Research Methods for Economics and Policy
Data and Methods

Lorenzo Neri[1]

[1]University of Rome “Tor Vergata”, CESifo and IZA

April-May 2025

What are we going to do?

1. Data and methods: different types of data
2. Research replicability
3. Data and methods: correlation vs causation, pre-registration, power analysis, etc
4. Research credibilty

Data sources

Data can be broadly categorised into primary and secondary data sources:

- Survey data (P)
- Experiments (P)
- Field and register data (S)
- Survey or exprimental data collected by opthers (S)

Depending on the type of data, access can be public or (partially or fully) restricted

- Particularly when thinking about a potential master thesis, you may want to consider datasets that are readily available
- Be creative: you can also construct a new dataset by putting together data from publicly available sources (e.g., webscraping)

Survey data and online repositories

Survey data typically cover a subsample of a population (e.g., Italy) and often cover different topics

- Examples: GSOEP (Germany), MCS, BHPS/US (UK), Labor Force Survey (EU), SHIW (Italy), World Value Survey, World Management Survey
- Advantage: cover a wealth of different topics, potentially in a very detailed way. Also, often they have de-identified versions that are available relatively easily for everybody
- Disadvantage: coverage can be an issue, depending on the study/methods

You can also create your own survey, of course; or, cooperate with existing organisations and include questions you are interested in

National statistics offices (e.g., Istat) and International organizations (e.g., World Bank, IMF, OECD, WTO) regularly publish and maintain series of datasets online

Administrative datasets

Administrative datasets are typically produced and maintain by institutions for different purposes

- National Pupil Database (England & Wales): population of children in state schools in England and Wales: observe characteristics, residence, school, test scores, etc
- Business Structure Database (UK): population of British companies: info on type, revenues, employment
- Land Registry (UK): Population of house transactions
- INPS Archive (Italy): social security system, pensions, labour market, income support measures and more generally welfare related issues
- Often not always: you have the population, but not as much detail as survey data

Other sources

Private companies sometimes maintain datasets that can purchased

- E.g., data on rents: Idealista (Italy and Spain), Rightmove (UK), Zoopla (UK)

Web-scraped data - need to be mindful of websites' policies though. Scraped data have become quite common...

Journals require authors to publish their data online. A great source of data to explore related questions

Online pre-registration sites

Online data resources: miscellanea

- National statistics offices, international organizations (worlds bank, IMF, OECD Statistics Portal, WTO trade statistics)
- Online panels/survey (GSOEP, LFS, SHIW, BHPS/US). Often they have different versions with different levels of access. For instance, many of them may have de-identified versions available relatively easily. Many of these websites also keep repositories of studies done with the data, which can be a source of inspiration.
- Large surveys such as the World Value Survey, World bank surveys, World Management Survey
- Lots of data can be scraped from the web + consider replication packages
- You can also create your own survey and/or add questions to existing surveys (the latter is unfeasible for your proposal!)

Online data resources: development economics

- Afrobarometer: https://www.afrobarometer.org/. Nationally representative attitudinal survey for 37 African countries.
- International Household Survey Network (IHSN): https://catalog.ihsn.org/index.php/catalog. Huge catalog of data from household surveys and censuses, mainly for developing countries. Provides basic information on the surveys as well as links and contact information
- J-pal dataverse: https://dataverse.harvard.edu/dataverse/jpal. Over 100 datasets from RCTs in developing countries.
- D-place: https://d-place.org/contributions. This is a great source for cultural datasets (a database of Places, Language, Culture, and Environment).

*From U Goteborg's course

Online data resources: other examples

- UK Censuses and other aggregate statistics: https://www.nomisweb.co.uk/
- FRED Economic Data: https://fred.stlouisfed.org/. US and international time series from various data sources.
- Oppportunity Insights: https://www.opportunityatlas.org/. Neighborhood statistics for the US
- Istat Databank: https://www.istat.it/en/data/databases/
- London Development Database: https://apps.london.gov.uk/land-development-database/
- FBREF Football data: https://fbref.com/en/
- A great list by Pietro Biroli: resources

Research replicability

Research replicability

- Research replicability is important to prevent misconduct (!), but more broadly to improve scientific credibility and make sure that policy decisions and interventions are not taken based on "false" findings

Research replicability

- Research replicability is important to prevent misconduct (!), but more broadly to improve scientific credibility and make sure that policy decisions and interventions are not taken based on "false" findings

- A recent example: the GDRI scandal uncovered by the Institute4Replication

Research replicability

- Research replicability is important to prevent misconduct (!), but more broadly to improve scientific credibility and make sure that policy decisions and interventions are not taken based on "false" findings
- A recent example: the GDRI scandal uncovered by the Institute4Replication
- A famous example: Reinhart and Rogoff, "Growth in a time of debt" (AER, 2010)
  - ⇒ Economic growth slows down when the debt/GDP ratio exceeds the threshold of 90% of GDP
  - ⇒ These results were powerful in a period where governments around the world were slashing spending to decrease public deficit and stimulate economic growth

Research replicability

- Research replicability is important to prevent misconduct (!), but more broadly to improve scientific credibility and make sure that policy decisions and interventions are not taken based on "false" findings
- A recent example: the GDRI scandal uncovered by the Institute4Replication
- A famous example: Reinhart and Rogoff, "Growth in a time of debt" (AER, 2010)
  - ⇒ Economic growth slows down when the debt/GDP ratio exceeds the threshold of 90% of GDP
  - ⇒ These results were powerful in a period where governments around the world were slashing spending to decrease public deficit and stimulate economic growth
- See (amongst others) the LSE Blog coverage of this case: Why we need open data in economics

Research replicability: Reinhart and Rogoff (2010)

- Thomas Herndon, Michael Ash and Robert Pollin from UMass tried to replicate the results of Reinhart and Rogoff and uncovered the following issues:

Research replicability: Reinhart and Rogoff (2010)

- Thomas Herndon, Michael Ash and Robert Pollin from UMass tried to replicate the results of Reinhart and Rogoff and uncovered the following issues:

  → **Coding errors**: due to a spreadsheet error five countries were excluded completely from the sample resulting in significant error of the average real GDP growth and the debt/GDP ratio in several categories

Research replicability: Reinhart and Rogoff (2010)

- Thomas Herndon, Michael Ash and Robert Pollin from UMass tried to replicate the results of Reinhart and Rogoff and uncovered the following issues:

  → **Coding errors**: due to a spreadsheet error five countries were excluded completely from the sample resulting in significant error of the average real GDP growth and the debt/GDP ratio in several categories

  → **Selective exclusion** of available data and data gaps: Reinhart and Rogoff exclude Australia (1946-1950), New Zealand (1946-1949) and Canada (1946-1950). This exclusion is alone responsible for a significant reduction of the estimated real GDP growth in the highest public debt/GDP category

Research replicability: Reinhart and Rogoff (2010)

- Thomas Herndon, Michael Ash and Robert Pollin from UMass tried to replicate the results of Reinhart and Rogoff and uncovered the following issues:
  - → **Coding errors**: due to a spreadsheet error five countries were excluded completely from the sample resulting in significant error of the average real GDP growth and the debt/GDP ratio in several categories
  - → **Selective exclusion** of available data and data gaps: Reinhart and Rogoff exclude Australia (1946-1950), New Zealand (1946-1949) and Canada (1946-1950). This exclusion is alone responsible for a significant reduction of the estimated real GDP growth in the highest public debt/GDP category
  - → **Unconventional weighting** of summary statistics: the authors do not discuss their decision to weight equally by country rather than by country-year, which could be arbitrary and ignores the issue of serial correlation.
- Implications:
  - → Countries with high levels of public debt experience only “modestly diminished” average GDP growth rates
  - → There is a wide range of GDP growth performances at every level of public debt among the twenty advanced economies in the survey of Reinhart and Rogoff

The replication crisis

- There's evidence of a high frequency of "false positives" generated and published in academic journals.
- False positive: finding a statistically significant effect when there is none (i.e., non-replicable results)
- Example: testing whether a new drug A saves more patients than existing drug B and erroneously finding that it does (due to benign or malign reasons)

The replication crisis

- There's evidence of a high frequency of "false positives" generated and published in academic journals.
- False positive: finding a statistically significant effect when there is none (i.e., non-replicable results)
- Example: testing whether a new drug A saves more patients than existing drug B and erroneously finding that it does (due to benign or malign reasons)
- How many false positives should we expect applying standard frequentist hypothesis testing with an alpha of 0.05?

The replication crisis

- There's evidence of a high frequency of "false positives" generated and published in academic journals.
- False positive: finding a statistically significant effect when there is none (i.e., non-replicable results)
- Example: testing whether a new drug A saves more patients than existing drug B and erroneously finding that it does (due to benign or malign reasons)
- How many false positives should we expect applying standard frequentist hypothesis testing with an alpha of 0.05?
- What do you think we see empirically?

Terminology

- **Reproduction (reproducibility)**: Can published results be reproduced based on the same methods, code, and data?

Terminology

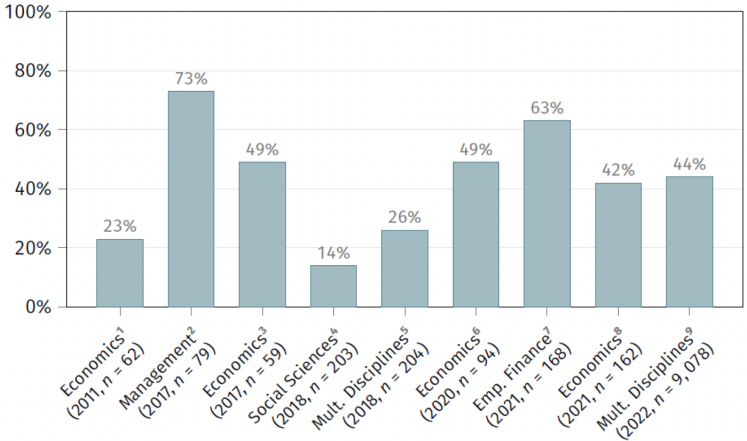
- **Reproduction (reproducibility)**: Can published results be reproduced based on the same methods, code, and data?

- **Direct Replication (replicability)**: Do results replicate if run the same experiment in the same way as the original experiment, i.e., ideally using exactly the same materials and software as in the original study?

Terminology

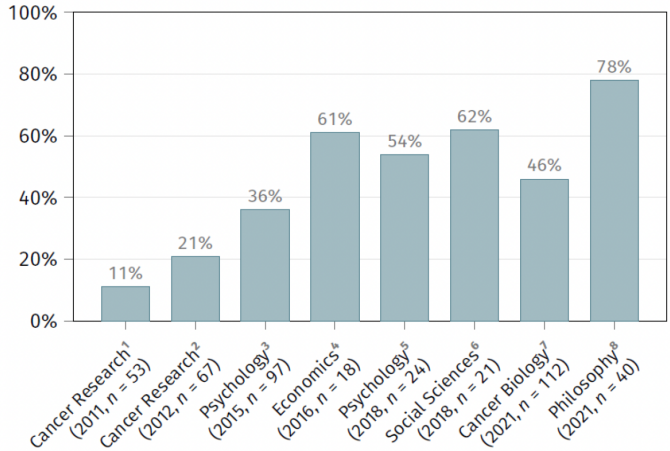
- **Reproduction (reproducibility)**: Can published results be reproduced based on the same methods, code, and data?
- **Direct Replication (replicability)**: Do results replicate if run the same experiment in the same way as the original experiment, i.e., ideally using exactly the same materials and software as in the original study?
- **Conceptual replication**: testing the same hypothesis as in the original study but using a different method/design?

Reproduction rates across disciplines



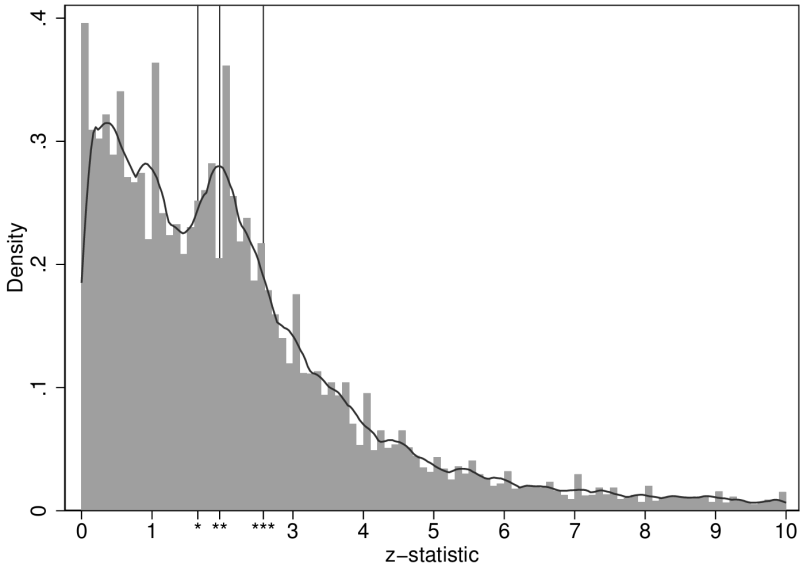
[1] McCullough et al. (2006), [2]Bergh et al. (2017), [3] Chang and Li (2017), [4] Gertler et al. (2018), [5]Stodden et al. (2018), [6]Villhuber (2020), [7]Pérignon et al. (2021), [8]Herbert et al. (2021), [9]Trisovic et al. (2022).

Replication rates across disciplines

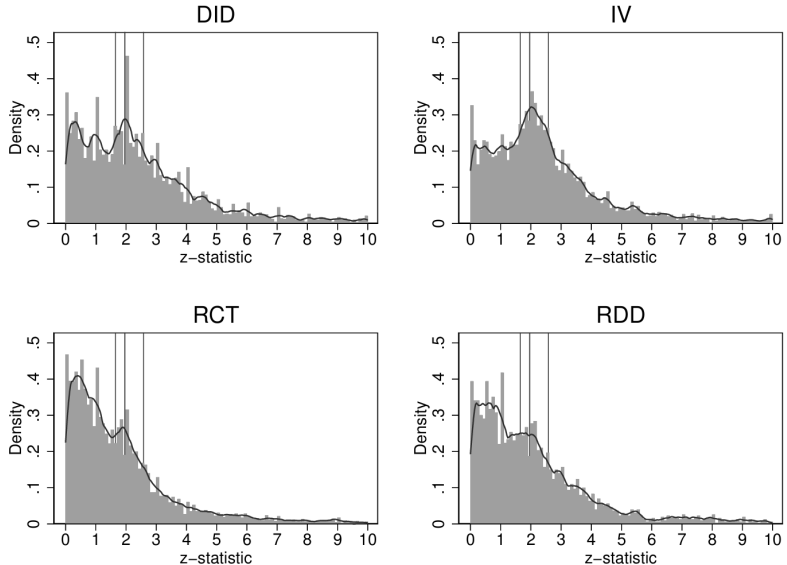


[1]Begley and Ellis (2011), [2] Prinz et al. (2012), [3]Open Science Collaboration (2015), [4]Camerer et al. (2015), [5]Klein et al. (2018), [6]Camerer et al. (2018), [7]Errington et al. (2021), [8]Cova et al. (202:

P-hacking (Brodeur, Cook, and Heyes, AER)



P-hacking (Brodeur, Cook, and Heyes, AER)



Why so many false positives?

- Technical factors (e.g., natural variability)
- Study design and inapt statistical practices (e.g., low power, multiple testing, p-hackingb. . .)
- Human factors (e.g., errors)
- Incentives (e.g., fraud, publication bias, selective reporting. . .)
- Institutional factors (e.g., paywalled access)

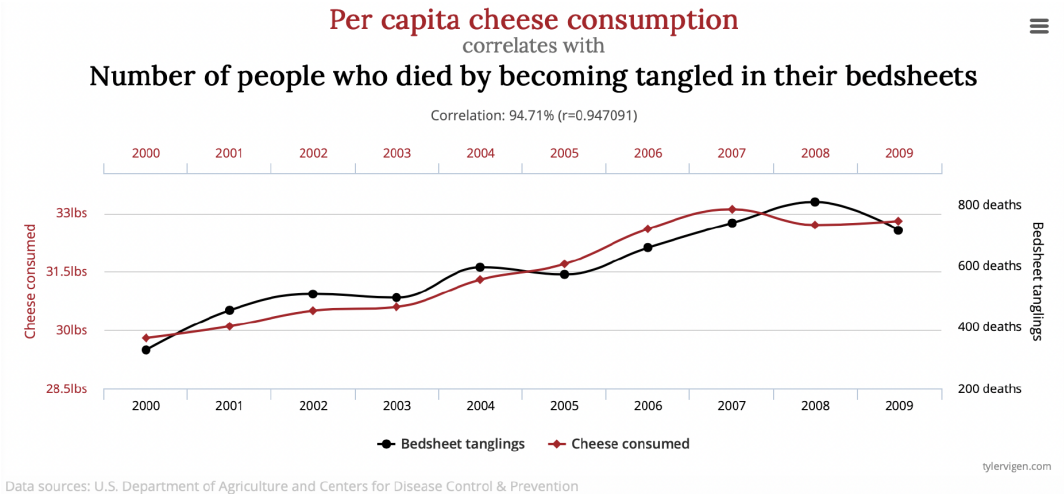
Correlation vs Causation

Correlation vs Causation

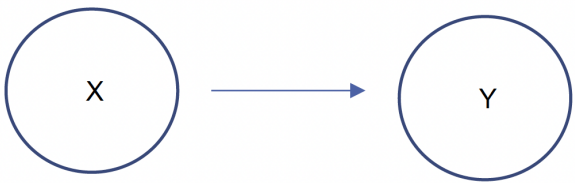
$\Rightarrow$ People who get a university education tend to earn more

$\Rightarrow$ Getting a university education will cause you to earn more

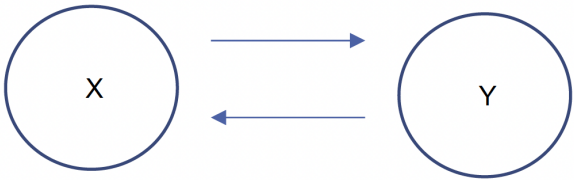
Spurious Correlation



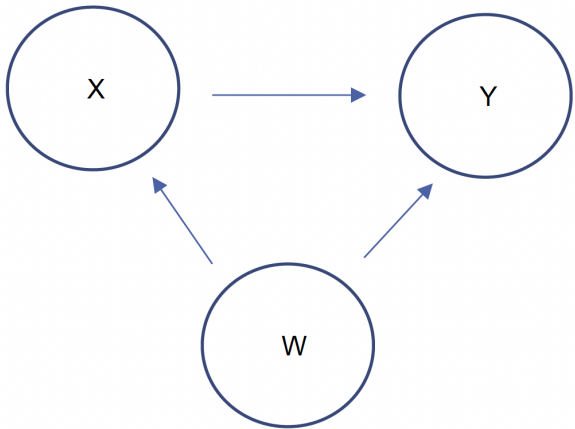
Causal Relationship



Non-causal Relationship: Reverse Causality



Non-causal Relationship: Omitted Variable



Causal Research Questions

- Does having health insurance make people healthier?
- Does the minimum wage lower employment?
- Does getting a university degree increase wages?
- Does smoking cause cancer?

How Do Economists Approach Causality?

- Causality is (often) our aim
- Economists look for situations where a variation is “as good as random” – in other words, contexts with *exogenous* variation –, thereby making it possible to *causally identify* the effect that you are after
  - Controlled experiments: control group arise through randomization
  - Natural experiments: “natural” events that credibly generate a random treatment and/or give rise to variations that “mimick” an experimental setting
  - Econometric methods for causal inference: RDD, IV, DID, Structural estimation, etc
- You should look for settings where you can tackle your question with an exogenous variation, using one of the approaches above
- This is not easy. Your proposal will hardly be perfect, but you should understand and acknowledge the limitations of your proposal and (!) discuss their potential consequences

Some Examples of Papers with Non-Causal Analyses I Like

- There isn't only causality in the world
- Violence against Women at Work (Adams-Prassl, Huttunen, Nix, Zhang)
- Measuring and Explaining Management Practices Across Firms and Countries (Bloom and Van Reneen, QJE 2007)
- The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility (Chetty, Friedman, Hendren, Jones, Porter, AER forth)
- The Organisational Economics of School Chains (Neri, Silva, Pasini, WP - of course :)!)

Statistical Inference

Statistical Inference

- Hypothesis testing one way to examine whether a particular proposition concerning a population is likely to hold (subject to sample variation)
- **Null hypothesis** (H0): the “default position” assumed to hold true until proven otherwise. E.g., the conjecture that there is no difference in in X between population A and population B (other than due to chance
- **Alternative hypothesis** (H1): contrasts the H0 with an opposing statement conjectured to be true instead of H0 (There is a difference in X between population A and B)

Statistical Inference

How do we infer whether an observed difference originates from the null or the alternative model?

- We use a hypothesis test (computing a p-value) to distinguish whether there is a significant effect or whether any difference is due to random variation

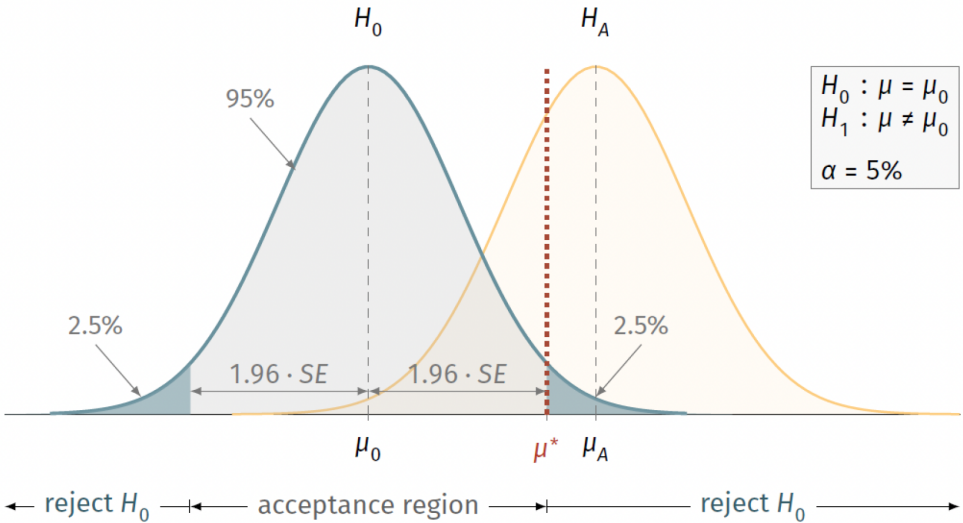
It is important to distinguish between two types of significance:

- Statistical significance
- Economic significance

P-value:

- The probability of obtaining a test result at least as “far away” from the null, given that the null is true.

Statistical Inference



Statistical Testing: Errors

Type I and Type II Errors

Type I errors: false positives

- An incorrect rejection of a true H0 (finding an effect that is not true)
- Given a correctly implemented hypothesis test, what is the probability of making a Type-1 error?

Type I errors: false positives

- An incorrect rejection of a true H0 (finding an effect that is not true)
- Given a correctly implemented hypothesis test, what is the probability of making a Type-1 error?
- Alpha (α) is the test's probability of making a type-I error
  - $p<0.05$ "magical" limit for statistical significance
  - Is there much of a difference betweena study finding $p=0.051$ and $p=0.049$?
  - We see many more studies with $p=0.049$ than $0.051\ldots$
- Lowering α leads to a lower probability of making a Type-I error, why do we not just lower $\alpha\ldots$?

Type II errors: false negatives

- Failure to reject a false H0 (failure to find an existing effect)
- We denote the probability of making a type-II error by β
- So why do we not just decrease α?

Type II errors: false negatives

- Failure to reject a false H0 (failure to find an existing effect)
- We denote the probability of making a type-II error by β
- So why do we not just decrease α?
- Because decreasing α (the probability of making a type-I error) would lead to an increase in the probability of making a type-II error...

Statistical Power

Statistical Power (1)

- Statistical power is the probability of a test to correctly reject a false null hypothesis
- In other words: we want to understand what "power" we have to identify an effect when there is one:

$$\pi = P(\textit{reject } H_0 \mid H_1 \textit{ is true})$$

- The statistical power of a test is defined as $\pi = 1 - \beta$
  - Higher power $\rightarrow$ lower false negative rate.

Statistical Power (2)

- Statistical power is an important aspect of a study
- A low power suggests that a study is not well adapted to detect the effect size it is looking for
  - I.e., Sample size is too small to detect an effect of the size you are looking for
- Studies with significant results but low power are unreliable
- When constructing your study, calculating power and knowing whether you can detect reasonable effect sizes is important

Statistical Power (2)

- Statistical power is an important aspect of a study
- A low power suggests that a study is not well adapted to detect the effect size it is looking for
  - I.e., Sample size is too small to detect an effect of the size you are looking for
- Studies with significant results but low power are unreliable
- When constructing your study, calculating power and knowing whether you can detect reasonable effect sizes is important

Ultimately, **power calculations** involve either:

- determining the effect size that can be detected given a set sample size and other parameters
- determining the sample size needed to detect the minimum detectable effect (MDE) given other parameters

Statistical Power (3)

Statistical power depends on:

- Significance: $\alpha\downarrow \rightarrow \ \pi\downarrow$
- Measurement accuracy: $SE\downarrow \rightarrow \ \pi\uparrow$
- Sample size used to detect the effect: $N\uparrow \rightarrow \ \pi\uparrow$
- The effect size: $D\uparrow \rightarrow \ \pi\uparrow$

Statistical Power (4)

Statistical power depends on:

- **Significance**: $\alpha\downarrow \rightarrow \ \pi\downarrow$
- Lower significance level $\rightarrow$ Smaller type-I error probability $\rightarrow$ Higher type-II error probability, lower statistical power

Statistical Power (4)

Statistical power depends on:

- **Significance**: $\alpha \downarrow \rightarrow \ \pi \downarrow$
- Lower significance level $\rightarrow$ Smaller type-I error probability $\rightarrow$ Higher type-II error probability, lower statistical power
- **Measurement accuracy**: $SE \downarrow \rightarrow \ \pi \uparrow$
- Higher measurement accuracy $\rightarrow$ Higher precision (less noise) $\rightarrow$ Lower type-II error probability, higher statistical power

Statistical Power (4)

Statistical power depends on:

- **Significance**: $\alpha\downarrow \rightarrow \ \pi\downarrow$
- Lower significance level $\rightarrow$ Smaller type-I error probability $\rightarrow$ Higher type-II error probability, lower statistical power
- **Measurement accuracy**: $SE\downarrow \rightarrow \ \pi\uparrow$
- Higher measurement accuracy $\rightarrow$ Higher precision (less noise) $\rightarrow$ Lower type-II error probability, higher statistical power
- **Sample size used to detect the effect**: $N\uparrow \rightarrow \ \pi\uparrow$
- Larger sample size N $\rightarrow$ More accurate measurement $\rightarrow$ Lower type-II error probability, higher statistical power

Statistical Power (4)

Statistical power depends on:

- **Significance**: $\alpha\downarrow \rightarrow \ \pi\downarrow$
- Lower significance level $\rightarrow$ Smaller type-I error probability $\rightarrow$ Higher type-II error probability, lower statistical power
- **Measurement accuracy**: $SE\downarrow \rightarrow \ \pi\uparrow$
- Higher measurement accuracy $\rightarrow$ Higher precision (less noise) $\rightarrow$ Lower type-II error probability, higher statistical power
- **Sample size used to detect the effect**: $N\uparrow \rightarrow \ \pi\uparrow$
- Larger sample size N $\rightarrow$ More accurate measurement $\rightarrow$ Lower type-II error probability, higher statistical power
- **The effect size**: $D\uparrow \rightarrow \ \pi\uparrow$
- Larger effect size $\rightarrow$ Clearer signal-noise separation $\rightarrow$ Lower type-II error probability, higher statistical power

Computing Statistical Power

- **Fixed sample size** (according to budget constraints or external factors – e.g., the number of eligible children in a partner's schools), power calculations determine the effect size the study is powered to detect (the MDE):

$$MDE = (z_{1-\alpha/2} + z_{1-\beta}) \cdot \sigma \cdot \sqrt{\frac{2}{n}}$$

where:

  - α: significance level
  - β type-II error rate
  - $z_{1-\alpha/2}$: z-score for two-tailed test at significance level (e.g., 0.05 for 95% confidence)
  - $z_{1-\beta}$: z-score for desired power (e.g., 0.84 for 80% power))
  - σ : standard deviation of the outcome
  - n: sample size per group (assuming equal group size, so total sample size is 2n)

- For a linear regression coefficient δ:

$$MDE = (z_{1-\alpha/2} + z_{1-\beta}) \cdot se(\hat{\delta})$$

remembering that for a univariate regression, $se(\hat{\delta}) = \frac{\sigma_\varepsilon}{\sqrt{n \cdot var(x)}}$

Issues Linked to Lack of Statistical Power (1)

- Low probability of finding true effects (large type-II error rate)
  - Simply by definition a power of, e.g., 20% as estimated in empirical economics, implies only 20% of true effects explored detected
  - The median statistical power in Economics is $\leq$ 18% (Ioannidis et al., 2017).

Issues Linked to Lack of Statistical Power (1)

- Low probability of finding true effects (large type-II error rate)
  - Simply by definition a power of, e.g., 20% as estimated in empirical economics, implies only 20% of true effects explored detected
  - The median statistical power in Economics is $\leq$ 18% (Ioannidis et al., 2017).
- High false discovery rate:
  - Define m=number of hypotheses, m_0 and m_1 the number of true null and true alternative
  - α is the significance level and β is type-II error rate - power is $1-\beta$. One can show that:

$$FDR \approx \frac{\alpha m_0}{\alpha m_0 + (1-\beta) m_1} \tag{1}$$

  - When $\beta \uparrow$ (i.e., power decreases) $\rightarrow$ FDR increases

Issues Linked to Lack of Statistical Power (2)

- Exaggerated effect size estimates for true positives
- Intuition:
  - Suppose a true effect exists. Call this effect S
  - With X% power to detect an effect of S, we expect to discover an effect of S on average in X% of the tests
  - Sampling variation and random error means effect size estimates will vary around the true effect - some will be smaller, some larger

Issues Linked to Lack of Statistical Power (2)

- Exaggerated effect size estimates for true positives
- Intuition:
  - Suppose a true effect exists. Call this effect S
  - With X% power to detect an effect of S, we expect to discover an effect of S on average in X% of the tests
  - Sampling variation and random error means effect size estimates will vary around the true effect - some will be smaller, some larger

  $\Rightarrow$ With low power, only large estimates will reach statistical significance, but not small results, leading to a systematic inflation of true positives

Issues Linked to Lack of Statistical Power (2)

- Exaggerated effect size estimates for true positives
- Intuition:
  - Suppose a true effect exists. Call this effect S
  - With X% power to detect an effect of S, we expect to discover an effect of S on average in X% of the tests
  - Sampling variation and random error means effect size estimates will vary around the true effect - some will be smaller, some larger

  $\Rightarrow$ With low power, only large estimates will reach statistical significance, but not small results, leading to a systematic inflation of true positives
- If you are “lucky enough” to detect a significant result with low power, chances are that such effect is inflated

Multiple hypothesis

Multiple hypothesis testing

- What is the problem? Suppose you have run an experiment with several treatments t, and you are interested in examining the effect on a range of outcomes

Multiple hypothesis testing

- What is the problem? Suppose you have run an experiment with several treatments t, and you are interested in examining the effect on a range of outcomes
- You estimate the following model:

$$y_i^x = \alpha_0 + \sum_{t=1}^{T} \beta_t treat_t + \alpha_1 W_i + \varepsilon_i$$

where $x = 1, \ldots Z$ is the number of outcomes. This implies $T \times Z$ hypothesis tests, and therefore $T \times Z$ p-values.

Multiple hypothesis testing

- What is the problem? Suppose you have run an experiment with several treatments t, and you are interested in examining the effect on a range of outcomes
- You estimate the following model:

$$y_i^x = \alpha_0 + \sum_{t=1}^{T} \beta_t treat_t + \alpha_1 W_i + \varepsilon_i$$

where $x = 1, \ldots Z$ is the number of outcomes. This implies $T \times Z$ hypothesis tests, and therefore $T \times Z$ p-values.

- Suppose that:
  - i None of the treatments have any effect on any outcome (all null hypotheses are true)
  - ii The outcomes are independent
- Consider a critical value of 0.05 and 0.10 and two scenarios: 1. $T = 5$ and $Z = 4$ and 2. $T = 5$ and $Z = 20$:
  - How many significant effects would you expect to find?

Multiple hypothesis testing

- What is the problem? Suppose you have run an experiment with several treatments t, and you are interested in examining the effect on a range of outcomes
- You estimate the following model:

$$y_i^x = \alpha_0 + \sum_{t=1}^{T} \beta_t treat_t + \alpha_1 W_i + \varepsilon_i$$

where $x = 1, \ldots Z$ is the number of outcomes. This implies $T \times Z$ hypothesis tests, and therefore $T \times Z$ p-values.

- Suppose that:
  - i None of the treatments have any effect on any outcome (all null hypotheses are true)
  - ii The outcomes are independent
- Consider a critical value of 0.05 and 0.10 and two scenarios: 1. $T = 5$ and $Z = 4$ and 2. $T = 5$ and $Z = 20$:
  - How many significant effects would you expect to find?
  - What is the probability to find at least one false effect?

Multiple hypothesis testing

- Consider scenario 1: $T = 5$ and $Z = 4$ imply $5 \times 4 = 20$ hypotheses.

Multiple hypothesis testing

- Consider scenario 1: $T = 5$ and $Z = 4$ imply $5 \times 4 = 20$ hypotheses. We'd expect 1 significant effect.

Multiple hypothesis testing

- Consider scenario 1: $T = 5$ and $Z = 4$ imply $5 \times 4 = 20$ hypotheses. We'd expect 1 significant effect.
- Then, if we just test the hypotheses one by one, the probability of at least one false rejection when using a critical value of 0.05 and 0.10 are:

$$1 - 0.95^{20} = 64\% \qquad 1 - 0.90^{20} = 88\%$$

Multiple hypothesis testing

- Consider scenario 1: $T = 5$ and $Z = 4$ imply $5 \times 4 = 20$ hypotheses. We'd expect 1 significant effect.
- Then, if we just test the hypotheses one by one, the probability of at least one false rejection when using a critical value of 0.05 and 0.10 are:

$$1 - 0.95^{20} = 64\% \qquad 1 - 0.90^{20} = 88\%$$

- Consider scenario 2: $T = 5$ and $Z = 20$ imply $5 \times 20 = 100$ hypotheses.

Multiple hypothesis testing

- Consider scenario 1: $T = 5$ and $Z = 4$ imply $5 \times 4 = 20$ hypotheses. We'd expect 1 significant effect.
- Then, if we just test the hypotheses one by one, the probability of at least one false rejection when using a critical value of 0.05 and 0.10 are:

$$1 - 0.95^{20} = 64\% \qquad 1 - 0.90^{20} = 88\%$$

- Consider scenario 2: $T = 5$ and $Z = 20$ imply $5 \times 20 = 100$ hypotheses. We'd expect 5 significant effects.

Multiple hypothesis testing

- Consider scenario 1: $T = 5$ and $Z = 4$ imply $5 \times 4 = 20$ hypotheses. We'd expect 1 significant effect.
- Then, if we just test the hypotheses one by one, the probability of at least one false rejection when using a critical value of 0.05 and 0.10 are:

$$1 - 0.95^{20} = 64\% \qquad 1 - 0.90^{20} = 88\%$$

- Consider scenario 2: $T = 5$ and $Z = 20$ imply $5 \times 20 = 100$ hypotheses. We'd expect 5 significant effects.
- Then, if we just test the hypotheses one by one, the probability of at least one false rejection when using a critical value of 0.05 and 0.10 are:

$$1 - 0.95^{100} = 99.4\% \qquad 1 - 0.90^{100} = 99.99\%$$

Visualising FDR; $\alpha = 0.05$

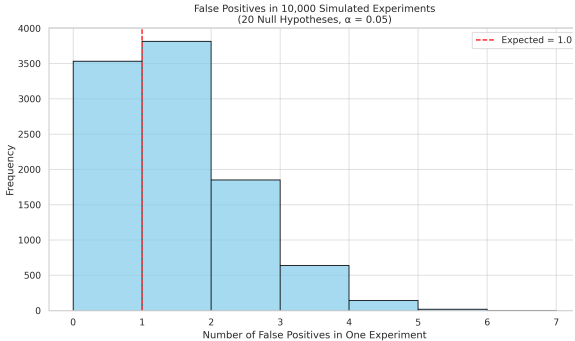


Figure 1. 20 hypotheses

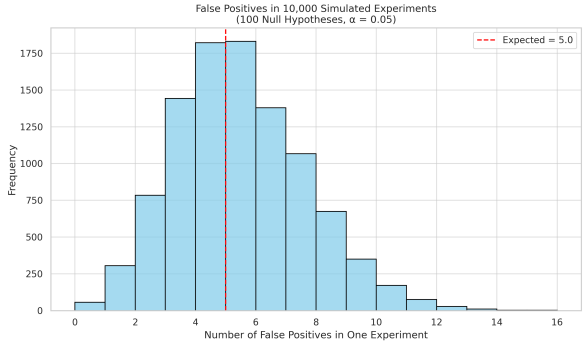


Figure 2. 100 hypotheses

Visualising FDR; $\alpha = 0.10$

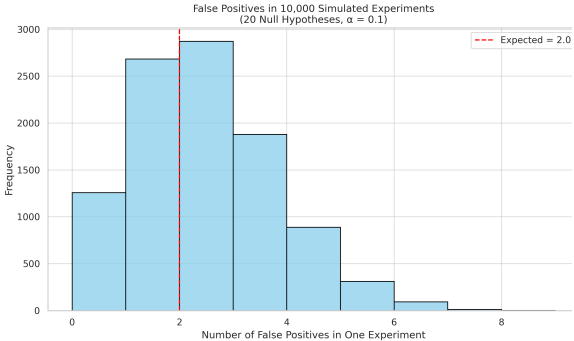


Figure 1. 20 hypotheses

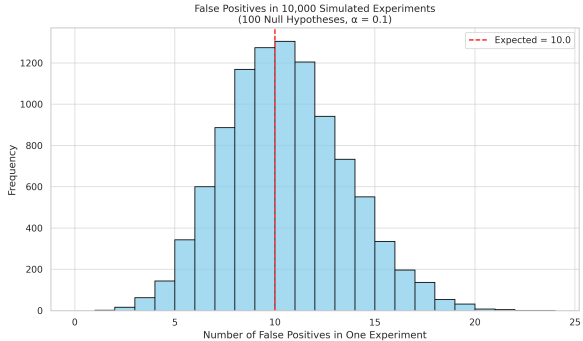


Figure 2. 100 hypotheses

Multiple Hypothesis Testing

- The implication is that we may end up finding significant effects and building stories out of nowhere

Multiple Hypothesis Testing

- The implication is that we may end up finding significant effects and building stories out of nowhere
- In order to reduce the likelihood of these false rejections, we want some way of adjusting for the fact that we are testing multiple hypotheses (m)

Multiple Hypothesis Testing

- The implication is that we may end up finding significant effects and building stories out of nowhere
- In order to reduce the likelihood of these false rejections, we want some way of adjusting for the fact that we are testing multiple hypotheses (m)
- There are different adjustments that have been proposed to deal with the Family-wise Error Rate (FWER) - the probability of making any type I error.

Multiple Hypothesis Testing: Adjustments

- **Bonferroni adjustment**: multiply the p-value by the number of tests (m). Reject hypothesis i if:

$$pvalue_i \cdot m \leq \alpha$$

Multiple Hypothesis Testing: Adjustments

- **Bonferroni adjustment**: multiply the p-value by the number of tests (m). Reject hypothesis i if:

$$pvalue_i \cdot m \leq \alpha$$

- **Holm-Bonferroni adjustment**. Start by ordering p-values from lowest to highest. At each step, compare your p-value as follows:

$$pvalue_k \leq \alpha/(m-k+1)$$

Continue until P_k is greater than the adjusted p-value (the null hypothesis for this k^{th} hypothesis is not rejected). All subsequent hypotheses are not significant

Multiple Hypothesis Testing: Adjustments

- **Bonferroni adjustment**: multiply the p-value by the number of tests (m). Reject hypothesis i if:

$$pvalue_i \cdot m \leq \alpha$$

- **Holm-Bonferroni adjustment**. Start by ordering p-values from lowest to highest. At each step, compare your p-value as follows:

$$pvalue_k \leq \alpha/(m-k+1)$$

  Continue until P_k is greater than the adjusted p-value (the null hypothesis for this k^{th} hypothesis is not rejected). All subsequent hypotheses are not significant

- Other methods: Anderson's sharpened False Discovery Rate q-value
  - (!) This FDR is the expected proportion of rejections that are type I errors (false rejections).

Research Credibility

Research Credibility

- We can now think more coherently about research credibility and – more in general – the interpretation of your and other papers' results

Research Credibility

- We can now think more coherently about research credibility and – more in general – the interpretation of your and other papers' results
- Beyond significance, you should be aware of the issues we just discussed

Research Credibility

- We can now think more coherently about research credibility and – more in general – the interpretation of your and other papers' results
- Beyond significance, you should be aware of the issues we just discussed
- In addition, always be mindful of the distinction between statistical and economic significance of an estimate

Research Credibility

- We can now think more coherently about research credibility and – more in general – the interpretation of your and other papers' results
- Beyond significance, you should be aware of the issues we just discussed
- In addition, always be mindful of the distinction between statistical and economic significance of an estimate
- When reading a paper you should think critically whether the actual results presented warrant the story or narrative that is "pushed" by the athors
- And of course, this is also true when setting up your own research projects
  - Doing a power calculation up front o show that your study makes, or does not make, sense increases the quality of your thesis
  - In general, be ready and discuss openly and clearly potential issues that you may face

Questionable Research Methods

- Researcher have many degrees of freedom: design, hypotheses, data collection, analysis, reporting etc. There's a myriad choices, some are arbitrary and that's fine

Questionable Research Methods

- Researcher have many degrees of freedom: design, hypotheses, data collection, analysis, reporting etc. There's a myriad choices, some are arbitrary and that's fine
- However, this opens up the door for questionable practices. For example:
  - Explore your data (measure/test many things but report significant results only/not correct for multiple testing)
  - Collect data until an effect is significant
  - Play around with controls in regression analysis
  - Test different exclusion criteria for outliers
  - Which conditions can be pooled, which should be compared?
  - Decide what was pilot data at the end instead of the beginning of the study
  - Run several pilots and adapt parameters, transform variables
- Hypothesizing After the Results are Known
  - Pre-registered reports (accepting studies based on design, result-blind
  - Pre-Analysis Plans (PAPs): Ex-ante publishing the analysis plan for a study specifying all important aspects of the study (hypothesis, data collection, variables, analysis etc.)
  - PAPs should include all important aspects of a study, be precise enough to be replicable, and exclude the possibilities to diverge