204A Research Design Class notes:

Mixed analysis - the use of both qualitative and quantitative data.
Experimental data - data produced in a controlled setting where the researcher is able to manipulate critical factors, including assignment to treatment and control groups.
Observational data (non-experimental) - data not gathered as part of a controlled manipulation, data produced by the world.

Unobtrusive data - gathered without becoming involved with the respondents or data-generating process (soil samples, published government statistics, etc.)
Intrusive data - requires interaction with respondents/process (risks perturbation of the data, Hawthorne effects) (experimental data is *necessarily* intrusive)

Evaluating measurement
Construct validity and reliability are two criteria used to evaluate the quality of measurement (good measures have both, both matter regardless of the type of data)

Construct validity - how well you go from the conceptual measure to actual data
Reliability - how much noise is in your data

Construct validity - most studies involve abstract theoretical constructs that we never directly observe: health, poverty, drug abuse, presidential approval, democracy.
Before measuring, we must operationalize abstract constructs - translate them into concrete, observable indicators ("connecting ideas with facts")

- Conceptualiation - define construct and deliniate necessary components/attributes (moving from a background concept to a systematized concept)
- Operationalization - mapping construct to observable indicators, scoring cases
- Adcock & Collier (**Figure 1-get**):
    1. Background Concept
    2. Systematized Concept - specific formulation used by particular scholars
    3. Indicators - measures and operationalizations, operational definitions that maps to a specific measure
    4. Scores for Cases - the scores for cases generated by a particular indicator
- Example: (1-2: Conceptualization, 3-4: Operationalization)
    1. Poverty
    2. "absolute poverty" (systematized) - inability to meet basic needs
    3. "Over the past year, how often have you and your family gone without…
    4. Respondent's answers to survey questions
- Example: Democracy
    1. Democracy
    2. "A regime in which the highest offices of government are filled through periodic competitive elections…
    3. Two executive turnovers
    4. Scores: 0 or 1, assigned to each country year
- **Construct Validity**: does our operationalization (measure) adequately capture our construct (systematized concept)?
    - Getting from level 2 to level 4
    - Does the indicator produce scores that adequately capture the systematized concept?
    - Subjective, not mechanical

- Strategies for assessing construct validity
  - **Translation validity** (conceptual): enumerate construct into key components, measure all key components, and nothing extra.
    - Face validity: operationalization is valid on its face. Makes sense to an intelligent observer not active in the field
    - Content validity: if we enumerate all the attributes of the systemadtized concept (construct), does our operationalization capture all of the key attributes *and nothing else*?
  - **Criterion-related validity** (data-based)
    - Data-based: evaluate a measure by comparing it with other measures
    - Post-measurement
    - Most useful in a context with strong theory and established measures
    - Predictive validity: the measure should predict what theory suggests it should predict
    - Concurrent validity: the measure should correlate with a validated measure of the same construct (a "gold standard")
      - You might be interested in replicating a validated measure more cheaply with a larger sample size, for instance.
    - Convergent validity: alternative measures of a construct should be strongly correlated (no "gold standard")
      - Example: poverty - measures for not enough food, not enough clean water, not enough shelter, etc. should all be correlated
      - If they don't correlate, that doesn't mean any particular measure is wrong. They could also all correlate, and all be measuring the wrong thing.
    - Discriminant validity: measures of different constructs should *not* correlate
      - Example: Partisanship and athletic ability
  - **Threats to construct validity**
    - Failing to begin with a systematized concept ("inadequate pre-operational explication of concepts.
    - Measuring too narrowly, missing parts of the construct ("mono-operation or mono-method bias")
    - Inability to isolate the effects of the treatment from the effects of other factors occurring with the treatment ("The bundle problem")
      - Interaction of other treatments
      - The mere fact of being tested ("testing threats") or being observed ("Hawthorne effects")
      - "White Land Rover effects - when the NGO goes out to a village to administer a program, and they're the center of the village's attention and it affects everyone).
    - Avoiding threats: think through concepts and operationalizations in advance, and evaluate by examining correlations between your measure and others: predicted covariates, validated measures, etc.
- **Reliability:** the extent to which measurements can be replicated
  - Test-retest reliability, interrater reliability
  - Reliable measures have lower random error (noise). The more random error, the less reliable the measure

    - Random error vs. systematic error: random error doesn't change the mean (noise), just the variance. Systematic error shifts the mean of the distribution (bias). (IMPORTANT)
    - One way to enhance reliability is to narrow down to something concrete, but this narrowing can come at the expense of broader validity - tradeoff
- See Ferree's dartboard slides re: reliability and validity (week 2)
- Assignment #1 - data manipulation AND analysis

Lecture Readings:
Trochim and Donnelly, Chapter 3 (skip 3-1d through 3-1g).
King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press, Chapter 5, section 5.1 on measurement error.
Cook, Thomas D. and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin. Pages 59-70.
Adcock, Robert and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3), September: 529-546.

Discussion Readings: (focus on measurement strategies)
Fenno, Richard. US House Members in Their Constituencies: An Exploration. Reprinted in Weisberg at. al.

> "The paper addresses itself to two questions left underdeveloped in the literature on representative-constituent relations. First, what does the representative see when he or she sees a constituency? Second, what consequences do these perceptions have for his or her behavior? The paper reverses the normal Washington-oriented view of representative-constituent relations and approaches both questions by **examining the representative in his or her constituency**. The paper's observations are **drawn from the author's travels with seventeen U.S. House members while they were working in their districts**. Member perceptions of their constituency are divided into the geographical, the reelection, the primary and the personal constituencies. Attention is then given to the **home style** of House members. Home style is treated as an amalgam of three elements – **allocation of resources, presentation of self, explanation of Washington activity**. An effort is made to relate home style to the various perceived constituencies. Some observations are made relating constituency-oriented research to the existing literature on representation."

From <https://www.cambridge.org/core/journals/american-political-science-review/article/us-house-members-in-their-constituencies-an-exploration/30E53EAE8422C4C333A5406DB8CD72E0>

Gastil, Raymond. 1986. Freedom in the World: Political Rights and Civil Liberties, 1985-1986. New York: Greenwood Press. Pp. 3-30.

> "This yearbook marks the thirteenth year of the Comparative Survey of Freedom and is the seventh edition in the Freedom House series of annual publications. In addition to the ratings and tables of the Comparative Survey, this volume contains an extensive discussion of the criteria for and definitions of freedom. For the first time ever, the yearbook includes the checklist of political rights and civil liberties that forms the basis of the Survey's ratings system. Summary discussions of the status of freedom in each

country and related territories are included. This edition also examines the continuing controversy over the role of and regulations appropriate to the news media in the ongoing struggle for greater political, social, and economic freedom. It reports the outcome of a Freedom House-sponsored conference on strengthening American support for liberalization in Eastern Europe. Finally, the volume includes an assessment of the American campaign for democracy in the world and considers the opportunities and strategies appropriate to it."

From <https://www.amazon.com/Freedom-World-Political-Rights-Liberties/dp/0878558527>

3. James H. Fowler and Sangick Jeon, 2008, "The Authority of Supreme Court Precedent," *Social Networks* 30: 16–30.

   "We construct the complete network of 30,288 majority opinions written by the U.S. Supreme Court and the cases they cite from 1754 to 2002 in the *United States Reports*. Data from this network demonstrates quantitatively the evolution of the norm of *stare decisis* in the 19th Century and a significant deviation from this norm by the activist Warren Court. We further describe a method for creating *authority scores* using the network data to identify the most important court precedents. This method yields rankings that conform closely to evaluations by legal experts, and even predicts which cases they will identify as important in the future. An analysis of these scores over time allows us to test several hypotheses about the rise and fall of precedent. We show that reversed cases tend to be much more important than other decisions, and the cases that overrule them quickly become and remain even more important as the reversed decisions decline. We also show that the Court is careful to ground overruling decisions in past precedent, and the care it exercises is increasing in the importance of the decision that is overruled. Finally, authority scores corroborate qualitative assessments of which issues and cases the Court prioritizes and how these change over time."

4. Cottler, Linda B., Shaun Ajinkya, Bruce A. Goldberger, Mohammad Asrar Ghani, David M Martin, Hui Hu, Mark S. Gold. 2014. "Prevalence of drug and alcohol use in urban Afghanistan: epidemiological data from the Afghanistan National Urban Drug Use Study (ANUDUS)." *Lancet Global Health 2014*; 2: e592-600.

   Background Previous attempts to assess the prevalence of drug use in Afghanistan have focused on subgroups that are not generalisable. In the Afghanistan National Urban Drug Use Study, we assessed risk factors and drug use in Afghanistan through self-report questionnaires that we validated with laboratory test confirmation using analysis of hair, urine, and saliva.

   Methods The study took place between July 13, 2010, to April 25, 2012, in 11 Afghan provinces . 2187 randomly selected households completed a survey, representing 19 025 household members. We completed surveys with the female head of the household about past and current drug use among members of their household . We also obtained hair, urine, and saliva samples from 5236 people in these households and tested them for metabolites of 13 drugs.

1. King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review* 107(2): 1-18.

   We offer the first large scale, multiple source analysis of the outcome of what may be the most
   extensive effort to selectively censor human expression ever implemented. To do this, we have
   devised a system to locate, download, and analyze the content of millions of social media

posts originating from nearly 1,400 different social media services all over China before the Chinese
government is able to find, evaluate, and censor (i.e., remove from the Internet) the subset they deem
objectionable. Using modern computer-assisted text analytic methods that we adapt to and validate in
the Chinese language, we compare the substantive content of posts censored to those not censored over
time in each of 85 topic areas. Contrary to previous understandings, posts with negative, even vitriolic,
criticism of the state, its leaders, and its policies are not more likely to be censored. Instead, we show
that the censorship program is aimed at curtailing collective action by silencing comments that represent,
reinforce, or spur social mobilization, regardless of content. Censorship is oriented toward attempting
to forestall collective activities that are occurring now or may occur in the future—and, as such, seem to
clearly expose government intent.

Sampling - choose a subset of all possible units in the population.

The size of the sample doesn't necessarily need to be bigger for a larger population (larger than 1k or 2k).

Goals:

1. Minimize bias - representative sample, avoid selection bias

- Design goals of sampling
  - Reduce systematic error (selection bias) for a more representative sample.
    - Biased samples are not representative and have lower external validity
  - Reduce random error for (sampling error - noise) more precision in measurement
- How to draw a representative sample:
  - Probability Sampling: For large samples (50+), use probability sampling (always use if you can)
    - Uses random selection where each element has a known, nonzero chance of being included.
    - Avoids selection bias (in theory)
    - Representative
    - Possible to draw clear inferences about population parameters from sample parameters.
    - Do not use if $n < 50$
  - Feasibility Sampling
    - Pick the unit based on what is feasible or convenient
    - Examples: snowball, volunteer and convenience sampling
    - Drawbacks: prone to selection bias, difficult to statistically characterize the relationship between sample and population. Not representative.
  - Purposive sampling
    - Pick the unit based on research goals (not random)

        - Prone to selectin bias, cannot statistically characterize the relationship between sample and population.
        - Preferred for small-n designs, shows up in some large sample designs too (as an early stage).
- Probability sampling
    - Necessary but not sufficient for a representative sample
        - Poor design/implementation can let bias back in
            - Sample frame problems
            - Unit non-response
    - How to improve precision (reduce random error)
        - Draw a bigger sample (diminishing returns, but bigger is always better)
        - Reduce variance in the quantities of interest (by stratification)
    - Groups in sampling
        - Theoretical population (Ex. - all of Afghanistan)
        - Study population (Ex. - urban areas in safe regions of Afghanistan)
        - Sampling frame ("list" of the study population)
        - Sample population (the actual households they interviewed)
    - Sample = sampling frame plus sampling procedure
        - Sampling procedure - selecting items from the sampling frame
    - Sampling Frame
        - The set of units with a positive probability of being included in the sample
            - The study population and sampling frame should match
                - Every element in the study population is present in the frame, and present only once.
                - There are no elements from outside the study (nothing else)
            - Ideally, the sampling frame should precisely overlap the study population, but this is rarely possible.
            - Problems:
                - Ineligible units: inefficiency but not bias
                - Under-covered areas: missing part of the population, problematic when not missing at random
                - Under-coverage introduces sampling bias
                    - How to fix:
                        - Fix the frame (supplement with additional information)
                        - Redefine study population
                        - Use post-hoc weighting scheme (requires additional assumptions, not ideal)
        - Types of frames
            - List/census
            - Area set
                - Card catalog - pull every 10th card
                - Random walk
                    - Easy to see how this could not be random, but widely used
                - Random selection of geographic coordinates
                - Area frames: if you only visit once, and at one time of day, people encountered will not be a random sample.
            - Activity frame
                - Defined by an activity - individuals who went somewhere or did something
                    - Visiting a hospital, buying a product, entering a store, voting

                - Everyone who did the activity has to be in the sample
                - Implemented with a process, not a list
                - Ex. Exit poll
            - What can go wrong:
                - Missing people (multiple exits)
                - Not sticking to random selection pattern
                - Time of day
                - Unit non-response (people ignore the surveys, introduces bias)
                    - Non-respondents are systematically different from respondents, so bias is common. Destroys representivity.
            - Common problems: the sampling frame is wrong to begin with, or nonresponse issues arise
- Types of probability samples
    - Simple random sample (SRS)
        - Select n elements randomly from a frame (with N total elements), one at a time <u>without replacement</u>
        - Epsem: every element has an equal probability of selection
        - Ignores sub-group structure
        - Random samples have clusters - they're not evenly distributed
    - Deviations: (require recalculation of sampling error)
        - Stratified sample
        - Clustered samples
        - Systematic sample
    - N: the number of observations in the sampling frame
    - n: the number of observations in the sample
    - f: the sampling fraction = n/N
    - Sampling error: the extent to which sample parameters deviate from true population parameters due to random chance. The random error generated by taking a sample rather than measuring the population itself. Tradeoff: efficiency gains, but introduces random error.
        - Less sampling error indicates more precision and greater confidence that our sample is similar to population.
        - Sampling distribution: the distribution of a statistic (e.g. mean) generated from infinite repeated sample of the same size (n) from the same population. Will be <u>normally distributed</u>, with mean equal to the mean of the true population. (hypothetical - can only be approximated)
        - Sampling error is the standard deviation of the sampling distribution.
            - But the sampling distribution is hypothetical
            - For a simple random sample without replacement:
                - SE = SQRT((N-n/N)(s^2/n) copy from her notes
                - The higher the variance of the outcome and the smaller your sample, the more error
                - To reduce error, collect more data
            - S^2 is the sample variance of the outcome
        - Finite Population Correction: (N-n)/n is the finite population correction (FPC), which adjusts for the fact that the population is finite and sampling is done without replacement. For n>2, it reduces the sampling error. As N get large, it approaches 1, so in very large samples it's almost irrelevant and is sometimes omitted.

- To reduce sampling error:
  - Reduce variance in outcome (s^2)
  - Increase sample size (n)
  - N is not significant except when very small
    - This means that if you have a huge population you don't need a bigger sample than if you only had a moderate population.
  - Effects are non-linear
  - See her slide on Z-scores and confidence intervals
  - See her slide on determining sample size
    - If you have to estimate variance, overshoot! Will result in more data collected.
  - Proportions: in the case of binary variables, the mean and variance are mathematically related. Variance is maximized when $p = .5$ (conservative estimate). That makes $s^2 = .25$
  - See her sample size calculator - tells you how big a sample you need given p and s^2
- Sampling error and margin of error are estimates of the random error introduced by sampling. Even perfectly-implemented sampling procedures will have random error. Sampling error decreases when:
  - Sample size increases
  - Variance of outcome decreases

- Other Sampling Strategies:
  - Stratified sampling:
    - Strata are population subgroups (Gender, income, education, region, sector, age, etc.)
    - They can be discrete or continuous
    - Strata are selected based on data characteristics, research goals, and availability
    - You cannot stratify on something if you don't have that relationship in your sampling frame.
      - Ex: Can't stratify on income if you don't have income data
      - We could stratify IRIS on one of the six measures, for instance
    - Basic method:
      - Divide sampling frame into exhaustive, mutually exclusive population sub-groups (strata).
      - Randomly select a specific number of units from each stratum
      - It is possible to stratify on multiple dimensions, but common practice is to use 1-2
    - Two types of stratified sampling:
      - Disproportionate stratified sampling: sample over-represents some groups and under-represents others. Use if small, important groups are insufficiently represented in the simple random sample. Over-sample the small group to get enough of them.
      - Proportionate stratified sample: sample will reproduce population proportions on stratified dimension. Reduces sampling error (noise).

- Drawing a sample: Sampling frame + sampling procedure
- N = number of observatiosn in the sampling frame
- n = the number of observations in the saple
- f = n/N = the sampling fraction
- Sampling error: random error introduced by studying a sample rather than the population
- Stratified Sampling:
    - Strata are population subgroups
        - Gender, income, education, race, ethnicity, region, religion, etc.
    - Variance reduction mechanisms:
    - Disproportionate stratified sampling: sample over-represents some groups and under-represents others
        - Sampling fraction (f) is different across strata
        - Higher for small group, lower for a big group
        - Not epsem! (equal probability of selection method)
        - When analyzing, use weights (inverse of sampling fractions) to correct for over-representation.
        - Use when a proportionate random sample would have vanishingly small numbers of important subgroups.
        - Distinct from post-hoc stratification (less accurate)
    - Proportionate stratified sampling:
        - Select number of units in each strata to be proportionate to the sampling frame
            - The sampling frame is the same for each stratum
        - The procedure is to stratify the sample and then take a random sample *within each category*, taking f from each one.
            - This eliminates one kind of sampling error - variation across strata
                - Effectively shrinks variance
            - We're engineering the sample to have the same proportions as the population on the stratified dimension.
        - Has the most impact when there is a lot of variation across groups but not much variation within groups.
            - A dimension that has a big impact on outcome
    - Sampling error for SRS - see her slide with formulae
    - It's only possible to stratify on variables for which we have data
    - Should be part of the data framing exercise before collecting data
        - "What variables are driving the variance?"
- Cluster Sampling:
    - A clustered design is less expensive and easier to implement than a simple random sample
    - Population subgroups (often geographic clusters)
    - From a frame of all clusters, randomly select some, then study *all* units within the cluster
        - Two-stage sampling: taking a simple random sample within the cluster,
    - Big operational advantages - cheaper to go to a few places than everywhere

    - Does not require a separate sample frame for every unit, just a frame of the clusters
    - But, it increases sampling error
        - This error is minimized when the clusters are similar to each other (interchangeable) and representative of population
        - Maximized when clusters are idiosyncratic (not interchangeable) and not representative of population (and when you don't use many clusters)
    - See her slide on sampling error for a clustered sample in an SRS
        - Variance of cluster means (sa^2)
    - See her chart on stratified vs. cluster sampling
- Systematic sampling
    - Easier to implement than a simple random sample
    - Systematically move through the sample frame, sampling every nth household
    - If the order is random, then sampling error is equivalent to SRS
    - If the list orders units on an underlying dimension, then systematic sampling actually is the same as stratification on a continuous variable, and sampling error can be calculated with proportionate stratified sampling formulae.
- Departures from the simple random sample
    - Very common
    - For practical reasons
        - List not available, cumbersome (systematic sample)
        - Random sample requires geographically dispersed data collection, is too costly (cluster sample)
    - For research goal purposes (FIND THE REST OF THIS SLIDE)
- Causal inference
    - Empirical evaluation assesses the impact of a treatment (x) on an outcome (y)
        - Also known as "independent variable" (x) and "dependent variable" (y)
    - Treatments can take many forms
        - Concrete, direct, deliberate, explicit (doctor treats a patient)
        - Generated by nature (disasters), or the aggregation of many decisions (civil war)
    - Correlation: has there been a change in outcomes where the treatment was given?
        - Positive correlation, negative correlation, and no correlation
        - Third variable problem (spurious correlation) - correlation can be driven by a third variable causing the first two to correlate (rather than causation within the two variables)
        - Endogeneity problems - x doesn't cause y, y causes x (reverse causality)
        - Selection into treatment - outcome is driven by selection (fast swimmers get more practice time)
    - Causation: did the treatment *cause* the change in outcomes?
        - Consider the counterfactual - what would have happened without treatment?
        - Approximate a counterfactual in the real world (design a valid comparison (control) group
        - Evaluate the empirical relationship between treatment and outcome. Is there one?
        - Rule out alternative explanations for the observed relationship
        - If the treatment works: is it a **treatment** effect or a **selection** effect?
    - The perfect research design
        - Involves the counterfactual - what the outcome (y) would have been in the absence of treatment (x)
        - Only x changes, nothing else
            - This rules out alternative explanations *by design*
            - Potential outcomes framework (Donald Rubin) - see her slide

        - Real research design: the world of second-bests
            - Can't study clones, study twins
            - Can't go back in time, study natural experiments (designating valid comparison groups)
    - **Valid comparison groups**
        - Treatment and control groups must be highly similar in the absence of treatment
        - Groups are highly similar at the baseline
            - Similar on observables *and* unobservables
        - Groups would change over time in highly similar ways absent treatment
        - Groups aren't differentially exposed to other interventions during the evaluation period
        - Similarity is really important because the more similar the groups are at the baseline, the more confident we can be that changes are due to treatment
    - Quasi-experimental design
        - Engineer treatment and control groups to be highly similar
            - Matching designs - match every treated unit with an untreated twin. Compare outcomes across all pairs, take average difference.
            - Statistical designs: use statistical tools like multivariate regression analysis to "hold constant" differences between treated and untreated units.
    - Experiments - rely on randomization to create highly similar groups
            - At least two comparison groups
            - Randomization (random assignment of treatment – confers internal validity)
                - This gives us groups that are equivalent on both observable and unobservable characteristics
            - Controlled by the researcher
            - Optional – random sample (confers external validity)
        - True experiments: two comparison groups, random assignment of treatment researcher controls process of randomization
        - Natural experiments: processes outside the researcher's control plausibly introduce randomization

10/26

- Fundamental problem of causal inference: we can't observe the counterfactual - what would have happened if the treatment didn't occur? This is a missing-data problem.
    - We need to designate comparison groups:
        - Equivalent groups: Same on observables (height, weight, location) and unobservables (beliefs, aptitude, competence, motivation, culture)
        - We need a design that factors out unobservables (like a natural experiment)
- Design Notation
    - O = observation/measurement
    - X = treatment
    - In theory, treatments can be binary or continuous (we start our analysis with binary treatments)
    - Time sequences move horizontally (R O X O) - we randomize, take a baseline, apply a treatment, measure again.
        - R is random, N is non-equivalent (same as non-randomized)
    - Each row is a group - groups move vertically (R O _ O) - control group
    - Dif-and dif (differences and differences design) (N O X O)
        - If this is done without a control/comparison group, it's a time-series

    - Subscript 0 indicates a control group
- Experiments
    - Requirements:
        - Two or more comparison groups
        - Large n (>100)
        - Each group is exposed to a particular condition
            - Classically, treatment and control
            - Control is not required, but different treatments to at least two groups *is* required
            - Assignment to groups is controlled by the researcher
            - Assignment is done randomly
                - In large samples, randomization creates equivalent groups, or *ex ante* symmetry
                    - Equivalent does not mean identical - there will be random differences
                    - A small but predictable fraction (1/20) of these differences will be statistically significant due to random chance.
                - Randomization *creates equivalent groups* which end up being similar (up to random error) on *both observable and unobservable characteristics*. It rules out alternative explanations *by design*. If we observe post-treatment differences, then we can confidently attribute this difference to treatment.
            - See Gertler 2011! World Bank resource. Then the Duflo handbook.
        - Baseline covariates - pretreatment measure
        - Balance table: High T-stats mean that the randomization procedure was deficient
            - We care because a high-T stat means that the treatment and the control group were different on that measure, and what if that measure is driving the treatment results? Then we no longer have causation.
            - You can only do a balance table if you do a pretreatment measure (O)
        - Random sample vs. random assignment:
            - Random sample: how you select units to study. Ensures representative sample and external validity (surveys)
            - Random assignment: how you assign treatment and control to units in your sample. Enables causal inference and ensures internal validity. (Mechanical Turk)
            - Studies with both external and internal validity do both (see her chart on Random Sample vs. Random Assignment slide)
        - Experiments in three setting:
            - Lab - highly controlled and contrived setting
                - Convenience samples (undergraduates in the US) (do not have external validity)
                - Prioritize internal validity over construct or external validity
                - The treatment is often different from the thing you're trying to measure (cooperation measured by the divider game - low construct validity, external validity).
            - Survey - treatment is different version of the survey
                - Often low construct validity (treatment you can deliver in a survey is often quite different)
                - Can get a representative sample
                - Can be administered to much larger n than lab experiments

        - Field experiments
            - In "natural" environment of participants
            - "Social science experiments" explore set of hypotheses via constructed treatments
            - Randomized controlled trials: explore impact of program by randomizing some aspect of it.
    - What is Blocking?
- Analysis
    - Stratification and blocking:
        - Divide sample into groups sharing the same or similar values of certain observable characteristics (strata, or blocks)
        - Randomize assignment within each stratum or block
            - “Block” is used only for discrete/nominal data
            - “Strata” is used for continuous variables
                - This distinction is only used in polisci - economists don't talk about "blocks"
            - Only possible if you have a pre-treatment measure
        - Rationale: to ensure that the randomization produces similar groups along the dimension being blocked (don't leave balance to chance).
        - Blocking is a noise reducer, it absorbs variation in the outcome, increasing precision of estimates.
        - Differences between blocked/stratified & one-shot randomizations tend to disappear with n>300, so blocking is most useful with a relatively small n.
        - Block on variables correlated with the outcome, where there is low variance within group, high variance across group. We are blocking out a lot of variance.
        - Frequently used: baseline measure of the outcome variable; geographic units.
        - See slide: stratification and blocking: method
        - When we do this ourselves for assignment 2, consider using blocking.
    - Factorial (Cross-cutting) Designs
        - Multiple factors are tested simultaneously with randomizations conducted independently so that treatment assignments are orthogonal (cross-cutting).
        - Useful for exploring multiple factors at once, interaction effects
        - See her slides - graphical layout
            - Null effects vs. main effects vs. interaction effects
        - "If you do this, you had better have a hypothesis about an interaction term!"
            - This means we’re studying an interaction, otherwise factorial design is unnecessary.
    - Heterogeneous treatment effects
        - Explore interactions between the treatment and an un-manipulated factor, like age, gender, height, etc.
        - Done post-hoc (not built into the design), risks spurious correlation
        - Specify in pre-analysis plan if possible
    - Cluster randomization
        - It is often natural (and cheaper) to implement randomization at a unit more aggregated than the true unit of analysis.
        - Also useful if you are worried about spillovers (contagion) between treated and untreated units.

- Examples: School-level randomization of education programs, hospital-level randomizations of medical practices.
- *Must* cluster standard errors in estimation of treatment effects. This will usually increase them, with effects increasing with the intracluster correlation (ICC) in the data.
- The power of the test has more to do with the number of units over which you randomize (schools, hospitals) than it does with the number of units in the study.
- Fewer clusters, high ICCC mean larger standard errors, harder to find effects.
- Downside – your n is now basically the larger-scale n (villages, not people)

- Compliance
  - Because most programs are voluntary, imperfect compliance is the norm.
    - People fail to "uptake" the treatment
  - All experimental treatments potentially suffer from compliance issues
    - Survey experiments don't often suffer from a compliance problem
      - Unless people stop listening
  - Imperfect Compliance - compliance can have heterogeneous effects – certain kinds of people can tend to comply while others don't.
    - Those who get treatment are not a good comparison group for those who don't get treatment.
      - Randomization creates equivalent groups; differences in compliance means the groups are no longer equivalent.
  - To accommodate imperfect compliance, instead of estimating the Average Treatment Effect (ATE), we measure the:
    - Intention to Treat Effect (ITT) – literally the effect of *offering the program* to the average person in the evaluation sample (results in a diluted effect)
      - Ignores uptake, just look at assignment: outcomes in group *assigned to treatment* vs. outcomes in group *assigned to control*.
      - No extra assumptions, pure product of experiment. Easy to calculate.
      - Default procedure for a lot of experimental design "when they don't say what they're doing, they're using an ITT"
        - Tends to underestimate effect
      - When compliance is very low, ITT may be very weak. Risk Type II errors.
    - Treatment-on-Treated effect (TOT) – the effect of getting the program *on the treated* (results in exaggerated effect) (rare in political science)
      - Wald estimator for *treatment on compliers* – ITT/compliance rate (see TOT slide) (instrumental variables estimator)
      - Y: outcome; Z: assignment indicator; T: uptake
      - Treatment on Compliers = TOT when non-compliance only affects the group assigned treatment
      - With perfect compliance, the denominator = 1, TOT=ITT=ATT
      - With imperfect compliance, TOT will be large than or equal to ITT (since we're dividing by 1 or less in the Wald estimator)
      - TOT requires extra assumptions:
        - Assumptions required to use an instrumental variables estimator:
        - Monotonicity – assignment to treatment makes subjects more likely to be treated (usually a safe assumption)

                - Assignment to treatment effects affects outcomes only through uptake of treatment (no spillover of treatment effects/contagion)
                - Spillover can happen when the exclusion restriction is violated:
                    - Within the experimental group (compliers to noncompliers – ex: herd immunity)
                    - From the experimental to the control group, or vice versa
        - When there is perfect compliance, ATE=ITE=TOT
        - Otherwise TOT>ATE>ITE
        - The safest thing to do is simply calculate the ITE
        - It's also a good idea to calculate a higher bound (if assumptions are met – looks like TOT) and a lower bound (if the assumptions don't apply, looks like ITE)
- Conclusion validity (usually learned in statistics classes)
    - A measure of how well we have assessed a correlation in our data. Are the conclusions reached about relationships in data reasonable?
        - If we say there *is* a relationship, is the conclusion reasonable?
        - If we say there *isn't* a relationship, is *that* conclusion reasonable?
    - Threats: Type I and Type II errors
        - Type I : Finding something that *isn't* there
            - The odds of making a type one error – alpha, or p-value
            - Confidence level: 1-alpha
            - Risks of Type I errors:
                - Setting the bar too low: .05 means one in 20 observations will correlate by random chance
                - Fishing/p-hacking – running your data multiple times with different p-values – violates the whole point of the p-value
        - Type II: Failing to find something that *is* there
            - The odds of finding no effect when there is one (Beta) k=1-beta; k="power level"
            - A beta of .20 implies a power level of .80: 80 times out of 100, when there is an effect, we will find it, but 20 times out of 100, we won't.
            - Risks:
                - Setting the bar too high: Lower p-values (alpha) are higher bards. The lower we set the p value, the higher the Type II errors.
                - Solution: collect more data.
                - Underpowered studies with null effects do not constitute a "meaningful zero"
        - Improving power:
            - Increase sample size
                - Note: clustering does the opposite, thereby reducing power
            - Reduce random error (noise) in the outcome variable
                - Stratify or block
            - For ITTs, improve compliance
            - Have treatment and control groups of similar size
            - More details: field of "Power Analysis" (see her slides on this)
                - Must know the expected treatment effect, the variance of outcomes, and the treatment uptake percentage.

- You can get away with a smaller sample size if you have:
    - High expected treatment effects
    - Low variance outcomes
    - Treatment & control groups of similar sizes (p=.5)
    - A willingness to accept low significance and power
- Interaction terms often require a lot more power to demonstrate than main effects (Gelman)
- The smaller the expected treatment effect and the higher the variance in the outcome, the more data required to find that effect.
- For experiments, more data is always better.
- Be wary of interactions.
- Rough rule of thumb: n under 500 is usually bad news unless effect size is large and outcome variance is small.
- Power is relevant because it helps to interpret your null results. If you observe no effects, sufficient power means that there actually weren't any effects.

- Threats to inferences in experimental design
    - Threats to conclusion n validity
        - Insufficient power (sample size not big enough)
            - Potential causes: too few units, cluster design with high ICC and small numbers of clusters, poor compliance and weak ITT.
            - Consequences: shows no effect, but impossible to conclude that there *wasn't* one.
    - Threats to internal validity:
        - Mortality/attrition: the failure to measure outcomes for certain subjects
        - If attrition is *random*, groups are still equivalent
        - Non-random, "differential attrition" is more common, and is a big problem. Treatment and control groups start off equivalent, but then become less so over time.
        - Mitigate: track participants closely, motivate completion, sensitivity tests.
        - "Social" threats (Trochim)
            - Diffusion or imitation of treatment (contagion, externalities)
                - Treatment is diffusing to the control group
                - Solution: model the contagion
            - Compensatory Rivalry (John Henry effects)
                - Individuals in the control group learn about the treatment and work harder to simulate the effect of the treatment.
            - Resentful demoralization
                - Individuals in the control group get demoralized by being controls and behave differently (stop working hard, etc.)
            - Compensatory equalization of treatment
                - People running the experiment feel bad about excluding the controls and extend treatment to them (cash handouts, etc.)
            - Solutions: model diffusion directly, segregate T and C, rigorously monitor staff/participants, placebos (eliminates Compensatory and Resentful).
    - Threats to construct validity (in experiments)
        - Treatment in the experiment may not match up well with theoretical treatment

                - Example – construct: ethnicity of candidate; experiment: Luo last name
            - Bundled treatments:
                - Hawthorne effects: the treatment group interacts with the researcher, but the control group does not, so is the effect from treatment or interaction?
                - Testing threat: see slide
    - Threats to external validity
        - Failure to draw a random sample
            - Study is not reperesntative, does not generalize
        - Gold-plated experiment problem:
            - Study was implemented in unique conditions (place, organization, program, resource level) that are unlikely to be replicable elsewhere.
        - Non-response
        - Solution: use random selection if possible, build external validity through <u>replication.</u>
- Basics of Survey Experiments
    - In survey experiments, the survey both:
        - Delivers the treatment, and
        - Measures the outcome(s)
    - Randomization ensures balanced groups
    - "Treatments" are different versions of the survey
        - Often different in minute but significant ways
    - Common objects of manipulation (factors)
        - Question wording
        - Question number or order
        - Response options
        - Additional informational content
    - Ideal treatments are small; clear, simple, discrete, they move one thing at a time
        - They also need to capture the construct in question
    - Common outcomes in survey experiments:
        - Behavior during the survey
        - Beliefs, perceptions, information, understanding, etc.
        - Hypothetical or projected behavior outside of the survey
-

MAKE SURE THE SUBJECTS ARE RANDOMLY ASSIGNED TO TREATMENT
Go for a 2x3 factorial (n=2000), not 3x3 (unless n=5000+) (400-500 per cell)
INCLUDE A BALANCE TABLE (NATURAL EXPERIMENT)

- Hallmarks of Natural Experiments
    - Two or more comparison groups
    - The researcher does not control the assignment to treatment
    - The researcher argues that the assignment was: (each weaker than preceding)
        - Random (uses explicit randomization device)

        - "As if random" (actual process is not precisely known, but appears to have been random."
        - Exogenous with regard to the outcome in question.
    - The credibility of natural experiments depends on the plausibility of claims about assignment. Never accept them at face value.
        - If there's clear, transparent randomization, this is a strong claim for randomization
        - Dubious arguments about as-if randomization or exogenous treatment presents a weaker claim for randomization.
    - To evaluate:
        - Consider the evidence about the process of assignment
        - Evaluate the balance (table) on pre-treatment characteristics
    - See her slide on Varieties of Experiments
    - Useful for studying history, since we can't do an experiment.
- Standard natural experiments – some process (not the researcher) assigns units to treatment.
    - Sometimes true randomization (lotteries)
    - "As-if" randomization or "exogenous to DV"
    - Instrumental variables – not a natural experiment type, but some process assigns units to a variable (Z), correlated with the treatment of interest (X). We exploit this to isolate the effect of X on Y. More controversial, requires additional assumptions. Z can only be related to Y via X. Z is randomly assigned by nature, not X (as in a standard natural experiment).
    - Regression discontinuity design (RDDs)– a threshold (possibly designated by the researcher). Everyone below [above] the threshold gets treatment, everyone above [below] does not.
        - Often natural experiments, but not always
        - Involves the treatment assignment process
        - Use a cutoff score, and examine whether there is a discontinuous jump in outcomes at cutoff.
            - Assignment to treatment is "as-if" random right around the cutoff. This simulates a natural experiment.
            - Collect a balance table of pre-treatment covariates around the cutoff point
        - Requires a continuous pre-program measure, called the assignment, forcing or running variable.
        - Can have strong internal validity but weak external validity (b/c treatment effects are measured only at the cutoff point, which is the only point where there's robust causal inference), and power issues
        - The entire untreated group is not a good comparison for the entire treatment group because they're different on many variables. Comparison should be at the cutoff point.
        - Design-based approach: estimate treatment effect as the simple difference in means of groups in narrow bands on either side of the cutoff.
            - The narrower the band, the more likely the assumption will hold, but the less data you have (tradeoff). Begin with an extremely narrow band, with low power but high internal validity, then expand the band out, improving power and reducing errors, but also weakening causal inference.

        - Model-based approaches: statistically model the relationship between the forcing variable and the outcome, and test for discontinuity in that relationship at the cutoff point (see her equation in "Two Approaches to Analysis" slide.
        - There must not be any spurious discontinuity in the relationship between the forcing variable and the outcome at the cutoff point (must be a continuous function). It's vital to use the correct functional form (see slides). Check many functional forms to confirm that the effect isn't model-dependent.
        - The effect is LATE (local average treatment effect) – limited in scope to the range of data around the cutoff point.
        - Four issues with RDDs
            - Sorting – evaluate at cutoff, no heaping (does data heap at a certain point when we graph across the variable? Means non-random, can't do an RDD.)
            - Balance – covariates should not jump at cutoff
            - Robustness – estimates are not sensitive to the particular model chosen
            - Placebo Tests – if the relationship between the forcing variable and the outcome is continuous in the absence of treatment, then there shouldn't be any discontinuities at fake cutoff points.
    - For all natural experiments:
        - Who or what is doing the assignment?
        - How credible are claims to randomization?
            - Find qualitative evidence confirming randomization
        - What evidence do the authors use to argue in favor of randomization?
    - Advantages to natural experiments:
        - No need to worry about treatment effects, human subjects problems
        - Can study history!
- Standard natural experiments
    - John Snow – cholera
- Quasi-experiments
    - Cross-section
        - Treated and untreated nits
        - Outcome observed post-treatment only
        - Ex: survey
        - N – non-equivalent groups (non-random, alternative to R in design notation)
        - Two groups, between subject design
        - One post-treatment measure
        - Assumption for valid comparison: independence of treatment assignment and outcome (despite non-randomness)
            - On all dimensions related to outcome, groups were highly similar pre-treatment.
        - Threats: any violation of the assumption that assignment was independent of Y.
            - T and C groups are different at baseline, and the difference correlates with outcome.
        - Neutralizing threats: create post-hoc independence between treatment and outcomes by conditioning on observables (X)
            - Two approaches:
            - OLS (control for X) and
            - Matching estimators (match treated and control units on X (observable characteristics), hold X constant for that unit, look at differences.

Advantage: less model-dependent, forces you to only use data for which there's a reasonable match. Avoids extrapolation.

- Drop units for which there's no match.
- "Find someone who looks just like the subject but who didn't get the treatment."
- If you have a lot of variables, a good match is hard to find. "The Curse of Multidimensionality"
    - Solution: propensity score matching (p(x))
        - Use observables to estimate a propensity score p(x): the probability that the unit is treated.
        - Estimate logit or probit model, calculate predicted fvalues for each observation, p(treatment).
        - Single number p(treatment) summarizes allobservables influencing the participation/treatment.
            - This makes multiple variables into a virtue, but is still model-dependent, like OLS.
        - Match treated and untreated units on their propensity score.
- Unobservable difference will still be a problem.

- Confounds must affect both the treatment process and the outcome
- Omitted variable bias – omitting a variable correlated with both treatment and outcome.
- Fixed Effects: compare members only from within the same group

- (interrupted) time series (within subject)
    - One unit, observed (at minimum) pre- and post-treatment
- Panel design (difference in differences)
    - Treated and untreated units, pre- and post-measures
    - Within and between subjects