Lecture 1

Introduction to the course

Basic Terms

Population/parameter vs. sample/statistic

- A population is a group or set of objects or people (or phenomena) about which we want to understand something. A population may be too large to make a measurement on feasibly. Populations can be somewhat abstract. They are always bigger than samples.
- A parameter is a measurement of that population (a mean, a proportion, a variance, etc.)
- A sample is a subset of a population, that is hopefully representative, on which we are able to take a measurement
- A statistic is an estimate of a parameter taken by making the same measurement on the sample.

If the sample is representative of the population, then the statistic will be a good estimate of the parameter (most of the time).

Conceptual or hypothetical populations may be descriptions of populations that don't really exist at the moment. Such as all future customers, or yet-to-be-manufactured widgets at a factory.

Data is the collective term for many observations. The "singular" is "datum" for a single observation. In the past, data was treated as a plural, and still is in some texts, but it can also be treated as a mass noun, like water, and take a singular verb.

Descriptive statistics vs. inferential statistics

Descriptive statistics – organizing and summarizing data (a dataset)

Inferential statistics – formal methods that use probability in combination with descriptive statistics to understand a population

This course is primarily frequentist in design. Another type of statistics is Bayesian statistics. We'll discuss that approach briefly at the end of the course.

Sampling methods

Simple random sample -- Any group of n individuals is equally likely to be chosen as any other group of n individuals if the simple random sampling technique is used

cluster sampling -- divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. The number of groups tends to be large. Often used to reduce cost to due to the geographic spread of the population.

stratified sampling -- divide the population into groups called strata and then take a proportionate number from each stratum (the number of groups tends to be small). Ethnic groups or genders are common strata. Used to ensure that people from each strata are included in the sample, to make it more likely to be representative.

systematic sampling -- randomly select a starting point and take every nth piece of data from a listing of the population. Commonly used to sample people in line, such as in a grocery queue or cars crossing a border.

convenience samples -- involves using results that are readily available. Good example: talk about the WEIRD acronym.

modern technologically derived samples

census – when the sample is equal to the population – no randomness, measure everything/everyone

Biases and errors

Sampling bias – created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen).

Non-sampling error -- A defective counting device can cause a non-sampling error.

Non-response – an error that occurs when people selected for a study refuse to respond. Can produce non-representative results.

voluntary response – people with strong opinions more likely to reply and therefore are not representative of the population, sometimes called a selection error

measurement error – a malfunctioning device, or a user who does not know how to use the device

data entry errors

biased survey questions – a type of measurement error

biased processing/decision making

inappropriate analysis conclusions (including ignoring data that doesn't fit preconceptions)

false information provided by respondents – sometimes subjects lie about things they are embarrassed about, they consider too personal, may be illegal; or they may forget things, or have an unrealistically optimistic view of their own future behavior

processing errors

coverage error (undercoverage) – underrepresentation of a group in a sample of interest

Sampling errors can also occur such as selecting from the wrong sampling frame or population.

Common problems in studies and statistics to look for:

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions
- Undue influence: collecting data or asking questions in a way that influences the response
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context

- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor

Variables:

Qualitative vs. quantitative

- Qualitative variables are variables that take on values that are not numbers (words usually), sometimes these are called categorical variables since the levels of the variables are categories. Examples are things like state names (place of birth), or level of agreement to a statement (strongly agree, etc.), and so forth. Sometimes numbers can be categorical if they stand in place of names or words and it doesn't make sense to calculate averages on them, such as credit card numbers or identification numbers.
- Quantitative variables are numerical values. Things like counts (how many children does a family have), or measurements (how tall are you?). It does make sense to calculate averages on such numbers.

level of measurement

- Nominal – these are categorical values that do not come in any fixed order – states where a person was born, or favorite color are some examples. Nominal means "name".
- Ordinal – these are usually categorical variables that come in a fixed order (could be reversed, but cannot be randomly organized). These include things like letter grades (A, B, C, etc.) or level of agreement to a statement (strongly agree, agree,…, strongly disagree, etc.). Sometimes these words can be replaced by numbers in an analysis, but the numbers don't represent a quantity, only the ordering. "ordinal" just means "ordered".
- Interval – these are numerical variables that do not have a true zero (the location of zero is arbitrary). The best way to identify interval values is to eliminate them from the fourth level: the ratio level. Examples are GPA and temperature.
- Ratio – these are numerical variables that do have a true zero, and where ratios of observations produce meaningful interpretations. Height or the number of children in a family are both ratio level variables. Someone who has 6 children has twice the number of children as a family with three children (6/3=2). A building that is 40 feet tall is 4 times taller than a building that is only 10 feet high (40/10=4). Temperature (Fahrenheit or Celsius at least) is not ratio because 60 degrees is not "twice as warm" as 30 degrees.

discrete/continuous

- These categories apply only to numerical values
- Discrete values usually only take on integer values
- Continuous variables can take on any value on the number line (including decimals).
- In practice, since we round everything, in some sense all observations are discrete, but think about the number of possible values we can get even then. Are there some values that we cannot observe? Then it's probably discrete. For example, the number of children in a family can't be anything but a whole number. You can't have half a child. Income, even if rounded to a whole number, can take on a much larger range of values, and rounding to a whole number is a choice: we could choose to round to a penny instead. This is continuous.

The type of variable we have determines which kinds of graphs we can make, how those graphs look, and some aspects of interpreting our data.

Categorize the following variables:

a. the number of pairs of shoes you own
b. the type of car you drive
c. the distance it is from your home to the nearest grocery store
d. the number of classes you take per school year.
e. the type of calculator you use
f. weights of sumo wrestlers
g. number of correct answers on a quiz
h. IQ scores

We will be looking at univariate, bivariate and multivariate data

Experimental design – to establish causation

Explanatory variable – the independent variable, the variable that can be obtained first or more easily, the variable that (might) be the causal variable
Response variable – the dependent variable, the variable that is harder to measure, that (might) be caused, the variable we want to be able to predict

The different values of the explanatory variable (especially if categorical or discrete) may be referred to as treatments.

Lurking variables -- generally, you want to try to control for these. They also affect the value of the response variable and may make measuring the response to the treatments more difficult to observe.

Placebo/nocebo

Experiments have a control group to which the experimental group can be compared. Comparable groups are established by random assignment (or sometimes in pairs of similar subjects). The control group is given a placebo, an inert treatment that measures what happens if you do nothing. The placebo exists because of the placebo effect – some people will feel better anyway but may attribute feeling better to something that has no causal effect. However, if you give them nothing, they may continue to report feeling bad because they believe they are not being treated. A nocebo does the opposite – it is an inert treatment that actually makes people feel worse (psychologically).

(double) blind

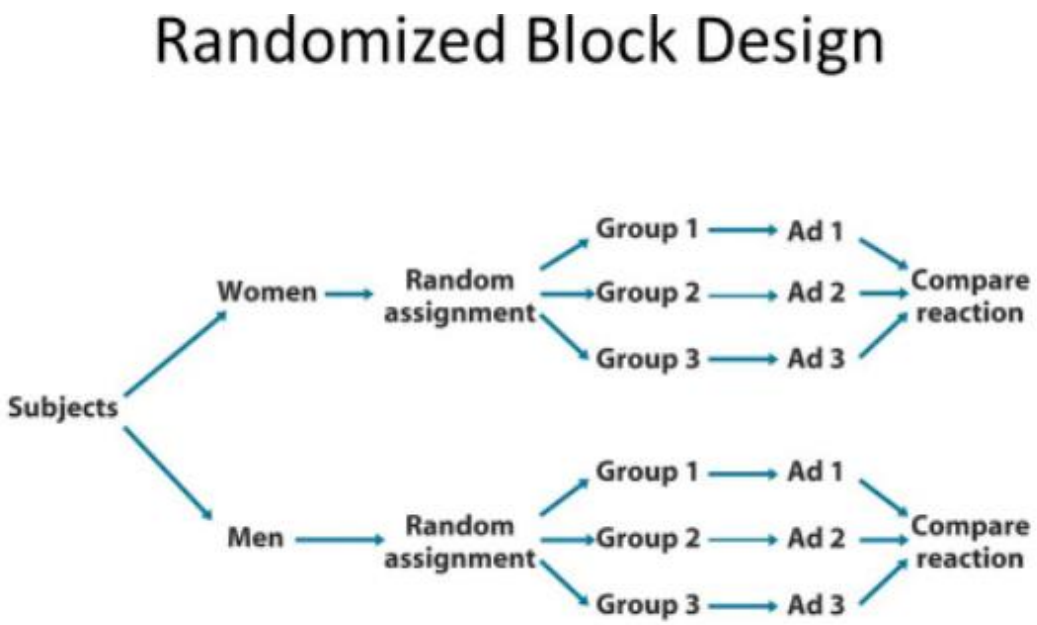
A study is blind if the patient receiving the treatment does not know what group (treatment or control) they are in.

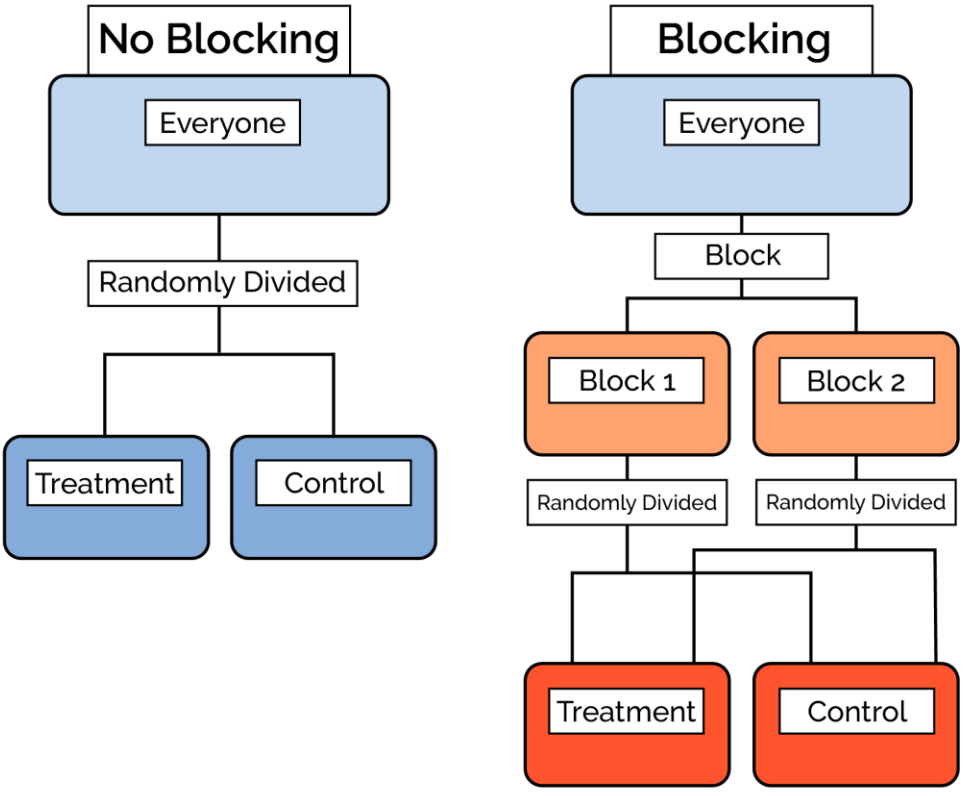
A study is double blind if the person giving the treatment to the patient also does not know which treatment their patient is receiving. To avoid body language subtly giving away the game and creating a psychological impact on results.

block design

Similar to a stratified sampling approach. In order to study a treatment on different demographic groups. The patients in the study are separated into groups and then within those groups, randomly assigned to treatments, to ensure equal numbers of each group are in both

the treatment and control groups, and to allow for study of how the treatment impacts the groups in possibly different ways as well as overall.





Ethical issues

Researchers can get into serious trouble by falsifying data:

- creating datasets, which largely confirmed the prior expectations
- altering data in existing datasets
- changing measuring instruments without reporting the change
- misrepresenting the number of experimental subjects.

But these are not the only ethical issues in statistical research
Institutional Review Boards oversee research to prevent ethical lapses

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy

References:

1. https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf
2. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
3. https://www.qualtrics.com/experience-management/research/sampling-errors/
4. https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch6/nse-endae/5214806-eng.htm
5. https://discovery.cs.illinois.edu/learn/Basics-of-Data-Science-with-Python/Experimental-Design-and-Blocking/
6. https://introductorystats.wordpress.com/2011/03/09/design-of-experiments/