CSE P 590

Building Data Analysis Pipelines

Fall 2024

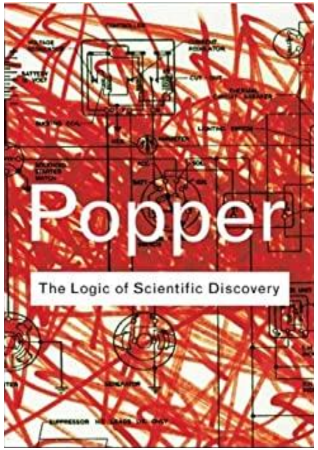
Analysis Design and Validity

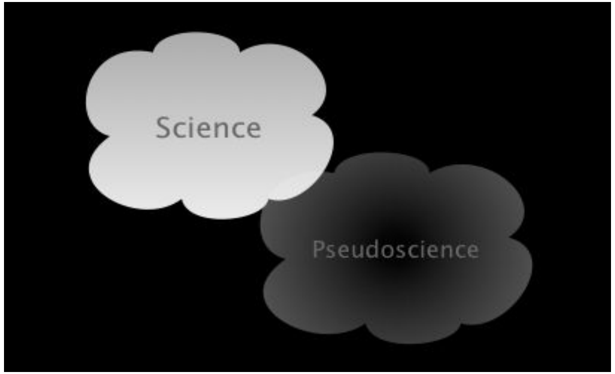
Today

- Objectivity in science
- Analysis design
- Confirmatory vs. exploratory analyses
- Analysis validity
- In-class exercise 1: R basics

Objectivity in science

The holy grail: objectivity in science

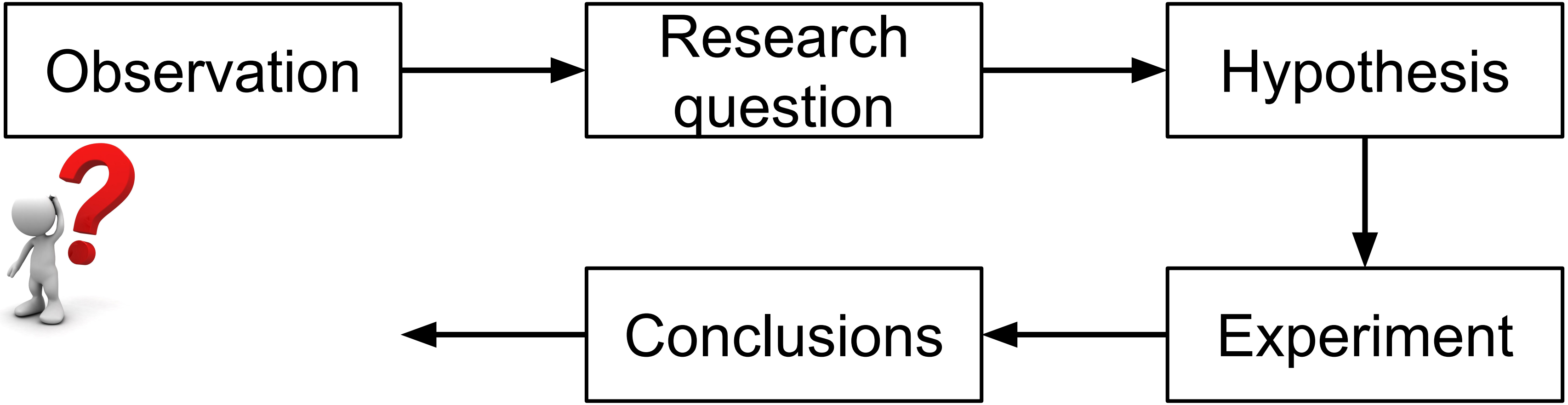




The holy grail: objectivity in science

Falsifiability and NHST are the solution, right?

- Scientific method: rigorous framework and easy to execute



Observation

Research question

Hypothesis

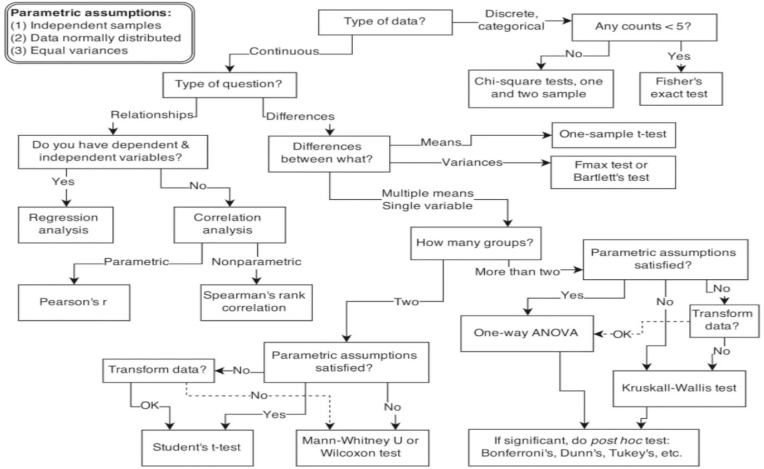
Experiment

Conclusions

The holy grail: objectivity in science

Falsifiability and NHST are the solution, right?

- Scientific method: rigorous framework and easy to execute
- Agreed-upon analysis methods and selection criteria



Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances

Type of data?
Discrete, categorical
Any counts < 5?
No
Chi-square tests, one and two sample
Yes
Fisher's exact test
Continuous
Type of question?
Relationships
Do you have dependent & independent variables?
Yes
Regression analysis
No
Correlation analysis
Parametric
Pearson's r
Nonparametric
Spearman's rank correlation
Differences
Differences between what?
Means
One-sample t-test
Variances
Fmax test or Bartlett's test
Multiple means Single variable
How many groups?
More than two
Parametric assumptions satisfied?
Yes
One-way ANOVA
No
Transform data?
OK
No
Kruskall-Wallis test
If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.
Two
Parametric assumptions satisfied?
No
Transform data?
OK
No
Yes
Student's t-test
No
Mann-Whitney U or Wilcoxon test

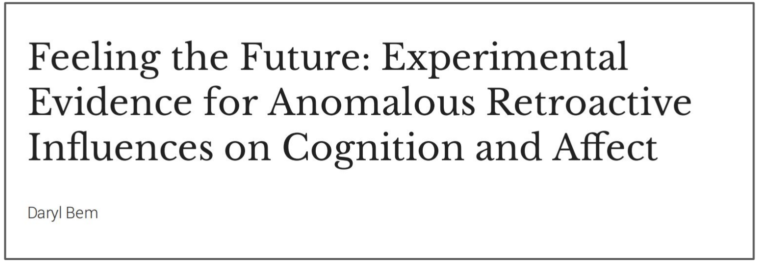
The holy grail: objectivity in science

Falsifiability and NHST are the solution, right?

- Scientific method: rigorous framework and easy to execute
- Agreed-upon analysis methods and selection criteria
- Mechanical and dichotomous decision making (p<0.05)

OH SO YOU READ IT IN A SCIENCE MAGAZINE?
TELL ME MORE ABOUT THE OBJECTIVITY AND INFALLIBILITY OF SCIENCE
quickmeme.com

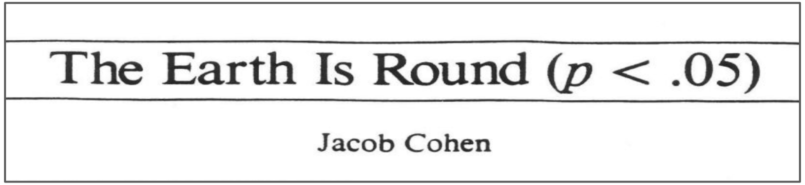
The holy grail: objectivity in science



Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

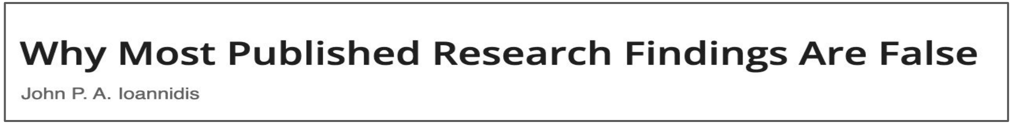
Daryl Bem

The holy grail: objectivity in science



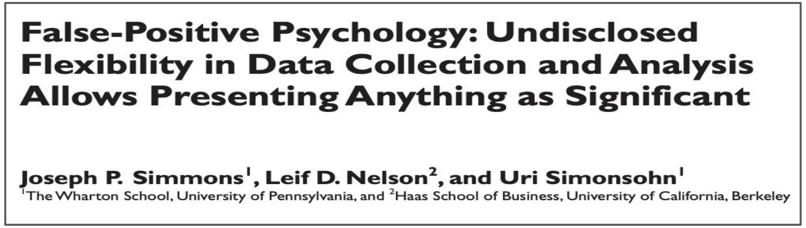
The Earth Is Round ($p < .05$)

Jacob Cohen



Why Most Published Research Findings Are False

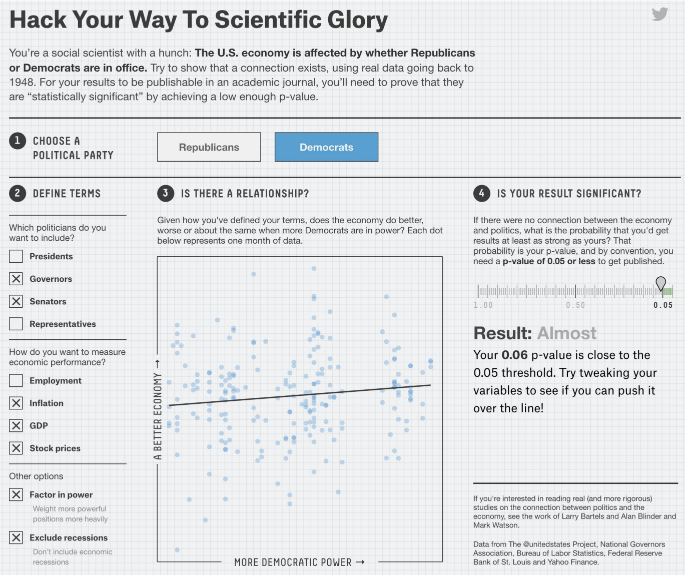
John P. A. Ioannidis



False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

The holy grail: objectivity in science



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**

How do you want to measure economic performance?

- □ **Employment**
- ☒ **Inflation**
- ☒ **GDP**
- ☒ **Stock prices**

Other options

- ☒ **Factor in power**
  Weight more powerful positions more heavily
- ☒ **Exclude recessions**
  Don't include economic recessions

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!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

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

[https://projects.fivethirtyeight.com/p-hacking]

Operationalization introduces subjectivity!

Science is subjective

Transparency and replication go a long way

Science is subjective

Science is subjective: ethics

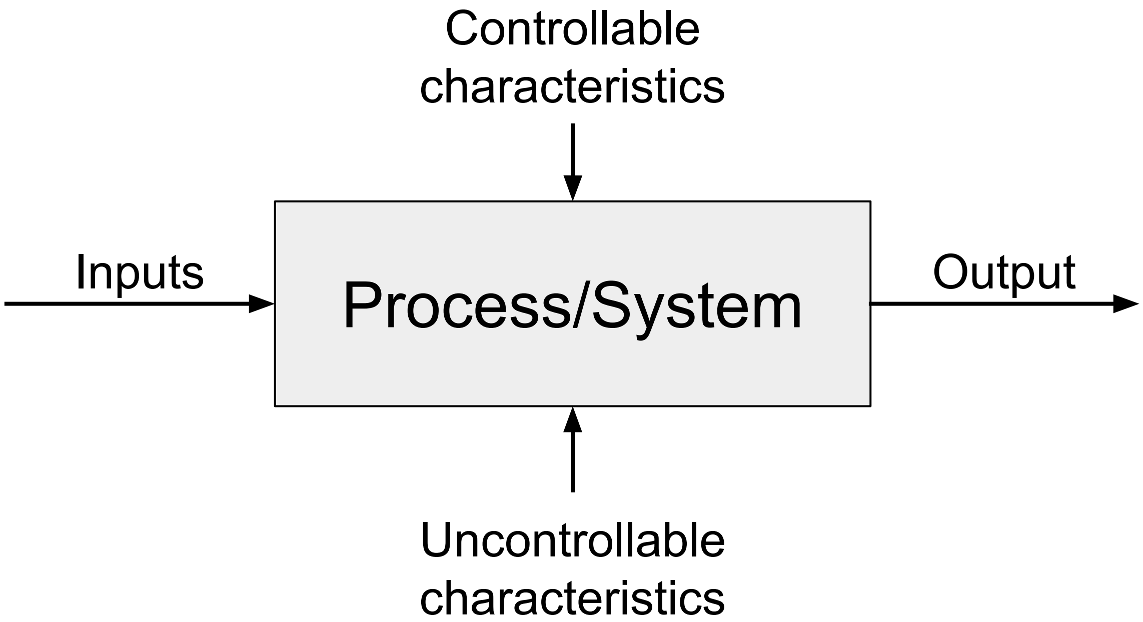
Four core values (e.g., APA's ethics framework)

- Risks and benefits
  - Do benefits outweigh risks?
- Responsibility and integrity
  - Representation of a scientific field
  - Public trust
- Justice and fairness
  - No biased selection of control/treatment
- Rights, and dignity
  - Awareness and consent
  - Privacy
  - Debriefing

This framework does not cover rigor and validity!

Analysis design

Analysis design: overview



Controllable characteristics

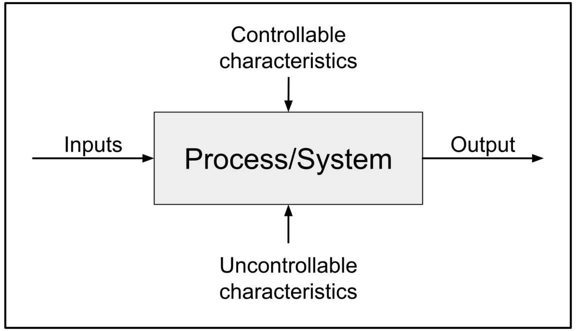
Inputs

Process/System

Output

Uncontrollable characteristics

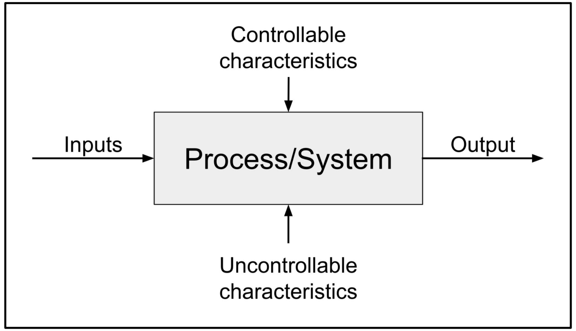
Kinds of variables



- **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?

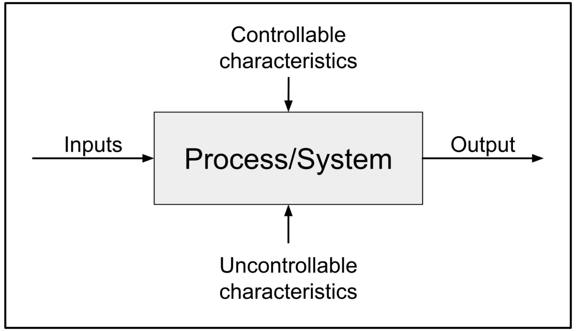
Types of variables



What other types of variables do we frequently encounter?

- **Continuous/Interval**
  - Ordered values (equidistant values)
  - Example: [0..100]

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.

Confirmatory vs. exploratory analyses

Data analysis

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

Confirmation

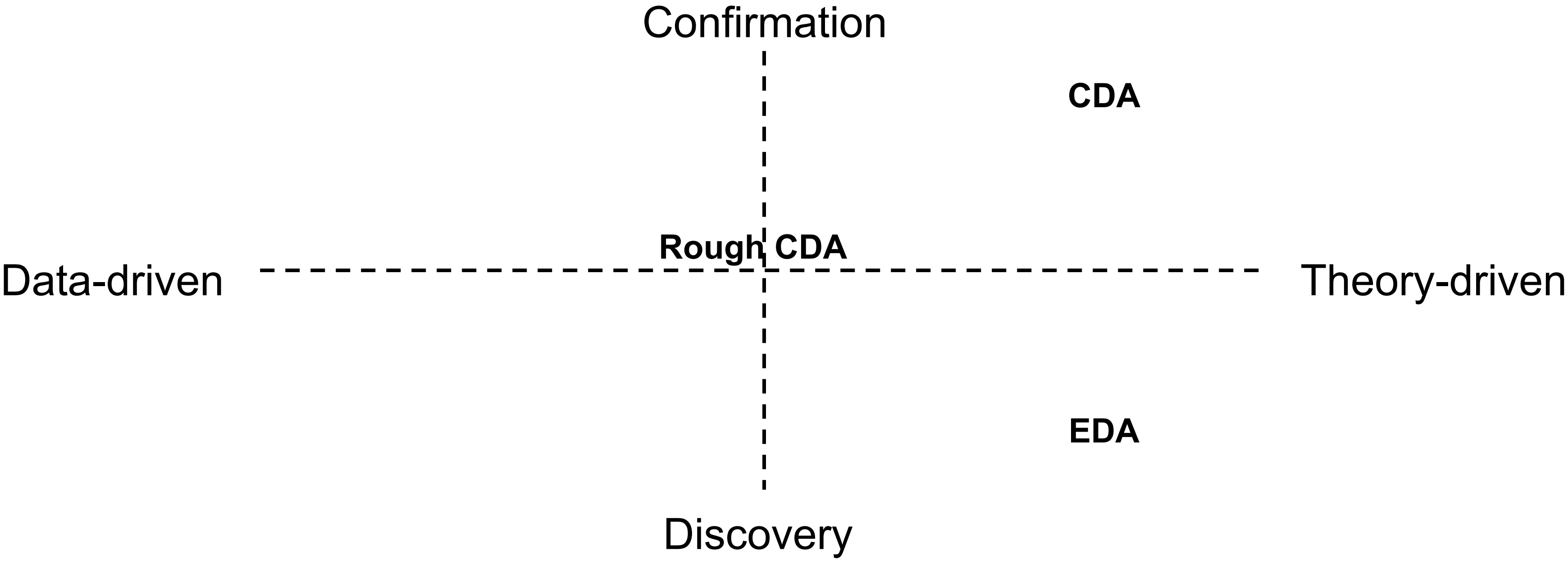
Data-driven

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

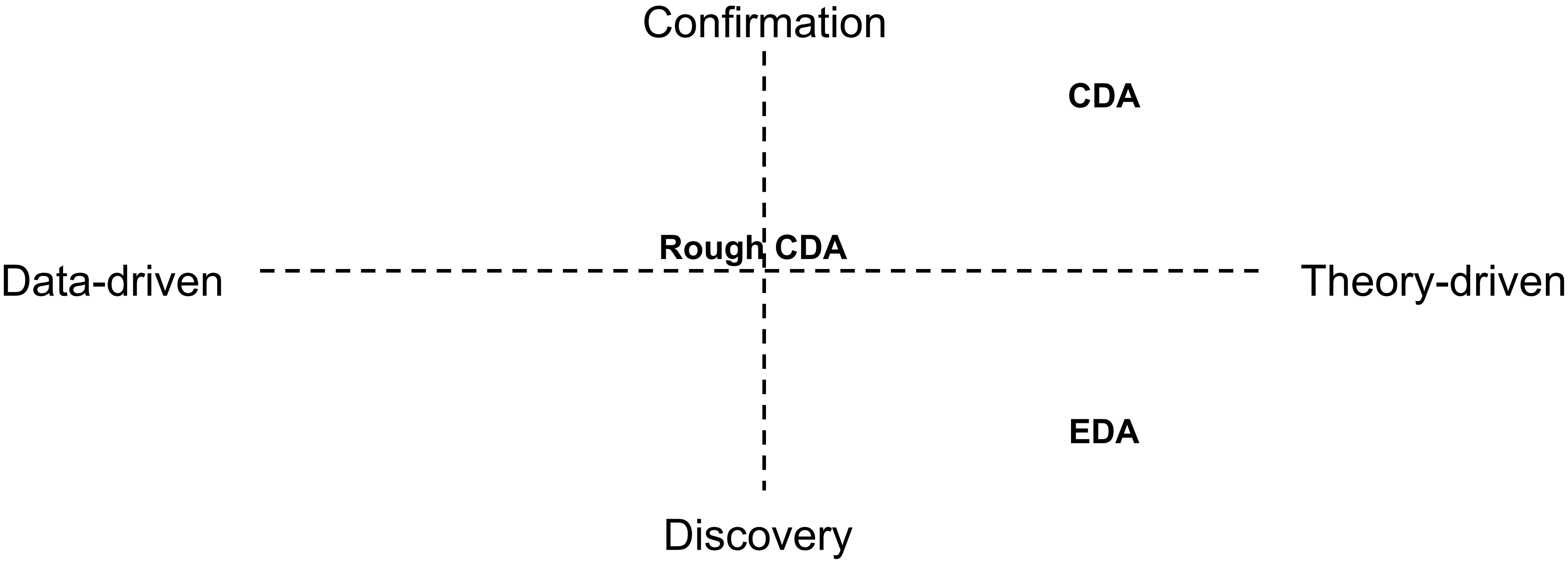
Exploratory data analysis (EDA)

- 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

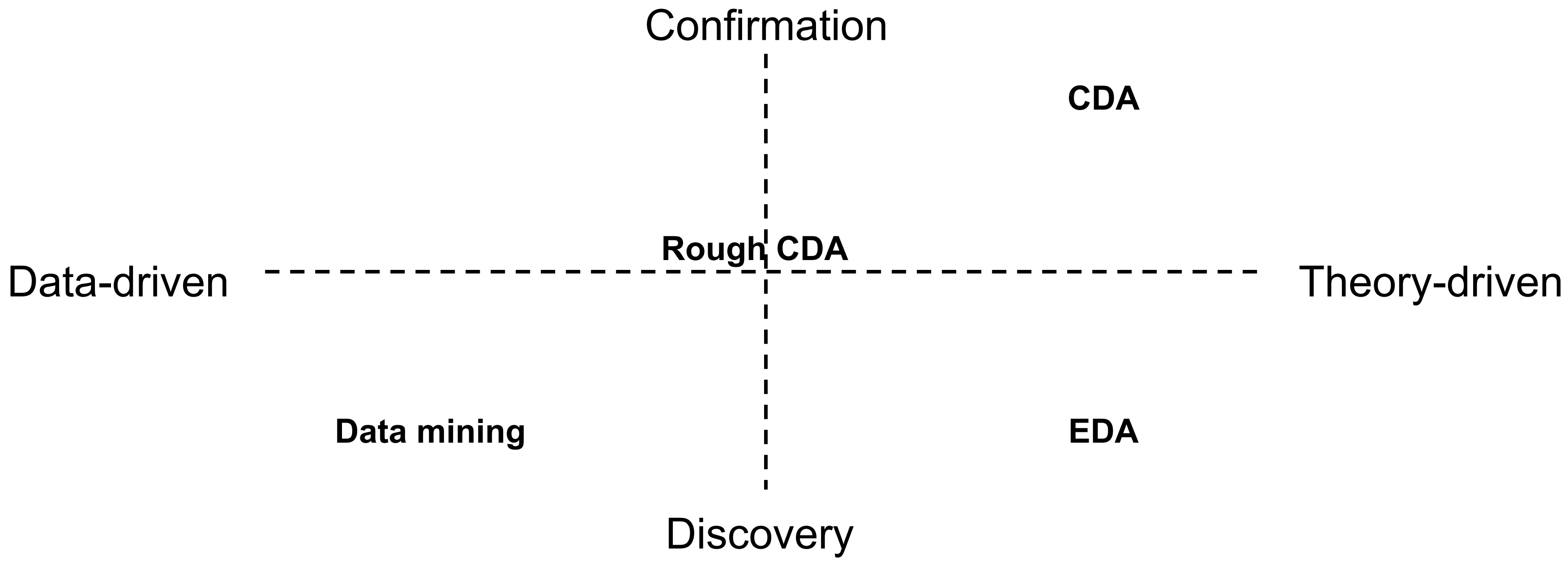
Theory-driven

EDA

Discovery

How/where does data mining fit in?

Data analysis



Confirmation

CDA

Rough CDA

Data-driven

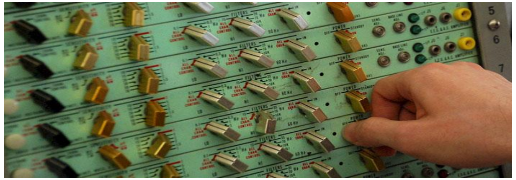
Theory-driven

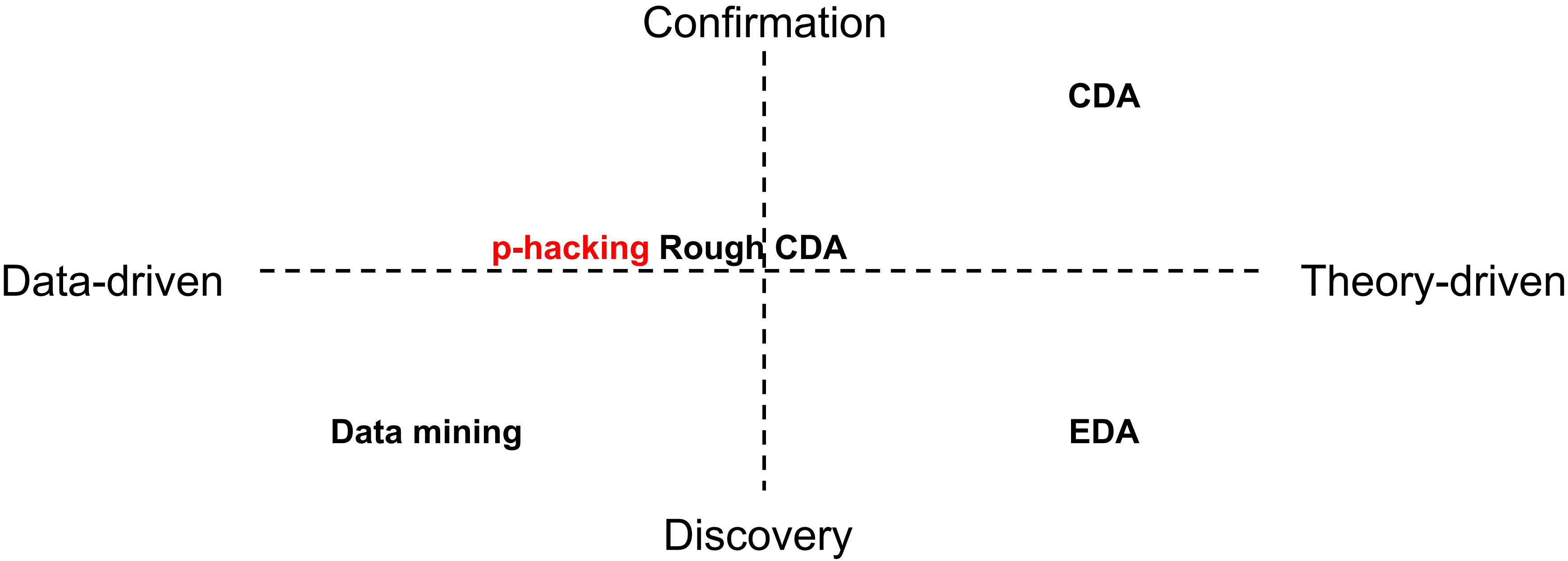
Data mining

EDA

Discovery

Data analysis: the dark side





Confirmation

CDA

p-hacking Rough CDA

Data-driven

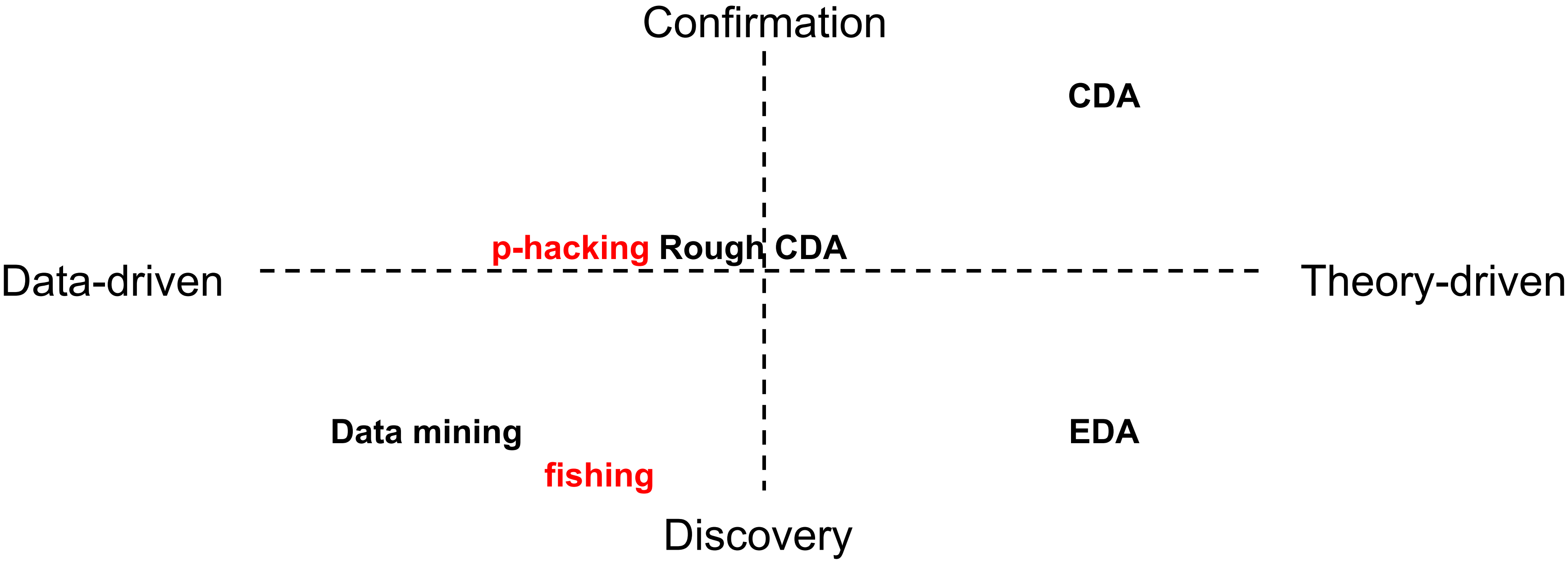
Theory-driven

Data mining

EDA

Discovery

Data analysis: the dark side



Confirmation

CDA

p-hacking Rough CDA

Data-driven

Theory-driven

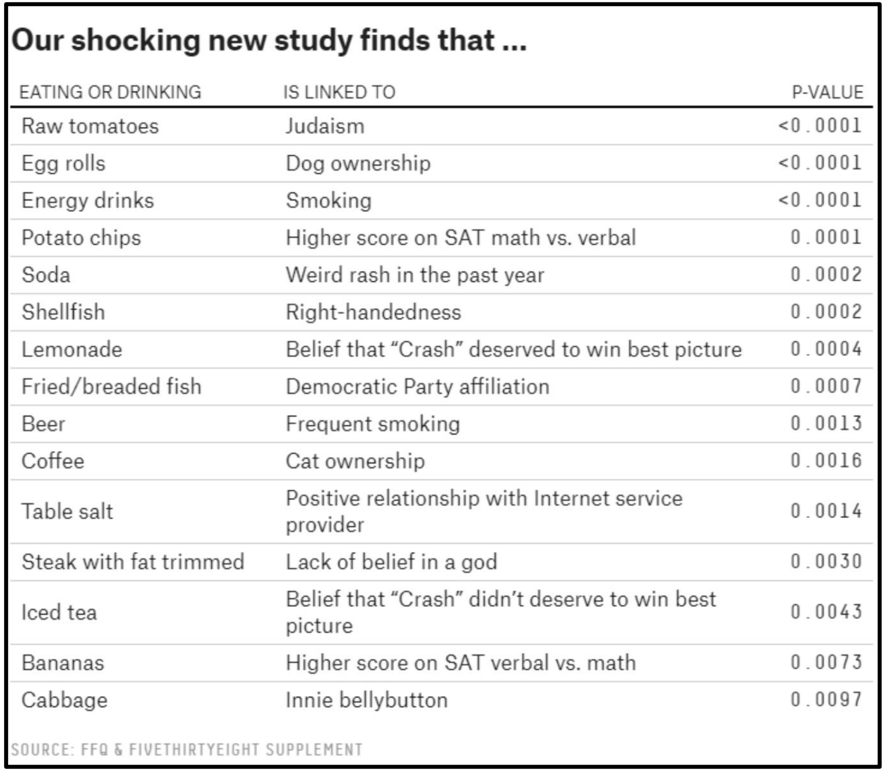
Data mining

fishing

EDA

Discovery

Data analysis: the dark side

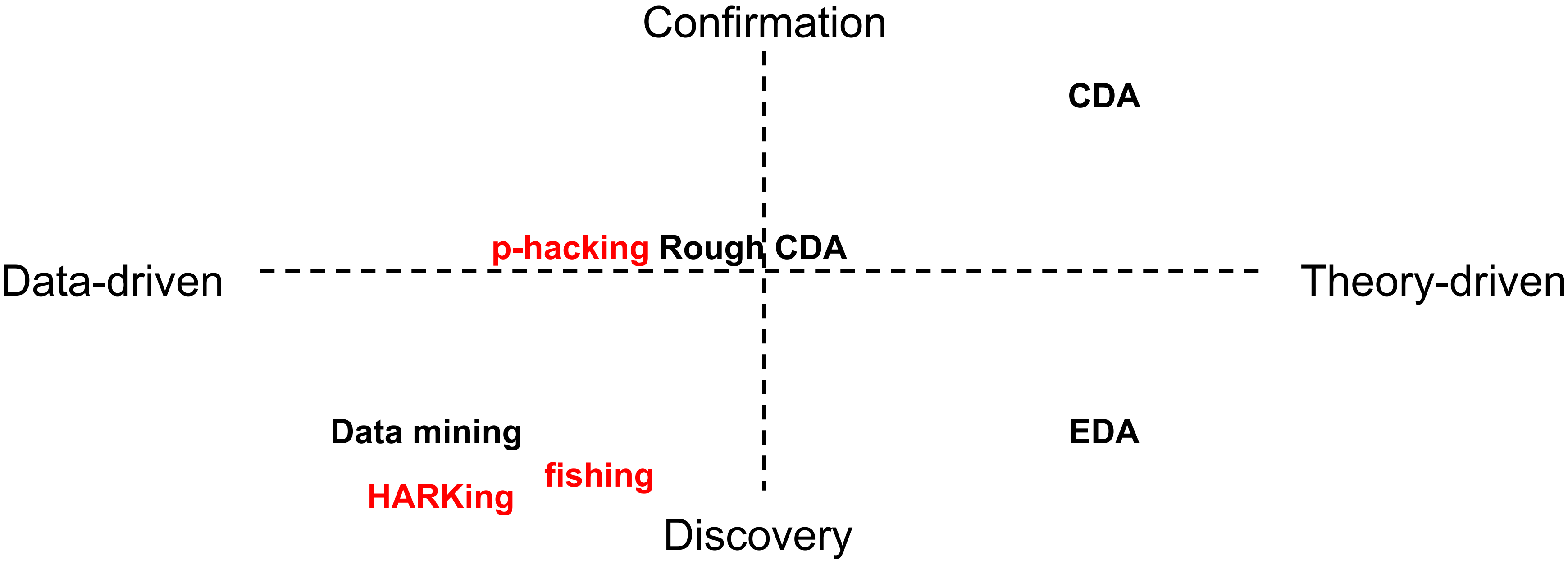


Our shocking new study finds that ...

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



Confirmation

CDA

p-hacking Rough CDA

Data-driven

Theory-driven

Data mining

EDA

fishing

HARKing

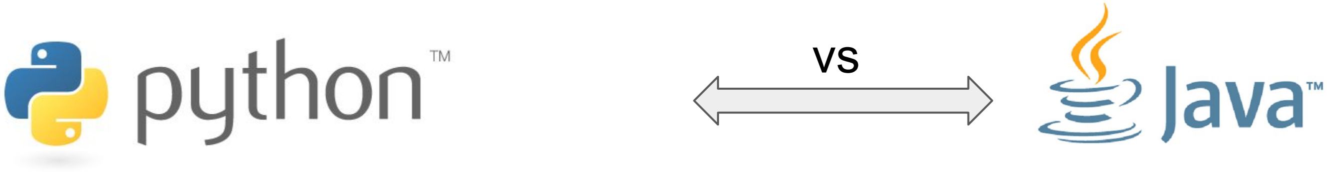
Discovery

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.



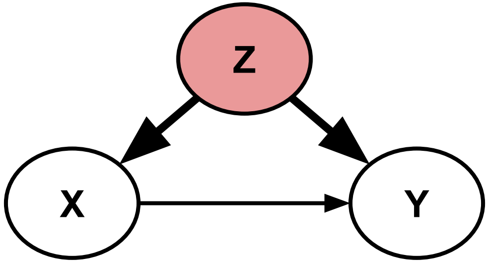
External, internal, and construct validity

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 tiered 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)?

In-class exercise 1: R basics