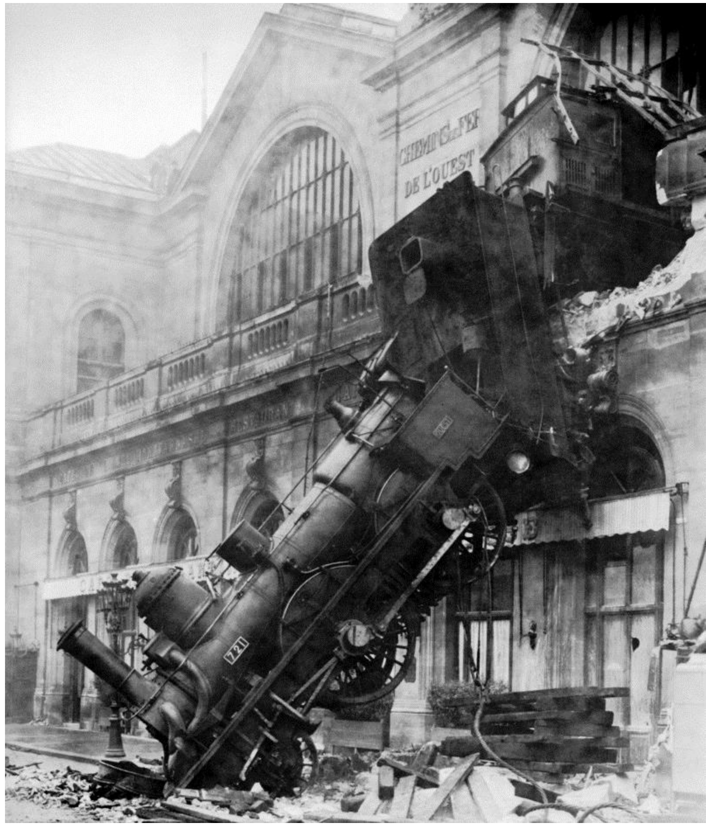
Significance, reproducibility, power analysis, *p*-hacking, harking and other disasters

Edoardo Saccenti

Laboratory of Systems and Synthetic Biology

Reproducible Microbiome 2018



Content

- Reproducibility in Science
- *P*-value
- Scientific method and
- Hypothesis testing
- 0.05 confidence level and publication bias
- *P*-hacking
- Power analysis in PCA
- The Texas sharpshooter fallacy
- Examples and scientific misconduct
- HARcking
- Multiple testing (correction)

Facts

- Most part of published results are false
- Most part of published results are misleading

Facts

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

ANALYSIS

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Psychological Methods
2004, Vol. 9, No. 2, 147–163

Copyright 2004 by the American Psychological Association
1082-989X/04/$12.00 DOI: 10.1037/1082-989X.9.2.147

The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies

Scott E. Maxwell
University of Notre Dame

Behavioral Ecology Vol. 15 No. 6: 1044–1045
doi:10.1093/beheco/arh107
Advance Access publication on June 30, 2004

A farewell to Bonferroni: the problems of low statistical power and publication bias

Shinichi Nakagawa
Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom

Facts

- Most part of scientific experiments are not reproducible

- How many?

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration[*,†]

+ Author Affiliations

†Corresponding author. E-mail: nosek@virginia.edu

Science 28 Aug 2015:
Vol. 349, Issue 6251,
DOI: 10.1126/science.aac4716

criteria, they find that about one-third to one-half of the original findings were also observed in the replication study.

This actually means that 50% to 67% of experiments **were not reproducible!!**

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration *,†

\+ Author Affiliations

†Corresponding author. E-mail: nosek@virginia.edu

Science 28 Aug 2015:
Vol. 349, Issue 6251,
DOI: 10.1126/science.aac4716

eLIFE
elifesciences.org

FEATURE ARTICLE

REPRODUCIBILITY IN CANCER BIOLOGY

Making sense of replications

Abstract The first results from the Reproducibility Project: Cancer Biology suggest that there is scope for improving reproducibility in pre-clinical cancer research.
DOI: 10.7554/eLife.23383.001

BRIAN A NOSEK AND TIMOTHY M ERRINGTON*

Attempt to reproduce 29 groundbreaking cancer research studies
https://osf.io/e81xl/wiki/home/

First results: 3 out 5 studies failed to be reproduced (~60%)

Why this?

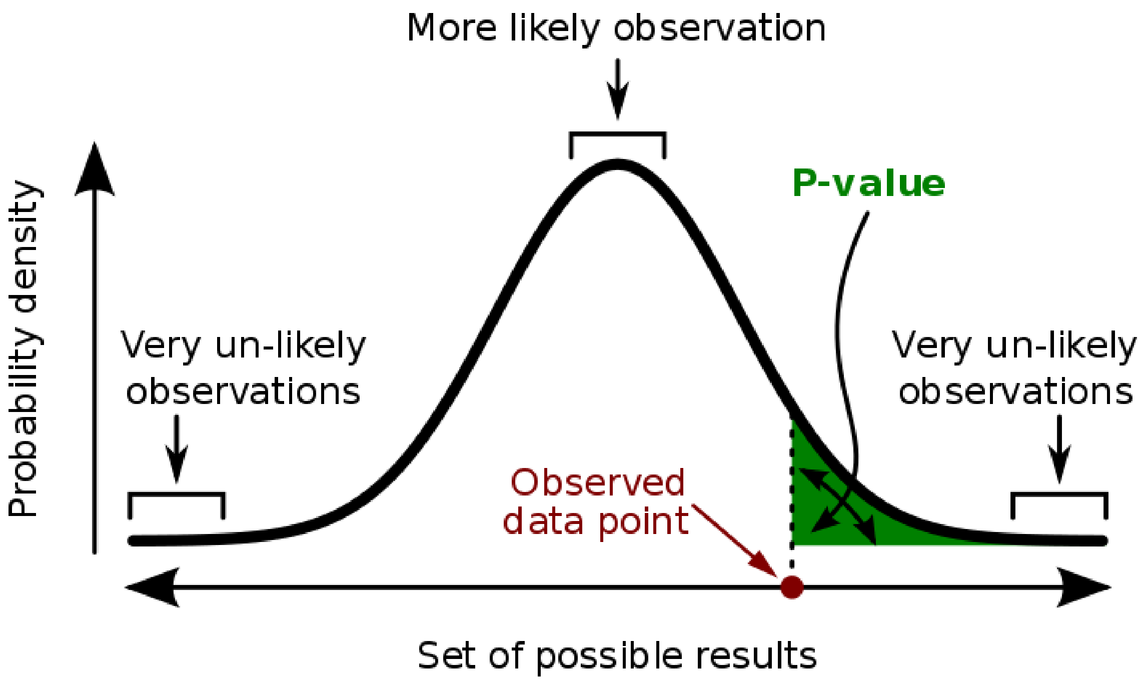
- Suboptimal/wrong experimental design / underpowering

- Publication bias towards p-value < 0.05

Quiz: what is a *p*-value?

1. The probability that my results are correct
2. The probability of observing a result as large or larger as the one observed if the null hypothesis is true
3. The probability that my results are given only by chance
4. The probability that the tested hypothesis is true
5. The probability of observing my data if the null hypothesis is true
6. A measure of the significance of the results
7. The likelihood of the null hypothesis being false given my data
8. None of the above

P-value



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Source: https://rationalwiki.org/wiki/File:P-value_in_statistical_significance_testing.svg

Scientific method

1. State a problem
2. Formulate a theory
3. Perform experiments
4. *Look for agreement between data and theory*
5. *If NOT: adjust theory*

Hypotheses

Null hypothesis

H_0: No effect or no difference

Assume true but look for evidence to disprove

Alternative hypotheses

H_A: Presence of an effect or difference

Try to prove

Hypotheses

Well-formulated hypotheses are quantifiable and testable

- ✓Common problem: hypotheses are too vague
- ✓What is the research question of interest?
- ✓Requires discussion and careful thought

Need to think about directionality (*e.g.*, what are you trying to prove?)

Hypothesis Testing

Make a decision to reject (or fail to reject) the H_0 by comparing what is observed to what is expected if the H_0 is true (i.e., *p*-values)

The hypotheses concern (unknown and unobservable) population parameters.

We make decisions about these parameters based on the (observable) sample statistics (experiments!)

Garbage in – garbage out!

Hypothesis Testing

Evidence is used to disprove hypotheses

We can prove the alternative hypothesis to some standard of proof.

We cannot prove the null hypothesis (we can only fail to reject it)
We cannot prove what we have already assumed to be true.

Thus we do not “accept” H_0. We simply fail to reject it.

What this as to do with reproducibility?

The 0.05 bias

We claim results to be significant if we obtain a p-value < 0.05

P-value

The probability of observing a result as large or larger as the one observed if the null hypothesis is true, the hypothesis is correctly stated and all test assumption are met

Stated otherwise: th *p*-value tells how well your data fit your (null) hypothesis

If $p < 0.05$ we say that here is no agreement between data an null hypothesis thus we fail to prove it

The problem is in this interpretation…P-value is not intended to be dicotomic measure

P-value

If the results are unlikely we should restart and perform experiments again because something went wrong*

Instead $p<0.05$ became a stopping point

Researchers are biased towards results with $p<0.05$

All publication business is $p<0.05$ biased

*This what actually Fisher's interpretation of the *p*-value

Founding fathers view

- *"We are inclined to think that as far as a particular hypothesis is concerned, no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis"*

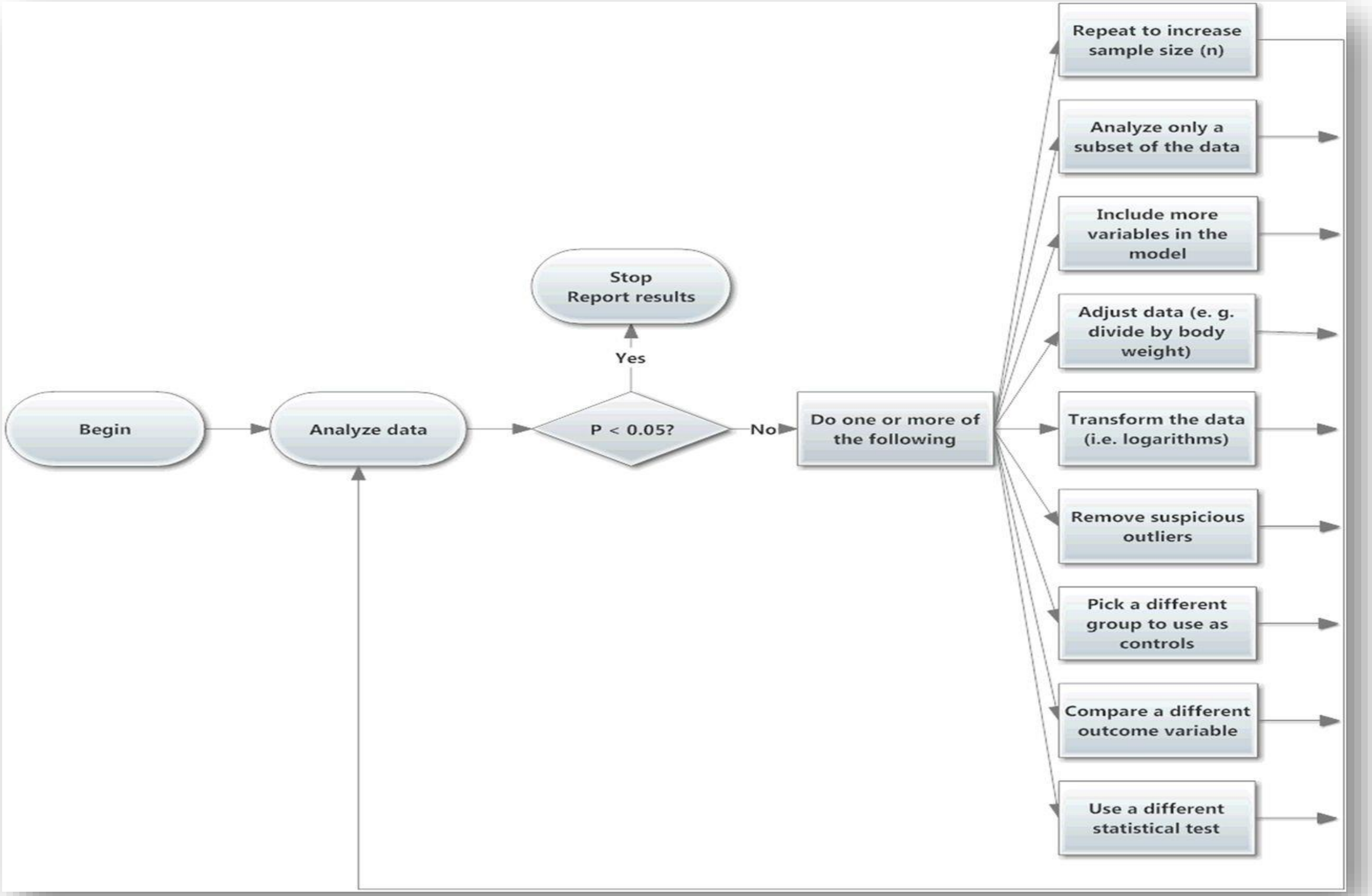
Karl Pearson (1857 – 1936)   Ronald Fisher (1890 – 1961)   Jerzy Neyman (1894 – 1981)

P-hacking*

Steering/manipulating/adjusting statistical analysis towards p<0.05 is called *p*-hacking*

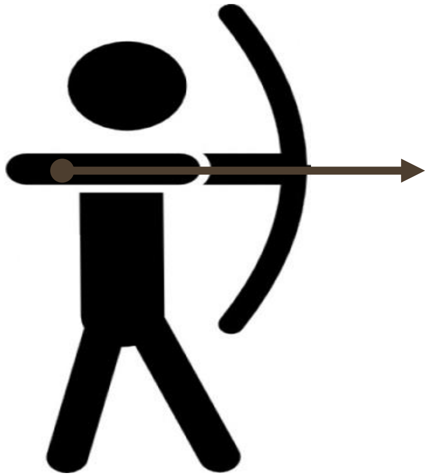
Data dredging, data massaging

Simmons JP, Nelson LD, and Simonsohn U (**2011**) *False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant*. Psychological Science22:1359–1366.



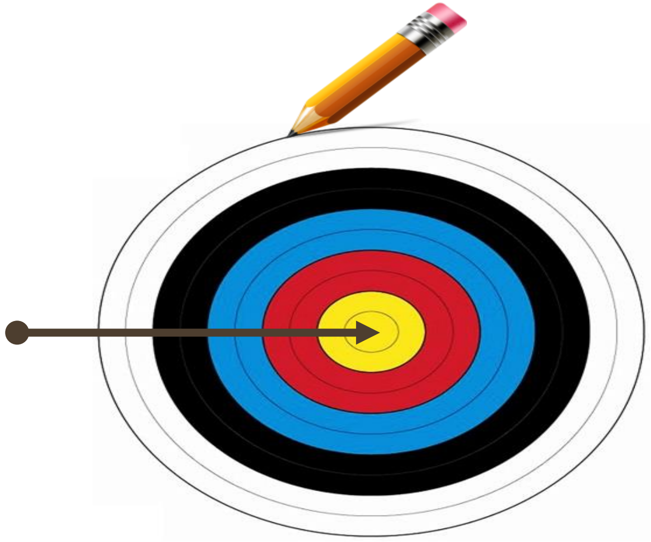
Harvey J. Motulsky J Pharmacol Exp Ther 2014;351:200-205

The Texas sharpshooter fallacy

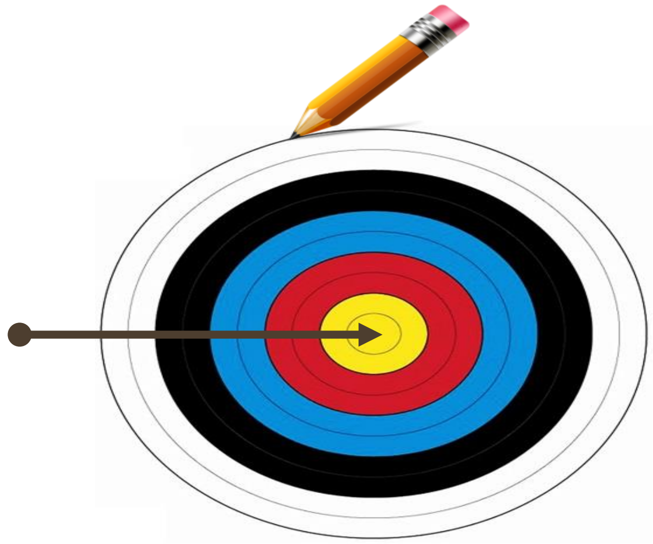


The Texas sharpshooter fallacy

The Texas sharpshooter fallacy

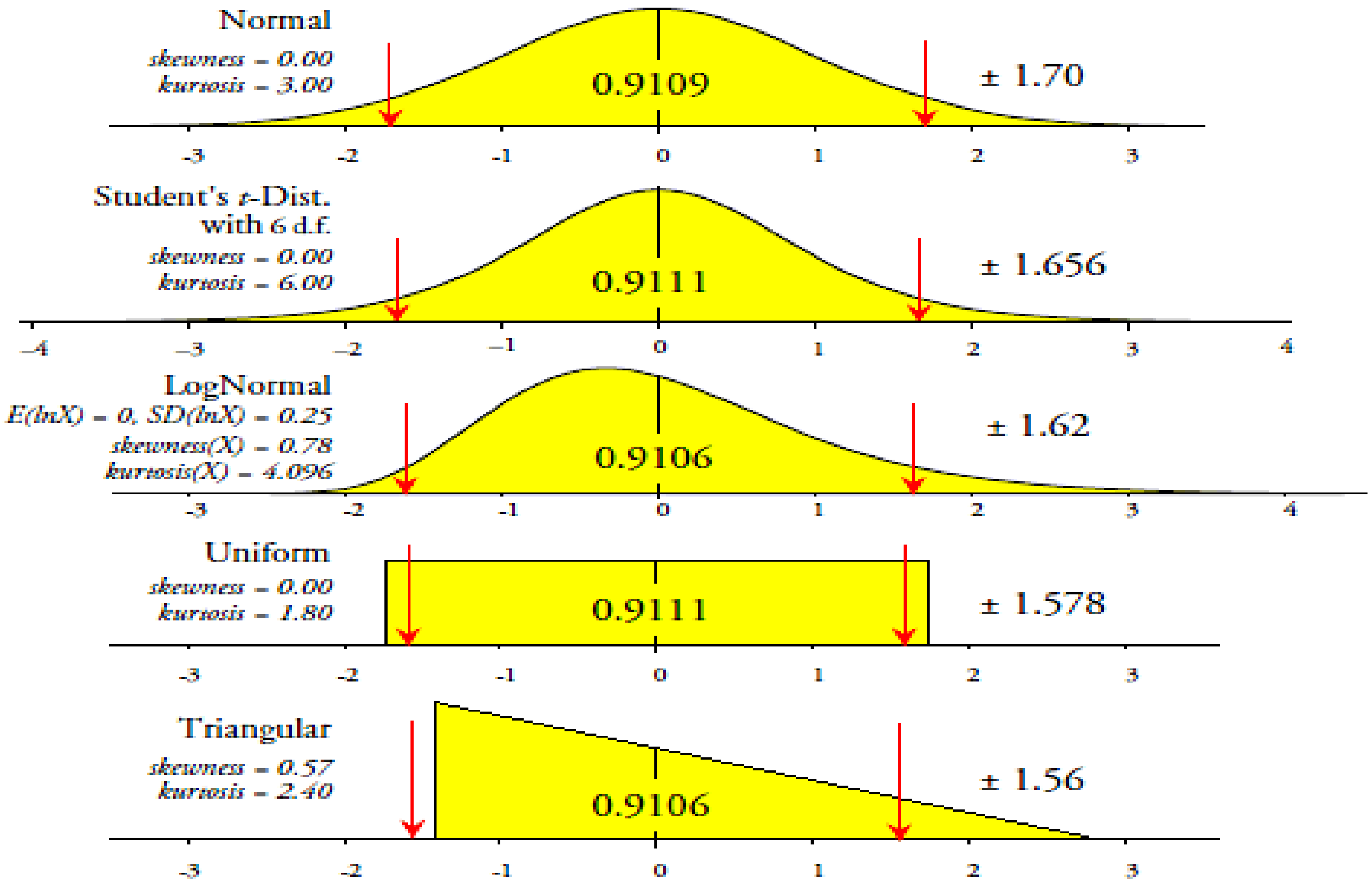


The Texas sharpshooter fallacy

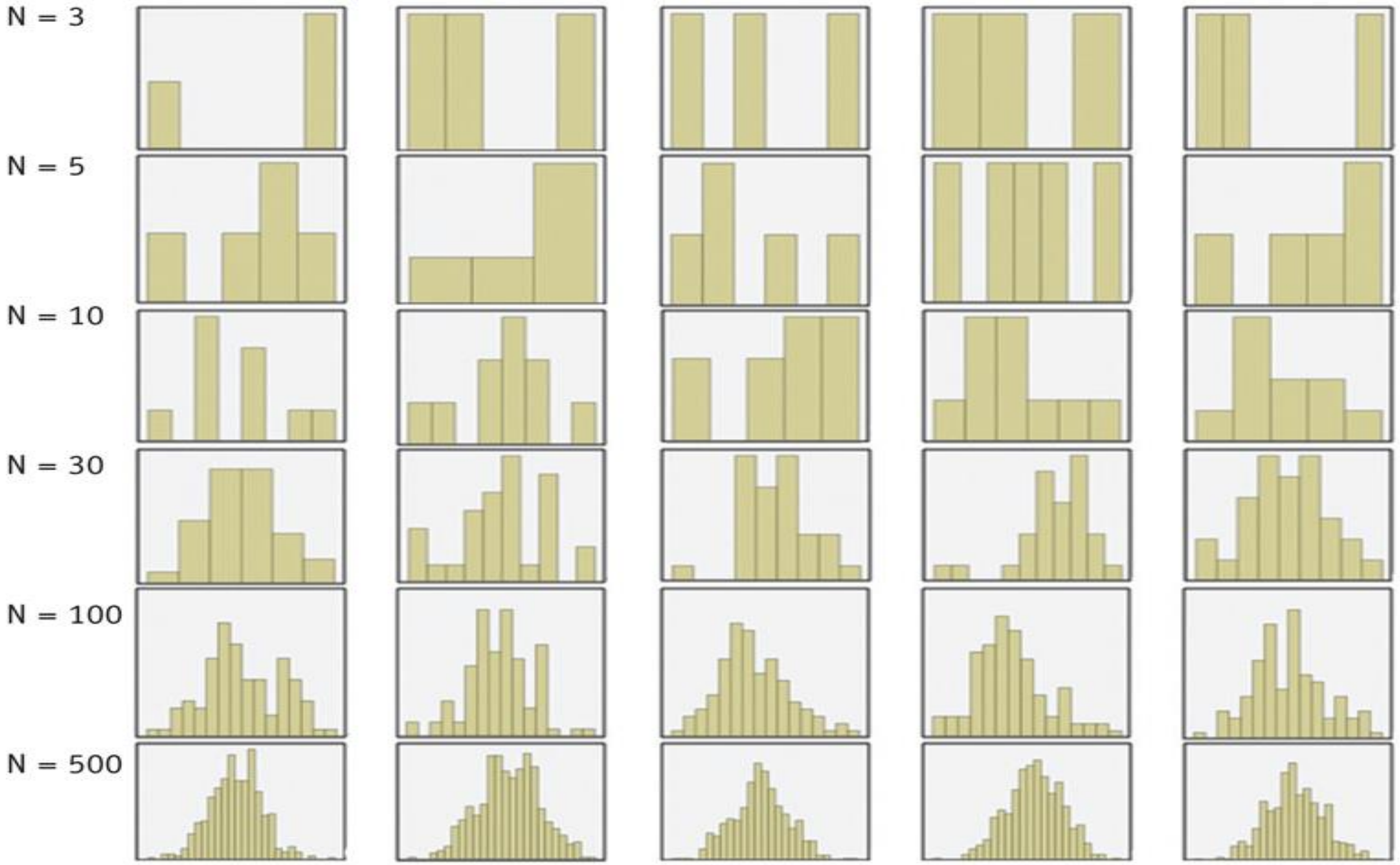


This proves that you are an excellent archer!

Example: normality



Which of these data samples are (not)normally distributed?



How to decide if data are (not) normally distributed?

Use a statistical test!

There are three kinds of lies: lies, damn lies and statistics

Anderson Darling test

H_0: data is *normally* distributed

H_1: data is *not normally* distributed

Anderson Darling test

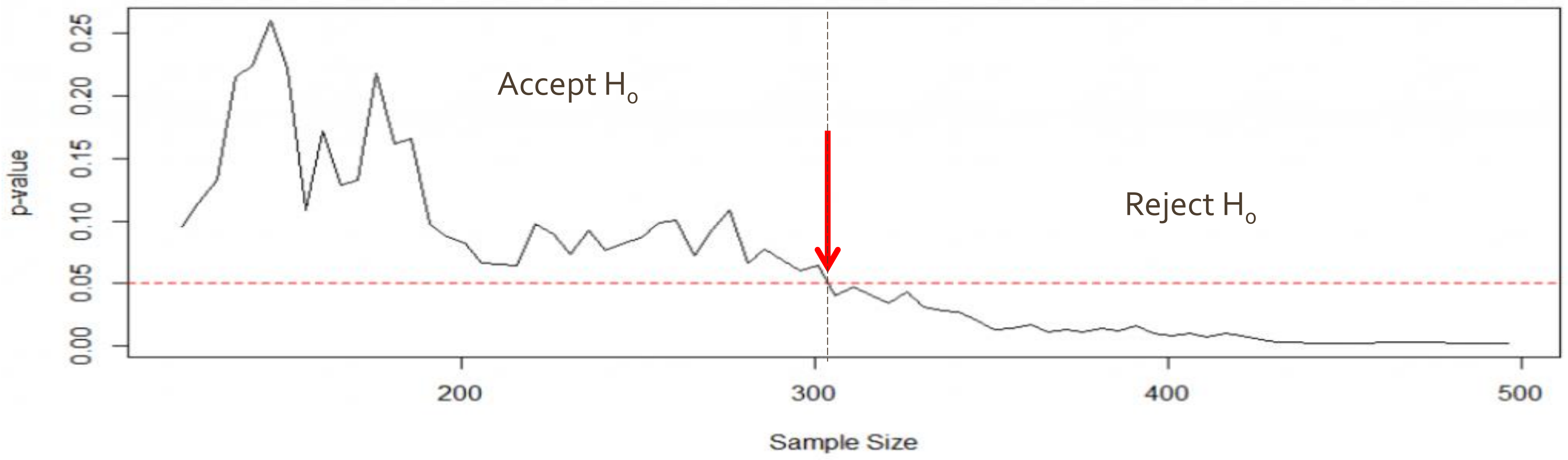
H_0: data is *normally* distributed

H_1: data is *not normally* distributed

With our data (500 observations)

we get *p-val* = 10^{-5}

Perform the test with different sample size



Adding or removing just a few samples changes the results of the test!

Sample size determination

i. Well developed theory for the univariate case:

   - Comparison of means
   - Correlations
   - Regression
   - Proportions

ii. Multivariate extensions (*i.e.* MANOVA)

Sample size determination / power analysis

$(\Delta, N, \alpha, 1\text{-}\beta)$

Generalizability

Reproducibility

Multivariate analysis of *omics* data

1. How many samples for a “good” PCA model?

2. How many samples for a “good” PLS model?

GOOD $\leftrightarrow$ GENERALIZABILITY

Multivariate case

1. Move from *exploratory* setting to *inferential* setting

2. Re-formulate the questions in a *power analysis context*

PCA case

1. How many samples needed for “stable” loading estimation?

   → Approach: numerical simulations

2. How many samples needed for correct dimensionality assessment?

   → Approach: Theoretical approach (RMT)

Journal of proteome research

Article

pubs.acs.org/jpr

Approaches to Sample Size Determination for Multivariate Data: Applications to PCA and PLS-DA of Omics Data

Edoardo Saccenti*,† and Marieke E. Timmerman‡

†Laboratory of Systems and Synthetic Biology, Wageningen University and Research Center, Dreijenplein 10, 6703 HB, Wageningen, The Netherlands

‡Department Psychometrics & Statistics, University of Groningen, Grote Kruissstraat 2/1, 9712 TS, Groningen, The Netherlands

PCA case

Define appropriate statistical model

Define effect(s)

Simulation scheme

Infer population parameters (loadings) from samples

Simulation scheme 2/2

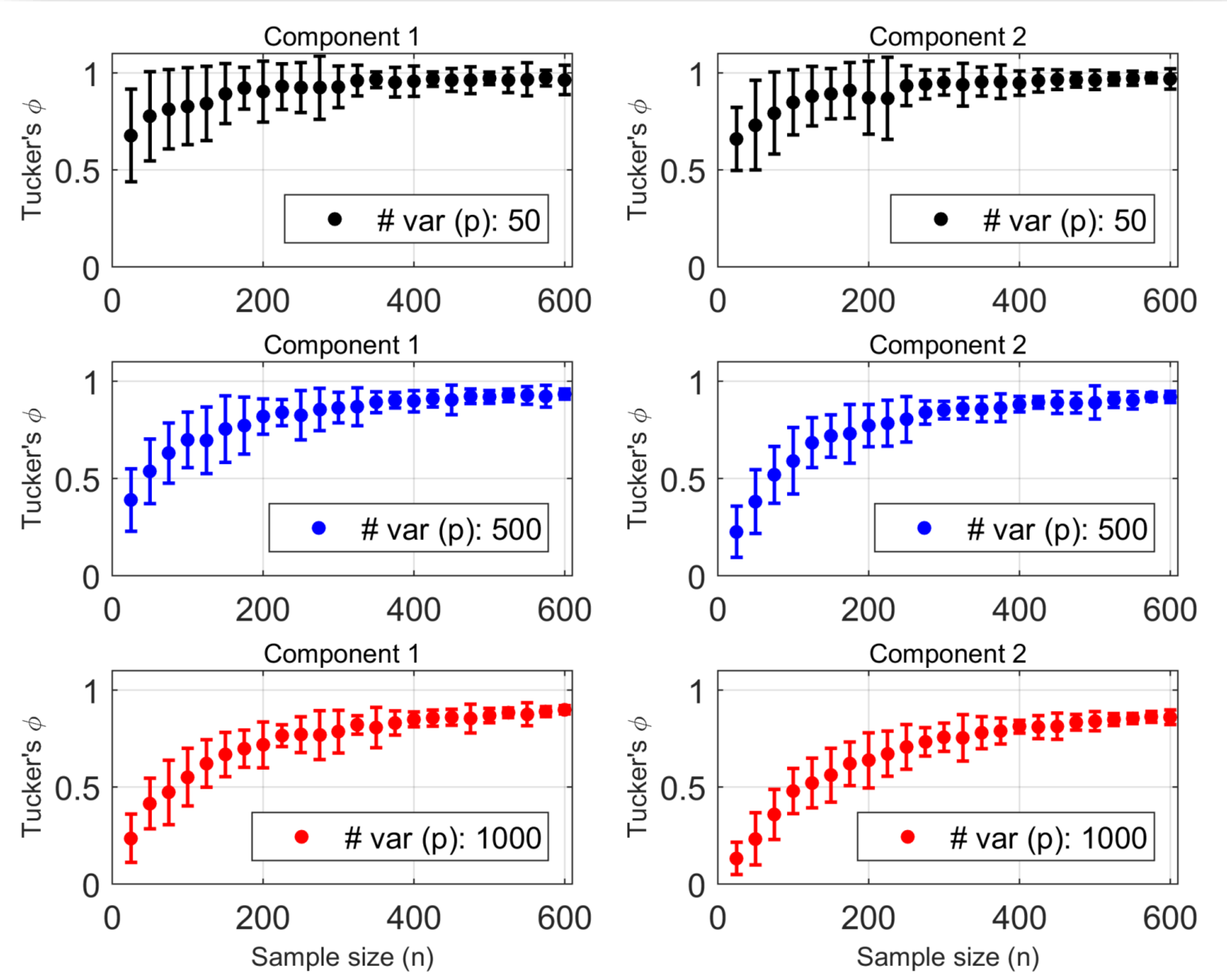
- ✓ Sample loadings converge to Population loadings for *N* large enough

- ✓ Large enough = ?

- ✓ Compare to population loadings (Tucker's φ: loading equivalence $\varphi > 0.9$)

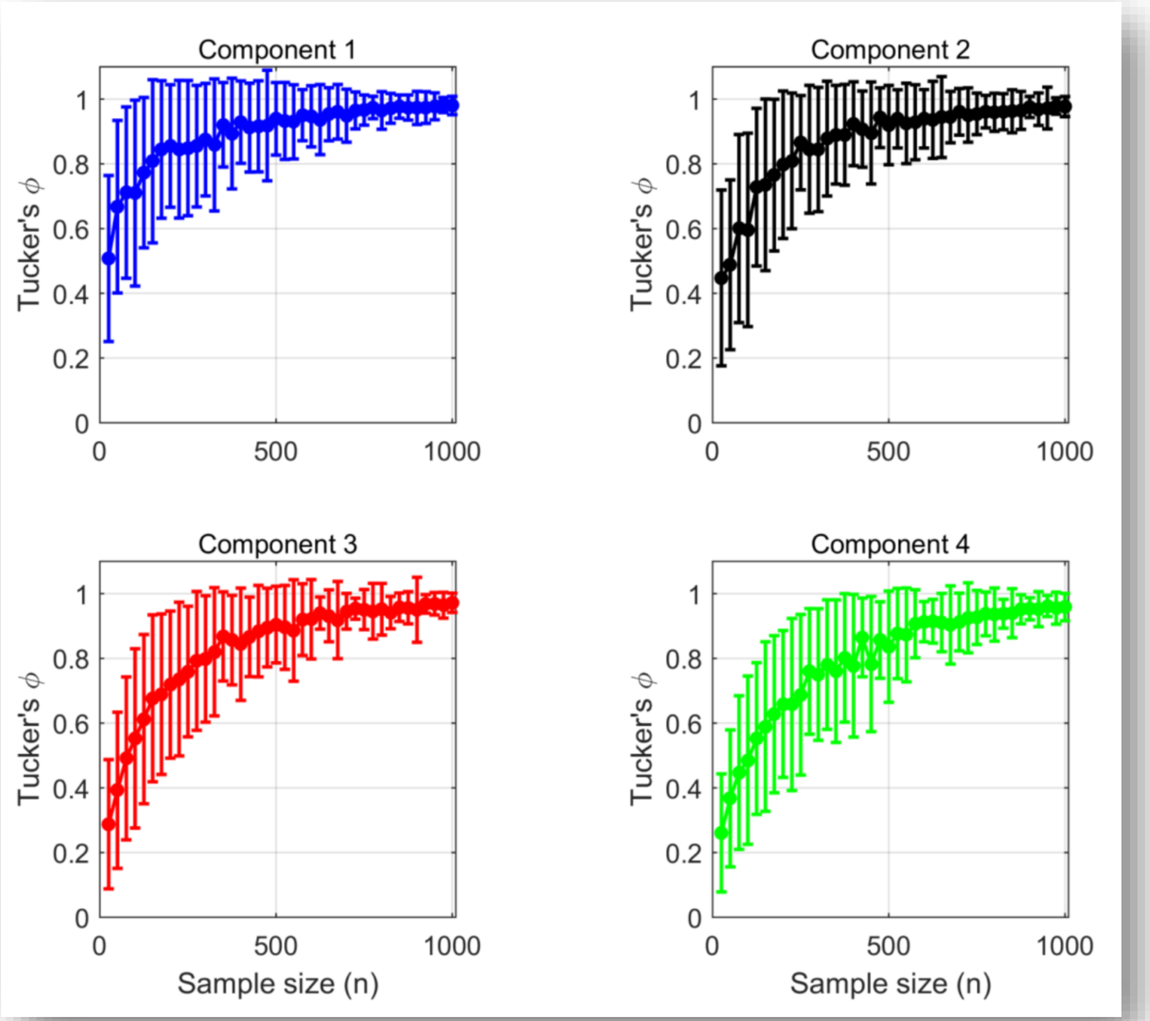
- ✓ Analyse variability and convergence

Congruence of sample loadings and population loadings

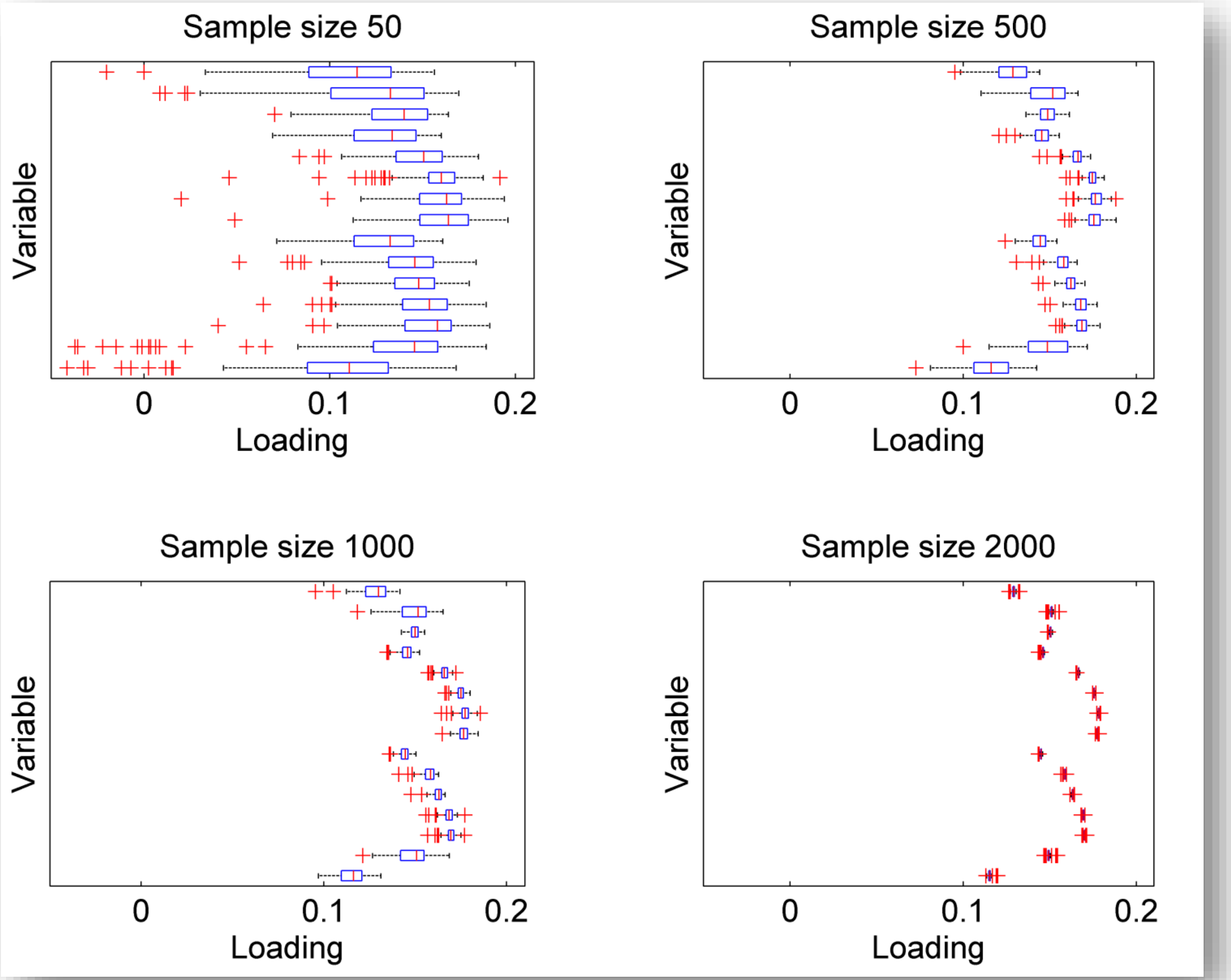


Simulated data under Spiked model

Congruence of sample loadings and population loadings



MS blood metabolomics data, 2139 × 133; Metabolights MTBLS93



Loading estimation and stability

	Serum metabolites MS (2139 × 133)		Serum metabolites qNMR (864 × 29)		Urine bucketed (733 × 1225)		Urine qNMR (343 × 62)	
Sample size *n*	ϕ	%*var*	ϕ	%*var*	ϕ	%*var*	ϕ	%*var*
5	0.331	44.5	0.825	77.5	0.651	62.9	0.783	86.1
25	0.507	20.8	0.977	71.9	0.893	41.8	0.744	68.8
50	0.657	19.2	0.99	70.1	0.941	39.3	0.798	61.8
100	0.759	17.7	0.996	70.3	0.979	37.8	0.855	57.7
150	0.795	16.5	0.998	70.2	0.988	37.9	0.95	57.8
200	0.876	16.1	0.998	70.4	0.989	37.3	0.967	56.8
250	0.871	15.5	0.999	70.6	0.994	37.2	0.989	56.3
300	0.919	15.6	0.999	70.6	0.995	37.3	0.996	56.2
350	0.915	15.1	0.999	70.7	0.996	37.2		
400	0.955	15.7	0.999	70.7	0.997	37.1		
450	0.933	14.9	0.999	70.5	0.998	37.1		
500	0.961	15.4	1	70.7	0.998	37.2		
550	0.97	15.3	1	70.6	0.999	37.1		
600	0.948	15.1	1	70.5	0.999	37		
650	0.974	15.1	1	70.6	1	37.1		
700	0.98	15	1	70.5	1	37		
750	0.978	15	1	70.6				
800	0.976	14.8	1	70.7				
850	0.983	15	1	70.6				
900	0.986	15.1	1	70.5				
950	0.989	15.1	1	70.6				
1000	0.988	14.9						
1050	0.987	15						

Tucker's congruence for different metabolomics data sets

Summary I

- $N \approx 200$ samples needed for stable loading estimation
- N increase with component order (less variance)

- Theoretical results for Covariance estimation:

$$N = O(p);\ N = O(p \times \log(p));\ N = O(\log(p))$$

  - (Vershynin, 2012; Adamczk, 2010; Rudelson, 1999; Gupta, 1987)

Retraction

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Retraction of "Unlocking Past Emotion: Verb Use Affects Mood and Happiness"

Psychological Science
1

DOI: 10.1177/0956797617692524
www.psychologicalscience.org/PS
SAGE

The following article has been retracted by the Editor and publishers of *Psychological Science*:

Hart, W. (2013). Unlocking past emotion: Verb use affects mood and happiness. *Psychological Science, 24*, 19–26. doi: 10.1177/0956797612446351

The retraction follows an investigation by the University of Alabama's Office for Research Compliance. That investigation found that a former graduate student in William Hart's lab altered the data in strategic ways. The investigation found that William Hart was unaware when the article was published that the data had been manipulated. William Hart cooperated in the investigation and agreed to this retraction.

Fox came to me to apologize after he admitted to the fabrication.

He described how and why he started tampering with data.

The first time it happened he had analyzed a dataset and the **results were just shy of significance**.

Fox noticed that if he duplicated a couple of cases and deleted a couple of cases, he could shift the p-value to below .05.

And so he did.

Fox recognized that the system rewarded him, and his collaborators, not for interesting research questions, or sound methodology, but for significant results.

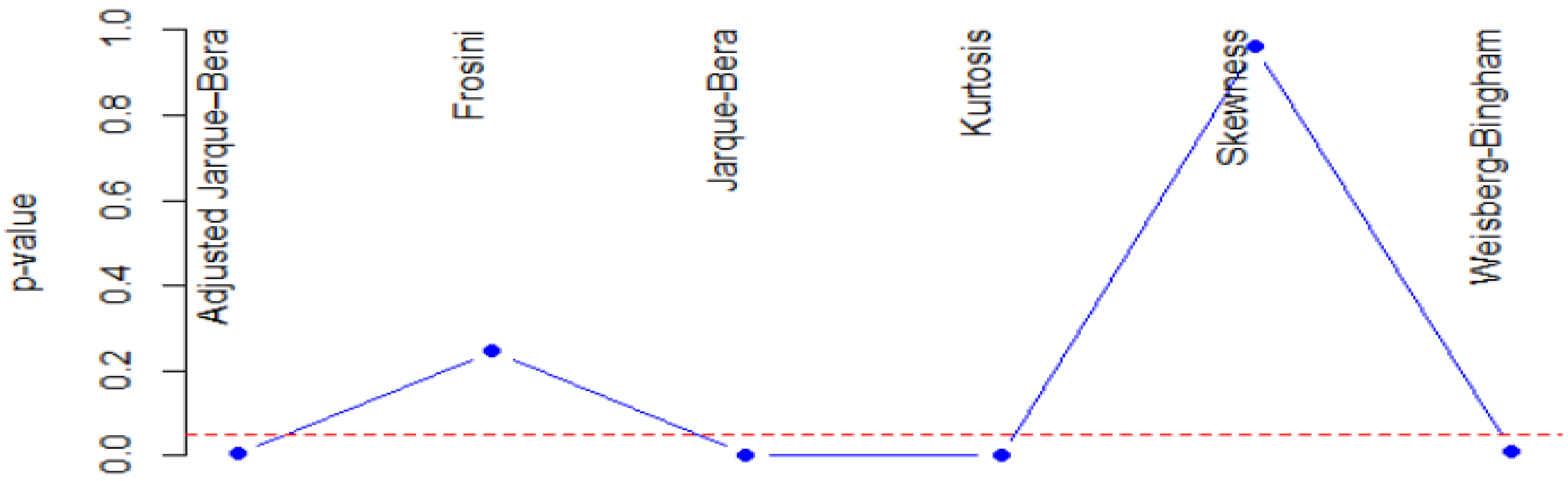
When he showed his collaborators the findings they were happy with them—and happy with Fox.

Goodhart's law

When a measure becomes a target, it ceases to be a good measure

Goodhart, C.A.E. (**1975**). *Problems of Monetary Management: The U.K. Experience*. Papers in Monetary Economics. Reserve Bank of Australia. **I**.

Change test to achieve significance



Keep changing testing procedure untill you find one that gives significant results

Important:

Pr (observation | hypothesis) ≠ Pr (hypothesis | observation)

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a “score” is committing an egregious logical error: **the transposed conditional fallacy.**

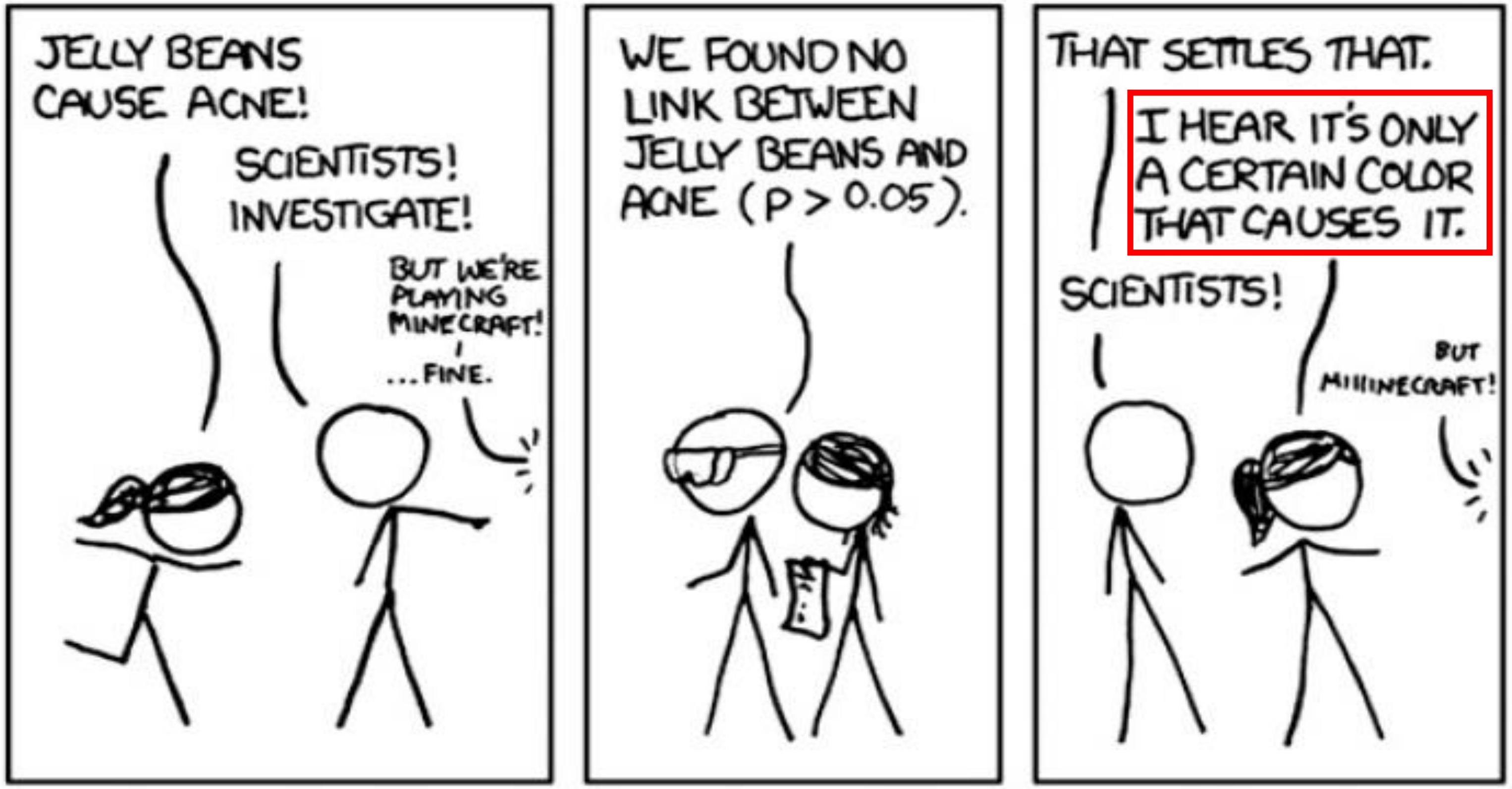
Source: https://rationalwiki.org/wiki/File:P-value_in_statistical_significance_testing.svg

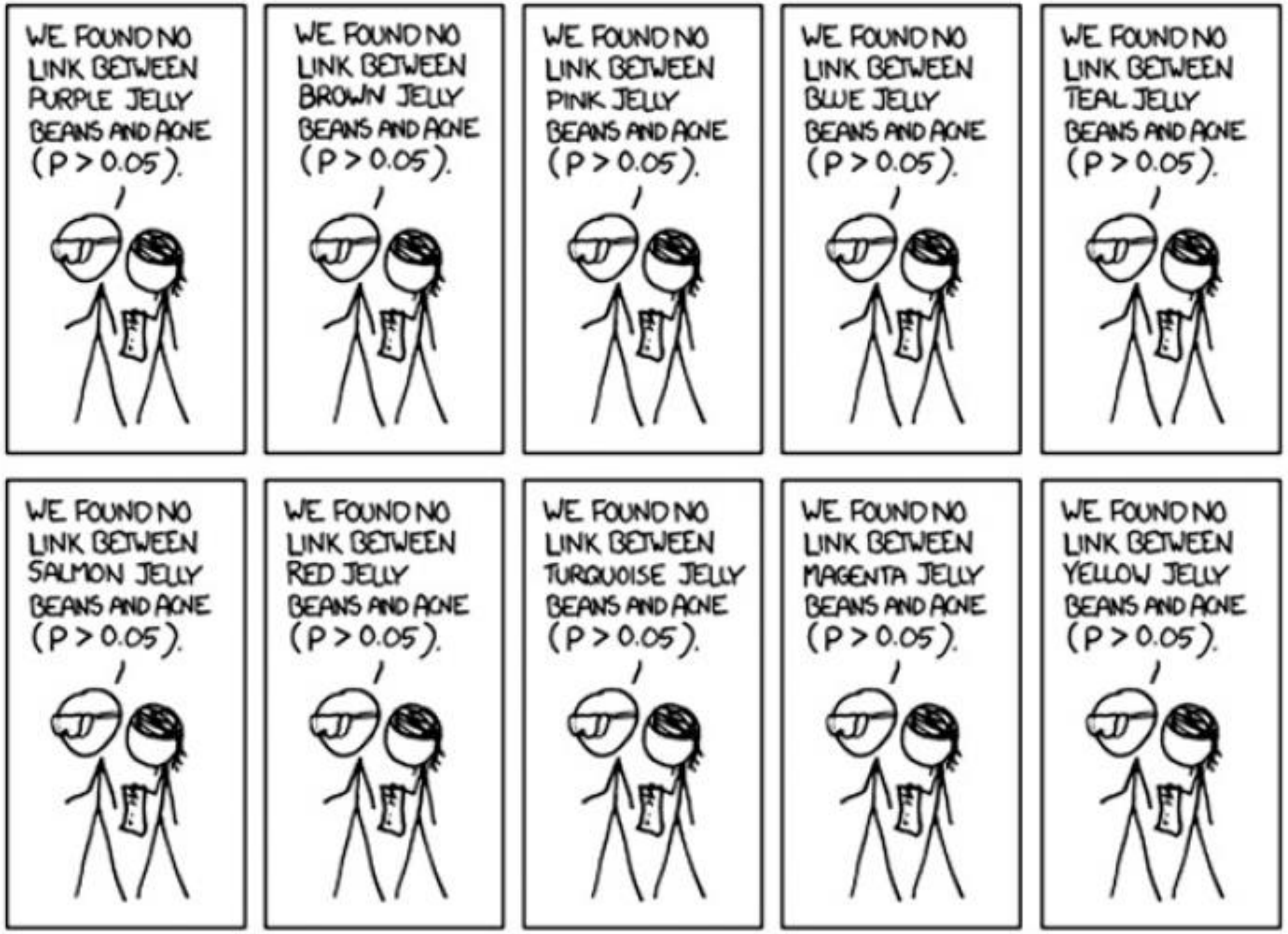
HARCKING

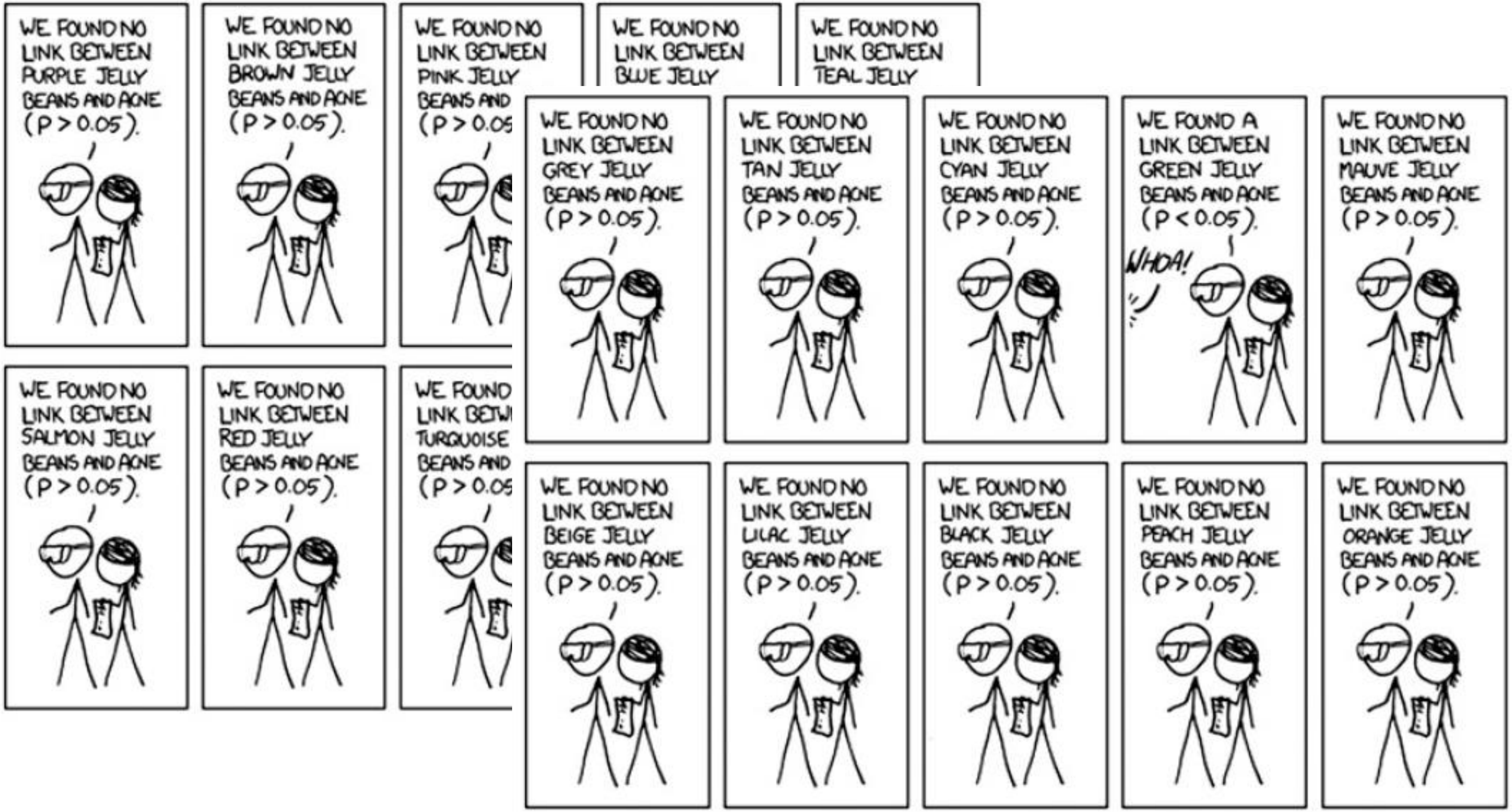
Hypothesiz***ing*** **A**fter the **R**esult Is **K**nown

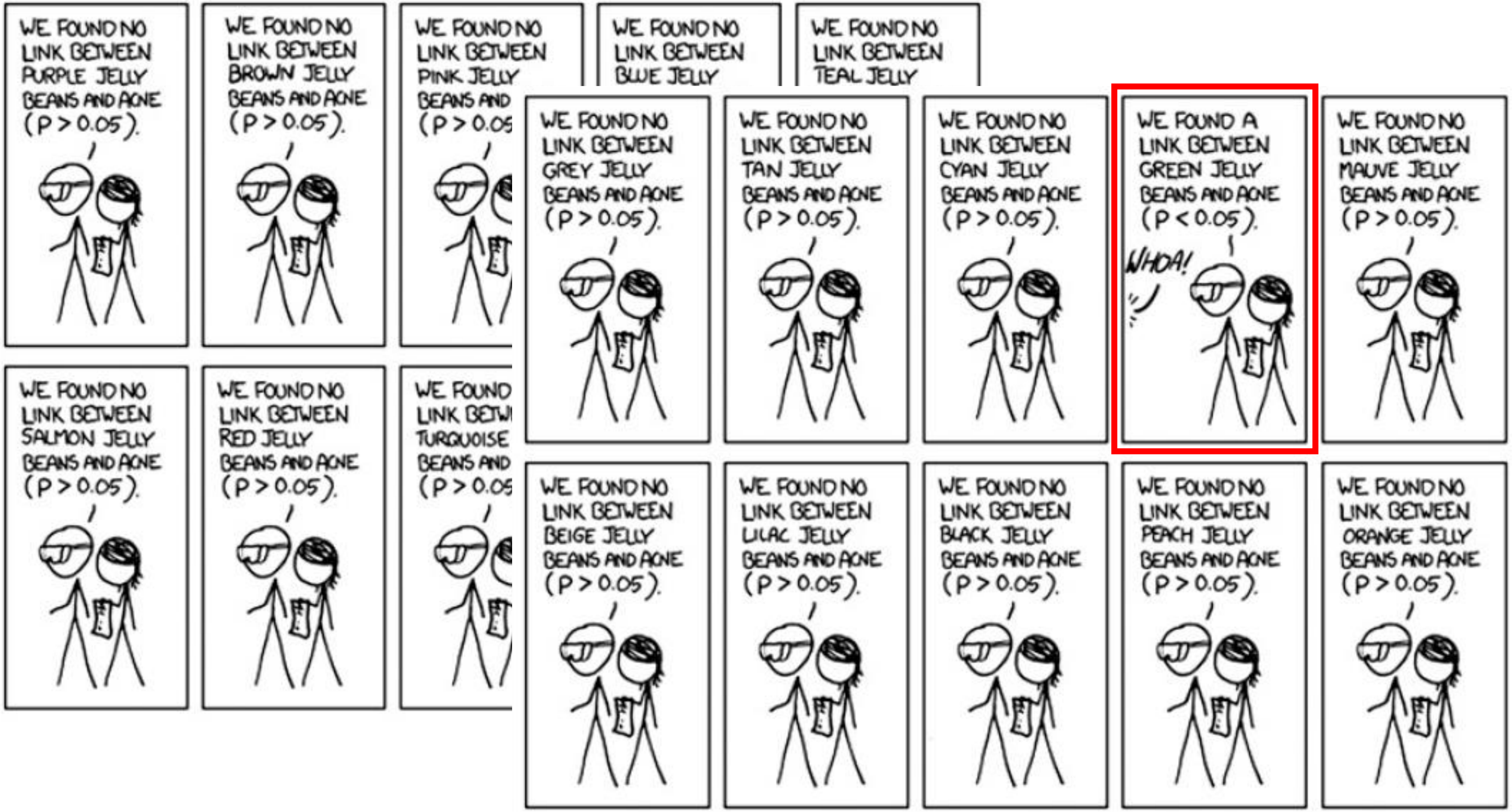
This is when you analyze the data many different ways (say different subgroups), discover an intriguing relationship and then publish the data so it appears that the hypothesis was stated before the data were collected

Kerr NL (**1998**) *HARKing: hypothesizing after the results are known*. Pers Soc Psychol Rev 2:196–217.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND
WE FOUND NO LINK BETWEEN BLUE JELLY
WE FOUND NO LINK BETWEEN TEAL JELLY
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND
TURQUOISE
BEANS AND
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE ($P < 0.05$).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE ($P > 0.05$).


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND (P > 0.05
WE FOUND NO LINK BETWEEN BLUE JELLY
WE FOUND NO LINK BETWEEN TEAL JELLY
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND LINK BETW TURQUOISE BEANS AND (P > 0.05
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

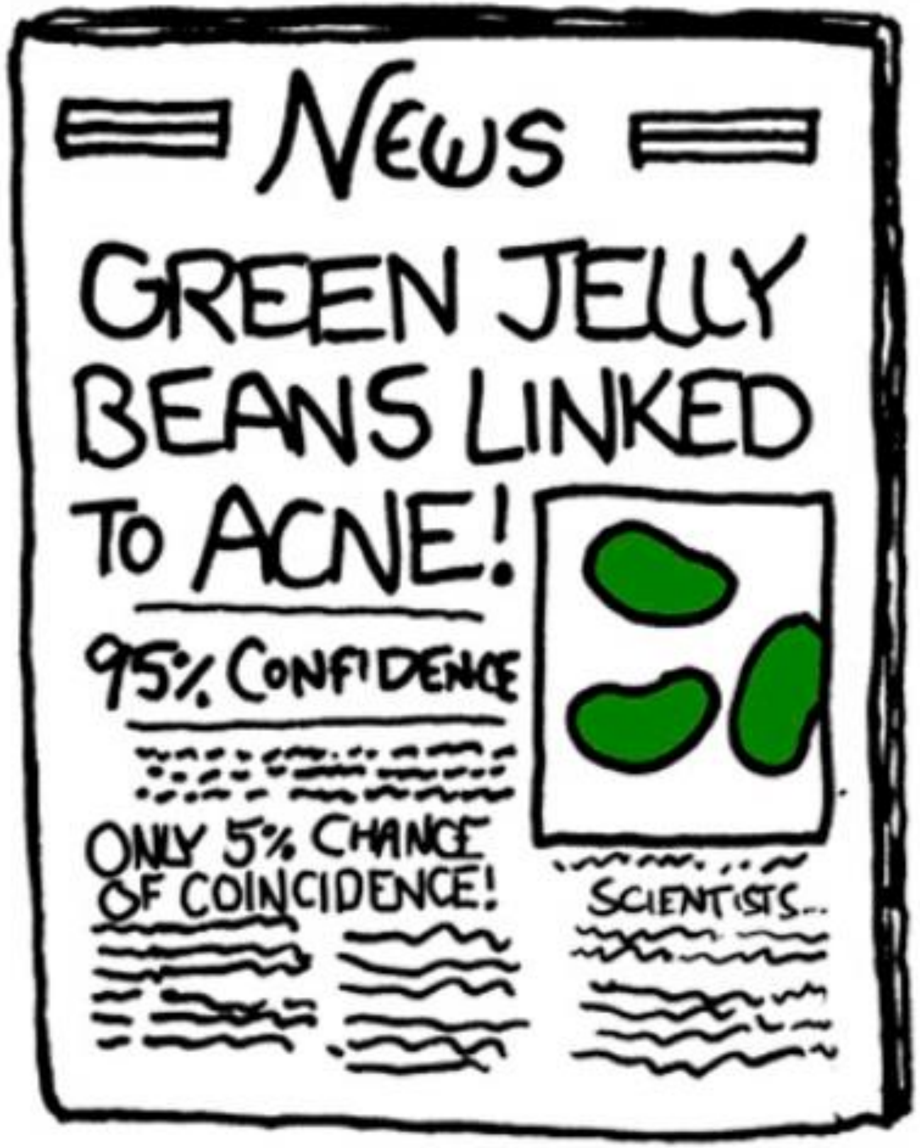


Fig. 3. The problem of HARKing. (Reprinted from http://xkcd.com/882 under the CC BY-NC 2.5 license.)

This is a multiple testing problem

If you decide to reject the null hypothesis at 0.05 level (5%)

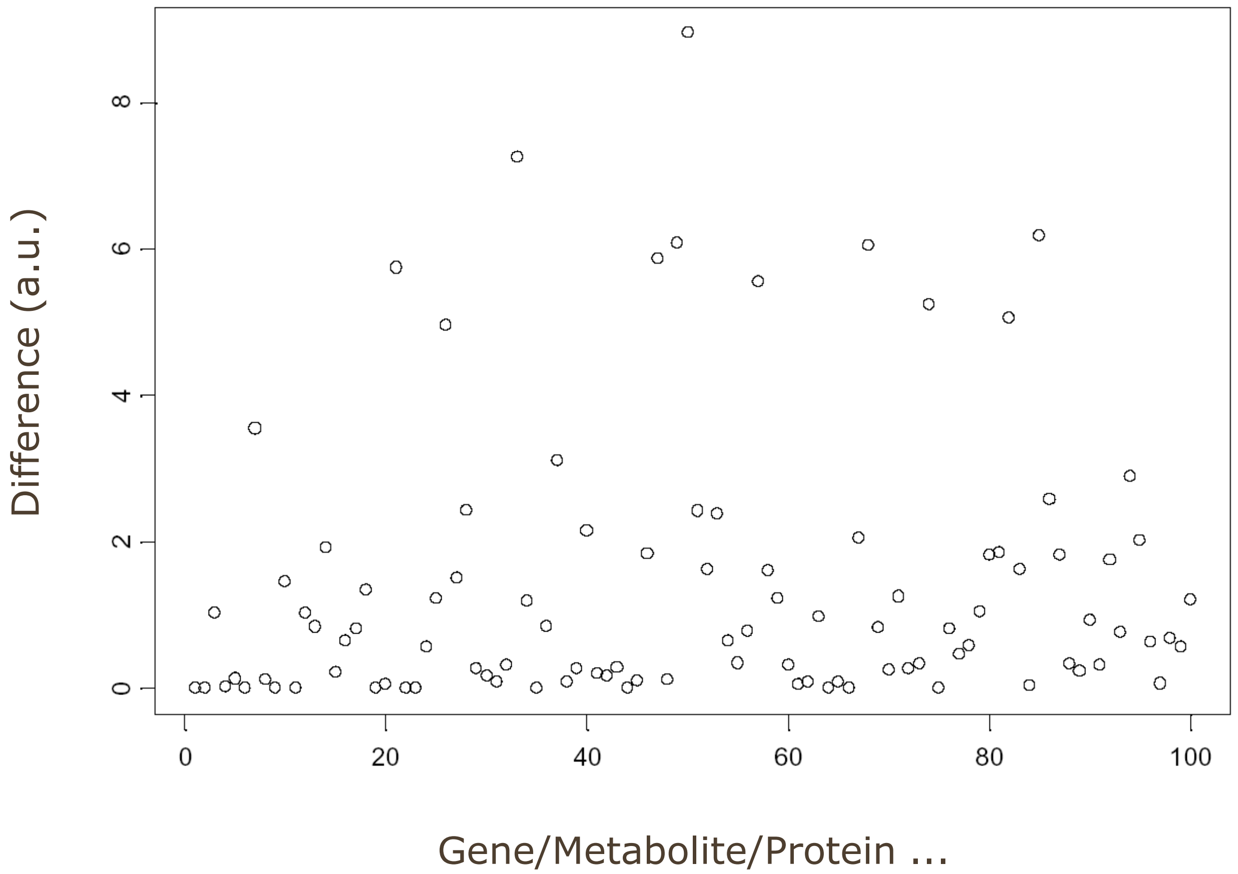
and you repeat the test 100 times (under H_0 true) you will fail to provide evidence for H_0 being true on average 5 times!

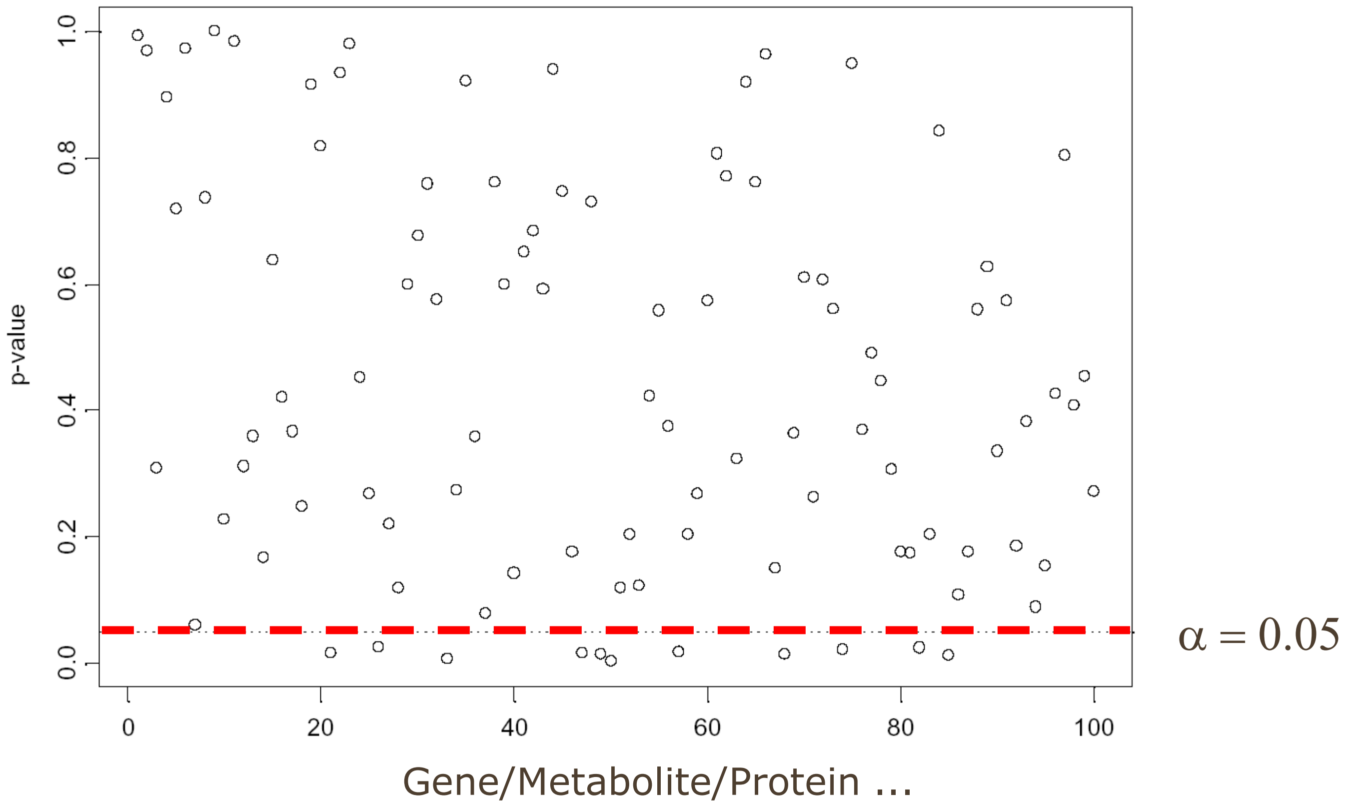
So keep testing until you find something.

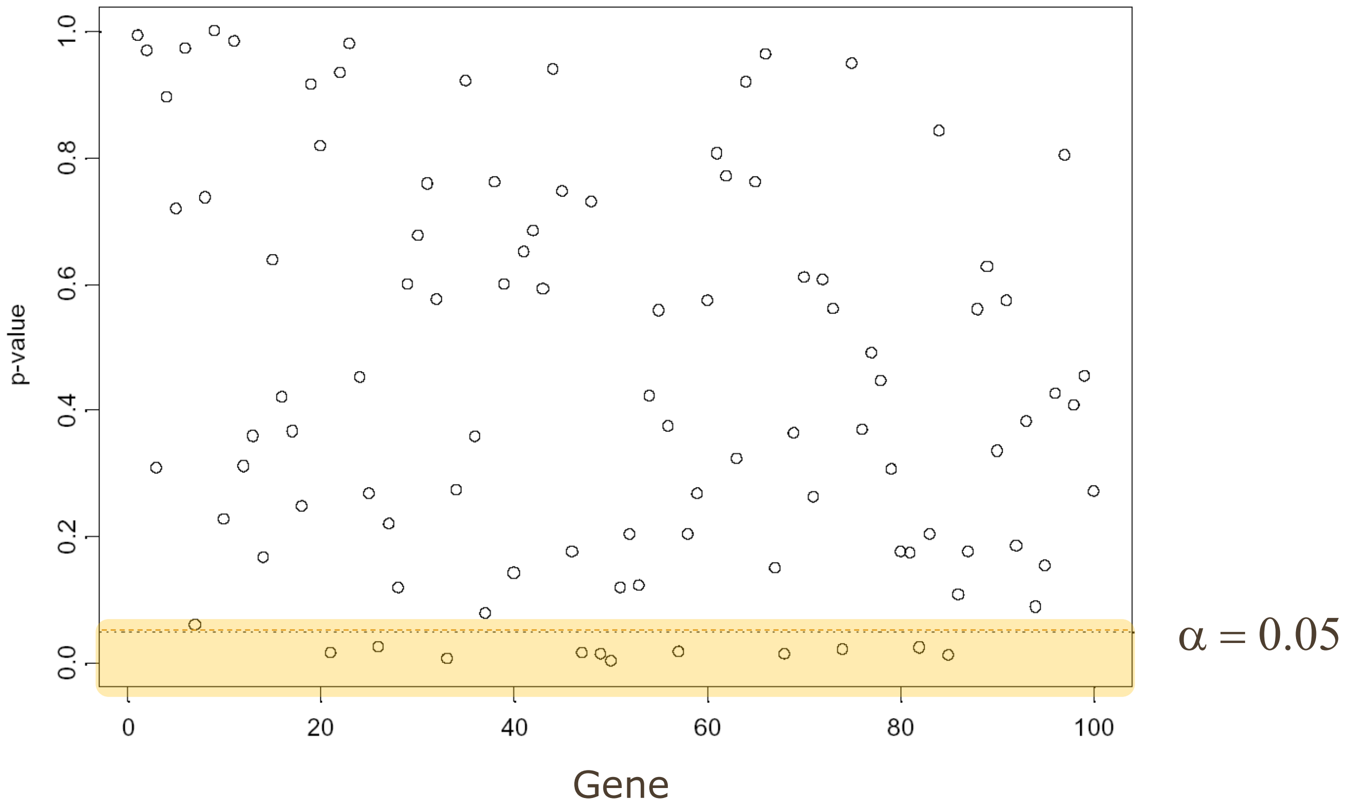
And stop there.

Example: Simulated experiment

- Screen of 100 genes/metabolites/proteins
- Control – treatment setting (2 groups)
- Simulate experiment results ***under null hypothesis*** (no difference)


Difference (a.u.)
0
2
4
6
8
0
20
40
60
80
100
Gene/Metabolite/Protein ...


p-value
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
Gene/Metabolite/Protein ...
$\alpha = 0.05$



Expect 0.05 x 100 = 5 rejections of null hypothesis *by chance*

Here 11!

Multiple testing

- Increased risk of false positives
- Claiming a difference which is not true
- Need to correct for this
- Must take this into account when planning your experiments

To finish

- 3 is almost never a magical number and is NOT an appropriate sample size