Block 1.4:
History & Methods of Psychology

Research Methods & Practices of Psychology

Oliver Lindemann
Erasmus University Rotterdam

Principles of Statistical Hypothesis Testing

Procedure of Empirical Research in Psychology

Theory → Hypothesis → Experiment → Data → Statistics

Procedure of Empirical Research in Psychology

Theory → **Hypothesis** → Experiment → **Data** → Statistics

Classical Statistics

Karl Pearson (1857 – 1936) Ronald Fisher (1890 – 1961) Jerzy Neyman (1894 – 1981)

Classical Statistics

Ronald Fischer

- terms “null-hypothesis” & “significant”
- urged the distinction between sample and population
- degrees of freedom
- suggested $p < .05$
- random assignment of conditions, random sampling

Neyman and Pearson

- formal decision logic of statistics.
- Power and Type II error
- Effect sizes

Empirical Behavioural Research

Procedure

Theory → **Hypothesis** → Experiment → **Data** → Statistics

Neyman & Pearson suggested decision rule

- following this rule, in the long run, we will not be often wrong
- error rate (α) of the decision process
- e.g. $p < .05$

Simple question or not?

What is a p-value?

Interpreting p

What is a p-value?

$$p(\text{Data}|H_0)$$

And what do most people what to know from the data?

$$p(H_1|\text{Data})$$

But

$$p(H_1|\text{Data}) \neq p(\text{Data}|H_0)$$

Thus, p tells us **nothing** about the likelihood of the hypothesis, neither H_1 nor H_0!

p-values and strength of evidence

Neyman-Pearson approach:
p-values interpretable as binary decision rule! (effect or not)

- Why can't we use p-values as measure of evidence?
- Why is a smaller p-value not more evidence for H_1?

- p-values are not consistent measures of evidence
  - It is relative to sample size
  - It is effected by sampling plan and other subjective elements

Hypothesis Testing with p-values in Practice

How good is the 5% decision rule?

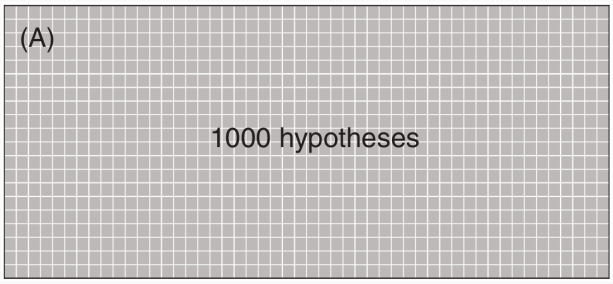
In psychology, we commonly use for the statistics:

- $\alpha = 0.05$
- $(1 - \beta) = 0.80$ (power of 80%)

If we strictly follow the rules above …

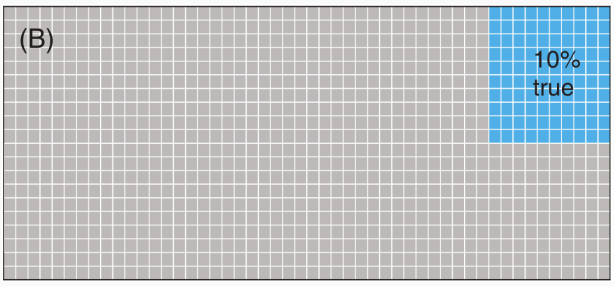
How many published research findings are then false?

How many research findings are false?



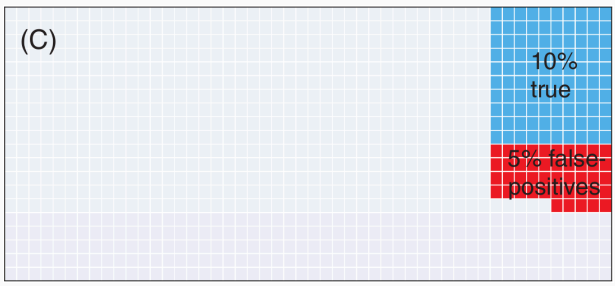
from Forstmeier et al., 2017

How many research findings are false?



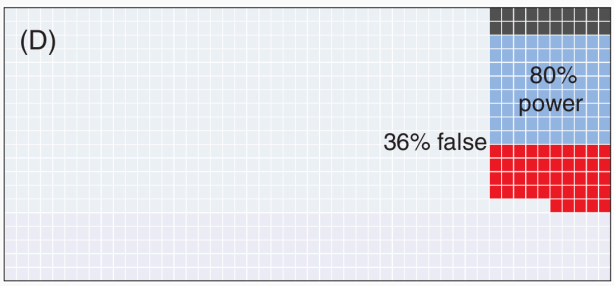
from Forstmeier et al., 2017

How many research findings are false?



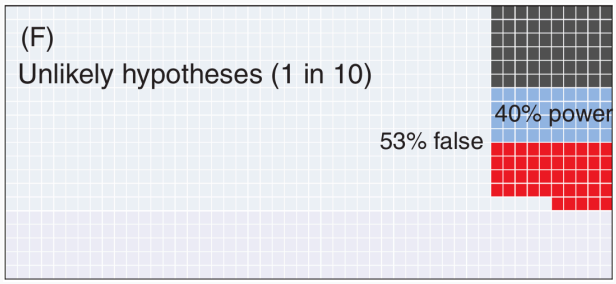
from Forstmeier et al., 2017

How many research findings are false?



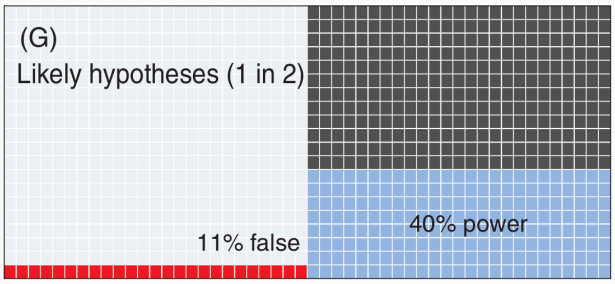
from Forstmeier et al., 2017

How many research findings are false?



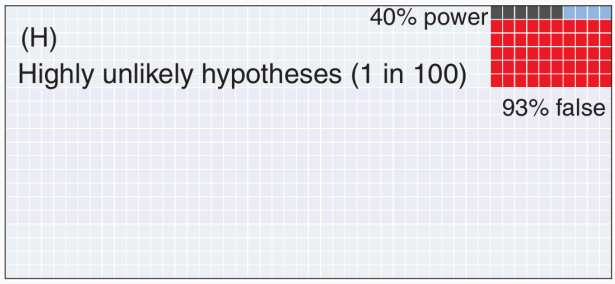
from Forstmeier et al., 2017

How many research findings are false?



from Forstmeier et al., 2017

How many research findings are false?

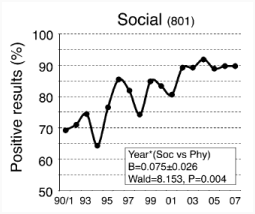


from Forstmeier et al., 2017

Scientific Publication Practice

How many Published Effects are Significant?

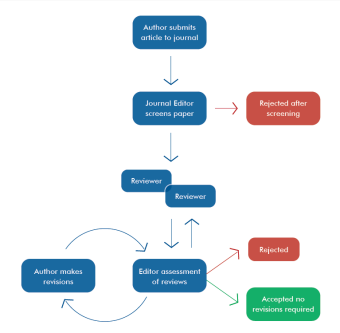
Study by Fanelli (2012)



- Negative results are disappearing from the literature
- this happens from most disciplines and countries

Why are most published effects significant?

Publishing Research in Journals: The Review Process



Which Studies and Results will be Published?

- Journals want to publish the most **exciting** and surprising findings
- Unfortunately, the review process and selection of the editor is affected by many non-scientific factors

Publication bias

- occurs when the publication of research results depends **not just on the quality** of the research but also on the **hypothesis tested**, and the **significance** and direction of effects detected.
- is usually a bias towards reporting significant results

"File Drawer Problem"

- Studies showing significant effects and supporting the hypotheses will be published.
- Studies with no effect and no support for the hypothesis end up in the researchers' file drawer.

As a result, the amount of significant results in most studies is overestimated.

Psychological Reasons for Publication Biases

Why tend researchers to focus on significant effects so strongly?

- Confirmation bias

- Problem of Incentives in Science
- Researchers evaluation and carrier depend mainly on the amount of publications
- This is a general problem that ultimately affects the quality of research.

"There is no cost in to getting things wrong, the cost is not getting them published"

Brain Nosek

Replicability

Skepticism and Replications

Demarcation criterion between science and non–science

Replicability

- the amount of consistency in results when scientific studies are repeated
- a basic element of critical scrutiny of claims
- an engine to the advancement of **self-correcting** science

Advantages

- confirms scientific findings
- specifies the conditions under which the effect is registered
- more accurate estimates of the strength of the effect (Brandt et al, 2013)

What does it look like in real science?

Meta-science study by Mackel, Pluncker & Hegarty (2012)

- Analysis of ALL articles in top 10 psychological journals from 1900
- The term "*replication*" occurred how many articles?

1.6%

- Analysis of 500 randomly chosen articles from this 1.6%:
  - 68% of articles using the term replication are designed to replicate

Replication Rate in Psychology?

Open Science Collaboration 2015

- 100 direct replications of experimental and correlational studies
- Direct replication = recreate conditions that are thought to suffice to obtain original effect
- Close to original studies (consultation of authors, use of original materials and internal review)
- Studies were matched with interests and expertise of replication team

Replication project 2015: Method Details

- Quasi-random sample
  - 2008 articles from 3 journals:
    Journal of Personality and Social Psychology (JPSP), Psychological Science (PSCI), Journal of Experimental Psychology: Learning, Memory and Cognition (JEP:LMC)
  - From chosen articles, one study was selected
  - from this study only 1 statistical result was tested
- No standard exists to assess replication success
- Used
  - Significance and P-values
  - Effect sizes (transformed into r)
  - Subjective assessments of replication success
  - Meta analysis of effect sizes

Results: Significance and p-Values

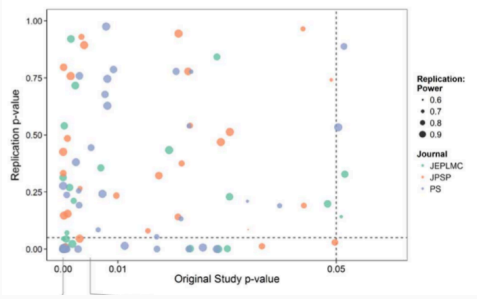
Replication effect tested against Null-hypothesis of no effect

- 97 studies originally significant
- Expected: 89 positive results

Only 35 studies could be replicated

→ a replication rate of **36%**

Results: Significance and p-Values



Differences between Subdisciplines

Replication success rate for

1. Social psychology: 25%
2. Cognitive psychology: 50%

Possible explanation

- Weaker original effects for social psychology
- Higher power of test in cognitive psychology (e.g., within-subject designs)

Reasons for Low Replication Rates?


nature International weekly journal of science

nature news home | news archive | specials | opinion | features | news blog | nature journal

comments on this story

Stories by subject
Brain and behaviour
Lab life

Stories by keywords
Diederik Stapel
Tilburg University
Academic fraud
Retractions
Social psychology

This article elsewhere
Blogs linking to this article
Add to Digg
Add to Facebook
Add to Newsvine

Published online 1 November 2011 | Nature 479, 15 (2011) | doi:10.1038/479015a
Updated online: 1 November 2011
Updated online: 6 December 2011

News

Report finds massive fraud at Dutch universities

Investigation claims dozens of social-psychology papers contain faked data.

Ewen Callaway

When colleagues called the work of Dutch psychologist Diederik Stapel too good to be true, they meant it as a compliment. But a preliminary investigative report (go.nature.com/tqmp5c) released on 31 October gives literal meaning to the phrase, detailing years of data manipulation and blatant fabrication by the prominent Tilburg University researcher.

"We have some 30 papers in peer-reviewed journals where we are actually sure that they are fake, and there are more to come," says Pim Levelt, chair of the committee that investigated Stapel's work at

Fraud & Fabricating Data

Example: The case of Diederik Stapel (Tilburg University)

- fabricated data for at least 30 publications.
- young researchers as the whistleblowers
- suspended from his duties as Professor and returned his Ph.D.

Several other cases are reported …

But fraud, as a general problem in society, is a very isolated problem.

Fraud does not explain the low replication rate

Questionable Research Practices (QRP)

"**p-hacking**"

Analyzing your data multiple ways and selectively reporting only those that result in $p < .05$.

1. outcome-dependent analysis
   - researchers degrees of freedom (see next slides)

- special case: *Optional stopping*
  - Peek into the data frequently and stop analysing if result is significant
  - Collecting more data until results become significant

Researchers Degrees of Freedom while Data Analysis

Researchers flexibility in

- Selecting dependent variables
- Selecting the participants
- Choosing covariates
- Analysis only subsets
- Exclude outliers selectively
  - Choosing to conduct analyses with different outlier criteria

Demo: Hack Your Way To Scientific Glory

Further Types of p-Hacking

2. **Selective reporting**

- Selectively reporting treatment groups and covariates
- Reporting only significant variables
- only reporting studies that show an effect (File drawer problem, problem 7)

3. **HARKing**: Hypothesizing After Results are Known

John Oliver on P-Hacking (YouTube, 1:44–7:55)

What shall we do?

Guidelines for Researchers (. . . and students writing theses)

- Be clear: Exploratory or confirmatory analysis
- Confirmatory research –> specify hypothesis **in advance**
- Report data collection practices
- Determine sample size in advance
  - Include at least 20 participants
- List all variables, experimental conditions and covariates
- Specify analysis procedure beforehand

→ **Study pre-registration!**

Preregistration

- Prevents most types of p-hacking
  - e.g., outcome switching, garden of forking paths, adaptive outlier dropping, exclusion of conditions
- Clear distinction between confirmatory & exploratory research
  - No HARKing
  - Make p-values
- Minimizes publication bias
  - Even if pre-registered studies are not published at the end, the registry can be searched

How to do it? → use websites such as aspredicted.org or osf.io

Questions?

Thank you very much