Chapter 1: The Nature of Probability and Statistics

Diana Pell

Section 1.1: Descriptive and Inferential Statistics

Statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

A variable is a characteristic or attribute that can assume different values.

A population consists of all subjects (human or otherwise) that are being studied. When data are collected from every subject in the population, it is called a census.

A sample is a group of subjects selected from a population. The information obtained from a statistical sample is said to be biased if the results from the sample of a population are radically different from the results of a census of the population.

Descriptive statistics consists of the collection, organization, summarization, and presentation of data.

Inferential statistics consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.

The area of inferential statistics called hypothesis testing is a decision-making process for evaluating claims about a population, based on information obtained from samples.

Exercise 1. Determine whether descriptive or inferential statistics were used.

a. The average jackpot for the top five lottery winners was $367.6 million.

b. Scientists at the University of Oxford in England found that a good laugh significantly raises a person’s pain level tolerance.

Exercise 2. Read the following on attendance and grades, and answer the questions.

A study conducted at Manatee Community College revealed that students who attended class 95 to 100% of the time usually received an A in the class. Students who attended class 80 to 90% of the time usually received a B or C in the class. Students who attended class less than 80% of the time usually received a D or an F or eventually withdrew from the class.
Based on this information, attendance and grades are related. The more you attend class, the more likely it is you will receive a higher grade. If you improve your attendance, your grades will probably improve. Many factors affect your grade in a course. One factor that you have considerable control over is attendance. You can increase your opportunities for learning by attending class more often.

a. What are the variables under study?

b. What are the data in the study?

c. Are descriptive, inferential, or both types of statistics used?

d. What is the population under study?

e. Was a sample collected? If so, from where?

f. From the information given, comment on the relationship between the variables.

Section 1.2: Variables and Types of Data

**Qualitative variables** are variables that have distinct categories according to some characteristic or attribute.

**Quantitative variables** are variables that can be counted or measured.

**Discrete variables** assume values that can be counted.

**Continuous variables** can assume an infinite number of values between any two specific values. They are obtained by measuring. They often include fractions and decimals.
Exercise 3. Classify each variable as a discrete variable or a continuous variable.

a. The highest wind speed of a hurricane

b. The weight of baggage on an airplane

c. The number of pages in a statistics book

d. The amount of money a person spends per year for online purchases

The boundary of a number, then, is defined as a class in which a data value would be placed before the data value was rounded.

Example 1. A recorded height of 73 inches could mean any measure from 72.5 inches up to but not including 73.5 inches. Boundaries are written for convenience as 72.5 - 73.5 but are understood to mean all values up to but not including 73.5.

Exercise 4. Find the boundaries of each variable.

a. 8.4 quarts

b. 138 mmHg

Measurement scales

The nominal level of measurement classifies data into mutually exclusive (non-overlapping) categories in which no order or ranking can be imposed on the data.

A sample of college instructors classified according to subject taught (e.g., English, history, psychology, or mathematics).

The ordinal level of measurement classifies data into categories that can be ranked; however, precise differences between the ranks do not exist.

Floats in a homecoming parade might be ranked as first place, second place, etc.

The interval level of measurement ranks data, and precise differences between units of measure do exist; however, there is no meaningful zero.

IQ is an example of such a variable.
The ratio level of measurement possesses all the characteristics of interval measurement, and there exists a true zero. In addition, true ratios exist when the same variable is measured on two different members of the population.

Examples of ratio scales are those used to measure height, weight, area, and number of phone calls received. Ratio scales have differences between units (1 inch, 1 pound, etc.) and a true zero.

**Exercise 5.** What level of measurement would be used to measure each variable?

a. The ages of patients in a local hospital

b. The ratings of movies released this month

c. Colors of athletic shirts sold by Oak Park Health Club

d. Temperatures of hot tubs in local health clubs

**Exercise 6.** Read the following information about the number of fatal accidents for the transportation industry in 2010, and answer each question.

<table>
<thead>
<tr>
<th>Industry</th>
<th>Number of fatalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highway accidents</td>
<td>968</td>
</tr>
<tr>
<td>Railway accidents</td>
<td>44</td>
</tr>
<tr>
<td>Water vehicle accidents</td>
<td>52</td>
</tr>
<tr>
<td>Aircraft accidents</td>
<td>151</td>
</tr>
</tbody>
</table>

*Source: Bureau of Labor Statistics.*

a. Name the variables under study.

b. Categorize each variable as quantitative or qualitative.

c. Categorize each quantitative variable as discrete or continuous.

d. Identify the level of measurement for each variable.

e. The railroad has the fewest fatalities for 2010. Does that mean railroads have fewer accidents than the other industries?
f. What factors other than safety influence a person’s choice of transportation?

g. From the information given, comment on the relationship between the variables.

**Section 1.3: Data Collection and Sampling Techniques**

Surveys can be done by using a variety of methods. Three of the most common methods are the telephone survey, the mailed questionnaire, and the personal interview. Data can also be collected in other ways, such as surveying records or direct observation of situations.

A **random sample** is a sample in which all members of the population have an equal chance of being selected.

Random samples are selected by using chance methods or random numbers. Statisticians generate random numbers with a computer or calculator.

A **systematic sample** is a sample obtained by selecting every $k^{th}$ member of the population where $k$ is a counting number.

A **stratified sample** is a sample obtained by dividing the population into subgroups or strata according to some characteristic relevant to the study. (There can be several subgroups.) Then subjects are selected from each subgroup.

A **cluster sample** is obtained by dividing the population into sections or clusters and then selecting one or more clusters and using all members in the cluster(s) as the members of the sample.

**Sampling error** is the difference between the results obtained from a sample and the results obtained from the population from which the sample was selected.

A **nonsampling error** occurs when the data are obtained erroneously or the sample is biased, i.e., nonrepresentative.
Exercise 7. State which sampling method was used.

a. Out of 10 hospitals in a municipality, a researcher selects one and collects records for a 24-hour period on the types of emergencies that were treated there.

b. A researcher divides a group of students according to gender, major field, and low, average, and high grade point average. Then she randomly selects six students from each group to answer questions in a survey.

c. The subscribers to a magazine are numbered. Then a sample of these people is selected using random numbers.

d. Every 10th bottle of Super-Duper Cola is selected, and the amount of liquid in the bottle is measured. The purpose is to see if the machines that fill the bottles are working properly.