QTM 101 Midterm 1 Cheat Sheet

Section 1, 10am-11:15am, Wed, Oct 4$^{\text{th}}$, 2023

Name: Jiuru Lyu

Types of Questions

- Descriptive: Something we can **see** in the data.
- Predictive: What the **future** might look like
- Causal: If X changes, will Y change?

Strength vs. Magnitude

- Covariance: retain units
- Correlation (r): standardized covariance $\Rightarrow r \in [-1, 1]$
  - Sign: direction
  - Absolute Value: strength
- Regression coefficient (β): magnitude $\to y = \beta x + \alpha$.

All of them should have the same sign: all positive if y increases when x increases; all negative if y decreases when x increases.

- Covariance isn't actually telling us anything because they retain units - they cannot be used in comparison!
- Determine which is the most **predictive**? (***strength**, so we should use correlation coefficient*)
- Determine which is the most effective **chance**? (***magnitude**, so we should consider the regression coefficient*)

Counterfactual Dependence

- We need to prove both: (1) when X happens, Y happens; and (2) when X does not happen, Y does not happen.
- Some notations:
  - $T = 0$: control, untreated
  - $T = 1$: treatment, treated
  - Y_1: outcome if treated
  - Y_0: outcome if untreated
- Fundamental problem of casual inference: We cannot observe ATT (Average Treatment Effect of Treated) anyhow!

$$\text{ATT} = \mathbf{E}[Y_1 \mid T = 1] - \mathbf{E}[Y_0 \mid T = 1],$$

  where $\mathbf{E}[Y_0 \mid T = 1]$ is not observable because we cannot have the same group be both treated and untreated.

- Resolution: Estimate the ATT:

$$\widehat{\text{ATT}} = \mathbf{E}[Y_1 \mid T = 1] - \mathbf{E}[Y_0 \mid T = 0],$$

  where both $\mathbf{E}[Y_1 \mid T = 1]$ and $\mathbf{E}[Y_0 \mid T = 0]$ are observable.

- An important equation

$$\underbrace{\boxed{\text{Estimate}}}_{\text{what we see in data}} = \underbrace{\boxed{\text{Estimand}}}_{\text{what we want to know}} + \text{Bias} + \text{Noise},$$

  where possible Estimand's are (remember that Estimand's are unobservable true values in our experiment)

  - $\text{ATE} = \mathbf{E}[Y_1] - \mathbf{E}[Y_0]$
  - $\text{ATT} = \mathbf{E}[Y_1 \mid T = 1] - \mathbf{E}[Y_0 \mid T = 1]$
  - $\text{ATU} = \mathbf{E}[Y_1 \mid T = 1] - \mathbf{E}[Y_0 \mid T = 0]$

  The only observable effect (Estimate) is

$$\widehat{\text{ATE}} = \widehat{\text{ATT}} = \widehat{\text{ATU}} = \mathbf{E}[Y_1 \mid T = 1] - \mathbf{E}[Y_0 \mid T = 0]$$

- Following up, we use ATT to represent our general Estimand and $\widehat{\text{ATT}}$ as our Estimate.

Identifying Inferential Errors

- Selecting on the DV: Only one **outcome** in data
- Confounder/Common Cause: **omitted variable** (We are not comparing apples to apples)

Bias and Noise

- Bias: Systematic error; Not apples-to-apples; Inaccuracy

$$\text{Bias} = \mathbf{E}[Y_0 \mid T = 1] - \mathbf{E}[Y_0 \mid T = 0]$$

  - To measure Bias, we give the two groups Placebo (however, bias is still unobservable).
  - Eliminating bias: **randomization**.
- Law of Large Numbers (LLN): the sample average can be arbitrarily close to the true population average by making the sample large enough.
- Noise: Random error; $\mathbf{E}[\text{Noise} = 0]$; Imprecision
  - Everything is noisy.
  - Measured by standard error

$$\text{SE} = \sqrt{\frac{p(1-p)}{N}}, \quad p = \text{porportion of outcome}$$

  - If we repeat our experiment for enough times, our Noise should be 0.
  - We are not super stressful with Noise since the solution is to increase sample size.

Confidence Interval and p-values

- Point Estimate (=$\widehat{\mathrm{ATT}}$) $\xrightarrow[\pm 2\times\mathrm{SE}]{\mathrm{MATH}}$ 95% CI
  - If we run the experiment identically 100 times, 95 of them should contain the true value of the estimand.
  - When we have 99% CI, then we should find larger intervals.
- Interpreting a CI: Sign and Magnitude (every important if the interval contains 0).
  - $\widehat{\mathrm{ATT}} = 10;\ 95\%\ \mathrm{CI} = [6, 14]$: If we run the experiment identically 100 times and calculation the confidence interval in each repetition, out of 95 times, we are confident that $[6, 14]$ captures the ATT. Furthermore, ATT is likely **positive** and has a **large magnitude**.
  - $\widehat{\mathrm{ATT}} = 3;\ 95\%\ \mathrm{CI} = [-2, 8]$: If we run the experiment identically 100 times, and we calculate the confidence interval in each repetition, out of 95 times, we are confident that the interval $[-2, 8]$ captures the ATT. However, ATT **sign is unclear** and has a **small magnitude**.
- Central Limit Theorem: If we repeat a study/analysis/experiment a zillion times, the value of many of the statistics we get will follow a normal distribution.
- Null Hypothesis: $\mathrm{ATT} = 0$. ($\Rightarrow$ Any nonzero $\widehat{\mathrm{ATT}}$ is due to noise)
  - **Fail to reject the null**: $\widehat{\mathrm{ATT}}$ could be due to noise
  - **Reject the null**: $\widehat{\mathrm{ATT}}$ is probably not due to noise
- p-values:
  - High p-value $\rightarrow$ **High** chance of observing if $\mathrm{ATT} = 0 \rightarrow$ **fail to reject** the null $\rightarrow$ ATT could **reasonably** be 0.
  - Low p-value $\rightarrow$ **Low** chance of observing if $\mathrm{ATT} = 0 \rightarrow$ **reject** the null $\rightarrow$ ATT is **unlikely** be 0.
- Threshold of p-values: 0.05
- Connecting p-values and CI's
  - $\widehat{\mathrm{ATT}} = 5;\ 95\%\ \mathrm{CI} = [-1, 11];\ p\text{-value} = 0.13$: **fail to reject** the null, ATT could **reasonably** be 0 (the result could be due to noise).
  - $\widehat{\mathrm{ATT}} = 5;\ 95\%\ \mathrm{CI} = [2, 8];\ p\text{-value} = 0.03$: **reject** the null, ATT is **unlikely** be 0 (the result is not due to noise).

Experiments

- Random Assignment
  - Observational studies: non-random assignment; based on observed treatments in the world
  - Randomization: a computer selecting random number to allocate treatment
  - Stratified Randomization: split subjects by one or few factors then randomize within those strata
- Types of experiment:
  - Lab: classical model of experiment
  - Field: experimenter manipulate a treatment in the real world. -*Randomized Control Trial (RCT)*
  - Lab in the Field
  - Quasi or Natural Experiments
- Fundamental Principle of Controlled Experiment: Actual outcomes among the control group should give the same counterfactual outcomes of the treated groups
- Internal Validity: how close $\mathbf{E}[Y_0 \mid T = 1]$ is to $\mathbf{E}[Y_0 \mid T = 0]$.
  - Chance imbalance: check the balance in observables.
    - throw out broken experiment
    - proceed as normal
    - compare the experimental effects within groups sharing the same level of unbalanced observable
  - Lack of statistical power: larger sample size $\rightarrow$ less noise $\rightarrow$ more statistical power (=ability to detect a true effect if one exists)
  - Non-compliance: subjects fail to take the treatment
    - Intent-to-treat effect (ITT)
  - Placebo effects: individuals know they are part of the experiment and will be recorded.
    - Double blinded experiments
    - Three groups: treatment, control, placebo; compare measure $(T - C)$ and $(P - C)$, then true effect is $(T - C) - (P - C)$. Or simply compare T to C.
  - Attrition: some subjects drop out of the analysis after randomization, and we cannot observe their outcomes.
    - confident that attrition is random: still estimate the average effect.
    - attrition is non-random but unrelated to treatment: estimate the average effect for those stayed in sample
    - attrition is non-random and related to treatment
  - Interference: control and treatment are in contact
    - Contamination
    - Spillovers