CHAPTER 1 Data Collection
Section 1.1 Introduction to the Practice of Statistics

Objectives
1. Define statistics and statistical thinking
2. Explain the process of statistics
3. Distinguish between qualitative and quantitative variables
4. Distinguish between discrete and continuous variables
5. Determine the level of measurement of a variable

Objective 1 Define statistics and statistical thinking
- **Statistics** is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

Data
- The information referred to in the definition is **data**.
- **Data** are a “fact or proposition used to draw a conclusion or make a decision.” Data describe characteristics of an individual.
- A key aspect of data is that they vary.
- One goal of statistics is to describe and understand sources of variability.

Objective 2 Explain the Process of Statistics
- A ________ consists of the entire group of individuals to be studied.
- A __________ is a subset of the population that is being studied.
- An __________ is a person or object that is a member of the population being studied.
- _____________ consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs.
- A ____________ is a numerical summary based on a sample.
- _____________ uses methods that take results from a sample, extends them to the population, and measures the reliability of the result.
- A __________ is a numerical summary of a population.

Parameter versus Statistic
*Example:*
Suppose the percentage of all students on our campus who have a job is 84.9%.
a) What is the population?

b) Does the value 84.9% represent a parameter or a statistic?

Suppose a sample of 250 students is obtained, and from this sample we find that 86.4% have a job. Does the value 86.4% represent a parameter or a statistic?

Illustrating the process of Statistics
Step 1: Identify the research objective.
Step 2: Collect the information needed to answer the question.
Step 3: Describe the data-Organize and summarize the information.
Step 4: Draw conclusions from the data.
Objective 3  Distinguish between Qualitative and Quantitative Variables

- ____________ is the characteristics of the individuals within the population.
- Key Point: Variables vary. Consider the variable height. If all individuals had the same height, then obtaining the height of one individual would be sufficient in knowing the heights of all individuals. Of course, this is not the case. As researchers, we wish to identify the factors that influence variability.

Variables and Types of Data
Variables can be classified as _______________________ or _________________________.
- _______________ variables allow for classification of individuals based on some attribute or characteristic.
- _______________ variables provide numerical measures of individuals. Arithmetic operations such as addition and subtraction can be performed on the values of the quantitative variable and provide meaningful results.

Objective 4 Distinguish between Discrete and Continuous Variables
Quantitative variables can be further classified into two groups.
- A _____________ variable is a quantitative variable that has either a finite number of possible values or a countable number of possible values. The term “countable” means the values result from counting such as 0, 1, 2, 3, and so on. (e.g. # of books, # of desks)
- A _____________ is a quantitative variable that has an infinite number of possible values it can take on and can be measured to any desired level of accuracy.

Classification of Variables

Example:
Classify each variable as qualitative or quantitative. If the variable is quantitative, further classify it as discrete or continuous.

a) The number of heads obtained after flipping a coin five times.

b) Weights of new born babies in a hospital

c) Eye colors of students in Math 227

- The list of observations a variable assumes is called ________.
- While gender is a variable, the observations, male or female, are data.
- Qualitative data are observations corresponding to a qualitative variable.
- Quantitative data are observations corresponding to a quantitative variable.
  - Discrete data are observations corresponding to a discrete variable.
  - Continuous data are observations corresponding to a continuous variable.

Objective 5 Determine the Level of Measurement of a Variable
- Variables can also be classified by how they are categorized, counted, or measured.
• The level of measurement of the data is useful in deciding what procedure to take to apply statistics to real problems.

Ex) Can the data be organized into specific categories, such as area of residence (rural, suburban, or urban)? Can the data values be ranked, such as first place, second place, etc.? Are the values obtained from measurement, such as heights, IQs, or temperature?
• Four common types of measurement scales are used to classify variables:

Levels of Measurement:
• _____________ — the values of the variable name, label, or categorize. In addition, the naming scheme does not allow for the values of the variable to be arranged in a ranked, or specific, order.
• _____________ — it has the properties of the nominal level of measurement and the naming scheme allows for the values of the variable to be arranged in a ranked, or specific, order.
• _____________ — it has the properties of the ordinal level of measurement and the differences in the values of the variable have meaning. A value of zero in the interval level of measurement does not mean the absence of the quantity. Arithmetic operations such as addition and subtraction can be performed on values of the variable.
• _____________ — it has the properties of the interval level of measurement and the ratios of the values of the variable have meaning. A value of zero in the ratio level of measurement means the absence of the quantity. Arithmetic operations such as multiplication and division can be performed on the values of the variable.

Example: Classify each as nominal-level, ordinal-level, interval-level, or ratio level data.
  a) Sizes of cars
  b) Nationality of each student
  c) IQ of each student
  d) Weight

Section 1.2 Observational Studies Versus Designed Experiments
Objectives
1. Distinguish between an observational study and an experiment
2. Explain the various types of observational studies

Objective 1  Distinguish between an Observational Study and an Experiment
• An _________________ measures the value of the response variable without attempting to influence the value of either the response or explanatory variables. That is, in an observational study, the researcher observes the behavior of the individuals in the study without trying to influence the outcome of the study.
• If a researcher assigns the individuals in a study to a certain group, intentionally changes the value of the explanatory variable, and then records the value of the response variable for each group, the researcher is conducting a ______________________
In both studies, the goal of the research was to determine if radio frequencies from cell phones increase the risk of contracting brain tumors. Whether or not brain cancer was contracted is the response variable (dependent variable). The level of cell phone usage is the explanatory variable (independent variable). In research, we wish to determine how varying the amount of an explanatory variable affects the value of a response variable.

______________________________ in a study occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory (independent) variable and the response (dependent) variable may be due to some other variable or variables not accounted for in the study.

A _______________ lurking variable is an explanatory variable that was not considered in a study, but that affect the value of the response variable in the study. In addition, lurking variables are typically related to any explanatory variables considered in the study.

**Example 1:**
Determine whether the study depicts an observational study or an experiment.

(a) Rats with cancer are divided into two groups. One group receives 5 mg of a medication that is thought to fight cancer, and the other receives 10 mg. After 2 years, the spread of the cancer is measured.

(b) Conservation agents netted 320 large-trout in a lake and determined how many were carrying parasites.

**Example 2:**
Identify the explanatory variable and the response variable for the following studies.

(a) Rats with cancer are divided into two groups. One group receives 5 mg of a medication that is thought to fight cancer, and the other receives 10 mg. After 2 years, the spread of the cancer is measured.

(b) A researcher wants to determine whether young couples who marry are more likely to gain weight than those who stay single.

**Objectives 2** Explain the various types of observational studies

______________________________ Observational studies that collect information about individuals at a specific point in time, or over a very short period of time.
These studies are **retrospective**, meaning that they require individuals to look back in time or require the researcher to look at existing records. In case-control studies, individuals who have certain characteristics are matched with those that do not.

A cohort study first identifies a group of individuals to participate in the study (the cohort). The cohort is then observed over a long period of time. Over this time period, characteristics about the individuals are recorded. Because the data is collected over time, cohort studies are **prospective**.

**EXAMPLE  Observation Study or Designed Experiment?**

Determine whether each of the following studies depict an observational study or an experiment.

a. Researchers wanted to assess the long-term psychological effects on children evacuated during World War II. They obtained a sample of 169 former evacuees and a control group of 43 people who were children during the war but were not evacuated. The subjects’ mental states were evaluated using questionnaires. It was determined that the psychological well being of the individuals was adversely affected by evacuation. (Source: Foster D, Davies S, and Steele H (2003) The evacuation of British children during World War II: a preliminary investigation into the long-term psychological effects. Aging & Mental Health (7)5.)

b. Xylitol has proven effective in preventing dental caries (cavities) when included in food or gum. A total of 75 Peruvian children were given milk with and without xylitol and were asked to evaluate the taste of each. Overall, the children preferred the milk flavored with xylitol. (Source: Castillo JL, et al (2005) Children’s acceptance of milk with xylitol or sorbitol for dental caries prevention. BMC Oral Health (5)6.)

c. A total of 974 homeless women in the Los Angeles area were surveyed to determine their level of satisfaction with the healthcare provided by shelter clinics versus the healthcare provided by government clinics. The women reported greater quality satisfaction with the shelter and outreach clinics compared to the government clinics. (Source: Swanson KA, Andersen R, Gelberg L (2003) Patient satisfaction for homeless women. Journal of Women’s Health (12)7.)

A **census** is a list of all individuals in a population along with certain characteristics of each individual.

**Section 1.3**

**Simple Random Sampling**

**Objective**

1. Obtain a Simple Random Sample

Random sampling

- **is the process of using chance to select individuals from a population to be included in the sample.**

- A sample of size n from a population of size N is obtained through simple random sampling if every possible sample of size n has an equally likely chance of occurring. The sample is then called a simple random sample.

**EXAMPLE  Illustrating Simple Random Sampling**

Suppose a study group of consists of 5 students:
Bob, Patricia, Mike, Jan, and Maria

2 of the students must go to the board to demonstrate a homework problem. List all possible samples of size 2 (without replacement).
Section 1.4 Other Effective Sampling Methods

Steps for Obtaining a Simple Random Sample

1) Obtain a frame that lists all the individuals in the population of interest. Number the individuals in the frame 1 – N. (A frame is a list of all the individuals within the population.)

2) Use a random number table, graphing calculator, or statistical software to randomly generate n numbers where n is the desired sample size.

Objectives
1. Obtain a Stratified Sample
2. Obtain a Systematic Sample
3. Obtain a Cluster Sample

A stratified sample is one obtained by separating the population into homogeneous, nonoverlapping groups called strata, and then obtaining a simple random sample from each stratum. The individuals within each stratum should be homogeneous (or similar) in some way.

A systematic sample is obtained by selecting every kth individual from the population. The first individual selected is a random number between 1 and k.

Steps in Systematic Sampling, Population Size Known

Step 1: Determine the population size, N.
Step 2: Determine the sample size desired, n.
Step 3: Compute N/n and round down to the nearest integer. This value is k.
Step 4: Randomly select a number between 1 and k. Call this number p.
Step 5: The sample will consist of the following individuals:
\[ p, p + k, p + 2k, \ldots, p + (n - 1)k \]

A cluster sample is obtained by selecting all individuals within a randomly selected collection or group of individuals.

Section 1.5 Bias in Sampling

Objectives
1. Explain the Sources of Bias in Sampling

Sources of Bias in Sampling
- If the results of the sample are not representative of the population, then the sample has bias.

Three Sources of Bias
1. Sampling Bias
2. Nonresponse Bias
3. Response Bias

________ means that the technique used to obtain the individuals to be in the sample tends to favor one part of the population over another.

________ is a type of sampling bias. __________ occurs when the proportion of one segment of the population is lower in a sample than it is in the population.
exists when individuals selected to be in the sample who do not respond to the survey have
different opinions from those who do.
Nonresponse can be improved through the use of callbacks or rewards/incentives.

exists when the answers on a survey do not reflect the true feelings of the respondent.

Response Bias - Types of Response Bias
1. Interviewer error
2. Misrepresented answers
3. Words used in survey question
4. Order of the questions or words within the question

Example 1: The survey has bias. Determine the type of bias.
(Sampling, non-response, response)
(a) To determine the public’s opinion of the police department, the police chief obtains a cluster sample of 15
census tracts within his jurisdiction and samples all households in the randomly selected tracts. Uniformed
police officers go door to door to conduct the survey.

(b) The village of Oak Lawn wishes to conduct a study regarding the income level of households within the
village. The village manager selects 10 homes in the southwest corner of the village and sends an
interviewer to the homes to determine household income.

Nonsampling errors are errors that result from sampling bias, nonresponse bias, response bias, or data-entry
error. Such errors could also be present in a complete census of the population.

Sampling error is error that results from using a sample to estimate information about a population. This type of
error occurs because a sample gives incomplete information about a population.