Notes | Design

For later in the semester

- Choose set of comparable texts for text analysis.
- Should be short and comparable, like bios, obits, 10k reports, plot summaries, etc

We talk about significance, probabilities, validity issues, research designs.

Significance

Hypothesis testing. Whether research is explicitly hypothesis testing or exploratory, we can think of it as coming down to examining the association between X and Y.

What is the concept of significance and p-values? what is inferential stats? We find patterns everywhere, but we to need guard against chance.

permutation test
- see excel file. Permutations go to permutations tab
- suppose in test, every boy scored better than every girl. Does gender affect knowledge?
- how many ways could this come out? 24 of these 17% have this property. so could easily happen by chance.
- this is a p-value
- difficult to calculate with more than 10 observations
- not really helpful for generalizing

classical hypothesis testing
- we have a population, such adult americans, and we are interested whether men or women know more. so we sample 100 people and look at the correlation between gender and knowledge. suppose the correlation is .327. it looks like gender affects knowledge.
- but what if the correlation in the population is zero, and we just happened to get a weird sample, where the men were knowledgeable than the women.
- classical stats proceeds by calculating the likelihood of getting a sample in which the observed statistic is as large .327, even though it is zero in the popular.
- to make tractable we have to assume random sampling or more generally a probability sample where we know the relative prob of each person being in the sample. many other assumptions as well. homoskedasticity, bivariate normality, linearity etc. When all the conditions are met, we

can mathematically derive the distribution of correlation coefficients across all possible samples, can then calculate just how rare a correlation of .327 should be, if x and y are uncorrelated in the pop.
- so the p-value says how likely your data are, given a particular model, which is the null hypothesis
- this as opposed to bayesian inference, where you work out the probabilities of a hypothesis given the data

p-hacking and harking

See harking handout

Collider variables

See collider handout

Validity

Inferential and construct validity. Studies and measurements

- Translation and criterion
- Internal and external

Consider types of studies in this context

- / stealing fire

Measurement: Distinguishing between reliability and validity

See the handouts

- Types of validity
- Validity and reliability

Types of research design

Elicit names:

experiments, quasi-experiments, natural experiments, field studies, observational studies, case-control studies, cohort studies, survey studies, ethnographies (part obs), pure observational, archival studies, longitudinal, cross-sectional, simulations, meta-analyses, mixed methods, qual, quant, interventions, exploratory and hypothesis testing

Elicit dimensions:
- manipulation. all studies X -> Y. some manipulate X and some don't. experiments do obs don;t
- temporal. cross/long.
- natural setting vs lab.
- continuous vs categorical X. irrelevant confound with experiments vs observational
- primary vs secondary data collection
- prospective vs retrospective

Some type of studies:

Experimental Studies, where the X variable is manipulated.

- RCT randomized control trials:**
    - Gold standard of causality-determining studies
    - treatment and control groups (X could also be continuous)
    - randomized assignment to treatment groups
        - Ensures equivalence of groups, trying to create counterfactual (where we see what would have happened if we hadn't applied the treatment)
        - no endogeneity issues -- the X variable has no (natural) cause. Purely determined by the experimenter
    - Usually double-blinded so neither experimenter nor subject knows which group they are in
- Quasi-Experiments:**
    - Similar to controlled experiments, but lacking random assignment. These studies are often used when randomization is impractical.

Observational/Field Studies where X is observed rather than manipulated

- survey studies
    - ask people to report on themselves. common method bias
    - others use "objective" or third party data.
    - example: are people who do yoga healthier?
    - issues establishing causality.
        sick people can't do yoga. mindset influences both yoga and eating
    - longitudinal survey studies
        - at least in the study, we can see if prior yoga affected later health

- archival studies -- no different from surveys
    - firms with female CEOs are more profitable
- case/control studies -
    - condition on the dependent variable
    - retrospective
- cohort studies
    - condition on the independent variable. like smoking
    - can be prospective or retrospective
    - might call this a natural experiment but we didn't manipulate the X

Simulations and Modeling

- **Computer Simulations:** Researchers create models that simulate real-world processes or systems to predict outcomes.
- **Mathematical Modeling:** Uses mathematical equations to represent complex systems. It’s often used to understand relationships between variables and to predict outcomes in a simplified, abstract manner.

Meta-Analysis and Systematic Reviews

- **Meta-Analysis:** Combines results from multiple studies to arrive at a comprehensive conclusion. It’s a quantitative approach to synthesizing research findings.
- **Systematic Review:** A rigorous summary of existing literature on a particular topic, often including meta-analysis. It follows a structured protocol to ensure comprehensive coverage and objectivity.

Mixed-Methods Research

- Combines qualitative and quantitative approaches in a single study or series of studies. The goal is to leverage the strengths of both methods to gain a more comprehensive understanding of the research problem.

qual vs quant

All interpretation is qualitative

Bot Data and Analysis can be qual or quant

In table below rows are types of analyses and columns are types of data

Data

Analysis	qual	Quant
qual	stealing fire hermeneutic analysis	Data visualization ● scatter plots ● stock market chart naming factors in a factor analysis
quant	Code texts for variables. Eg. code comments online for level of anger then correlate anger with gender of writer	predicting heart health as a function of cholesterol

Rows and columns below are different from table above

Obituaries

	qual	Quant
Data	content	Number of words, age at death
Analysis	Searching for themes	Predicting # of words

- analysis: when to use qual and quant
  - qual: theory and hyp generation; interpreting statistical results
  - quant: testing hypotheses. making precise predictions.

The monty hall problem

- Monty hall handout