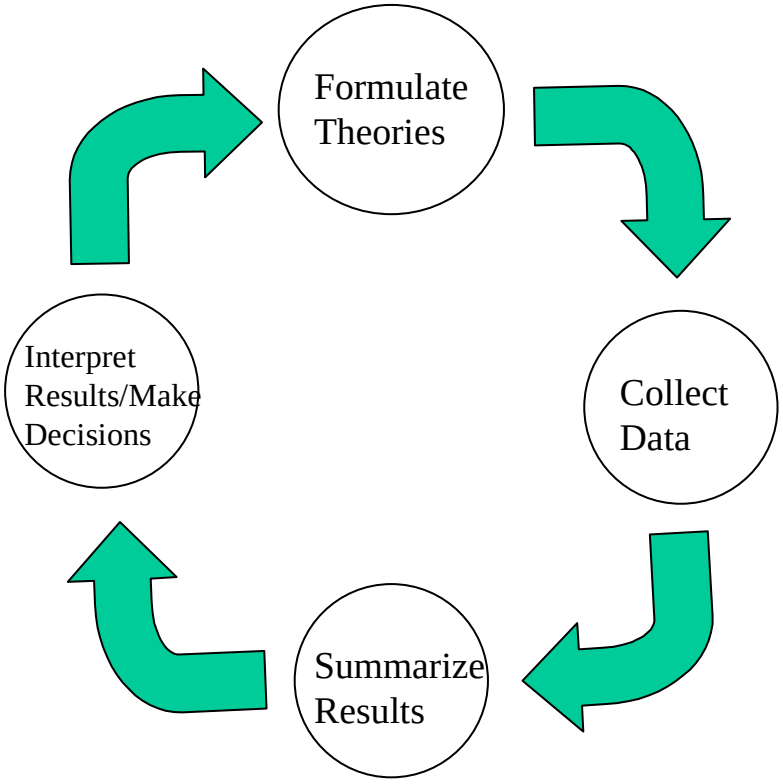
Chapter Goals

To introduce you to data collection

You will

- learn to think critically about the data collected or presented
- learn various methods for selecting a sample



Why Sample?

Definition *A **census** is a sample of the entire population*

FINISHED FILES ARE THE RESULT OF YEARS OF SCIENTIFIC STUDY COMBINED WITH THE EXPERIENCE OF MANY YEARS

The Language of Sampling

- A **population** or **universe** is the total set of elements of interest for a given problem.
- A **finite population** consists of a finite number of elements. Population sizes are usually represented by N.
- An **infinite population**consists of an indefinitely large number of elements.
- A **sample** is a part of the population under study selected so that inferences can be drawn from it about the population. Sample sizes are usually represented by n.
- **Sampling error (variation)** is the difference between the result obtained from a sample and the result that would be obtained from a census conducted by using the same procedures as in the sample.
- **Parameters** are numerical descriptive measures of populations / processes.
- **Statistics** are numerical descriptive measures of samples (more generally, a quantity computed from the observations in a sample).

Let's do it! 2.1

Exercise *Nine percent of the US population has Type B blood. In a sample of* 400 *individuals from the US population,* 12.5% *were found to have Type B blood. Circle your answer:*

- *In this particular situation, the value* 9% *is a (parameter, statistic)*
- *In this particular situation, the value* 12.5% *is a (parameter, statistic)*

Good Data?

Definition *A sampling method is **biased** if it produces results that systematically differ from the truth about the population.*

Example ***Convenience samples** and **volunteer samples** generally lead to biased samples.*

Definition ***Selection bias** is the systematic tendency on the part of the sampling procedure to exclude or include a certain part of the population*

Definition ***Nonresponse bias** is the distortion that can arise because a large number of units selected for the sample do not respond.*

Definition ***Response bias** is the distortion that arises because of the wording of a question or the behavior of the interviewer.*

For a sample to be representative, we must take care in how it is selected. Errors in a sampling design can lead to errors in the conclusions drawn from the sample data. The best defense against bias is **randomization**, in which each individual is given a fair, random, chance of selection.

Example *In the election of* 1936 *the Literary Digest magazine predicted that challenger Alf Landon would beat the incumbent, Franklin Roosevelt. They based their prediction on a survey of ten million citizens taken from lists of car and telephone owners, of whom over* 2.3 *million responded. This was the largest response to any poll in history, and based on this, the Literary Digest predicted that Landon would win* 57% *to* 43%. *In reality, Roosevelt won* 62% *to* 38%. *What went wrong?At the same time, a young man known as George Gallup surveyed* 50,000 *people and correctly predicted that Roosevelt would win the election.*

Let's do it! 2.3

A study was conducted to estimate the average size of households in the US. A total of 1000 people were randomly selected from the population and they were asked to report the number of people in their household. The average of these 1000 responses was found to be 4.6.

1. What is the population of interest?

2. What is the parameter of interest?

3. An average computed in this manner tends to be larger than the true average size of households in the US. True or false? Explain.

Probability Sampling Methods

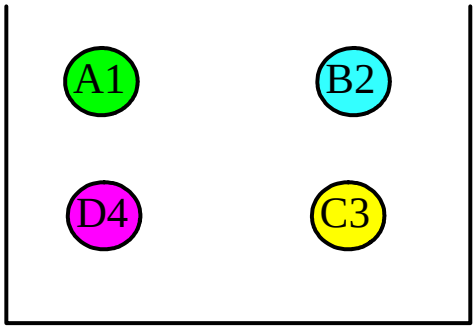
Definition *A sampling method that gives each unit in the population a known, non-zero chance of being selected is called a* ***probability sampling method****.*

Probability Sampling Methods:

- Simple random sampling
- Stratified random sampling
- Systematic sampling
- Cluster sampling
- Multistage sampling

Simple Random Sampling

Sample size $n = 1$: A method of sampling that gives each item in the population an equal chance of being selected.



Sample size $n > 1$: A method of sampling that gives each possible sample combination an equal probability of being chosen. Then, the possible combinations are (using sampling with replacement)

1^{st} Observation	2^{nd} Observation			
	A	B	C	D
A	AA	AB	AC	AD
B	BA	BB	BC	BD
C	CA	CB	CC	CD
D	DA	DB	DC	DD

If random sampling is employed, what is the chance of choosing a particular combination of 2 items?

Simple Random Sampling (cont.)

1. Finite versus Infinite Populations

2. Sampling with and without replacement

3. Ways of choosing a simple random sample
   a. Random number tables

   b. Calculator/Computer

   c. Other

4. Advantages and disadvantage of taking simple random samples

Let's do it! 2.4

Form a group of $N = 10$ students. Now choose a random sample of $n = 3$ from this population.

Steps:

1. Identify and label your population
2. Select your sample by randomly choosing from the available labels

Now, what do we do with this population...

Stratified Random Sampling

Definition *A **stratified random sample** is selected by dividing the population into **mutually exclusive** subgroups, and then taking a simple random sample from each subgroup. The simple random samples are then combined to give the full sample.*

Stratified random sampling

- allows us to obtain information about each subgroup
- can be more efficient than simple random sampling

Example

Systematic Sampling

Definition *For a* **1-*in-k* systematic sample**, *you order the units of the population in some way and randomly select one of the first k units in the ordered list. This selected unit is the first unit to be included in the sample. You continue through the list selecting every* k^{th} *unit from then on.*

Advantages and Disadvantages

- Convenient
- Fast
- Could be biased

Cluster Sampling

Definition *In **cluster sampling**, the units of the population are grouped into clusters. One or more clusters are then selected at random. If a cluster is selected, that all units of that cluster are part of the sample.*

Example

Think about it

- Is a cluster sample a simple random sample?
- Is a cluster sample a stratified random sample?
- Were you to form clusters, how should the variability of the units within each cluster compare to the variability between the clusters?
- Is this criterion the same as in stratified random sampling?

Let's do it! 2.13

Identify the sampling method for each of the following scenarios:

1. A shipment of 1000 3 oz. bottles of cologne has arrived to a merchant. These bottles were shipped together in 50 boxes with 20 bottles in each box. Of the 50 boxes, 5 boxes were randomly selected. The average content for these 100 bottles was obtained.

2. A faculty member wishes to take a sample from the 1600 students in the school. Each student has an ID number. A list of ID numbers is available. The faculty member selects an ID number at random from the first 16 ID numbers in the list, and then every sixteenth number on the list from then on.

3. A faculty member wishes to take a sample from the 1600 students in the school. The faculty member decides to interview the first 100 students entering her class next Monday morning.

Multistage Sampling

Definition *M**ultistage sampling**, uses different sampling techniques at various stages to create a sample.*

Example *A college has* 15,000 *students. Additional information about the students at this college is as follows:*

	Women	*Men*	*Total*
Undergraduate	5,683	5,817	11,500
Graduate	2,117	1,383	3,500
Total	7,800	7,200	15,000

We are to obtain a sample of $n = 20$ *students from this population using each of the sampling methods discussed in this chapter. For each sample, the number of undergraduate women, graduate women, undergraduate men, and graduate men will be reported.*