Chapter 2- Collecting Data Sensibly

Selin Kalaycıoğlu

Math 106-Spring 2010

February 10, 2010

Important Terms

Variable: A variable is any characteristic whose value may change from one individual to another.
Examples:

- Household income
- SAT score
- Height of a building
- Number of students in a class
- Time spent sleeping ...

Two type of statistical studies:

1. **Observational study**
   A study is observational if the investigator observes characteristics of a subset of one or more populations.
   Goal: To draw conclusions about the population or about the differences between populations.
2. **Experimental study**
   A study is experimental if the investigator observes how a response variable behaves when the researcher manipulates one or more explanatory variables, called factors.
   Goal: To determine the effect of the manipulated factors on the response variable.

The main difference

In an observational study it is impossoble to draw cause-and-effect conclusions because we cannot rule out the possibility that the observed effect is due to some variable other than the factor being studied. Such variables are called **confounding variables**.

Example: The American Heart Ass: Heart rate variability was higher for those who owned a dog than for those who did not.
Can we conclude that the heart rate variability can be explained by dog ownership?
What is a possible confounding variable?

Bias in Sampling-Types of Bias

- Selection bias
- Measurement or response bias
- Nonresponse bias

Selection Bias

Selection Bias is the tendency for samples to differ from the corresponding population as a result of systematic exclusion of some part of the population.

Example: Taking a sample of opinion in a community by selecting participants from phone numbers in the local phone book would systematically exclude people who choose to have unlisted numbers, people who do not have phones, and people who have moved into the community since the telephone directory was published.

Measurement or Response Bias

Tendency for samples to differ from the corresponding population because the method of observation tends to produce values that differ from the true value.

Example: Taking a sample of weights of a type of apple when the scale consistently gives a weigh that is 0.2 ounces high.

Nonresponse Bias

Tendency for samples to differ from the corresponding population because data are not obtained from all individuals selected for inclusion in the sample

Example: In a study that ask questions of a personal nature, many individuals that are selected might refuse to answer the survey questions. This occurs quite often when the questions are of a highly personal nature or when the individual feels that certain response might prove personally damaging.

Very important fact about Bias

Bias is introduced by the way in which a sample is selected so that increasing the size of the sample does nothing to reduce the bias.

Sampling Methods-Random Sampling

A **Simple Random Sample** (SRS) of size *n* is a sample that is selected in a way that ensures that every different possible sample of the desired size has the same chance of being selected.
A common method of selecting a random sample is to first create a list, called a **sampling frame** of the individuals in the population. Each item on the list can then be identified by a number, and a table random digits or a random number generator can be used to select the sample.

Example

We want SRS of 10 employees who work at a design company. For the sample to be an SRS each of the many subsets of 10 employees must be equally likely to be the one selected.
What about if we take a sample from only full-time employees?

Caution!!!

SRS $\rightarrow$ every individual has an equal chance of being selected.
However, the fact that every individual has an equal chance of selection is not enough to guarantee that the sample is an SRS.

Example: A class of 100 students: 60 Female and 40 male. Can a sample of 6 female and 4 male be regarded as an SRS?

Sampling with replacement

Sampling with replacement means that after each successive item is selected for the sample, the item is“replaced” back into the population and may therefore be selected again.

Example: Choose a sample of 5 digits by spinning a spinner and choosing the number where the pointer is directed.

Sampling without replacement

Sampling without replacement means that after an item is selected for the sample it is removed from the population and therefore cannot be selected again.

Example: A hand of “five card stud” poker is dealt from an ordinary deck of playing cards. Typically, once a card is dealt it is not possible for that card to appear again until the deck is reshuffled and dealt again.

Stratified Random Sampling

An entire population is divided into subpopulations called **strata**.
Stratified sampling entails selecting a separate simple random sample from each of the strata.

Example: Teachers in a large urban school district are given tenure by subject. The sample is taken by choosing random samples from each of the tenure areas.

Advantage: It allows us to make more accurate inferences about a population than does SRS.

Cluster Sampling

An entire population is divided into non-overlapping subgroups called **clusters**.
Cluster sampling entails selecting clusters at random and all individuals in the selected clusters are included in the sample.

Example: In a large university, a professor wanting to find out about student attitudes randomly selects a number of classes to survey and he includes all the students in those classes.

Systematic Sampling

Systematic sampling is a procedure that can be employed when it is possible to view the population of interest as consisting of a list or some other sequential arrangement. A value k is specified. The one of the first k individuals is selected at random, and then every kth individual in the sequence is selected to be included in the sample.

Example: In a large university, a professor wanting to select a sample of students to determine the student's age, might take the student directory and randomly choose one of the first 100 students and then take every 100th student from that point on.

Convenience Sampling—Do not go there

Convenience sampling is using an easily available or convenient group to form a sample.

Example: A “voluntary response sample” is often taken by television news programs. Viewers are encouraged to go to a website and “vote” yes or no on some issue. The commentator then would announce the results of the survey. It is highly unlikely that the responses would be accurately representative of the opinion of the public at large.

Class Discussion

Suppose a popular fashion magazine asked its married female readers to fill out a survey form in its September issue on the subject of fidelity. Suppose 10,000 married women responded, and of those, 48% reported having an affair. The blurb on the cover of the October issue states: “Nearly Half of the Married Woman Have Had an Affair!" In the local barber shop, Bob says “ Yep, I believe it,’" and decides to hire a private detective to follow his wife. Fred is not convinced of the validity of the results. What argument might Bob make to defend his position, and what argument might Fred make? Who do you believe?

Class Discussion

A Kenyon student developed a sampling scheme to study the drinking habits of the student population of Kenyon College. The student proposed that 10% of the students in each dorm be chosen at random, and each of the selected students answer an anonymous questionnare. The sampling scheme was approved by the student's advisor. As a research project, the student wished to conduct the same study at nearby Ohio State University, using the same sampling scheme. However this time the advisor rejected the proposed OSU study, citing a faulty sampling scheme. Can you think of reasons why the advisor might have rejected the OSU study?

Part II-Simple Comparative Experiments

We might have questions about the effect of certain explanatory variables on some response.
What happens when.....?
What is the effect of.....?

An **experiment** is a planned intervention undertaken to observe the effects of one or more explanatory variabled, called **factors**, on a response variable.
Example: Scores on the stats test vs. Room temperature

An **experimental unit** is the smallest unit to which a treatment is applied.

Blocking

An **extraneous factor** is one that is not of interest in the current study but is thought to affect the response variable (Some of them can be controlled directly).

Example: Textbook, instructor
A process known to filter out the effects of some extraneous factors. The factors that are addressed through blocking are called **blocking factors**.

Example: Instructor I- Section 1 in 65°, Section 2 in 75°
Instructor II- Section 3 in 65°, Section 4 in 75°

Two factors are confounded

Two factors are confounded if their effects on the response variable cannot be distinguished from one another.

Example: Instructor 1- both sections in 65°
Instructor 2- both sections in 75°

Randomization

So we can control instructor, textbook use via different methods.
How do we control for example student ability?
Such extraneous factors are handled by the use of random assignment to experimental groups- a process called **randomization**.

Replication

Replication is the design strategy of making multiple observations for each experimental condition, in other words each treatment is applied to more than one experiemntal unit.

Control group

When we are interested in determining whether some treatment has an effect, it is important to include an experimental group that does not receive the treatment. Such a group is called a **control group**.

A **placebo** is something identical to the treatment received by the treatment group, except that it contains no active ingredients.

Single-blind and double-blind experiment

An experiment in which subjects do not know what treatment they have received is described as **single-blind**.

To ensure that the person measuring the response does not let personal beliefs influence the way in which the response is recorded, the researchers should make sure that the measurer does not know which

A **double-blind** experiment is one in which neither the subjects nor the measurer know which treatment was received.