INTRODUCTION TO STATISTICS

Learning Objectives

After reading and studying this chapter, you should be able to do the following:

- Define samples and populations
- Define descriptive statistics and inferential statistics
- Define scales of measurement
- Provide two or three examples of each scale of measurement
- Identify numeric characteristics of each scale of measurement

Why is statistics important? What do you get out of it? Before you spend a semester suffering through a statistics course, the least I can do is to shed some light on these two urgent questions. Statistics is an essential process behind how people make scientific advances and sound management decisions based on data. Statistics can be applied to many fields such as research, medicine, engineering, social sciences, education, management, as well as our daily life. Having knowledge in statistics provides you with necessary tools and conceptual framework in quantitative thinking to extract useful information out of raw data. The job market for people with statistics skills is growing. Many of my former students told me that statistics knowledge and skills set them apart from their coworkers and is the reason for their steady paychecks.

Alright! Consider yourself sufficiently buttered up. Now, please endure some basic but important stuff. Statistics is a science that deals with collecting, organizing, summarizing, analyzing, and interpreting numerical data. It is especially useful when using numerical data from a small group (i.e., a sample) to make inferences about a large group (i.e., a population). Learning statistics is similar to learning a new language, and understanding the vocabulary, meaning, and structure is critical from the very beginning. Therefore, definitions, numerical characteristics, and symbols for both samples and populations are clearly presented and explained in this book. Scales of measurement are introduced to show how each scale measures and describes different numerical attributes. After reading this chapter, you should have learned how to differentiate samples from populations as well as understand specific numerical attributes of scales of measurement.

WHAT IS STATISTICS?

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data. **Data** are defined as factual information used as a basis for reasoning, discussion, or calculation. The use of statistics originated in collecting data about states or communities as administrators studied their social, political, and economic conditions in the mid- to late 17th century. Statistics are used to describe the numerical characteristics of large groups of people. Understanding statistics is useful in answering questions such as the following:

- What kind of data need to be collected?
- How many data are needed to provide sufficient guidance?
- How should we organize the data?
- How can we analyze the data and draw conclusions?
- How can we assess the strength of the conclusions?

Population Versus Sample

Learning statistics is very similar to learning a new language. Before you can master a new language, there are basic vocabularies you need to memorize and understand. The basic vocabulary of statistics starts with terms describing the numerical characteristics of a **population** and numerical characteristics of a **sample**. A population is defined as the entire collection of everyone or everything that researchers are interested in measuring or studying. A sample is defined as a subset of the population, from which measures are

actually obtained. For example, a Midwest urban university wants to gauge students' interest in the possibility of establishing a football team on campus. The population, in this example, is the number of students in the entire student body, $N = 17{,}894$. N, uppercase, refers to the size of the population; a lowercase n is used for sample size. It would be extremely time-consuming and very expensive to ask every student in the university his or her interest in having a university football team. No matter how hard you tried, it was highly likely that some people would simply refuse to answer survey questions.

To obtain information efficiently, the university administration assigned the task to the Student Government Association. This association put a survey on the university website and collected responses from 274 students. This group of 274 students was the sample from where measures were actually obtained. The answers from these 274 students formed the data. Data are required to run statistical analyses. Collecting and sorting data usually happen before calculating statistics. Both a population and a sample contain numerical attributes that researchers want to investigate. The numerical attributes of a population are called **parameters**, and they are usually denoted by Greek letters. Numerical attributes of a sample are called **sample statistics**, and they are usually denoted by ordinary English letters. Table 1.1 shows the basic vocabulary to describe populations and samples. It is important to master the basic vocabulary, so as to make effective communication and clear understanding of statistical concepts achievable. In this chapter, we cover the definitions of some basic population parameters and sample statistics. The corresponding mathematical formulas will be covered in the following chapters.

Descriptive Statistics and Inferential Statistics

Statistical procedures can be generally classified into two categories: (1) *descriptive statistics* and (2) *inferential statistics*. **Descriptive statistics** are statistical procedures used to describe, summarize, organize, and simplify relevant characteristics of data. They are simply ways to describe and understand the data, but they do not go beyond the data. Mean, standard deviation, and variance are examples of descriptive statistics.

TABLE 1.1 ⬢ The Basic Vocabularies to Describe Numerical Characteristics of a Population and a Sample

Meaning of the Symbol	Population Parameters	Sample Statistics
Mean	μ (pronounced as 'mew')	$\bar{X}$
Standard deviation	σ (pronounced as 'sigma')	s
Variance	σ^2	s^2
Size, number of observations	N	n

Inferential statistics are statistical procedures that use sample statistics to generalize to or make inferences about a population. As a rule, inferential statistics are more complicated than descriptive statistics, and they will be introduced in Chapter 7 of this book. Such statistical procedures usually require samples to be unbiased representatives of the population. This can be achieved if samples are randomly selected from the population. As there are different ways to slice a pie, there are different ways to sample a population.

Sampling a Population

Let's start this section with examples of a population. For instance, all the cars built by an automobile manufacturer, the entire student body in a university, and every citizen in a particular country are all examples of a population. As you can imagine, the number of members in a population can get enormous and become impossible to measure or study. Sometimes, it is simply not feasible to use the population for research purposes. You understand why an automobile manufacturer does not use every car they make to conduct the crash safety test. There will only be crashed cars left! Therefore, scientists have to conduct their research with a carefully selected sample. For example, a small number of cars used to conduct a crash safety test, an online survey administrated to evaluate students' preference of living on campus versus off campus, public polling on a politician's popularity, and a survey on the most memorable advertisement during a Super Bowl: All of these are examples of conducting research using samples. The reason to use a sample to conduct research instead of using the population is that it is neither possible nor feasible to gain access to every member in the population and have everyone consent to participate in the research. The most obvious difference between a population and a sample is its size. Many samples can be selected from a population.

Two basic ways of sampling a population are nonprobability and probability sampling—examples are *convenience samples* and *random samples*, respectively. A **convenience sample** is one in which researchers use anyone who is willing to participate in the study. A convenience sample is created based on easy accessibility. For example, a student completes an assignment on measuring people's perceived job satisfaction by begging her Facebook friends to fill out a job satisfaction questionnaire. Various talent shows (e.g., *American Idol*, *The Voice*, *Dancing With the Stars*, *America's Got Talent*, etc.) on television ask the audience to vote for their favorite contestants, so they can move on to the next level of the competition. A convenience sample is a nonprobability sample because not all members in the population have an equal chance of being included in the sample. Well, you know that not everyone uses Facebook and not everyone watches talent shows.

A **random sample** is an ideal way to select participants for scientific research. A random sample is defined as being one in which every member in the population has an equal chance of being selected. Because of the random nature of the selection process, it creates an unbiased representation of the population. However, this ideal is easier said than done.

Can you think of an example of a random sample? This is a question I pose to my students in Introductory Statistics classes. An overwhelming majority of students' answers actually fall in the category of convenience samples, such as exit polling during an election, spot surveying in a shopping mall/library/cafeteria, using students in a class, and so on.

Random sampling does not just happen. It actually requires thoughtful planning and careful execution. How do we select a random sample? I'm glad you asked!

POP QUIZ

1. A human resource manager created an employee job satisfaction survey and posted the link on the company's website to invite employees to participate in the survey. This manager is likely to get a _____sample.
   a. convenience
   b. random
   c. probability
   d. population

RANDOM SAMPLING METHODS

There are four commonly used methods to select a random sample: (1) simple random sampling, (2) systematic sampling, (3) stratified sampling, and (4) cluster sampling. The definition of each method and a practical example of each method will be provided in the following subsections.

Simple Random Sampling

A **simple random sample** of a sample size n is created in a way that all samples with the same sample size have the same chance of being selected. A simple random sample has a stronger requirement than a random sample. Each individual is chosen randomly, entirely by chance, and each subset of n individuals has the same probability of being chosen as any other subset of n individuals. For instance, a university was considering switching courses from four credit hours to three credit hours and the university administration sought students' opinions before making such changes. The targeted population is the entire student body in the university (N = 17,500). The administration would like to select 2% of the students to participate in a survey (n = 350). A simple random sample can be obtained by listing all students' names and giving each a unique identification number ranging from 1 to 17,500. A computer can then be used to randomly generate 350 numbers between 1 and 17,500. This creates a sample with n = 350. Such a

procedure can be repeated many times to create many samples. Any one of these samples with $n = 350$ has an equal chance of being selected to participate in the survey.

Systematic Sampling

A **systematic sample** is obtained by selecting a sample from a population using a random starting point and a fixed interval. Typically, every "*k*th" member is selected from the population to be included in the sample. Systematic sampling is still thought of as being random, as long as the interval is determined beforehand and the starting point is selected at random. To choose a 2% systematic sample in our previous example, the university needs to randomly select a number between 1 and 50 as its starting point and then choose every 50th number to be included in the sample. For instance, to create a random sample with 350 students, Number 28 was randomly chosen as the starting point. The sample is obtained by picking every 50th student after that—meaning that we chose the 28th, 78th, 128th, 178th, 228th, and so on until 350 students are selected in the sample.

Stratified Sampling

Stratified sampling works particularly well when there are large variations in the population characteristics. Stratification is the process of grouping members of the population into relatively homogeneous subgroups before sampling. These homogeneous subgroups are called "strata." The strata need to be mutually exclusive: Every member in the population must be assigned to only one stratum. The strata should also be collectively exhaustive—meaning that no member in the population is excluded. A random sample from each stratum is independently taken in the same proportion as the stratum's size relative to the population. These subsets of the strata are then pooled to form a random sample. Therefore, the distribution of the key characteristics is the same as that in the population. For example, there are 60% female students and 40% male students in the university; to create a sample size of 350, a stratified sampling will randomly choose $60\% \times 350 = 210$ female students and $40\% \times 350 = 140$ male students.

Cluster Sampling

Cluster sampling works best when "natural" grouping (clustering) occurs in the population. Random sampling is conducted to select particular clusters to include in the sample. Once a cluster is selected in the sample, all individuals in the cluster are included in the sample. In the previous university example, there are, say, 900 courses offered in any given semester. A list of all 900 classes could be obtained from the registrar's office. A random sampling procedure is conducted to select 2% of classes to be included in the sample ($n = 18$ classes). Once the 18 classes are selected, every student in each of these classes is included in the sample. In this particular example, students need to be reminded not to answer the survey multiple times if more than one of the classes in which they are enrolled get selected in the sample.

Although all four random sampling methods are designed to obtain random samples in the university's effort to seek opinions from students, it is important to recognize that not everyone who is selected in the sample completes the survey. A low response rate ruins the nature of probability sampling, especially when there are systematic differences between students who take time to answer the survey versus students who either neglect or refuse to answer the survey.

POP QUIZ

2. When a large school district randomly selected three classes to conduct a learning environment study, every student in these three classes participated in a face-to-face interview with the school psychologist. This is an example of _____

   a. simple random sampling.
   b. systematic sampling.
   c. stratified sampling.
   d. cluster sampling.

SCALES OF MEASUREMENT

Researchers study physical or psychological characteristics by measuring them or asking questions about the attributes. In statistics, a **variable** refers to a measurable attribute. These measures have different values from one person to another, or the values change over time. Different values of a variable provide information for researchers. If there is no variation in the measures, the variable does not contain any information. For example, when I studied the variables that might be related to students' performance in a statistics course, I collected answers to the following questions: student's grade point average (GPA), number of mathematical courses taken at the college level, number of classes missed in the statistics course, number of tutoring sessions attended for the statistics course, and gender. The answers to these questions reflected the attributes of the students, and they usually varied from one individual to the next. I did not have to ask, "Are you currently enrolled in a statistics course?" Enrolling in a statistics course is an unnecessary question because only students who enrolled in a statistics course were selected for this study. A question producing identical answers from everyone provides no information at all. Only when answers to questions change from one person to the next do they provide information that we do not have before the questions are answered. Variables have different numerical values that can be analyzed, and hopefully, meaningful information can be extracted from them.

Scales of measurement provide a way to think systematically about the numerical characteristics of variables. **Scales of measurement** specifically describe how variables are defined and measured. Each scale of measurement has certain mathematical properties that determine appropriate applications of statistical procedures. There are four scales of measurement: (1) *nominal*, (2) *ordinal*, (3) *interval*, and (4) *ratio*. I will discuss them in the order from the simplest to the most complex. Remember that a higher level scale of measurement contains all the mathematical properties from a lower level scale of measurement plus something more.

Nominal Scale

In **nominal scales**, measurements are used strictly as identifiers, such as your student identification number, phone number, or social security number. The numbers on athletes' jerseys, for example, are simply used as identifiers. Among other things, jersey numbers allow referees to identify which player just committed a personal foul so a penalty can be properly assessed. In social sciences and behavioral sciences, nominal variables such as gender, race, religion, socioeconomic status, marital status, occupation, sexual orientation, and so on are often included in research. Nominal scaled variables allow us to figure out whether measurements are the same or different. The mathematical property of the nominal scale is simply $A = B$ or $A \neq B$.

Ordinal Scale

In **ordinal scales**, measurements not only are used as identifiers but also carry information about ordering in a particular sequence. The numbers are ranked or sorted in an orderly manner such as from the lowest to the highest or from the highest to the lowest. For example, the first time you are invited to a friend's house for dinner, at the outset, you need to find the house. Luckily, you have the address. You notice that houses on one side of the street have odd numbers, and houses on the other side have even numbers. The house numbers either increase or decrease as you walk down the street. Yes! House numbers are arranged in order, so they are ordinal. Therefore, you know 2550 is located in between 2500 and 2600. Can you imagine how confusing it would be trying to find a house for the first time if house numbers were nominal and displayed in random order? Another example is that at a swimming meet, gold, silver, and bronze medals are awarded to swimmers who finish in first, second, and third place. Although we don't know the time difference between the gold medalist and the silver medalist or the time difference between the silver medalist and the bronze medalist, we are sure that the gold medalist is faster than the silver medalist, and the silver medalist is faster than the bronze medalist.

Likert scales are often used to measure people's opinions, attitudes, or preferences. Likert scales measure attributes along a continuum of choices such as 1 = *strongly disagree*, 2 = *somewhat disagree*, 3 = *neutral*, 4 = *somewhat agree*, or 5 = *strongly agree* with each

statement. The ratings from Likert scales belong in the category of ordinal scales. We know the rankings of the responses, but we are not sure whether the difference between 1 and 2 is the same as the difference between 2 and 3. For example, Melissa rates her satisfaction with an online purchase she made a week ago: 1 stands for *very unhappy*, and 5 stands for *very happy*. Melissa's answers are reflected by the bolded numbers in the table.

Characteristic	Very Unhappy	Somewhat Unhappy	Neutral	Somewhat Happy	Very Happy
Quality of product	1	2	3	4	**5**
Delivery time	1	2	3	**4**	5
Competitive price	1	2	**3**	4	5
Customer service	1	2	3	**4**	5

In this particular case, we learn that Melissa is *very happy* with the product quality, *somewhat happy* with the delivery time and customer service, and *neutral* on the pricing. But we don't know how much happier Melissa is with the product quality than with its price. It is not possible to quantify the difference between two numbers from an ordinal scale.

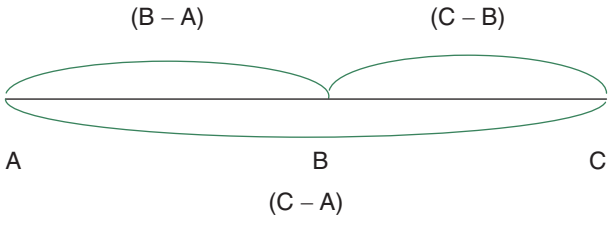
In summary, order contains information. Order gives us the direction of the rankings between two values. The mathematical properties of ordinal scales include everything a nominal scale has (i.e., identifiers) and something more: the direction of the rankings. The mathematical property of the ordinal scale is expressed as if A > B and B > *C*, A > C.

Interval Scale

As measurements evolve to be more sophisticated, scientists go beyond merely figuring out the direction of the rankings. Being able to calculate the amount of the difference becomes the primary objective. The interval scale is designed to fulfill that particular objective. **Interval scales** can be used as identifiers, showing the direction of the rankings and something more: equal units. Interval scales not only arrange observations according to their magnitudes but also distinguish the ordered arrangement in equal units. When measurements come with equal units, they allow us to calculate the amount of difference or the distance between two measurements. Fahrenheit and Celsius temperature scales are two of the most commonly mentioned examples of interval scales, for which equal units exist and zero is an arbitrarily assigned measurement. We know that 0 degrees Celsius equals 32 degrees Fahrenheit. Zero in an interval scale simply represents a measurement.

Equal units have very important implications in statistics. Calculation of means and standard deviations, which are introduced in Chapter 3, require the variables to have equal units. Many measures in social science are interval scales, such as age, test scores, or standardized intelligence quotient (IQ) scores.

FIGURE 1.1 ⬢ Calculating Differences Among Three Values of an Interval Scale



Interval scales allow us to do mathematical operations on the amount of difference between two values. The mathematical property of the interval scale is expressed as when A < B < C, the difference between A and C is the sum of the difference between B and A plus the difference between C and B, (C – A) = (B – A) + (C – B), as shown in Figure 1.1.

For example, age is an interval variable. The ages of three siblings in a family are 8, 13, and 17. The age difference between the youngest and the oldest is 17 – 8 = 9, which equals the sum of the differences between the oldest and the middle one, 17 – 13 = 4, plus the difference between the middle one and the youngest, 13 – 8 = 5.

Ratio Scale

The ratio scale is the most sophisticated scale of measurement. The **ratio scale** contains everything that lower level scales of measurement have, such as identifiers, direction of rankings, and equal units, but it also has something more: an **absolute zero**. Absolute zero means a complete absence of the attribute that you are measuring. Zero is not an arbitrarily assigned number. Most physical measurements, such as height and weight, belong in the ratio scales. Income is a good example of a ratio scale in the social sciences. Zero income means a complete absence of income.

Absolute zero makes it possible to measure the absolute amount of the variable, and it allows us to compare measures in terms of ratios. This unique mathematical property of ratio scales is expressed as if A = 2B and B = 2C, then, A = 4C.

After discussing scales of measurement, sometimes students have questions about why exam scores and IQ scores are classified as interval scales instead of ratio scales. Let's discuss the reasons why exam scores and IQ scores are classified as interval scales. The purpose of an exam is to assess your knowledge in a specific topic. The purpose of an IQ test is to measure and quantify intelligence. When an exam or IQ score turns out to be zero, it does not mean that the person completely lacks knowledge or intelligence. There might be other reasons to explain the zero, such as the test is in a language foreign

to the test taker, the test taker is provided a wrong scantron for the test, or the test taker marks the answers on the wrong side of the scantron. The point is that exam or IQ scores do not have an absolute zero to show complete absence of the attribute; therefore, they belong in the category of interval scales.

It is useful to summarize all four scales of measurement with their mathematical properties in Table 1.2. This table also illustrates that the higher level scales of measurement have every attribute that the lower level scales have, plus something more. It is very common to label variables with nominal or ordinal scales of measurement as categorical variables and variables with interval or ratio scales of measurement as continuous or numerical variables.

TABLE 1.2 ⬢ Mathematical Properties of Scales of Measurement

Scale of Measurement	Identifier	Direction of Ranking	Equal Units	Absolute Zero
Nominal	X			
Ordinal	X	X		
Interval	X	X	X	
Ratio	X	X	X	X

POP QUIZ

3. Travel speed is measured by the distance traveled divided by the time traveled. This is an example of a(n) _____ scale of measurement.
   a. nominal
   b. ordinal
   c. interval
   d. ratio
4. The answers to a marital status question are usually married, single, divorced, or cohabiting. This is an example of a(n) _____ scale of measurement.
   a. nominal
   b. ordinal
   c. interval
   d. ratio

VARIABLE CLASSIFICATIONS

There are other ways to classify variables. Some are related to the scales of measurement, but others are related to the research designs. We will cover the essential ones.

Discrete Versus Continuous Variables

Discrete variables are made of values that have clear separation from one number to the next. The answers for discrete variables can only be integers (i.e., whole numbers). For example, how many cars do you own? The answer can simply be counted as the number of cars that are registered under your name. The answers are likely to be 0, 1, 2, 3, and so on. You can't own 2.56 cars. The same goes with the answer to "How many pets do you have?" I hope, for your sake, you don't own 1.25 cats. Clearly, nominal and ordinal scales can also be classified as discrete variables.

Continuous variables are composed of values that do not have clear separation from one number to the next. There are numerous possible answers between two adjacent integers for continuous variables. Continuous variables are usually expressed with decimals or fractions. For example, the current world record in men's 50-meter freestyle was 20.16 seconds. Official swim time is measured at 100th of a second. Physical measures such as height and weight are similarly continuous variables. As a person who is "vertically challenged" (a politically correct way to say "short"), it is very important for me to list my height with as many numbers after the decimal point as possible given the limits of the preciseness of the ruler. Interval and ratio scales can also be classified as continuous variables.

Independent Variables Versus Dependent Variables

The distinction between independent variables and dependent variables is very important in experimental research. Generally speaking, there are three different types of empirical research: (1) *experimental*, (2) *quasi-experimental*, and (3) *nonexperimental*. **Experimental research** is usually conducted in a tightly controlled environment (i.e., research laboratories). Researchers deliberately choose variables to manipulate at different levels so as to investigate their effects on the variables that researchers are really interested in. The variables that are deliberately manipulated by researchers are called **independent variables**, predictor variables, or explanatory variables. **Dependent variables** are what researchers are really interested in studying. Dependent variables are also called criterion variables or response variables. Measuring and/or observing the changes in dependent variables as a result of the deliberate manipulation of the independent variables is the foundation of experimental research.

For example, researchers studied the effect of caloric consumption on the longevity of rats. Researchers selected a sample of 20 pairs of newborn rats in a lab. Each pair of rats was from the same mother. The researchers randomly assigned one in each pair to the control group and the other to the experimental group. The control group was fed a normal amount of calories every day. The experimental group was fed only 70% of the normal amount every day. At the end, researchers compared the life span of rats in the

control group versus that of the experimental group. In this experiment, the rats' living environment was strictly controlled and kept constant by the researchers in terms of light, temperature, moisture, and sound. The independent variable was the amount of calories provided every day, which was deliberately manipulated by the researchers to set at 100% for the control group versus 70% for the experimental group. The dependent variable was the life span of the rats, which was the focus of this study. Researchers were really interested in finding out how variation in the amount of calories affected rats' life span by comparing rats in the control group with those in the experimental group.

In summary, experimental research is scientific research that has achieved three important features: (1) control to keep everything else constant, (2) manipulation of the independent variable, and (3) random assignment of participants to different conditions.

"Quasi" means "almost but not really." **Quasi-experimental research** has some but not all of the features of experimental research. More specifically, if one or more important features are not feasible but others remain intact, the research becomes quasi-experimental. For example, a school teacher investigated the effects of two instruction methods on sixth graders' science proficiency. One class was assigned to use inquiry through which students were supposed to find answers to their own questions while the instructor helped them design ways to do so. The other class was assigned to use standard lectures to teach science. Let's compare this study with the calories on rats' longevity study. The school study occurred in a school setting, not a research laboratory. These two groups of students were exposed to the particular instruction methods only during the science course instead of 24 hours a day. Both studies had manipulations on the independent variables. However, in the school study, it was not feasible to randomly assign every student to either the student-centered inquiry or the standard lecture group. This was done by assigning entire classes to either one of the conditions. Thus, it was clear that some of the control, manipulation, and random assignments were not achieved in the school study, but others remained intact; therefore, the school study was quasi-experimental. The independent variable in this research was the instruction methods, and the dependent variable was sixth graders' science proficiency scores.

Nonexperimental research is composed of studies in which none of the control, manipulation, or random assignment is attempted. Nonexperimental research happens in a natural environment where the research participants' behaviors, thoughts, opinions, interests, or preferences are observed and recorded. Surveys and public opinion polling belong in the category of nonexperimental research, and so do observational studies and cross-sectional studies. For example, an executive of a company was curious to find out whether its employees were happy with their work conditions. To satisfy this curiosity, an industrial and organizational psychologist was hired to design an employee job satisfaction survey to investigate how employees felt about their work conditions without any

attempt to control, manipulate, or randomly assign employees to any particular condition. There was no real distinction between independent variables and dependent variables. Nonexperimental research is about observing and studying research participants in their natural settings without deliberately controlling the environment or manipulating their behavior or preferences.

There is no one single best way to conduct an empirical study. The selection of research design depends on the purpose of the research. If the identification of a causal relationship is desired, experimental research is more likely to achieve the purpose. The *extraneous variables* that exist with quasi-experimental and nonexperimental research designs will prevent a causal relationship being asserted as a result of the research. The **extraneous variables** are variables that are not included in the study but might have an impact on the relationship between variables in the study. Extraneous variables can be attributable to a lack of strict control of environmental variables or individual differences that research participants bring along with them. On the other hand, if participants' behavior in their natural habitat is the focus of the research, nonexperimental research is the right choice.

POP QUIZ

A company explored the effect of its compensation (pay) structure on job performance of salespeople. The company randomly chose three stores to conduct the study. Every salesperson in Store A was compensated with salary, every salesperson in Store B was compensated by sales commissions, and every salesperson in Store C was compensated by a low base salary plus an additional sales-based bonus. Six months later, job performance across all three stores was measured and reported.

5. What is the independent variable of this study?
6. What is the dependent variable of this study?
7. What type of research design is used in this study?

REQUIRED MATHEMATICAL SKILLS FOR THIS COURSE

What kinds of mathematical knowledge and skills are required for this statistics course? The answers to this question will help you go through this course smoothly and successfully. First, you need to master the order of operations. A statistical formula usually contains multiple parts of different math operations such as addition, subtraction, multiplication, division, exponentiation, and parentheses. Knowing which part of a mathematical operation to do first is the key to arrive at the correct answers. There is only

one correct answer and an unlimited number of wrong answers for a statistical question. Knowing the order of operations provides a standard way to simplify and find the correct answer. There are several priority levels in order of operations shown in Table 1.3. The principle is to perform higher priority operations before moving on to lower priority ones. Many middle school math teachers use the phrase "please excuse my dear Aunt Sally" to help students memorize the order of operations: P-E-MD-AS. When dealing with many operations at the same level of priority, simplify the operations in the order they appear from left to right. Multiplication and division are at the same level of priority. Addition and subtraction are at the same level of priority. The strict rule of order of operations provides a standard way and the only correct way for everyone to interpret and solve statistical formulas containing complex operations. Therefore, when faced with the operations $8/2 \times 3 - 7$, you will be able to come up with the correct answer (i.e., 5). Order of operations is a fundamental math principle that needs to be strictly followed at all times. If you come up with wrong answers in your statistical exams, quizzes, or homework assignments, then chances are that you have messed up the order of operations somewhere. If I get a quarter every time any of my students makes a mistake on the order of operations, I will be a rich woman.

Second, learn to use your calculator correctly. When you don't use your calculator correctly, you will get wrong answers and lose points on homework, quizzes, and exams. Here is a simple check. Use your calculator to find the square of –3. Now, I will *pause* for your answer. Tell me the answer first before you move on to the next sentence.

If your answer is –9, you need figure out how to use your calculator correctly. If your answer is 9, you have passed the calculator test.

TABLE 1.3 ⬢ Priority Levels in Order of Operations

Priority Level: Perform Higher Priority Operations First Before Moving on to Lower Priority Ones	Math Operations	Example
First priority	Parentheses: Simplify whatever is inside the parentheses before doing anything else	$(X - 1)$, $(X^2 - 1)$ or $(3 - 1) \times (4 + 2)$
Second priority	Exponentiation	X^2, $\sqrt{x}$ or 8^2
Third priority	Multiplication or division	XY, $\frac{X}{Y}$ or $2 \times 4/5$
Fourth priority	Addition or subtraction	$X_1 + X_2 + X_3$, $n - 1$ or $5 - 7 + 8$

Third, refresh your knowledge of basic algebra. Sometimes you need to know the basic algebra to solve unknowns in the equations. Solving equations means figuring out a solution set for the equations. A very basic algebraic knowledge is that one linear equation can solve one unknown, two linear equations can solve two unknowns, and so on.

Statistical Notation

It is very common to add a set of values in statistics. To make it easier to communicate such a computation, a special notation is used to refer to the sum of a set of values. The uppercase Greek letter $\sum$ (pronounced sigma) is used for summation. Here is our first summation operation:

$$\sum_{i=1}^{n} X_i$$

This summation operation means to calculate the sum of all values of X, starting at the first value, X_1, and ending with the last value, X_n.

In a summation operation, there are several components. They are clearly explained one by one in the following list.

1. The summation sign, $\sum$, means performing addition in this operation.
2. X is the variable being added.
3. The subscript i, which is placed next to X, indicates that there are i values of X.
4. There is a starting point and an ending point of the index i. This tells us how many values need to be added. Usually the starting point is $i = 1$, which is placed under the $\sum$, and the ending point is n, which is placed on top of the $\sum$.

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \cdots + X_n$$

Sometimes, for the sake of simplicity, summation is expressed as $\sum X$.

$$\sum X = X_1 + X_2 + X_3 + \cdots + X_n$$

From now on this simple form of $\sum$ will express summation. You have to do summation from the first value to the last value given to you in a problem statement. Let's go through several examples of the summation question to get you used to the simple form of $\sum$.

EXAMPLE 1.1

$X = 3, 5, 7,$ and 9. Find $\sum X$, $(\sum X)^2$, and $\sum X^2$.

There are four values in the X variable, and this problem has three sub-questions (1) $\sum X$, (2) $(\sum X)^2$, and (3) $\sum X^2$. You need to figure out the correct order of operations in each sub-question.

The first sub-question is $\sum X$, so there is only one operation involved.

$$\sum X = 3 + 5 + 7 + 9 = 24$$

The second sub-question $(\sum X)^2$ involves three operations: parentheses, addition, and exponentiation. Let's break it down. According to the order of operations, solving inside the parentheses is the first priority and there is an addition $(\sum X)$ inside the parentheses. Then you have to square the answer of $\sum X$. Therefore, you need to figure out $\sum X$ first.

$$(\sum X)^2 = (3 + 5 + 7 + 9)^2 = (24)^2 = 576$$

The third sub-question $\sum X^2$ involves two operations: exponentiation and addition. Order of operations mandates that the exponentiation has to be done before the addition. Therefore, you figure out the value of each X^2 and then add them up.

$$\sum X^2 = 3^2 + 5^2 + 7^2 + 9^2 = 9 + 25 + 49 + 81 = 164$$

EXAMPLE 1.2

$X = 2, 4, -6,$ and -9. Find $\sum X$, $(\sum X - 7)^2$, and $\sum X^2$.

There are four values in the X variable, and this problem has three sub-questions: (1) $\sum X$, (2) $(\sum X - 7)^2$, and (3) $\sum X^2$. The difference between Examples 1.1 and 1.2 is that some of the values are negative. Keep the negative sign consistent as stated in the problem.

The first sub-question is $\sum X$, so there is only one operation involved.

$$\sum X = 2 + 4 + (-6) + (-9) = -9$$

The second sub-question $(\sum X - 7)^2$ involves four different operations: parentheses with both addition and subtraction inside of it then exponentiation. Order of operations tells us to simplify whatever is inside the parentheses first. Therefore, you need to figure out $\sum X$ first then do the subtraction inside the parentheses before conducting the exponentiation.

$$(\sum X - 7)^2 = (-9 - 7)^2 = (-16)^2 = 256$$

The third sub-question $\sum X^2$ involves in two operations: exponentiation and addition. Order of operations mandates that the exponentiation has to be done before the addition. Therefore, you figure out the value of each X^2 and then add them up.

$$\sum X^2 = 2^2 + 4^2 + (-6)^2 + (-9)^2 = 4 + 16 + 36 + 81 = 137$$

You have just witnessed the order of operations and statistical notation of $\sum$ in action. It is obvious that the correct order of operations must be followed to arrive at the correct answer. There will be many statistical formulas involving different math operations, so make sure you have a solid understanding on solving these types of questions. There will be more practices like this in the Exercise Problems.

POP QUIZ

$X = 3, 6, 9, 12,$ and 15.

8. Find $\sum(X - 9)^2$.
9. $\sum X^2 - 9$.

EXERCISE PROBLEMS

1. Researchers studied the effect of driving while texting on driving mistakes. College students were recruited to participate in the study. Due to risk of actual driving on the road, researchers conducted the research in a lab with a driving simulator. Students were randomly assigned to one of the two groups: (1) "driving without texting" or (2) "driving while texting." Driving mistakes were recorded by the simulator, which include driving more than 10 miles above or below the speed limit and failing to stay within the lane. What type of research was this study? What was the independent variable in this study? What was the dependent variable in this study?
2. Researchers investigated the gender gap in pay for physicians. An online salary survey was sent out to the members of the American Medical Association with an electronic link. What type of research was this study? What kind of sample was likely to be obtained by this online survey?
3. Where $X = 1, 2, 3,$ and 4, and $Y = 2, 4, 6,$ and 8, compute $\sum(X - 3)(Y - 5)$ and $\sum X^2Y^2$.

Solutions

1. The research is conducted in lab with a driving simulator. There are two conditions that are manipulated by the researchers. Research participants are randomly assigned to one of the two conditions. All three important features, namely (1) control, (2) manipulation, and (3) random assignment, are achieved in this study. Therefore, this study is an experimental research design. The independent variable measures the two driving conditions: (1) "driving without texting" and (2) "driving while texting." The dependent variable is the driving mistakes.

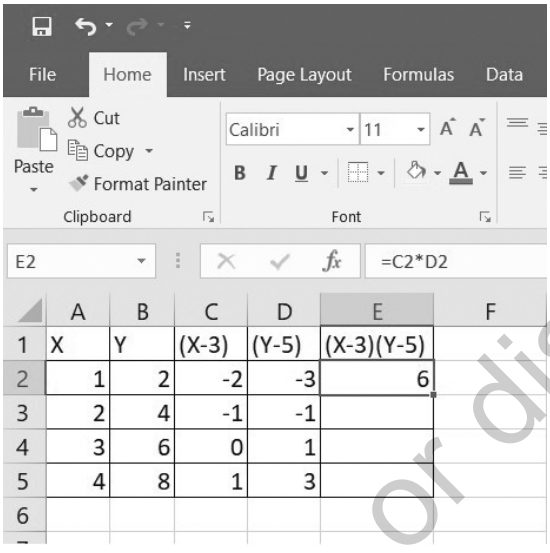
2. An online survey does not attempt to control, manipulate, or randomly assign participants to different conditions. Therefore, it is a nonexperimental design. Not all physicians belong to the American Medical Association. Therefore, sending a survey out to its members can't reach a random sample of physicians. The survey was voluntarily answered by people who received the link. Some physicians would simply ignore the email. Therefore, it was a convenience sample.

3. It is easy to create a table of X and Y to figure what need to be computed to solve the problems.

$$X = 1, 2, 3, \text{ and } 4;\ Y = 2, 4, 6, \text{ and } 8; \text{ compute } \sum (X-3)(Y-5) \text{ and } \sum X^2Y^2$$

You may solve this problem by using your calculator. I am going to show you how to use Excel to solve computation problems. Excel is a great tool to construct such a table to conduct mathematical operations. All instructions regarding Excel use Microsoft 365 Excel 2019. Most programming of mathematical operations is identical across different versions of Excel such as 2019, 2016, 2013, and 2010. Excel allows you to program mathematical operation formulas in each column. It is a great skill to have to be able to program math operations in Excel. Let's start by entering X values in column **A** and Y values in column **B**. Label **C1** as (X–3). Move the cursor to **C2** (column C, row 2), and type the formula that you want to create. In Excel, a function starts with a "=" which is a command to calculate. Excel is based on location of the variable. Therefore, to create the first operation of $(X-3)$, grab the first value of X located in **A2** (column A, row 2) and subtract 3, in Excel language "=**A2–3**". The beauty of Excel is that once you create a formula for the first value of the variables involved, you may simply copy and paste the formula for the rest of the values in the sequence. Copy and paste can be done by moving the cursor to the lower right-hand corner of **C2** where the formula is successfully created, until a solid + shows up. Hold the left click on the mouse and drag to the last value of the variable, and then let go of the left click. You should see that the formula is copied with all cell references. Next, do the same procedure to create $(Y-5)$ by labeling **D1** as (Y–5). Move the cursor to **D2**, and type "=**B2–5**", then hit **Enter**. Move the cursor to the lower right-hand corner of **D2** where the formula is successfully created, until a solid + shows up. Hold the left click on the mouse and drag to the last value of the Y, and then let go of the left click. According to the question, you need to find $\sum X-3)(Y-5)$, so you need to create $(X-3)(Y-5)$. To create $(X-3)(Y-5)$, label **E1** as (X–3)(Y–5), simply type "=**C2*D2**" in **E2**, and then hit **Enter**; the answer 6 appears in **E2** as shown in Figure 1.2.

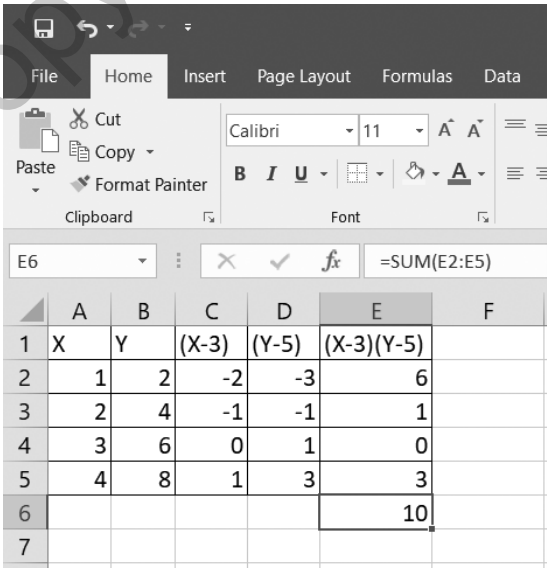
FIGURE 1.2 ⬢ Programming (*X* – 3) (*Y* – 5) in Excel



	A	B	C	D	E	F
1	X	Y	(X-3)	(Y-5)	(X-3)(Y-5)	
2	1	2	-2	-3	6	
3	2	4	-1	-1		
4	3	6	0	1		
5	4	8	1	3		
6						

Perform the same copy and paste for the rest of values for $(X-3)(Y-5)$ to the last value in the sequence. Now move the cursor to **E6** where you want the answer of $\Sigma(X-3)(Y-5)$ to show up typing "**=SUM(E2:E5)"**, then hit **Enter**. The answer for **Σ (X–3)(Y–5)** appears in **E6** as shown in Figure 1.3.

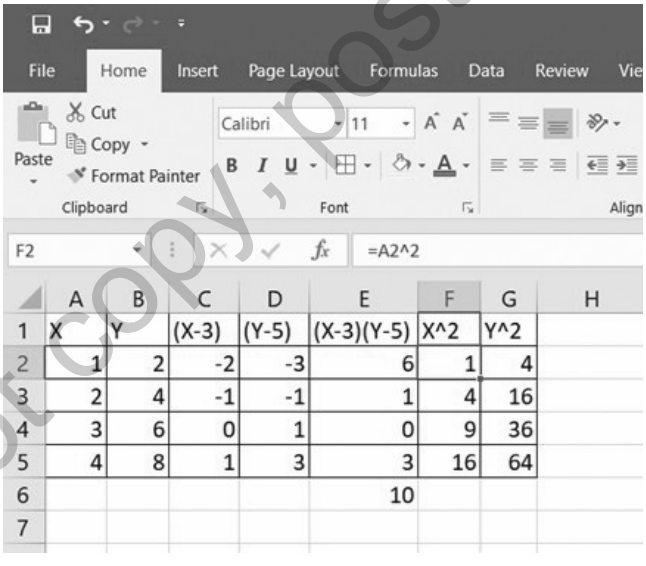
FIGURE 1.3 ⬢ Programming $\sum$ (*X* – 3) (*Y* – 5) in EXCEL



	A	B	C	D	E	F
1	X	Y	(X-3)	(Y-5)	(X-3)(Y-5)	
2	1	2	-2	-3	6	
3	2	4	-1	-1	1	
4	3	6	0	1	0	
5	4	8	1	3	3	
6					10	
7						

Next, you need to figure out how to perform necessary operations to solve $\sum X^2Y^2$ in Excel. Let's break it down. $\sum X^2Y^2$ involves exponentiation, multiplication, and addition. Order of operations dictates that exponentiation needs to be done first, next multiplication, and then addition. Each step is shown in a column of the table in Figures 1.4. In **F1**, label it as X^2. Excel calculates the square by using ^2. You may find the ^ symbol by holding shift and 6 simultaneously on your keyboard. In **F2**, type "=**A2^2**", which means grab the first value of *X* and square it. Hit **Enter** and the answer 1 appears in **F2**. Move the cursor to the lower right-hand corner of **F2** where the formula is successfully created, until a solid + shows up. Hold the left click on the mouse and drag to the last row of the variable, and then let go of the left click. You should see that the formula is copied with all cell references. The formula's operations are performed on the first row of the variable all the way to the last row of the variable. In **G1**, label it as Y^2. In **G2**, type "=**B2^2**", which means grab the first value of *Y* and square it. Copy and paste the formula to the last value of *Y* as shown in Figure 1.4.

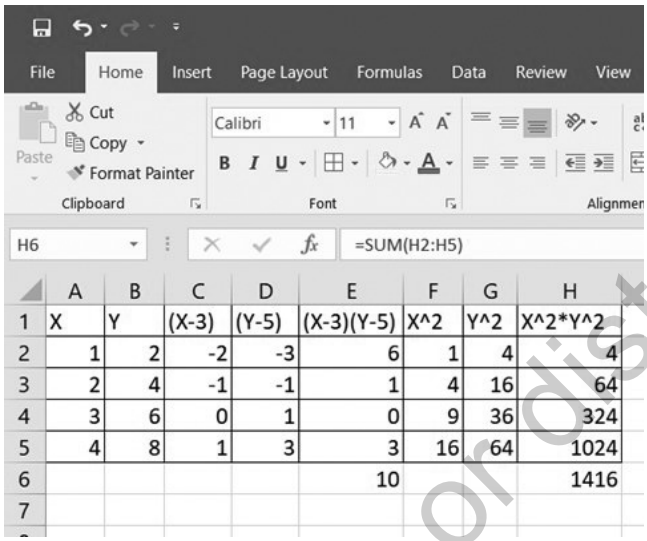
FIGURE 1.4 ⬢ Programming X^2 and Y^2 in Excel



	A	B	C	D	E	F	G	H
1	X	Y	(X-3)	(Y-5)	(X-3)(Y-5)	X^2	Y^2	
2	1	2	-2	-3	6	1	4	
3	2	4	-1	-1	1	4	16	
4	3	6	0	1	0	9	36	
5	4	8	1	3	3	16	64	
6					10			
7								

Now, the next step to solve this question is to create a formula for the multiplication of X^2Y^2. Label **H1** as X^2*Y^2. Move the cursor to **H2**. Type "=**F2*G2**", which means grab the first values of X^2 and Y^2 and multiply them. Copy and paste the function to the rest of the variables. Move the cursor to **F6** and conduct a summation by typing in "=**SUM(H2:H5)**". Hit **Enter** and the answer 1,416 appears as seen in Figure 1.5.

FIGURE 1.5 ⬢ Programming $\sum X^2Y^2$ in Excel



	A	B	C	D	E	F	G	H
1	X	Y	(X-3)	(Y-5)	(X-3)(Y-5)	X^2	Y^2	X^2*Y^2
2	1	2	-2	-3	6	1	4	4
3	2	4	-1	-1	1	4	16	64
4	3	6	0	1	0	9	36	324
5	4	8	1	3	3	16	64	1024
6					10			1416
7								

What You Learned

In this chapter, you have learned some of the basic vocabulary of statistics. Three major objectives of statistics are as follows:

1. To describe information contained in a sample
2. To design a sampling process, so that the selected sample is an unbiased representation of the population
3. To make inferences about a population from information contained in a sample

A brief definition of all the statistical terms mentioned in the chapter will be listed in the section "Key Words." You also need to sharpen some basic math skills such as order of operations and basic algebra to be able to follow formulas and solve equations. The correct order of operations is P-E-MD-AS.

Note: Each column of the Excel spreadsheet shows a particular step in the order of operations. Using Excel clearly shows the step-by-step procedures that lead to the correct answers. Excel is also widely available without incurring additional expenses for you. Excel is a great learning tool that solidifies the understanding order of operations in carrying out the calculation of statistical formulas.

Key Words

Absolute zero: Absolute zero means a complete absence of the attribute that you are measuring. It is not an arbitrarily assigned number. 10

Cluster sampling: Cluster sampling works best when "natural" grouping (clustering) occurs in the population. Random sampling is conducted to select which clusters are included in the sample. Once a cluster is selected in the sample, all individuals in the cluster are included in the sample. 6

Continuous variables: Continuous variables are values that do not have separation from one integer to the next. Continuous variables usually are expressed with decimals or fractions. 12

Convenience sample: A convenience sample is one in which researchers use anyone who is willing to participate in the study. A convenience sample is created based on easy accessibility. 4

Data: Data are defined as factual information used as a basis for reasoning, discussion, or calculation, so that meaningful conclusions can be drawn. 2

Dependent variable: A dependent variable is the variable that is the focus of researchers' interests and is affected by the different levels of an independent variable. 12

Descriptive statistics: Descriptive statistics are statistical procedures used to describe, summarize, organize, and simplify relevant characteristics of sample data. 3

Discrete variables: Discrete variables are values that have clear separation from one integer to the next. The answers for discrete variables can only be integers (i.e., whole numbers). 12

Experimental research: Experimental research is usually conducted in a tightly controlled environment (i.e., research laboratories). The three important features in experimental research are (1) control, (2) manipulation, and (3) random assignment. 12

Extraneous variable: The extraneous variables are variables that are not included in the study but might have an impact on the relationship between variables included in the study. 14

Independent variable: An independent variable is the variable that is deliberately manipulated by the researchers in a research study. 12

Inferential statistics: Inferential statistics are statistical procedures that use sample statistics to generalize to or make inferences about a population. 4

Interval scale: An interval scale not only arranges observations according to their magnitudes but also distinguishes the ordered arrangement in equal units. 9

Likert scales: Likert scales are often used to measure people's opinions, attitudes, or preferences. Likert scales measure attributes along a continuum of choices such as 1 = *strongly disagree*, 2 = *somewhat disagree*, 3 = *neutral*, 4 = *somewhat agree*, or 5 = *strongly agree* with each individual statement. 8

Nominal scale: In a nominal scale, measurements are used as identifiers, such as your student identification number, phone number, or social security number. 8

Nonexperimental research: Nonexperimental research is conducted to observe and study research participants in their natural settings without deliberately controlling the environment or manipulating their behaviors or preferences. 13

Ordinal scale: In an ordinal scale, measurements are used not only as identifiers but also to carry orders in a particular sequence. 8

Parameters: Parameters are defined as numerical characteristics of a population. 3

Population: A population is defined as an entire collection of everything or everyone that researchers are interested in studying or measuring. 2

Quasi-experimental research: Quasi-experimental research has some but not all of the features of experimental research. More specifically, if one or more of the control, manipulation, and random assignment features are not feasible but others remain intact, the research becomes quasi-experimental. 13

Random sample: A random sample is an ideal way to select participants for scientific research. A random sample occurs when every member in the population has an equal chance of being selected. 4

Ratio scale: A ratio scale contains every characteristic that lower level scales of measurement have, such as identifiers, direction of ranking, equal units, and something extra: an absolute zero. 10

Sample: A sample is defined as a subset of the population from which measures are actually obtained. 2

Sample statistics: Sample statistics are defined as numerical attributes of a sample. 3

Scales of measurement: Scales of measurement illustrate different ways that variables are defined and measured. Each scale of measurement has certain mathematical properties that determine the appropriate application of statistical procedures. 8

Simple random sample: A simple random sample is a subset of individuals (a sample) chosen from a larger set (a population). Each individual is chosen randomly and entirely by chance, and each subset of k individuals has the same probability of being chosen for the sample as any other subset of k individuals. 5

Statistics: Statistics is a science that deals with the collection, organization, analysis, and interpretation of numerical data. 2

Stratified sampling: Stratified sampling is the process of grouping members of the population into relatively homogeneous subgroups before sampling. A random sample from each stratum is independently taken in the same proportion as the stratum's size to the population. These subsets of the strata are then pooled to form a random sample. 6

Systematic sample: A systematic sample is achieved by selecting a sample from a population using a random starting point and a fixed interval. Typically, every "kth" member is selected from the total population for inclusion in the sample. 6

Variable: A variable refers to a measurable attribute. These measures have different values from one person to another, or the values change over time. 7

Learning Assessment

Multiple Choice: Circle the Best Answer in Every Question

1. A researcher was interested in the sleeping habits of college students. A group of 50 students were selected at random and interviewed. The researcher found that these students slept an average of 6.7 hours per day. For this study, the 50 students are an example of a _____.
   a. parameter
   b. statistic
   c. population
   d. sample

2. A quantity, usually an unknown numerical value that describes a population, is a _____.
   a. parameter
   b. statistic
   c. population
   d. sample

3. What additional characteristic is required on a ratio scale compared with an interval scale?
   a. Whether the measurements are the same or different
   b. The order of the magnitudes
   c. An absolute zero
   d. Scores with equal units

4. For $X = 0, 1, 6, 3$, what is $(\sum X)^2$?
   a. 20
   b. 46
   c. 64
   d. 100

5. Gender, religion, and ethnicity are measurements on a(n) _____ scale.
   a. nominal
   b. ordinal
   c. interval
   d. ratio

6. The measure of temperature in Fahrenheit is an example of a(n) _____ scale of measurement.
   a. nominal
   b. ordinal
   c. interval
   d. ratio

7. Which of the following pairs is usually unknown parameters of the population?
   a. $\bar{X}$ and μ
   b. s and σ
   c. s^2 and σ^2
   d. μ and σ
8. The measure of income is an example of a(n) _____ scale of measurement.
   a. nominal
   b. ordinal
   c. interval
   d. ratio

Free Response Questions

9. $X = 3, 4, 5,$ and 7; compute $\sum X$, $(\sum X)^2$, and $\sum X^2$
10. $X = -3, 0, 1,$ and 2; compute $\sum X$, $\sum (X - 1)^2$, and $\sum X^2 - 3$
11. $X = 3, 4, 5,$ and 7; $Y = -1, 0, 1,$ and 2; compute $\sum XY$ and $(\sum XY)^2$
12. $X = 3, 4, 5,$ and 7; $Y = -1, 0, 1,$ and 2; compute $\sum X^2Y^2$ and $\sum (X - 2)(Y - 3)$
13. $X = 4, 5, 6,$ and 9; $Y = -1, -1, 1,$ and 2; compute $\sum XY)^2$ and $\sum (X - 5)(Y + 1)$

Answers to Pop Quiz Questions

1. a
2. d
3. d
4. a
5. Pay structure
6. Job performance of salespeople
7. Quasi-experimental research design with cluster sampling
8. $\sum (X - 9)^2 = 90$
9. $\sum X^2 - 9 = 486$