1/21/2023

Introduction to the course
Definitions, Variables
Sampling Methods
Experimental Design

What is statistics?
The study of data. We want to understand something about the data we collected, and ideally, infer information about data we have not collected.

Descriptive Statistics: describes the data we have. Numerical calculations (eg. Average), proportions, make graphs, etc.

Inferential Statistics: We want to use the data we have to infer information (approximately) about a larger population that we can't directly measure.

Population vs. Sample

The population is the whole group about which we want to know something. The sample is a subset of the population that (hopefully) is representative of that population.

Samples have statistics. Statistics infer to a parameter.

We use the sample statistics to infer the population parameter.

Census is a sample that is equivalent to the whole population. (Hardly ever work with these.)

What is data?

Data is a measure of some aspect of a sample or population. The answer to a question about a sample.
Variables: the responses to a particular question.
Variables have more than one possible response.

Probability: what does it mean to be fair? What proportion of the population has a particular characteristic?

Types of Variables

Categorical Data (Qualitative Data) vs. Numerical (Quantitative Data)

Numerical Data is a number… and it makes sense to calculate an average.
A count or a numerical measurement.

How many kids are living in the household? A count.
How tall are you? A measurement.

Not all numbers are numerical (quantitative) data.

Pain scale: even though the values are from 1-10, standing in for verbal descriptions of pain
Employee number: replace names with numbers for anonymity
Credit card number, football jersey numbers

Categorical is everything else: most of the time these are responses in words

What is your favorite color? What state were you born in?

Within numerical data we can have discrete or continuous data.
Discrete data has a finite number of distinct values. (it can't be a decimal or a fraction)
Continuous data can take one (almost) any value in a range.

Discrete data is typically a count. Continuous data has lots of values and rarely repeats.

Level of Measurement

4 levels of measure:
Nominal: the name of a thing (what is your favorite color? Jersey numbers) -- categorical
Ordinal: things that are ordered (pain scale, or agreement scale) -- categorical
Interval: (not ratio) temperature or GPA -- numerical
Ratio: has a real zero (height) and scale proportionately: ratios are real -- numerical

Sampling Methods

Simple Random Sample (SRS) – everyone in the population has an equal chance of being selected for the sample

Stratified Sample – separate the population into groups (demographic groups usually), and then take a simple random sample within each group (strata), small number of large groups

Cluster Sampling – large number of small groups, and you sample from the clusters (select a cluster or group of clusters as your sample), everyone in the selected cluster is in the sample.

Systematic Sampling – survey every nth person. (in a line)

Probability sampling: weight the subgroups to reflect the weight in the general population.

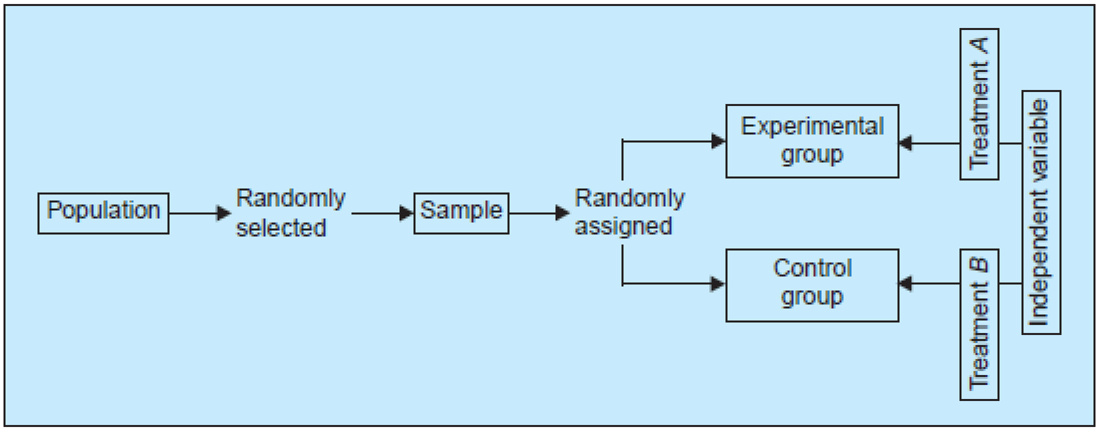
Convenience Sample: this is a bad sampling method! WEIRD

Biases and Errors
Sampling and non-sampling errors
Sampling errors are errors in sampling (the sample is not representative…)
Over- or underrepresentation, non-response, voluntary response
Non-sampling errors: typos, mismeasuring

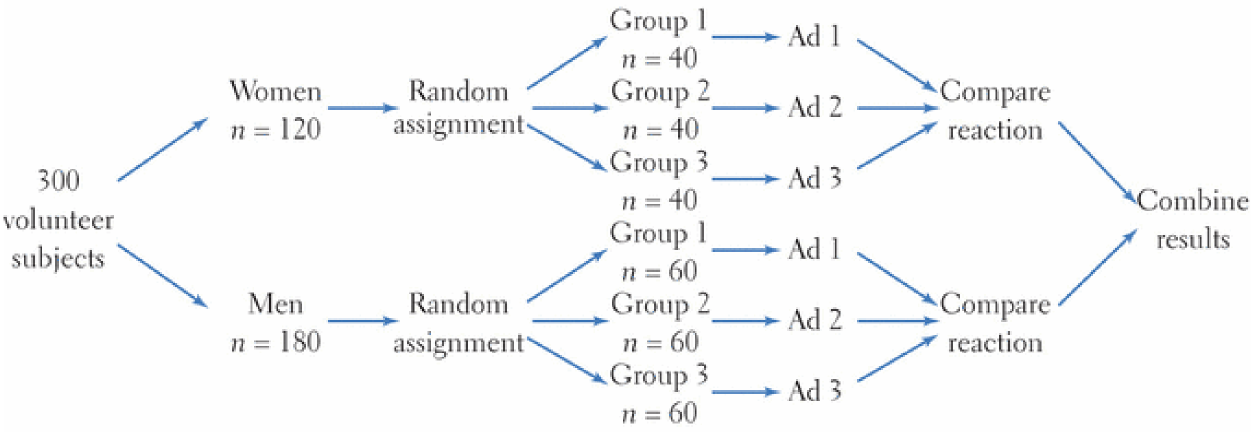
When we do inference, the errors and biases are not accounted for.

Pg 20 of the textbook (Critical Evaluation)

Experimental Design:
Understand causation
We have an explanatory variable (input), and a response variable (output)



Block design



Experiments require us to consider ethics
Institutional Review Boards
Informed consent
Blinding, double-blinding