Chapter 1:Introduction to Statistics

Chapter Outline

1.1 Statistics, Science, and Observations
- Definitions of Statistics
- Populations and Samples
- Variables and Data
- Parameters and Statistics
- Descriptive and Inferential Statistical Methods
- Statistics in the Context of Research

1.2 Data Structures, Research Methods, and Statistics
- Individual Variables
- Relationships between Variables
- Statistics for the Correlational Method
- Limitations of the Correlational Method
- Statistics for Comparing Two (or More) Groups of Scores
- Experimental and Nonexperimental Methods
- The Experimental Method
- Terminology in the Experimental Method
- Nonexperimental Methods: Nonequivalent Groups and Pre-Post Studies

1.3 Variables and Measurement
- Constructs and Operational Definitions
- Discrete and Continuous Variables
- Scales of Measurement
- The Nominal Scale
- The Ordinal Scale
- The Interval and Ratio Scales

1.4 Statistical Notation
- Scores
- Summation Notation

Learning Objectives and Chapter Summary

1. Define the terms population, sample, parameter, and statistic, and describe the relationships between them.

   The term statistics is used to refer to methods for organizing, summarizing, and interpreting data.

   Scientific questions usually concern a population, which is the entire set of individuals one wishes to study. Usually, populations are so large that it is impossible to examine every individual, so most research is conducted with samples. A sample is a group selected from a population, usually for purposes of a research study.

A characteristic that describes a sample is called a statistic, and a characteristic that describes a population is called a parameter. Although sample statistics are usually representative of corresponding population parameters, there is typically some discrepancy between a statistic and a parameter.

2. Define descriptive and inferential statistics and describe how these two general categories of statistics are used in a typical research study.

   Statistical methods can be classified into two broad categories: descriptive statistics, which organize and summarize data, and inferential statistics, which use sample data to draw inferences about populations.

3. Describe the concept of sampling error and explain how this concept creates the fundamental problem that inferential statistics must address.

   The naturally occurring difference between a statistic and a parameter is called sampling error.

4. Differentiate correlational, experimental, and nonexperimental research and describe the data structures associated with each.

5. Define independent, dependent, and quasi-independent variables and recognize examples of each.

6. Explain why operational definitions are developed for constructs and identify the two components of an operational definition.

   The correlational method examines relationships between variables by measuring two different variables for each individual. This method allows researchers to measure and describe relationships, but cannot produce a cause-and-effect explanation for the relationship.

   The experimental method examines relationships between variables by manipulating an independent variable to create different treatment conditions and then measuring a dependent variable to obtain a group of scores in each condition. The groups of scores are then compared. A systematic difference between groups provides evidence that changing the independent variable from one condition to another also caused a change in the dependent variable. All other variables are controlled to prevent them from influencing the relationship. The intent of the experimental method is to demonstrate a cause-and-effect relationship between variables. The experimental method examines relationships between variables by manipulating an independent variable to create different treatment conditions and then measuring a dependent variable to obtain a group of scores in each condition. The groups of scores are then compared. A systematic difference between groups provides evidence that changing the independent variable from one condition to another also caused a change in the dependent variable. All other variables are controlled

to prevent them from influencing the relationship. The intent of the experimental method is to demonstrate a cause-and-effect relationship between variables.

Nonexperimental studies also examine relationships between variables by comparing groups of scores, but they do not have the rigor of true experiments and cannot produce cause-and-effect explanations. Instead of manipulating a variable to create different groups, a nonexperimental study uses a preexisting participant characteristic (such as male/female) or the passage of time (before/after) to create the groups being compared.

7. Describe discrete and continuous variables and identify examples of each.

8. Differentiate nominal, ordinal, interval, and ratio scales of measurement.

   A discrete variable consists of indivisible categories, often whole numbers that vary in countable steps. A continuous variable consists of categories that are infinitely divisible and each score corresponds to an interval on the scale. The boundaries that separate intervals are called real limits and are located exactly halfway between adjacent scores.

   A measurement scale consists of a set of categories that are used to classify individuals. A nominal scale consists of categories that differ only in name and are not differentiated in terms of magnitude or direction. In an ordinal scale, the categories are differentiated in terms of direction, forming an ordered series. An interval scale consists of an ordered series of categories that are all equal-sized intervals. With an interval scale, it is possible to differentiate direction and magnitude (or distance) between categories. Finally, a ratio scale is an interval scale for which the zero point indicates none of the variable being measured. With a ratio scale, ratios of measurements reflect ratios of magnitude.

9. Identify what is represented by each of the following symbols: X, Y, N, n and $\sum$.

10. Perform calculations using summation notation and other mathematical operations following the correct order of operations.

    The letter X is used to represent scores for a variable. If a second variable is used, Y represents its scores. The letter N is used as the symbol for the number of scores in a population; n is the symbol for a number of scores in a sample.

    The Greek letter sigma (Σ) is used to stand for summation. Therefore, the expression ΣX is read “the sum of the scores.” Summation is a mathematical operation (like addition or multiplication) and must be performed in its proper place in the order of operations; summation occurs after parentheses, exponents, and multiplying/dividing have been completed.

Other Lecture Suggestions

1. Early in the first class, acknowledge that:
   - Most students are not there by choice. (No one picked statistics as an elective because it looked like a fun class.)
   - Many students have some anxiety about the course.

   However, try to reassure them that the class will probably be easier and more enjoyable (less painful) than they would predict, *provided* they follow a few simple rules:
   - Keep Up. In statistics, each bit of new material builds on the previous material.As long as you have mastered the old material, then the new stuff is just one small step forward.On the other hand, if you do not know the old material, then the new stuff is totally incomprehensible.(For example, try reading Chapter 10 on the first day of class.It will make no sense at all.However, by the time we get to Chapter 10, you will have enough background to understand it.)Keeping up means coming to class, asking questions, and doing homework on a regular basis.If you are getting lost, then get help immediately.

   - Test Yourself. It is very easy to sit in class and watch an instructor work through examples.Also, it is very easy to complete homework assignments if you can look back at example problems in the book.Neither activity means that you really know the material.For each chapter, try one or two of the end-of-chapter problems without looking back at the examples in the book or checking your notes.Can you really do the problems on your own?If not, pay attention to where you get stuck in the problem, so you will know exactly what you still need to learn.

2. Give students a list of variables, for example items from a survey (age, gender, education level, income, occupation) and ask students to identify the scale of measurement most likely to be used and whether the variable is discrete or continuous.

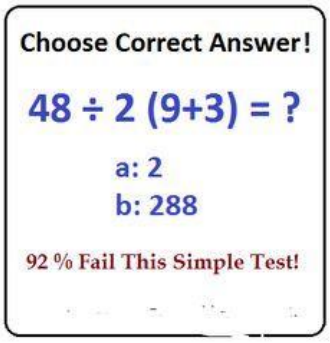
3. Describe a non-experimental or correlational study and have students identify reasons that you cannot make a cause-and-effect conclusion from the results.For example, a researcher finds that children in the local school who regularly eat a nutritious breakfast have higher grades than students who do not eat a nutritious breakfast.Does this mean that a nutritious breakfast *causes* higher grades?For example, a researcher finds that employees who regularly use the company's new fitness center have fewer sick days than employees who do not use the center.Does this mean that using the fitness center *causes* people to be healthier?

   In either case, describe how the study could be made into an experiment by:
   a. beginning with equivalent groups (random assignment).
   b. manipulating the independent variable (this introduces the ethical question of forcing people to eat a nutritious breakfast).
   c. controlling other variables (the rest of the children's diet).

4. After introducing some basic applications of summation notation, present a simple list of scores (1, 3, 5, 4) and a relatively complex expression containing summation notation, for

example, $\Sigma(X - 1)^2$.Ask the students to compute the answer.You are likely to obtain several different responses.

Note that this is not a democratic process - the most popular answer is not necessarily correct.There is only one correct answer because there is only one correct sequence for performing the calculations.Have the class identify the step by step sequence of operations specified by the expression.(First, subtract 1 from each of the scores.Second, square the resulting values.Third, sum the squared numbers.)Then apply the steps, one by one, to compute the answer.As a variation, present a list of steps and ask students to write the mathematical expression corresponding to the series of steps.



Alternatively, there are frequently social media posts that test knowledge of the order of operations. Google "social media order operations" and click on "images" to see recent ones. Present several to students to review the order of operations. One that claims a certain percentage of people get it wrong will allow an opportunity to discuss the misuse of statistics as well.

5. Invite students to explore how they come into contact with statistics in their everyday lives. Use an article like Statistics in Everyday Life (http://www.isixsigma.com/community/blogs/statistics-everyday-life/) to stimulate discussion. Invite the students to find an article online or in a newspaper about a topic of interest to them and to review how that article uses (or misuses) statistics. Ask them to consider the implications of not understanding statistics and their use.

Answers to Even-Numbered Problems

2. The population is the entire group of individuals (or scores) of interest for a particular research study. A sample is a group selected from a population that usually is used to represent the population in a research study. A parameter is a characteristic, usually a numerical value, that describes a population. A statistic is a characteristic, usually numerical, that describes a sample.

4. Sampling error is the naturally occurring difference between a sample and the population from which the sample is obtained. Specifically, the statistics obtained for a sample will be different from the corresponding parameters for the population and the statistics will differ from one sample to another. This is a problem for inferential statistics because any difference found between two treatment conditions may be explained by the treatments but it also may be explained by sampling error.

6. The goal of an experiment is to demonstrate the existence of a cause-and-effect relationship between two variables. To accomplish the goal, an experiment must manipulate an independent variable and control other, extraneous variables.

8. This is not an experiment because no independent variable is manipulated. The researchers are comparing two preexisting groups of individuals.

10. a. The dependent variable is comprehension of the passage, which is measured by the test scores.
    b. Knowledge or comprehension is continuous.
    c. ratio scale (zero means no correct answers)

12. a. The independent variable is taking the Tai Chi course versus not taking the course.
    b. Nominal scale
    c. The dependent variable is the amount of arthritis pain experienced.
    d. The amount of pain probably is measured with an interval or a ratio scale.

14. a. An ordinal scale provides information about the direction of difference (greater or less) between two measurements.
    b. An interval scale provides information about the magnitude of the difference between two measurements.
    c. A ratio scale provides information about the ratio of two measurements.

16. Honesty is an attribute or personality characteristic that cannot be observed or measured directly. Shyness could be operationally defined by identifying and observing external behaviors associated with being shy. Or, participants could be given a questionnaire asking how they behave or feel in situations for which shyness might have an influence.

18. a. $\Sigma X = 10$
    b. $\Sigma X^2 = 38$
    c. $\Sigma X + 1 = 11$
    d. $\Sigma(X + 1) = 14$

20. a. $\Sigma X = 0$
    b. $\Sigma X^2 = 50$
    c. $\Sigma(X + 3) = 15$

22. a. $(\Sigma X)^2$
    b. ΣX^2
    c. $\Sigma(X - 2)$
    d. $\Sigma(X - 1)^2$

Chapter 2: Frequency Distributions

Chapter Outline

Learning Objectives and Chapter Summary

1. Describe the basic elements of a frequency distribution table and explain how they are related to the original set of scores.

> The goal of descriptive statistics is to simplify the organization and presentation of data. One descriptive technique is to place the data in a frequency distribution table or graph that shows exactly how many individuals (or scores) are located in each category on the scale of measurement.

2. Calculate the following from a frequency table: ΣX, ΣX^2, and the proportion and percentage of the group associated with each score.

> A frequency distribution table lists the categories that make up the scale of measurement (the *X* values) in one column. Beside each *X* value, in a second column, is the frequency (*f*) or number of individuals in that category. The table may include a proportion (*p*) column showing the relative frequency for each category and it may include a percentage column showing the percentage (*%*) associated with each X value.

3. Identify when it is useful to set up a grouped frequency distribution table, and explain how to construct this type of table for a set of scores.

> It is recommended that a frequency distribution table have a maximum of 10–15 rows to keep it simple. If the scores cover a range that is wider than this suggested maximum, it is

customary to divide the range into sections called class intervals. These intervals are then listed in the frequency distribution table along with the frequency or number of individuals with scores in each interval. The result is called a grouped frequency distribution.The guidelines for constructing a grouped frequency distribution table are as follows:

a. There should be about 10 intervals.
b. The width of each interval should be a simple number (e.g., 2, 5, or 10).
c. The bottom score in each interval should be a multiple of the width.
d. All intervals should be the same width, and they should cover the range of scores with no gaps.

4. Describe how the three types of frequency distribution graphs - histograms, polygons, and bar graphs - are constructed and identify when each is used.

5. Describe the basic elements of a frequency distribution graph and explain how they are related to the original set of scores.

6. Explain how frequency distribution graphs for populations differ from the graphs used for samples.

A frequency distribution graph lists scores on the horizontal axis and frequencies on the vertical axis. Thetype of graph used to display a distribution depends on the scale of measurement used. For interval or ratio scales, you should use a histogram or a polygon. For a histogram, a bar is drawn above each score so that the height of the bar corresponds to the frequency. Each bar extends to the real limits of the score, so that adjacent bars touch. For a polygon, a dot is placed above the mid- point of each score or class interval so that the height of the dot corresponds to the frequency; then lines are drawn to connect the dots. Bar graphs are used with nominal or ordinal scales. Bar graphs are similar to histograms except that gaps are left between adjacent bars.

7. Identify the shape - symmetrical, and positively or negatively skewed - of a distribution in a frequency distribution graph.

Shape is one of the basic characteristics used to describe a distribution of scores. Most distributions can be classified as either symmetrical or skewed. A skewed distribution that tails off to the right is positively skewed. If it tails off to the left, it is negatively skewed.

8. Define percentiles and percentile ranks.

9. Determine percentiles and percentile ranks for values corresponding to real limits in a frequency distribution table.

The cumulative percentage is the percentage of individuals with scores at or below a particular point in the distribution. The cumulative percentage values are associated with the upper real limits of the corresponding scores or intervals.

Percentiles and percentile ranks are used to describe the position of individual scores within a distribution. Percentile rank gives the cumulative percentage associated with a particular score. A score that is identified by its rank is called a percentile.

10.Estimate percentiles and percentile ranks using interpolation for values that do not correspond to real limits in a frequency distribution table.

When a desired percentile or percentile rank is located between two known values, it is possible to estimate the desired value using the process of interpolation. Interpolation assumes a regular linear change between the two known values.

11. Describe the basic elements of a stem and leaf display and explain how the display shows the entire distribution of scores.

A stem and leaf display is an alternative procedure for organizing data. Each score is separated into a stem (the first digit or digits) and a leaf (the last digit). The display consists of the stems listed in a column with the leaf for each score written beside its stem. A stem and leaf display is similar to a grouped frequency distribution table, however the stem and leaf display identifies the exact value of each score and the grouped frequency distribution does not.

Other Lecture Suggestions

1. Begin with an unorganized list of scores as in Example 2.1, and then organize the scores into a table. If you use a set of 20 or 25 scores, it will be easy to compute proportions and percentages for the same example.

2. Present a relatively simple, regular frequency distribution table (for example, use scores of 5, 4, 3, 2, and 1 with corresponding frequencies of 1, 3, 5, 3, 2. Ask the students to determine the values of N and ΣX for the scores. Note that ΣX can be obtained two different ways: 1) by computing and summing the fX values within the table, 2) by retrieving the complete list of individual scores and working outside the table.

Next, ask the students to determine the value of ΣX^2. You probably will find a lot of wrong answers from students who are trying to use the fX values within the table. The common mistake is to compute $(fX)^2$ and then sum these values. Note that whenever it is necessary to do complex calculations with a set of scores, the safe method is to retrieve the list of individual scores from the table before you try any computations.

3. It sometimes helps to make a distinction between graphs that are being used in a formal presentation and sketches that are used to get a quick overview of a set of data. In one case, the graphs should be drawn precisely and the axes should be labeled clearly so that the graph can be easily understood without any outside explanation. On the other hand, a sketch that is intended for your own personal use can be much less precise. As an instructor, if you are expecting

precise, detailed graphs from your students, you should be sure that they know your expectations.

4. Introduce interpolation with a simple, real-world example. For example, in Buffalo, the average snowfall during the month of February is 30 inches. Ask students, how much snow they would expect during the first half of the month. Then point out that the same interval (February) is being measured in terms of days and in terms of inches of snow. A point that is half-way through the interval in terms of days should also be half-way through the interval in terms of snow.

5. Refer to Box 2.1 The Use and Misuse of Graphs and discuss common misuses. For more examples, refer to the subtly-named How to Lie with Data Visualization (http://data.heapanalytics.com/how-to-lie-with-data-visualization/). Challenge students to bring in examples of misleading graphs they find online or in print. (Hint: The more stridently a website advocates for or against a particular point of view on a social, political or other controversial issue, the more likely you are to find misrepresentation of data.)

Answers to Even-Numbered Problems

2.

X	*f*	*p*	*%*
9	1	0.05	5%
8	0	0.00	0%
7	1	0.05	5%
6	2	0.10	10%
5	4	0.20	20%
4	2	0.10	10%
3	3	0.15	15%
2	5	0.25	25%
1	2	0.10	10%

4. a. $n = 14$
   b. $\Sigma X = 44$
   c. $\Sigma X^2 = 168$

6.

X	*f*
60-64	1
55-59	2
50-54	2
45-49	1
40-44	2
35-39	3
30-34	3
25-29	5
20-24	8
15-19	3

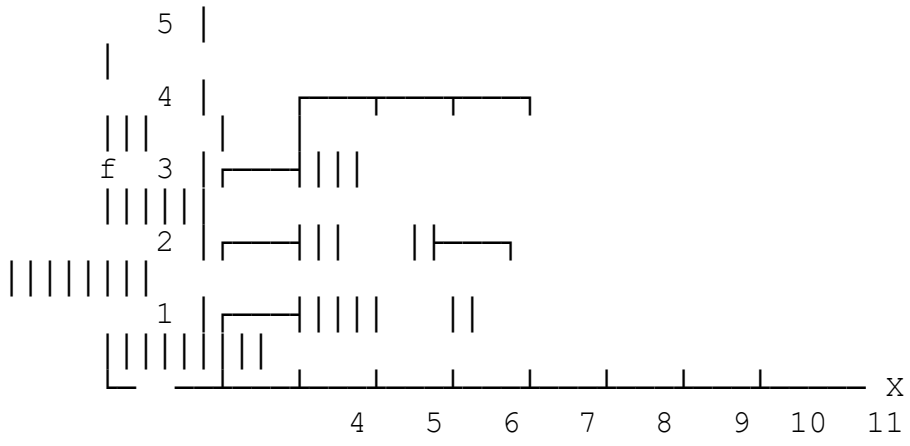
Younger drivers, especially those 20 to 29 years old, tend to get more tickets.

8. A regular table reports the exact frequency for each category on the scale of measurement. After the categories have been grouped into class intervals, the table reports only the overall frequency for the interval but does not indicate how many scores are in each of the individual categories.

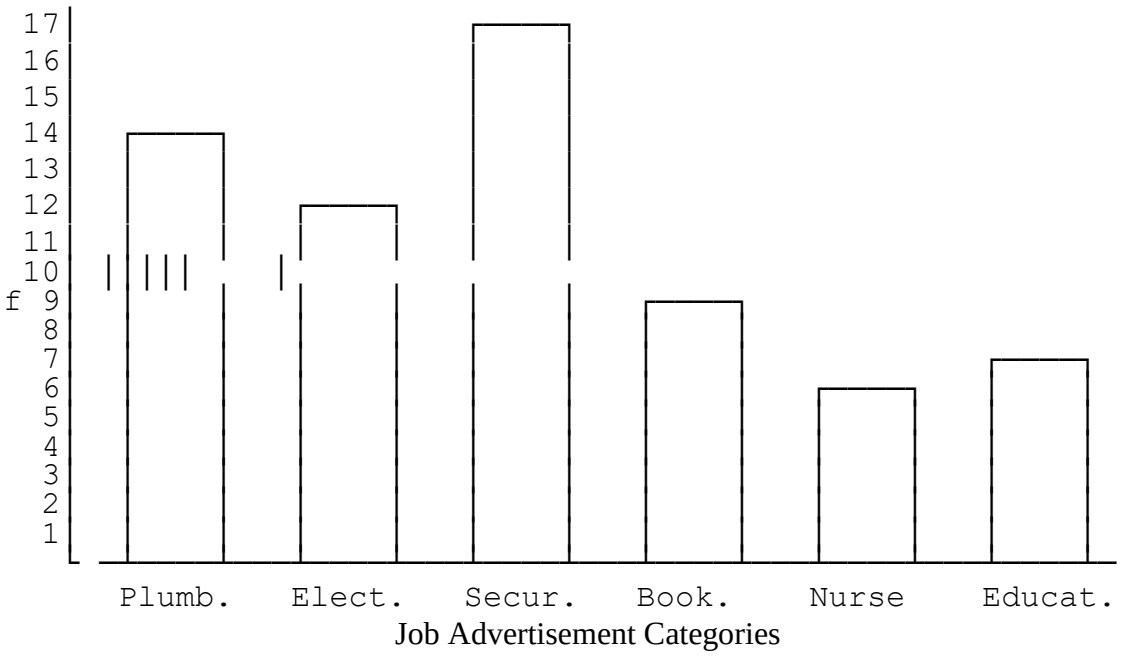
10. a.

X	*f*
10	2
9	4
8	4
7	4
6	3
5	2
4	1

b.



12.



Job Advertisement Categories

14.

X	*f*
15	5
14	6
13	4
12	2
11	2
10	1

The distribution is negatively skewed.