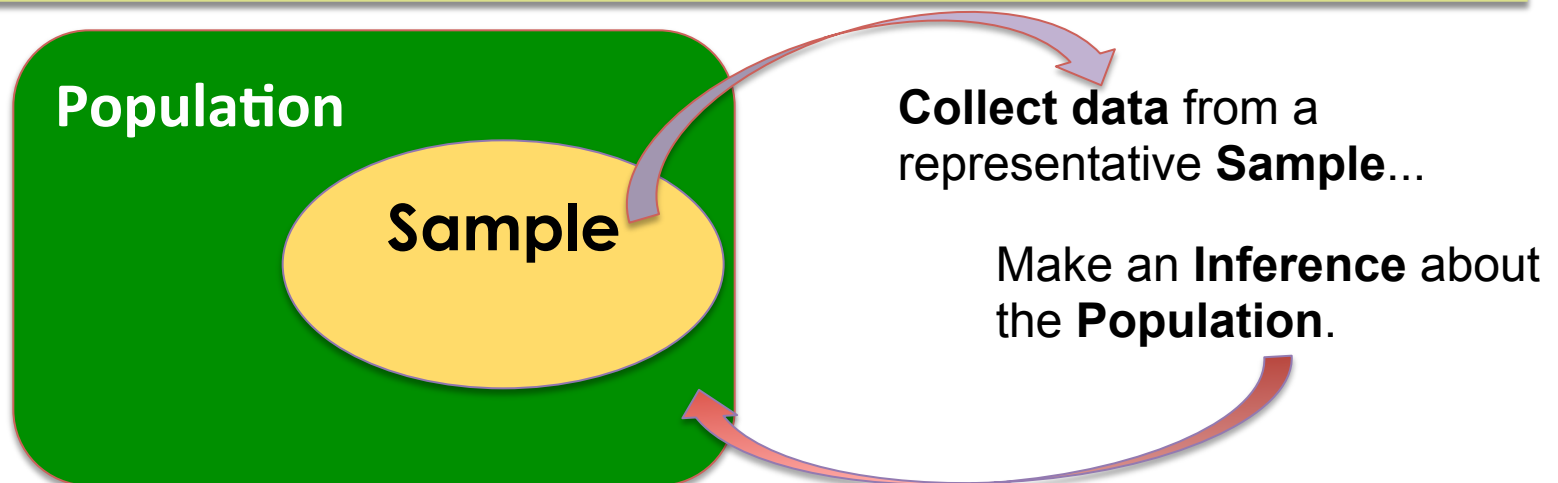


# Population and Sample

## Chapter 8. Producing data: Sampling

The **population** in a statistical study is the entire group of individuals about which we want information.

A **sample** is the part of the population from which we actually collect information. We use information from a sample to draw conclusions about the entire population.



**Ex. Describe the population and the sample.**

For each of the following situations, describe the population and the sample.

(a) A survey of 17,096 students in U.S. four-year colleges reported that 19.4% were binge drinkers.

Population: college students. Sample: 17,096 students.

(b) In a study of work stress, 100 female restaurant workers were asked about the impact of work stress on their personal lives.

Population: female restaurant workers. Sample: 100 workers.

(c) A tract of forest has 584 longleaf pine trees. The diameters of 40 of these trees were measured.

Population: longleaf pine trees. Sample: 584 trees.

**Ex.** A political scientist wants to know how college students feel about the Social Security system. She obtains a list of the 3456 undergraduates at her college and mails a form to 250 students selected at random. Only 104 forms are returned.

- The population is all college students.
- The sample is the 104 students at the researcher's college who returned the questionnaire.

# Sample survey

- The first step in planning a **sample survey** is to say exactly *what population* we want to describe.
- The second step is to say exactly *what we want to measure*, that is, to give exact definitions of our variables.
- The final step in planning a sample survey is the sampling design. A sampling design is a specific method for choosing a sample from the population.

# How to Sample Badly

The design of a sample is **biased** if it systematically favors certain outcomes.

Choosing individuals who are easiest to reach results in a **convenience sample**.

A **voluntary response sample** consists of people who choose themselves by responding to a general appeal. Voluntary response samples show bias because people with strong opinions (often in the same direction) are most likely to respond.

**Ex.** You see a woman student standing in front of the student center, now and then stopping other students to ask them questions. She says that she is collecting student opinions for a class assignment.

It is a **convenience sample**.

She is only getting opinions from students who are at the student center at a certain time of day.

This might underrepresent some group:  
commuters, graduate students, etc.

**Ex.** A sample of mall shoppers is fast and cheap. But people at shopping malls tend to be more prosperous than typical Americans. They are also more likely to be teenagers or retired. Moreover, unless interviewers are carefully trained, they tend to question well-dressed, respectable-looking people and avoid poorly dressed or tough-looking individuals. In short, mall interviews will not contact a sample that is representative of the entire population.

Interviews at shopping malls will almost always overrepresent middle-class and retired people and underrepresent the poor. This is **bias**: the outcomes of mall surveys will repeatedly miss the truth about the population in the same ways.

Suppose, for example, that a news show asks viewers to participate in an on-line poll.

This would be a voluntary sample. The sample is chosen by the viewers, not by the survey administrator.

**Voluntary response samples** are always biased: they only include people who choose to volunteer, whereas a random sample would need to include people whether or not they choose to volunteer.

Often, voluntary response samples oversample people who have strong opinions and undersample people who don't care much about the topic of the survey.



**Ex.** A study in El Paso, Texas, looked at seat belt use by drivers. Drivers were observed at randomly chosen convenience stores. After they left their cars, they were invited to answer questions that included questions about seat belt use. In all, 75% said they always used seat belts, yet only 61.5% were wearing seat belts when they pulled into the store parking lots. Explain the reason for the bias observed in responses to the survey.

**Do you expect bias in the same direction in most surveys about seat belt use?**

People likely claim to wear their seat belts because they know they should. Such bias is likely in most similar surveys.

**Ex. Why the random samples report public opinion more reliably than the online poll?**

In general, online polls, call-in polls, and voluntary response polls tend to attract responses from those who have strong opinions on the subject, and therefore, such polls are often not representative of the population as a whole.

# Simple Random Samples

**Random sampling**, the use of chance to select a sample, is the central principle of statistical sampling.

A **simple random sample (SRS)** of size  $n$  consists of  $n$  individuals from the population chosen in such a way that every set of  $n$  individuals has an equal chance to be the sample actually selected.

In practice, people use random numbers generated by a computer or calculator to choose samples. If you don't have technology handy, you can use a **table of random digits**.

# How to Choose a SRS

A **table of random digits** is a long string of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with these properties:

- Each entry in the table is equally likely to be any of the 10 digits 0–9.
- The entries are independent of each other. That is, knowledge of one part of the table gives no information about any other part.

Table B at the back of the book is a table of random digits. To make the table easier to read, the digits appear in groups of five and in numbered rows. The groups and rows have no meaning — the table is just a long list of randomly chosen digits.

## How to Choose an SRS Using Table B

**Step 1: Label.** Give each member of the population a numerical label of the *same length*.

**Step 2: Table.** To choose an SRS, read from Table B successive groups of digits of the length you used as labels.

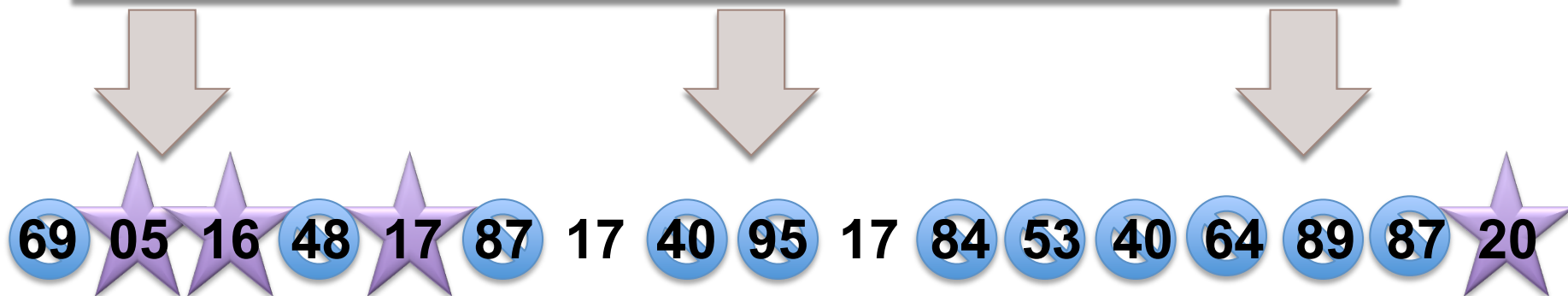
**Your sample contains the individuals whose labels you find.**

# SRS Example

Use the random digits provided to select an SRS of four hotels.

01 Aloha Kai	08 Captiva	15 Palm Tree	22 Sea Shell
02 Anchor Down	09 Casa del Mar	16 Radisson	23 Silver Beach
03 Banana Bay	10 Coconuts	17 Ramada	24 Sunset Beach
04 Banyan Tree	11 Diplomat	18 Sandpiper	25 Tradewinds
05 Beach Castle	12 Holiday Inn	19 Sea Castle	26 Tropical Breeze
06 Best Western	13 Lime Tree	20 Sea Club	27 Tropical Shores
07 Cabana	14 Outrigger	21 Sea Grape	28 Veranda

69051 64817 87174 09517 84534 06489 87201 97245



**Our SRS of four hotels for the editors to contact is: 05 Beach Castle, 16 Radisson, 17 Ramada, and 20 Sea Club.**

The walk to your statistics class takes about 10 minutes, about the amount of time needed to listen to three songs on your iPod. You decide to take a simple random sample of songs from a Billboard list of Rock Songs. Here is the list:

Miss Murder	Animal I Have Become	Steady, As She Goes	Dani California
The Kill (Bury Me)	Original Fire	When You Were Young	MakeD—Sure
Vicarious	The Diary of Jane		

Select the three songs for your iPod using a simple random sample.

Line #104 in the Table B:

52711      38889      93074      60227      40011      85848      48767      52573

# Random sampling

**Random sampling**, the use of impersonal chance to avoid bias, is the basic idea of sampling design. Online polls and mall interviews also produce samples, but these samples are chosen in ways that invite bias. *The first question to ask about any sample is whether it was chosen at random.*

## Why should we rely on random sampling?

1. To eliminate bias in selecting samples from the list of available individuals.
2. The laws of probability allow trustworthy inference about the population.
  - Results from random samples come with a **margin of error** that sets bounds on the size of the likely error.
  - Larger random samples give better information about the population than smaller samples.

**Ex.** Just before a presidential election, a national opinion-polling firm increases the size of its weekly sample from the usual 1500 people to 4000 people. Why do you think the firm does this?

With the election close at hand, the polling organization wants to increase the accuracy of its results.

**Larger samples provide better information about the population.**

We need an accurate and complete list of the population. Because such a list is rarely available, most samples suffer from some degree of ***undercoverage***.

**Ex.** A sample survey of households, for example, will miss not only homeless people but prison inmates and students in dormitories.

**Ex.** An opinion poll conducted by calling landline telephone numbers will miss households that have only cell phones as well as households without a phone. The results of national sample surveys therefore have some bias if the people not covered differ from the rest of the population.



A more serious source of bias in most sample surveys is ***nonresponse***, which occurs when a selected individual cannot be contacted or refuses to cooperate.

**Ex.** Nonresponse to sample surveys often exceeds 50%, even with careful planning and several callbacks. If the people contacted differ from those who are rarely at home or who refuse to answer questions, some bias remains.

Most national sample surveys are carried out by telephone, using **random digit dialing** to choose residential telephone numbers at random. Call screening is increasing nonresponse to such surveys, and the rise of cell-phone-only households is increasing undercoverage.

In addition, the behavior of the respondent or the interviewer can cause **response bias** response bias in sample results.

People know that they should take the trouble to vote, for example, so many who didn't vote in the last election will tell an interviewer that they did.

Answers to questions that ask respondents to recall past events are often inaccurate.

For example, many people “telescope” events in the past, bringing them forward in memory to more recent time periods. “Have you visited a dentist in the last 6 months?” will often draw a “Yes” from someone who last visited a dentist 8 months ago.

- Careful training of interviewers and careful supervision to avoid variation among interviewers can reduce response bias.
- Good interviewing technique is another aspect of a well-done sample survey.

The **wording of questions** wording effects is the most important influence on the answers given to a sample survey. Confusing or leading questions can introduce strong bias, and changes in wording can greatly change a survey's outcome. Even the order in which questions are asked matters.

**Ex.** Ask a sample of college students these two questions:  
“How happy are you with your life in general?” (Answers on a scale of 1 to 5)  
“How many dates did you have last month?”

The correlation between answers is  $r = -0.012$  when asked in this order. It appears that dating has little to do with happiness.

Reverse the order of the questions, however, and  $r = 0.66$ .

Asking a question that brings dating to mind makes dating success a big factor in happiness.

# Cautions About Sample Surveys

Good sampling technique includes the art of reducing all sources of error.

**Undercoverage** occurs when some groups in the population are left out of the process of choosing the sample.

**Nonresponse** occurs when an individual chosen for the sample can't be contacted or refuses to participate.

A systematic pattern of incorrect responses in a sample survey leads to **response bias**.

The **wording of questions** is the most important influence on the answers given to a sample survey.

## Exercises.

1. Using a local telephone book to select a simple random sample could introduce what type of bias?

- a) Under-coverage bias
- b) Non-response bias
- c) Response bias
- d) Question wording bias

2. If people tend to respond differently to a question depending on whether the interviewer is male or female, which type of bias is present?

- a) Under-coverage bias
- b) Non-response bias
- c) Response bias
- d) Question wording bias

## Exercises.

A 45,000-pound truckload of potatoes is considered for purchase by a potato chip company. The company selects 150 pounds of potatoes from 5 points in the shipment for inspection. If the fraction of acceptable potatoes is high enough in the 150-pound selection of potatoes, the shipment will be purchased.

3. What is the population in this example?

- a) all potatoes in the world
- b) all potatoes in the U.S.
- c) all potatoes in the truckload
- d) all potatoes in the 150-pound selection

4. What is the sample in this example?

- a) all potatoes in the world
- b) all potatoes in the U.S.
- c) all potatoes in the truckload
- d) all potatoes in the 150-pound selection

5. The Web site [www.twiigs.com](http://www.twiigs.com) allows you to vote on polls that interest you or to post one of your own. Once you have found a poll of interest, you just click on “Vote” and your response becomes part of the sample. One of the questions in July 2010 was “How many times have you been pulled over by the police?” Of the 780 people responding, 70% said “1–5 times.” You can conclude that

- (a) about 70% of Americans have been pulled over by the police “1–5 times.”
- (b) the poll uses voluntary response, so the results tell us little about the population of all adults.
- (c) more people still need to vote on the question, as a larger sample is required to reduce bias.

6. A sample of households in a community is selected at random from the telephone directory. In this community, 4% of households have no telephone, 10% have only cell phones, and another 25% have unlisted telephone numbers. The sample will certainly suffer from

- (a) nonresponse.
- (b) undercoverage.
- (c) false responses.

## 7. Simple Random Sample

Select an SRS of 3 packages of peanuts from a case containing 30 packages of peanuts. Label the packages 01 to 30 and use the line from the table of random digits shown below.

14459 26056 31424 80371 65103 62253 22490 61181

Your SRS contains the packages of peanuts labeled

- a) 14, 44, 45
- b) 14, 26, 05
- c) 14, 45, 92
- d) 14, 03, 10

8. The Pew Research Center survey asked a random sample of 1500 adults, “Do you think the use of marijuana should be made legal, or not?” In the entire sample, 41% said, “Yes, legal.” But only 24% of the Republicans in the sample said, “Yes, legal.” Which of these two sample percents will be more accurate as an estimate of the truth about the population?

- (a) The result for Republicans is more accurate because it is easier to estimate a proportion for a smaller group.
- (b) The result for the entire sample is more accurate because it comes from a larger sample.
- (c) Both are equally accurate because both come from the same sample.