Lecture 6

Types of research questions

inference/ association

- What is the relationship between the explanatory variables and the response?
- Do we have evidence for a relationship between these variables?
- How well can we predict the response?
- I have a lot of variables available. Which ones should I focus on?

prediction

model selection

Prediction vs. inference

Question: how might your model choices differ when your goal is *prediction* (predicting the response) vs. *inference/association* (modeling and testing the relationship with particular explanatory variables)?

Prediction: ROC, AUC, sensitivity, specificity etc.

Inference: group means, effectsize, test statistics, p-values

Prediction vs. inference

Prediction:

- Care about *predictive ability* of model
- Often less interested in model interpretation
- Model selection useful
- Assumptions less important

Prediction vs. inference

Inference/association:

- Generally not good to test hypotheses after performing model selection
- Models and hypothesis tests should address specific research questions
- Valid inference requires assumptions
- Variables of interest have to be in the model!

Strengths and limitations of model selection

Situations in which model selection is appropriate:

- You care more about prediction than inference
- You are doing a preliminary/exploratory study to identify potentially important variables
- Your research question does not concern specific explanatory variables

Strengths and limitations of model selection

Problems with model selection:

- Resulting model might not be interpretable
- Model selection does not fix violations of assumptions
- Do not do inference with the same data used for model selection

Developing a statistical analysis plan

1. What are the research questions/objectives?
2. What are the variables?
3. What is the study design / how was the data collected?
4. How will you explore/summarize the data?
5. What will be examined statistically?
6. What alternative strategies should be considered?
7. What statistical results will be presented?

Class activity

https://sta712-f23.github.io/class_activities/ca_lecture_6.html

Class activity: The variables

What variables should the researchers use to investigate the research question? How are those variables measured?

well being:
- PA
- Happiness
- SCS

etc.

Ill being:
- NA
- Loneliness

etc.

Class activity: The study design

What information is recorded for each individual in the study? What are the three treatment groups, and how was treatment assigned?

- Treatment group (randomly assigned)
- Demographics (age, # pets, year in school)
- outcome variables

Class activity: The study design

The researchers randomly assigned participants to the three treatment groups. The benefit of random assignment is that we no longer need to worry about confounding variables, because no explanatory variable can be systematically associated with the treatment. So why do the researchers collect demographic information about their participants, and compare the demographics for the three groups in Table 1?

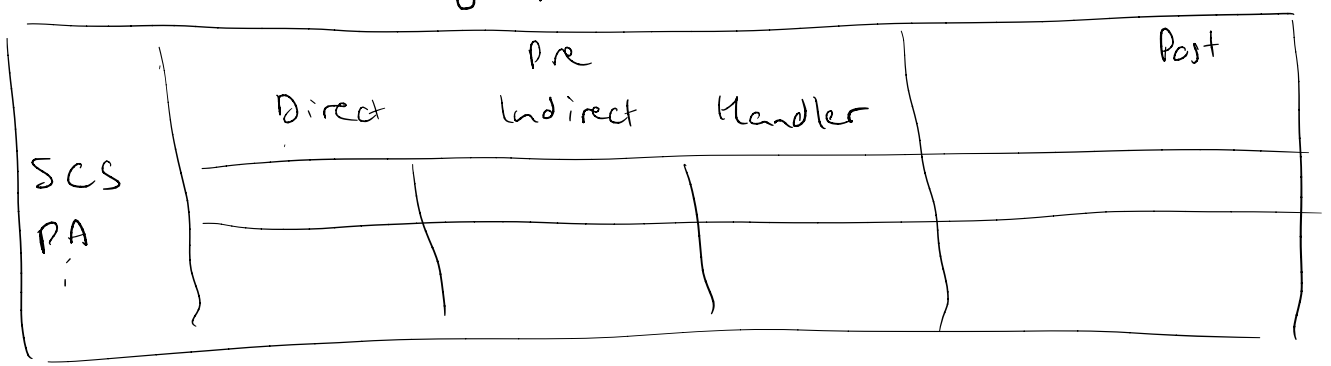
- describe the study population
- testing for balance between treatment group

Class activity: Data exploration

How could you summarize the data, and any relevant relationships between variables?

Here are some options:

- mean & standard deviation of each outcome, for each treatment group:



- Pre/post difference in means

Class activity: Statistical analysis

What statistical method(s) will you use to address the research question?

To address Hypothesis 1, could use paired-sample t-tests

Class activity: Alternative strategies

Are there any assumptions we need to check for the statistical methods we have chosen? What will we do if those assumptions are violated?

t - tests assume that a t distribution is appropriate for the test statistic. We're looking at the average change in wellbeing / illbeing for each treatment grap, so a t - test is reasonable given a sufficiently large sample size.

Otherwise, nonparametric tests cald be considered as an alternative

Class activity: Statistical results

What statistical results will be presented? (e.g. p-values, confidence intervals, test statistics, etc.)

① average change for each outcome, for each treatment

② test statistics for testing a null hypothesis of no difference, for each comparison in ①

③ p-values for the tests in ②