Author: Brenda Gunderson, Ph.D., 2015

Stat 250 Gunderson Lecture Notes
2: Sampling, Surveys and Gathering Useful Data

> Do not put faith in what statistics say until you have carefully considered what they do not say. -- *William W. Watt*

So far we have mainly studied how to summarize data - exploratory data analysis - with graphs and numbers. The knowledge of ***how the data were generated*** is one of the key ingredients for translating data intelligently. We will next discuss sampling, how to conduct surveys, how to make sure they are representative, and what can go wrong.

Collecting and Using Sample Data Wisely

There are two main types of statistical techniques that can be applied to data.

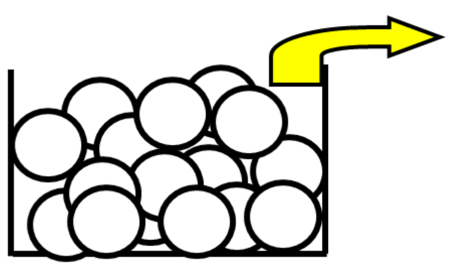
> ***Definitions:***
> **Descriptive Statistics:** Describing data using numerical summaries (such as the mean, IQR, etc.) and graphical summaries (such as histograms, bar charts, etc.).
>
> **Inferential Statistics:** Using sample information to make conclusions about a larger group of items/individuals than just those in the sample.

In most statistical studies, the objective is to use a small group of units (the sample) to make an inference (a decision or judgment) about a larger group (the population).

> ***Definitions:***
> **Population**: The entire group of items/individuals that we want information about, about which inferences are to be made.
> **Sample**: The smaller group, the part of the population we actually examine in order to gather information.
> **Variable**: The characteristic of the items or individuals that we want to learn about.

One way to view these terms is through a *Basket Model*:

Population=
basket of balls,
1 ball for each unit
in population.



Sample =
a few balls selected
from the basket.

***X* = variable (value of variable is**
recorded on each ball as small *x*)

Fundamental Rule for Using Data for Inference:
Available data can be used to make inferences about a much larger group *if the data can be considered to be representative with regard to the question(s) of interest.*

One principal way to guarantee that sample data represents a larger population is to use a (simple) random sample.

Try It! Fundamental Rule?

For each situation explain whether or not the Fundamental Rule holds.

a. **Research Question:** Do a majority of adults in state support lowering the drinking age to 19?
Available Data: Opinions on whether or not the legal drinking age should be lowered to 19 years old, collected from a random sample of 1000 adults in the state.

b. **Research Question:** Do a majority of adults in state support lowering the drinking age to 19?
Available Data: Opinions on whether or not the legal drinking age should be lowered to 19 years old, collected from a random sample of parents of high school students in the state.

c. **Available Data:** Pulse rates for smokers and nonsmokers in a large stats class at a major university.
Research Question: Do college-age smokers have higher pulse rates than college-age nonsmokers?

Sample versus Census

Why can't we learn about a population by just taking a **census** (measure every item in the population)? Takes too long, costs too much, measuring destroys the item. So, we often rely on a special type of statistical study called a **sample survey**, in which a subgroup of a large population is questioned on a set of topics.

Sample surveys are often used to estimate the proportion or percentage of people who have a certain trait or opinion. If you use proper methods to sample 1500 people from a population of many millions, you can almost certainly gauge the percentage of the entire population who have a certain trait or opinion to within 3%. The tricky part is that you have to use a proper sampling method.

Bias: How Surveys Can Go Wrong

While it is unlikely that the sample value will equal the population value exactly, we do want our surveys to be unbiased. Results based on a survey are **biased** if the method used to obtain those results would consistently produce values that are either too high or too low.

Selection bias occurs if the method for selecting the participants produces a sample that does not represent the population of interest.

Nonparticipation bias (nonresponse bias) occurs when a representative sample is chosen for a survey, but a subset cannot be contacted or does not respond.

Biased response or response bias occurs when participants respond differently from how they truly feel. The way questions are worded, the way the interviewer behaves, as well as many other factors might lead an individual to provide false information.

Try It! Type of Bias

Which type of bias **do you think would be introduced if …**

a. A magazine sends a survey to a random sample of its subscribers asking them if they would like the frequency of publication reduced from biweekly to monthly, or would prefer that it remain the same.

b. A random sample of registered voters is contacted by phone and asked whether or not they are going to vote in the upcoming presidential election.

Margin of Error, Confidence Intervals, and Sample Size

Sample surveys are often used to estimate the proportion or percentage of all people who have a certain trait or opinion (p). Newspapers and magazines routinely survey only one or two thousand people to determine public opinion on current topics of interest.

When a survey is used to find a proportion based on a sample ($\hat{p}$) of only a few thousand individuals, one question is ***how close that proportion comes to the truth for the entire population***. This measure of accuracy in sample surveys is a number called the **margin of error**.

The margin of error provides an upper limit on the amount by which the sample proportion $\hat{p}$ is expected to differ from the true population proportion p , and this upper limit holds for at least 95% of all random samples. To express results in terms of percents instead of proportions, simply multiply everything by 100.

Conservative (approximate 95%) Margin of Error = $1/\sqrt{n}$ **where *n* is the sample size.**

We will see where this formula for the conservative margin of error comes from when we discuss in more detail confidence intervals for a population proportion. For now we will consider an **approximate 95% confidence interval for a population proportion** to be given by:

Approximate 95% Confidence Interval for *p*:
sample proportion $\pm\ 1/\sqrt{n}$ or expressed as $\hat{p} \pm 1/\sqrt{n}$

Try It! School Quality

A survey of 1,250 adults was conducted to determine *How Americans Grade the School System*. One question: *In general, how would you rate the quality of American public schools?*

Frequency Distribution of School Quality Responses

Excellent	462
Pretty Good	288
Only Fair	225
Poor	225
Not Sure	50

a. What type of response variable is *school quality*?

b. What graph is appropriate to summarize the distribution of this variable?

c. What proportion of sampled adults rated the quality of public schools as excellent?

d. What is the conservative 95% margin of error for this survey?

e. Give an approximate 95% (conservative) confidence interval for the population proportion of all adults that rate the quality of public schools as excellent.

Interpretation Note:
Does the interval in part (e) of 34.2% to 39.8% actually contain the population proportion of all adults that rate the quality of public schools as excellent?

It either does or it doesn't, but we don't know because we don't know the value of the population proportion. (And if we did know the value of *p* then we would not have taken a sample of 1250 adults to try to estimate it).

The 95% confidence level tells us that in the long run, this procedure will produce intervals that contain the unknown population proportion *p* about 95% of the time.

f. ***Bonus #1:*** What (approximate) sample size would be necessary to have a (conservative 95%) margin of error of 2%?

g. ***Bonus #2:*** How does the margin of error for a sample of size 1000 from a population of 30,000 compare to the margin of error for a sample of size 1000 from a population of 100,000?

Sampling Methods

There are good sampling designs and poor ones.

- **Poor:** volunteer, self-selected, convenience samples, often biased in favor of some items over others.
- **Good:** involve random selection, giving all items a non-zero change of being selected.

Most of our inference methods require the data be considered a …

______________________ __________________________.

This implies that the responses are to be ***independent and identically distributed (iid)***. We will make this more formal later after probability, but here are the basic ideas between these two properties.

Independent =the response you will obtain from one individual
the response you will get from another individual.

Identically distributed = all of the responses ______________________________.

Many sampling designs are discussed in your text (SRS, stratified, cluster, etc). We will not cover the details of these various methods, nor work with a random number table. However, we will expect you to think about whether the data available can be considered a random sample, based on the fundamental rule for using data for inference.

We will also discuss various graphs that sometimes can be used for checking assumptions, one of which is a time plot for assessing the identically distributed property of a random sample (if the data are collected over time).

Difficulties and Disasters in Sampling

This section presents some of the problems that can arise even when a sampling plan has been well designed. It talks about sampling from the wrong population, relying on volunteer response, and meaningless polls.

How to Ask Survey Questions

The wording and presentation of questions can significantly influence the results of a survey. Here is one example of a pitfall that is a possible source of response bias in a survey.

Asking the Uninformed

People do not like to admit that they don't know what you are talking about when you ask them a question. Crossen (1994, p. 24) gives an example: "When the American Jewish Committee studied Americans' attitudes toward various ethnic groups, almost 30% of the respondents had an opinion about the fictional Wisians, rating them in social standing above a half-dozen other real groups, including Mexicans, Vietnamese and African blacks."

Try It! Consider the following two questions:

1. *Considering that research has shown that exposure to cigarette smoke is harmful, do you think smoking should be allowed in all public restaurants or not?*
2. *Considering it is not against the law to smoke, do you agree that smoking should be allowed in all public restaurants?"*

Here are the two results:

- *30% favored banning smoking*
- *70% favored banning smoking*

Which question (1 or 2) produced the 30%, which the 70%?
A more neutral and unbiased question might be:
Do you believe smoking should or should not be allowed in all public restaurants?

Types of Studies

Two Basic Types of Research Studies: Observational or Experimental

Definitions:
Observational Studies: The researchers simply observe or measure the participants (about opinions, behaviors, or outcomes) and do not assign any treatments or conditions. Participants are not asked to do anything differently.

Experiments: The researchers manipulate something and measure the effect of the manipulation on some outcome of interest. Often participants are ***randomly assigned*** to the various conditions or treatments.

Most studies, observational or experimental, are interested in learning of the effect of one variable (**explanatory variable**) on another variable (**response** or **outcome variable**).

A **confounding variable** is a variable that both affects the response variable and also is related to the explanatory variable. The effect of a confounding variable on the response variable cannot be separated from the effect of the explanatory variable.

Confounding variables are especially a problem in observational studies. Randomized experiments help control the influence of confounding variables.

Try It! Student's Health Study

A researcher at the University of Michigan believes that the number of times a student visits the Student Health Center (SHC) is strongly correlated with the student's type of diet and their amount of weekly exercise. The researcher selected a simple random sample of 100 students from a total of 3,568 students that visited SHC last month and first recorded the number of visits made to the SHC for each selected student over the previous 6 months. After recording the number of visits, he looked into their records and classified each student according to the type of diet (Home-Cooked/Fast Food) and the amount of exercise (None/Twice a Week/Everyday).

a. Is this an observational study or a randomized experiment?

b. What are the explanatory and response variables?

Try It! External Clues Study

A study examined how external clues influence student performance. Undergraduate students were randomly assigned to one of four different forms for their midterm exam. Form 1 was printed on blue paper and contained difficult questions, while Form 2 was also printed on blue paper but contained simple questions. Form 3 was printed on red paper, with difficult questions, and Form 4 was printed on red paper with simple questions. The researchers were interested in the impact that color and type of question had on exam score (out of 100 points).

a. This research is based on: ***an observational study*** ***a randomized experiment***

b. Complete the following statements by circling the appropriate answer.

   i. The color of the paper is a(n) ***response*** ***explanatory*** variable
   and its type is (circle one) ***categorical*** ***quantitative***.

   ii. The exam score is a(n) ***response*** ***explanatory*** variable
   and its type is (circle one) ***categorical*** ***quantitative***.

c. Fill in the blank. Suppose students in the "blue paper" group were mostly upper-classmen and the students in the "red paper" group were mostly first and second-year students.

   The variable "class rank" is an example of a(n) ___________________________ variable.

A Little More about Studies:
Hawthorne Effect, Placebo Effect, Randomized Studies, Control Groups, and Blinding

The Hawthorne Effect – In early studies from 1924-1932 at the Hawthorne Works (a Western Electric factory outside Chicago), investigators studied how various changes to the production process could increase production. In general, they observed that no matter what "production changes" were adapted, overall production levels increased. However, when the observations and recordings stopped, then production levels slumped back to what they had been before. Simply said, when someone observes and records a particular behavior, that behavior may improve during the observation period, but then return to usual behavior levels thereafter. To understand more about the phenomena called the Hawthorne effect see the first few pages of: http://en.wikipedia.org/wiki/Hawthorne_effect

The Placebo Effect – The placebo effect refers to the phenomenon in which some people experience some type of benefit after the administration of a placebo (a substance with no known medical benefit, e.g., a *sugar pill* or a saline solution). In short, a placebo is a fake treatment that in some cases can produce a real and positive response. For more info see: http://psychology.about.com/od/f/placebo-effect.htm

A Randomized Study (or Experiment) – These experiments involve the comparison of at least two treatments or methods (say Treatment A versus Treatment B). A group of study participants (or subjects) is randomized to receive either Treatment A or Treatment B using a "randomization schedule" which may involve a series of "random digits" or flips of a coin. The randomization is usually 1:1, that is, an equal number of subjects per treatment, or balanced; although some studies have been conducted using a 2:1 randomization where twice as many subjects are assigned to one treatment compared to the other. To learn more, see "Explorable Psychology Experiments" website: https://explorable.com/randomization

Blinding – In an experiment where Treatment A is compared to Treatment B, it is quite common to stipulate that the design be "single blinded", that is, the subjects are completely unaware of which treatment they are receiving. This blinding is found in pharmaceutical studies in which the pills or capsules *appear* exactly the same.

A study is said to be "double blinded" if not only are the subjects receiving the treatment 'blinded', but also the study personnel who recruit the subjects or who guide the subjects through the various procedures are also "blinded" as to actual treatment the subject received. This is especially true for the study personnel who gather and record the data, especially the measurements regarding how each subject is responding to treatment. Such study personnel having knowledge of which patients are given the various treatments has the potential to bias the various efficacy measurements.

Pharmaceutical companies quite often insist on a "triple-blind" study design in which personnel at the company itself also remain unaware to the treatment assignment of the subjects until all the data has been obtained and *cleaned* (carefully examined to insure the consistency and correctness of each value in the database).

Blinding can help reduce the potential for bias in studies. https://explorable.com/randomization

A Placebo-Controlled Study – Studies that compare the response of an experimental treatment with a placebo are called placebo-controlled studies.

An Active-Controlled Study – An active control is a treatment that has already been shown to be an efficacious product by several previous investigations and is so recognized by the medical community. Studies that compare the response of an experimental treatment with an active-control are called active-controlled studies. For more information Trial Design in http://en.wikipedia.org/wiki/clinical-trial

Where are we going?

- We have a population (a basket) that we cannot examine but we want to learn something about it - so we will take a sample - preferably it will be a random sample.
- We will use the sample to *estimate* the things we wanted to know about the population - we will use the sample results to test theories about the population and make some decisions.
- Since the sample is just a part of the population there will be some uncertainty about the estimates and decisions we make. To measure and quantify that uncertainty we turn to **PROBABILITY**!

Additional Notes

A place to … jot down questions you may have and ask during office hours, take a few extra notes, write out an extra problem or summary completed in lecture, create your own summary about these concepts.