**Lecture**

**Internal Validity- Design**

<https://neptune.ai/blog/exploratory-data-analysis-natural-language-processing-tools>

The process of establishing causality by ruling out everything else

* In this study… → if constructs are related
* Is that traceable to events in the world or traceable to sth else
* Establishing causality
  + Process of ruling out as many plausible alternatives - alternative hypotheses - as possible
  + Temporal precedence - cause must happen before effect
  + Plausibility - top note

Threats to internal validity

* These are the plausible alternatives you should be worried about
  + Think using these categories
  + Confound- treatment or effect claim to make the call
  + Actually causing sth else
  + Maturation confound - processes within respondent operating as a function of passage of time perse (growing older, growing more tired..)
  + Testing confound - act of measurement is changing the measurement
    - Learning from previous question - changing future measurement
  + Instrumentation confound - systematic bias,
  + Statistical regression confound - selecting group by sth interesting, bc of the way that you selected group they regress to them - players might have rlly good day so average they dont do that well
  + Experimental mortality - any processes that cause to people drop out the experiment - most depressed ppl might stop posting
  + Mutli-group threats- control for lot of confounds by using multiple groups
  + Selection - x threats: two groups mature in different rates
  + Social threats:

One-group pretest-protest study

* Randomization:

Covariance designs

* Noise reduction: Ancova
  + Controlling noise

Nonequailvent experiment- compare two separate groups like anxiety vs depression

Regression point displacement design- only one sector different response to remote work than other sectors

Multi-time series - facebook with pro vaccination vs anti vaccination - compare trends

Design that mitigate these threats

**Does covid affect post trends from mental illness related subreddits?**

**Mental health-related subreddits exhibit a shift in post trends as well as the number of engagements based on post texts, number of comments and upvote ratios, between periods pre-covid, mid-covid, and post-covid**

For post content, we will gauge association with distribution over words, and apply word clustering into topics which we will use LDA for in this case. This will allow us to compare before, during and after covid.

As scraping reddit posts are limited to 1000 posts, we instead utilize the upvoted ratio and ave number of comments as more appropriate metrics for assessing interaction tendencies rather than quantity of posts. Also, unique user ids to analyze shifts in user dynamics across different time intervals.

By focusing on top posts which are already trending during each time interval, post trends are more likely to be identified.

**Questions**

1. **Describe in detail the causal theory underlying your analysis. What are the proposed causes and what are the proposed effects?**

Cause: covid

Effect: shift of post trends on subreddit

What you do →

What you see →

Alternative cause : some other reasons?

1. **Show pilot data demonstrating some statistical relationship, including temporal precedence and covariance between your proposed cause and effect.**

Temporal Precedence –The cause must happen BEFORE the effect • Covariation of Cause and Effect –The effect DOES happen WITH the cause ANDdoes NOT happen (as

Temporal precedence - covid pandemic

**Statistical relationship graph**

* **Frequency of upvoted ratio**
* **Frequency of num comments**
* **Top posts num**
* **Scores**
* **Unique id**

1. **For each possible type of threat to internal validity, indicate whether your analysis is susceptible.**
2. **For each threat for which your analysis is susceptible, indicate whether that threat is plausible.**

Threats to internal validity - plausible alternatives you should worried about

* **History confound : instead of covid, personal situation might serves as a factor or significant world event**
* **Instrumentation: reddit changed top posts algorithm over time give us bias**
* **Maturation: users gets old which less likely to interact on social media**
* **Experimental mortality: most depressed people might not stop posting - user dynamic cannot be measure or just being inactive**
* **Statistical regression: selection bias, need multiple groups instead of reddit**

Statistical regression - to see improvement of poor atheltes over training, and improved is it bc of training? Some people might get lucky when data collection, on that particular measurement they did rlly well; they might be worse with multiple times

→ form of selection bias

1. **Choose an experimental design to rule out plausible threats, assuming you can manipulate all units of analysis and conduct random assignment. Your design should allow you to rule out all plausible confounds.**
2. **Can you do random assignment of subjects? If not, why not?**

**Our project is not random assignment with multiple measurement we can do quasi experiment - such as time series**

**→ pre covid posts is control group and post covid posts is treatment group**

**→ using 9 different mental illness related subreddits to measure**

**No, we have collected the data manually which we cannot control the dataset for our sake**

1. **Given your answer to the previous question, which plausible threats to internal validity are you unable to address?**
2. **Propose a combination of quasi-experimental designs you might use to mitigate the remaining threats to internal validity.**
3. **Discuss how you would rule out other measures of the size of these remaining threats.**

**Time series**

**nonequivalent dependent variable**

* **Group: reddit**
* **9 different subreddit groups**
* **Ex: covid affected not only depression subreddit but also other subreddit as well**
  + **mitigate :**
* **Several observations for time series**
  + **Post content, engagement → post text, num of comments, upvoted ratio**
* **Weakness: sensitive to history, instrumentation**
* **Difficult to generalize tested sample , cross validation is ideal esp. Effect is delayed**
* **X must be specified prior to examine the outcome to avoid attributing effect to change**

1. **Demonstrate your proposed experimental or quasi-experimental design on pilot data.**

**Design would be:**

**O1 → o2 → o3 → o4 → o5 → x → o6 → o7 → o8 → o9 → o10**

**Depression subreddit**

| **num of comments pre covid** | **upvoted ratio pre covid** | **unique id pre covid** | **Post text topic modeling pre covid** | **Post text length pre covid** | **x** | **num of comments post covid** | **upvoted ratio post covid** | **unique id post covid** | **Post text topic modeling post covid** | **Post text length post covid** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

**num of comments pre covid → upvoted ratio pre covid → unique id pre covid → Post text topic modeling pre covid → Post text length pre covid → x → num of comments post covid → upvoted ratio post covid → unique id post covid → Post text topic modeling post covid → Post text length post covid**

**REPEAT this to each subreddits**