

CSE 599K

Empirical Research Methods

Winter 2025

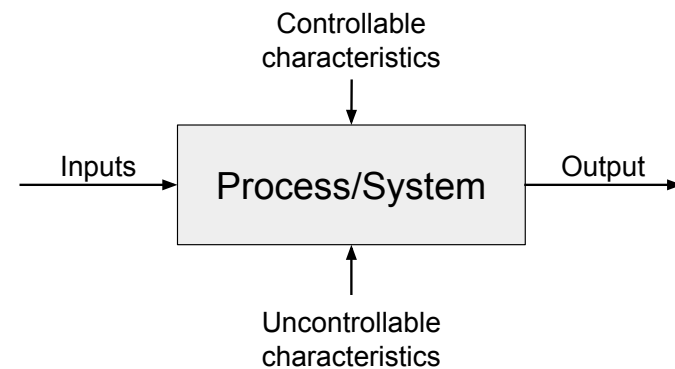
Study design and validity

Today

- Analysis design
- Confirmatory vs. exploratory analyses
- Analysis validity

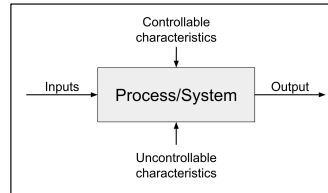
Analysis design

Analysis design: overview



Kinds of variables

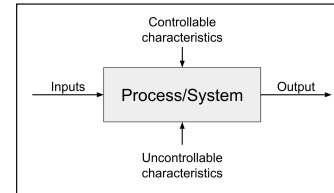
- **Dependent variable**
 - Outcome variable -- the measured response.
- **Independent variable**
 - Experimental variable -- systematically manipulated/controlled.
- **Covariate**
 - Experimental variable -- measurable but not controllable.



What are examples for covariates?

Types of variables

- **Categorical (nominal)**
 - Unordered set of values
 - Example: [HCI, PLSE, Robotics, UbiComp]
- **Dichotomous** (dichotomized or “natural” dichotomy)
 - Categorical with exactly two possible values
 - Example: [Day, Night]
- **Ordinal**
 - Ordered set of values (no assumption about equidistant values)
 - Example: [low, medium, high]
- **Continuous/Interval**
 - Ordered values (equidistant values)
 - Example: [0..100]



Kinds of studies

Experiment

- Independent **variable(s)** are **directly manipulated**/controlled.
- Repeatable with a testable hypothesis.
- Randomization (e.g., counterbalancing for within-subjects designs).

What is a quasi-experiment?

Kinds of studies

Experiment

- Independent **variable(s)** are **directly manipulated**/controlled.
- Repeatable with a testable hypothesis.
- Randomization (e.g., counterbalancing for within-subjects designs).

Observational study

- **Variables** are **not manipulated**/controlled.
- Useful if an experiment is impractical/unethical.
- Greater risk of spurious correlations.

Can you think of an example where an experiment would be impractical/unethical?

Kinds of studies

Experiment

- Independent **variable(s)** are **directly manipulated**/controlled.
- Repeatable with a testable hypothesis.
- Randomization (e.g., counterbalancing for within-subjects designs).

Observational study

- **Variables** are **not manipulated**/controlled.
- Useful if an experiment is impractical/unethical.
- Greater risk of spurious correlations.

Case study

- Focus on one particular subject (“deep dive”).
- Useful for qualitative analyses and interpretation of results.

Study designs

Between subjects design

- Independent variable(s) take on exactly one value for each subject.

Within subjects design

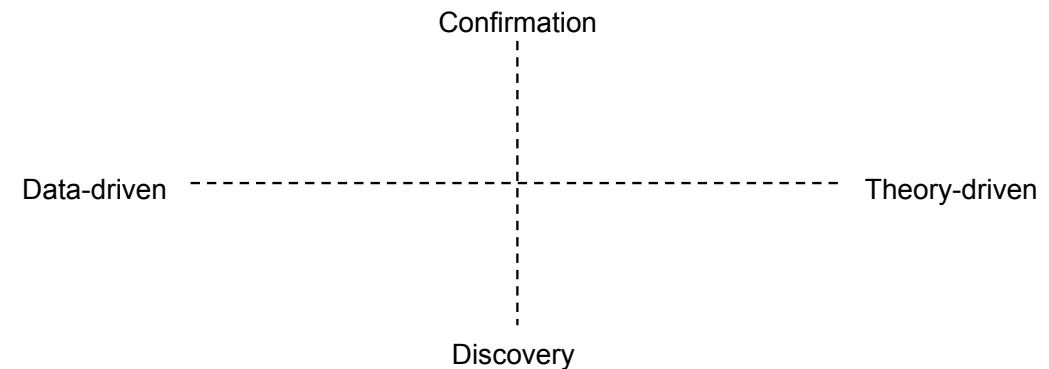
- Independent variable(s) take on multiple/all possible values for each subject.
- Repeated measures design.

Mixed design

- A mixed design of between-subjects variables and within-subjects variables.

Data analysis

Confirmatory vs. exploratory analyses



Data analysis

- Test a hypothesis (once)
- Specify all data collection and analysis aspects in advance
- Preregistration

Confirmation

Data-driven

Theory-driven

- Unknown hypothesis
- Open-ended exploration

Discovery

Data analysis

Confirmation

CDA

Rough CDA

Data-driven

Theory-driven

EDA

Discovery

Data analysis

Confirmatory data analysis (CDA)

- Theory-driven confirmation of a hypothesis
- Pre-specified data analysis

Exploratory data analysis (EDA)

- Theory-driven discovery
- Flexible data analysis
- New hypotheses or models may emerge

Rough CDA

- Theory- and data-driven confirmation of a hypothesis
- Flexible data analysis (researcher degrees of freedom)
- All design decisions and tests are reported

Data analysis

Confirmation

CDA

Rough CDA

Data-driven

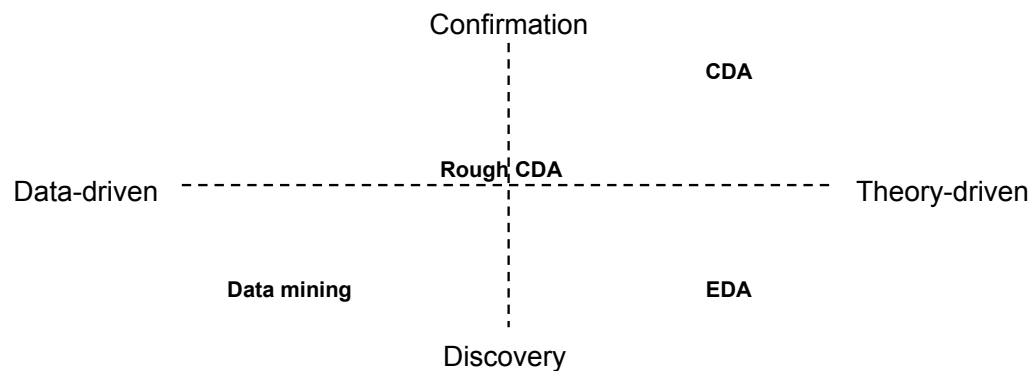
Theory-driven

EDA

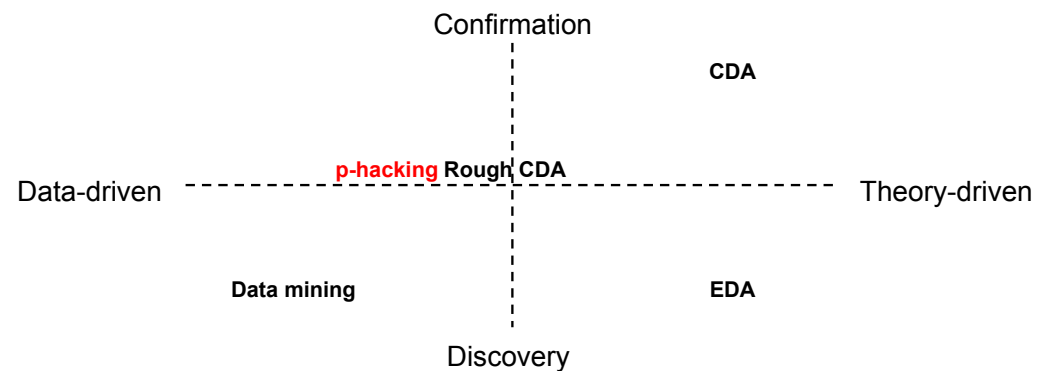
Discovery

How/where does data mining fit in?

Data analysis



Data analysis: the dark side



Data analysis: the dark side



Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY: Republicans Democrats

2 DEFINE TERMS: Which politicians do you want to include?
 Presidents
 Governors
 Senators
 Representatives

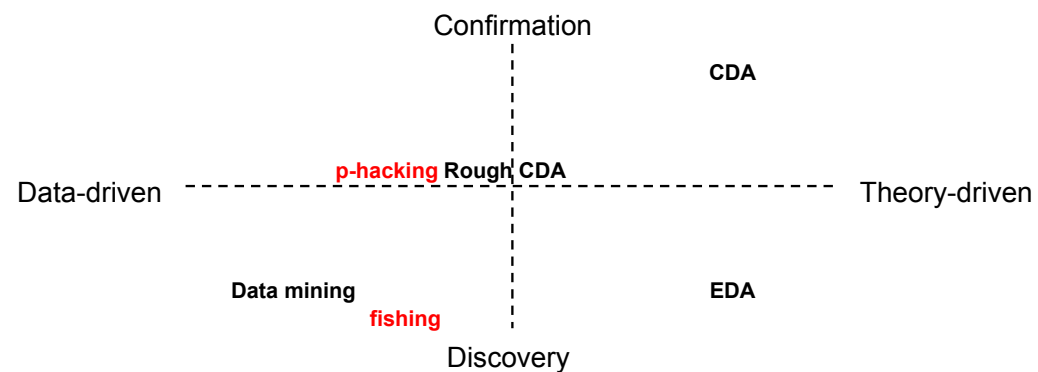
3 IS THERE A RELATIONSHIP? Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.

4 IS YOUR RESULT SIGNIFICANT? If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.

Result: Almost
 Your 0.06 p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

Data from The Quahogian Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Data analysis: the dark side



Data analysis: the dark side

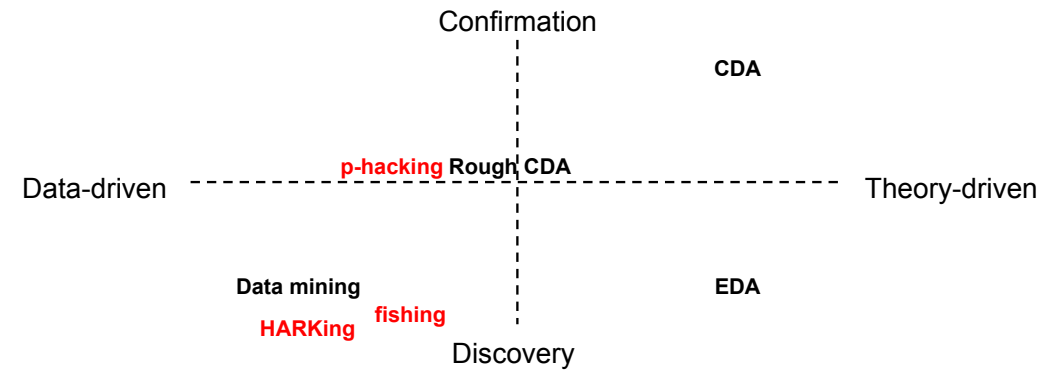


Our shocking new study finds that ...

| EATING OR DRINKING | IS LINKED TO | P-VALUE |
|------------------------|--|---------|
| Raw tomatoes | Judaism | <0.0001 |
| Egg rolls | Dog ownership | <0.0001 |
| Energy drinks | Smoking | <0.0001 |
| Potato chips | Higher score on SAT math vs. verbal | 0.0001 |
| Soda | Weird rash in the past year | 0.0002 |
| Shellfish | Right-handedness | 0.0002 |
| Lemonade | Belief that "Crash" deserved to win best picture | 0.0004 |
| Fried/breaded fish | Democratic Party affiliation | 0.0007 |
| Beer | Frequent smoking | 0.0013 |
| Coffee | Cat ownership | 0.0016 |
| Table salt | Positive relationship with Internet service provider | 0.0014 |
| Steak with fat trimmed | Lack of belief in a god | 0.0030 |
| Iced tea | Belief that "Crash" didn't deserve to win best picture | 0.0043 |
| Bananas | Higher score on SAT verbal vs. math | 0.0073 |
| Cabbage | Innie bellybutton | 0.0097 |

SOURCE: FFQ & FIVETHIRTYEIGHT SUPPLEMENT

Data analysis: the dark side



Analysis validity

External, internal, and construct validity

External validity

- Does the experiment generalize (to larger population, other subjects, etc.)?
- How representative is the sample?
- Be aware of **WEIRD** subjects!
 - For example: studying mostly **Western, Educated** people from **Industrialized, Rich, and Democratic** countries.



vs



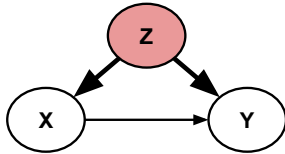
External, internal, and construct validity

External validity

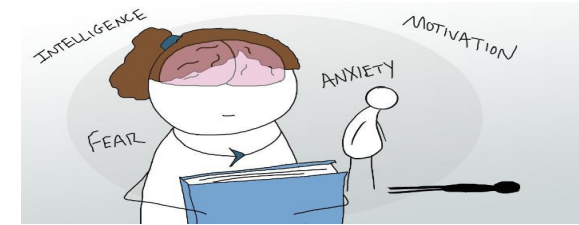
- Does the experiment generalize (to larger population, other subjects, etc.)?
- How representative is the sample?

Internal validity

- Does the experiment isolate the variable(s) of interest?
- Does the experiment control for confounders and unwanted effects?
- Be aware of **carry-over effects** (within-subjects designs)!
 - For example: order of tasks (subjects get accustomed to or tired of a task).



External, internal, and construct validity



Construct validity

- Does the experiment measure what it claims to measure?
- Do the proxy measures and tools adequately measure the concept of interest?
- Be aware of **interactions (being tested vs. treatment) and bias!**
 - For example: subjects may perform better/worse under test conditions.

External, internal, and construct validity

External validity

- Does the experiment generalize (to larger population, other subjects, etc.)?
- How representative is the sample?

Internal validity

- Does the experiment isolate the variable(s) of interest?
- Does the experiment control for confounders and unwanted effects?

Construct validity

- Does the experiment measure what it claims to measure?
- Do the proxy measures and tools adequately measure the concept of interest?

Statistical concepts

(Statistical) conclusion validity

- Are the conclusions valid based on the chosen statistical test and sample size?
- Are the conclusions valid based on the observed significance (p value)?

Types of errors

- Type I error (false positive): rejecting a true null hypothesis
- Type II error (false negative): not rejecting a false null hypothesis

Analysis validity: open discussion

External validity

- Does the experiment generalize (to larger population, other subjects, etc.)?
- How representative is the sample?

Internal validity

- Does the experiment isolate the variable(s) of interest?
- Does the experiment control for confounders and unwanted effects?

Construct validity

- Does the experiment measure what it claims to measure?
- Do the proxy measures and tools adequately measure the concept of interest?

(Statistical) conclusion validity

- Are the conclusions valid based on the chosen statistical test and sample size?
- Are the conclusions valid based on the observed significance (p value)?