On the abundance of false positives in psychology

Matthijs J Warrens

M&S Colloquium, September 23, 2013

Introduction

- Clear cases of scientific misconduct (fraud)
- Debate on how to do proper data analysis
- Ongoing in M&S research group
- Spring colloquium by Hemmo

To continue the debate, sources

- Book "A statistical guide for the ethically perplexed" Hubert and Wainer (2013)
- Paper "Why most published research findings are false" Ioannidis (2005)
- Paper "Misunderstanding analysis of covariance" Miller and Chapman (2001)

Colloquial series

Two concerns

- Questionable research practices
- Many published research findings are probably not true

Two topics for colloquial series

1. Uncritical use of covariates in quasi-experimental studies
2. Abundance of false positives in psychology

Hypothesis testing

Statistical hypothesis testing

- Main method for statistical inference
- Procedure
  1. Given initial research question two hypotheses are formulated H_0 (no effect) and H_A
  2. Test statistic is chosen and distribution under H_0 is determined
  3. Level of significance is chosen (α, type I error rate)
  4. Using data observed test statistic is computed
  5. Associated p-value under H_0 is determined
  6. H_0 is rejected if $p < \alpha$
- Hybrid form of Fisher and Neyman-Pearson paradigms

Type I error rate

Type I error rate denoted by α

- A type I error is the incorrect rejection of a true H_0

Statistical significance criterion

- A type I error is a false positive:
  conclude a research finding exists when in fact it doesn't

Type II error rate

Type II error rate denoted by β

- A type II error is the failure to reject a false H_0

Power $1 - \beta$

- High power, less likely type II error occurred
- High power, more likely that stat. significance reflects a true relationship (Ioannidis 2005)

Power $1 - \beta$, depends on

- Type I error rate α
- Sample size
- Effect size

This talk: false positives

In research, errors are inevitable

Most costly error is probably a false positive (type I error)

- Once in the literature they are persistent
  1. Failure to replicate previous finding are never conclusive
  2. Journals do not publish null results
- Waste of resources
- Field with many false positives loses credibility

General belief that if $\alpha = 0.05$, maximum false-positive rate is 5%

Outline

- Tendency is psychology to report more positive outcomes Fanelli (2010)
- Modeling framework for false positive findings Ioannidis (2005): $(1 - \beta)R > \alpha$
- Consequences framework for psychology
  1. Power in psychological research
  2. Characteristic ratio R in psychology
  3. Inflated type I error rate

Comparing sciences

Idea of hierarchy of sciences (controversial, 200 years old)

1. Physical sciences
2. Biological sciences
3. Social sciences

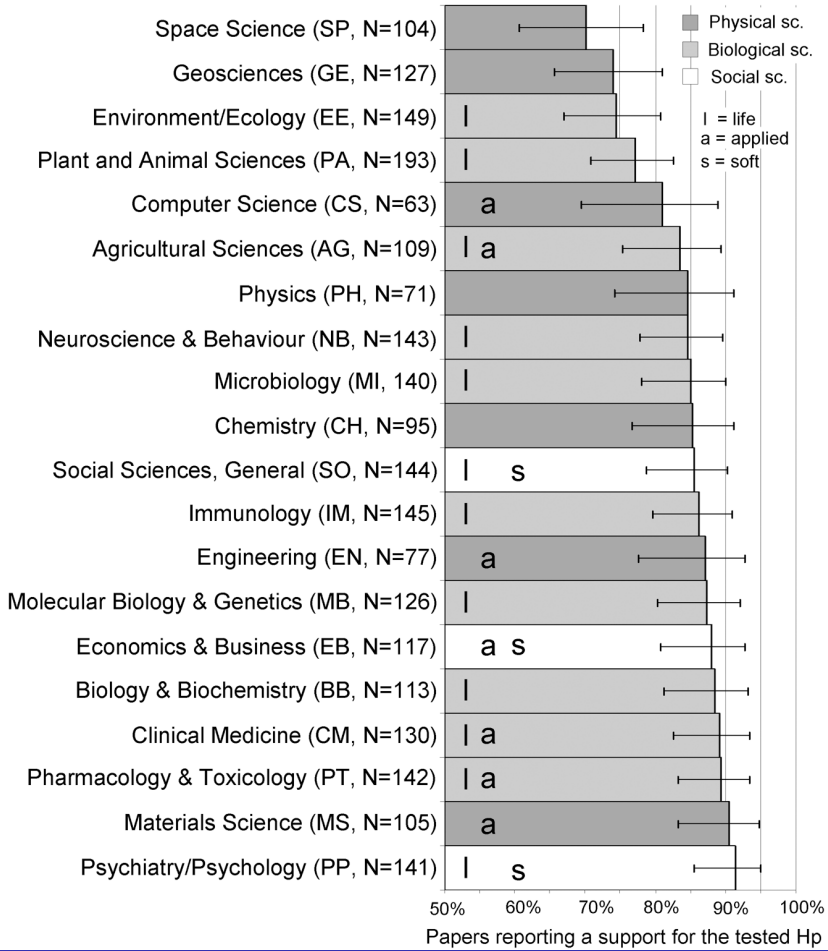
Modern view: three dimensions

1. hard/soft (corresponds roughly to above hierarchy)
2. pure/applied
3. life/non-life

Dimensions have been validated in subsequent studies

Fanelli (2010)

- All scientists have tendency to confirm expectations
- Papers that declared to have tested a hypothesis
- Sampled at random from 10837 journals
- Total of 22 disciplines
- Total of 2434 papers
- No hypothesis testing in Mathematics



Ioannidis (2005)

- "Why most published research findings are false" Ioannidis (2005)
- Provides probability that a research finding is true

- Modeling framework for false positives
- Research finding: any relationship reaching formal statistical significance, e.g.
  1. effective intervention
  2. informative predictor
  3. association

Characteristic ratio

For all the relationships tested/probed in a field, define the ratio

$$R = \frac{\#\text{true relationships}}{\#\text{no relationships}}$$

- R characteristic for a field
- High if only highly likely relationships are targeted
- Low if there are only a few relationships among many hypotheses that may be postulated

Probabilities

For all the relationships tested/probed in a field

$$R = \frac{\#\text{true relationships}}{\#\text{no relationships}}$$

Pre-study probability of a relationship being true is then

$$\frac{R}{R+1} = \frac{\#\text{true relationships}}{\#\text{probed relationships}}$$

- prob. of finding a true relationship is reflected by $1 - \beta$ (power, 1 minus type II error rate)
- prob. of claiming a true relationship when none truly exists is reflected by α (type I error rate)

Expected values

Let c denote the number of relationships that are probed

The expected values of the 2×2 table are

True Relationship	Research Finding Yes	No	Total
Yes	$\frac{c(1-\beta)R}{R+1}$		$\frac{cR}{R+1}$
No	$\frac{c\alpha}{R+1}$		$\frac{c}{R+1}$
Total			c

Expected values

Let c denote the number of relationships that are probed

The expected values of the 2×2 table are

True Relationship	Research Finding: Yes	No	Total
Yes	$\frac{c(1-\beta)R}{R+1}$	$\frac{c\beta R}{R+1}$	$\frac{cR}{R+1}$
No	$\frac{c\alpha}{R+1}$	$\frac{c(1-\alpha)}{R+1}$	$\frac{c}{R+1}$
Total	$\frac{c((1-\beta)R+\alpha)}{R+1}$	$\frac{c(1-\alpha+\beta R)}{R+1}$	c

PPV

After a research finding has been claimed (statistical significance) the post-study prob. it is true is the positive predictive value

$$\text{PPV} = \frac{(1-\beta)R}{(1-\beta)R + \alpha}.$$

Research finding more likely true than false if $\text{PPV} > \frac{1}{2}$ or

$$(1-\beta)R > \alpha = 0.05$$

Corollaries

$(1 - \beta)R > \alpha$

- small sample size, small power
  $\rightarrow$ research findings more likely to be true in fields that undertake large studies

- small effect size, small power
  $\rightarrow$ research findings more likely to be true in fields with large effects
  example large effect: impact of smoking on cancer or cardiovascular disease

Corollaries

$(1 - \beta)R > \alpha$

$$R = \frac{\#\text{true relationships}}{\#\text{no relationships}}$$

- research findings more likely to be true for large R
  1. confirmatory studies or meta-analyses
  2. independent teams addressing the same research questions
- research findings less likely to be true if R is small
  1. hypothesis generating experiments
  2. single teams focusing on isolated discoveries
- the hotter a scientific field (with more teams involved) the less likely the research findings are to be true

Impact on psychology

Probability research finding is true

$$\text{PPV} = \frac{(1-\beta)R}{(1-\beta)R + \alpha}.$$

Three components

- Power $1-\beta$
- Characteristic ratio R
- Type I error rate α

Ratio R difficult to determine (publication bias)

Power in psychological research

Cohen (1962) surveyed 70 studies done in 1960 in 'Journal of Abnormal and Social Psychology'

- Small effect size (mean = 0.18): mean power 0.18
- Medium effect size (mean = 0.48): mean power 0.48
- Large effect size (mean = 0.82): mean power 0.83

Low power due to small samples

Cohen (1990) "Things I have learned (so far)"

- In psychology, studies have low power
- Power should be 0.80

Power in psychology

Sedlmeier and Gigerenzer (1989) repeated Cohen's study
power had actually declined over the years

Rossi (1990) included articles from other journals from 1980
power had not improved over 20 years

Similar trends found by
Maddock and Rossi (2001), Maxwell (2004)

Bakker et al (2012): average power is 0.35
(two independent samples comparison)

Button et al (2013)
in neuroscience, median power 0.18

Publication bias

For all the relationships tested/probed in a field

$$R = \frac{\#\text{true relationships}}{\#\text{no relationships}}$$

Number of true relationships is unknown

Publication process biased in favor of statistically significant results (file drawer problem)

Cooper et al (1997) asked 33 US psychology researchers to describe the fate of 159 studies approved by their departmental review committee

- 2/3 completed studies did not results in published summaries

The type I error rate

In many studies multiple testing is done

- if for 1 test $\alpha = 0.05$
- overall false positive rate is usually higher
- correct for multiple testing (Bonferonni)

Hubert and Wainer (2013)

- Do not do a Bonferonni post hoc, i.e.
  find a set of tests that lead to statistical significance
  and then apply Bonferonni to this subset
- Salami science: a single research study is subdivided into least publishable units

Simmons et al (2011)

When collecting and analyzing data decisions have to be made

- should more data be collected?
- should some observations be excluded?
- should conditions be combined?
- which control variables should be considered?

Unusual to make all these decisions beforehand

Instead, common practice to explore various analytic alternatives and report only results with statistical significance

Flexibility situations

Authors performed a simulation study

Flexibility situations were

1. Two dependent variables ($r = .5$), 3 tests
2. Collect 20 observations per condition,
   keep adding 10 observations till statistical significance
3. Controlling for gender, 3 tests
4. Three conditions, 3 pairwise tests

Results

Assuming $\alpha = 0.05$ for each test separately

Flexibility situation	type I error rate
Situation 1: 2 dep. variables	0.095
Situation 2: adding 10 more	0.077
Situation 3: add control variable	0.117
Situation 4: 3 pairwise tests	0.126
Situations 1 and 2	0.144
Situations 1,2 and 3	0.309
Situations 1,2,3 and 4	0.607

John et al (2012)

Interested in questionable research practices

- Failing to report all dependent variables
- Deciding to collect more data after lack of statistical significance
- Reporting $p < 0.05$ when in fact $p = 0.054$
- Deciding to exclude data after looking at the impact on the results
- Reporting an unexpected finding as having been predicted from the start

Survey

Authors did a survey

- e-mailed survey to 5964 academic psychologists in US
- survey was anonymous
- 2155 respondents to the survey (36%)
- they asked for
  1. self-admission
  2. defensibility (0 = no, 1 = possibly, 2 = yes)
  3. ...

Results

Items	admission rate	defensibility rate
Failing to report all dependent variables	63.4	1.84
Collecting more data when no stat. sign.	55.9	1.79
Failing to report all conditions	27.7	1.79
Stop collecting data when stat. sign.	15.6	1.76
Reporting $p < 0.05$ when in fact $p = 0.054$	22.0	1.68
Selectively reporting studies that worked	45.8	1.66
Exclude data after looking at impact	38.2	1.61
Report unexpected finding as expected	27.0	1.50
Claim that demographic vars not important	3.0	1.32
Falsify data	0.6	0.16

Conclusion

Probability research finding is true

$$\text{PPV} = \frac{(1-\beta)R}{(1-\beta)R + \alpha}.$$

Since in psychology (and other social sciences)

- power usually low
- R small ($< \frac{1}{10}$?)
- α much higher than 0.05

it is very likely that many research findings are false

Discussion

Despite that many research findings are probably not true difficult to change modern publication system

Various authors have proposed guidelines

- how to do improve experimental studies
  (larger sample size, multiple groups working together)
- how to improve data analysis
  (define rules before data is gathered)
- how to reform publication system
  (open access to used data,
  publication guarantee before study is performed)

Teaching

Statistical hypothesis testing (usually) taught as

- binary decision process involving only α
- power is optional

Teach framework Ioannidis (2005) to students?