ANALYSIS
EXAM REVIEW, MORE HYPOTHESIS
testing when using samples

- Exam review
- More on hypothesis testing with samples


## EXAM

- Wednesday: Exam \#2
- Bring a calculator (no phone etc.)
- Allowed to bring one single-page letter-size ( $8.5 \times 11$ ) sheet with you. Front page only. What you put on it is up to you, but it has to be your own sheet (we'll collect it)
- If you take exam at CDR, please sign up now!
- No new problem set posted this week
- Problem set 7 due on Friday
- Good idea to complete before exam
- No quiz for Wednesday
- Material covered
- Everything from Oct 2 (slides: 09-sampling_2) to Nov 1 (16-hyptest_sample_1)
- For Nov 1 slides:
- Only things on bivariate hypothesis testing
- Hypothesis testing with samples will not be on this exam

RESEARCH PROCESS

- Formulate research question
- Propose explanation/theory, hypotheses
- Data collection process
- Use data to evaluate hypotheses
- Reassess explanation



## STUDY GUIDE

- Random sampling error
- What is the standard error? Where does it come from and how can we compute it?
- What is a 95\% confidence interval? How can we compute it? How do we interpret it?

RECAP
$\mathrm{SE}=\frac{s}{\sqrt{n}}$

- SE: Standard error of the sample mean
- A measure of how much random sampling error we have
- s: sample standard deviation
- n : sample size

RECAP

$$
95 \% \mathrm{Cl}=\bar{x} \pm(1.96 \times \mathrm{SE})
$$

- We draw a large number of random samples from population
- Do the confidence interval for each
- $95 \%$ of those intervals will contain the population mean


## STUDY GUIDE

- Formulate research question
- Propose explanation/theory, hypotheses
- Data collection process
- Use data to evaluate hypotheses
- Reassess explanation



## HYPOTHESES AND THEORY

- Explanation/Theory: (Simplified) description of how social reality works
- Hypotheses: Statements what, if the theory is true, we should observe in our data



## GOOD HYPOTHESES

- Involves two variables
- dependent and independent variable
- Relationship between the variables is clearly specified and measurable
- Unit of analysis is clear
- Hypothesis is testable
- falsifiable


## CAUSALITY

- Most of our theories: relationship between a single cause (independent variable) and a single effect (dependent variable)
- simple"bivariate" relationship (involves 2 variables)


## HURDLES TO CAUSALITY

- Is there a credible causal mechanism that connects X to Y ?
- Can we rule out the possibility that Y could cause X ?
- Is there covariation between $X$ and $Y$ ?
- Have we controlled for all confounding variables $(Z)$ that might make the association between $X$ and $Y$ spurious?


## EMPIRICAL STUDIES

- Two ways to do empirical studies:
- Qualitative, small N
- Quantitative, large $\mathbf{N}$
- $\mathbf{N}=$ number of observations


## QUALITATIVE STUDIES

- Talked about two forms of qualitative studies
- Case study
- Comparative case study


## CASE STUDY

- Key technique: "Process tracing"
- Method to identify the causal relationship in a particular case though detailed examination of each step in the causal chain

COMPARATIVE CASE STUDY

- Method of difference
- Cases where dependent variable is different between cases
- Identify independent variable that is different among cases in the same way as DV is
- Method of agreement
- Cases where dependent variable is same between cases
- Identify independent variable that is also the same among cases


## QUANT AND QUAL

- Strength and weakness of small-n studies relative to large-n studies
- Internal validity
- External validity

QUANT: BIVARIATE RELATIONSHIPS

## Independent Variable



CROSS-TABULATIONS

## Independent Variable

|  | $\begin{gathered} \text { IV } \\ \text { Value } 1 \end{gathered}$ | $\begin{gathered} \text { IV } \\ \text { Value } 2 \end{gathered}$ | Total |
| :---: | :---: | :---: | :---: |
| DV Value 1 | \% In Column (\# Cases) | \% In Column (\# Cases) | \% Of Total (\# In Row) |
| $\begin{gathered} \mathrm{DV} \\ \text { Value } 2 \end{gathered}$ | $\%$ In Column (\# Cases) | \% In Column <br> (\# Cases) | \% Of Total (\# In Row) |
| Total | $\begin{gathered} 100 \% \\ \text { (\# In Column) } \end{gathered}$ | 100\% <br> (\# In Column) | 100\% <br> (\# Total) |

## CROSS-TABULATIONS

## Gender



## TERMINOLOGY

- Zero-order relationship: relationship between two variables, without controlling for any other factors


## ZERO-ORDER RELATIONSHIP

## Gender



## BIVARIATE RELATIONSHIPS

## Independent Variable

> Nominal/Ordinal Interval

## MEAN COMPARISON TABLE

Average of DV
Frequency

| IV <br> Value 1 | Mean of DV for <br> IV Value 1 | \# Cases <br> IV Value 1 |
| :---: | :---: | :---: |
| IV <br> Value 2 | Mean of DV for <br> IV Value 2 | \# Cases <br> IV Value 2 |
| Total | Mean of DV <br> overall | \# Cases <br> overall |

## MEAN COMPARISON

|  | Mean Thermometer <br> Score | Frequency |
| :---: | :---: | :---: |
| Female | 62.8 | 58 |
| Male | 51.6 | 30 |
| Total | 62.4 | 88 |

## ZERO-ORDER RELATIONSHIP

Mean Thermometer<br>Score

Frequency

Female

Male

Total


## BIVARIATE RELATIONSHIPS

## Independent Variable

Nominal/Ordinal Interval

## CORRELATION








## JOE BIDEN



## $r=-0.29$

## DONALD TRUMP


$r=0.61$

## REGRESSION LINE



## REGRESSION LINE



- Intercept: 17


## REGRESSION LINE



- Slope $=$ Rise over run $=2.9 / 1=2.9$


## REGRESSION LINE



- Thermometer Score $=17+2.9$ * Hours/Day


## REGRESSION EQUATION

- Thermometer score $=17+2.9$ * Hours/Day
- General form: $\mathbf{y}=\mathrm{a}+\mathrm{b}$ * x
- y: dependent variable
- a: intercept
- b: slope
- $x$ : independent variable
- $y=a+b$ * $x$
- Interpretation of slope: For every one unit increase in $x, y$ changes by $b$ units
- Interpretation of intercept: When $x=0, y$ takes the value a


## REGRESSION EQUATION

- Thermometer score $=17+2.9$ * Hours/Day
- Interpretation of slope: For every hour that a student studies more per day, their evaluation of the Supreme Court increases by 2.9 points
- Interpretation of intercept: Someone who studies 0 hours per day is expected to have a thermometer score of 17


## EXERCISE

- Thermometer score $=17+2.9$ * Hours/Day
- What is the expected thermometer score of someone who studies 6 hours per day?
- Solution will be on last slides (don't peek)


## EXAM

- What to do and bring
- Be a few minutes early
- Bring calculator and pen/pencil
- Bring your cheat sheet
- Show your steps
- Remember time management
- Write legibly


## STUDY GUIDE

- Questions?


## TODAY

- Exam review
- More on hypothesis testing with samples


## USING SAMPLES

- Bivariate relationship between two variables in sample
- Is this a real relationship that we would find in the population as well, or is it something that only shows up in our sample?


## HYPOTHESIS

- $\mathrm{H}_{0}$ : In the population, there is no relationship between dependent and independent variable - If there is a difference in the sample, it is due to random sampling error
- $H_{A}$ : There is a relationship between the independent and dependent variable in the population


## ERRORS

There Is A Relation In The Population

There Is No Relation In The Population

We Conclude There Is A Relation

We Conclude There Is No Relation

| $\boldsymbol{\sim}$ | $\boldsymbol{x}$ <br> Type I |
| :---: | :---: |
| $\boldsymbol{x}$ |  |
| Type II |  |

## ERRORS

There Is A Relation There Is No Relation In The Population In The Population

We Conclude There Is A Relation

We Conclude There Is No Relation


- We use a procedure designed to ensure we make a Type I error at most $5 \%$ of the time
- We start out thinking $\mathrm{H}_{0}$ is true
- No difference between men and women population
- We have a sample that shows some difference
- Do we reject $\mathrm{H}_{0}$ ?
- If we do, we want to do so wrongly at most $5 \%$ of time
- Ask: If $\mathrm{H}_{0}$ is true (no difference in population), what is the probability ( $p$ ) of observing a difference as large (or greater) as we did in our sample?
- If less than $5 \%$ ( $p<0.05$ ): we reject $\mathrm{H}_{0}$
- If more than $5 \%$ ( $\mathbf{p}>0.05$ ): we don't reject $\mathrm{H}_{0}$


## SIGNIFICANCE

- If less than $5 \%$ ( $p<0.05$ ): we reject $\mathrm{H}_{0}$
- "Statistically significant difference between men and women in support for Biden"
- If more than 5\% ( $p>0.05$ ): we don't reject $\mathrm{H}_{0}$
- "Difference between men and women in support for Biden is not statistically significant"


## REJECTING HO

- So: High bar before we reject $H_{0}$ that $X$ has no effect on $Y$
- We are conservative and need a lot of evidence before we are willing to reject $\mathrm{H}_{0}$


## REJECTING HO





## PODCAST

Longform Podcast
\#315: Elizabeth Kolbert

$>$
00:00

Elizabeth Kolbert, author of Field Notes from a Catastrophe: Man, Nature, and Climate Change and The Sixth Extinction: An Unnatural History, is a staff writer at The New Yorker.

- longform.org/posts/longform-podcast-315-elizabeth-kolbert


## REJECTING HO



PODCAST
Longform Podcast
\#543: Jeff Goodell

Jeff Goodell is a climate change writer for Rolling Stone and the author of seven books. His new book is The Heat Will Kill You First: Life and Death on a Scorched Planet.

- https://longform.org/posts/longform-podcast-543-jeff-goodell

