«Honni soit qui mal y science» A little stroll through science, bad science... and statistics

Guy Tremblay
Professeur titulaire
Département d'informatique

http://www.labunix.uqam.ca/~tremblay_gu

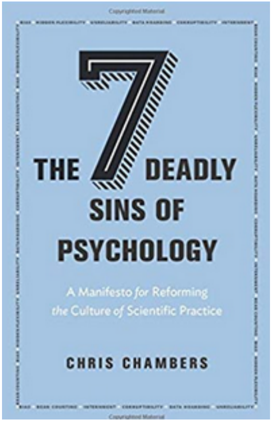
Dept. of CS & SE
Concordia University
October 29, 2019

Have you ever noticed that all the instruments searching for intelligent life are pointed away from earth?

Outline

1. Why this seminar ?
2. Is science in crisis ?
3. Some basic statistical concepts
4. Scientific method and statistical inference
5. Some causes of the crisis
   - Focus on «positive» and «novel» results (aka. «Publication bias»)
   - Flexibility in choosing experiment protocols and analyses
   - Other aspects
6. Conclusion : Some possible solutions ?

Three interesting books published in recent years. . .

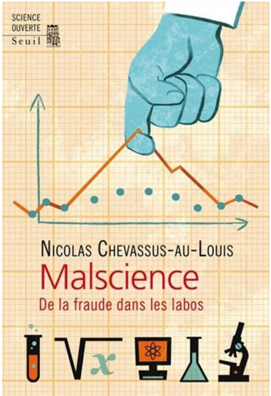


THE 7 DEADLY SINS OF PSYCHOLOGY

A Manifesto for Reforming the Culture of Scientific Practice

CHRIS CHAMBERS

(Chambers, 2017)



SCIENCE OUVERTE
Seuil

Nicolas Chevassus-au-Louis

Malscience

De la fraude dans les labos

(Chevaussus-au-Louis, 2016)

THE IRREPRODUCIBILITY CRISIS OF MODERN SCIENCE

Causes, Consequences, and the Road to Reform

DAVID RANDALL AND CHRISTOPHER WELSER
NATIONAL ASSOCIATION OF SCHOLARS
APRIL 2018
ISBN: 978-0-9986635-5-5

NAS

(NAS, 2018)

«Malscience» = «Badscience»

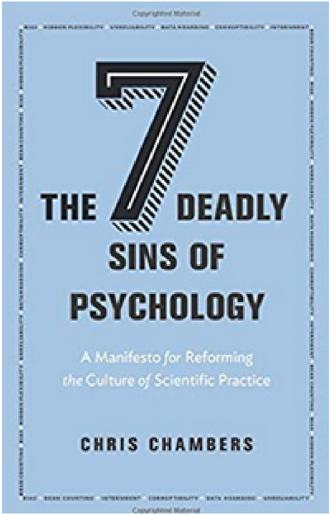
This talk will discuss «malscience» . . . not necessarily fraud

THE IRREPRODUCIBILITY CRISIS OF MODERN SCIENCE

Causes, Consequences, and the Road to Reform

DAVID RANDALL AND CHRISTOPHER WELSER
NATIONAL ASSOCIATION OF SCHOLARS
APRIL 2018
ISBN: 978-0-9986635-5-5

NAS



THE 7 DEADLY SINS OF PSYCHOLOGY

A Manifesto for Reforming the Culture of Scientific Practice

CHRIS CHAMBERS

What’s in it for CS/SE researchers ?

- In the last 15–20 years, the field of *Empirical Software Engineering* has been blossoming
  - *Empirical Software Engineering* (Journal, 1996)
  - *Evaluation and Assessment in Software Engineering* (Conférence, 1996)
  - *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (Conférence, 2007)
  - Guéhéneuc YG., Khomh F. (2019) Empirical Software Engineering. In : Cha S., Taylor R., Kang K. (eds) *Handbook of Software Engineering*. Springer, Cham,

⇒ More frequent use of «experimentations»

What’s in it for CS/SE researchers ?

- In the last 15–20 years, the field of *Empirical Software Engineering* has been blossoming
  - *Empirical Software Engineering* (Journal, 1996)
  - *Evaluation and Assessment in Software Engineering* (Conférence, 1996)
  - *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (Conférence, 2007)
  - Guéhéneuc YG., Khomh F. (2019) Empirical Software Engineering. In : Cha S., Taylor R., Kang K. (eds) *Handbook of Software Engineering*. Springer, Cham,

⇒ More frequent use of «experimentations»

- Experimentations

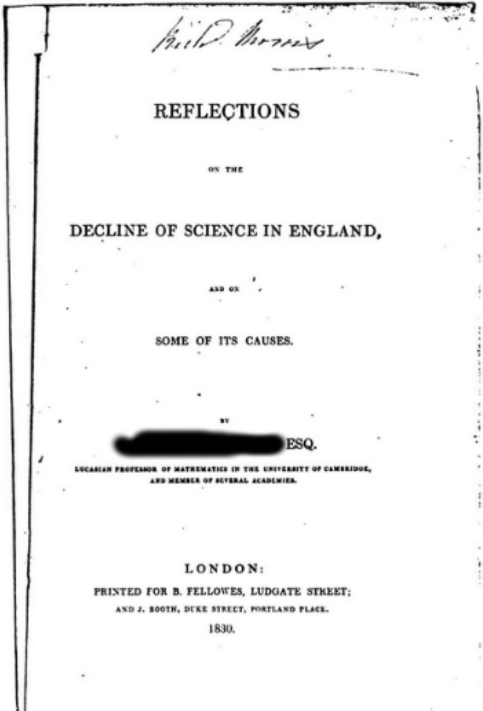
  ⇒ Irregular or random phenomena (people, contexts, etc.)

  + Experimental errors

  + Use of samples

  ⇒ Use of statistical methods and inferences

Did you know there is a (very !) old book on «malscience» written by a «*computer scientist*» ?



REFLECTIONS

ON THE

DECLINE OF SCIENCE IN ENGLAND,

AND ON

SOME OF ITS CAUSES.

BY

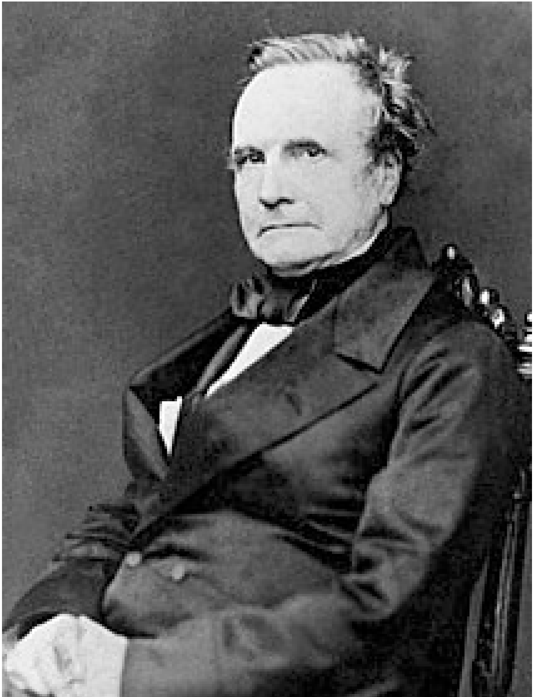
[redacted] ESQ.

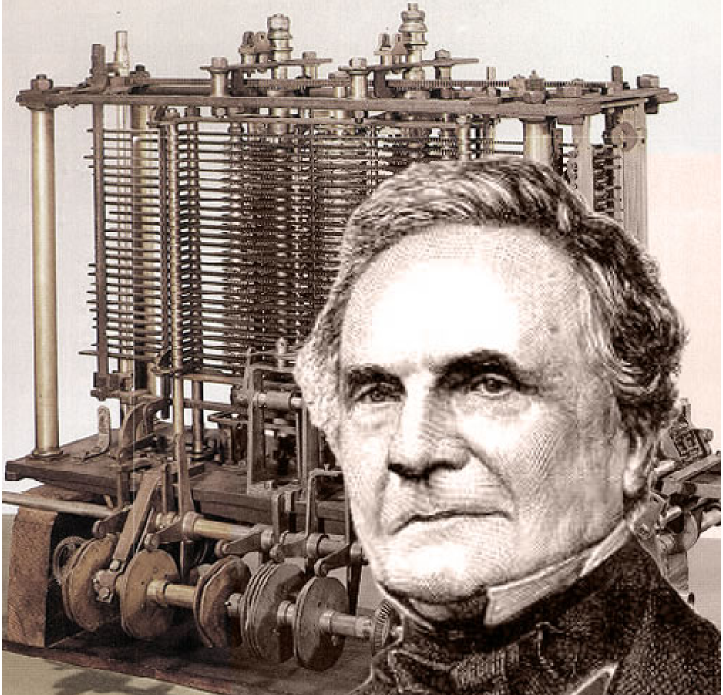
LUCASIAN PROFESSOR OF MATHEMATICS IN THE UNIVERSITY OF CAMBRIDGE,
AND MEMBER OF SEVERAL ACADEMIES.

LONDON:
PRINTED FOR B. FELLOWES, LUDGATE STREET;
AND J. BOOTH, DUKE STREET, PORTLAND PLACE.
1830.

Four «species» of «bad science»

1. Hoaxing
2. Forging (data)
3. Trimming (data)
4. Cooking (data)





REFLECTIONS

ON THE

DECLINE OF SCIENCE IN ENGLAND,

AND ON

SOME OF ITS CAUSES.

BY

CHARLES BABBAGE, ESQ.

LUCASIAN PROFESSOR OF MATHEMATICS IN THE UNIVERSITY OF CAMBRIDGE,
AND MEMBER OF SEVERAL ACADEMIES.

Outline

1. Why this seminar ?
2. Is science in crisis ?
3. Some basic statistical concepts
4. Scientific method and statistical inference
5. Some causes of the crisis
   - Focus on «positive» and «novel» results (aka. «Publication bias»)
   - Flexibility in choosing experiment protocols and analyses
   - Other aspects
6. Conclusion : Some possible solutions ?

Outline

1. Why this seminar ?
2. Is science in crisis ?
3. Some basic statistical concepts
4. Scientific method and statistical inference
5. Some causes of the crisis
   - Focus on «positive» and «novel» results (aka. «Publication bias»)
   - Flexibility in choosing experiment protocols and analyses
   - Other aspects
6. Conclusion : Some possible solutions ?

The Irreproducibility Crisis Report

Causes, Consequences, and the Road to Reform

A reproducibility crisis afflicts a wide range of scientific and social-scientific disciplines, from epidemiology to social psychology. [. . .] Many supposedly scientific results cannot be reproduced reliably in subsequent investigations, and offer no trustworthy insight into the way the world works.

National Association of Scholars, 2018

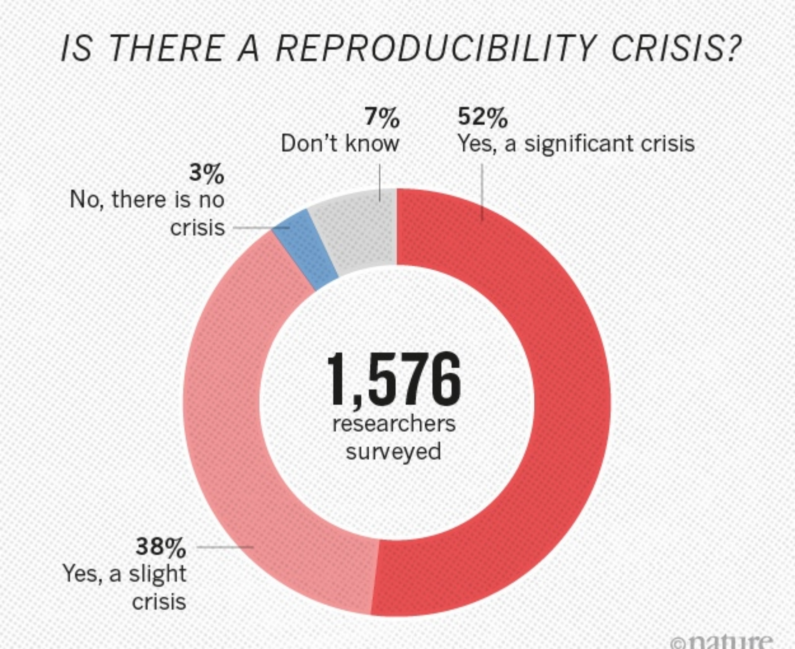
THE IRREPRODUCIBILITY
CRISIS OF MODERN SCIENCE

Causes, Consequences, and the Road to Reform

DAVID RANDALL AND CHRISTOPHER WELSER
NATIONAL ASSOCIATION OF SCHOLARS
APRIL 2018
ISBN: 978-0-9986635-5-5

Survey conducted by *Nature* (2016)

https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970



2005 : Paper on «false» research results

Occur often in the medical field according to the author’s analysis

Open access, freely available online

Essay

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

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may

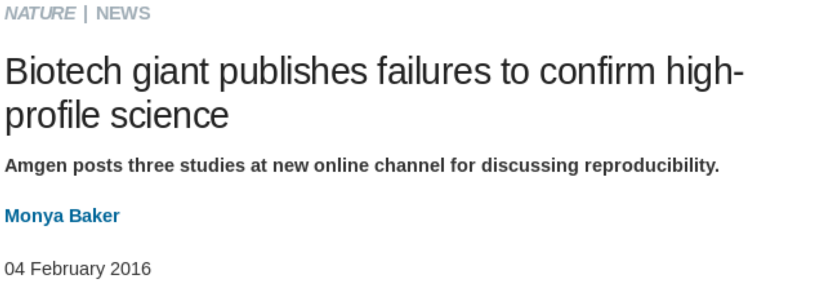
---

«*Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true.*

[. . .]

[This is in part because of the] 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***.*»

2012 : Paper on non reproducibility of cancer studies



NATURE | NEWS

Biotech giant publishes failures to confirm high-profile science

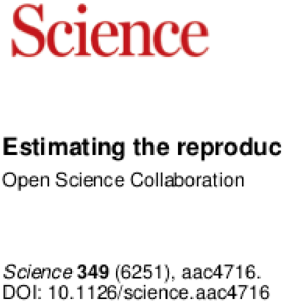
Amgen posts three studies at new online channel for discussing reproducibility.

Monya Baker

04 February 2016

Amgen researchers made headlines when they declared that they had been unable to reproduce the findings in 47 of 53 «landmark» [cancer and hematology] papers.

2015 : Paper on non reproducibility of psychology studies



Science

Estimating the reproducibility of psychological science

Open Science Collaboration

Science **349** (6251), aac4716.
DOI: 10.1126/science.aac4716

«Aarts et al. describe the replication of 100 experiments reported in papers published in 2008 in three high-ranking psychology journals. [. . .] they find that about one-third to one-half of the original findings were also observed in the replication study [donc 50–60% non reproductibles].»

Note that reproducibility is also an issue in software engineering… although often ignored ☹

> *Routinely, we are told Tool X or Technique Y is a panacea to many of software engineering's problems, but where is the accompanying empirical evidence that can stand scrutiny, that has been verified by an independent research team?*

«*Replication's Role in Software Engineering*», Brook et al., Chap. 14 [SSS08]

2016 : B. Wansik’s «Disastrous blog post»

- Former Cornell professor — nutrition science, consumer behavior
- Former USDA Center for Nutrition Policy and Promotion Executive Director
- Over 20 000 citations !
- But. . .

2016 : B. Wansik's «Disastrous blog post»

- Former Cornell professor — nutrition science, consumer behavior
- Former USDA Center for Nutrition Policy and Promotion Executive Director
- Over 20 000 citations !
- But since 2017 : 17 papers were retracted by journals, including 6 (in a single day) by the *Journal of the American Medical Association*

2016 : B. Wansik’s «Disastrous blog post»

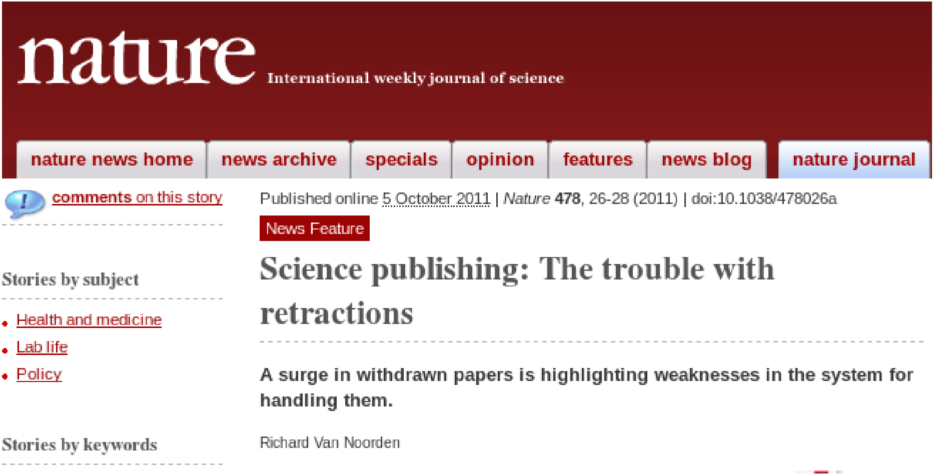
When [this graduate student] arrived, I gave her a data set of a [. . .] ***failed study*** *which had null results [. . .]. I said, “This cost us a lot of time and our own money to collect. There’s got to be something here we can salvage because it’s a cool (rich & unique) data set.”*

I had three ideas for potential Plan B, C, & D directions (since Plan A [the one-month study with null results] had failed). I told her what the analyses should be and what the tables should look like. [. . .] Six months after arriving, . . . [she] had one paper accepted, two papers with revision requests, and two others that were submitted (and were eventually accepted).

www.benitaepstein.com
© Benita Epstein
Benita Epstein

"I already wrote the paper.
That's why it's so hard to
get the right data."

Another symptom : Increase in the number of retracted papers



Published online 5 October 2011 | *Nature* **478**, 26-28 (2011) | doi:10.1038/478026a

News Feature

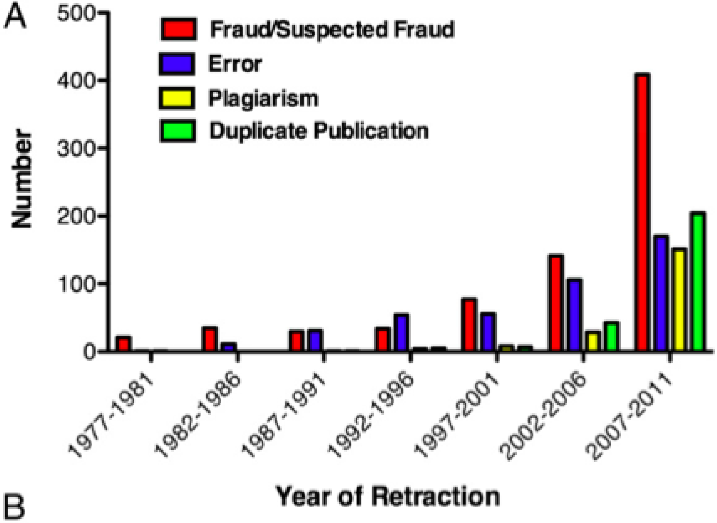
Science publishing: The trouble with retractions

A surge in withdrawn papers is highlighting weaknesses in the system for handling them.

Richard Van Noorden

- Number of retracted papers ≈ 10–12 times more !
- Prestigious journals (e.g., Science, Nature, Cell) are the most affected by this phenomena !

Another symptom : Increase in the number of retracted papers

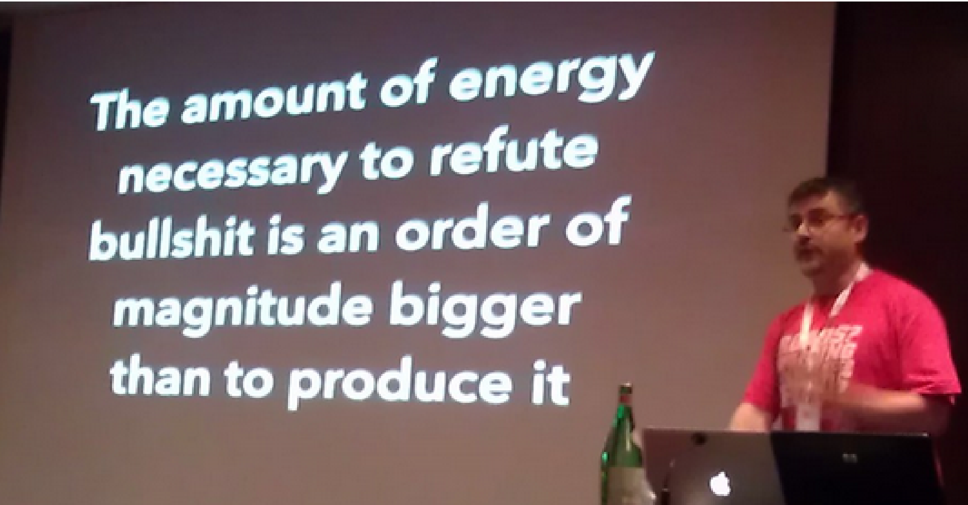


A key problem = Retracting a paper generally has. . . little impact ☹

Brandolino’s law = ?

A key problem = Retracting a paper generally has. . . little impact ☹

Brandolino’s law = *Bullshit asymetry principle*

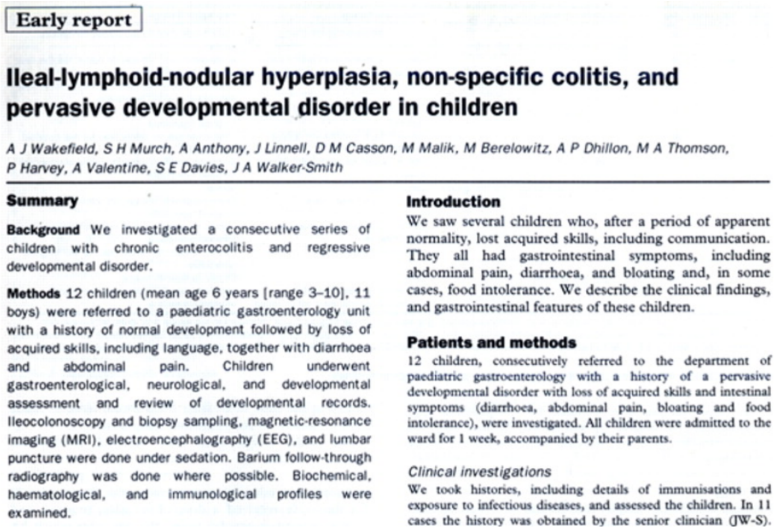


The amount of energy necessary to refute bullshit is an order of magnitude bigger than to produce it

Any example in mind?

A famous example : Lancet's paper (1998) on links between autism and MMR vaccine

MMR = Measles, Mumps, and Rubella



Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

Summary

Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Introduction

We saw several children who, after a period of apparent normality, lost acquired skills, including communication. They all had gastrointestinal symptoms, including abdominal pain, diarrhoea, and bloating and, in some cases, food intolerance. We describe the clinical findings, and gastrointestinal features of these children.

Patients and methods

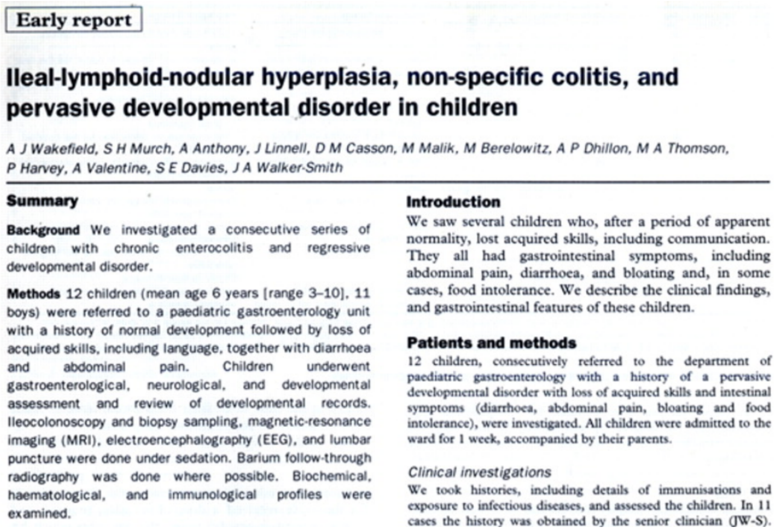
12 children, consecutively referred to the department of paediatric gastroenterology with a history of a pervasive developmental disorder with loss of acquired skills and intestinal symptoms (diarrhoea, abdominal pain, bloating and food intolerance), were investigated. All children were admitted to the ward for 1 week, accompanied by their parents.

Clinical investigations

We took histories, including details of immunisations and exposure to infectious diseases, and assessed the children. In 11 cases the history was obtained by the senior clinician (JW-S).

A famous example : Lancet's paper (1998) on links between autism and MMR vaccine

MMR = Measles, Mumps, and Rubella



Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

Summary

Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Introduction

We saw several children who, after a period of apparent normality, lost acquired skills, including communication. They all had gastrointestinal symptoms, including abdominal pain, diarrhoea, and bloating and, in some cases, food intolerance. We describe the clinical findings, and gastrointestinal features of these children.

Patients and methods

12 children, consecutively referred to the department of paediatric gastroenterology with a history of a pervasive developmental disorder with loss of acquired skills and intestinal symptoms (diarrhoea, abdominal pain, bloating and food intolerance), were investigated. All children were admitted to the ward for 1 week, accompanied by their parents.

Clinical investigations

We took histories, including details of immunisations and exposure to infectious diseases, and assessed the children. In 11 cases the history was obtained by the senior clinician (JW-S).

- Cited more than 700 times (upto 2000)

The paper was retracted in 2010

- Paper was retracted following an investigation (2004–10 !) by B. Deer, a Sunday Times journalist

- Among the 12 children mentioned in the paper :
  - 3 had no autism symptoms
  - 5 developed the symptoms before receiving the vaccine

- Key info omitted from paper : All tests on presence of measle ARN (made by Wakefield's assistant) were negative !

And now (2019). . .

A. Wakefield

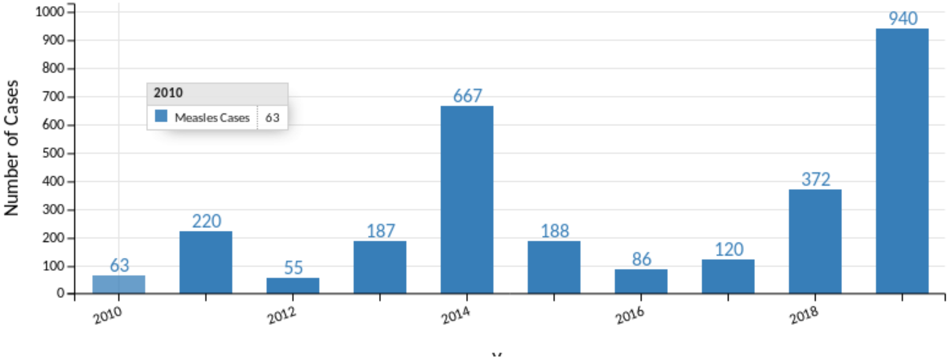
- United Kingdom : Banned from medical practice
- USA : Works as medical advisor for anti-vaccine associations

And now (2019). . .

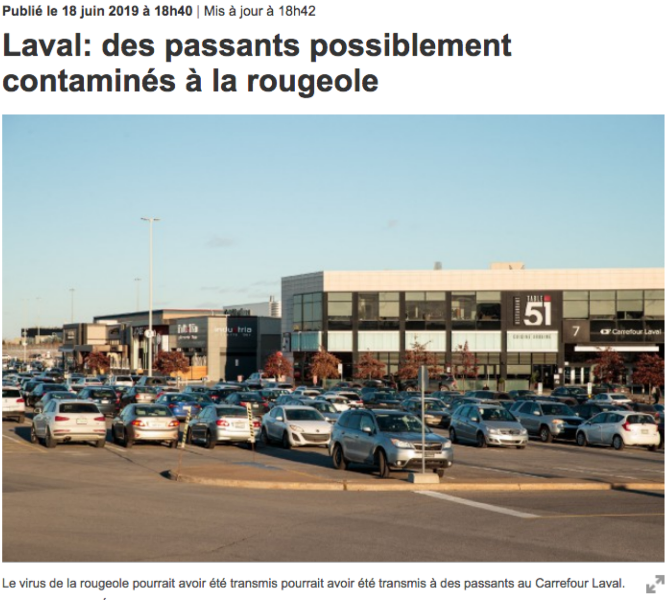
Number of cases in USA — Similar trend in many other countries ☹

Number of Measles Cases Reported by Year

2010-2019**(as of May 24, 2019)



And now (2019) : La Presse, 18 juin 2019



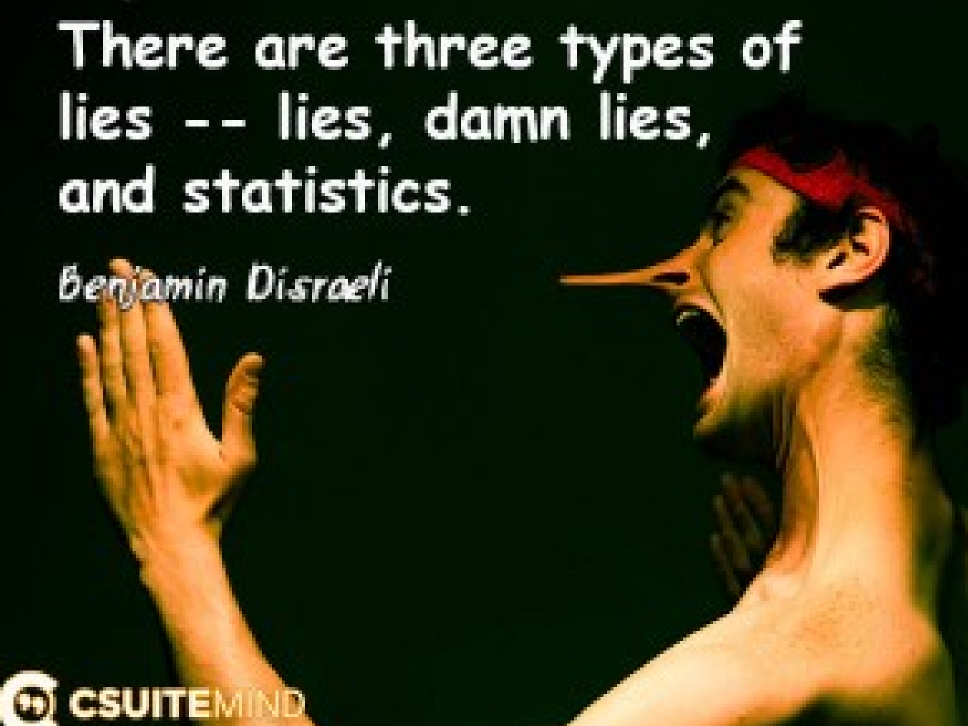
Publié le 18 juin 2019 à 18h40 | Mis à jour à 18h42

Laval: des passants possiblement contaminés à la rougeole

Le virus de la rougeole pourrait avoir été transmis pourrait avoir été transmis à des passants au Carrefour Laval.

Outline

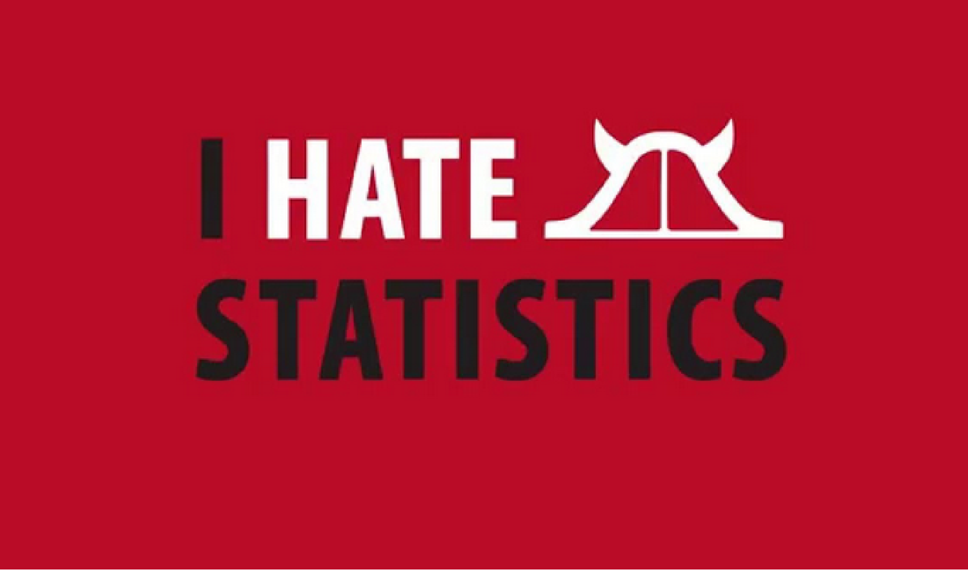
1. Why this seminar ?
2. Is science in crisis ?
3. **Some basic statistical concepts**
4. Scientific method and statistical inference
5. Some causes of the crisis
   - Focus on «positive» and «novel» results (aka. «Publication bias»)
   - Flexibility in choosing experiment protocols and analyses
   - Other aspects
6. Conclusion : Some possible solutions ?



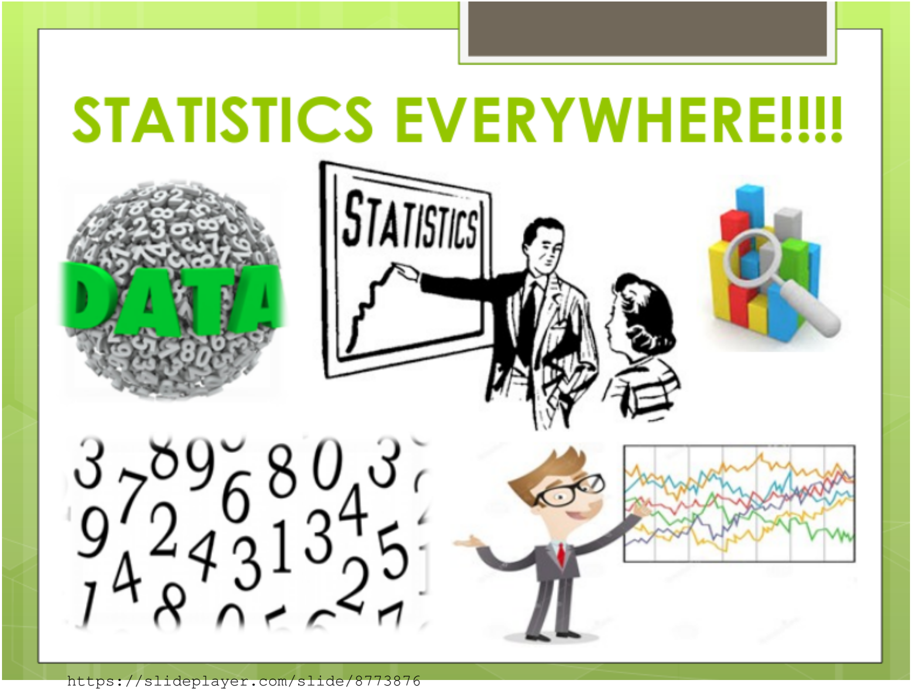
There are three types of lies -- lies, damn lies, and statistics.

Benjamin Disraeli

Do you like statistics？



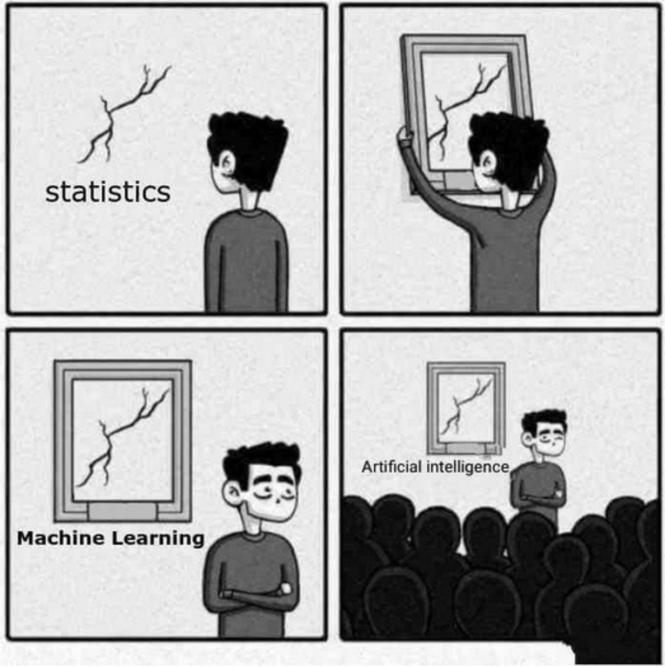
https://www.youtube.com/watch?v=ldy9RiRRZ3Y



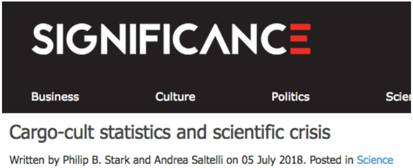
STATISTICS EVERYWHERE!!!!

https://slideplayer.com/slide/8773876

http://towardsdatascience.com

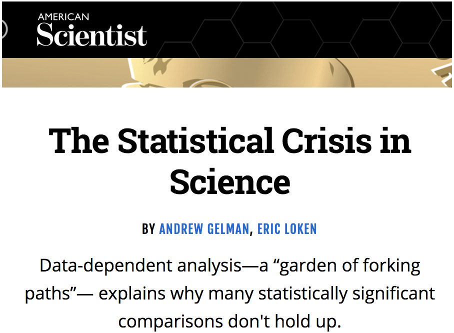


The use — or bad use ! ? — of statistics plays a key role in the crisis in science



Cargo-cult statistics and scientific crisis

Written by Philip B. Stark and Andrea Saltelli on 05 July 2018. Posted in Science



The Statistical Crisis in Science

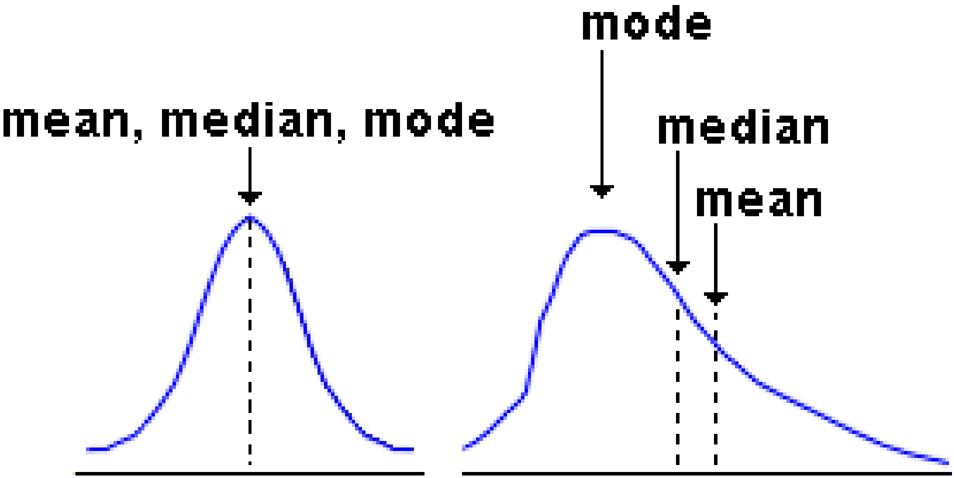
BY ANDREW GELMAN, ERIC LOKEN

Data-dependent analysis—a “garden of forking paths”— explains why many statistically significant comparisons don't hold up.

Central tendency measures

Central tendency measure = Value around which most data is centered

https://vula.uct.ac.za



Central tendency measure = Value around which most data is centered ⋆

Mean

Let $xs = \{x_0, x_1, \ldots, x_{n-1}\}$ (multiset !)

$$Mean(xs) = \frac{\sum_{i=0}^{n-1} x_i}{n}$$

Family income in USA

$\text{Mean} \approx 0.9 \times 34\,074\$ + 0.1 \times 312\,536\$ = 61\,920\$$

Average U.S. Household Income In 2015

The top 10 percent averaged more than nine times as much income as the bottom 90 percent. Americans in the top 1 percent averaged over 40 times more income than the bottom 90 percent.

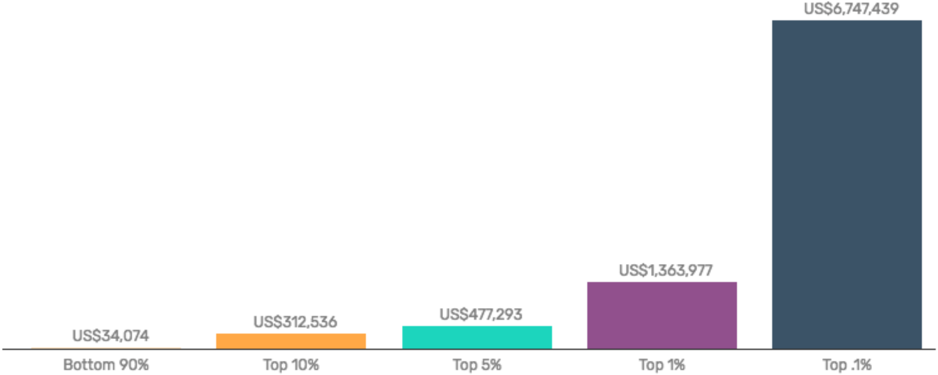
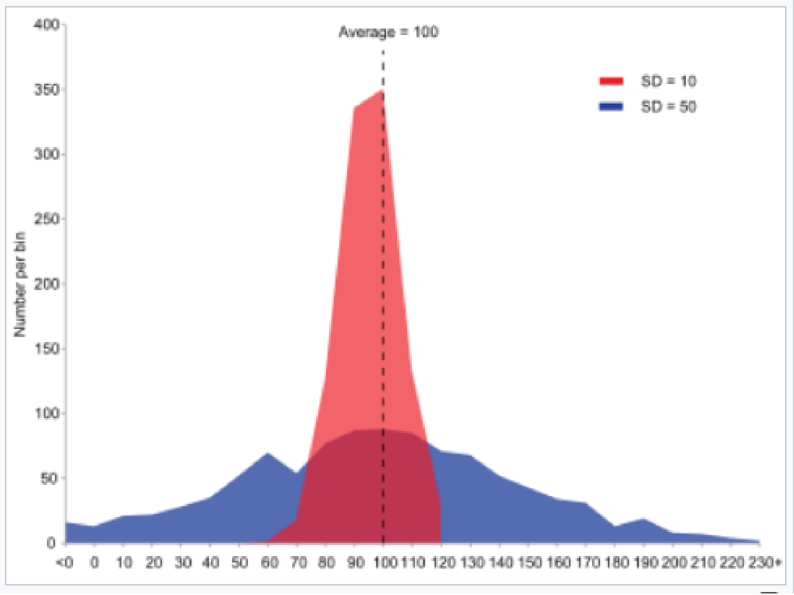


Chart: The Balance • Source: Inequality.org

Dispersion measures

Dispersion measure = Describes variability among the various values

https://en.wikipedia.org/wiki/Statistical_dispersion



Dispersion measure = Describes variability among the various values

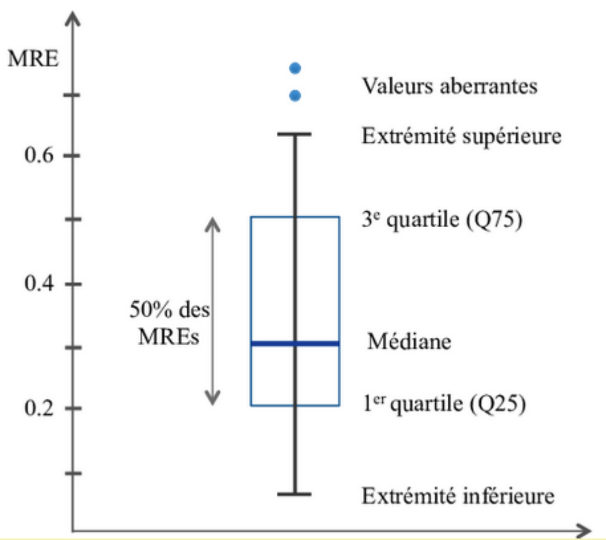
Standard deviation

Let $xs = \{x_0, x_1, \ldots, x_{n-1}\}$ and $m = \text{Mean}(xs)$

$$\text{Sd}(xs) = \sqrt{\frac{\displaystyle\sum_{i=0}^{n-1}(x_i - m)^2}{n-1}}$$

Representation that combine central tendency, dispersion, and distribution

The Boxplot


MRE
0.6
0.4
0.2
Valeurs aberrantes
Extrémité supérieure
3e quartile (Q75)
50% des MREs
Médiane
1er quartile (Q25)
Extrémité inférieure

Association measure

Often used assocation measure = Linear regression coefficient

Describes the correlation between two measures

«*standardized way of describing the amount by which [two measures] covary*»

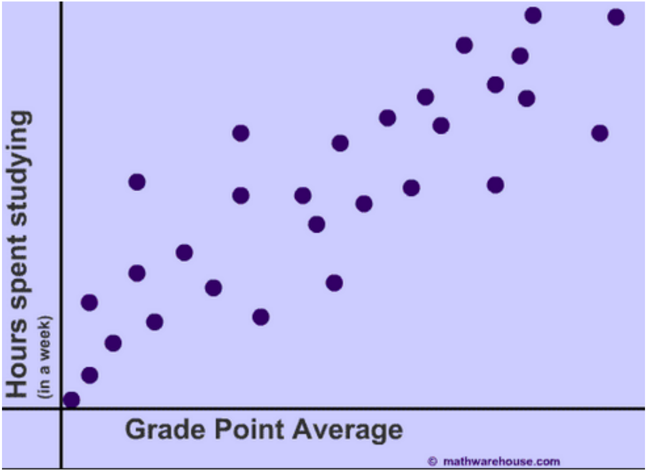
«*Statistical Methods and Measurement*», J. Rosenberg [SSS08]

Correlation examples — positive

Number of hours of study vs. academic result

https://www.mathwarehouse.com/statistics/correlation-coefficient/

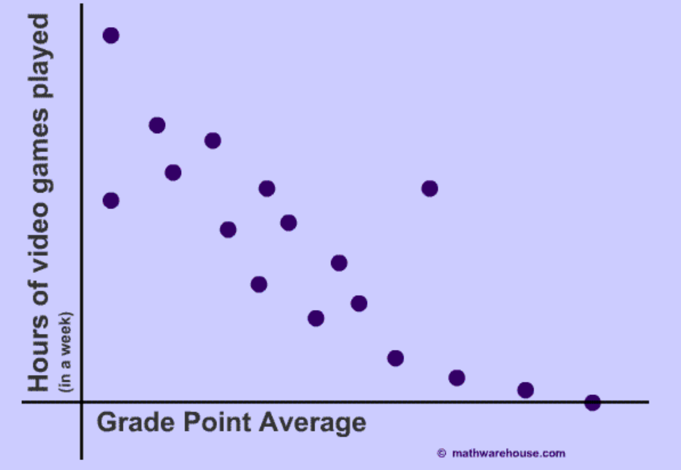
how-to-calculate-correlation-coefficient.php



Correlation examples — negative

Number of hours of video game play vs. academic result

https://www.mathwarehouse.com/statistics/correlation-coefficient/how-to-calculate-correlation-coefficient.php



Pearson correlation coefficient

Pearson correlation coefficient between two data series

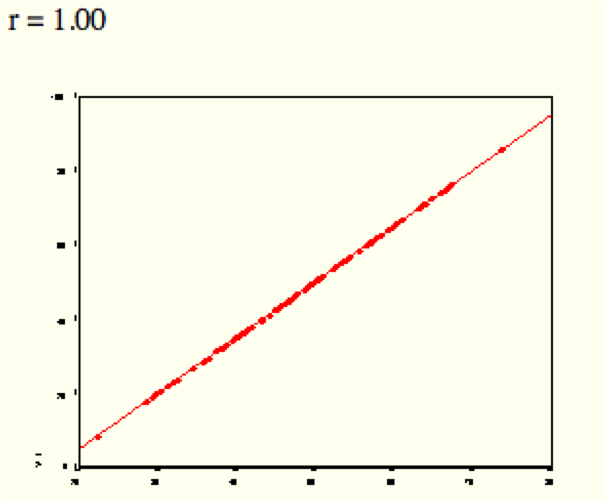
Let $xs = [x_0, x_1, \ldots, x_{n-1}]$
Let $ys = [y_0, y_1, \ldots, y_{n-1}]$

correlation(*xs*, *ys*) = degree of linear relationship between *xs* and *ys*

$$\text{correlation}(xs, ys) = \frac{\displaystyle\sum_{i=0}^{n-1} \frac{(x_i - m_x)}{sd_x} \frac{(y_i - m_y)}{sd_y}}{n - 1}$$

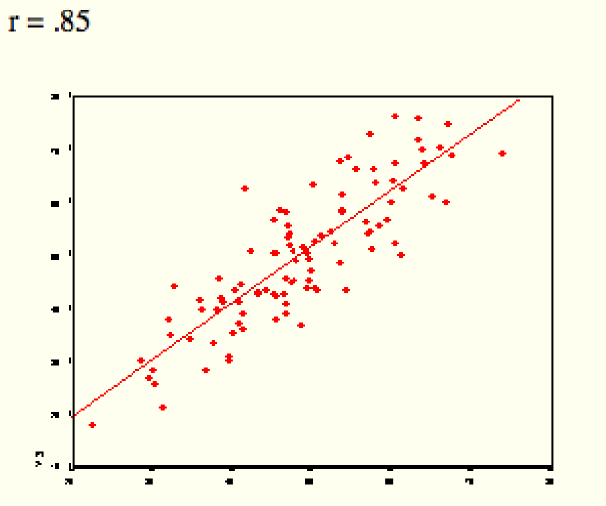
The correlation coefficient varies from -1.0 to $+1.0$

Source: `http://faculty.cbu.ca/~erudiuk/IntroBook/sbk17.htm`



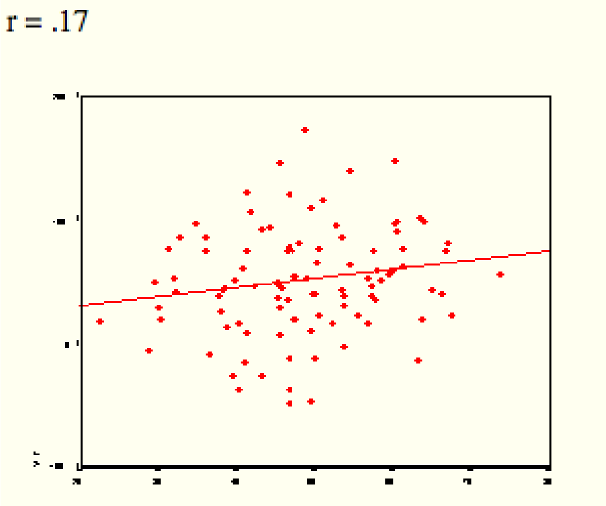
The correlation coefficient varies from -1.0 to $+1.0$

Source: `http://faculty.cbu.ca/~erudiuk/IntroBook/sbk17.htm`

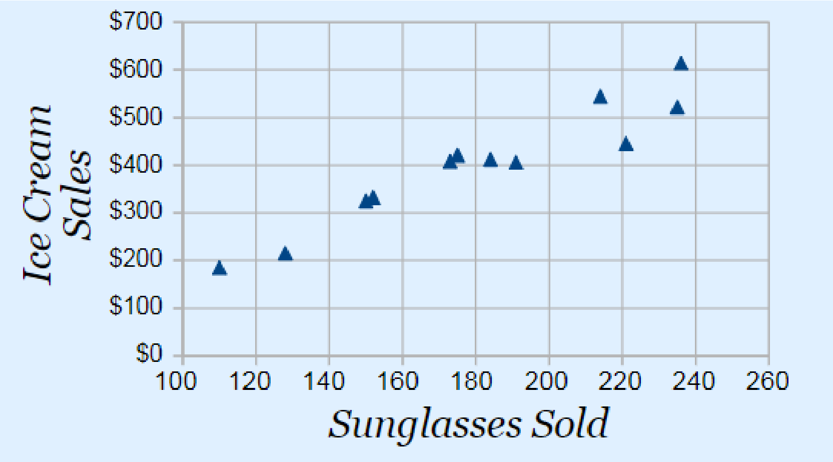


The correlation coefficient varies from -1.0 to $+1.0$

Source: `http://faculty.cbu.ca/~erudiuk/IntroBook/sbk17.htm`

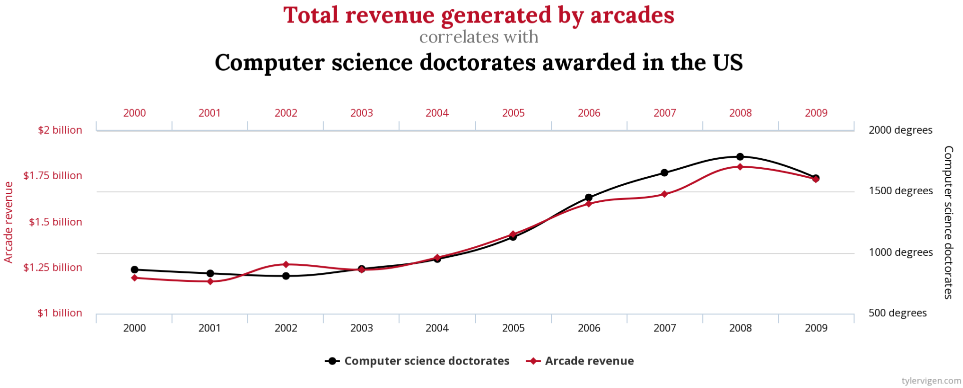


Correlation does not mean causality !



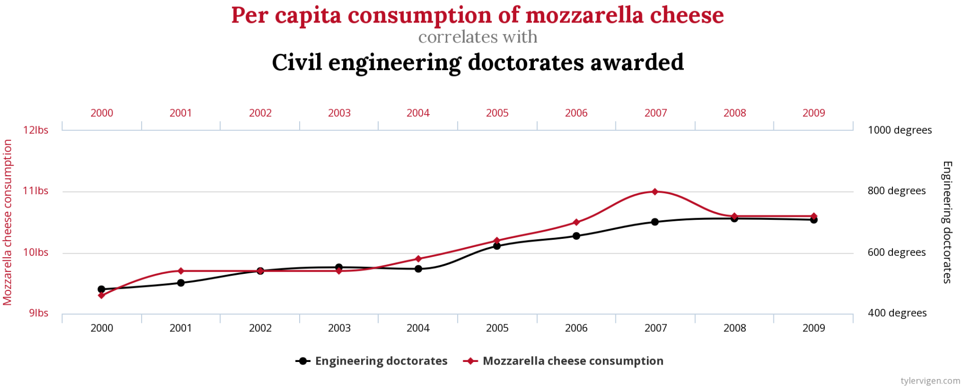
By looking long enough, one can find numerous correlations !

http://www.tylervigen.com/spurious-correlations



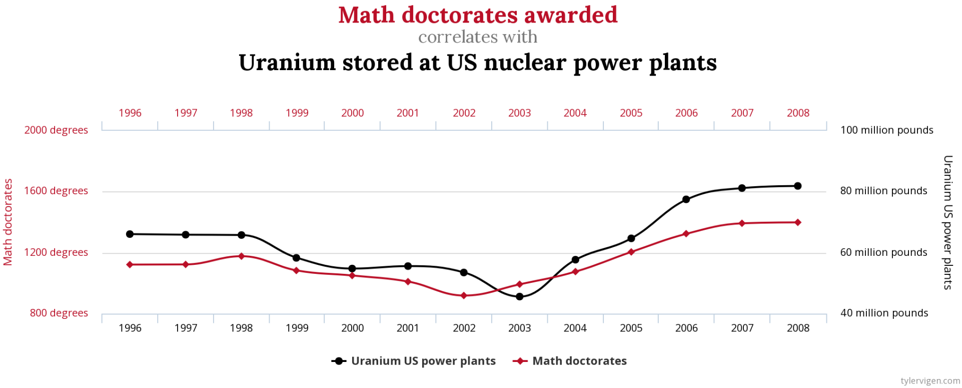
By looking long enough, one can find numerous correlations !

http://www.tylervigen.com/spurious-correlations



By looking long enough, one can find numerous correlations !

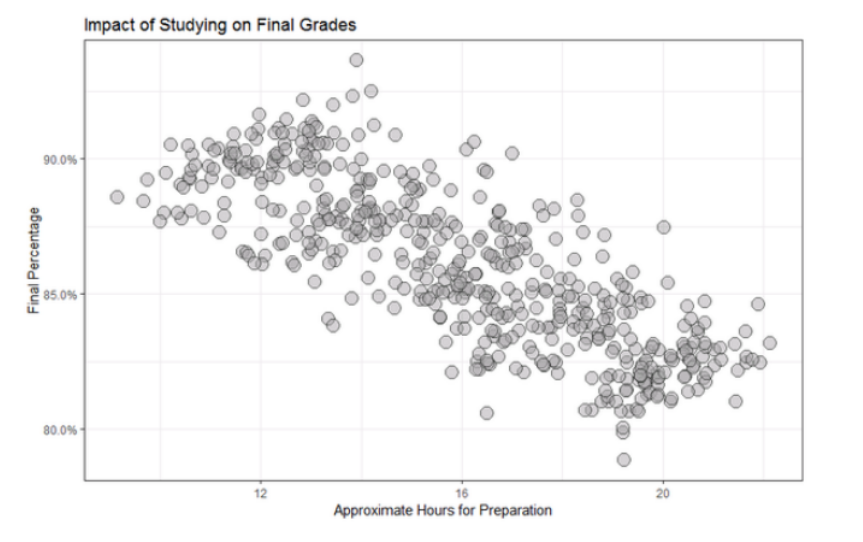
http://www.tylervigen.com/spurious-correlations



Correlation and Simpson's paradox

⋆

Source: `https://www.quora.com/What-is-Simpsons-paradox`

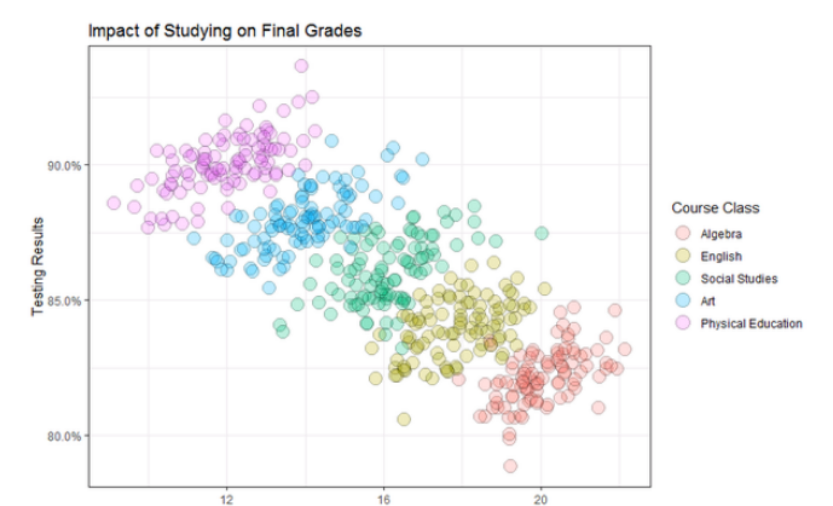


Correlation and Simpson's paradox

Negative correlation for the whole dataset, but positive for various subsets

Source: `https://www.quora.com/What-is-Simpsons-paradox`

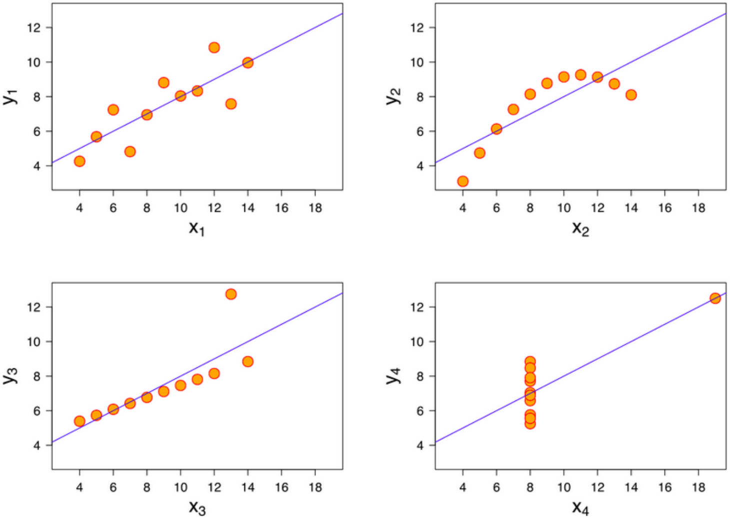
Source: `https://www.quora.com/What-is-Simpsons-paradox`



Data distribution

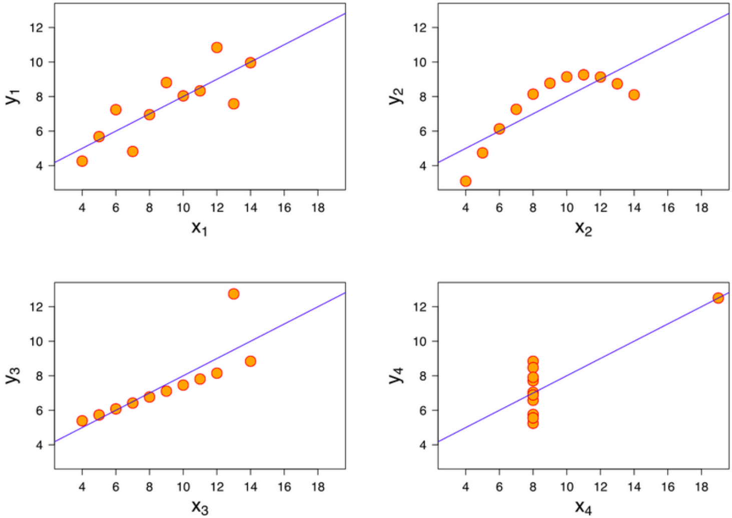
The measures are useful. . . but often misleading

What do these 4 dataset have in common (*Anscombe Quartet*, 1973) ?



The measures are useful. . . but often misleading

What do these 4 dataset have in common (*Anscombe Quartet*, 1973) ?



Same mean, standard deviation, and correlation coefficient (+0.816)

The measures are useful. . . but often misleading ⋆

Twelve datasets with same mean, standard deviation, and correlation coefficient (+0.32)

«*Stat Stats, Different Graphs : Generating Datasets with Varied Appearances and Identical Statistics through Simulated Annealing*», Metjka et Fitzmaurice, 2017

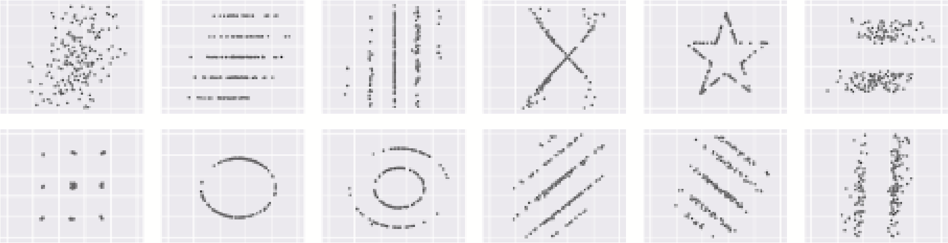


Figure 1. A collection of data sets produced by our technique. While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places. (*$\bar{x}$ =54.02, $\bar{y}$ = 48.09, sd_x = 14.52, sd_y = 24.79, Pearson's r = +0.32*)

The measures are useful. . . but often misleading ⋆

Twelve datasets with same mean, standard deviation, and correlation coefficient (+0.32)

«*Stat Stats, Different Graphs : Generating Datasets with Varied Appearances and Identical Statistics through Simulated Annealing*», Metjka et Fitzmaurice, 2017

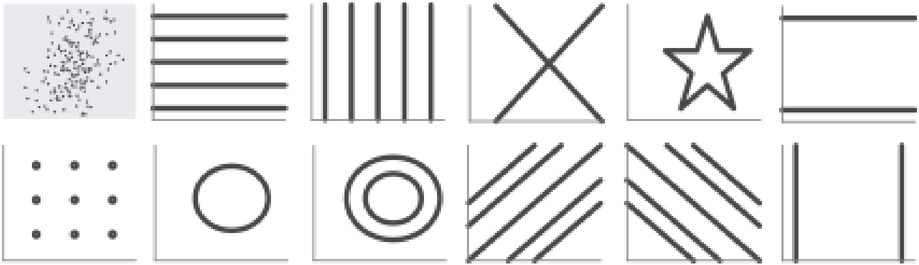
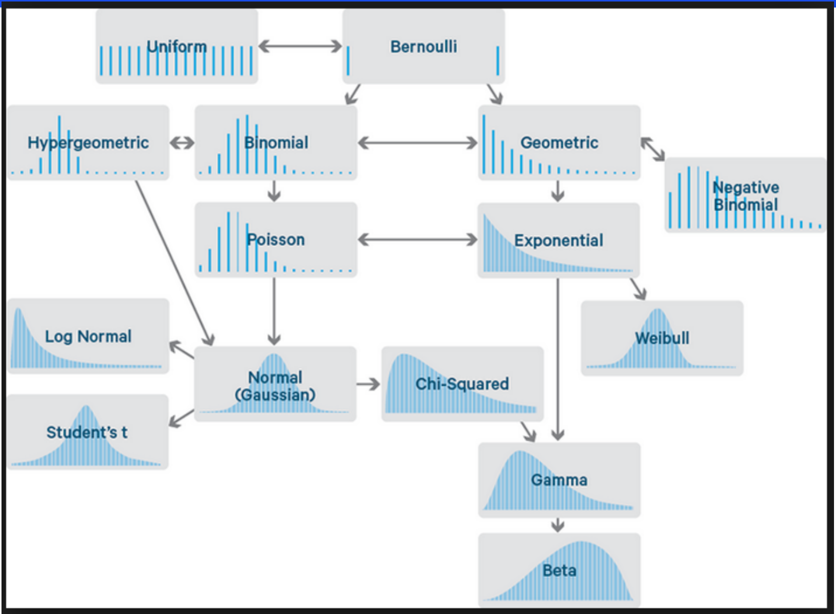
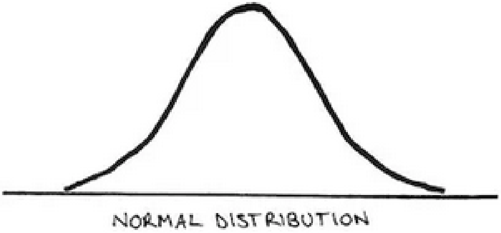


Figure 3. The initial data set (top-left), and line segment collections used for directing the output towards specific shapes. The results are seen in Figure 1.

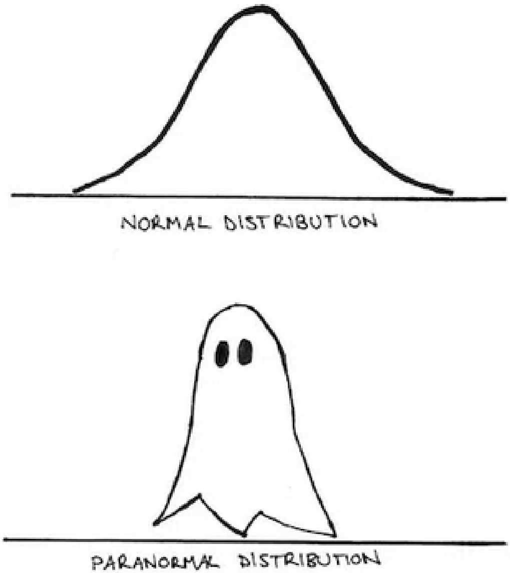
There are many different data distribution



An often seen distribution = Normal (Gaussian) distribution

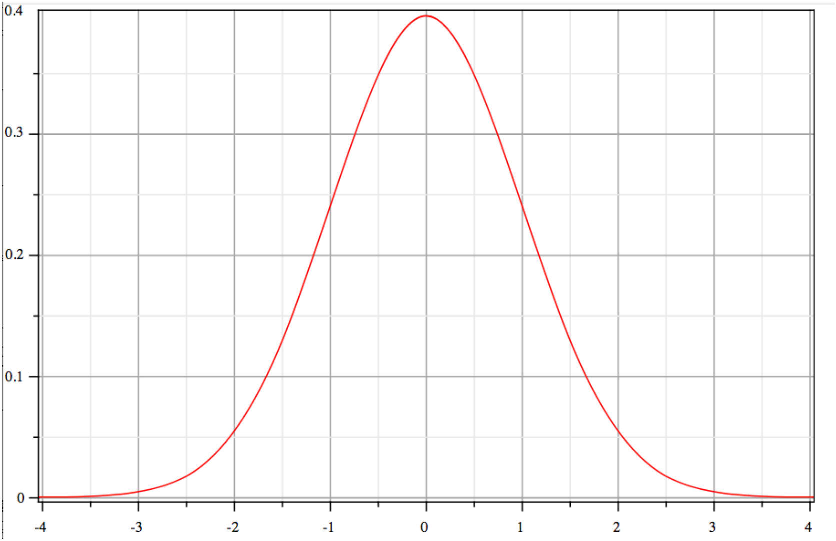


An often seen distribution = Normal (Gaussian) distribution

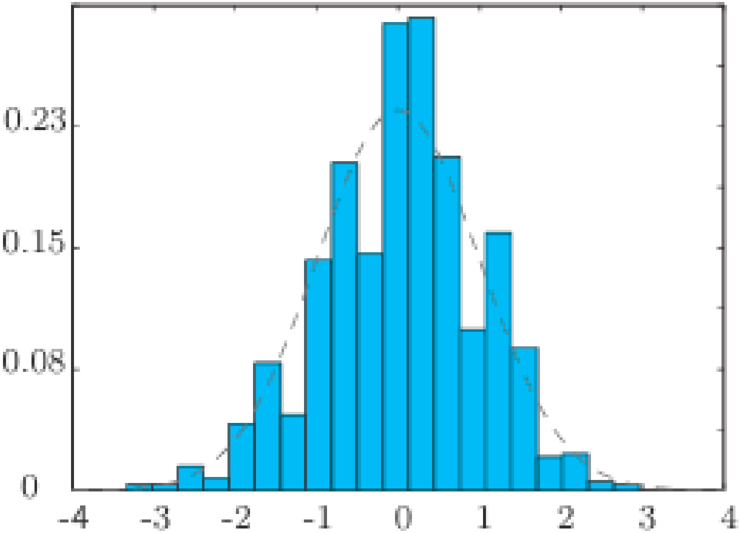


Normal distribution (continuous) : $\mathcal{N}(0, 1)$

https://upload.wikimedia.org/wikipedia

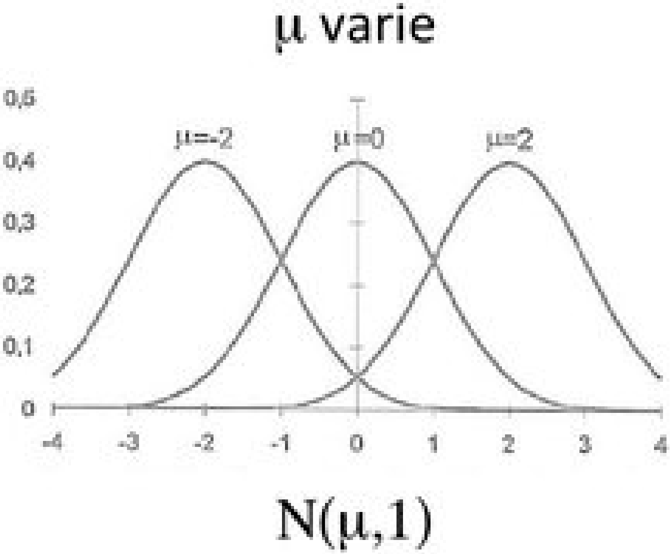


Normal distribution (discrete)



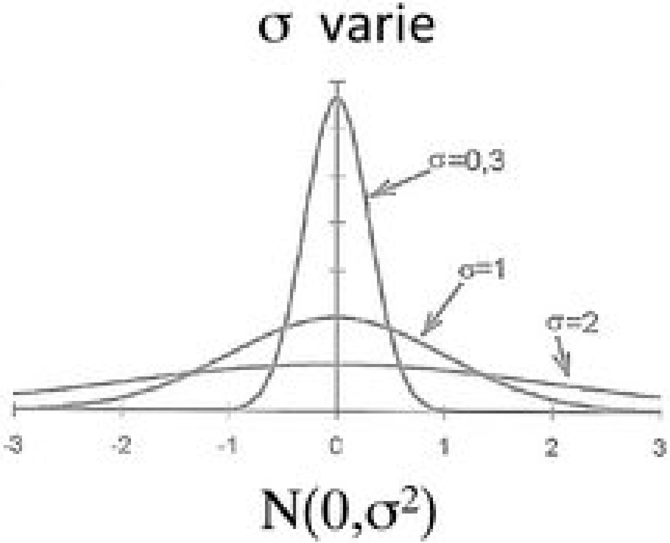
Normal distribution : Varying μ

https://upload.wikimedia.org/wikipedia



Normal distribution : Varying σ

https://upload.wikimedia.org/wikipedia



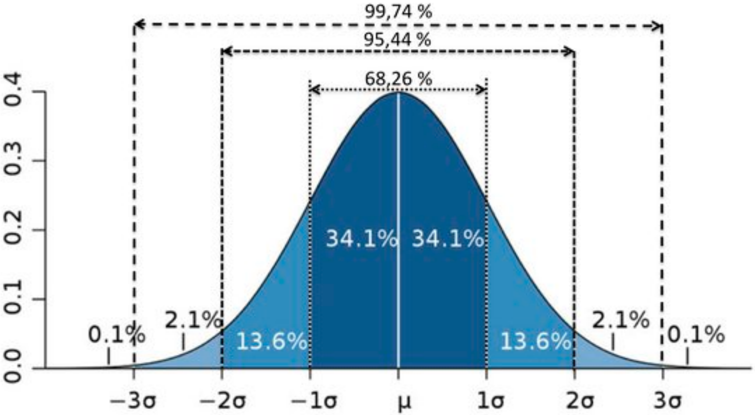
Normal distribution : $\mathcal{N}(\mu, \sigma^2)$

http://www.ilovestatistics.be/probabilite/loi-normale.html

What information does σ provide ?

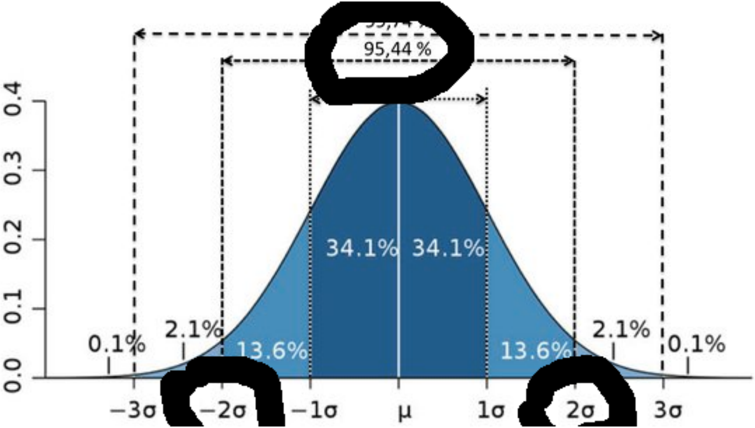
Normal distribution : $\mathcal{N}(\mu, \sigma^2)$

http://www.ilovestatistics.be/probabilite/loi-normale.html



Normal distribution : $\mathcal{N}(\mu, \sigma^2)$

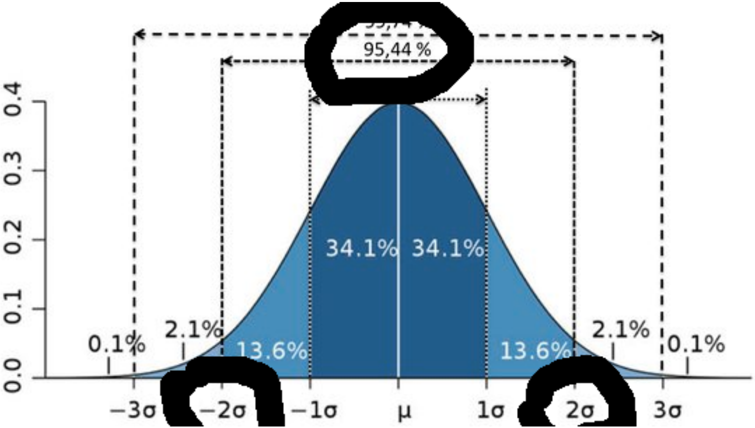
http://www.ilovestatistics.be/probabilite/loi-normale.html



$$P(X \in [\mu - 2\sigma, \mu + 2\sigma]) = 95.44\%$$

Normal distribution : $\mathcal{N}(\mu, \sigma^2)$

http://www.ilovestatistics.be/probabilite/loi-normale.html



$$P(X \in [\mu - 1.96\sigma, \mu + 1.96\sigma]) = 95.00\%$$

$$P(X \notin [\mu - 1.96\sigma, \mu + 1.96\sigma]) = 5.00\%$$

Distribution of the sample mean = Normal distribution

Also known as the “Central Limit Theorem”

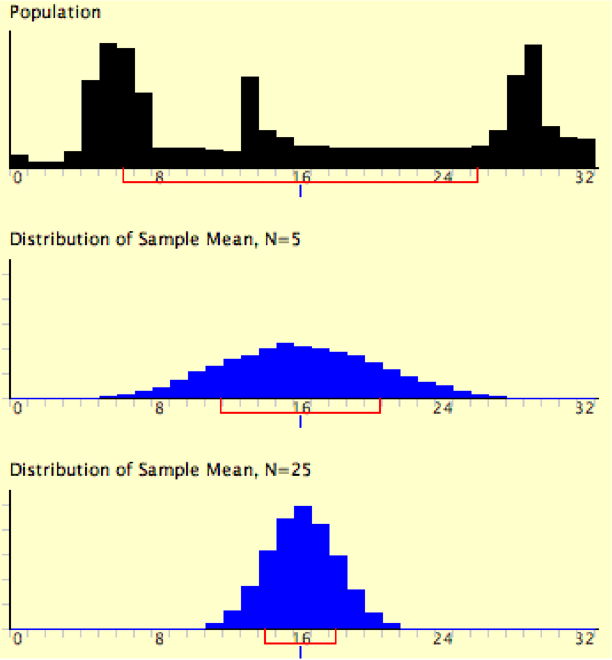
Key statistical property of sampling

Let *P* be a population with mean μ and variance σ^2.

If we take samples of size *N* from *P* and compute their means, then these various means follow a normal distribution

$$\mathcal{N}(\mu, \frac{\sigma^2}{N})$$

Note : *P* does not have to follow a normal distribution. *N* simply has to be large enough = «Law of large numbers».



Source : http://onlinestatbook.com/2/sampling_distributions/samp_dist_mean.html

Outline

1 Why this seminar ?

2 Is science in crisis ?

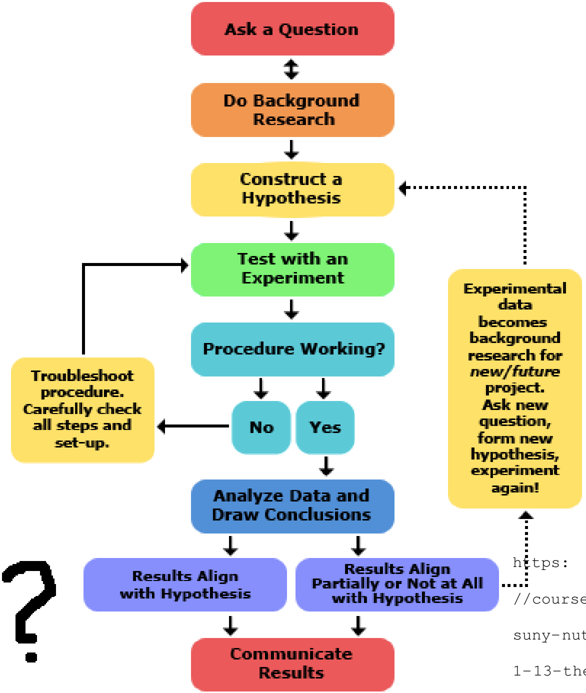
3 Some basic statistical concepts

4 Scientific method and statistical inference

5 Some causes of the crisis
- Focus on «positive» and «novel» results (aka. «Publication bias»)
- Flexibility in choosing experiment protocols and analyses
- Other aspects

6 Conclusion : Some possible solutions ?

The scientific method



Ask a Question

Do Background Research

Construct a Hypothesis

Test with an Experiment

Procedure Working?

No

Yes

Troubleshoot procedure. Carefully check all steps and set-up.

Analyze Data and Draw Conclusions

Results Align with Hypothesis

Results Align Partially or Not at All with Hypothesis

Experimental data becomes background research for *new/future* project. Ask new question, form new hypothesis, experiment again!

Communicate Results

https://courses.lumenlearning.com/suny-nutrition/chapter/1-13-the-scientific-method/

Why are statistics often used ?

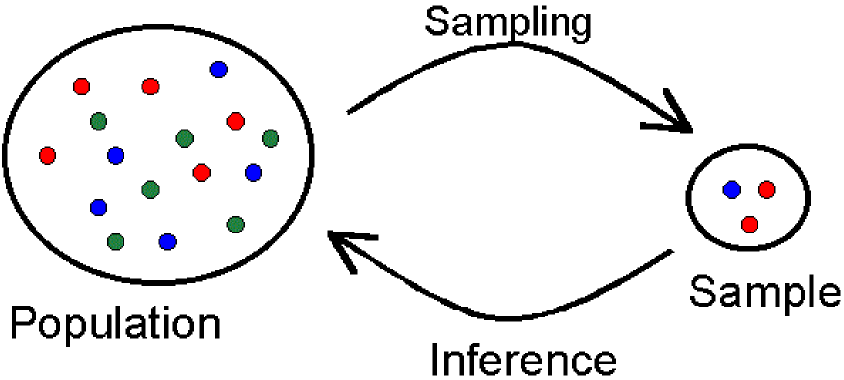
“Data don’t make any sense,
we will have to resort to statistics.”

Why are statistics often used ?

- Irregular, random phenomena, . . .
- Imprecise experimental measures
- Reasoning with samples
- Etc.

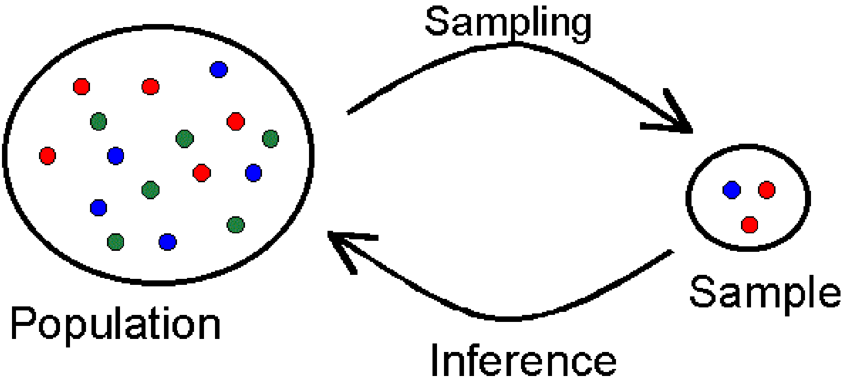
Why are statistics often used？

http://palin.co.in/difference-between-population-and-sampling-with-example



Why are statistics often used ?

http://palin.co.in/difference-between-population-and-sampling-with-example



Goal of statistical inference

Allow to state, with reasonable «confidence», that a phenomena (effect) **is not entirely due to randomness**

An (imaginary) example related with the teaching of software engineering

Context description

Course INF3456 uses programming language *L*

- Undergraduate course offered for the last 9 semesters
- $\approx$ 30–40 students per semester
- Programming language used = *L*
- No IDE available for *L* but…

Context description

Course INF3456 uses programming language *L*

- Undergraduate course offered for the last 9 semesters
- $\approx$ 30–40 students per semester
- Programming language used = *L*
- No IDE available for *L* but…

New IDE for *L*

- Prof. *P* designed and implemented a new IDE for *L*
- Prof. *P* would like to know if using this IDE helps students learn *L*

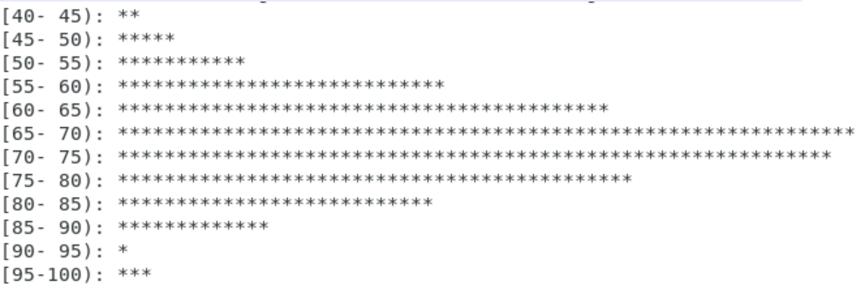
Experiment description

Known data ≈ Population

Known data

- Results from the previous 9 semesters (300 students) :

⇒ average = 69.8 % (std. dev. = 9.7)



```
[40- 45): **
[45- 50): *****
[50- 55): ***********
[55- 60): ****************************
[60- 65): ******************************************
[65- 70): ***************************************************************
[70- 75): *************************************************************
[75- 80): ********************************************
[80- 85): ***************************
[85- 90): *************
[90- 95): *
[95-100): ***
```

Experiment description

Winter 2019 results = Sample

Results obtained when new IDE was used (winter 2019)

- Number of students = 30
- average = 73.2 % (std. dev. = 14.1)

```
[35- 40): *
[40- 45):
[45- 50): *
[50- 55):
[55- 60): **
[60- 65): **
[65- 70): ******
[70- 75): *******
[75- 80): **
[80- 85): ****
[85- 90): *
[90- 95): **
[95-100): **
```

What can we conclude regarding the use of the IDE ?

Results without IDE
(300 students)

- Average = 69.8 %
- Std. dev. = 9.7

Results with IDE
(30 students)

- Average = 73.2 %
- Std. dev. = 14.1

What can we conclude regarding the use of the IDE ?

Results without IDE
(300 students)

- Average = 69.8 %
- Std. dev. = 9.7

Results with IDE
(30 students)

- Average = 73.2 %
- Std. dev. = 14.1

1. Helps students ?
   (average is larger ≈ +5%)

What can we conclude regarding the use of the IDE ?

Results without IDE
(300 students)

- Average = 69.8 %
- Std. dev. = 9.7

Results with IDE
(30 students)

- Average = 73.2 %
- Std. dev. = 14.1

1. Helps students ?
   (average is larger ≈ +5%)
2. Helps some students, but hinders others ?
   (std. dev. is larger ≈ +45%)

What can we conclude regarding the use of the IDE ?

Results without IDE
(300 students)

- Average = 69.8 %
- Std. dev. = 9.7

Results with IDE
(30 students)

- Average = 73.2 %
- Std. dev. = 14.1

1. Helps students ?
   (average is larger ≈ +5%)
2. Helps some students, but hinders others ?
   (std. dev. is larger ≈ +45%)
3. No effect ?
   (differences are purely «random» (sampling effect))

NHST approach to statistical inference (on mean)

Null Hypothesis Significance Testing

We state the hypothesis that we would like to verify

- H : Using the IDE increases the average

NHST approach to statistical inference (on mean)

Null Hypothesis Significance Testing

We state the hypothesis that we would like to verify

- H : Using the IDE increases the average

We state a null hypothesis (no effect = **it's only randomness !**)

- H_0 : Using the IDE… has no effect on the average

NHST approach to statistical inference (on mean)

Reductio ad unlikely

We use «reasoning to absurdity» (*reductio ad absurdum*) but using statistics

- Suppose the null hypothesis (it’s only randomness) is true

NHST approach to statistical inference (on mean)

Reductio ad unlikely

We use «reasoning to absurdity» (*reductio ad absurdum*) but using statistics

- Suppose the null hypothesis (it’s only randomness) is true
- Is it “surprising” to obtain the observed results ?

NHST approach to statistical inference (on mean)

Reductio ad unlikely

We use «reasoning to absurdity» (*reductio ad absurdum*) but using statistics

- Suppose the null hypothesis (it's only randomness) is true
- Is it "surprising" to obtain the observed results ?

  - If the result **is not surprising**,
    then we do not reject the null hypothesis :

    Our action do not seem to have any impact ☹

    Randomness makes the result reasonable and expectable !

NHST approach to statistical inference (on mean)

Reductio ad unlikely

We use «reasoning to absurdity» (*reductio ad absurdum*) but using statistics

- Suppose the null hypothesis (it's only randomness) is true
- Is it "surprising" to obtain the observed results ?
  - If the result is not surprising,
    then we do not reject the null hypothesis :
    Our action do not seem to have any impact ☹
  - If the result is «very» **«surprising!»**,
    then **we reject** the null hypothesis :
    Our action seems to have some impact ☺

Distribution of the sample mean

Statistical property of sampling

Let *P* be a population with mean μ and variance σ^2.

If we take samples of size *N* from *P* and compute their means, then they follow a normal distribution

$$\mathcal{N}(\mu, \frac{\sigma^2}{N})$$

Note : *P* does not have to follow a normal distribution. *N* simply has to be large enough = «Law of large numbers».

NHST approach applied to our example (IDE for *L*)

Population characteristics with H_0

Assume a population with :

- Average = 69.78%
- Std. dev. = 9.72

Distribution of the sample mean for $N = 30$

If we take samples of size 30 from this population, then the means follow a normal distribution

$$\mathcal{N}(69.78, \frac{9.72^2}{30}) = \mathcal{N}(69.78, 1.77^2)$$

Is it surprising for a sample of size 30 to have a mean = 73.22 — given $\mu = 69.78$ and $\sigma = 9.72$?

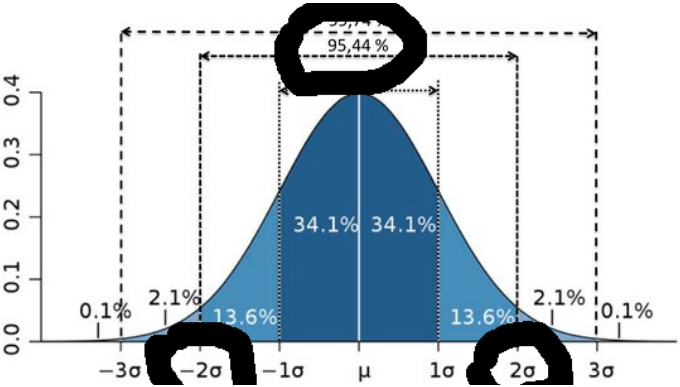
$X \sim \mathcal{N}(69.78, 1.77^2)$

Is it surprising for a sample of size 30 to have a mean = 73.22 — given $\mu = 69.78$ and $\sigma = 9.72$?

$X \sim \mathcal{N}(69.78, 1.77^2)$

$\Rightarrow$

$\Rightarrow$

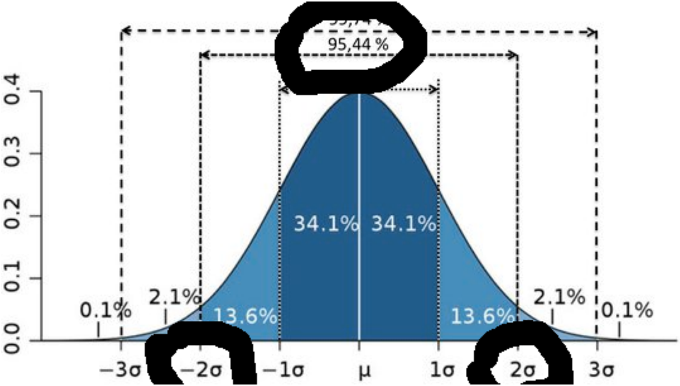


Is it surprising for a sample of size 30 to have a mean = 73.22 — given $\mu = 69.78$ and $\sigma = 9.72$?

$X \sim \mathcal{N}(69.78, 1.77^2)$

$\Rightarrow$ $P(X \in [69.78 - 2\sigma, 69.78 + 2\sigma]) = 95.44\%$

$\Rightarrow$ $P(X \in [66.24, 73.32]) = 95.44\%$



Is it surprising to obtain a sample whose mean differs by than 1.94σ or more from the population mean ?

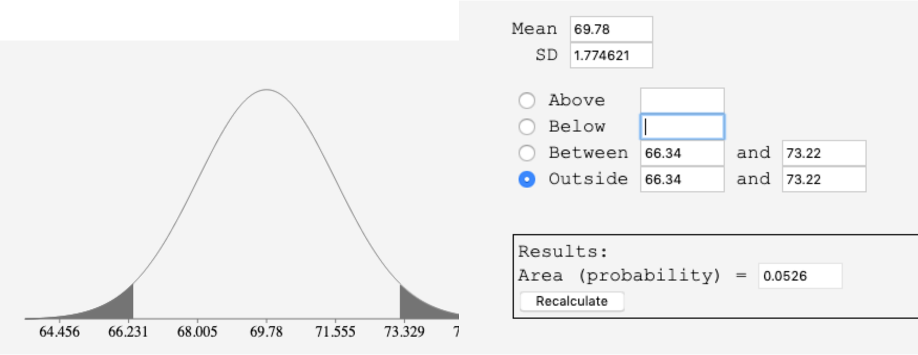
Is it surprising to obtain a sample whose mean differs by than 1.94σ or more from the population mean ?

$$X \sim \mathcal{N}(69.78, 1.77^2)$$

$$\Rightarrow P(X \in [69.78 - 1.94\sigma, 69.78 + 1.94\sigma]) = 94.74\%$$

$$\Rightarrow P(X \in [66.34, 73.22]) = 94.74\%$$

$$\Rightarrow \boxed{P(X \notin [66.34, 73.22]) = 5.26\%}$$



Mean 69.78

SD 1.774621

○ Above

○ Below

○ Between 66.34 and 73.22

● Outside 66.34 and 73.22

Results:

Area (probability) = 0.0526

Is it surprising to obtain a sample whose mean differs by than 1.94σ or more from the population mean ?

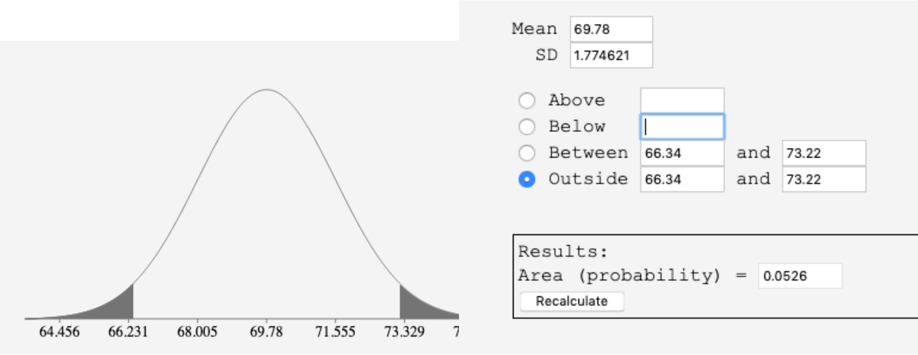
$1.94\ \sigma$ or more $\Rightarrow$ **p-value =** $0.0526 > 0.05$ ☹

$$X \sim \mathcal{N}(69.78, 1.77^2)$$

$$\Rightarrow P(X \in [69.78 - 1.94\sigma, 69.78 + 1.94\sigma]) = 94.74\%$$

$$\Rightarrow P(X \in [66.34, 73.22]) = 94.74\%$$

$$\Rightarrow \boxed{P(X \notin [66.34, 73.22]) = 5.26\%}$$ ⇒ **Not surprising !**



Mean 69.78

SD 1.774621

○ Above

○ Below

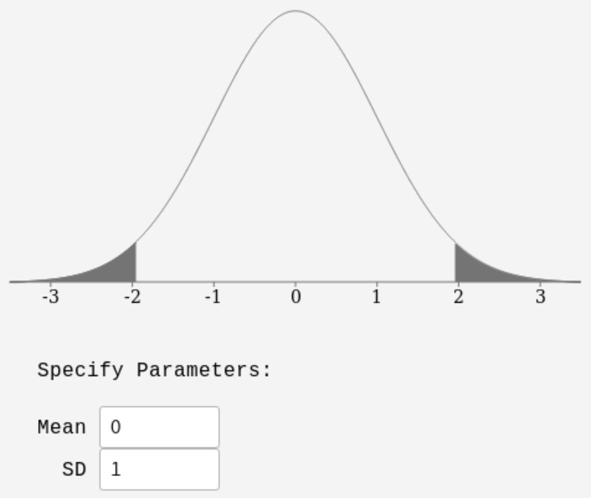
○ Between 66.34 and 73.22

● Outside 66.34 and 73.22

Results:

Area (probability) = 0.0526

When can we conclude that a result is indeed «surprising» ? Standard answer = $p < 0.05$!

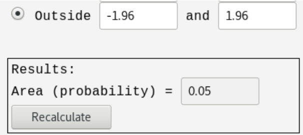


Specify Parameters:

Mean 0

SD 1

Case $\mathcal{N}(0, 1)$

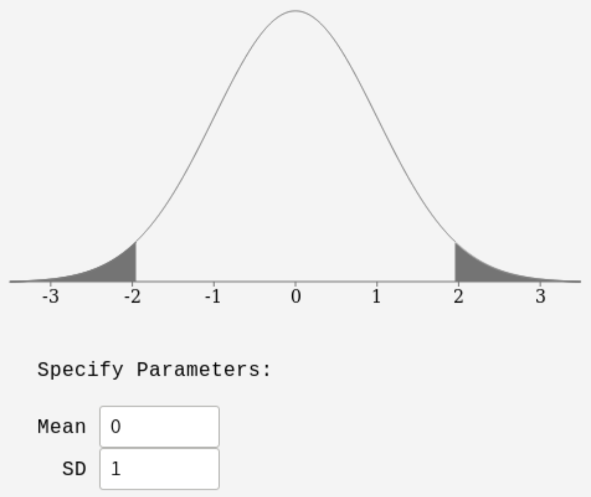


● Outside -1.96 and 1.96

Results:

Area (probability) = 0.05

When can we conclude that a result is indeed «surprising» ? Standard answer = $p < 0.05$!

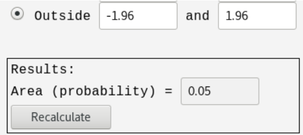


Specify Parameters:

Mean 0

SD 1

Case $\mathcal{N}(0,1)$


● Outside -1.96 and 1.96

Results:

Area (probability) = 0.05

For $X \sim \mathcal{N}(\mu, \sigma^2)$: If it's only randomness, then $X \in [\mu - 1.96\sigma, \mu + 1.96\sigma]$ 19 times out of 20

Résultat d’un sondage présenté sur le site Web de La Presse

Publié le 24 mai 2019 à 06h26 | Mis à jour à 06h26

Ontario : Doug Ford et son parti en chute libre

Les intentions de vote du Parti progressiste-conservateur de l’Ontario dégringolent et le taux d’insatisfaction envers le premier ministre Doug Ford n’a jamais été aussi élevé selon un sondage Recherche Mainstreet réalisé mardi et mercredi derniers.

[. . .]

Le sondage Mainstreet a été réalisé auprès de 996 personnes en Ontario. Sa marge d’erreur est de plus ou moins 3,1 %, **19 fois sur 20**.

Does «19 times out of 20» ring any bell ?

Does «19 times out of 20» ring any bell ?

Election survey results presented on the Gazette’s web site

Marian Scott, Montreal Gazette Updated : October 8, 2019

Election 2019 : New poll puts Conservatives ahead
A new poll taken after Monday’s federal leaders’ debate suggests that rising support for the Bloc Québécois in Quebec could put the Conservatives in power.

The telephone survey of 1,013 Canadians by Forum Research Inc. has the Tories leading with 35 per cent of voter intentions, while the Liberals are trailing with 28 per cent.
[. . .]
Results of the poll are considered to be accurate within three percentage points, **19 times out of 20.**

https://montrealgazette.com/news/local-news/poll-predicts-conservative-minority

Why do we use $p < 0.05$?

Suggestion by R.A. Fisher (1890–1962)

- A suggestion… which has become a convention — almost a «dogma!» — in many domains :
  - Biomedical sciences
  - Psychology
  - Social sciences
  - Surveys

Why do we use $p < 0.05$?

Suggestion by R.A. Fisher (1890–1962)

- A suggestion… which has become a convention — almost a «dogma!» — in many domains :
  - Biomedical sciences
  - Psychology
  - Social sciences
  - Surveys

«*Statistical errors*», R. Nuzzo, Nature, 2014

The irony is that when UK statistician Ronald Fisher introduced the P-value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the old-fashioned sense : ***worthy of a second look****.*

Some domains use values much smaller than 0.05 !

High-energy particle physics

High-energy physics requires even lower p-values to announce evidence or discoveries. The threshold for "evidence of a particle," corresponds to p=0.003, and the standard for "discovery" is p=0.0000003.

In our experiment with the IDE for *L*, let's see what happens when we change a single data. . .

We decide to review the marking. . . and change a single mark :

$3\underline{3}.9 \rightarrow 3\underline{5}.9$

In our experiment with the IDE for *L*, let's see what happens when we change a single data…

We decide to review the marking… and change a single mark :

$3\underline{3}.9 \rightarrow 3\underline{5}.9$

$\Rightarrow$ Sample mean : $73.22 \rightarrow 73.32 \Rightarrow 1.9948\ \sigma$ (from 69.78)

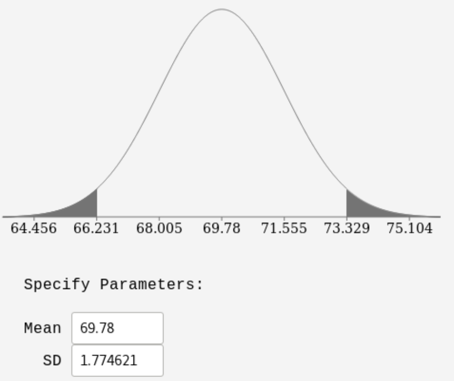
In our experiment with the IDE for *L*, let's see what happens when we change a single data. . .

We decide to review the marking. . . and change a single mark :

$3\underline{3}.9 \rightarrow 3\underline{5}.9$

$\Rightarrow$ Sample mean : $73.22 \rightarrow 73.32 \Rightarrow 1.9948\ \sigma$ (from 69.78)

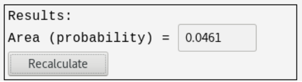
$\Rightarrow$ $\boxed{P(X \notin [66.24, 73.32]) = 4.61\%}$



Specify Parameters:

Mean 69.78

SD 1.774621



Results:

Area (probability) = 0.0461

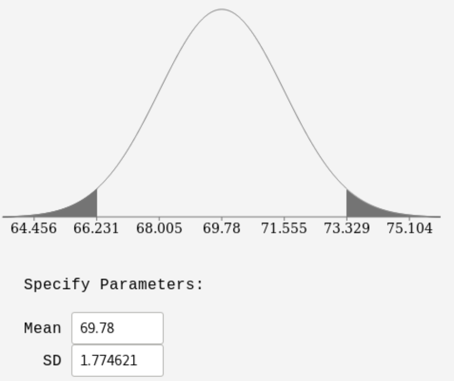
In our experiment with the IDE for *L*, let's see what happens when we change a single data. . .

We decide to review the marking. . . and change a single mark :

$3\underline{3}.9 \to 3\underline{5}.9$

$\Rightarrow$ Sample mean : $73.22 \to 73.32 \Rightarrow 1.9948\,\sigma$ (from 69.78)

$\Rightarrow$ $\boxed{P(X \notin [66.24, 73.32]) = 4.61\%}$ $\Rightarrow$ **Surprising !**

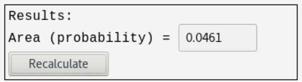


Specify Parameters:

Mean 69.78

SD 1.774621

Now $p < 0.05$, so we can claim that our result is «**statistically significant**»



Results:

Area (probability) = 0.0461

Outline

1. Why this seminar ?
2. Is science in crisis ?
3. Some basic statistical concepts
4. Scientific method and statistical inference
5. Some causes of the crisis
    - Focus on «positive» and «novel» results (aka. «Publication bias»)
    - Flexibility in choosing experiment protocols and analyses
    - Other aspects
6. Conclusion : Some possible solutions ?

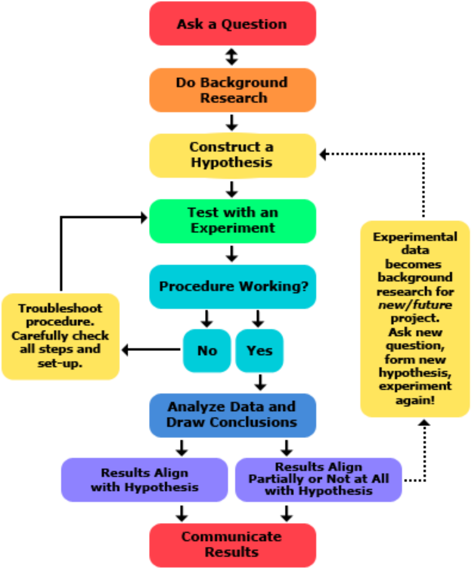
The crisis is not mainly due to «frauds»

Outright fraud is almost certainly just a small part of that problem, but high-profile examples have exposed a greyer area of bad or lazy scientific practice that many had preferred to brush under the carpet.

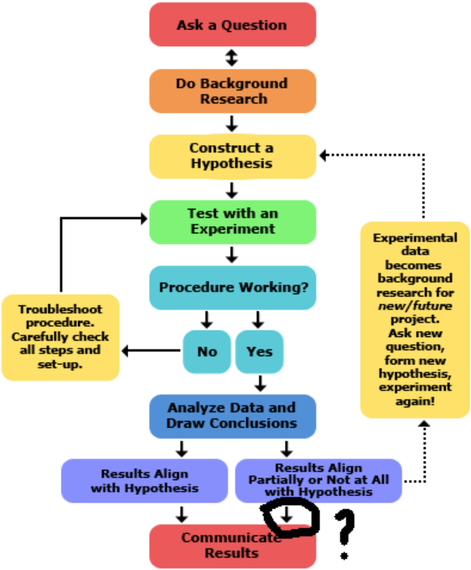
«False positives : Fraud and misconduct are threatening scientific research», A. Jha, The Guardian, 2012

5.1 Focus on «positive» and «novel» results (aka. «Publication bias»)

Can all results be published ?

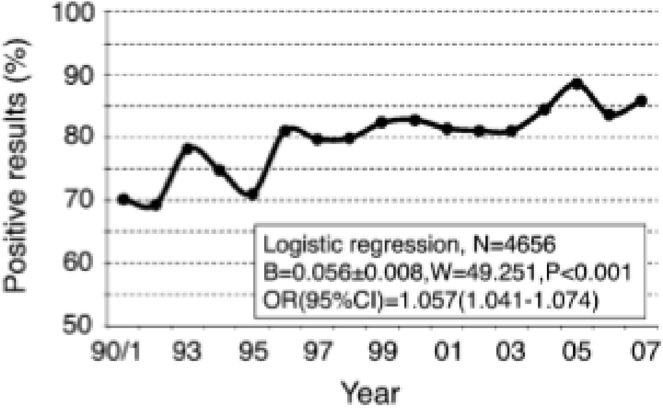

Ask a Question
Do Background Research
Construct a Hypothesis
Test with an Experiment
Procedure Working?
No
Yes
Troubleshoot procedure. Carefully check all steps and set-up.
Analyze Data and Draw Conclusions
Results Align with Hypothesis
Results Align Partially or Not at All with Hypothesis
Experimental data becomes background research for *new/future* project. Ask new question, form new hypothesis, experiment again!
Communicate Results

Can all results be published ?


Ask a Question
Do Background Research
Construct a Hypothesis
Test with an Experiment
Procedure Working?
No
Yes
Troubleshoot procedure. Carefully check all steps and set-up.
Analyze Data and Draw Conclusions
Results Align with Hypothesis
Results Align Partially or Not at All with Hypothesis
Experimental data becomes background research for *new/future* project. Ask new question, form new hypothesis, experiment again!
Communicate Results
?

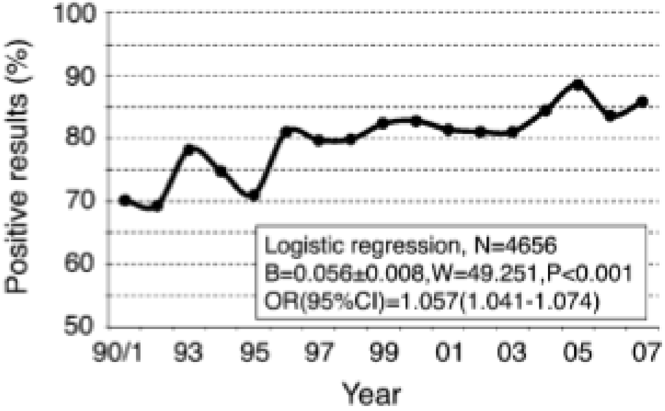
Percentage of published articles claiming positive results

- Fanelli (2010) : 2000 papers in various domains (bio, psycho, physique, chimie, etc.) — *space science* : 70%, . . . , psycho : 91%.



Percentage of published articles claiming positive results

- Fanelli (2010) : 2000 papers in various domains (bio, psycho, physique, chimie, etc.) — *space science* : 70%, . . . , psycho : 91%.
- Another study : molecular biology and clinical studies : 100%



Some people think scientists exclaim
Eureka!
when doing experiments.
But they're way more likely to say...
Bollocks!
oh...Sh*t!
F*ck!
Arse!
Stupid piece-of-crap machine!
I hate Science!
twisteddoodles.com

Scientific papers tell a story, not the real thing

Pour le béotien qui l’aborde, la littérature scientifique étonne en effet par son étonnante efficacité. Exceptionnels sont les articles qui décrivent un échec, une fausse piste, une impasse. Tout se passe comme si les chercheurs n’avaient toujours que de bonnes idées. Supposés interroger la nature, leurs expériences ont presque toujours le bon goût de confirmer l’hypothèse qui avait conduit à leur élaboration.

«Malscience — De la fraude dans les labos»,
N. Chevassus-au-Louis (2016)

Journals that only publish papers with negative results

«Le côté sombre de la science», S. Larivée, Revue de psychoéducation, 2017

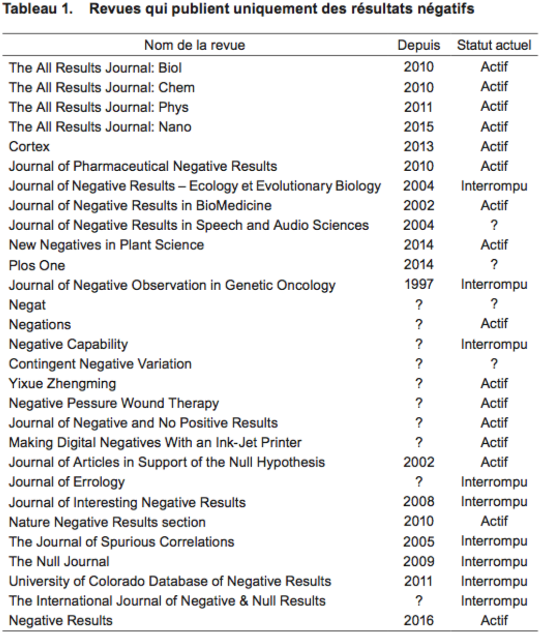
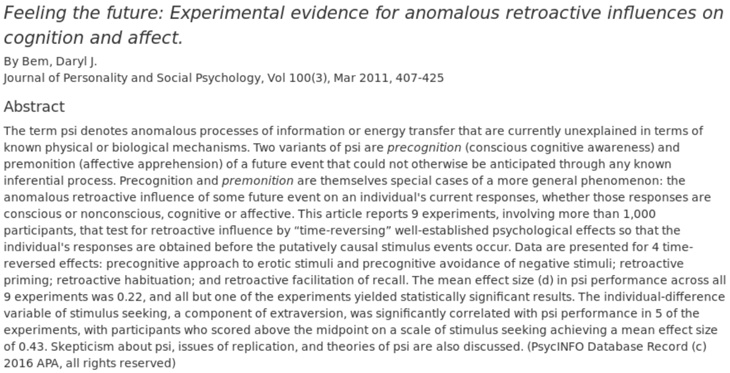


Tableau 1. Revues qui publient uniquement des résultats négatifs

Nom de la revue	Depuis	Statut actuel
The All Results Journal: Biol	2010	Actif
The All Results Journal: Chem	2010	Actif
The All Results Journal: Phys	2011	Actif
The All Results Journal: Nano	2015	Actif
Cortex	2013	Actif
Journal of Pharmaceutical Negative Results	2010	Actif
Journal of Negative Results – Ecology et Evolutionary Biology	2004	Interrompu
Journal of Negative Results in BioMedicine	2002	Actif
Journal of Negative Results in Speech and Audio Sciences	2004	?
New Negatives in Plant Science	2014	Actif
Plos One	2014	?
Journal of Negative Observation in Genetic Oncology	1997	Interrompu
Negat	?	?
Negations	?	Actif
Negative Capability	?	Interrompu
Contingent Negative Variation	?	?
Yixue Zhengming	?	Actif
Negative Pessure Wound Therapy	?	Actif
Journal of Negative and No Positive Results	?	Actif
Making Digital Negatives With an Ink-Jet Printer	?	Actif
Journal of Articles in Support of the Null Hypothesis	2002	Actif
Journal of Errology	?	Interrompu
Journal of Interesting Negative Results	2008	Interrompu
Nature Negative Results section	2010	Actif
The Journal of Spurious Correlations	2005	Interrompu
The Null Journal	2009	Interrompu
University of Colorado Database of Negative Results	2011	Interrompu
The International Journal of Negative & Null Results	?	Interrompu
Negative Results	2016	Actif

Very difficult to publish negative results : An «interesting» example



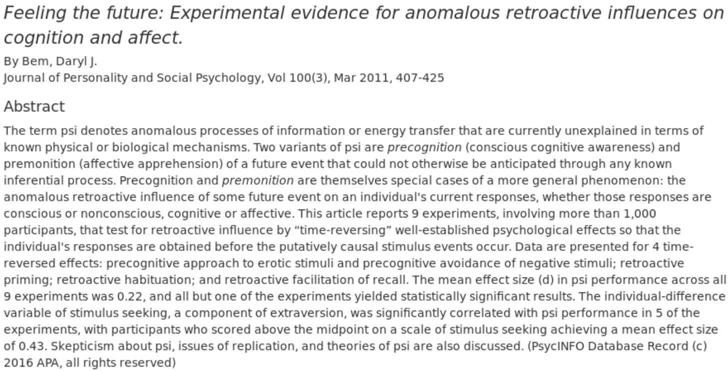
Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect.

By Bem, Daryl J.
Journal of Personality and Social Psychology, Vol 100(3), Mar 2011, 407-425

Abstract

The term psi denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and premonition (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and *premonition* are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by “time-reversing” well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (d) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed. (PsycINFO Database Record (c) 2016 APA, all rights reserved)

Very difficult to publish negative results : An «interesting» example



Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect.

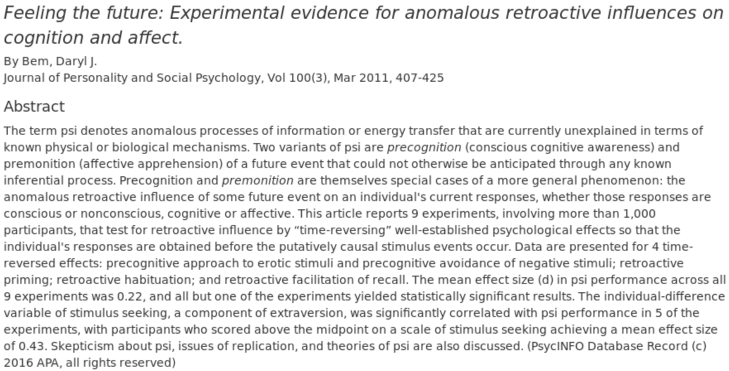
By Bem, Daryl J.
Journal of Personality and Social Psychology, Vol 100(3), Mar 2011, 407-425

Abstract

The term psi denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and premonition (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and *premonition* are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by “time-reversing” well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (d) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed. (PsycINFO Database Record (c) 2016 APA, all rights reserved)

- A team tried (3 times !) to reproduce Bem’s experiment & results. . . to no avail ☹

Very difficult to publish negative results :
An «interesting» example



Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect.

By Bem, Daryl J.
Journal of Personality and Social Psychology, Vol 100(3), Mar 2011, 407-425

Abstract

The term psi denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and premonition (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and *premonition* are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (d) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed. (PsycINFO Database Record (c) 2016 APA, all rights reserved)

- A team tried (3 times !) to reproduce Bem's experiment & results. . . to no avail ☹
- Answer from *Journal of Pers. and Soc. Psy.* : «*[we do] not publish replication studies, whether successful or unsuccessful*» !

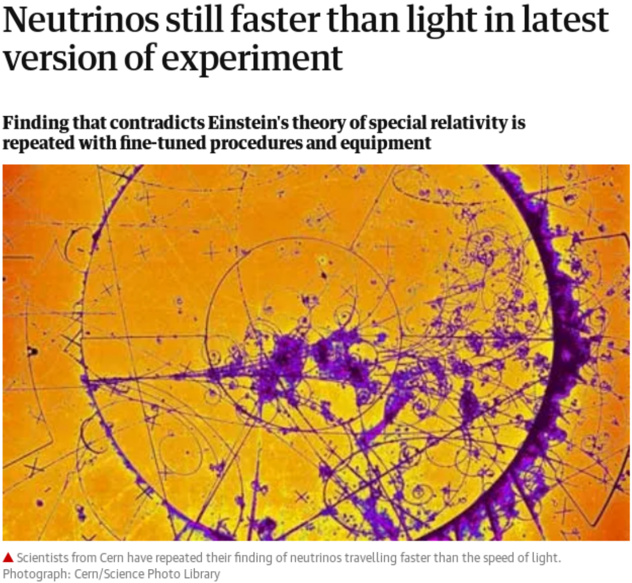
Replication is essential to «confirm» that a result is significant

But (non-)replication is also essential to «**refute**» a result

did not find
I ~~FOUND~~ THE EFFECT THAT MY COLLEAGUE FOUND YESTERDAY!

Can neutrinos travel faster than light?

2011



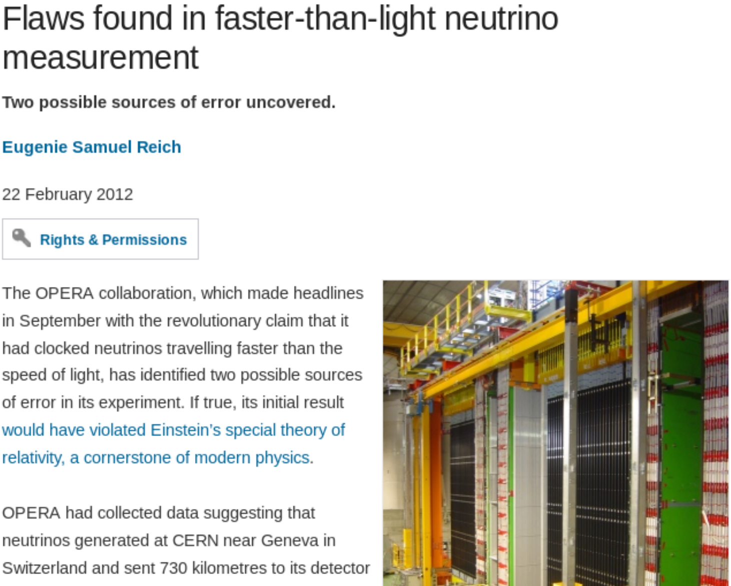
Neutrinos still faster than light in latest version of experiment

Finding that contradicts Einstein's theory of special relativity is repeated with fine-tuned procedures and equipment

▲ Scientists from Cern have repeated their finding of neutrinos travelling faster than the speed of light. Photograph: Cern/Science Photo Library

Can neutrinos travel faster than light ? No !

2012 : Error due to a loose fiber-optic cable !



Flaws found in faster-than-light neutrino measurement

Two possible sources of error uncovered.

Eugenie Samuel Reich

22 February 2012

Rights & Permissions

The OPERA collaboration, which made headlines in September with the revolutionary claim that it had clocked neutrinos travelling faster than the speed of light, has identified two possible sources of error in its experiment. If true, its initial result would have violated Einstein's special theory of relativity, a cornerstone of modern physics.

OPERA had collected data suggesting that neutrinos generated at CERN near Geneva in Switzerland and sent 730 kilometres to its detector

Focus on positive result positifs ⇒ Cobra Effect ? ⋆

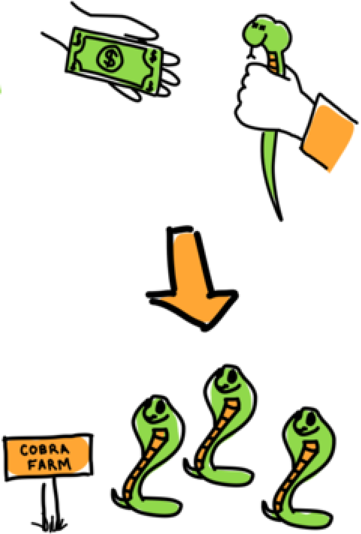
If researchers are rewarded for publications and positive results are generally both easier to publish and more prestigious than negative results, then researchers who can obtain more positive results—whatever their truth value—will have an advantage.

«The natural selection of bad science», *P.E. Smaldino & R. McElreath (2016)*

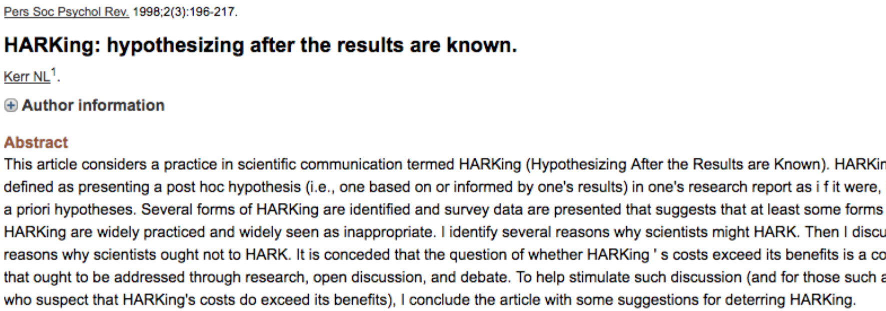
COBRA EFFECT

An attemped solution to a problem actually makes the problem worse.

Named for an anecdote to reduce dangerous cobras where villagers were paid a bounty for dead cobras, and people began 'farming' the snakes to collect more bounty.



Focus on positive results can lead to «dubious» practices



Pers Soc Psychol Rev. 1998;2(3):196-217.

HARKing: hypothesizing after the results are known.

Kerr NL[1].

⊕ **Author information**

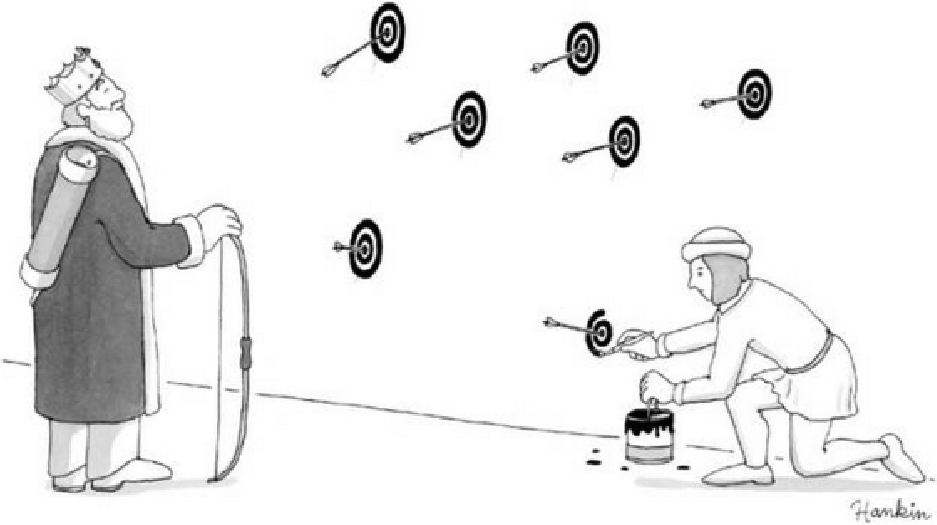
Abstract

This article considers a practice in scientific communication termed HARKing (Hypothesizing After the Results are Known). HARKin
defined as presenting a post hoc hypothesis (i.e., one based on or informed by one's results) in one's research report as i f it were,
a priori hypotheses. Several forms of HARKing are identified and survey data are presented that suggests that at least some forms
HARKing are widely practiced and widely seen as inappropriate. I identify several reasons why scientists might HARK. Then I discu
reasons why scientists ought not to HARK. It is conceded that the question of whether HARKing ' s costs exceed its benefits is a co
that ought to be addressed through research, open discussion, and debate. To help stimulate such discussion (and for those such a
who suspect that HARKing's costs do exceed its benefits), I conclude the article with some suggestions for deterring HARKing.

HARKing

«[P]resenting a post hoc hypothesis in the introduction of a research report as if it were an a priori hypothesis.»

Note : *Hark ! = Listen ! (Oxford Dictionary)*


Hankin

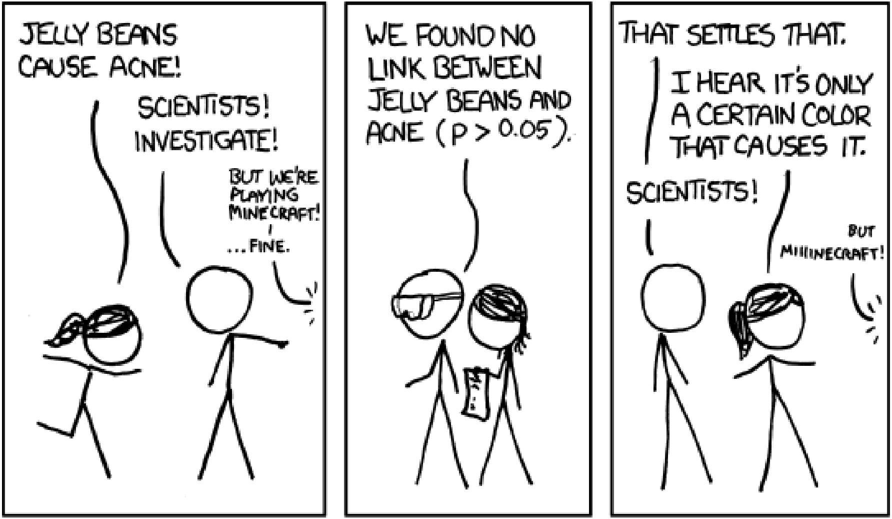
Self-Admission Rates of HARKing in Self-Report Surveys

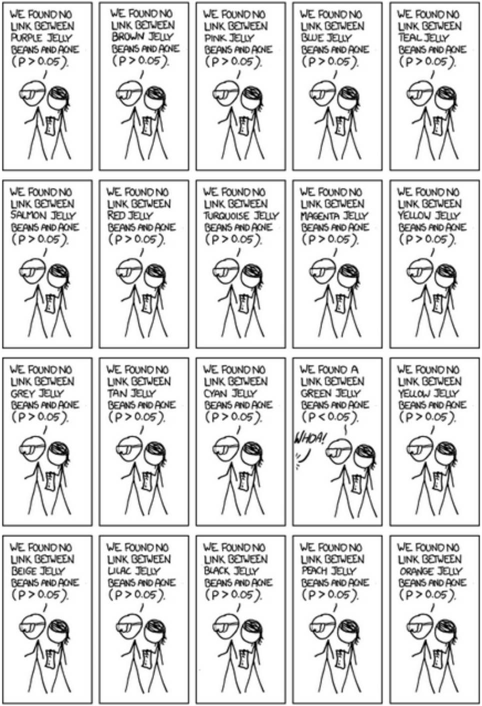
Survey	Population	Survey Item	*N*	Self-Admission Rate
John, Loewenstein, and Prelec (2012)	USA psychologists	"In a paper, reporting an unexpected finding as having been predicted from the start."	2,155	27.0%
Agnoli, Wicherts, Veldkamp, Albiero, and Cubelli (2017)	Italian psychologists	"In a paper, reporting an unexpected finding as having been predicted from the start."	277	37.4%
Bosco, Aguinis, Field, Pierce, and Dalton (2016, Study 1)	Researchers who published in *Personnel Psychology* and the *Journal of Applied Psychology* during 2005 to 2010	"whether any changes in hypotheses had occurred between the completion of data collection and subsequent publication."	53	38%
Fiedler and Schwarz (2016)	German psychologists	"Reporting an unexpected finding as having been predicted from the start."	1,138	47%
Banks et al. (2016, Studies 1 & 2)	Management researchers	"selectively reported hypotheses on the basis of statistical significance…and presented a post hoc hypothesis as if it were developed a priori."	749	50%
Motyl et al. (2017, Study 1)	Personality and social psychologists from Australian, European, and the USA	"Report that unexpected findings were expected."	1,166	58%
			Mean	43%

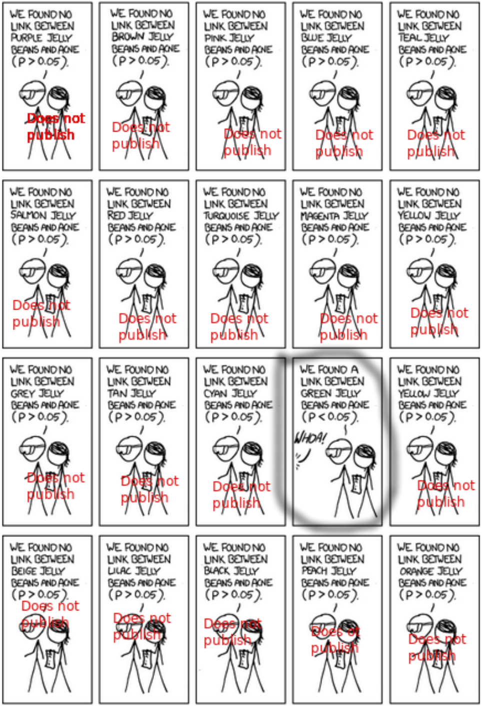
Note. Self-admission rates are for undertaking the stated behavior "at least once." Self-admission rates are likely to be underestimates because researchers tend to underreport practices that they perceive to be undesirable (Agnoli et al., 2017).

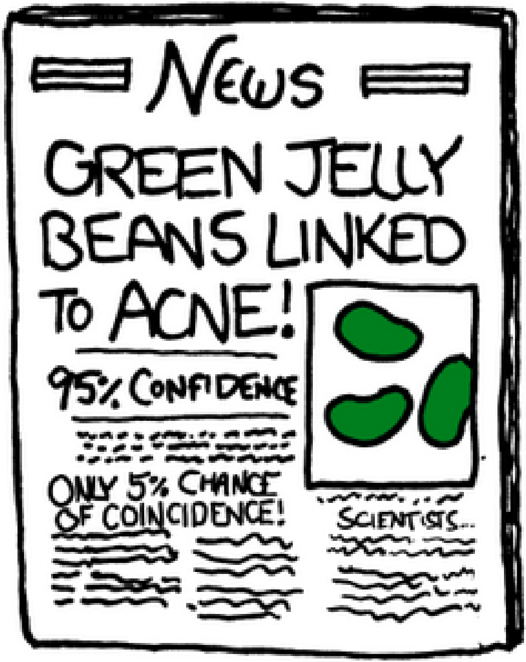
«For what is improbable does happen, and therefore it is probable that improbable things **will** happen.»

Aristotle

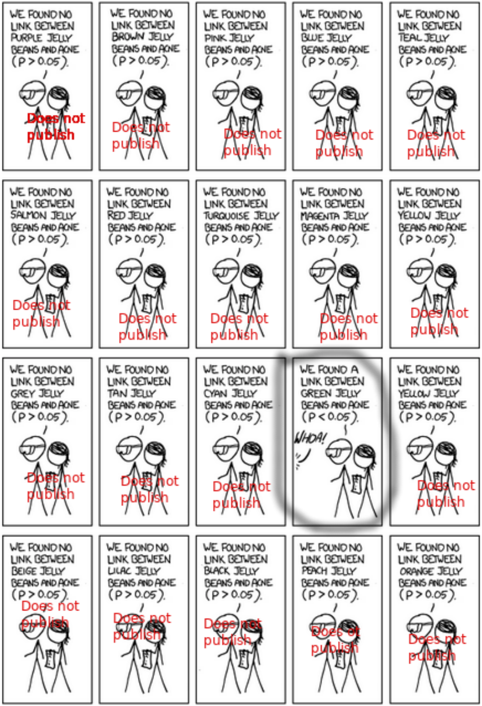

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


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 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).
WHOA!
WE FOUND A LINK BETWEEN GREEN 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 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).
Does not publish
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
Does not publish



NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
Does not publish
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
Does ot publish
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
Does not publish

The same can also happen if 20 different teams are researching the same topic, performing *similar* experiments !

SCIENCE

A Waste of 1,000 Research Papers

Decades of early research on the genetics of depression were built on nonexistent foundations. How did that happen?

ED YONG MAY 17, 2019

SEAN NEL / SHUTTERSTOCK

A Waste of 1,000 Research Papers

In 1996, a group of European researchers found that a certain gene, called SLC6A4, might influence a person's risk of depression.

It was a blockbuster discovery at the time. [. . .] Over two decades, this one gene inspired at least 450 research papers.

But a new study—the biggest and most comprehensive of its kind yet—shows that this seemingly sturdy mountain of research is actually a house of cards, built on nonexistent foundations.

[. . .]

Between them, these 18 genes have been the subject of more than 1,000 research papers, on depression alone. And for what ? If the new study is right, these genes have nothing to do with depression. "This should be a real cautionary tale," Keller adds. "How on Earth could we have spent 20 years and hundreds of millions of dollars studying pure noise ?"

https://www.theatlantic.com/science/archive/2019/05/waste-1000-studies/589684/

We must distinguish between exploratory vs. descriptive vs. causal research ⋆

Exploratory vs. descriptive vs. explanatory research

> *[HARKing] would be innocuous if the researcher acknowledged the exploratory nature of the study and sought to confirm the findings in another set of data (or if he or she used cross validation techniques). It becomes a problem when researchers pretend that they had the hypothesis a priori and that the study was done to confirm it, hiding the exploratory nature of the study and conferring more strength to the results than they actually have.*

https://academia.stackexchange.com/questions/60401/are-p-hacking-and-hypothesising-after-results-are-known-considered-misconduct-in

5.2 Flexibility in choosing experiment protocols and analyses

Researchers, when performing their experiments and analyses, have a wide range of choices and options

- Excluding some values/participants (*outliers*) . . .
  or not ?
- Terminating early the data collection. . .
  or not ?
- Using some statistical analysis statistique. . .
  or an other ?

IF YOU TORTURE THE DATA LONG ENOUGH, IT ***WILL*** CONFESS.

One well-known method of «torture» = *p-hacking*

One well-known method of «torture» = *p-hacking*

P-hacking

[p-hacking] occurs when researchers collect or select data or statistical analyses until **nonsignificant** *results become* **significant**.

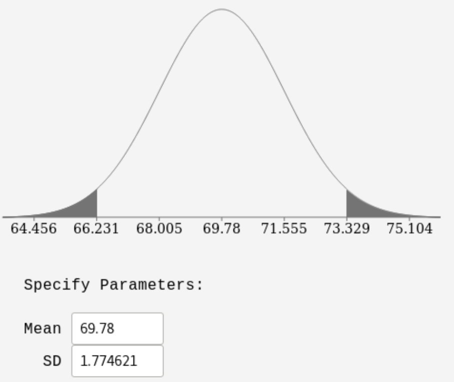
«The Extent and Consequences of P-Hacking in Science», *Head et al. (2015)*

Remember the experiment on the use of an IDE for *L*

Revised marking with a single (1) mark changed :

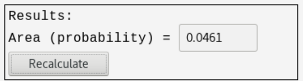
$33.9 \rightarrow 35.9 \Rightarrow$ Average : $73.2 \rightarrow 73.3$

- Before : $p = 0.0526 > 0.05$ ☹
- **After :** $p = 0.0461 < 0.05$ ☺



```
Specify Parameters:

Mean  69.78
  SD  1.774621
```

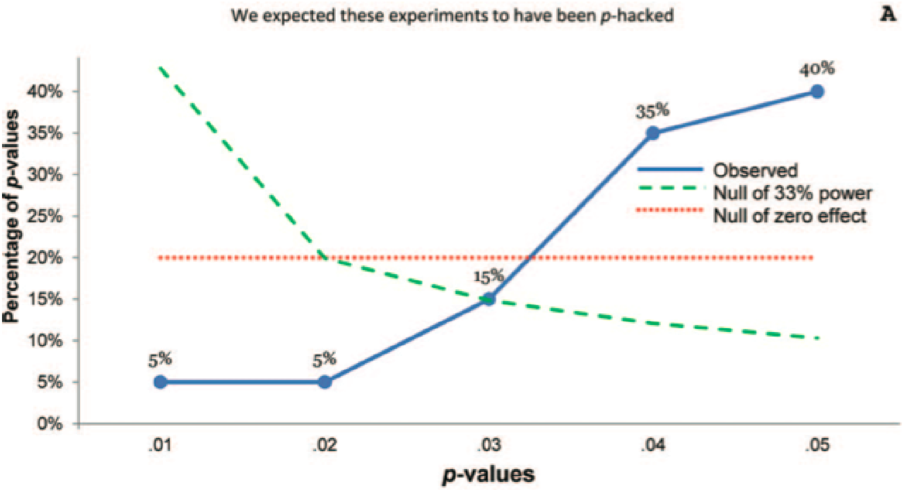


```
Results:
Area (probability) =  0.0461
```

Is this kind of tinkering common？

Is this kind of tinkering common ?

Yes !



Performing different analyses on the same data can lead to quite different results !

https://www.youtube.com/watch?v=vBzEGSm23y8

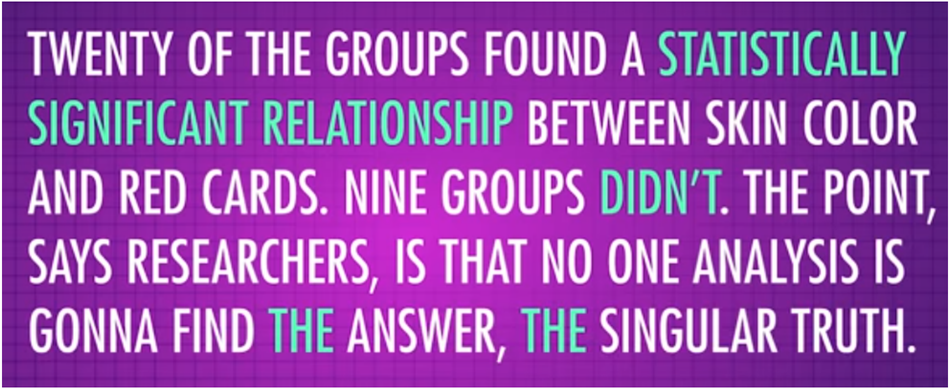
Question : Do referees give more penalties to players with dark skin than to those with light skin ?

The Replication Crisis: Crash Course Statistics #31

Performing different analyses on the same data can lead to quite different results !

https://www.youtube.com/watch?v=vBzEGSm23y8

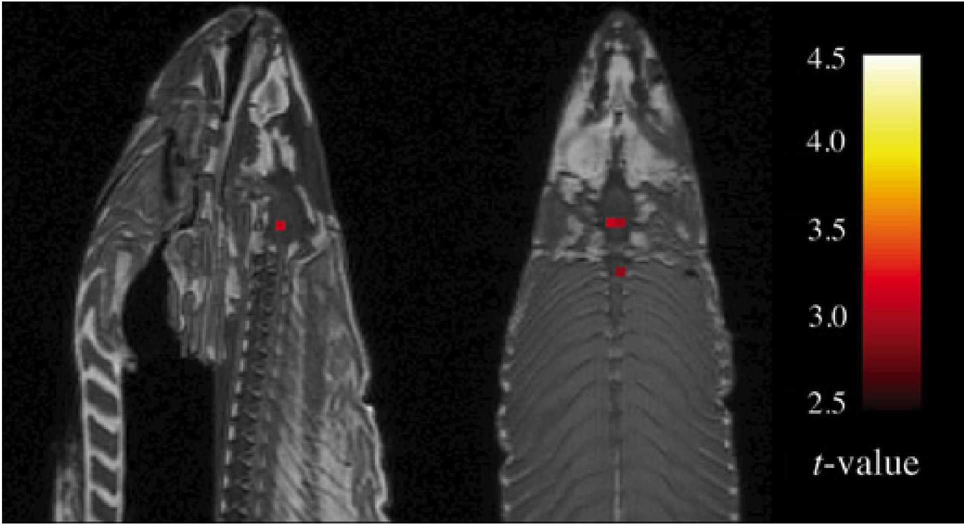
Question : Do referees give more penalties to players with dark skin than to those with light skin ?



TWENTY OF THE GROUPS FOUND A STATISTICALLY SIGNIFICANT RELATIONSHIP BETWEEN SKIN COLOR AND RED CARDS. NINE GROUPS DIDN'T. THE POINT, SAYS RESEARCHERS, IS THAT NO ONE ANALYSIS IS GONNA FIND THE ANSWER, THE SINGULAR TRUTH.

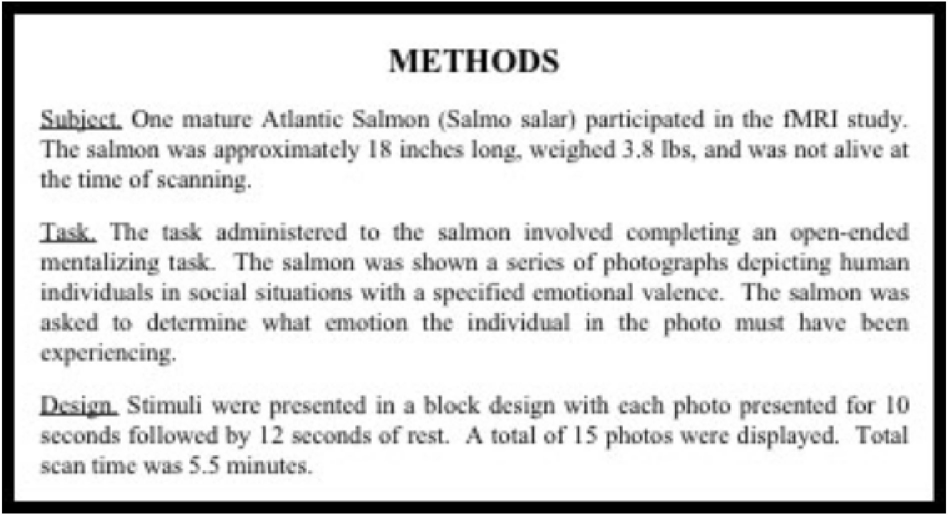
An example of *result fishing* : A salmon that reacts to photos of humans expressing various emotions

Experiments based on *Functional Magnetic Resonance Imaging* (fMRI)



An example of *result fishing* : A salmon that reacts to photos of humans expressing various emotions

Experiments based on *Functional Magnetic Resonance Imaging* (fMRI)



METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

And let's not forget the perils of *data mining*!

Data mining explicitly capitalizes on one of the key principles of both cherry-picking and question trolling—i.e., that if a researcher looks at enough sample results, he or she is bound to eventually find something that looks interesting. [. . .]

«*HARKing : How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results ?*», K.R. Murphy & H. Aguinis, J. of Bus. and Psy., 2017.

Not surprisingly, machine learning can amplify errors and distortions. Inconsistent training methods and poorly designed statistical frameworks lead to patterns and correlations that have no validity or link to causality in the real world.

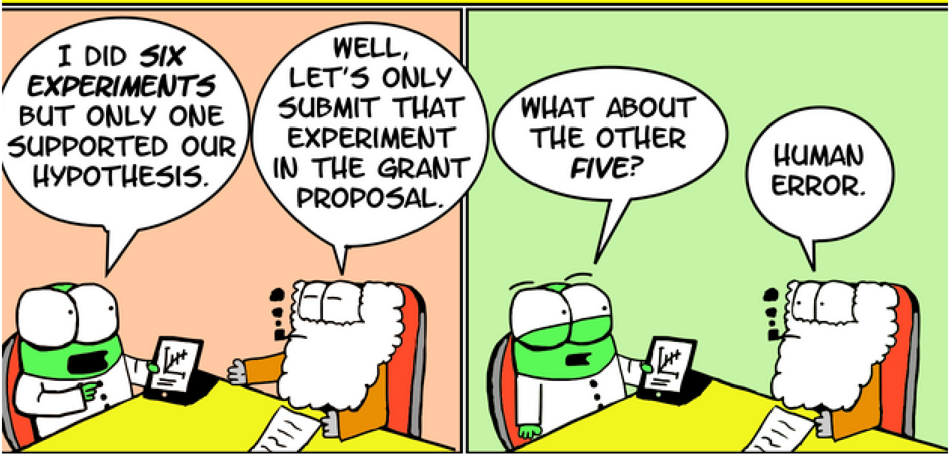
«*An Inability to Reproduce*», S. Greengard, Comm. of the ACM, Sept. 2019.

5.3 Other aspects

Confirmation bias

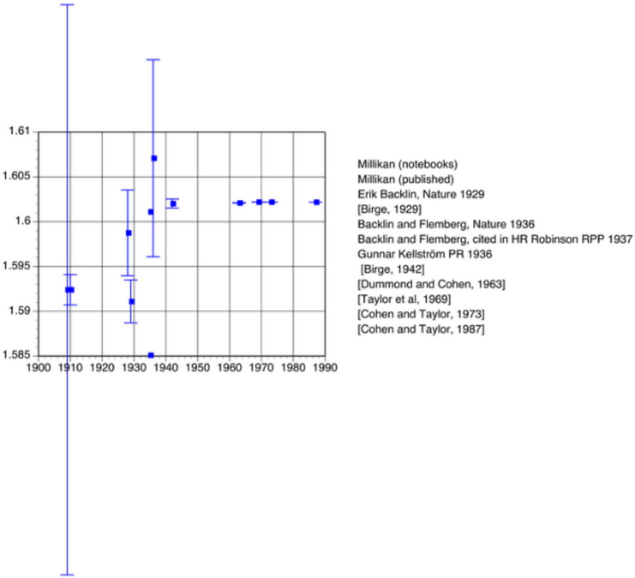
CONFIRMATION BIAS

FAVOURING EVIDENCE THAT SUPPORTS YOUR PRE-EXISTING BELIEFS WHILE IGNORING EVIDENCE THAT DOESN'T.


I DID *SIX* *EXPERIMENTS* BUT ONLY ONE SUPPORTED OUR HYPOTHESIS.
WELL, LET'S ONLY SUBMIT THAT EXPERIMENT IN THE GRANT PROPOSAL.
WHAT ABOUT THE OTHER *FIVE?*
HUMAN ERROR.

Elementary charge of the electron and the role of «negative» results (non-replication)

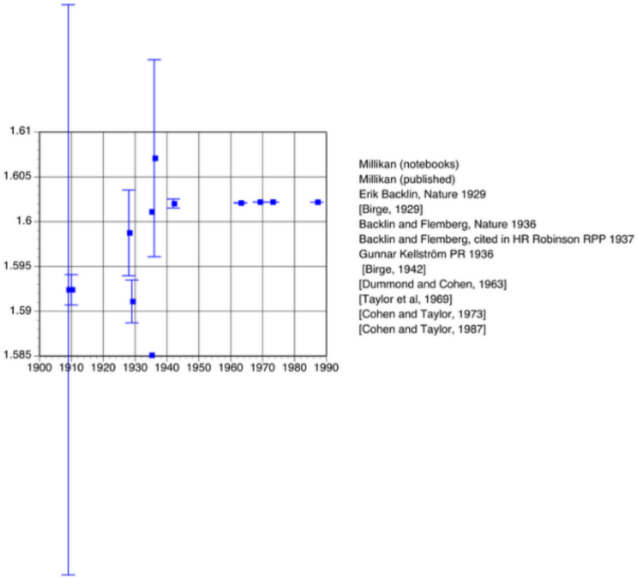
Initial work by R.A. Milikan ⇒ Nobel prize in Physics (1923)



But. . .

Elementary charge of the electron and the role of «negative» results (non-replication)

Initial work by R.A. Milikan ⇒ Nobel prize in Physics (1923)



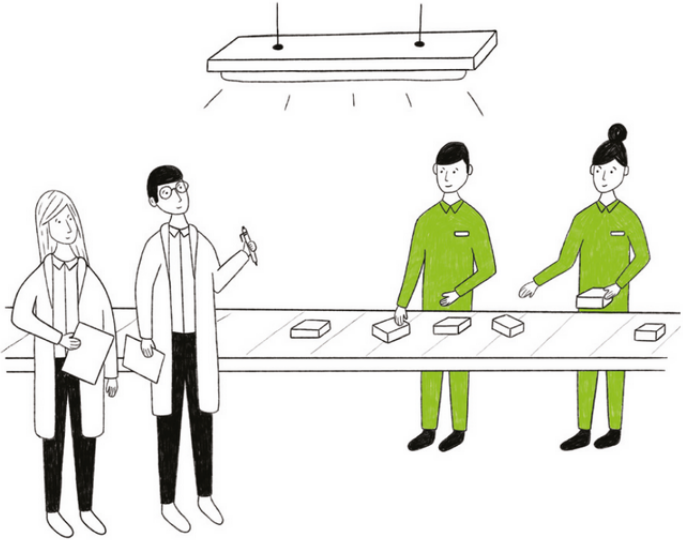
But. . .

«*Finding out that something does not work isn't going to win you a Nobel prize*»

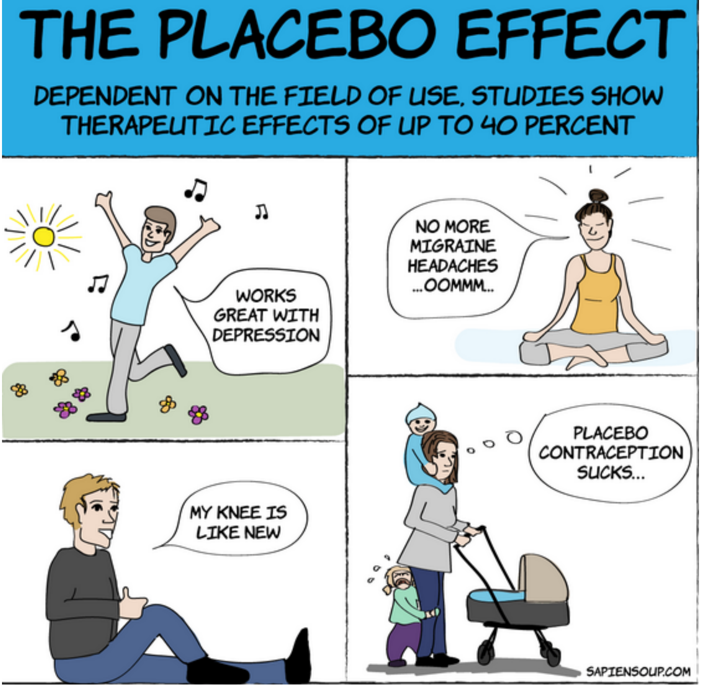
Experiments involving human subjects and Hawthorne effect

Hawthorn Effect ≈ Observer effect

https://www.geckoboard.com/learn/data-literacy/statistical-fallacies/hawthorne-effect/



Experiments involving human subjects and placebo effect


THE PLACEBO EFFECT
DEPENDENT ON THE FIELD OF USE, STUDIES SHOW THERAPEUTIC EFFECTS OF UP TO 40 PERCENT
WORKS GREAT WITH DEPRESSION
NO MORE MIGRAINE HEADACHES ...OOMMM...
MY KNEE IS LIKE NEW
PLACEBO CONTRACEPTION SUCKS...
SAPIENSOUP.COM

https://sapiensoup.com/placebo-homeopathy

PHARMACOLOGICAL DRUG TRIAL RESULTS
OUR TRIALS SHOW THAT THE NEW DRUG PERFORMS NO BETTER THAN PLACEBO
MAYBE WE SHOULD INVEST IN PLACEBOS
CHRIS MADDEN

https://drnancymalik.wordpress.com/2012/12/11/medicine-placebo-effect/

DRUGS
PLACEBOS
EX-STRENGTH PLACEBOS
DAVE CARPENTER...

Outline

1 Why this seminar ?

2 Is science in crisis ?

3 Some basic statistical concepts

4 Scientific method and statistical inference

5 Some causes of the crisis
   - Focus on «positive» and «novel» results (aka. «Publication bias»)
   - Flexibility in choosing experiment protocols and analyses
   - Other aspects

6 Conclusion : Some possible solutions ?

Conclusion : Some possible solutions ?

- Encourage replication studies

Conclusion : Some possible solutions ?

- Encourage replication studies
- Use tools to detect «dubious» results
  - GRIM/GRIMMER (Wansik !)
  - SPRITE

Conclusion : Some possible solutions ?

- Encourage replication studies
- Use tools to detect «dubious» results
  - GRIM/GRIMMER (Wansik !)
  - SPRITE
- Use open data… and require them (for publishing)

Conclusion : Some possible solutions ?

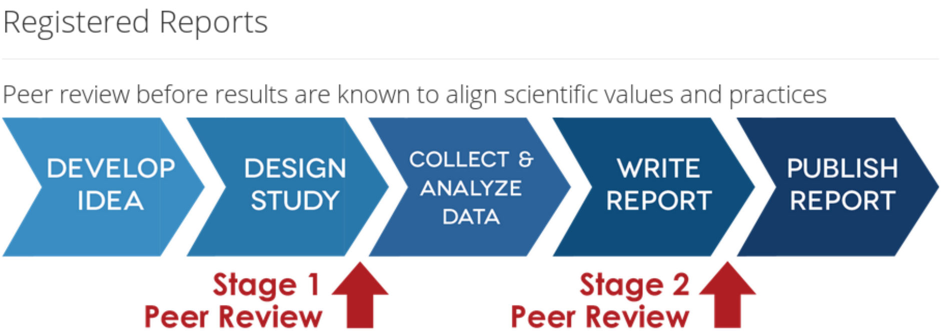
- Encourage replication studies
- Use tools to detect «dubious» results
  - GRIM/GRIMMER (Wansik !)
  - SPRITE
- Use open data. . . and require them (for publishing)
- Use $p < 0.01$ or $p < 0.005$

Conclusion : Some possible solutions ?

- Encourage replication studies
- Use tools to detect «dubious» results
  - GRIM/GRIMMER (Wansik !)
  - SPRITE
- Use open data. . . and require them (for publishing)
- Use $p < 0.01$ or $p < 0.005$
- Drop the use of NHST — Bayesian statistics ?

Conclusion : Some possible solutions ?

- Encourage replication studies
- Use tools to detect «dubious» results
  - GRIM/GRIMMER (Wansik !)
  - SPRITE
- Use open data. . . and require them (for publishing)
- Use $p < 0.01$ or $p < 0.005$
- Drop the use of NHST — Bayesian statistics ?
- Encourage «Registered reports»



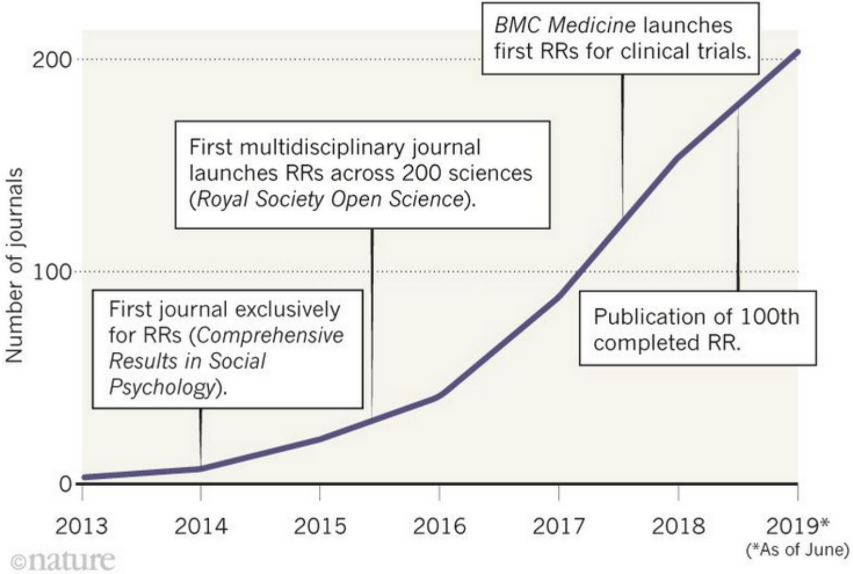
Registered Reports

Peer review before results are known to align scientific values and practices

Source: *Center for Open Science* : `https://osf.io/8mpji/wiki/home/`

Source: `https://www.nature.com/articles/d41586-019-02674-6`

Since 2013, the number of journals offering Registered Reports (RRs) has risen to more than 200 titles.



First journal exclusively for RRs (*Comprehensive Results in Social Psychology*).

First multidisciplinary journal launches RRs across 200 sciences (*Royal Society Open Science*).

BMC Medicine launches first RRs for clinical trials.

Publication of 100th completed RR.

Source: C. Chambers

To learn more about this. . .

N. Chevassus-au Louis.
Malscience — De la fraude dans les labos.
Éditions du Seuil, 2016.

C. Chambers.
The seven deadly sins of psychology : A manifesto for reforming the culture of scientific practice.
Princeton University Press, 2017.

N. Gauvrit.
Statistiques — Méfiez-vous !
Ellipses, 2007.

S. Greengard
An Inability to Reproduce.
Comm. of the ACM, 62(9) :13-15, 2019.

R.R. Haccoun and D. Cousineau.
Statistiques—Concepts et applications (Deuxième édition revue et augmentée).
Les Presses de l'Université de Montréal, 2010.

To learn more about this. . .

J.P.A. Ioannidis.
Why most published research findings are false.
PLoS Medicine, 2(8) :e124, 2005.

J.P.A. Ioannidis.
What have we (not) learnt from millions of scientific paper with p values ?
The American Statistician, 73(S1) :20–25, 2019.

D. Randall and C. Welser.
The irreproducibility crisis of modern science—Causes, consequences, and the road to reform.
Technical report, National Association of Scholars, 2018.

F. Shull, J. Singer, and D.I.K. Sjoberg, editors.
Guide to Advanced Empirical Software Engineering.
Springer, 2008.

R.L. Wasserstein and N.A. Lazar.
The ASA's statement on p-values : Context, process, and purpose.
The American Statistician, 70(2) :129–133, 2016.

A. Zeller, T. Zimmermann, and C. Bird.
Failure is a four-letter word : A parody in empirical research.
In *Proc. of the 7th Int. Conf. on Predictive Models in Software Engineering*. ACM, 2011.

Comments ?

Questions ?