The crisis of modern science

Modern Inference

Aldo Solari

Outline

1. **Scientific studies**

2. **The likelihood of false discoveries**

3. **The likelihood of replicating discoveries**

The crisis of modern science

It is not difficult to find stories of a crisis in modern science, either in the popular press or in the scientific literature

October 19, 2013

Media

Scientific Studies: Last Week Tonight with John Oliver (HBO)

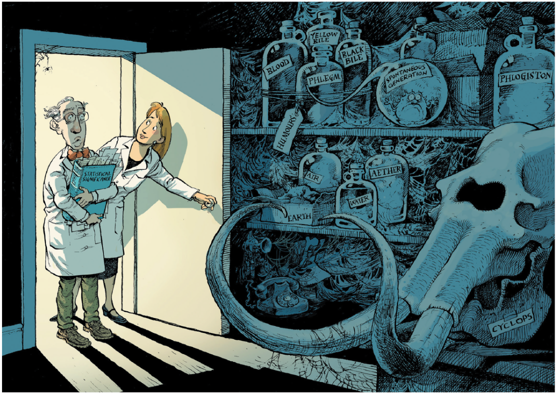
May 8, 2016 (≈16M views)

Who is guilty?

- ☐ I don't know
- ☐ Scientists
- ☐ Statistical methodology
- ☐ Scientists and statistical methodology
- ☐ Another

It's the *p*-value fault

ILLUSTRATION BY DAVID PARKINS



Retire statistical significance

Nature, March 21, 2019

Psychology journal bans *p*-values

Statistical community

2016 The American Statistical Association (ASA) Statement on p-Values: Context, Process, and Purpose

- Opens: the p-value "can be useful"
- Then comes: a list of "Don't"

2019 American Statistician Special Issue: Statistical Inference in the 21st Century: A World Beyond $p < 0.05$

- Opens: "Don't" is not enough
- It concludes: Let's do it. Let's move beyond p-values

Which statistical measure can replace the p-value?

- ☐ Point estimate
- ☐ Likelihood ratio
- ☐ Bayes factor
- ☐ Confidence interval
- ☐ Credibility interval
- ☐ Prediction interval
- ☐ Another
- ☐ A statistical measure not yet invented
- ☐ None, any statistical measure that would be used as frequently as the p-value will be misunderstood and abused as much

Outline

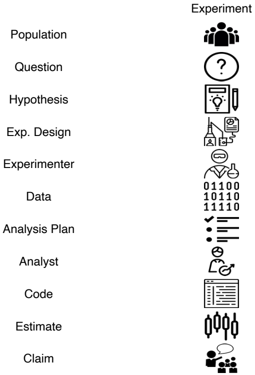
1. **Scientific studies**

2. The likelihood of false discoveries

3. The likelihood of replicating discoveries

Scientific study

Consists of document(s) specifying



False discovery

Publication: Making a public claim on the basis of a scientific study

False discovery: The claim at the conclusion of a scientific study is not equal to the claim you would make if you could observe all data from the population given your hypothesis, experimental design, and analysis plan

- **Population**: All adult males in the Italy at current time
- **Question**: Is Italian male average height more, less or equal to 178 cm?
- **Hypothesis/conjecture**: Italian male average height is < 178
- **Experimental Design**: collect a random sample of $n = 500$ males and measure height
- **Data**: The measured heights $x_1, \ldots, x_n$ in our sample
- **Analysis Plan**: Compute sample average $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

  Conduct a one-sample t-test of the null hypothesis $H_0 : \mu \geq 178$ and compute 95% one-sided confidence interval
- **Code**: `t.test(x, mu=178, alternative="less")`
- **Estimate**: $\bar{x} = 176.9$, $\mathrm{CI} = (-\infty, 177.6]$, $p = 0.005$
- **Claim**: Italian male average height is less than 178 cm

Our discovery seems a true discovery: the average height of Italian male is 176.5 cm (although self-reported height is 177.2 cm; source: `https://en.wikipedia.org/wiki/Average_human_height_by_country`).

p-hacking: Suppose we truly desire to state that the average male height is > 178 cm. We could rewrite our code to continually manipulate the data (drop observations, transform observations, use different statistical tests) until we are able to make this claim with statistical significance

File-drawer problem: If our statistical test does not produce a significant *p*-value, we will disregard our study and move on to a new one that has a better hope for a significant result

Garden of Forking Paths: Suppose we do not fix our assumptions and analysis plan before we observe our data, and based on the distribution of our sample we choose to run a nonparametric test instead of a t-test. If we were to take another sample that appeared normally distributed, we may choose to apply a t-test and get different results

Homework 1

- Does the height of plants depend on the type of fertilization?
  Darwin's plants experiment
- How things fall?
  Galileo's inclined plane experiment

Reproducibility and replicability

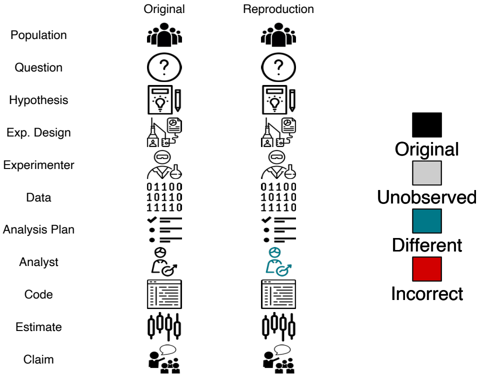
Reproducibility and replicability are fundamental characteristics of scientific studies

Reference:

Patil, P., Peng, R.D. and Leek, J. (2016) A statistical definition for reproducibility and replicability
BioRxiv
`https://www.biorxiv.org/content/10.1101/066803v1.full.pdf`

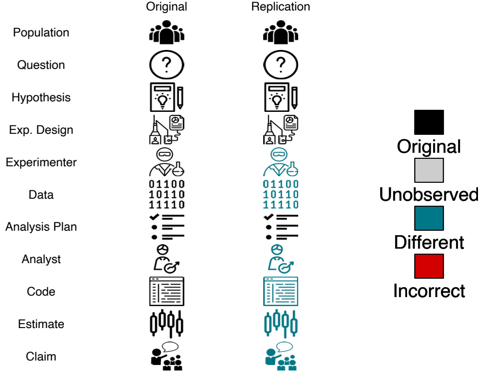
Reproducibility

Reproducibility is defined as reperforming the same analysis with the same code using a different analyst



Replicability

Replicability is defined as reperforming the experiment and collecting new data



Outline

1. Scientific studies
2. **The likelihood of false discoveries**
3. The likelihood of replicating discoveries

The likelihood of false discoveries

- In most scientific fields the acceptable risk of a false discovery (false positive or type I error) is pre-specified for all researchers
- It is conventionally set to $\alpha = 5\%$, which implies that 19 out of 20 times that a researcher performs an experiment the result should not be a false discovery
- This may seem to imply that 19 out of 20 published scientific results are reliable

Reference:

Goeman (2016)
Randomness and the Games of Science (Sections 1-3)
The Challenge of Chance, pp. 91–109
Springer

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Reference:

Ioannidis (2005)
Why most published research findings are false
PLoS medicine
`https://journals.plos.org/plosmedicine/article/file?id=10.1371/journal.pmed.0020124&type=printable`
($\approx$ 8K citations)

The file drawer problem

- Even if 95% of the time researchers produce results that are not false discoveries, this does not mean that 95% of all scientific publications are not false discoveries
- This is because negative results, being less newsworthy, are seldom published
- Looking only at published results, the proportion of false discoveries is likely to be much higher than 5%

Example

- Suppose that 200 experiments have been carried out by researchers in a certain field of science in a certain period of time
- Sometimes the conjecture the researchers set out to prove was correct, sometimes it was not
- For some experiments the researchers accumulated enough evidence to prove the conjecture; for others they were not
- Based on these two dichotomies we can summarize these 200 experiments in a 2×2 contingency table

Scenario A

- Of 200 experiments, 100 are correct conjectures, 100 are wrong conjectures
- Type I error rate = 5%
- Power = 80%

	Correct conjecture	Wrong conjecture	Total
Evidence for conjecture	**80**	**5**	**85**
No evidence for conjecture	20	95	115
Total	100	100	200

- As readers of the scientific literature we only see the 85 published results
- The % of false discoveries in publications is $5/85 = 6\%$

Scenario B

- 200 experiments, 20 correct conjectures, 180 wrong conjectures
- Type I error rate = 5%
- Power = 80%

	Correct conjecture	Wrong conjecture	Total
Evidence for conjecture	**16**	**9**	**25**
No evidence for conjecture	4	71	175
Total	20	180	200

- 25 published results
- 9/25 = 36% false discoveries

Scenario C

- 200 experiments, 100 correct conjectures, 100 wrong conjectures
- Type I error rate = 5%
- Power = 30%

	Correct conjecture	Wrong conjecture	Total
Evidence for conjecture	**30**	**5**	**35**
No evidence for conjecture	70	95	165
Total	100	100	200

- 35 published results
- 5/35 = 14% false discoveries

Outline

1. Scientific studies
2. The likelihood of false discoveries
3. **The likelihood of replicating discoveries**

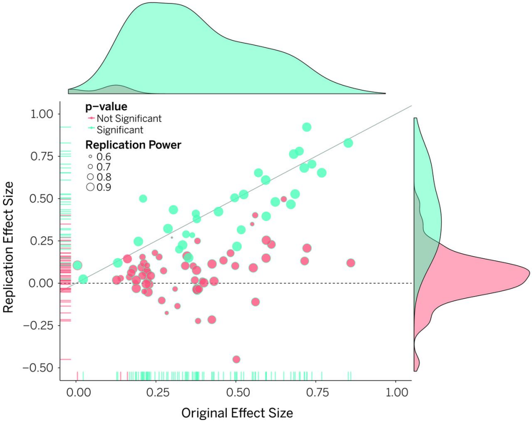
The likelihood of replicating discoveries

- Psychologists replicated a representative sample of 100 studies published in 2008 in three top psychology journals
- 64% of the replication studies did not find statistically significant results as the original studies

Reference:

Hung and Fithian (2020)
Statistical methods for replicability assessment
arXiv: arXiv:1903.08747

Reproducibility of psychological science

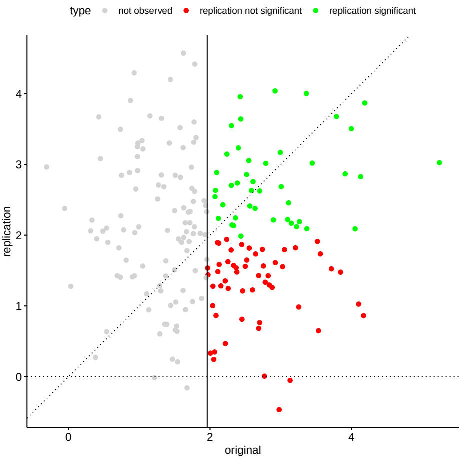


Reference:

Open Science Collaboration (2015)
Estimating the reproducibility of psychological science
Science

Simulation

- All experiments (both original and replication) are identical
- Testing zero effect ($H_0 : \mu = 0$) vs positive effect ($H_1 : \mu > 0$)
- Test statistic $Z \sim N(\mu, 1)$ rejects H_0 if $Z > 1.96$, or equivalently, $p < 0.05$ (type I error rate 5%)
- Suppose that true $\mu = 2$ (power $\approx 51.6\%$)
- We observe only study pairs for which the original experiment is significant (i.e published)



- We select only original experiments that are significant (97/200)
- $54/97 = 55.6\%$ of the replication experiments did not find statistically significant results as the (selected) original experiments

Selection is ubiquitous

- Experiments are selected for publication. But selection increases false discoveries
- Experiments are selected for replication. But selection decreases replicability

Disclaimer

The course material contains images obtained by routine Google Images searches. Some of these images may perhaps be copyright. They are included here for educational noncommercial purposes and are considered to be covered by the doctrine of Fair Use. In any event they are easily available from Google Images. It's not feasible to give full scholarly credit to the creators of these images. We hope they can be satisfied with the positive role they are playing in the educational process.