STATS 210P

Lecture 1

Sevan Koko Gulesserian

University of California, Irvine

Statistical Modeling

Examples:

- Do students with higher GPA have a better chance of getting into medical school?
- Do financial incentives help volunteers lose weight?
- How dosage of a medicine (in milligrams) correspond to an increase in a wellness factor?

Statistical Modeling

Statistical vs. deterministic models.

- Statistical models account for variation in the response given an explanatory variable. This is unlike a deterministic mathematical model.
- Example of a deterministic relationship: In Physics, force of an object in motion is equal to it's mass times acceleration.
- Two objects with the same mass and the same acceleration will have exactly the same force.
- Example of a statistical relationship: We want to see how a person's weight, height, age, and sodium intake help determine their blood pressure (systolic).
- It can be the case where two people have the same exact weight, height, age, and sodium intake but different blood pressure measurements.

Some Definitions

Statistical terminology.

- *Population*: All of the individual units about which we want information. For example all diabetic patients in the USA, all birds migrating in the winter, all 4 door sedans in use in Europe, etc.
    - *Parameter*: A quantity, usually unknown, that applies to the whole population. Example: the average blood pressure of a diabetic patient in the US or the fuel efficiency (in kilometer per liter) of a 4 door sedan in Europe.

Some Definitions

- *Sample*: Units for which we obtain data on. E.g. 100 diabetic patients sampled from the US, 15000 birds sampled during migration periods in winter, 250 4 door sedans in Europe sampled based on registration records.
  - *Statistic*: A quantity computed from a sample. E.g. the average blood pressure of 100 diabetic patients sampled in the US.
  - The idea is that the sample is a piece of the population (where the population may be impossible or infeasible to sample all of it). Then will use the statistic to obtain inference on the true parameter.
- *Variable*: Something we measure (for a sample) or could possibly measure (for a population) on each unit.

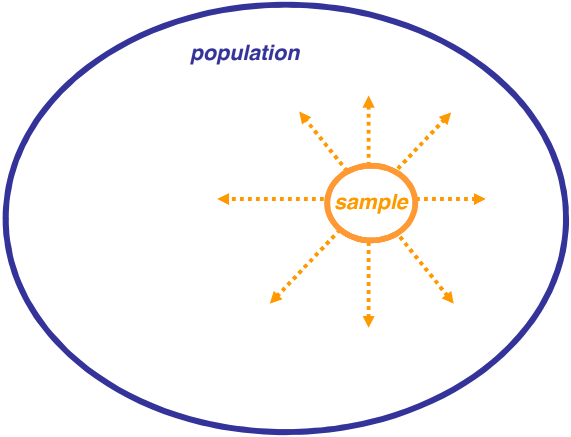
Population and Sample

Population
(data for all individuals)

→ parameter

Sample
(data for some individuals)

→ statistic



Variables

Two types of variables with respect to how they contain information.

- *Categorical*: Data consists of category names.
    - Examples are what political party you affiliate with (Democratic, Republican, or Independent). Or what blood type you are (A,B, AB, or O).
- *Quantitative*: Data consists of numerical values (discrete or continuous) where ordinary arithmetic makes sense.
    - Examples are how tall you are (in inches, continous). Or how many days a week you go to campus (0-7 days, discrete).

Variables

Within quantitative variables, there are two kinds: *continuous* and *discrete*

- Continuous quantitative variables are ones whose measured value can be any real number in a given range.
  - Height can be any positive number. E.g. 67.5 inches, 70 inches, 58.245 inches.
- Discrete quantitive variables are ones whose measured value must be a whole integer value.
  - How many days you go to campus must be a whole number, like 3 days or 5 days. Cannot have values such as 5.25 days.

Variables

With respect to the role a variable plays in a model, there are a few kinds of variables.

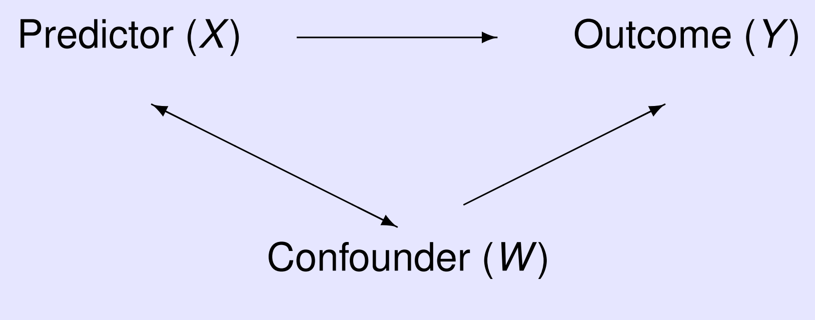
- *Response* (or *dependent* or *outcome*) variable.
    - The variable of interest that we are trying to model. Referring to slide 2, the response variables would be being accepted to med school (yes/no) and how much weight someone loses.
- *Explanatory* (or *independent* or *predictor* or *covariate*) variable.
    - The variable (or variables) that may explain or may cause differences in the response variable (or outcome or dependent variable). Referring to slide 2, the explanatory variables are GPA and amount of financial incentive.

Variables

- Within explanatory variables, we also have what is called a *confounding* variable.
  - Associated with the response variable and is also related to the explanatory variable.
  - It could be what is driving the association between the response and explanatory variables.
  - Referring to the GPA example on slide 2, a possible confounding variable can be the level of extracurricular activities a student participates in.
  - This variable is related to the explanatory variable of GPA (usually students with high levels of extracurricular activities have high GPAs) and is associated with the response variable (medical schools tend to accept students with high levels of extracurricular activities).

Variables

Confounding variable.



Here, the true association (or possibly causal) pathway is X to W to Y, not just X to Y.
Thus if we want to modify or influence what the value of Y could be, then we would need to adjust or tune what W is (not what X is).

Variables

Confounding variable example.

Before there was widespread knowledge of the harmful effects of cigarettes, a study was conducted where it was found that the higher the consumption of mints and gums was directly related to the higher rates of lung cancer.

Does this imply that the more mints and gums someone consumes, the higher the chances of lung cancer?

Think of a confounder being the amount of cigarettes smoked daily by the subject.

The Two Types of Statistics

Two types of statistics.

- *Descriptive* statistics.
  - Using numerical and graphical summaries to characterize a data set (and only that data set).
- *Inferential* statistics
  - Using sample information to make conclusions about a population.
  - *Model*: Used to approximate the population relationship between two (or more) variables.

Definitions of Types of Studies

Two types of studies.

- *Observational* study.
  - Researchers observe or question participants about opinions, behaviors, or outcomes.
  - Participants not asked to do anything different than what they would have done normally.
  - Referring to GPA example, this is an observational study. Cannot assign students to have a certain GPA.

Definitions of Types of Studies

- *Experimental* study.
    - Researchers manipulate something and measure the effect of the manipulation on some outcome of interest.
    - In a *randomized experiment*, participants are randomly assigned to participate in one condition (called treatment) or another.
    - Referring to weight loss example, this is an experimental study. Subjects are assigned to receive different levels of financial incentives.

Definitions of Types of Studies

- Extending results to a population:
  - This can be done if the data are representative of a larger population for the question of interest.
  - Safest to use a random sample (that is to say you randomly select units from the population for the sample).
- Cause and effect conclusion:
  - Can only be made if data are from a randomized experiment, not from an observational study.

Definitions of Types of Studies

The reason we can make a cause-and-effect relationship only in an experimental study is because of the potential for the existence of a confounding variable. The potential confounders are not adjusted for in an observational study.

- In an observational study, the relationship between the response variable and the explanatory variables could be because of a confounder.
- This would mean the association between the explanatory variable and the response variable is *spurious*.
- In an experimental study, by randomly assigning the experimental units (i.e. subjects) to the different levels of the explanatory variable, the issue of potential confounding should be negated.

Definitions of Types of Studies

Confounder example:

The goal of the study is to investigate if having a high fiber diet will lower the number of colon polyps someone develops (colon polyps are growths on the colon that are generally accepted to be unhealthy as they increase the risk of colon cancer).

Definitions of Types of Studies

Confounder example:

- In an observational study, a group of people will be followed through time and have their daily fiber intake measured, and how many colon polyps they develop.
- It is very likely we will see a negative association between fiber intake and number of polyps (higher the fiber, the lower the number of polyps).
- It could very well be that the subjects that have a high fiber diet also have a healthier lifestyle in general (more active, better diet, etc.).
- Thus, it is not possible to separate the effect of these possible confounders and the effect of the high fiber diet on the number of polyps being developed.

Definitions of Types of Studies

- In an experimental study, each subject in the study will be randomly assigned how much fiber they should intake daily (say we have 3 levels: low, medium, and high).
- The effect of the confounders should be mitigated because we will now have a variety of people in each of the 3 fiber groups because of the random assignment.
    - Will have people with good and bad health habits in each of the fiber groups.
    - This will allow the researcher to isolate the effect of the treatment (fiber level) on the risk of developing colon polyps.

Definitions of Types of Studies

Confounder example:

A study might show a negative correlation between exercise levels and stress. But, sleep patterns could act as a confounder, as individuals who exercise more might also have better sleep, which in turn could lower stress levels.

Confounder example:

Research indicates that students who attend classes regularly have better grades. However, a student intrinsic motivation to learn could be a confounding variable, as these students might not only attend class but also study more outside of class.

Definitions of Types of Studies

Referring to slide 2.

- Medical school study is an observational study based on a random sample
  - We cannot make a cause and effect conclusion.
  - We can extend results to the population.
- Weight loss study is a randomized experiment but used volunteers.
  - We can make a cause and effect conclusion.
  - We cannot extend results to the population of all people trying to lose weight, only to those similar to the volunteers.

Definitions of Types of Studies

> "Children exposed to lead are more likely to suffer tooth decay..."
> USA Today

- Observational study involving 24,901 children.
- Explanatory variable: Level of lead exposure.
- Response variable: Extent child has missing/decayed teeth.
- Possible confounding variables: Income level, diet, time since last dental visit.

This study is an observational study. We cannot conclude that lead exposure causes tooth decay.

Definitions of Types of Studies

"After the 8 week period of patch use, 46% of the nicotine group had quite smoking, while 20% of the placebo group (no nicotine patch) had quit smoking." Newsweek

- Randomized experiment, 240 smokers recruited
- Explanatory variable: Nicotine patch yes/no.
- Response variable: Quit smoking yes/no
- Each subject randomized to nicotine patch or placebo (equal chance of each).

This study is a randomized experiment.
We can conclude that nicotine patches cause people to quit smoking. Potential confounding variables should be similar in the placebo and nicotine patch groups because of random assignment.

Summary of Studies

- Observational study: Data are recorded without manipulating or influencing any of the variables.
- Statistical experiment: One or more of the explanatory variables are assigned/controlled for all experimental units.

Should use an experiment if the goal is to confirm a cause and effect relationship.
Cannot conclude a cause and effect relationship from an observation study.