Chapter 4: Designing Studies

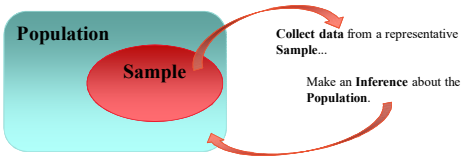
Section 4.1
Samples and Surveys

The Practice of Statistics, 4th edition – For AP*
STARNES, YATES, MOORE

Population, Census, and Sample

Definition:
The **population** in a statistical study is the entire group of individuals we want information about. A **census** attempts to contact everyone in the entire population.

A **sample** is a subset of individuals in the population from which we actually collect data.



Example 1: Identify the population and the sample in each of the following settings.

a) A furniture maker buys hardwood in large batches. The supplier is supposed to dry the wood before shipping (wood that isn't dry won't hold its size and shape). The furniture maker chooses five pieces of wood from each batch and tests their moisture content. If any piece exceeds 12% moisture content, the entire batch is sent back.

The population is all the pieces of hardwood in a batch.

The sample is the five pieces of wood that are selected from that batch and tested for moisture content.

b) Each week, the Gallup Poll questions a sample of about 1500 adult U.S. residents to determine national opinion on a wide variety of issues.

Gallup's population is all adult U.S. residents.

Their sample is the 1500 adults who actually respond to the survey questions.

The Idea of a Sample Survey

We often draw conclusions about a whole population on the basis of a sample. Choosing a sample from a large, varied population is not that easy.

Step 1: Define the *population* we want to describe.

Step 2: Say exactly *what we want to measure.*

A "sample survey" is a study that uses an organized plan to choose a sample that represents some specific population.

Step 3: Decide how to choose a sample from the population.

How to Sample Badly

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

Definition:
The design of a statistical study shows **bias** if it systematically favors certain outcomes. Avoid at all costs.

AP EXAM TIP: If you're asked to describe how the design of a study leads to bias, you're expected to do two things: (1) identify a problem with the design, and (2) explain how this problem would lead to an underestimate or overestimate. Suppose you were asked, "Explain how using your statistics class as a sample to estimate the proportion of all high school students who own a graphing calculator could result in bias." You might respond, "This is a convenience sample. It would probably include a much higher proportion of students with a graphing calculator than in the population at large because a graphing calculator is required for the statistics class. So this method would probably lead to an overestimate of the actual population proportion."

Definition:
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.

Example 2: Former CNN commentator Lou Dobbs doesn't like illegal immigration. One of his shows was largely devoted to attacking a proposal to offer driver's licenses to illegal immigrants. During the show, Mr. Dobbs invited his viewers to go to loudobbs.com to vote on the question "Would you be more or less likely to vote for a presidential candidate who supports giving driver's licenses to illegal aliens? The result: 97% of the 7350 people who voted by the end of the show said, "Less likely." What type of sample did Mr. Dobbs use in his poll? Explain how this sampling method could lead to bias in the poll results.

Mr. Dobbs used a voluntary response sample: people chose to go online and respond. Those who voted were viewers of Mr. Dobbs's program, which means that they are likely to support his views. The 97% poll result is probably an extreme overestimate of the percent of people in the population who would be less likely to support a presidential candidate with this position.

How to Sample Well: Random Sampling

Random sampling involves using a chance process to determine which members of a population are included in the sample.

Definition:
A **simple random sample** (SRS) of size *n* is chosen in such a way that every group of *n* individuals in the population has an equal chance to be selected as the sample.

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.**

When you think of an SRS, picture drawing names from a hat to remind yourself that an SRS doesn't favor any part of the population. That's why an SRS is a better method of choosing samples than convenience or voluntary response samples. But writing names on slips of paper and drawing them from a hat doesn't work as well if the population is large.

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 through 9.
*The entries are independent of each other. That is, knowledge of one part of the table gives no information about any other part.

HOW TO CHOOSE AN SRS USING TABLE D
Step 1: Label. Give each member of the population a numerical label with the same number of digits. Use as few digits as possible.

Step 2: Randomize. Read consecutive groups of digits of the appropriate length from left to right across a line in **Table D**. Ignore any group of digits that wasn't used as a label or that duplicates a label already in the sample. Stop when you have chosen *n* different labels.

Your sample contains the individuals whose labels you find.

Example 3: The school newspaper is planning an article on family-friendly places to stay over spring break at a nearby beach town. The editors intend to call 4 randomly chosen hotels to ask about their amenities for families with children. They have an alphabetized list of all 28 hotels in the town. Use Table D at line 130 to choose an SRS of 4 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

69 05 16 48 17 87 17 40 95 17 84 53 34 64 89 87 20

Our SRS of 4 hotels for the editors to contact is: 05 Beach Castle, 16 Radisson, 17 Ramada, and 20 Sea Club.

Other Sampling Methods
One of the most common alternatives to an SRS involves sampling groups (**strata**) of similar individuals within the population separately. Then these separate "subsamples" are combined to form one **stratified random sample**.

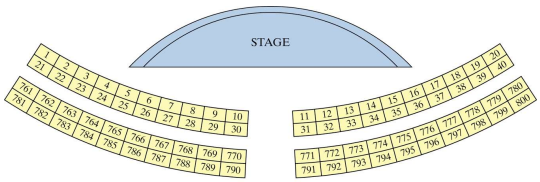
Definition:
To get a **stratified random sample**, start by classifying the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the sample.

Choose the strata based on facts known before the sample is taken. For example, in a study of sleep habits on school nights, the population of students in a large high school might be divided into freshman, sophomore, junior, and senior strata.

To get a **cluster sample**, start by classifying the population into groups of individuals that are located near each other, called **clusters**. Then choose an SRS of the clusters. All individuals in the chosen clusters are included in the sample.

In a cluster sample, some people take an SRS from each cluster rather than including all members of the cluster.

Example 4: The student council wants to conduct a survey during the first five minutes of an all-school assembly in the auditorium about use of the school library. They would like to announce the results of the survey at the end of the assembly. The student council president asks your statistics class to help carry out the survey. There are 800 students present at the assembly. A map of the auditorium is shown below. Note that students are seated by grade level and that the seats are numbered from 1 to 800.



9th grade: Seats 601–800	10th grade: Seats 401–600
11th grade: Seats 201–400	12th grade: Seats 1–200

Describe how you would use each of the following sampling methods to select 80 students to complete the survey.

a) Simple random sample

To take an SRS, we need to choose 80 of the seat numbers at random. Use randInt(1,800) on your calculator until 80 different seats are selected. Then give the survey to the students in those seats.

Write all 800 numbers on identical slips of paper, place them in a hat, mix it up well, choose 80 slips of paper and interview the students in those seats.

b) Stratified random sample

The students in the assembly are seated by grade level. Because students' library use might be similar within grade levels but different across grade levels, we'll use the grade level seating areas as our strata. Within each grade's seating area, we'll select 20 seats at random. For the 9th grade, use randInt(601,800) to select 20 different seats. Use randInt(401,600) to pick 20 different sophomore seats, randInt(201,400) to get 20 different junior seats, and randInt(1,200) to choose 20 different senior seats. Give the survey to the students in the selected seats.

We'll use the grade level seating areas as our strata. Within each grade's seating area, we'll select 20 seats at random. For the 9th grade, we will place the numbers 601 – 800 in a hat, mix it up well, choose 20 slips of paper and interview the students in those seats. For the 10th grade, we will place the numbers 401 – 600 in a hat, mix it up well, choose 20 slips of paper and interview the students in those seats. For the 11th grade, we will place the numbers 201 – 400 in a hat, mix it up well, choose 20 slips of paper and interview the students in those seats. For the 12th grade, we will place the numbers 1 – 200 in a mix it up well, choose 20 slips of paper and interview the students in those seats.

c) Cluster sample

With the way students are seated, each column of seats from the stage to the back of the auditorium could be used as a cluster. Note that each cluster contains students from all four grade levels, so each should represent the population well. Because there are 20 clusters, each with 40 seats, we need to choose 2 clusters at random to get 80 students for the survey. Use randInt(1,20) to select two clusters, and then give the survey to all 40 students in each column of seats.

With the way students are seated, each column of seats from the stage to the back of the auditorium could be used as a cluster. Note that each cluster contains students from all four grade levels, so each should represent the population well. Because there are 20 clusters, each with 40 seats, we need to choose 2 clusters at random to get 80 students for the survey. Write the numbers 1 – 20 on identical slips of paper, place them in a het, mix it up well, and choose 2 slips of paper. Interview all the students in each cluster.

Example 5: The manager of a sports arena wants to learn more about the financial status of the people who are attending an NBA basketball game. He would like to give a survey to representative sample of the more than 20,000 fans in attendance. Ticket prices for the game vary in a great deal: seats near the court cost over $100 each, while seats in the top rows of the arena cost $25 each. The arena is divided into 30 numbered sections, from 101 to 130. Each section has rows of seats labeled with letters from A (nearest to the court) to ZZ (top row of the arena).

a) Explain why it might be difficult to give the survey to an SRS of 200 fans.

You would have to identify 200 different seats, go to those seats in the arena, and find the people who are sitting there, which would take a lot of time.

b) Which would be a better way to take a stratified random sample of fans: using the lettered rows or the numbered sections as strata? Explain.

It is best to create strata where the people within a stratum are very similar to each other but different than the people in other strata. In this case, it would be better to take the lettered rows as the strata because each lettered row is the same distance from the court and so would contain only seats with the same (or nearly the same) ticket price.

c) Which would be a better way to take a cluster sample of fans: using the lettered rows or the numbered sections as clusters? Explain.

It is best if the people in each cluster reflect the variability found in the population. In this case, it would be better to take the numbered sections as the clusters because they include all different seat prices.

Inference for Sampling

The purpose of a sample is to give us information about a larger population. The process of drawing conclusions about a population on the basis of sample data is called **inference** because we infer information about the population from what we know about the sample.

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.

Sample Surveys: What Can Go Wrong?

Most sample surveys are affected by errors in addition to sampling variability.

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

Undercoverage occurs when some members of the population cannot be chosen in a sample.

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

A systematic pattern of inaccurate answers in a survey leads to **response bias**.

The **wording of questions** 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.