1 Statistics

1.3 Data Collection

Data Collection

Because it is generally impossible to study an entire population (every individual in a country, all college students, every medical patient, etc.), researchers typically rely on *sampling* to acquire the information, or *data,* needed.

It is important to obtain “good data” because the inferences ultimately made will be based on the statistics obtained from these data.

These inferences are only as good as the data.

Data Collection

Although it is relatively easy to define “good data” as data that accurately represent the population from which they were taken, it is not easy to guarantee that a particular sampling method will produce “good data.”

We need to use sampling (**data collection**) methods that will produce data that are representative of the population and not *biased.*

Sampling method The process of selecting items or events that will become the sample.

Data Collection

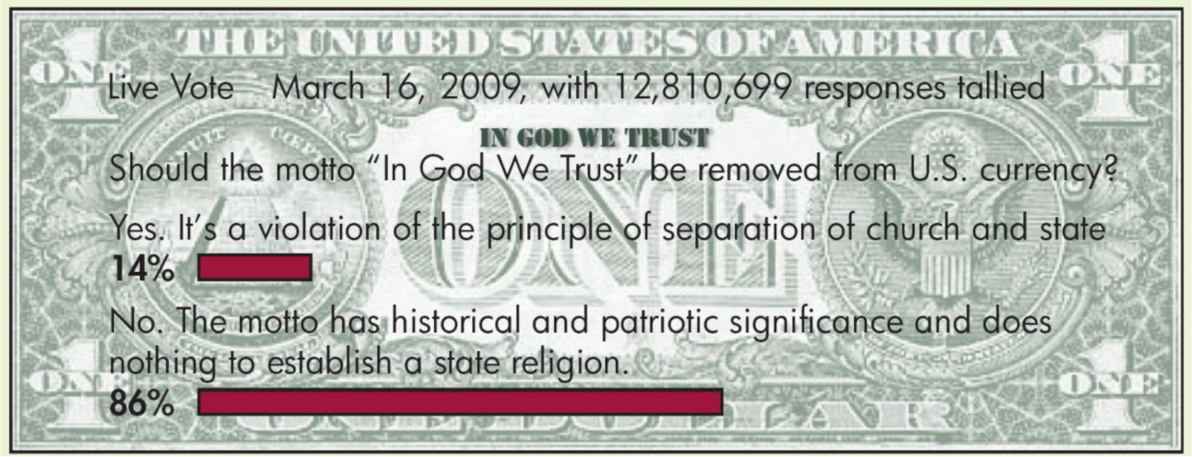
Biased sampling method A sampling method that produces data that systematically differ from the sampled population. Repeated sampling will not correct the bias.

Unbiased sampling method A sampling method that is not biased and produces data that are representative of the sampled population.

Applied Example 7 – *A Modern High-Tech Volunteer Sample*

Public Survey—**Let’s surprise NBC!**

In December 2008, NBC posted the question below on their website to survey the public.



Live Vote March 16, 2009, with 12,810,699 responses tallied

Should the motto “In God We Trust” be removed from U.S. currency?

Yes. It’s a violation of the principle of separation of church and state

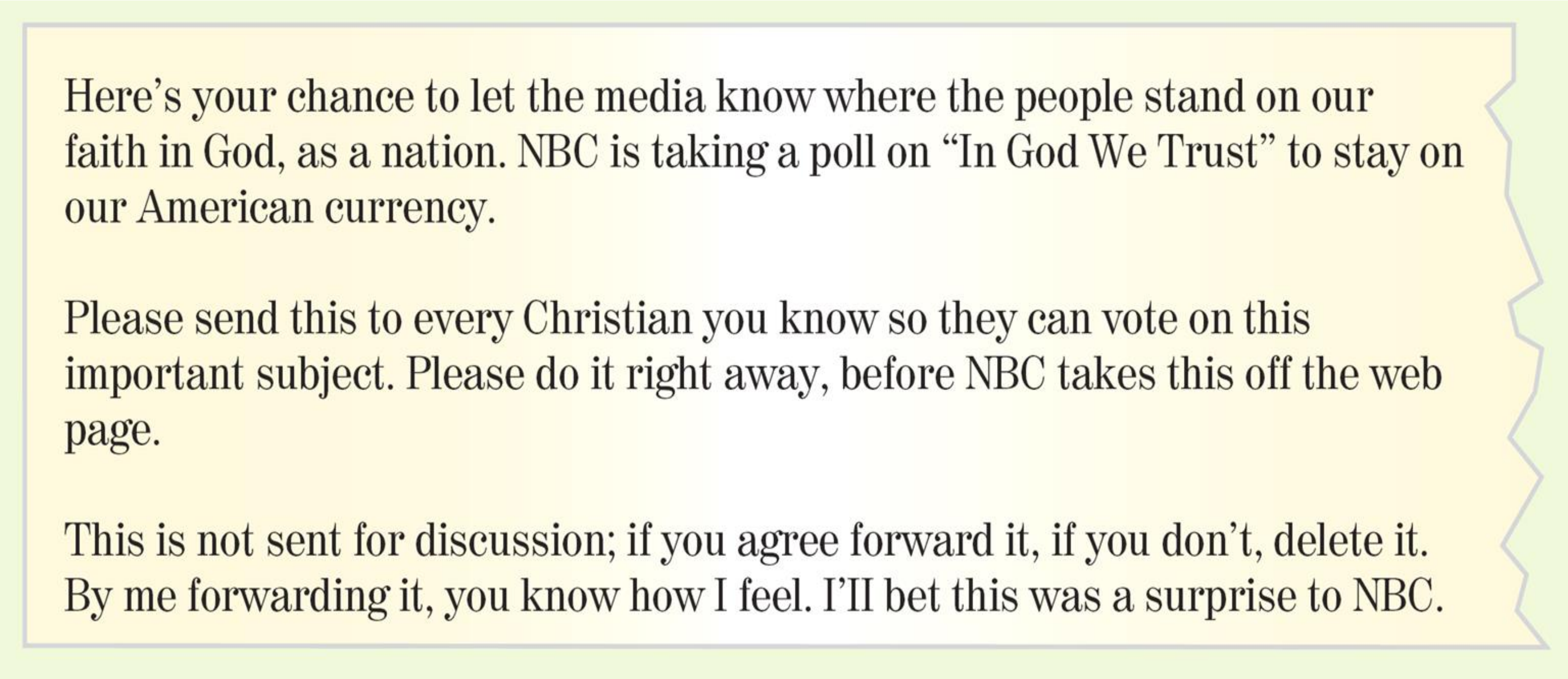
14%

No. The motto has historical and patriotic significance and does nothing to establish a state religion.

86%

Applied Example 7 – *A Modern High-Tech Volunteer Sample* cont’d

At the same time, the e-mail below was being circulated to help “get out the vote.”



Here’s your chance to let the media know where the people stand on our faith in God, as a nation. NBC is taking a poll on “In God We Trust” to stay on our American currency.

Please send this to every Christian you know so they can vote on this important subject. Please do it right away, before NBC takes this off the web page.

This is not sent for discussion; if you agree forward it, if you don’t, delete it. By me forwarding it, you know how I feel. I’ll bet this was a surprise to NBC.

Applied Example 7 – *A Modern High-Tech Volunteer Sample*

cont'd

No meaningful statistical conclusions can be drawn from this survey. The sampling process is severely flawed, and the results were very likely to be strongly biased and not representative of the American population.

Data Collection

Two commonly used sampling methods that often result in biased samples are the *convenience* and *volunteer samples.*

A **convenience sample,** sometimes called a *grab* sample, occurs when items are chosen arbitrarily and in an unstructured manner from a population, whereas a **volunteer sample** consists of results collected from those elements of the population that chose to contribute the needed information on their own initiative.

Data Collection

Did you ever buy a basket of fruit at the market based on the “good appearance” of the fruit on top, only to later discover that the rest of the fruit was not as fresh?

It was too inconvenient to inspect the bottom fruit, so you trusted a convenience sample.

Has your teacher used your class as a sample from which to gather data?

Data Collection

As a group, the class is quite convenient, but is it truly representative of the school’s population?
(Consider the differences among day, evening, and/or weekend students; type of course; etc.)

Have you ever mailed back your responses to a magazine survey?

Under what conditions did (would) you take the time to complete such a questionnaire?

Data Collection

Most people’s immediate attitude is to ignore the survey.

Those with strong feelings will make the effort to respond; therefore, representative samples should not be expected when volunteer samples are collected.

The Data-Collection Process

The Data-Collection Process

The collection of data for statistical analysis is an involved process and includes the following steps:

1. Define the objectives of the survey or study.
   Examples: compare the effectiveness of a new drug to the effectiveness of the standard drug; estimate the average household income in the United States.

2. Define the variable and the population of interest.
   Examples: length of recovery time for patients suffering from a particular disease; total income for households in the United States.

The Data-Collection Process

3. Define the data collection and data-measuring schemes. This includes sampling frame, sampling procedures, sample size, and the data-measuring device (questionnaire, telephone, and so on).

4. Collect your sample. Select the subjects to be sampled and collect the data.

5. Review of the sampling process upon completion of collection.

The Data-Collection Process

Often an analyst is stuck with data already collected, possibly even data collected for other purposes, which makes it impossible to determine whether the data are "good."

Using approved techniques to collect your own data is much preferred.

Applied Example 8 – *Population and Variable of Interest*

The admissions dean at our college wishes to estimate the current “average” cost of textbooks per semester, per student.

The population of interest is the “currently enrolled student body,” and the variable is the “total amount spent for textbooks” by each student this semester.

The Data-Collection Process

Two methods commonly used to collect data are *experiments* and *observational studies.*

In an **experiment**, the investigator controls or modifies the environment and observes the effect on the variable under study.

We often read about laboratory results obtained by using white rats to test different doses of a new medication and its effect on blood pressure.

The Data-Collection Process

The experimental treatments were designed specifically to obtain the data needed to study the effect on the variable.

In an **observational study**, the investigator does not modify the environment and does not control the process being observed.

The data are obtained by sampling some of the population of interest. **Surveys** are observational studies of people.

The Data-Collection Process

If every element in the population can be listed, or enumerated, and observed, then a **census** is compiled.

However, censuses are seldom used because they are often difficult and time-consuming to compile, and therefore very expensive.

Imagine the task of compiling a census of every person who is a potential client at a brokerage firm.

In situations similar to this, a *sample survey* is usually conducted.

The Data-Collection Process

When selecting a sample for a survey, it is necessary to construct a *sampling frame.*

Sampling frame A list, or set, of the elements belonging to the population from which the sample will be drawn.

Ideally, the sampling frame should be identical to the population, with every element of the population included once and only once.

In this case, a census would become the sampling frame. In other situations, a census may not be so easy to obtain because a complete list is not available.

The Data-Collection Process

Lists of registered voters or the telephone directory are sometimes used as sampling frames of the general public.

Depending on the nature of the information being sought, the list of registered voters or the telephone directory may or may not serve as an unbiased sampling frame.

Because only the elements in the frame have a chance to be selected as part of the sample, it is important that the sampling frame be **representative** of the population.

The Data-Collection Process

Once a representative sampling frame has been established, we proceed with selecting the sample elements from the sampling frame.

This selection process is called the **sample design**. There are many different types of sample designs; however, they all fit into two categories: *judgment samples* and *probability samples.*

Judgment samples Samples that are selected on the basis of being judged "typical."

The Data-Collection Process

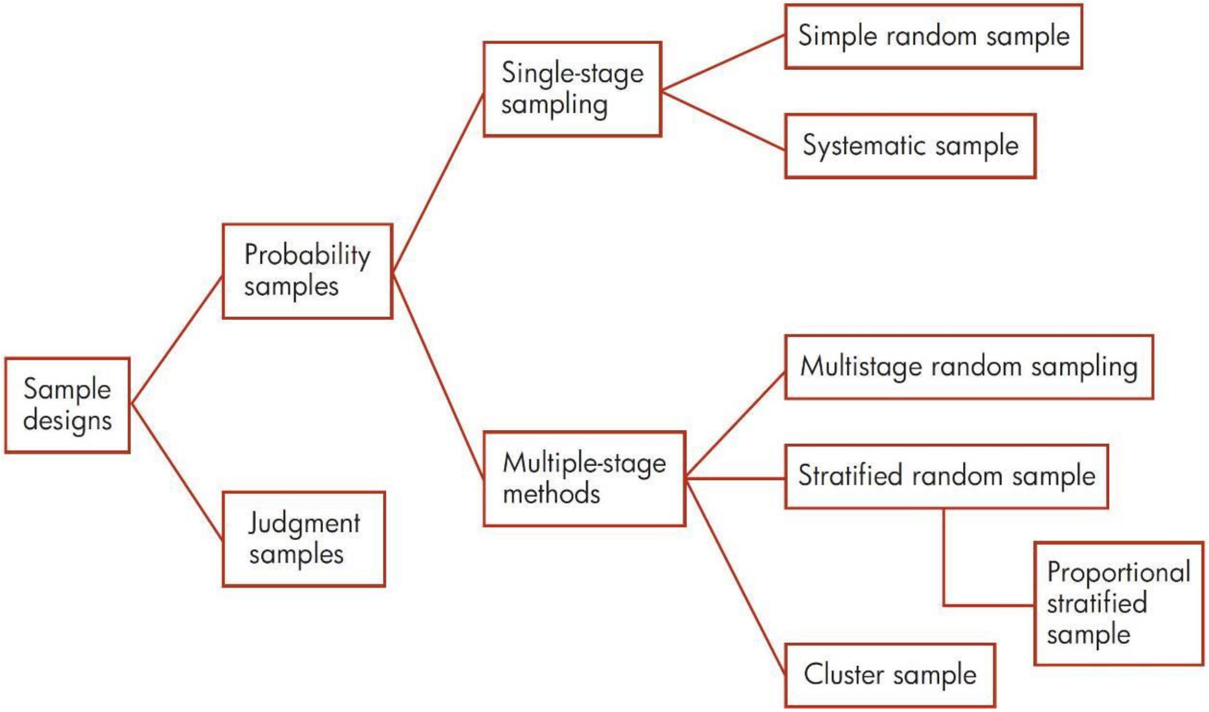
When a judgment sample is collected, the person selecting the sample chooses items that he or she thinks are representative of the population.

The validity of the results from a judgment sample reflects the soundness of the collector’s judgment. This is not an acceptable statistical procedure.

Probability samples Samples in which the elements to be selected are drawn on the basis of probability. Each element in a population has a certain probability of being selected as part of the sample.

The Data-Collection Process

There are many ways to design probability samples. We will look at two of them, single-stage methods and multistage methods, and learn about a few of the many specific designs that are possible.



Single-Stage Methods

Single-Stage Methods

Single-stage sampling A sample design in which the elements of the sampling frame are treated equally and there is no subdividing or partitioning of the frame.

One of the most common single-stage probability sampling methods used to collect data is the *simple random sample.*

Simple random sample A sample selected in such a way that every element in the population or sampling frame has an equal probability of being chosen. Equivalently, all samples of size *n* have an equal chance of being selected.

Single-Stage Methods

Note

Random samples are obtained either by sampling with replacement from a finite population or by sampling without replacement from an infinite population.

Inherent in the concept of randomness is the idea that the next result (or occurrence) is not predictable.

When a random sample is drawn, every effort must be made to ensure that each element has an equal probability of being selected and that the next result does not become predictable.

Single-Stage Methods

The proper procedure for selecting a simple random sample requires the use of random numbers.

Mistakes are commonly made because the term *random* (equal chance) is confused with **haphazard** (without pattern).

To select a simple random sample, first assign an identifying number to each element in the sampling frame.

This is usually done sequentially using the same number of digits for each element.

Single-Stage Methods

Then using random numbers with the same number of digits, select as many numbers as are needed for the sample size desired.

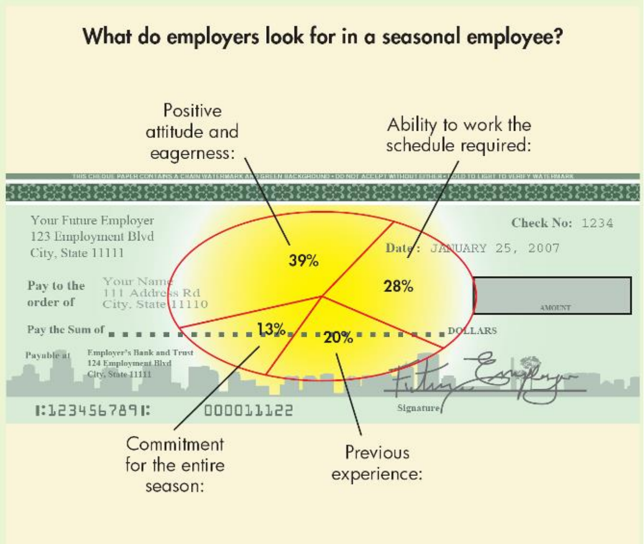
Each numbered element in the sampling frame that corresponds to a selected random number is chosen for the sample.

A simple random sample is our first step toward an unbiased sample.

Without a random design, the conclusions we draw from the statistical procedures may not be reliable.

Example 11 – *Process For Collecting Data*

Consider the graphic “Employers look for positive attitude” and the five steps of the data–collection process. –aa



Employers Look for Positive Attitude

Source: SnagAjob.com survey of 1,043 hiring managers.
Margin of error: ±3 percentage points

Example 11 – *Process For Collecting Data* cont'd

1. *Define the objectives of the survey or experiment.* Determine the opinion of employers regarding what qualities they look for when hiring seasonal employees.

2. *Define the variable and the population of interest.* The variable is the opinion or response to a question concerning qualities or characteristics. The population of interest is all U.S. hiring managers.

Example 11 – *Process For Collecting Data* cont’d

3. *Define the data-collection and data-measuring schemes.* Based on the graphic itself, it can be seen that the source for the percentages presented was SnagAJob.com. Upon further investigation, IPSOS Public Affairs, a third-party research firm, conducted the survey on behalf of the “hourly job website” SnagAJob.com between February 20 and 25, 2009.

   It was an online survey of 1043 hiring managers with responsibility for hiring summer and seasonal employees by the hour.

Example 11 – *Process For Collecting Data* cont'd

4. *Collect the sample.* The information collected from each hiring manager was his or her single "most" essential quality/characteristic that a seasonal employee should possess.

5. *Review the sampling process upon completion of collection*. Since the sampling process was an online survey, were only hiring managers that conduct their business online aware of this survey?

   Were various areas of the country and types of businesses represented?

   Perhaps you can think of additional concerns.

Single-Stage Methods

In concept, the simple random sample is the simplest of the probability sampling techniques, but it is seldom used in practice because it often is an inefficient technique.

One of the easiest-to-use methods for approximating a simple random sample is the *systematic sampling method.*

Systematic sample A sample in which every *k*th item of the sampling frame is selected, starting from a first element, which is randomly selected from the first *k* elements.

Single-Stage Methods

To select an *x* percent (%) systematic sample, we will need to randomly select 1 element from every $\frac{100}{x}$ elements.

After the first element is randomly located within the first $\frac{100}{x}$ elements, we proceed to select every $\frac{100}{x}$ th item thereafter until we have the desired number of data values for our sample.

For example, if we desire a 3% systematic sample, we would locate the first item by randomly selecting an integer between 1 and 33 ($\frac{100}{x} = \frac{100}{3}$ = 33.33, which when rounded becomes 33).

Single-Stage Methods

Suppose 23 was randomly selected. This means that our first data value is obtained from the subject in the 23rd position in the sampling frame.

The second data value will come from the subject in the 56th (23 + 33 = 56) position; the third, from the 89th (56 + 33); and so on, until our sample is complete.

The systematic technique is easy to describe and execute; however, it has some inherent dangers when the sampling frame is repetitive or cyclical in nature.

Single-Stage Methods

For example, a systematic sample of every *k*th house along a long street might result in a sample disproportional with regard to houses on corner lots.

The resulting information would likely be biased if the purpose for sampling is to learn about support for a proposed sidewalk tax. In these situations the results may not approximate a simple random sample.

Multistage Methods

Multistage Methods

Multistage random sampling A sample design in which the elements of the sampling frame are subdivided and the sample is chosen in more than one stage.

Multistage sampling designs often start by dividing a very large population into subpopulations on the basis of some characteristic. These subpopulations are called **strata**.

These smaller, easier-to-work-with strata can then be sampled separately. One such sample design is the *stratified random sampling method.*

Multistage Methods

Stratified random sample A sample obtained by stratifying the population, or sampling frame, and then selecting a number of items from each of the strata by means of a simple random sampling technique.

A stratified random sample results when the population, or sampling frame, is subdivided into various strata, usually some already occurring natural subdivisions, and then a subsample is drawn from each of these strata.

These subsamples may be drawn from the various strata by using random or systematic methods.

Multistage Methods

The subsamples are summarized separately first and then combined to draw conclusions about the entire population.

When a population with several strata is sampled, we often require that the number of items collected from each Stratum be proportional to the size of the strata; this method is called a *proportional stratified sampling.*

Proportional stratified sample A sample obtained by stratifying the population, or sampling frame, and then selecting a number of items in proportion to the size of the strata from each strata by means of a simple random sampling technique.

Multistage Methods

A convenient way to express the idea of proportional sampling is to establish a quota.

For example, the quota “1 for every 150” directs you to select 1 data value for each 150 elements in each strata.

That way, the size of the strata determines the size of the subsample from that strata.

The subsamples are summarized separately and then combined to draw conclusions about the entire population.

Multistage Methods

Another sampling method that starts by stratifying the population, or sampling frame, is a *cluster sample.*

Cluster sample A sample obtained by stratifying the population, or sampling frame, and then selecting some or all of the items from some, but not all, of the strata.

The cluster sample is a multistage design. It uses either random or systematic methods to select the strata (clusters) to be sampled (first stage) and then uses either random or systematic methods to select elements from each identified cluster (second stage).

Multistage Methods

The cluster sampling method also allows the possibility of selecting all of the elements from each identified cluster. Either way, the subsamples are summarized separately and the information then combined.

To illustrate a possible multistage random sampling process, consider that a sample is needed from a large country. In the first stage, the country is divided into smaller regions, such as states, and a random sample of these states is selected.

Multistage Methods

In the second stage, a random sample of smaller areas within the selected states (counties) is then chosen.

In the third stage, a random sample of even smaller areas (townships) is taken within each county.

Finally in the fourth stage, if these townships are sufficiently small for the purposes of the study, the researcher might continue by collecting simple random samples from each of the identified townships.

This would mean that the entire sample was made up of several “local” subsamples identified as a result of the several stages.

Multistage Methods

Sample design is not a simple matter; many colleges and universities offer separate courses in sample surveying and experimental design.

The topic of survey sampling is a complete textbook in itself. It is thus intended for the preceding information to provide you with an overview of sampling and put its role in perspective.