AP Statistics

SAMPLING
CHAP 11

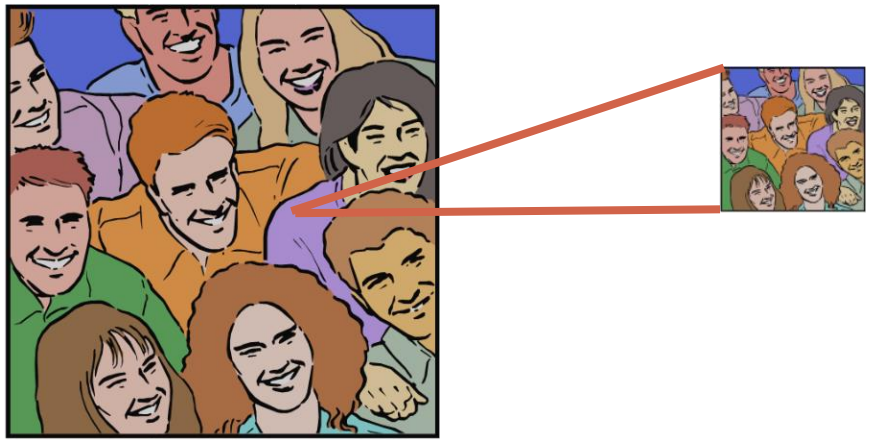
The idea that the examination of a relatively small number of randomly selected individuals can furnish dependable information about the characteristics of a vast unseen universe is an idea so powerful that only familiarity makes it cease to be exciting.

Helen Mary Walker (1891 - 1983)

What is Sampling?

We want to know something about a population (a mean or proportion, for example), but it isn't possible to go out and obtain the information from every member of the population (called a census).

So we take a "random sample" from the population, obtain information about the sample, and then use the sample information to estimate what we would get if we could reach the entire population.

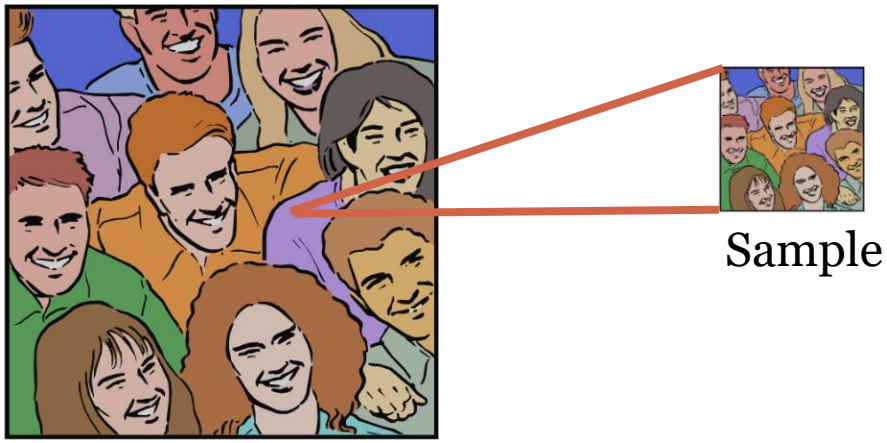


What is Sampling?

The information we want to know about the **population** is called a **parameter**.

The information we get from the **sample** is called a **statistic**.

We use the **statistic** to provide an **estimate** of the **parameter**.



Notation

Name	Statistic	Parameter
Mean	$\bar{y}$	μ (mu)
Standard Deviation	s	σ (sigma)
Correlation	r	ρ (rho)
Regression coefficient	b	β (beta)
Proportion	$\hat{p}$	π (pi -- but we will use p)

Surveys

A survey is a common way to get information from a population of people. Follow your favorite news source and you will hear nearly every day the phrase "In a recent survey …"

A survey generally consists of questions asked of a selected group (the sample) in order to obtain an estimate of the opinions of the population (U.S. voters, for example).

Sampling Variability

Different samples will produce different results. Statisticians refer to these different results as sampling variability (or sampling errors).

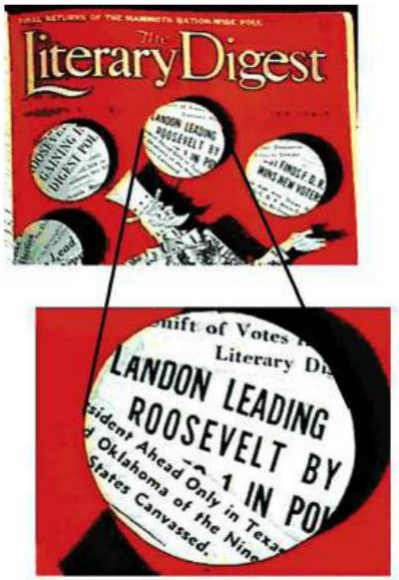
We will learn 2nd semester how to take just a single sample and estimate what the sampling variability would be if we could take more than one sample.

Bias

In an ideal world, the sample statistic will be very close to the actual population parameter we are trying to estimate. Unfortunately, the "Real World" is not ideal …

Bias

In 1936, incumbent President Franklin Roosevelt was running for re-election. A magazine, *The Literary Digest*, attempted to predict the outcome of the election by polling 10 million potential voters. They received answers from 2.4 million of those polled and predicted the challenger, Alf Landon, to win in a landslide (you studied "President" Alf Landon in your U.S. History class, didn't you????)



The questionnaires sent out by the Literary Digest went to people listed in telephone directories, motor vehicle registries, and country club memberships.

Remember the year is 1936. What problems do you see in the potential sample?

Bias

The sample obtained by the Literary Digest did not represent the population very well. Remember that in 1936, during the Great Depression, telephones[1], automobiles, and country club memberships were luxuries that few could afford.

This is a form of selection bias called **undercoverage** – the sample was selected from a subset of the population that did not allow for a "representative sample."

Meanwhile, George Gallup[2] queried a mere 50,000 U. S. voters obtained in a random manner from across the entire spectrum of voters and correctly predicted Roosevelt to win.

[1] *It wasn't until 1986 that enough homes had telephones to make this a viable method of surveying! And now, cell phones are making this method less effective.*

[2] *http://www.gallup.com/home.aspx*

Random Selection

Our best defense against selection bias is to obtain a random sample selected from a sampling frame (a list of the entire population).

In addition, random sampling will make possible the more formal methods of analysis we will learn next semester.

How Big a Sample?

Bigger is better, right?

Yes and no.

A large sample provides more information about the population than a small sample. But a large sample costs more to obtain.

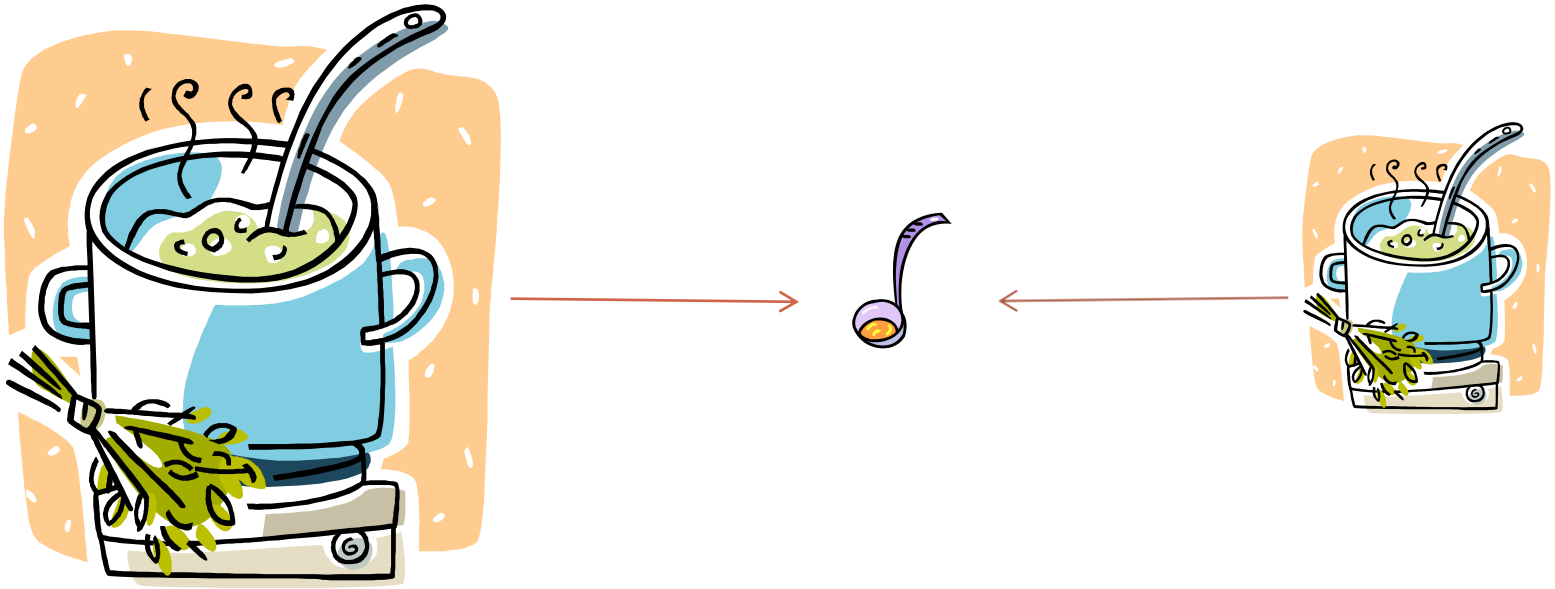
It turns out, a sample of say, 1000 residents of Denton will provide as much information about Denton as a sample of 1000 residents of the United States will provide about the United States!

It's *the size of the sample*, not the fraction of the population that is important. (More on this next semester.)

How Big a Sample?

It's like this …

If you are making soup, no matter how big the pot, a single spoonful is enough of a taste to know if the seasonings are to your liking.



Sampling Methods

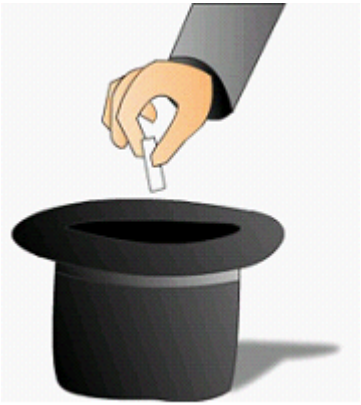
Census

- Measure or observe every member of a population.
- May be too expensive.
- May not be practical.
- The population is always changing (new items manufactured, births, deaths) so your count won't be exact.
- In the U.S. census, undercount is a problem. Sometimes entire groups of people are missed.

Sampling Methods

Simple Random Sample (SRS)

- Requires a sampling frame
- *Every sample of a given size has the same chance of selection.**
- A good mental picture is “draw from a hat.”

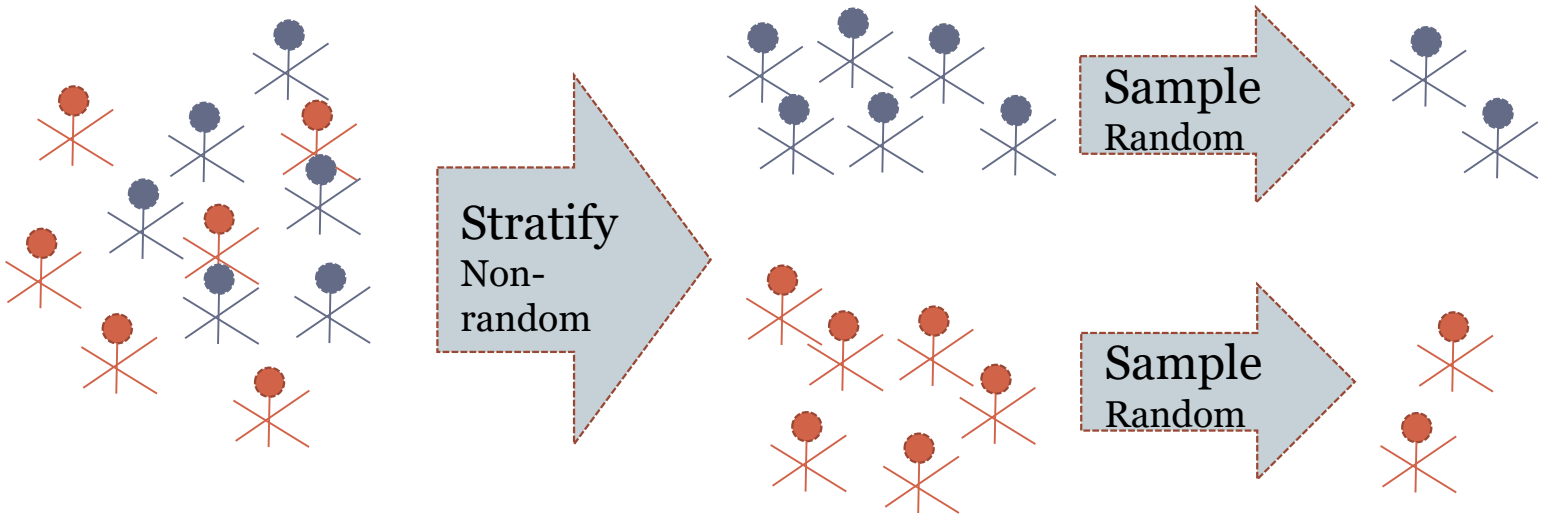


**This is important.*

Sampling Methods

Stratified Random Sample

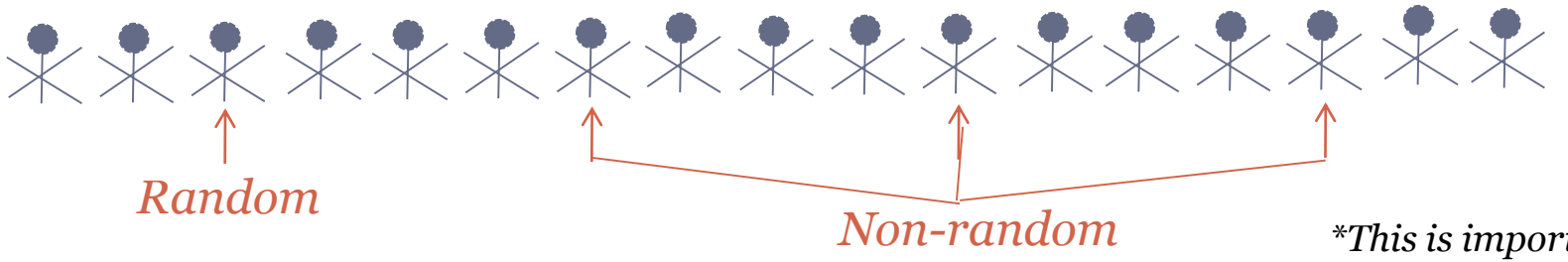
- Population is first organized into homogeneous groups called strata.
- Random sample then taken from each of the strata.
- Improves the "representativeness" of a sample when members of the population are recognized to belong to particular groups that are related to the variables of interest.



Sampling Methods

Systematic Random Sample

- It is not practical to randomly select individuals from a list.
- *The population is not organized in any particular manner with respect to the variable of interest.**
- When appropriate, this method of sampling can be cheap and easy to do.
- Imagine the population is "lined up."
  - Estimate the population size N.
  - If you want a sample of size n, randomly select a number from 1 to N/n and then take every (N/n)th member of the population.



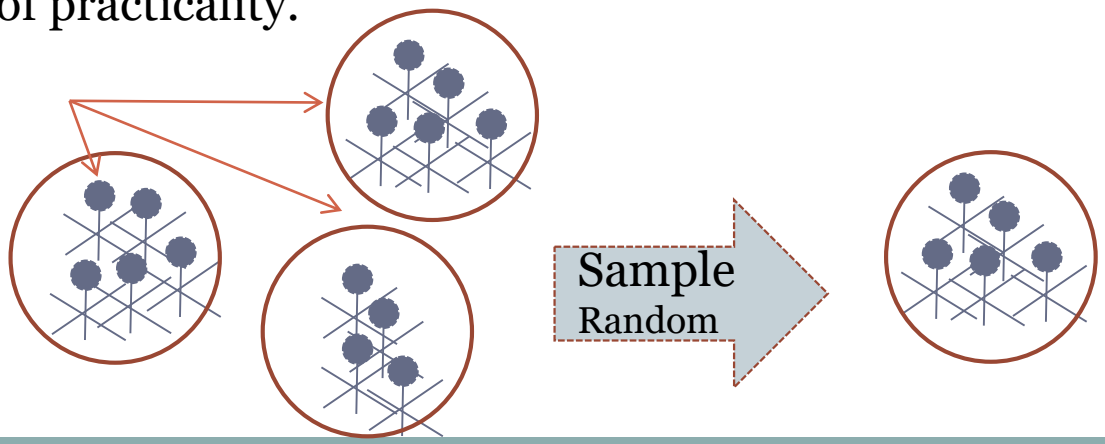
**This is important.*

Sampling Methods

Cluster Random Sample

- It is not practical to select individuals from a list.
- The population is divided into logical clusters (family groups, animal nests, classrooms, apartment buildings, etc.).
- Randomly select clusters and measure/observe every member of the cluster.
- Unlike homogeneous strata, clusters are roughly heterogeneous (i.e. each cluster more or less resembles the entire population). A cluster sample is a matter of practicality.

Non-random

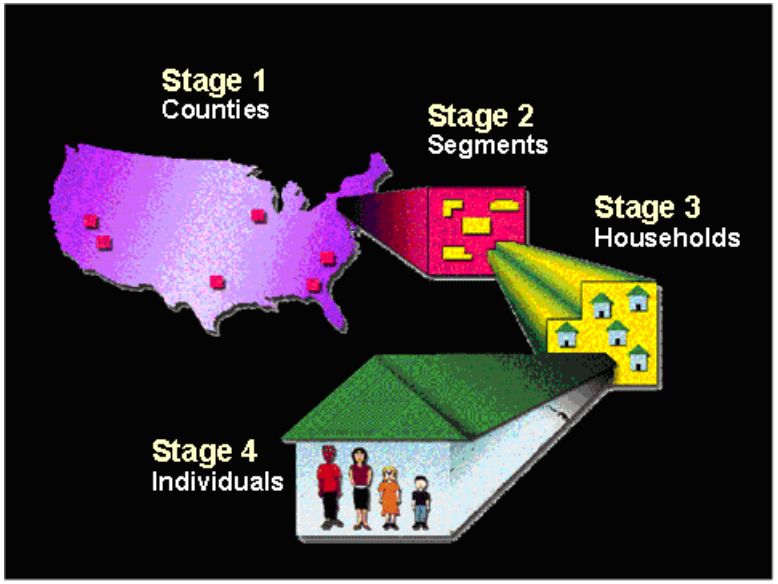


Sampling Methods

Multi-stage Random Sample

- Sampling methods may be repeated (stage 1, stage 2, etc.) or combined with other methods.

Example: The U.S. is divided into approximately 3000 counties that are labeled rural, suburban, and urban. A nation-wide sample might start with this stratification of counties to ensure that some of each type are selected. Then within each selected county, individual towns and cities can be selected (cluster sampling). Within each town or city, individual voting precincts can be selected (another cluster sampling), etc.



(Really) Bad Sampling

Voluntary Response Sample

- Ask people to participate
- Call in radio polls, website polls, etc.
- Tends to draw strong, non-representative opinions

(Really) Bad Sampling

Convenience Sample

- Sample obtained from easily observed members of the population.
- Sample not likely to be representative of the population of interest.
- Example: An opinion survey about shopping conducted at the local mall only reaches people who go to that mall.
- Example: A school newspaper wants to estimate the percent of seniors who are likely to go to college, so the reporter asks everyone in her AP classes about future plans.

(Really) Bad Sampling

Incomplete Sampling Frame

- A sampling frame that does not include all of the population of interest will lead to a non-representative sample.
- What do you think about the following sampling frames? Who might be left out, and why would it matter?
  - Telephone book.
  - Voter registration lists.
  - Vehicle registration lists.

Sources of Bias

Undercoverage Bias

- Certain groups within the population are underrepresented in the sample
- For example, a telephone survey conducted during the day will tend to miss people who work
- This is a problem if the people who work differ in an important way from people who don't work
- Pay close attention to the sample design to minimize this problem.

Sources of Bias

Nonresponse Bias

- Selected individuals are not available or choose not to respond
- This is important because the people who do not respond may have a different opinion or answer than the people who do respond.
- To encourage a better response:
  - Keep the survey short
  - Offer an incentive for participating

Sources of Bias

Response Bias

The questions being asked tend to lead towards particular answers because of the choice of wording or how the question is asked (if the survey is conducted through live interview):

- "In light of the problems in the current economy, are you opposed to raising property taxes to pay for new schools?"
- "Do you favor or oppose raising property taxes to pay for new schools?"

Sources of Bias

Response Bias

The order of the questions may influence the response:

- "Do you think a communist country like the Soviet Union should allow U.S. reporters to enter their country and freely report the news to the readers back home?"
- "Do you think the U.S. should allow reporters from a communist country like the Soviet Union to enter our country and freely report the news to the readers back home?"

When these questions were asked in a survey in the 1970's, the percent who answered the questions "Yes" varied depending on the order of the questions.

Example

The police set up a roadblock to check cars for up-to-date registration, insurance, and safety inspections. They stop every 10th car that passes.

Population	All cars in the jurisdiction of the police
Parameter	Proportion of cars with up-to-date registration, etc.
Sampling Frame	The cars on the road when they set up the roadblock.
Sample	Every 10th car that is stopped.
Sampling Method	Systematic Random Sample
Possible bias?	The time of day or location of the roadblock may not lead to a representative sample of cars. Otherwise, this is probably a pretty good method of collecting the data.

Assignment

Read Chapter 11

Exercises #1-11 odd, 17, 18, 23-29 odd, 31, 37

www.causeweb.org
John Landers