). What is the burglary rate for each state? Can you think of any reason for the differences between the two states?
4. **Creating a frequency table in SPSS:** We want to create a frequency table for the variable Murdercat. To do this select analyze->descriptives->frequencies and put Murdercat (which is listed as "Murder rate categorical [Murdercat]) into the box on the right. Looking at the output window, what proportion of states have 0–3 murders per 100,000? What proportion have fewer than 9 murders per 100,000?
   a. **Sorting data in SPSS:** Let's identify which state(s) have the highest murder rate by sorting the Murdercat variable. To do this, go into data view and scroll until you find the variable Murdercat. Right click on the variable name for Murdercat and select "sort descending," which will sort the data so high values are on top. You can also sort cases by selecting data->sort cases and then placing Murdercat in the box. Now scroll left. What state(s) have the highest murder rates?

EXCEL EXERCISES

Data for Exercise	
Data Set	**Description**
2012 states data.xls	This data set compiles official statistics from various official sources, such as the census, health department records, and police departments. It includes basic demographic data, crime rates, and incidence rates for various illnesses and infant mortality for entire states.

Variables for Exercise	
Variable Name	**Description**
State	The state from which data were collected.
Murdercat	An indicator for the rate at which murders occur.

Variables for Exercise	
Variable Name	**Description**
perindpoverty	The proportion of individuals below the poverty line.
BurglaryRt	An indicator for the rate at which burglaries occur.
MedianIncome	The median income for a state.

1. In the main screen, find Alabama and Alaska (rows 2 and 3). What is the burglary rate for each state? Can you think of any reason for the differences between the two states?

2. **Creating a frequency table in Excel:** We want to create a frequency table for the variable Murdercat. To get a frequency table in Excel, one must know the values used to assign categories. In our case, that is 0, 1, 2, 3, and 4. Now, enter these values as individual rows to the right of your data. This is required so Excel knows what values it is looking for to summarize. You may also wish to create column heads for organizational purposes. There are two ways to get a frequency table in Excel. First is the point and click method. To do this, click the "Formulas" tab, then "More Function", hover over "Statistical", and then scroll down and click "FREQUENCY". When you do this, there are two empty boxes titled "Data_array" and "Bins_array". Data array is the data you wish to summarize and bins array are the categories Excel uses to categorize the data frequencies. So, for our example, Murdercat is the data array and 0-4 is the bins array. In "Data_array" you want to include all of the data for the column "Murdercat" do this by clicking on the first row of Murdercat (cell AD2 in our case), click and hold in the "Shift" button, and scroll down to the last data entry for Murdercat (cell AD51). The Data array row should say "AD2:AD51". Next, click "Bins_array" and click on the cell containing "0", click and hold "Shift" and click on the last bins cell containing "4". Bins tab should now contain the cells you placed the categories in; for our example above it is "AF2:AF6". Then click "OK". The second method is to click on the box next to the first categorizing number ("0" in our case) and type "=FREQ" which will automatically bring up "FREQUENCY". Click the "tab" key. Highlight the cells in the column you wish to analyze like before for the Data array part. Enter a comma ",". Then the bins array similar to before and conclude with entering a parenthesis ")". Screenshotted output is available online. Click enter.

   Now, we may want percentages and cumulative percentages to go along with these frequencies. To do so, under the "Percent" column we created, click on the cell to the right of "17". Type in "=", and click on the cell containing 17 (cell AG2 in our example). We have a sample of 50 so now type in "/50" to divide 17 by 50 to get the percent. Click enter. Conveniently, Excel does not require us to redo this operation for each row. Simply click on the cell containing ".34", go to the bottom right where there is a square, click on it without releasing, and drag it down to the row containing "4" and then release. Excel applied the percentage equation to each row for us.

   Now for cumulative percent, type in .34 as that is the cumulative percent thus far. Next, click on the cell to the right of ".42" (cell AI3 in our example). Type in "=", click on the cell to the left which would be cell "AH3", type in "+", and finally click on the cell above containing ".34" (cell AI2). Click enter. Click enter. Similar to before, click on the cell containing the cumulative percent of .76, click and drag it to the row containing "4". This is the beauty of Excel, once you apply a formula you can simply click and drag it to apply to the rest of the column.

   a. **Sorting data in Excel:** Let's identify which state(s) have the highest murder rate by sorting the Murdercat variable. To do this, find the variable Murdercat and click the lettered box above it to highlight the entire column. Click on the "Data" tab, and then click the button with A on top and Z on the bottom. The data are in descending order. Note, however, that you must know what "4" was coded as to understand what it represents. Now scroll left. What state(s) have the highest murder rates?

STATA EXERCISES

Data for Exercise	
Data Set	**Description**
2012 states data.dta	This data set compiles official statistics from various official sources, such as the census, health department records, and police departments. It includes basic demographic data, crime rates, and incidence rates for various illnesses and infant mortality for entire states.

Variables for Exercise	
Variable Name	**Description**
State	The state from which data were collected.
Murdercat	An indicator for the rate at which murders occur.
perindpoverty	The proportion of individuals below the poverty line.
BurglaryRt	An indicator for the rate at which burglaries occur.
MedianIncome	The median income for a state.

1. First, take a look at the description and variable information. What is the unit of analysis for this data?
2. Find the following variables and look at the variable and value labels. For each variable, identify if it is a (1) quantitative or qualitative measure, (2) its level of measurement, and (3) if appropriate, the unit that it is measured in:
   a. State
   b. Murdercat
   c. perindpoverty
   d. BurglaryRt
   e. MedianIncome
3. In data editor, find Alabama and Alaska (rows 1 and 2). What is the burglary rate for each state? Can you think of any reason for the differences between the two states?
4. **Creating a frequency table in Stata:** We want to create a frequency table for the variable Murdercat. To get the frequency statistics in Stata, simply type into the command line "tabulate" and "Murdercat". Looking at the output window, what proportion of states have 0–3 murders per 100,000? What proportion have fewer than 9 murders per 100,000?
   a. **Sorting data in Stata:** Let's identify which state(s) have the highest murder rate by sorting the Murdercat variable. To do this, go into data editor and scroll until you find the variable Murdercat. Right click on the variable name for Murdercat and select "Data", "Sort data" which will sort the data in ascending order. Now scroll left. What state(s) have the highest murder rates?

STUDENT STUDY SITE

Access the eFlashcards, data sets, and software output for SPSS, Excel, and Stata at **edge.sagepub.com/bachmansccj5e**.