TERMS

Population The entire group of individuals or instances about whom we hope to learn. (p. 278)

Sample A (representative) subset of a population, examined in hope of learning about the population. (p. 278)

Sample survey A study that asks questions of a sample drawn from some population in the hope of learning something about the entire population. Polls taken to assess voter preferences are common sample surveys. (p. 279)

Bias Any systematic failure of a sampling method to represent its population is bias. Biased sampling methods tend to over- or underestimate parameters. It is almost impossible to recover from bias, so efforts to avoid it are well spent. (p. 279) Common errors include

- relying on voluntary response,
- undercoverage of the population,
- nonresponse bias, and
- response bias.

Randomization The best defense against bias is randomization, in which each individual is given a fair, random chance of selection. (p. 280)

Sample size The number of individuals in a sample. The sample size determines how well the sample represents the population, not the fraction of the population sampled. (p. 281)

Census A sample that consists of the entire population is called a census. (p. 281)

Population parameter A numerically valued attribute of a model for a population. We rarely expect to know the true value of a population parameter, but we do hope to estimate it from sampled data. For example, the mean income of all employed people in the country is a population parameter. (p. 282)

Statistic, sample statistic Statistics are values calculated for sampled data. Those that correspond to, and thus estimate, a population parameter, are of particular interest. For example, the mean income of all employed people in a representative sample can provide a good estimate of the corresponding population parameter. The term "sample statistic" is sometimes used, usually to parallel the corresponding term "population parameter." (p. 282)

Representative A sample is said to be representative if the statistics computed from it accurately reflect the corresponding population parameters. (p. 283)

Simple random sample (SRS) A simple random sample of sample size n is a sample in which each set of n elements in the population has an equal chance of selection. (p. 283)

Sampling frame A list of individuals from whom the sample is drawn is called the sampling frame. Individuals who may be in the population of interest, but who are not in the sampling frame, cannot be included in any sample. (p. 284)

Sampling variability The natural tendency of randomly drawn samples to differ, one from another. Sometimes, unfortunately, called *sampling error*, sampling variability is no error at all, but just the natural result of random sampling. (p. 284)

Stratified random sample A sampling design in which the population is divided into several subpopulations, or **strata.** Random individuals are then drawn from each stratum so that the sample includes individuals from each, often in a representative proportion. If the strata are homogeneous, but are different from each other, stratified sampling can reduce variability in results. (p. 284)

Cluster sample A sampling design in which entire groups, or **clusters,** are chosen at random. Cluster sampling is usually selected as a matter of convenience, practicality, or cost. Clusters are heterogeneous, and a random sample of clusters should be representative of the population. (p. 286)

Multistage sample Sampling schemes that combine several sampling methods are called multistage samples. For example, a national polling service may stratify the country by geographical regions, select a random sample of cities from each region, and then interview a cluster of residents in each city. (p. 288)

Systematic sample — A sample drawn by selecting individuals systematically from a sampling frame. When there is no relationship between the order of the sampling frame and the variables of interest, a systematic sample can be representative. (p. 288)

Pilot survey — A small trial run of a survey to check whether questions are clear. A pilot survey can reduce errors due to ambiguous questions. (p. 291)

Voluntary response bias — Bias introduced to a sample when individuals can choose on their own whether to participate in the sample. Samples based on voluntary response are always invalid and cannot be recovered, no matter how large the sample size. (p. 292)

Convenience sample — A convenience sample consists of the individuals who are conveniently available. Convenience samples often fail to be representative because every individual in the population is not equally convenient to sample. (p. 293)

Undercoverage — A sampling scheme that biases the sample in a way that gives a part of the population less representation than it has in the population suffers from undercoverage. (p. 293)

Nonresponse bias — Bias introduced when a large fraction of those sampled fails to respond. Those who do respond are likely to not represent the entire population. Voluntary response bias is a form of nonresponse bias, but nonresponse may occur for other reasons. For example, those who are at work during the day won't respond to a telephone survey conducted only during working hours. (p. 294)

Response bias — Anything in a survey design that influences responses falls under the heading of response bias. One typical response bias arises from the wording of questions, which may suggest a favored response. Voters, for example, are more likely to express support of "the President" than support of the particular person holding that office at the moment. (p. 294)

ON THE COMPUTER

Sampling

Computer-generated pseudorandom numbers are usually good enough for drawing random samples. But there is little reason not to use the truly random values available on the Internet.

Here's a convenient way to draw an SRS of a specified size using a computer-based sampling frame. The sampling frame can be a list of names or of identification numbers arrayed, for example, as a column in a spreadsheet, statistics program, or database:

1. Generate random numbers of enough digits so that each exceeds the size of the sampling frame list by several digits. This
2. Assign the random numbers arbitrarily to individuals in the sampling frame list. For example, put them in an adjacent column.
3. Sort the list of random numbers, carrying along the sampling frame list.
4. Now the first n values in the sorted sampling frame column are an SRS of n values from the entire sampling frame.