1. Role of statistical models

Statistical models

Statistical model

- Model is by definition a simplification of (a complex) reality.
- Possible uses of a statistical model (not mutually exclusive, from easy to hard):
  - Description. Ex: Describe how income depends on years of schooling, race, gender, region of residence.
  - Prediction. Ex: Predict the chance that a released convict will be rearrested, based on age, gender, nr of previous arrests, type of crime for which imprisoned.
  - Causal analysis: Ex: Does participation of a prisoner in an educational program lower the risk of being rearrested?
- In all of the above, we also want to know the precision of the estimates.

3 / 22

Statistical model

- In all cases, we examine the relation between a single *dependent variable* Y and one or more *independent variables* $X_1, \ldots, X_k$.
- Identify dependent and independent variables in the examples on the previous slide.
- Other names for dependent variable: response, outcome
- Other names for independent variables: predictor variables, explanatory variables, regressor variables, covariates, covariables

4 / 22

Populations and samples

Goal: to know a parameter of a population

- We often want to know a *parameter* of a *population*. Examples:
  - average income of people in Switzerland
  - average increase in income with every year of additional education for people in Switzerland
- It is infeasible to contact everybody and ask about their income.
- So we will never know the population parameters exactly.

Solution: use a sample

- Solution: use a sample
  - We collect data on a random sample of people.
  - We use the average income in the sample to *estimate* the average income in the population.
  - The estimate is random: taking a new sample would lead to a different estimate.
  - Estimate = population parameter + random error.
  - In order to draw conclusions from our estimate about the population parameter, we need to know properties of the estimator:
    - How large is the error we can expect?
    - How does the error depend on the sample size?
  - Therefore we will spend a large part of this class studying the distribution of regression estimates

An ideal sampling method

- Identify population
- List all individuals in the population
- Draw random sample with a probability method (meaning that you know the probability for each person to be included in the sample)
- Then the results of the sample are generalizable to the population

Sampling in reality

- Example:
  - We want to test efficacy of two different teaching methods.
  - We randomize the students in a certain high school class to either method.
  - We find that method A is significantly better.
  - You teach at another high school. Do you switch to method A?
    - Technically we cannot generalize beyond the specific class at that specific high school.
    - But if your class at the other high school is 'similar', it is reasonable to assume that the results will hold there as well. So then we would switch.

Sampling in reality

- Example:
  - A medical study wants to test efficacy of a drug
  - They ask for volunteers, and randomize these to receiving the drug or a placebo
  - Study finds a significant difference between the two groups
  - What should the authorities decide?
    - Volunteers may be different from general population
    - Compare several characteristics of the study group to the general population to check this
    - If they seem pretty similar, approve drug for population
    - If they seem very different, approve drug for subgroup, or do further study

10 / 22

Convenience samples and generalization

- Convenience samples are often used. Ex: students at nearby school, patients at specific hospital.
- We often want to generalize beyond the population from which we sampled.
  - This is reasonable if the population from which you sampled is similar to the population to which you want to generalize.
- In this class, we always assume that we have a representative sample from the population, with each person having equal probability of being in the sample.

11 / 22

Confounding

Example: prison data

- Does participation in educational program lower the chance of getting rearrested?
- Dependent variable: getting rearrested (1=yes, 0=no). Independent variable: participation (1=yes, 0=no).
- Example data:

	participated	did not participate
rearrested	10	50
not rearrested	40	50
total	50	100

- Among the people who participated, 20% were rearrested.
  Among the people who did not participate, 50% were rearrested.

What does this mean?

- Does participation in educational program lower the chance of getting rearrested?
- It depends on the study:
  - If the prisoners decided whether or not to participate in the study - no.
    - Difference can be due to the fact that people who choose to participate are systematically different from those who choose not to do so.
    - Think of: types of crime committed, motivation for reintegration in society, etc.
  - If the prisoners where randomly assigned to participate or not - probably yes. But not absolutely certain:
    - For example, it may be that the guards behaved differently towards the two groups.

Observational vs. experimental study

- Key difference:
  - Observational study: the subjects decide about treatment assignment (ex: smokers vs. non-smokers, diet choices)
  - Experimental study: the investigators decide about treatment assignment (ex: many medical studies)
- See overhead about different types of studies

Solutions in observational study

- Compare subgroups that are similar except for the factor you are interested in. Example:
  - Compare motivated prisoners who participated to motivated prisoners who did not participate
  - Compare non-motivated prisoners who participated to non-motivated prisoners who did not participate
- This is called *controlling for* the factor motivation.
- In regression, we can control for a factor by putting it in the model. (We will come back to this.)
- Problem: We can never be sure that we controlled for every possible relevant factor.
- But this is not enough to discredit every observational study. To discredit such a study, you need to argue persuasively that a specific factor could cause the pattern.

Solutions in experimental study

- Make sure that treatment assignment is done at random
- Use blinding if possible:
  - blinding of participants
  - blinding of evaluators/investigators

17 / 22

Confounding factor

- A factor such as motivation in the prisoners example is called a *confounding factor*.
- Definition:
  - the factor *influences* the dependent variable/outcome
  - *and* the factor is *related* to the independent variables that are the focus of the study
- If both conditions are met, then the effect of the confounding factor and the independent variables of interest are confounded (mixed up). We cannot determine anymore what causes the effect.
- See plant example on overhead

18 / 22

Back to prisoner's example

- Prisoner example:
  - Motivation influences chance of getting rearrested
  - Motivation is related to participation in the educational program (the people who participate are more motivated).
- So:
  - The group of prisoners who are highly motivated and participated in the program are rarely rearrested.
  - The group of prisoners who are non-motivated and did not participate in the program are often rearrested.
- We don't know whether the difference in rearrest rate is caused by motivation or by participation in the program. These effects are confounded = mixed up.

Example: Canadian refugees

- Do judges decide similarly about refugees' requests for leave?
- Canadian refugee data (Fox, Table 1.1, page 8)

Judge	Leave granted	Leave not granted
Pratte	9%	91%
Desjardins	49%	51%

- These data became the basis for a court case contesting the fairness of the Canadian refugee determination process.
- Dependent variable: leave granted (yes/no).
  Independent variable: judge (Pratte/Desjardin).

Is gender a confounding factor?

- Scenario 1: Judges are more likely to grant leave to women, and Desjardins had a higher proportion of women applicants.
- Scenario 2: Judges are more likely to grant leave to women, and both judges had about the same proportion of women applicants.
- Scenario 3: Gender of the applicant does not influence the decisions of the judges, and Desjardins had a higher proportion of women applicants.

Randomized experiments

- Confounding factors are not a problem in randomized experiments
- Why?
- Independent variable indicates the treatment group. By randomizing, the treatment groups will be about the same in all respects. Hence, the second condition of the definition of a confounding factor is never met.
- So we would always like to do a randomized experiment.
- But this is not always possible or moral.
  Examples: cigarette smoking, climate change.