BUS 232 chapter 3 - producing data

Summary

Usually we want to know something about the population. If we can't measure the population directly, we can use a sample to represent the population. This sample must be selected carefully to avoid bias and one way to do that is to use the Simple Random Sample. We can do the selecting of these with random numbers.

Terminology

Nonresponse:

- People who have something to hide
- People who don't like their privacy being invaded

Response bias:

- Lying when you think you shouldn't tell the truth
- People also may not remember

Undercoverage:

- Parts of the population are left out in the process of choosing the sample
- Not possible to get a certain group of people

Experimental units

- the individuals in an experiment
- if they are human we call them subjects

Treatment / factor

- In an experiment we do something to the subject and measure the response. The something we do is called a treatment or factor

Control

- a situation where no treatment is administered

Placebo

- a fake treatment, such as a suger pill

alternating allocation

- randomly choose one. give to group one.
- randomly choose another. give to group two.

block / stratefied design.

- subjects divided into groups prior to experiments to test hypotheses.
- eg) split into men and women. Because there may be differences

Ways to get data

Available Data

- library, internet
- our own and others experience
    - Anecdotal evidence

Producing Data

- Sampling
- Experiments

Population versus sample

Population

- the entire group of individuals in which we are interested.
- eg) **all** humans, **all** working-age people in california

Sample

- The part of the population we actually examine and for which we do have data.
- how good this is depends on the sample design.

Sampling

We try to make a reliable sample that is not biased to represent the population.

Simple random samples (SRS)

Take a samle of size *n* from a population in such a way that each group of size *n* is equally likely to be chosen.

To get the sample sizes we have to use the formula for counting combinations.
NCn

Stratified Sample

- divide population into similar groups called stratas
- take SRS from each strata
- eg) 100 students, 60 male, 40 female. Better to use stratified sample because we want proportionally as many male as female. We would take 6 males and 4 females.

Multistage Sampling

- sample in stages
- to sample the country, we can select a state (stage1) then select a county (stage2) then select a region (stage3)...

Technical issue: Selecting a random sample

1_ Use the random sample table. This is *table b*.
Or
2_ Use the computer as a random number generator.

1_
We need to know how many digits our ID number should have.
1 should use 2 digits (cannot have 1 digit, consider using 01 02 03 ...)
20 should use 2 digits
123 should use 3 digits

You have to make sure each number has the same number of digits:
001, 002, 003, ... , 500

eg) want to select 5 students from a class of 20.
Can start from any line.
In the example we started from row 103 for no particular reason.
Look at the row, then split each row into groups of digits of the same length as the number of digits you need.
If the number is larger than any ID number or the number has already been selected than we disregard it.
If you run out of the row, then go to the next one.
Go until you have 5 numbers

These numbers will give the id's that we will use.

Observation vs Experiment

observational studies based on sampling give **information** about a population but don't establish **causality**.

Properly designed experiments **can** establish **causality**

Comparative experiments

experiments are comparative in nature.

Completely randomized designs

Individuals are randomly assigned to groups, then the groups are randomly assigned to treatments.

- select 15 rats out of 30, this way you only need to look for 15 numbers

Bias and variability

Samples are not perfect information so there will be a percent error.

Bias and variability are not the same.
Bias is the difference to the actual value.

- problem with the sample

Variability is how different the outcome is.

- not enough data
- if sample size is small, then variability is high