

AP Statistics – Lesson Notes - Chapter 12: Sample Surveys

Sample of a population

We have the ability to analyze data that has been collected. Often, this data is a **representative sample** of a larger **population**.

Population:

An entire group of individuals.

Sample:

A subset of the population.

Census:

Data for the entire population

Poll:

Data for a sample of the population.

Parameters:

examples : μ, σ, r, β, p

Statistics:

examples : x, s_x, r, b, \hat{p}

Sampling Frame:

List of individuals from which the sample is drawn.

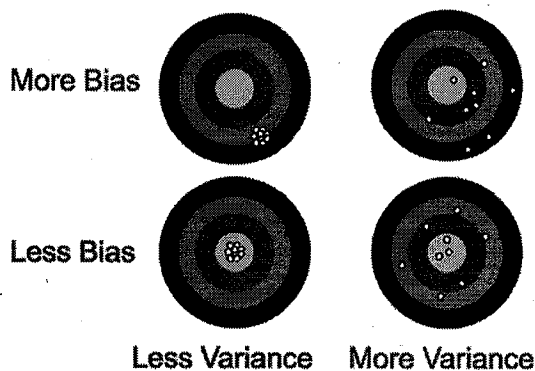
Idea 1: You can examine a sample, instead of the entire population

It is often impractical to examine an entire population. For example, imagine that you are making a pot of vegetable soup, and you want to taste the soup to make sure it is good. You wouldn't want to eat the entire pot of soup. It would be sufficient to taste a portion of the pot, as long as that portion (sample) is **representative** of the entire pot.

It is more difficult than it sounds to take a representative sample. The goal in selecting a sample is to avoid the two 'enemies' of statistical analysis: **bias** and **variability**:

Bias = The sample is not representative of the entire population you are studying and to which your conclusions will apply.

Variability = If multiple samples are taken, variability is the amount that some measured statistic varies from sample to sample.



Both bias and variability make it more difficult for statistical analysis to reach conclusions and for those conclusions to apply generally.

Kinds of Bias...

Undercoverage bias: Some portion of the population is not sampled at all.

Example: 1936 presidential campaign (Alf Landon, F.D.Roosevelt) a poll was conducted asking, "Who will you support for president?" Alf Landon projected the winner 57% to 43%. Actually, F.D.R. won 62% to 37%. The poll was conducted by telephone, but in 1936 only wealthy households had phones, so a large portion of the country was not sampled.

Response bias: Anything in the survey design which influences the responses of the subjects.

Examples:

- A survey may be designed which includes language or contextual references which are unfamiliar to some of the population.
- If a survey is conducted face-to-face and the interviewer is physically attractive, a participant may try to give the response they feel is what the interviewer 'wishes to hear'.
- If a participant's responses are not confidential, they may not respond honestly (due to fear of retaliation, for example from his/her boss).

Voluntary response bias: If sampling people, and you ask for volunteers to participate you may get more participation from subgroups.

Example: Ann Landers asks parents to write in an answer, "If you had it to do over again, would you have children?". 70% of 10,000 respondents said no. But a more careful verbal telephone survey showed that 90% of parents are happy with their decision to have children. When people needed to volunteer to write in, mostly parents with a particular point of view were motivated to respond.

Nonresponse bias: Usually, if people are polled, participation in the survey is not mandatory. Some people usually decide not to participate (nonresponse). But it may be that subsets of the population are more likely to be nonresponsive.

Voluntary Response Bias: Researchers don't choose the sample - subjects volunteer to be in the sample.

Nonresponse Bias: Researchers select the sample, but then some subject opt-out.

Judgment bias (not in book): If a sample is created by allowing the researchers to 'judge' who should and should not be included.

Example: A researchers including 2 people from each of many religions. (They will not be represented in the sample in the same proportions they appear in the population).

This last phrase is not really a bias in the same sense...

Confirmation bias (not in book): People evaluate all information based upon their current 'mindset' based upon previous experience. There are psychological reasons why humans tend to accept things that match their pre-conceived notions and reject as invalid things which do not.

...but is about how people are or are not convince by results.

Idea 2: Randomization produces the best sample

Most people think that if you are surveying people, you should try to select people from specific groups to make sure one of everything is included. But this approach does not produce a representative sample because the sizes of the subgroups are not equal (it tends to over-represent smaller subgroups). (In Chapter 11 we saw that our intuitions are not always correct).

In the soup analogy, if we want to taste a representative sample, we can just taste a spoonful, as long as any spoonful would be equally likely to contain all the components in the soup. If we, for example, just added salt to the top of the soup, or maybe another chef added peas that we weren't aware of and we didn't 'stir the soup' our spoonful might not contain a representative sample. But if we **randomize** (stir the soup) then it is likely that a single spoonful represents the entire population (including containing peas which we might not even have been aware were in the soup).

Idea 3: Sample size, not percentage of population, matter

It seems as though the larger the population, the larger the sample size needs to be to be representative (a percentage of the population), but it turns out **it isn't percentage but simply the size of the sample that is important.**

In our soup analogy, suppose instead of making a small pot of soup for your family you were making a giant pot of soup for a banquet. Would you need a gigantic spoon to get a huge sample to taste? No, in both cases the spoon only has to be large enough to get a representative sample (big enough to include the broth and typical numbers of vegetables).

In chapter 19, we'll learn more about how to determine appropriate sample sizes, but for now, just know that:

- 1) It is the sample size (not percentage of population size) that is important.
- 2) The size has to be large enough to be representative of the population.

Different types of sampling

1) Simple Random Sample (SRS)

- The entire population is available, individually, for selection.
- Select n individuals from the entire population using a random process.

Example: Number all the students in a school. Use `randInt(1,3100,50)` to randomly select 50 of these students.

For an SRS, it must be true that...

- Every individual has the same probability of being selected.
- Every combination of individuals has the same probability of being selected.

2) Random Sample

- Randomization is used in some way to select from the population.

Example: Flip a coin. If heads, use `randInt(1,#boys,50)` to randomly select 50 boys. If tails, use `randInt(1,#girls,50)` to randomly select 50 girls.

In a Random Sample...

- Every individual has the same probability of being selected.
- **Not** every combination of individuals has the same probability of being selected.

3) Cluster sampling

For large populations it may be difficult to 'number' all the individuals for random selection. We can divide the population in **clusters** which are each representative of the entire population, and then randomly select a cluster or clusters and **select all the elements in the selected clusters as the sample.**

Example: If we are collecting data on the number of words per sentence in a book, it is impractical to number each sentence and select sentence numbers at random. Instead, we can divide the book into pages (cluster = a page), randomly select a cluster (page) and then choose all the sentences on that page as the sample.

Note that even though we believe that each cluster is representative of the entire population, it is a subgroup (which could have unique properties, for example, perhaps sentence length is shorter in the beginning of the book and longer later in the book) so **cluster sampling is not an SRS of the population.**

The River Problem Activity

4) Stratified Random Sample

Sometimes, we suspect that there might be differences between subgroups in a population and we want to be able to highlight or analyze these differences.

We can divide the population into groups that are different in some way (gender, grade in school, amount of education, etc.) called **strata** and then use an SRS within each stratum before the results are combined.

This allows for analysis of each strata separately, or analysis of the entire population by combining the samples into a single sample for the population. Combining results into a sample to represent the entire population results in a sample which is guaranteed to have some representation from each stratum group but the result is **not an SRS**.

Our estimates are more precise / less variable if we use a well-chosen stratified sample.

Note: In the River Problem, sampling using a SRS is not incorrect, and does not produce biased results. But the SRS sampling method results in more **variability**. A stratified random sample is both unbiased and less variable, which improves the power of the statistical analysis to make conclusions (more about this in 2nd semester).

In the River Problem, we stratified on the variable 'distance from the river'.

Consider this question: There is a new proposal to double the amount taken out of each workers check to fund the medicare system (a system which pays for some of the medical expenses of people who have retired from working).

If we decided to do a Stratified Random Sample in order to make sure and include representation from important groups which might differ in some way, what variable should we stratify on?

5) Multistage sampling

Stratified, cluster, and SRS can be combined in **multistage sampling**.

Example: Number of words in sentences of a book

- **Stratify** a book into portions: 'beginning', 'middle', 'end'. Then randomly select a chapter within each stratum.
- **Cluster**: we could then divide each selected chapter from each stratum into pages (cluster = a page) and randomly select a page (cluster) for each stratum.
- **SRS**: finally, we could number and randomly select the sentences on the selected pages.

Example: Survey residents in the U.S.

- **Stratify** the country into geographic regions. Then randomly select a city or group of cities within each geographic region.
- **Cluster**: divide each city into regions (NW, SW, City center, NE, SE) and number all addresses in each cluster.
- **SRS**: finally, randomly select addresses by number in each cluster.

6) Systematic sampling

Using some systematic method, not randomness, to select items in a population.

Example: To select sentences to measure in a book, instead of numbering each sentence and selecting randomly, you could start with the first sentence and count sentence, systematically selecting every 100th sentence (selecting the 100th, 200th, 300th, etc. sentence).

Systematic sampling is not an SRS because not all samples have equal likelihood of being selected.

7) Convenience sampling

Selecting individuals in some way that is simply convenient to perform.

Example: Asking survey questions of the next 100 people who exit a grocery store.

Convenience sampling is not an SRS and is almost guaranteed to produce a biased result.

Example: Select a sample of students from a high school

Simple Random Sample

Slips of paper, each containing ID number of a student, for all students, into a box, randomly draw out students.

Things that could happen:

Every possible combination of students is equally likely to be selected.

Freshmen • •	Sophomores • • •	Juniors • •	Seniors •
Freshmen • • • •	Sophomores	Juniors • •	Seniors
Freshmen	Sophomores • • • • •	Juniors	Seniors

Cluster Sampling

Randomly select a class, and use the entire class as the sample.

Things that could happen:

Every student is equally likely to be selected, but you would only want to use cluster sampling if you believed there is no difference between the clusters.

Freshmen	Sophomores	Juniors	Seniors • • • • •
Freshmen • • • • •	Sophomores	Juniors	Seniors
Freshmen	Sophomores • • • • •	Juniors	Seniors

Stratified Random Sampling

Randomly select part of the sample from each class.

Things that could happen:

You are guaranteed to include members of each class. Especially useful if you believe there are differences between the strata.

Freshmen • •	Sophomores • •	Juniors • •	Seniors • •
-----------------	-------------------	----------------	----------------