

Statistics Done Wrong

Pitfalls in Experimentation

Alex Reinhart

www.statisticsonewrong.com

October 16, 2014

Statistics Done Wrong

- Most studies have inadequate sample sizes.
- Most studies test many different hypotheses.
- Most studies use tests even when not needed.

Statistics Done Wrong

- Most studies have inadequate sample sizes.
- Most studies test many different hypotheses.
- Most studies use tests even when not needed.

This means most statistically significant results will be either false positives or exaggerations.

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p -value less than 0.05. Research is not most appropriately represented and summarized by p -values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that

is characteristic of the field vary a lot depending on which field targets highly likely relationships or searches for only one or true relationships among thousands and millions of hypotheses to be postulated. Let us also consider for computational simplicity circumscribed fields where there is only one true relationship among many that can be hypothesized. If the power is similar to find several existing true relationships, the pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (or the Type II error rate). The probability of claiming a relationship when truly exists reflects the Type I error rate, α . Assuming that c relationships are being searched in the field

Example: Research question

“Do fonts alter user disclosure of sensitive information?” (*Psychological Science* 2009)

- If I could get into a movie without paying and be sure I was not seen, I would probably do it.

vs.

- *If I could get into a movie without paying and be sure I was not seen, I would probably do it.*

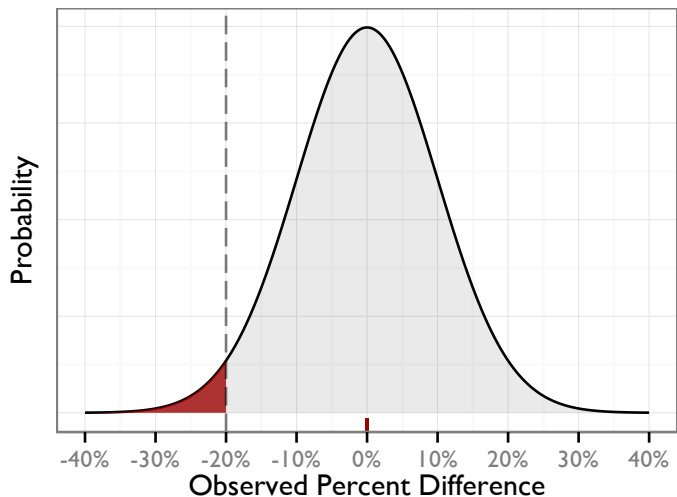
How do we answer this question?

- Look at percentage of questions for which they admitted undesirable activity
- Start with the question: “What would you expect to see if the font made no difference?”
 - Roughly equal percentages for each font
 - But with some random variation

How do we answer this question?

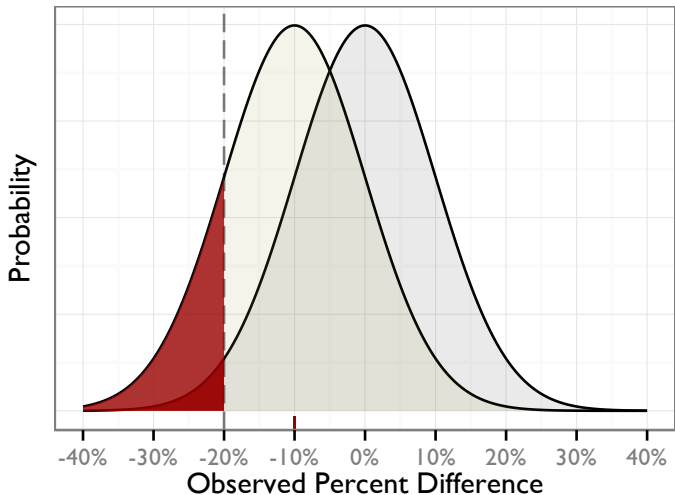
- Look at percentage of questions for which they admitted undesirable activity
- Start with the question: “What would you expect to see if the font made no difference?”
 - Roughly equal percentages for each font
 - But with some random variation
- How does my result compare to this?
- If my result would rarely happen, then it's *statistically significant*

If there's no difference between fonts



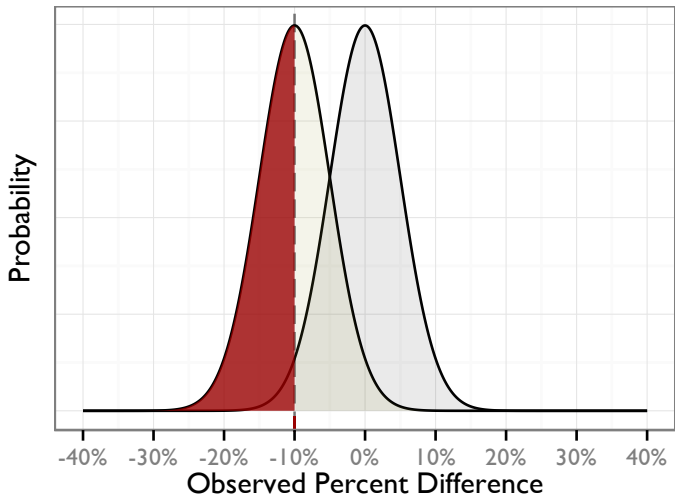
Effects we would see if we run this experiment many times.

If there's a 10% difference between fonts



Effects we would see if small fonts cause a 10% less disclosure. 7 / 33

If we quadruple our sample size



Larger sample sizes make it easier to get significant results.

Statistical power

- Power is the probability we will detect the effect, assuming it exists
- Depends on:
 - Sample size: Larger sample means larger power
 - Effect size: Larger effect means larger power
 - Threshold: Stricter significance threshold means *lower* power

A font replication attempt

LASER, last year:

- Replication attempt using four different methods
 - Online (Mechanical Turk): 390 participants
 - Tablets: 93
 - Written survey: 80
 - Exact replication: 59
- No statistically significant results found, but...

A font replication attempt

LASER, last year:

- Replication attempt using four different methods
 - Online (Mechanical Turk): 390 participants
 - Tablets: 93
 - Written survey: 80
 - Exact replication: 59
- No statistically significant results found, but...
- Power not calculated in advance
- Power varied from 99.5% to 40%

A font replication attempt

- What sample size would be needed for 80% power? About 150.

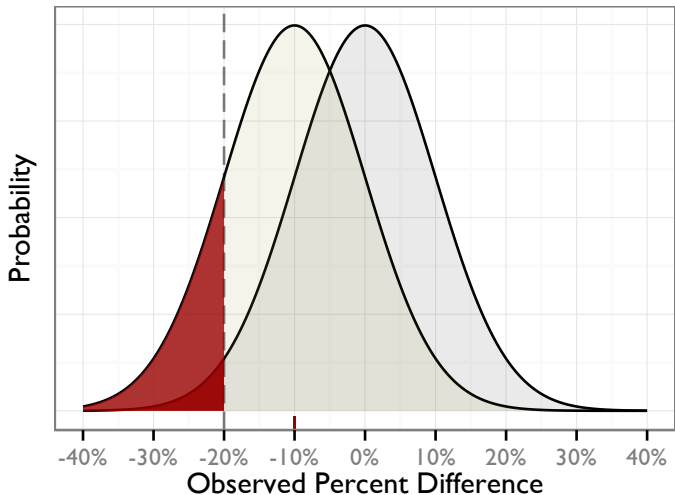
A font replication attempt

- What sample size would be needed for 80% power? About 150.
- Original study had $n = 33$!
- Reanalysis showed they used the wrong statistical test
- Their results weren't significant after all

Problem: Truth inflation

When your sample size is too small, all statistically significant results will be overestimates.

Truth inflation in action



The only statistically significant results are overestimates.

Truth inflation in action

- Why we see papers like “Beautiful Parents Have More Daughters” (*Journal of Theoretical Biology* 2005)
- Biologically plausible effect is 0.3%, papers claimed 20%
- Even with $n = 3000$, statistically significant effects exaggerate truth by factor of 20


Truth inflation in action

- Why we see papers like “Beautiful Parents Have More Daughters” (*Journal of Theoretical Biology* 2005)
- Biologically plausible effect is 0.3%, papers claimed 20%
- Even with $n = 3000$, statistically significant effects exaggerate truth by factor of 20
- If a paper makes a surprisingly large discovery with a surprisingly small sample, be wary

You're gonna need a bigger sample.

- About 20% of users ignore malware warnings
- We think a scarier warning will cut the rate in half
- What sample size do we need for 80% power?



 **Reported Web Forgery!**

This web page at www.itisatrap.org has been reported as a web forgery and has been blocked based on your security preferences.

Web forgeries are designed to trick you into revealing personal or financial information by imitating sources you may trust.

Entering any information on this web page may result in identity theft or other fraud.

[Get me out of here!](#) [Why was this page blocked?](#)

[Ignore this warning](#)

You're gonna need a bigger sample.

- About 20% of users ignore malware warnings
- We think a scarier warning will cut the rate in half
- What sample size do we need for 80% power?

400.

Solution: Power calculations

- A good sample size can be calculated in advance
- Just Google “statistical power calculator” or use R, SPSS, SAS, or Stata
- Power calculations have become mandatory in *Nature* and in reporting guidelines

Solution: Power calculations

- A good sample size can be calculated in advance
- Just Google “statistical power calculator” or use R, SPSS, SAS, or Stata
- Power calculations have become mandatory in *Nature* and in reporting guidelines
- (though power calculators don't mention truth inflation)

Problem: The keep-looking bias

- Collecting enough data can be expensive
- Why not start small and collect more data until we get statistical significance?
- Saves time if you get a significant result early

Problem: The keep-looking bias

- Collecting enough data can be expensive
- Why not start small and collect more data until we get statistical significance?
- Saves time if you get a significant result early
- ...but it also makes false positives and exaggeration more likely

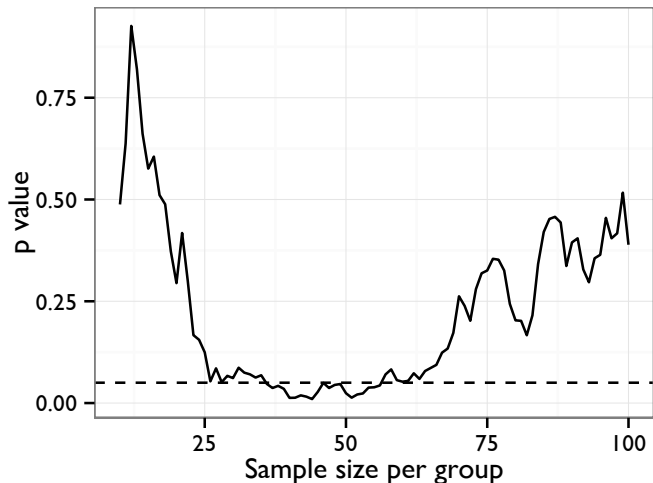
Problem: The keep-looking bias

- Suppose we start with 10 people per group
- If the test isn't significant, recruit one more to each group
- Repeat until we're out of money or have significant results

Problem: The keep-looking bias

- Suppose we start with 10 people per group
- If the test isn't significant, recruit one more to each group
- Repeat until we're out of money or have significant results
- ...and suppose our new warning is no scarier than the old

If we keep collecting more data



Keep increasing your sample size and you'll achieve significance.

Problem: The keep-looking bias

- Applies even if you don't have infinite time or data
- Most scientists do this, but admit it's indefensible
- If there is no published power analysis, this could easily have happened

Problem: The keep-looking bias

- Applies even if you don't have infinite time or data
- Most scientists do this, but admit it's indefensible
- If there is no published power analysis, this could easily have happened
- Entire field of sequential analysis built to solve this problem for medical trials

Problem: Multiple comparisons

The more significance tests you run, the more opportunities for false positives.

Problem: Multiple comparisons

The more significance tests you run, the more opportunities for false positives.

An example: (*SOUPS* 2014)

- Experiment about privacy options on business networking sites, like LinkedIn
- Users rated sensitivity of 27 different personal questions, trustworthiness of 9 categories of people (family, colleagues, students, etc.)

Problem: Multiple comparisons

The more significance tests you run, the more opportunities for false positives.

An example: (*SOUPS* 2014)

- Experiment about privacy options on business networking sites, like LinkedIn
- Users rated sensitivity of 27 different personal questions, trustworthiness of 9 categories of people (family, colleagues, students, etc.)
- 3 types of user: never heard of business networking sites, heard of them, current member
- Do types of users respond differently?

A multiple comparisons example

27 personal questions + 9 categories of people
= 36 tests.

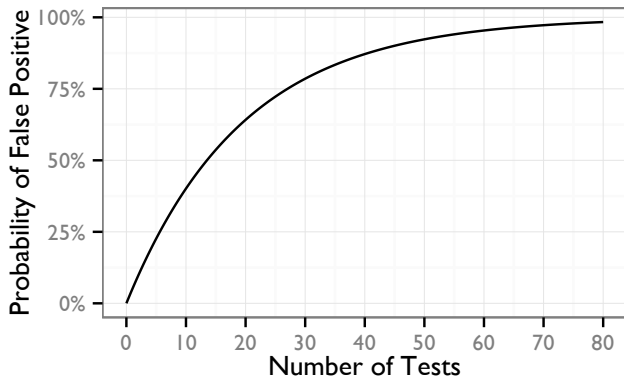
A multiple comparisons example

27 personal questions + 9 categories of people
= 36 tests.

When there are no differences between groups:

- 84% chance of at least one false positive
- On average, 1.8 significant results
- They had one significant result.

Calculating error rates



$P(\text{false positive}) = 1 - (1 - \alpha)^n$, where

- α is your significance level (0.05)
- n is the number of tests

Is this common?

Other papers at SOUPS '14:

- Several papers with > 100 hypothesis tests
- Many with > 20
- ...and this is probably an underestimate
- Reasons for choosing sample sizes not specified
- No mention of power

Solution: Multiple comparison correction

There are methods to correct for multiple testing:

- Bonferroni correction: set significance level to $0.05/n$ (and lose power)
- False discovery rate control

Solution: Multiple comparison correction

There are methods to correct for multiple testing:

- Bonferroni correction: set significance level to $0.05/n$ (and lose power)
- False discovery rate control
- Carefully choose your research hypotheses

Solution: Multiple comparison correction

There are methods to correct for multiple testing:

- Bonferroni correction: set significance level to $0.05/n$ (and lose power)
- False discovery rate control
- Carefully choose your research hypotheses
- ...and don't use tests when you really want effect size estimates!

Example: Right turns on red

- Right turns on red legalized during '70s oil crisis
- Safety studies showed small but statistically insignificant increase in accidents
- Reported as “no significant hazard”

Example: Right turns on red

- Right turns on red legalized during '70s oil crisis
- Safety studies showed small but statistically insignificant increase in accidents
- Reported as “no significant hazard”
- More data reveals 60% more pedestrians, twice as many bicyclists hit at right turns

Example: Right turns on red

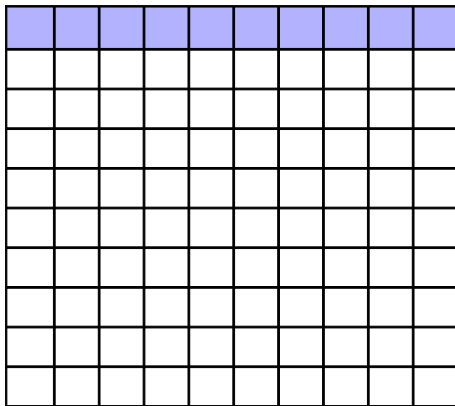
- “Statistically insignificant” does not mean “no significant hazard”!
- A *confidence interval* suggests an upper bound on the size of the hazard
- This could be used in a cost-benefit analysis

Poor power and multiple comparisons

If we have low power and make many comparisons, what happens?

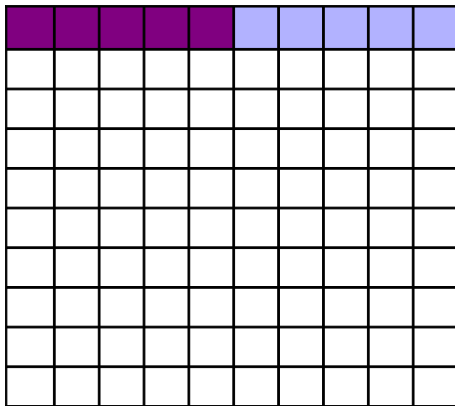
- Suppose we're testing 100 potential drugs
- We have 50% power
- Only 10 of the drugs actually work

The impact of multiple comparisons



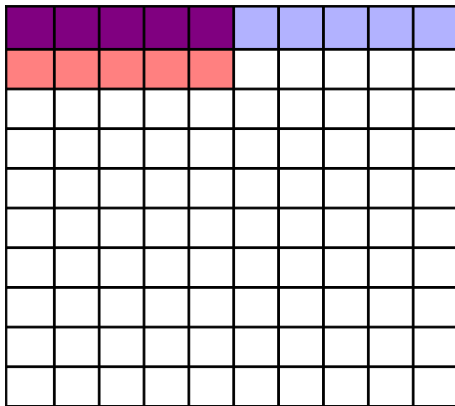
10 out of 100 drugs are truly effective.

The impact of multiple comparisons



But we have 50% power, so we miss 5 good drugs.

The impact of multiple comparisons



And we get 5 false positives in the process.

Our just deserts

- We are putting out results which do not stand up to scrutiny.
- Replication is rare, and errors will be cited as truth for years.
- Even contradicted results are still cited and used.

Enforcing good statistics

- We must learn from other fields, like medicine
- Sample size and statistical analyses must be planned in advance
- Talk to a statistician

Enforcing good statistics

- We must learn from other fields, like medicine
- Sample size and statistical analyses must be planned in advance
- Talk to a statistician
- Funders should require plans
- Publish study protocols in advance?

Presenting the evidence

- Adopt checklists for reporting of sample sizes, statistical tests, and all other important details
- *Nature* has a checklist, and CONSORT has been widely adopted by medical journals
- Use these checklists as a part of peer review
- Make statisticians available during review
- Make analysis code and data available

Think first, ask questions later

- Sample size matters. Calculate it in advance.
- Plan your analysis in advance.
- Otherwise, your results will be exaggerations or false positives.

Think first, ask questions later

- Sample size matters. Calculate it in advance.
- Plan your analysis in advance.
- Otherwise, your results will be exaggerations or false positives.
- I don't want to write another book.