# An Overview of Statistics

## Where You've Been

You are already familiar with many of the practices of statistics, such as taking surveys, collecting data, and describing populations. What you may not know is that collecting accurate statistical data is often difficult and costly. Consider, for instance, the monumental task of counting and describing the entire population of the United States. If you were in charge of such a census, how would you do it? How would you ensure that your results are accurate? These and many more concerns are the responsibility of the United States Census Bureau, which conducts the census every decade.

Terms and concepts you will learn from this prior to starting the course :

- How to distinguish between a population and a sample
- How to distinguish between a parameter and a statistic
- How to distinguish between descriptive statistics and inferential statistics
- How to distinguish between qualitative and quantitative data
- How to classify data with respect to the four levels of measurement: Nominal, Ordinal, Interval and Ratio
- How to distinguish between an observational study and an experiment
- Determine the sample technique used from either random sampling, simple random sampling, stratified sampling, cluster sampling, or systematic sampling and how to identify a biased sample.

## DATA SETS

There are two types of data sets you will use when studying statistics. These data sets are called **populations** and **samples.**

**DEFINITION**

A **population** is the collection of *all* outcomes, responses, measurements, or counts that are of interest.

A **sample** is a subset, or part, of a population.

A sample should be representative of a population so that sample data can be used to draw conclusions about that population. Sample data must be collected using an appropriate method, such as *random sampling.* When sample data are collected using an *inappropriate* method, the data cannot be used to draw conclusions about the population.

*In Exercises 1–4, identify the population and the sample.*

1. A survey of 1503 U.S. adults found that 78% favor government policies requiring better fuel efficiency for vehicles. *(Source: Pew Research Center)*
2. Thirty-eight nurses working in the San Francisco area were surveyed concerning their opinions of managed health care.
3. A survey of 2311 U.S. adults found that 84% have seen a health care provider at least once in the past year. *(Source: Harris Interactive)*
4. A survey of 186 U.S. adults ages 25 to 29 found that 76% have read a book in the past 12 months. *(Source: Pew Research Center)*

Two important terms that are used throughout this course are **parameter** and **statistic.**

**DEFINITION**

A **parameter** is a numerical description of a *population* characteristic.

A **statistic** is a numerical description of a *sample* characteristic.

It is important to note that a sample statistic can differ from sample to sample whereas a population parameter is constant for a population.

## BRANCHES OF STATISTICS

The study of statistics has two major branches: **descriptive statistics** and **inferential statistics.**

**DEFINITION**

**Descriptive statistics** is the branch of statistics that involves the organization, summarization, and display of data.

**Inferential statistics** is the branch of statistics that involves using a sample to draw conclusions about a population. A basic tool in the study of inferential statistics is probability.

*In Exercises 5–8, determine whether the numerical value is a parameter or a statistic. Explain your reasoning.*

**5.** In 2012, Major League Baseball teams spent a total of $2,940,657,192 on players' salaries. *(Source: USA Today)*

**6.** In a survey of 1000 U.S. adults, 65% plan to be awake at midnight to ring in the new year. *(Source: Rasmussen Reports)*

**7.** In a recent study of math majors at a university, 10 students were minoring in physics.

**8.** Fifty percent of a sample of 1025 U.S. adults say that the best years for the United States are behind us. *(Source: Gallup)*

**9.** Which part of the survey described in Exercise 3 represents the descriptive branch of statistics? Make an inference based on the results of the survey.

**10.** Which part of the survey described in Exercise 4 represents the descriptive branch of statistics? Make an inference based on the results of the survey.

## TYPES OF DATA

When doing a study, it is important to know the kind of data involved. The nature of the data you are working with will determine which statistical procedures can be used. In this section, you will learn how to classify data by type and by level of measurement. Data sets can consist of two types of data: **qualitative data** and **quantitative data.**

### DEFINITION

**Qualitative data** consist of attributes, labels, or nonnumerical entries.

**Quantitative data** consist of numerical measurements or counts.

*In Exercises 11–14, determine whether the data are qualitative or quantitative. Explain your reasoning.*

**11.** The ages of a sample of 350 employees of a software company

**12.** The zip codes of a sample of 200 customers at a sporting goods store

**13.** The revenues of the companies on the Fortune 500 list

**14.** The marital statuses of all professional golfers

## LEVELS OF MEASUREMENT

Another characteristic of data is its level of measurement. The level of measurement determines which statistical calculations are meaningful. The four levels of measurement, in order from lowest to highest, are **nominal, ordinal, interval,** and **ratio.**

### DEFINITION

Data at the **nominal level of measurement** are qualitative only. Data at this level are categorized using names, labels, or qualities. No mathematical computations can be made at this level.

Data at the **ordinal level of measurement** are qualitative or quantitative. Data at this level can be arranged in order, or ranked, but differences between data entries are not meaningful.

Data at the **interval level of measurement** can be ordered, and meaningful differences between data entries can be calculated. At the interval level, a zero entry simply represents a position on a scale; the entry is not an inherent zero.

Data at the **ratio level of measurement** are similar to data at the interval level, with the added property that a zero entry is an inherent zero. A ratio of two data entries can be formed so that one data entry can be meaningfully expressed as a multiple of another.

**Summary of Four Levels of Measurement**

| | Example of a data set | Meaningful calculations |
|---|---|---|
| **Nominal level** (Qualitative data) | *Types of Shows Televised by a Network*<br>Comedy, Documentaries, Drama, Cooking, Reality Shows, Soap Operas, Sports, Talk Shows | *Put in a category.*<br>For instance, a show televised by the network could be put into one of the eight categories shown. |
| **Ordinal level** (Qualitative or quantitative data) | *Motion Picture Association of America Ratings Description*<br>G General Audiences<br>PG Parental Guidance Suggested<br>PG-13 Parents Strongly Cautioned<br>R Restricted<br>NC-17 No One 17 and Under Admitted | Put in a category and *put in order.*<br>For instance, a PG rating has a stronger restriction than a G rating. |
| **Interval level** (Quantitative data) | *Average Monthly Temperatures (in degrees Fahrenheit) for Denver, CO*<br>Jan 30.7, Feb 32.5, Mar 40.4, Apr 47.4, May 57.1, Jun 67.4, Jul 74.2, Aug 72.5, Sep 63.4, Oct 50.9, Nov 38.3, Dec 30.0<br>*(Source: National Climatic Data Center)* | Put in a category, put in order, and *find differences between values.*<br>For instance, $72.5 - 63.4 = 9.1$°F. So, August is 9.1°F warmer than September. |
| **Ratio level** (Quantitative data) | *Average Monthly Precipitation (in inches) for Orlando, FL*<br>Jan 2.35, Feb 2.38, Mar 3.77, Apr 2.68, May 3.45, Jun 7.58, Jul 7.27, Aug 7.13, Sep 6.06, Oct 3.31, Nov 2.17, Dec 2.58<br>*(Source: National Climatic Data Center)* | Put in a category, put in order, find differences between values, and *find ratios of values.*<br>For instance, $\frac{7.58}{3.77} \approx 2$. So, there is about twice as much precipitation in June as in March. |

*In Exercises 15–18, determine the level of measurement of the data set. Explain your reasoning.*

**15.** The daily high temperatures (in degrees Fahrenheit) for Sacramento, California, for a week in July are listed. *(Source: National Climatic Data Center)*

96 77 75 84 87 94 101

**16.** The vehicle size classes for a sample of sedans are listed.

Minicompact Subcompact Compact Mid-size Large

**17.** The four departments of a printing company are listed.

Administration Sales Production Billing

**18.** The total compensations (in millions of dollars) of the top ten CEOs in the United States are listed. *(Source: Forbes)*

131 67 64 61 56 52 50 49 44 43

**17.** The four departments of a printing company are listed.

Administration Sales Production Billing

**18.** The total compensations (in millions of dollars) of the top ten CEOs in the United States are listed. *(Source: Forbes)*

131 67 64 61 56 52 50 49 44 43

A statistical study can usually be categorized as an observational study or an experiment. In an **observational study,** a researcher does not influence the responses. In an **experiment,** a researcher deliberately applies a treatment before observing the responses. Here is a brief summary of these types of studies.

- In an **observational study,** a researcher observes and measures characteristics of interest of part of a population but does not change existing conditions. For instance, an observational study was performed in which researchers observed and recorded the mouthing behavior on nonfood objects of children up to three years old. *(Source: Pediatrics Magazine)*
- In performing an **experiment,** a **treatment** is applied to part of a population, called a **treatment group,** and responses are observed. Another part of the population may be used as a **control group,** in which no treatment is applied. (The subjects in the treatment and control groups are called **experimental units.**) In many cases, subjects in the control group are given a **placebo,** which is a harmless, fake treatment, that is made to look like the real treatment. The responses of the treatment group and control group can then be compared and studied. In most cases, it is a good idea to use the same number of subjects for each group. For instance, an experiment was performed in which diabetics took cinnamon extract daily while a control group took none. After 40 days, the diabetics who took the cinnamon reduced their risk of heart disease while the control group experienced no change. *(Source: Diabetes Care)*

*In Exercises 19 and 20, determine whether the study is an observational study or an experiment. Explain.*

**19.** Researchers conduct a study to determine whether a drug used to treat hypothyroidism works better when taken in the morning or when taken at bedtime. To perform the study, 90 patients are given one pill to take in the morning and one pill to take in the evening (one containing the drug and the other a placebo). After 3 months, patients are instructed to switch the pills. *(Source: JAMA Internal Medicine)*

**20.** Researchers conduct a study to determine the number of falls women had during pregnancy. To perform the study, researchers contacted 3997 women who had recently given birth and asked them how many times they fell during their pregnancies. *(Source: Maternal and Child Health Journal)*

## SAMPLING TECHNIQUES

A **census** is a count or measure of an *entire* population. Taking a census provides complete information, but it is often costly and difficult to perform. A **sampling** is a count or measure of *part* of a population, and is more commonly used in statistical studies. To collect unbiased data, a researcher must ensure that the sample is representative of the population. Appropriate sampling techniques must be used to ensure that inferences about the population are valid.

A **random sample** is one in which every member of the population has an equal chance of being selected. A **simple random sample** is a sample in which every possible sample of the same size has the same chance of being selected.

***Stratified Sample*** When it is important for the sample to have members from each segment of the population, you should use a stratified sample. Depending on the focus of the study, members of the population are divided into two or more subsets, called *strata*, that share a similar characteristic such as age, gender, ethnicity, or even political preference. A sample is then randomly selected from each of the strata. Using a stratified sample ensures that each segment of the population is represented.

***Cluster Sample*** When the population falls into naturally occurring subgroups, each having similar characteristics, a cluster sample may be the most appropriate. To select a cluster sample, divide the population into groups, called *clusters*, and select all of the members in one or more (but not all) of the clusters. Examples of clusters could be different sections of the same course or different branches of a bank.

***Systematic Sample*** A systematic sample is a sample in which each member of the population is assigned a number. The members of the population are ordered in some way, a starting number is randomly selected, and then sample members are selected at regular intervals from the starting number. (For instance, every 3rd, 5th, or 100th member is selected.)

A type of sample that often leads to biased studies (so it is not recommended) is a **convenience sample.** A convenience sample consists only of members of the population that are easy to get.

*In Exercises 23–28, identify the sampling technique used, and discuss potential sources of bias (if any). Explain.*

**23.** Using random digit dialing, researchers ask 1003 U.S. adults their plans on working during retirement. *(Source: Princeton Survey Research Associates International)*

**24.** A student asks 18 friends to participate in a psychology experiment.

**25.** A pregnancy study in Cebu, Philippines, randomly selects 33 communities from the Cebu metropolitan area, then interviews all pregnant women in these communities. *(Source: Cebu Longitudinal Health and Nutrition Survey)*

**26.** Law enforcement officials stop and check the driver of every third vehicle for blood alcohol content.

**27.** Twenty-five students are randomly selected from each grade level at a high school and surveyed about their study habits.

**28.** A journalist interviews 154 people waiting at an airport baggage claim and asks them how safe they feel during air travel.