1 Collecting Data

1.1 Planning a Study

In order to better understand the characteristics of a population, statisticians and researchers often use a sample from that population and make inferences based on the summary results from the sample.

A population is the entire group we want information from.

A sample is a part of the population we actually examine.

A census collects data from every individual in the population.

An observational study observes individuals and measures variables of interest but does not attempt to influence the responses.

An experiment deliberately imposes some treatment on individuals to measure their responses.

It is only appropriate to make generalizations about a population based on samples that are renadomly selected or otherwise representative of that population.

A convenience sample uses subjects that are readily avaliable.

A voluntary response sample is a sample obtained by allowing subjects to decide whether or not to respond.

A simple random sample consists of n individuals from the popultion chosen in such a way that every set of n individuals has an equal chance in the sample selected.

A stratified random sampling is when you divide the population into groups of similar individuals then select a SRS within each strata. Combine the SRSs from each strata to form your full sample.

Cluster sampling is dividing the population into sections (clusters) then randomly choose a few of these clusters. Every member of the cluster becomes your sample.

Systematic random samplling is one randomly selects an arbitrary starting point and then select every kth member of the population.

When an item from a population can only be selected once, this is called without replacement. When it can be selected more than once, it is called with replacement.

Samples are biased if they are systematically not representative of the desired population.

Voluntary response is when a sample is comprised entirely of volunteers or people who choose to participate, the sample will typically not be representative of the population.

Undercoverage occurs when some groups in the population are left out of the process of choosing a sample.

Non-response occurs when an individual chosen for a sample can't be contacted or refuses to respond.

Response bias is bias cuased by the behavior of the respondent or interviewer.

Untruthful answers occur when people give untruthful answers for several reasons.

Ignorance is when people will give silly answers just so that they appear to know something about the subject.

Lack of Memory is giving a wrong answer simply because the respondent doesn't remember the correct answer.

Timing is when a survey is taken can have an impact on answers.

Phrasing is subtle differences that can make a large difference in results.

When drawing a sample, two types of errors may occur:

Sampling Error: The difference between a sample result and the true population result. This error results from chance variation.

Non-sampling Error: Occurs when the sample data are incorrectly collected, recorded, or analyzed. Usually occurs when the sample is selected in a non-random fashion.

1.2 Selecting a Random Sample

The Hat Method:

- Write down all the names or numbers on their own slip of paper. Then put all the pieces of paper into a hat, mix well in-between selections, and pull out the desired number of slips.

Calculator Random Number Generator:

- MATH - PROB - 5: randInt(lower, upper, n)

Random Digit Table:

To choose SRS with a random digit table, you must label, identify how any digits you will take at a time, indicate when you want to stop sampling, and use the random numbers to identify subjects to be selected from your population.

An observational study observes individuals and measures variables of interest but does not attempt to influence the reponse.

An experiment deliberately imposes some treatment on individulas to measure their responses.

Experimental Unit: the things on which the experiment is done

Subjects: experimental units that are human beings

Treatment: a specific experimental condition applied to the units

1.3 Experimental Design

Factor: The explatory variables in an experiment

Level: the various groups the factors take

Principles of Experimental Design

1. Comparison - we need to make sure we are using a design that compares two or more treatments

2. Randomization - Randomization produces groups of experimental units we expect to be similar in all respects before treatment is applied.

3. Control - Control group is treated identically in all respects to the group receiving the treatment except that the members of the control group do not receive the treatment.

4. Replication - Use enough experimental units in each group so that any difference in the effects of the treatment can be distinguished from change differences between groups.

Experimental terms

- Placebo: a "dummy" treatment
- Placebo Effect: subjects receiving the placebo have a response that is similar to what we would expect if they received the treatment
- Single Blind: when the subject does not know what treatment they are receiving to remove the power of suggestion
- Double Blind: experiments in which the subject and administrator do not know who receives the treatment

In a poorly designed experiment, it might be difficult to tell if the explatory variable causes a change or if it was another variable that wasn't measured.

Confounding variables are variables that might affect the outcome, but we did not control or account for them in our experiment.

One way that we have seen already of removing the effect of any confounding variables is to randomly assign subjects to the treatment or control group. This way any possible bias in population should be evenly spread among the treatment and control groups. Sometimes instead of relying on randomization to make the groups as even as possible we actually force the groups to be similar.

An extraneous variable is one that is not an explanatory variable in the study but is thought to affect the response variable.

Confounding variable refers to another variable that may affect the response and is in some way tied together with the factor under investigation. It leaves us unable to tell which of the two variables caused the observed response.

Lurking variable refers to a variable that drives each of the two variables under investigation, making it appear there's some association between them.

Inference is drawing conclusions beyond the data at hand.

Random selections of individuals allows inference for the population and random assignment in an experiment allows inference for cause and effect.

Both random sampling and random assignment introduce chance variation into a statistical study.

To write an experiment:

1. Determine what type of design is best for your experiment.
2. Diagram your experiment.
3. Tell exactly how you will randomly assign variables.
4. Explain exactly what you are comparing once you gather the data.

Completely Randomized Designed:

- The experimental units are assigned to treatments completely by chance.
- Treatment groups and control groups will be about equal in size in a completely randomized design.
- There are mathematical reasons for having groups of equal sizes.

Randomized Block Design:

- When groups of experimental units are similar, it's often a good idea to gather them together into blocks.
- Blocking isolates the variability due to the differences between the blocks so that we can see the differences due to the treatments more clearly.
- When randomization occurs only within the blocks, we call the design a randomized block design.
- Control what you can, block on what you can't control, and randomize to create compatible groups.

Matched Pairs Design:

- These are experimental designs in which either the same individual or two matched individuals are assigned to receive the treatment and the control.
- Often the "pair" in a matched pairs design is just one experimental unit which serves as its own control.
- In the case where an individual receives both the treatment and the control, the order in which this happens should be random.