3/22/2022

1.1 Definitions
1.2 Data and Sampling
1.3 Frequency, Tables, Levels of Measurement
1.4 Experimental Design and Ethics

Statistics
Descriptive statistics vs. inferential statistics

Descriptive statistics numerical characterizations of data
Summary of the data = descriptive statistic

Inferential statistics take a sample of all possible observations and infer information about observations we are not able to make.

Probability
Fairness: what do we mean when we talk about a fair coin? Or a fair die?

Samples and descriptions of data

Population vs. sample

Population is the entire group of objects about which we want to know something. For example: All Americans' view of the economy, All American women, All children age 6, All volcanoes, etc.

Sample is a subset of the population. The goal in sampling is to make that sample as representative of the whole as possible.

The Law of Large Numbers: the bigger sample the closer to the true (population) value that the sample will produce.

Parameter vs. statistic

A parameter is a measure of the whole population (ex. The true mean height of all American women)

A statistic is an estimate of a parameter based on a sample.

As we collect data, we collect different types of data depending on what we want to measure.

There is qualitative (or categorical) data. There is quantitative (or numerical) data.

Qualitative data is generally written in words.
What is your favorite color?

Quantitative data is generally written in numbers: it measures something.
How tall are you?

Sometimes data forms can be misleading.
Qualitative data that is written as a number: does it make sense to average this number?
A football jersey number.
A credit card number.

Categorical variables, their main descriptive statistic is a proportion.
What proportion of the sample had red as their favorite color?

Quantitative (numerical) variables, their main descriptive statistic is a mean (average).
What is the mean test score for the class? What is the mean height of American women?

Variable: is just used for any quality of an observation.
Anything that has more than one possible response.

Within the quantitative variables, there are two subgroups:
Discrete, and the continuous kind.

Discrete variable is one that has integer outcomes only (fixed responses that are separated numerically).
Counting variables. It's a whole number. Number of children. Age (usually).

Continuous variables can take on any value in a range.
Height. Salary.

Think about the main difference as the number of possible values.

Level of Measurement

4 levels of measurement:
Nominal (name/noun) -- inherently unordered – your favorite color, name-replacement numbers (like jerseys or social security numbers, because they are stand-ins for names)
Ordinal (ordered) – are categorical (not numbered usually) but do have an inherent order – the strength of agreement with a statement: strongly disagree, disagree, neutral, agree, strongly agree

Interval (it has no natural zero, ratios are meaningless) – numerical variable that is "arbitrary", like temperature (F/C), GPA
Ratio (here there is a natural zero and ratios are meaningful) – height, distance, time

Time between volcano eruptions:
Is this quantitative or qualitative? Quantitative
Is this discrete or continuous? Continuous
What is the level of measurement? Ratio

The kind of lava coming out of the volcano:
Is this quantitative or qualitative? Qualitative
What is the level of measurement? Nominal

Do you approve of the state of the economy? Yes/No?: I like it, I am neutral, I hate
Is it quantitative or qualitative? Qualitative

What is the level of measurement? Nominal or Ordinal depending on the response options.

Sampling Methods:

Most common and simplest sampling method is called a simple random sample.
Make a list of the population that you want to sample from, and then randomly select people from the list.
This is usually executed in the real world through telephone lists.

The list of the population is called a sampling frame. The sampling frame for a telephone survey is just a list of all the phone numbers assigned in the United States.

A stratified sample. Very common when subsets of the population are also of interest, not just the whole. The population is divided into groups (usually demographic), the number of groups is usually relatively small. The groups are called the "strata" = layer. And then a simple random is selected from within each group.

This ensures that the demographic make-up of the whole sample is representative. It also allows for the subgroups to be analyzed separately.

Cluster sampling. Very common when dealing with geographic regions. Divides the population into many small groups (like towns). Then they sample from the groups (randomly select groups). Then survey everyone within the group.

Systematic sampling. Every kth person is selected from the sampling frame. Only the first person is randomized. These are often done in lines of people: maybe in a grocery store checkout, or at the border going through TSA.

These are all considered good sampling methods (the four above) that can get you representative samples.

There are bad ways to sample.
Convenience sample: is just surveying people that are accessible (convenient) for the surveyor.
Psychology 101 profs sampling their students for psychological studies.
The sample may not be representative of what you are making inferences about.

WEIRD – Western Educated Industrialized Rich and Democratic

Experiments and Ethics

Experiments are determining the relationship between an explanatory variable and a response variable (input and output). Input might be a drug for a medical experiment, the output might be time to cure.

Trying to establish a cause-and-effect relationship.

Input is referred as a treatment.

Deals with eliminating the effects of lurking or confounding variables (usually done with randomization).

Experiments have a control that receives placebo (this is some an inert treatment that doesn't actually do anything). This is to control for the effects of the drug over the effects of people thinking they are taking a beneficial drug.

Studies are blind or double blind to prevent study participants from receiving unconscious clues from the people delivering the treatment. Blind is when the patient doesn't know, but the doctor giving the medication does. Double blind is when neither the patient nor the doctor knows.

Ethics in research studies are controlled by Institutional Review Boards.
Makes sure that no harm is done to anyone participating in the study.
Informed consent.

Next time, we talk about frequency tables. And we'll look at statistical graphs.