Notes

- Statistics is the *science* of collecting, organizing, and interpreting data.
- The **population** in a statistical study is the complete set of people or things being studied.
- **Population parameters** are specific characteristics of the population.
- A **sample** is a subset of the population from which data are actually obtained.
- The actual measurements or observations collected from the sample constitute the **raw data.**
- **Sample statistics** are characteristics of the sample found by consolidating or summarizing the raw data.

Example 1: The U.S. Labor Department defines the *civilian labor force* as all those people who are either employed or actively seeking employment. Each month, the Labor Department reports the unemployment rate, which is the percentage of people actively seeking employment within the entire civilian labor force. To determine the unemployment rate, the Labor Department surveys 60,000 households. For the unemployment reports, describe the a. population b. sample c. raw data d. sample statistics
e. population parameters

Answer: The *population* is the group that the Labor Department wants to learn about, which is all the people who make up the civilian labor force. The *sample* consists of all the people among the 60,000 households surveyed. The *raw data* consist of all the information collected in the survey. The *sample statistics* summarize the raw data for the sample. In this case, the relevant sample statistic is the percentage of people in the sample who are actively seeking employment. The *population parameters* are the characteristics of the entire population that correspond to the sample statistics. In this case, the relevant population parameter is the actual unemployment rate.

The *margin of error* in a statistical study is used to describe the range of values, or confidence interval, likely to contain the population parameter. We find this confidence interval by adding and subtracting the margin of error from the sample statistic obtained in the study. That is, the range of values likely to contain the population parameter is

from (sample statistic – margin of error)

to (sample statistic + margin of error)

The margin of error is usually defined to give a 95% confidence interval, meaning that 95% of samples of the size used in the study would contain the actual population parameter (and 5% would not).

Example 2: The Pew Research Center for the People and the Press interviewed 1,546 adult Americans about their attitudes toward the future. Asked whether humans would land on Mars within the next 50 years, 76% of these 1,546 people said either *definitely yes* or *probably yes*. The margin of error for the poll was 3 percentage points. Describe the population and the sample for this survey, and explain the meaning of the sample statistic of 76%. What can we conclude about the percentage of the population that thinks humans will land on Mars within the next 50 years?

Answer: The *population* is all adult Americans and the sample consists of the 1,546 people who were interviewed. The sample statistic of 76% is the *actual* percentage of people in the sample who answered that humans would definitely or probably land on Mars in the next 50 years. The 76% sample statistic and the margin of error of 3 percentage points tell us that the rang of values

from 76% – 3% = 73%

to 76% + 3% = 79%

is likely (with 95% confidence) to contain the population parameter, which in this case is the true percentage of all adult Americans who think humans will definitely or probably land on Mars within the next 50 years.

Example 3: The same Pew survey also asked people whether they believed that humans would be cloned within the next 50 years. On this question, 51% answered either *definitely yes* or *probably yes*. Again, the margin of error was 3 percentage points. Can we be confident that a majority of adult Americans think that humans will be cloned in the next 50 years?

No. The Confidence interval gives a range of values from 48% to 54%. Because this range includes values on both sides of 50%, we cannot be confident that the majority (that is, greater than 50%) of adult Americans think that humans will be cloned in the next 50 years.

Basic Steps in a Statistical Study

Step 1. State the goal of your study precisely; that is, determine the population you want to study and exactly what you'd like to learn about it.

Step 2. Choose a sample from the population. (Be sure to use an appropriate sampling technique, as discussed in the next section.)

Step 3. Collect raw data from the sample and summarize these data by finding sample statistics of interest.

Step 4. Use the sample statistics to make inference about the population.

Step 5. Draw conclusions; determine what you learned and whether you achieved your goal.

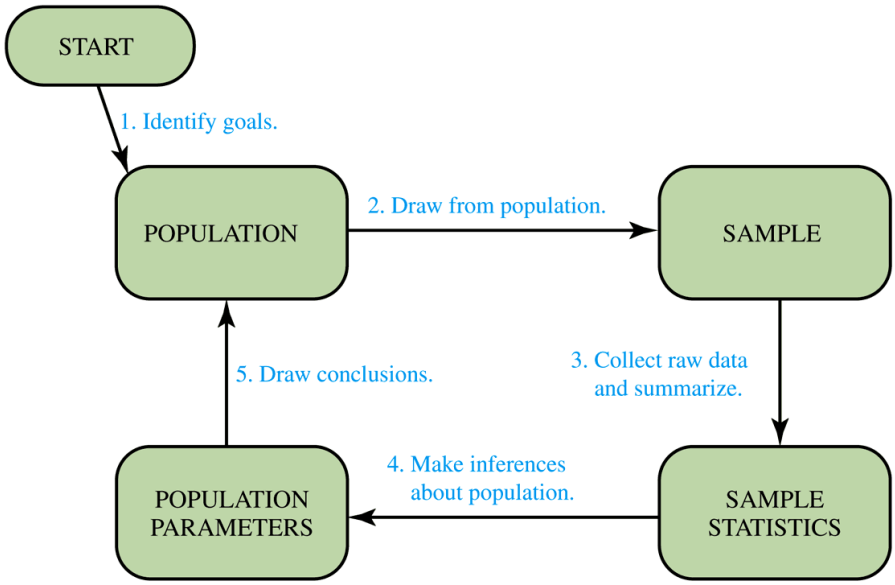

Figure 1.2 The process of a statistical study.

The Purpose of Statistics

Statistics has many uses, but perhaps its most important purpose is to help us make good decisions about issues that involve uncertainty.

SAMPLING:

A **representative sample** is a sample in which the relevant characteristics of the sample members are generally the same as the characteristics of the population.

A statistical study suffers from **bias** if its design or conduct tends to favor certain results.

Sampling Methods

- Simple Random Samples

A **random sample** is one in which every member of the population has an equal chance of being selected to be part of the sample.

With **simple random sampling** every possible sample of a particular size has an equal chance of being selected.

- Systematic Sampling

A type of sampling in which we use a system such as choosing every 50^{th} member of a population.

- Convenience Samples

The sample is chosen for convenience rather than by a more sophisticated procedure.

- Cluster Samples

Cluster sampling involves the selection of *all* members in randomly selected groups, or *clusters*.

- Stratified Samples

Stratified sampling involves randomly selecting members from each stratum.

Example: The U.S. Labor Department surveys 60,000 households each month to compile its unemployment report. To select these households, the Department first groups cities and counties into about 2000 geographic areas. It then randomly selects households to survey within these geographic areas. How is this an example of stratified sampling? What are the strata? Why is stratified sampling important in this case?

Solution:

The unemployment survey is an example of stratified sampling because it first breaks the population into subgroups. Stratified sampling is important in this case because unemployment rates in rural Kansas may be very different from those in Silicon Valley. By using stratified sampling, the Labor Department ensures that its sample fairly represents all geographic regions.

Summary: Common Sampling Methods

- Simple random sampling: We choose a sample of items in such a way that every sample of the same size has an equal chance of being selected.
- Systematic sampling: We use a simple system to choose the sample, such as selecting every 10th or every 50th member of the population.
- Convenience sampling: We use a sample that happens to be convenient to select.
- Cluster sampling: We first divide the population into groups, or clusters, and select some of these clusters at random. We then obtain the sample by choosing all the members within each of the selected clusters.
- Stratified sampling: We use this method when we are concerned about differences among subgroups, or strata, within a population. We first identify the strata and then draw a random sample within each stratum. The total sample consists of all the samples from the individual strata.

Types of Statistical Studies

The subjects of a study are the people, animals (or other living things), or objects chosen for the sample; if the subjects are people, they may also be called the participants in the study.

Two Basic Types of Statistical Study

There are two basic types of statistical study:

1. In an observational study, researchers observe or measure characteristics of the subjects but do not attempt to influence or modify these characteristics.
2. In an experiment, researchers apply some treatment and observe its effects on the subjects of the experiment.

Statistical studies—whether observations or experiments—generally are attempts to measure what we call variables of interest.

A **variable** is any item or quantity that can vary or take on different values. The **variables of interest** in a statistical study are the items or quantities that the study seeks to measure. When cause and effect may be involved, an **explanatory variable** is a variable that may explain or cause the effect, while a response variable is a variable that responds to change in the explanatory variable.

Variation on Observational Studies

The most familiar observational studies are those in which data are collected all at once (or as close to that as possible). Two variations on observational studies are also common:

1. A **retrospective** (or **case-control**) **study** uses data from the past, such as official records or past interviews.
2. A **prospective** (or **longitudinal**) **study** is set up to collect data in the future from groups that share common factors.

Treatment and Control Groups

The **treatment group** in an experiment is the group of subjects who receive the treatment being tested. The **control group** in an experiment is the group of subjects who do not receive the treatment being tested. In most cases, it is important to choose the members of the two groups by random selection from the available pool of subjects.

A study suffers from **confounding** if the effects of different variables are mixed so we cannot determine the specific effects of the variables of interest. The variables that lead to the confusion are called **confounding variables**.

Strategies for Selecting Treatment and Control Groups

1. **Select groups at random**. Make sure that the subjects of the experiment are assigned to the treatment or control group at random, meaning that each subject has an equal chance of being assigned to either group.

2. **Use sufficiently large groups**. Make sure that the treatment and control groups are both sufficiently large that they are unlikely to differ in a significant way (aside from the fact that one group gets the treatment and the other does not).

A **placebo** lacks the active ingredients of a treatment being tested in a study, but looks or feels like the treatment so that participants cannot distinguish whether they are receiving the placebo or the real treatment.

The **placebo effect** refers to the situation in which patients improve simply because they believe they are receiving a useful treatment.

Note: Although participants should not know whether they belong to the treatment or control group, for ethical reasons it is very important that participants be told that some of them will be given a placebo, rather than the real treatment.

An **experimenter effect** occurs when a researcher or experimenter somehow influences subjects through such factors as facial expression, tone of voice, or attitude.

An experiment is **single-blind** if the participants do not know whether they are members of the treatment group or members of the control group, but the experimenters do know.

An experiment is **double-blind** if neither the participants nor any experimenters know who belongs to the treatment group and who belongs to the control group.

In a **meta-analysis**, researchers review many past studies. The meta-analysis considers these studies as a combined group, with the aim of finding trends that were not evident in the individual studies.

Should You Believe a Statistical Study?

Eight Guidelines for Critically Evaluating a Statistical Study

1. Identify the goal of the study, the population considered, and the type of study.
2. Consider the source, particularly with regard to whether the researcher may be biased. Statistical studies are supposed to be objective, but the people who carry them out and fund them may be biased. it is therefore important to consider the source of a study and evaluate the potential for biases that might invalidate the study's conclusions.
3. Examine the sampling method to decide whether it is likely to produce a representative sample. A statistical study cannot be valid unless the sample is representative of the population under study. **Selection bias** (or a **selection effect**) occurs whenever researchers select their sample in a biased way. **Participation bias** occurs any time participation in a study is voluntary. A **self-selected survey** (or **voluntary response survey**) is one in which people decide for themselves whether to be included in the survey.
4. Look for problems in defining or measuring the variables of interest, which can make it difficult to interpret any reported results.
5. Watch out for confounding variables that can invalidate the conclusions of a study.
   Often, variables that are *not intended* to be part of the study can make it difficult to interpret results properly.
6. Consider the setting and wording in surveys or polls, looking for anything that might tend to produce inaccurate or dishonest responses.
7. Check that results are fairly represented in graphics and concluding statements, because both researchers and media often create misleading graphics or jump to conclusions that the results do not support.
8. Stand back and consider the conclusions. Did the study achieve its goals? Do the conclusions make sense? Do the results have any practical significance?