Règles de l'art de l'analyse statistique de données

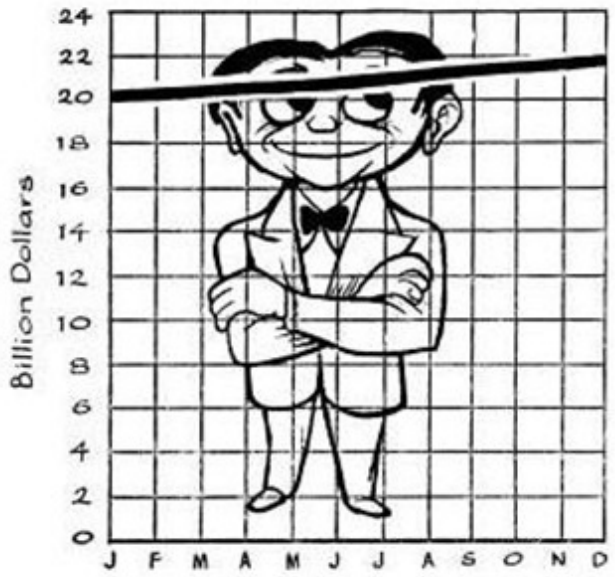
G. Haesbroeck

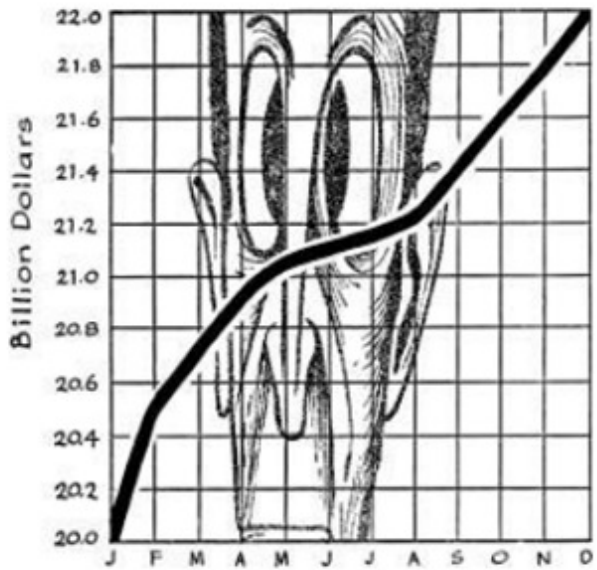
Colloque annuel de la Société Royale des Sciences de Liège

25 novembre 2022

Dans la vie de tous les jours, la statistique n'a pas toujours bonne presse

There are three kinds of lies: lies, damned lies, and statistics
(dicton popularisé par Marc Twain)
Les chiffres disent toujours ce que souhaite l'homme habile qui sait en jouer (Thomas Babington Macaulay)





Source: How to lie with statistics, Darrel Huff (1954).

La statistique telle qu'exploitée en recherche pose également question

Open access, free

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

PLoS Medicine | www.plosmedicine.org

August 2005 | Vc

Biostatistics (2014), **15**, 1, *pp.* 1–12
doi:10.1093/biostatistics/kxt007
Advance Access publication on September 25, 2013

R

An estimate of the science-wise false discovery rate and application to the top medical literature

LEAH R. JAGER

Department of Mathematics, United States Naval Academy, Annapolis, MD 21402, USA

JEFFREY T. LEEK*

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

jleek@jhsph.edu

Coupable désigné: la *p*-valeur

Scientific method: Statistical errors

_P_ values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

Regina Nuzzo

12 February 2014

La *p*-valeur

Sir R. A. Fisher

Plan d'expérience

A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested. [...] [It] consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of that the test will consist, namely, that she will be asked to taste eight cups, that these shall be four of each kind [...]. — Fisher, 1935.

Hypothèse nulle: la dame n'a pas la capacité de distinguer les tasses;
Résultat de l'expérience: elle sélectionne les 4 bonnes tasses.
p-valeur: $\boldsymbol{P}(4 \text{ choix corrects} \mid H_0) = 1/70 = 0.014$

Dans son livre *The design of experiments* (1935), Fisher introduit la notion d'*hypothèse nulle* et l'idée que celle-ci ne peut jamais être prouvée, mais bien réfutée par des données. Il définit la *p*-valeur comme la probabilité d'obtenir, sous l'hypothèse nulle, un résultat égal ou plus extrême que celui observé.

Le concept était apparu avant

Article de Karl Pearson publié en 1900 dans le *Philosophical Magazine*:

X. *On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling.* By KARL PEARSON, *F.R.S.*, *University College, London**.

Illustration III.

In the case of runs of colour in the throws of the roulette-ball at Monte Carlo, I have shown* that the odds are at least 1000 millions to one against such a fortnight of runs as occurred in July 1892 being a random result of a true roulette. I now give χ^2 for the data printed in the paper referred to, *i. e.*:

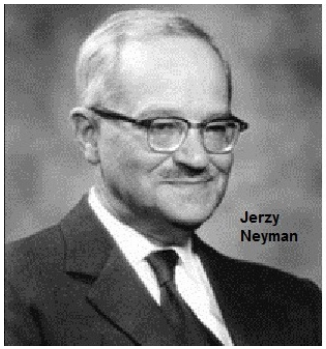
4274 Sets at Roulette.

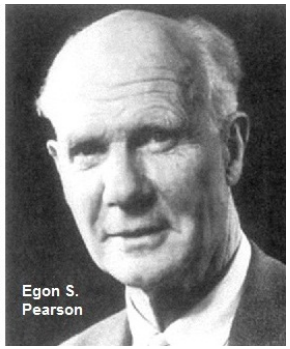
Runs	1	2	3	4	5	6	7	8	9	10	11	12	Over 12
Actual ...	2465	945	333	220	135	81	43	30	12	7	5	1	0
Theory ...	2137	1068	534	267	134	67	33	17	8	4	2	1	0

From this we find $\chi^2 = 172\,43$, and the improbability of a series as bad as or worse than this is about $14{\cdot}5/10^{30}$! From this it will be more than ever evident how little chance had to do with the results of the Monte Carlo roulette in July 1892.

Statistique "moderne": *Null Hypothesis Statistical Testing*

J. Neyman et E. S. Pearson





On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I

Author(s): J. Neyman and E. S. Pearson

Source: *Biometrika*, Jul., 1928, Vol. 20A, No. 1/2 (Jul., 1928), pp. 175-240

Published by: Oxford University Press on behalf of Biometrika Trust

Stable URL: https://www.jstor.org/stable/2331945

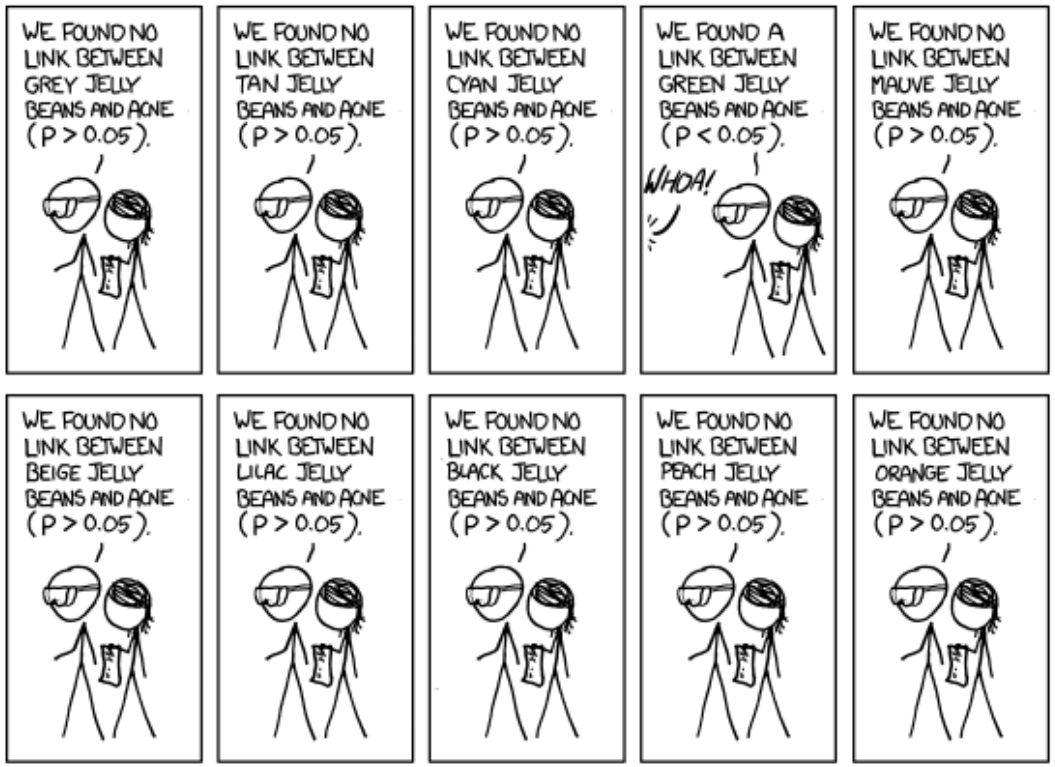
Introduction d'une hypothèse alternative et proposition de *contrôler* les erreurs de type I (α) et de type II (β):

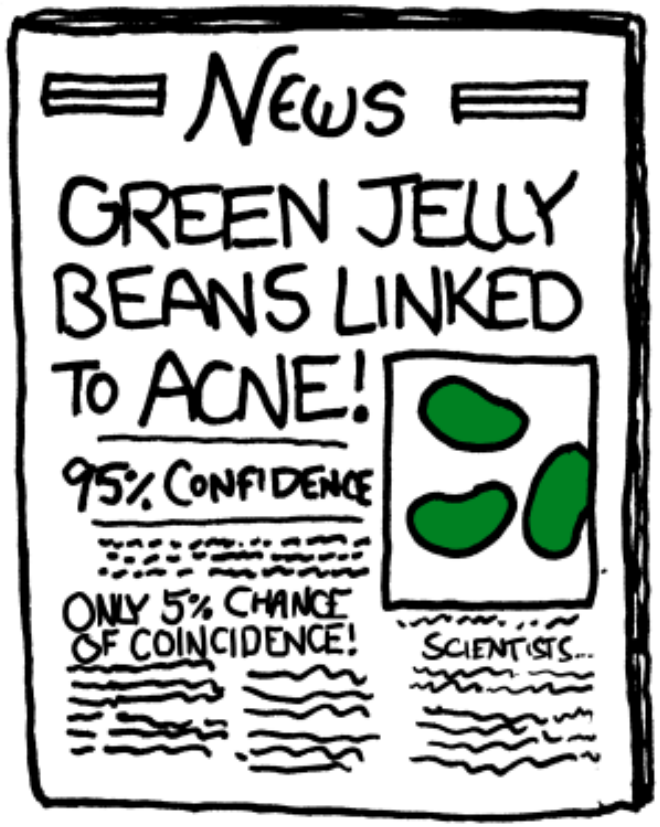
$$\alpha = \boldsymbol{P}(RH_0 \mid H_0) \text{ et Puissance}=1-\beta = \boldsymbol{P}(RH_0 \mid H_1)$$

	Décision	
Réalité	Rejeter H_0	Ne pas rejeter H_0
H_0 est vraie	x	v
H_0 est fausse	v	x

Cette approche a transformé l'outil flexible de Fisher en un *algorithme rigoureux*, dont l'exploitation *à tout prix* a mené à certaines dérives...

Emergence de pratiques telles que le *p*-hacking, "cherry picking"...





Ce qui a donné lieu, notamment, à un biais de publication: les recherches montrant un "effet significatif" sont publiées plus facilement et plus rapidement que les recherches ne mettant par en évidence un résultat significatif.

Réactions diverses

SOCIAL SELECTION *Popular articles on social media*

Psychology journal bans *P* values

A controversial statistical test has met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values, because the values were too often used to support lower-quality research.

Authors are still free to submit papers to *BASP* with *P* values and other statistical measures that form part of 'null hypothesis significance testing' (NHST), but the numbers will be removed before publication. "Basic and Applied Social Psychology just went science rogue and banned NHST from their journal. Awesome," tweeted Nerisa Dozo, a PhD student in psychology at the University of Queensland in Brisbane, Australia. But Jan de Ruiter, a cognitive scientist at Bielefeld University in Germany, tweeted: "NHST is really problematic", adding that banning all inferential statistics is "throwing away the baby with the p-value".

Basic Appl. Soc. Psych. 37, 1–2 (2015)

Based on data from altmetric.com. Altmetric is supported by Macmillan Science and Education, which owns Nature Publishing Group.

NATURE.COM
For more on popular papers:
go.nature.com/ynfi49

5 MARCH 2015 | VOL 519 | NATURE | 9

Comments from the New Editor

Published online by Cambridge University Press: **29 January 2018**

Jeff Gill

Article Metrics

In addition, *Political Analysis* will no longer be reporting p -values in regression tables or elsewhere. There are many principled reasons for this change—most notably that in isolation a p -value simply does not give adequate evidence in support of a given model or the associated hypotheses. There is an extremely large, and at times self-reflective, literature in support of that statement dating back to 1962. I could fill all of the pages of this issue with citations. Readers of *Political Analysis* have surely read the recent American Statistical Association report on the use and misuse of p -values, and are aware of the resulting public discussion. The key problem from a journal's perspective is that p -values are often used as an acceptance threshold leading to publication bias. This in turn promotes the poisonous practice of model mining by researchers. Furthermore, there is evidence that a large number of social scientists misunderstand p -values in general and consider them a key form of scientific reasoning. I hope other respected journals in the field follow our lead.

Réaction également de la très renommée "American Statistical Association"

AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

Statement on *p*-values

The statement's six principles, many of which address misconceptions and misuse of the *p*-value, are the following:

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

Edition spéciale de la revue *The American Statistician* (mars 2019): Statistical Inference in the 21st Century: A World Beyond $p < 0.05$

The *p*-Value Requires Context, Not a Threshold ›

Rebecca A. Betensky

Abandon Statistical Significance ›

Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert & Jennifer L. Tackett

Why is Getting Rid of *P*-Values So Hard? Musings on Science and Statistics ›

Steven N. Goodman

Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of *p*-Values ›

John L. Kmetz

Will the ASA's Efforts to Improve Statistical Practice be Successful? Some Evidence to the Contrary ›

Raymond Hubbard

Quelle est donc la bonne approche?

Il y a de nombreuses publications présentant *des règles de l'art* pour l'application de la statistique en science.

PLOS | COMPUTATIONAL BIOLOGY

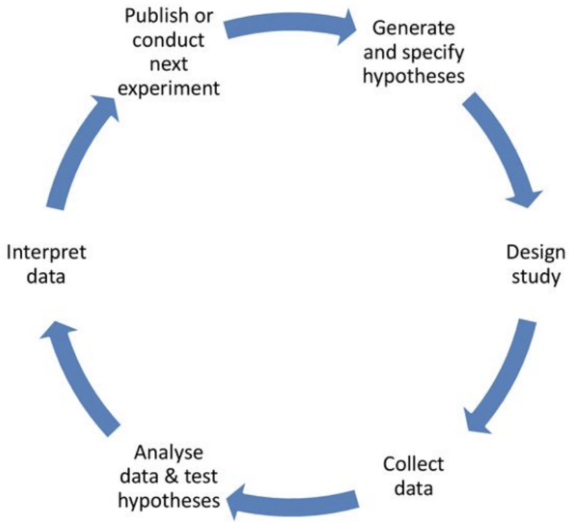
Ten Simple Rules for Effective Statistical Practice

Robert E. Kass, Brian S. Caffo, Marie Davidian, Xiao-Li Meng, Bin Yu, Nancy Reid

Published: June 9, 2016 • http://dx.doi.org/10.1371/journal.pcbi.1004961

1. Statistical methods should enable data to answer scientific questions
2. Signals always come with noise
3. Plan ahead, really ahead
4. Worry about data quality
5. Statistical analysis is more than a set of computations
6. Keep it simple
7. Provide assessments of variability
8. Check your assumptions
9. When possible, replicate!
10. Make your analysis reproducible

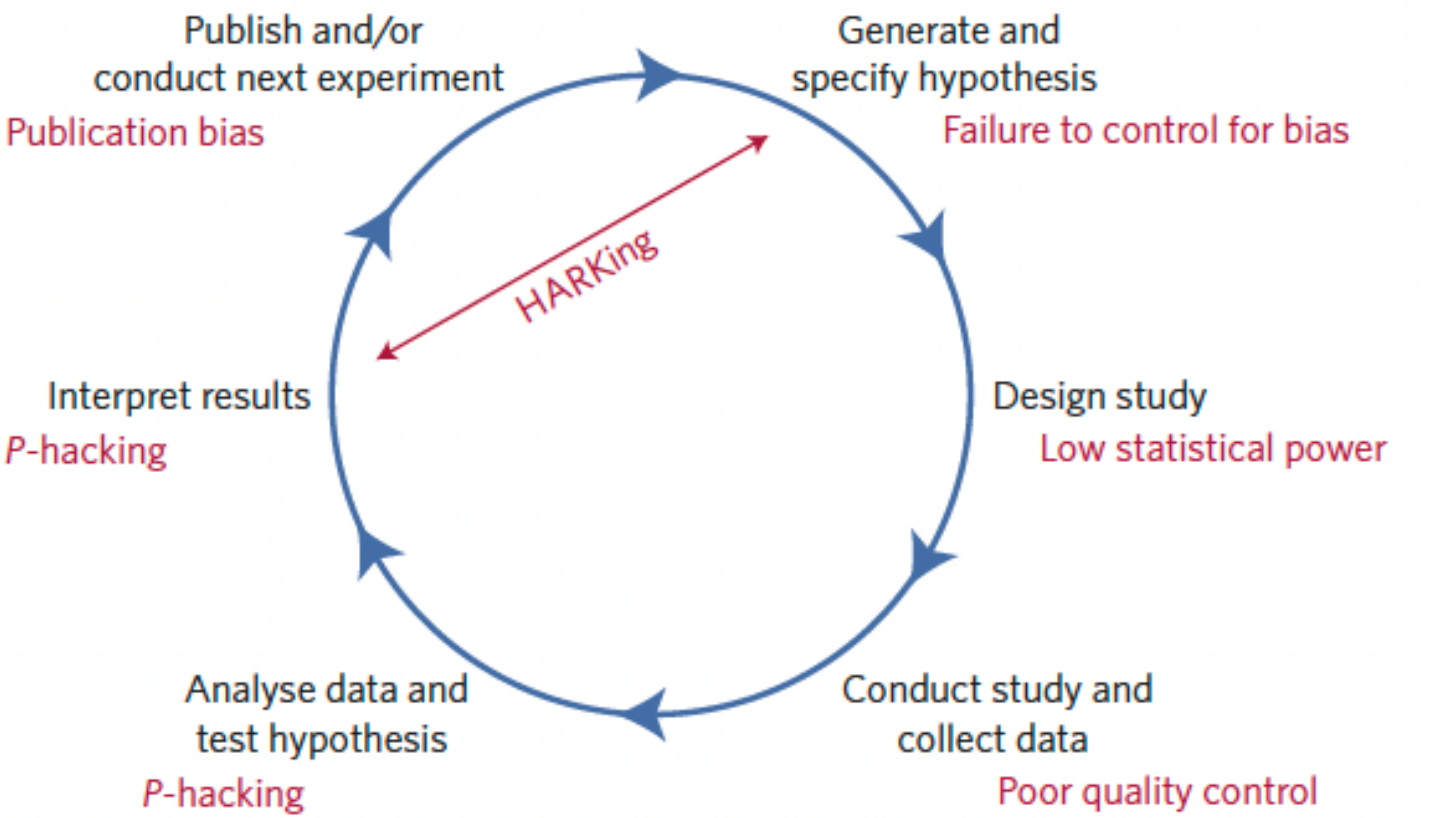
The "Ideal":
Hypothetico-Deductive Model



Barbara A. Spellman, Challenges for journals: Encouraging sound science.

Disposer d'un guide de bonnes pratiques n'est cependant par suffisant

Le modèle idéal est "menacé" à chaque étape:



Source: A manifesto for reproducible science, Munafo *et al* (*Nature Human Behavior*, 2017).

Il faut développer le raisonnement statistique

Cela implique la capacité de déjouer les pièges classiques:

- Non prise en compte de *facteurs confondants*
  - Paradoxe de Simpson dans les études observationnelles
  - Corrélation et causalité
- Violation d'hypothèses (normalité, indépendance...)
- Non compréhension de l'impact de la taille de l'échantillon et/ou du "curse of high dimensionality" sur les méthodes statistiques
- ...

Source: Statistical Inference Enables Bad Sciences; Statistical Thinking Enables Good Sciences, Tong (*The American Statistician*, 2019).

Paradoxe de Simpson dans les études observationelles

BRITISH MEDICAL JOURNAL VOLUME 292 29 MARCH 1986

Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy

C R CHARIG, D R WEBB, S R PAYNE, J E A WICKHAM

Abstract

This study was designed to compare different methods of treating renal calculi in order to establish which was the most cost effective and successful. Of 1052 patients with renal calculi, 350 underwent open surgery, 350 percutaneous nephrolithotomy, 328 extracorporeal shockwave lithotripsy (ESWL), and 24 both percutaneous nephrolithotomy and ESWL. Treatment was defined as successful if stones were eliminated or reduced to less than 2 mm after three months. Success was achieved in 273 (78%) patients after open surgery, 289 (83%) after percutaneous nephrolithotomy, 301 (92%) after ESWL, and 15 (62%) after percutaneous nephrolithotomy and ESWL. Comparative total costs to the NHS were estimated as £3500 for open surgery, £1861 for percutaneous nephrolithotomy, £1789 for ESWL, and £3210 for both ESWL and nephrolithotomy. ESWL caused no blood loss and little morbidity and is the cheapest and quickest way of returning patients to normal life.

Le traitement “open surgery” (notons le A) est moins performant que le traitement “percutaneous nephrolithotomy” (B).

Informations complémentaires au sein de l'article

Les patients étaient répartis en deux groupes selon la taille des pierres (diamètre $<$ ou $\geq$ 2 cm)

	Traitement A		Traitement B	
	$<$ 2cm	$\geq$ 2cm	$<$ 2cm	$\geq$ 2cm
Nombre de cas	87	263	270	80
Nombres de succès	81	192	234	55
Taux de succès	93%	73%	87%	69%

Le traitement A est plus performant que le traitement B pour les patients du groupe 1 et pour les patients du groupe 2, mais est moins performant globalement. C'est dû au fait que parmi les patients traités avec le traitement A, 75% présentaient des gros cailloux; alors que le pourcentage de patients dans ce cas-là et traités par B était égal à 23%!

En étant optimiste, on peut se dire que ce genre d'erreurs n'est plus d'actualité dans la littérature...

Et pourtant: article dans le Futura Santé du 27 juillet 2021

C'est une statistique assez perturbante dont se nourrissent les antivax : entre le 1er février et le 21 juin 2021 en Angleterre, il y a eu davantage de morts de la Covid-19 parmi les personnes vaccinées et testées positives au variant Delta que parmi les personnes non vaccinées, atteste un rapport du gouvernement. Plus exactement, sur les 257 décès, 45 personnes avaient reçu une dose, 118 avaient reçu deux doses, et 92 n'avaient reçu aucune injection. La conclusion hâtive qu'il pourrait être tiré de ces statistiques est que le vaccin n'est pas efficace, voire dangereux.

Données du rapport du gouvernement

Table 5. Attendance to emergency care and deaths by vaccination status among Delta confirmed cases (sequencing and genotyping) including all confirmed Delta cases in England, 1 February 2021 to 21 June 2021

	Age group (years)**	Total	Cases with specimen date in past 28 days	Vaccination status unknown	<21 days post dose 1	≥21 days post dose 1	Received 2 doses	Unvaccinated
Delta cases	All cases	123,620	63,707	14,359	8,562	17,933	10,834	71,932
	<50	111,008	57,673	12,900	8,453	13,391	5,600	70,664
	≥50	12,404	5,957	1,252	109	4,542	5,234	1,267
Deaths within 28 days of positive specimen date	All cases	257	N/A	2	1	44	118	92
	<50	26	N/A	-	-	3	2	21
	≥50	231	N/A	2	1	41	116	71

Effet global $\longleftrightarrow$ effet local

En ne prenant que les catégories `Vaccinés 2 doses` et `Non Vaccinés`:

	`Vaccinés 2 doses`	`Non Vaccinés`
Nombre de cas	10834	71932
Nombre de décès	118	92
Risque	1%	0,1%

Mais, en tenant compte de l'âge

	`Vaccinés 2 doses`		`Non Vaccinés`	
	<50 ans	≥ 50 ans	<50 ans	≥ 50 ans
Nombre de cas	5600	5234	70644	1267
Nombre de décès	2	116	21	71
Risque	0,03%	2,2%	0,03%	5,6%

on constate que les risques de décès sont égaux ou plus importants pour les non vaccinés par rapport aux vaccinés dans chaque "tranche d'âge".

La catégorie plus âgée est "sous-représentée" dans la catégorie des non vaccinés, ce qui change l'effet (paradoxe de Simpson).

Facteur confondant dans la phase d'interprétation et confusion entre corrélation et causalité

Exemple "réputé" : *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, F. H. Messerli, N Engl J Med 2012

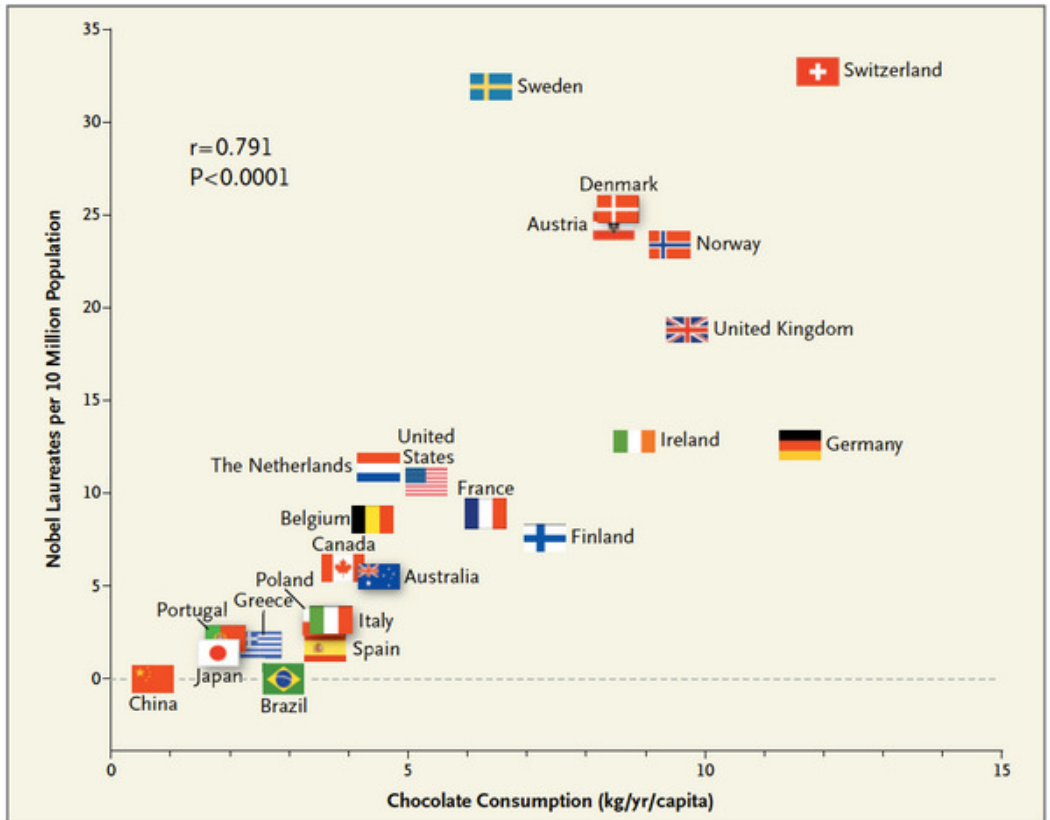


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Correlation of chocolate consumption with Nobel Laureates (Image credit: New England Journal of Medicine)

Facteur confondant

rtbf.be ACCUEIL VIDÉO AUDIO MON CHOIX CHAÎNES THÉMATIQUES PLUS

Consommation de chocolat et Prix Nobel: "aucun lien", selon des chercheurs

13 juin 2013 à 11:53 - mise à jour 13 juin 2013 à 13:18 · 1 min

Par Belga News

L'étude parue l'an dernier avançait que les flavonoïdes, de puissants antioxydants contenus en grande quantité dans le cacao, le thé et le vin rouge amélioraient les fonctions mentales.

Ils ont également établi une forte corrélation entre les magasins IKEA d'un pays et son nombre de Prix Nobel, pour démontrer par l'absurde que "*le rapport entre deux variables ne signifie pas qu'un lien réel existe entre elles*". Il serait en effet "*farfelu de supposer que la nécessité de comprendre les instructions de montage des meubles IKEA augmente le niveau d'intelligence d'un pays*", expliquent les scientifiques, pour qui le lien entre deux variables "*dépend surtout d'une troisième variable*".

Dans le cas du chocolat et du Prix Nobel, la troisième variable serait le PIB du pays. Ce qu'a observé le docteur Franz Messerli, auteur de l'essai paru en 2012, "*pourrait donc être en partie expliqué par le fait que le chocolat est un produit de luxe et serait davantage consommé dans les pays où le développement économique est optimal*" et donc propice à l'obtention de Prix Nobel, concluent les chercheurs.

Calculer des corrélations n'est pas suffisant pour montrer un lien de cause à effet.

Rousseuw Prize for Statistics (1 million de dollars)

First Rousseeuw Prize for Statistics awarded for pioneering research on causal inference

13 Oct 2022

The biennial Rousseeuw Prize for Statistics rewards excellence in statistical research that significantly impacts everyday life. During a ceremony at KU Leuven, five laureates and their pioneering research were recognised and celebrated in the presence of His Majesty King Philippe of Belgium.

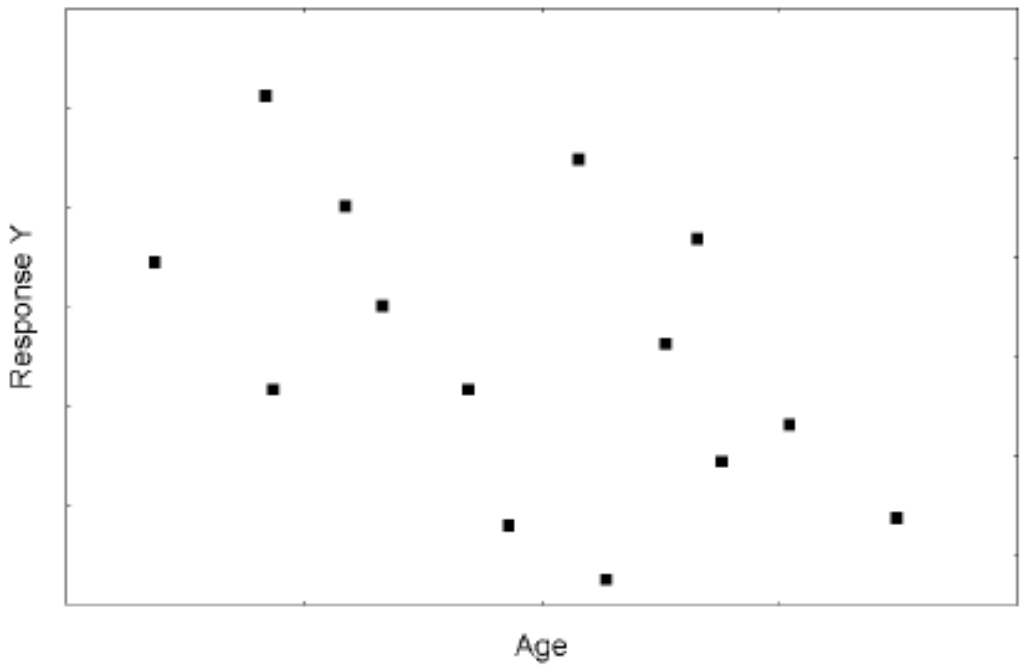
This new scientific prize, worth 1 million dollars, was established by Peter Rousseeuw, emeritus professor of Statistics at KU Leuven, and will be awarded biennially by the King Baudouin Foundation. It aims to reward excellence in statistical research that significantly impacts everyday life.

Cause and effect: difficult to detect

Robins and his collaborators are recognised for their pioneering research on causal inference with applications in medicine and public health. Their insights and newly developed statistical methods made it possible to better distinguish cause from correlation.

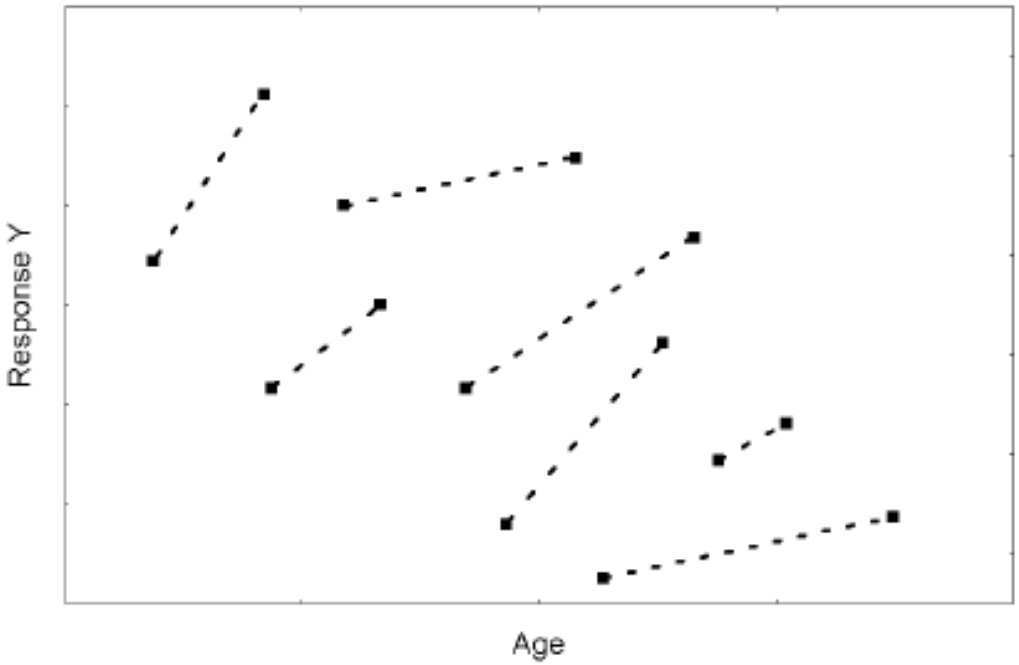
Validation des hypothèses: indépendance (iid)

On s'intéresse à la relation entre une variable réponse Y et l'âge. On dispose des données suivantes:



Violation de l'indépendance

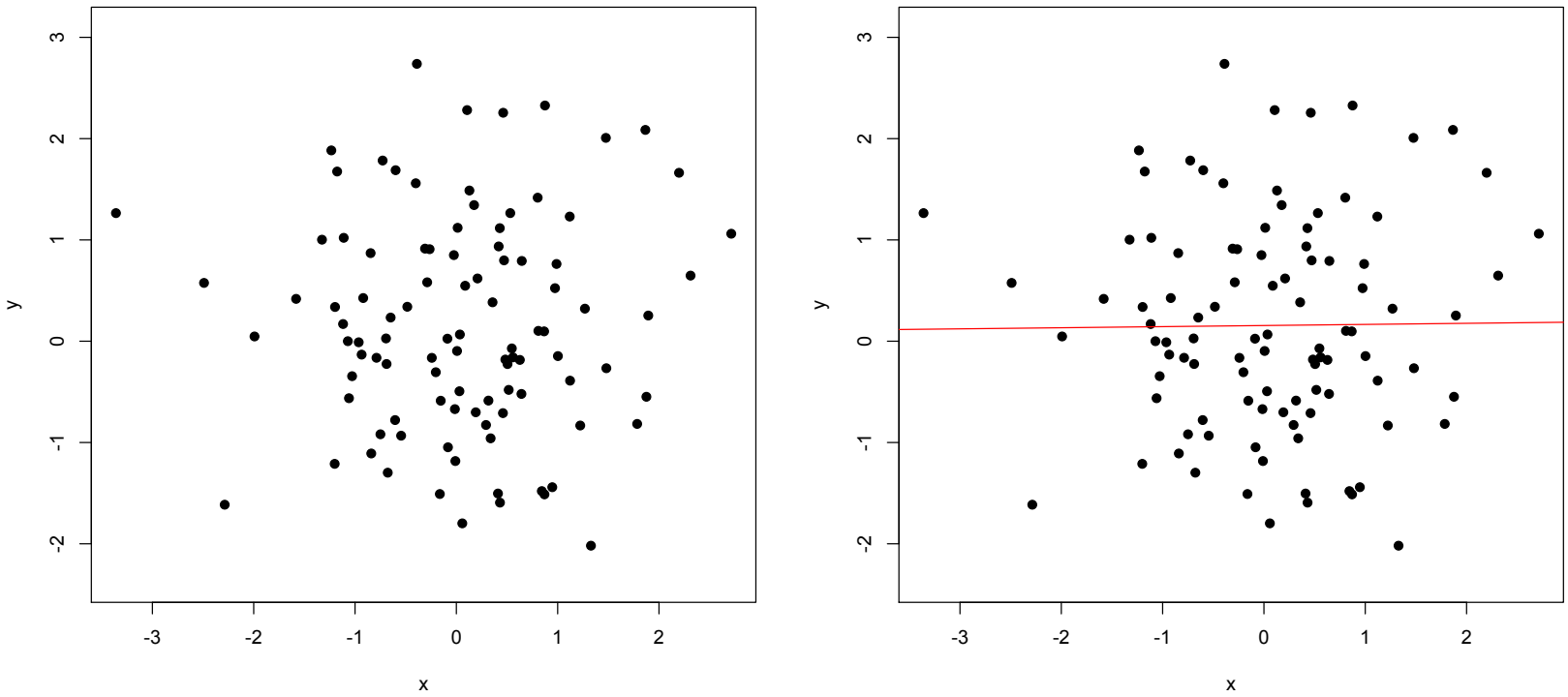
Si on précise maintenant qu'il s'agit de deux mesures répétées sur chaque individu, que devient la relation?



Il faut tenir compte de la dépendance entre les mesures!
Non seulement l'effet peut changer mais en plus, la taille de l'échantillon doit être divisée par 2!

Validation des hypothèses: normalité

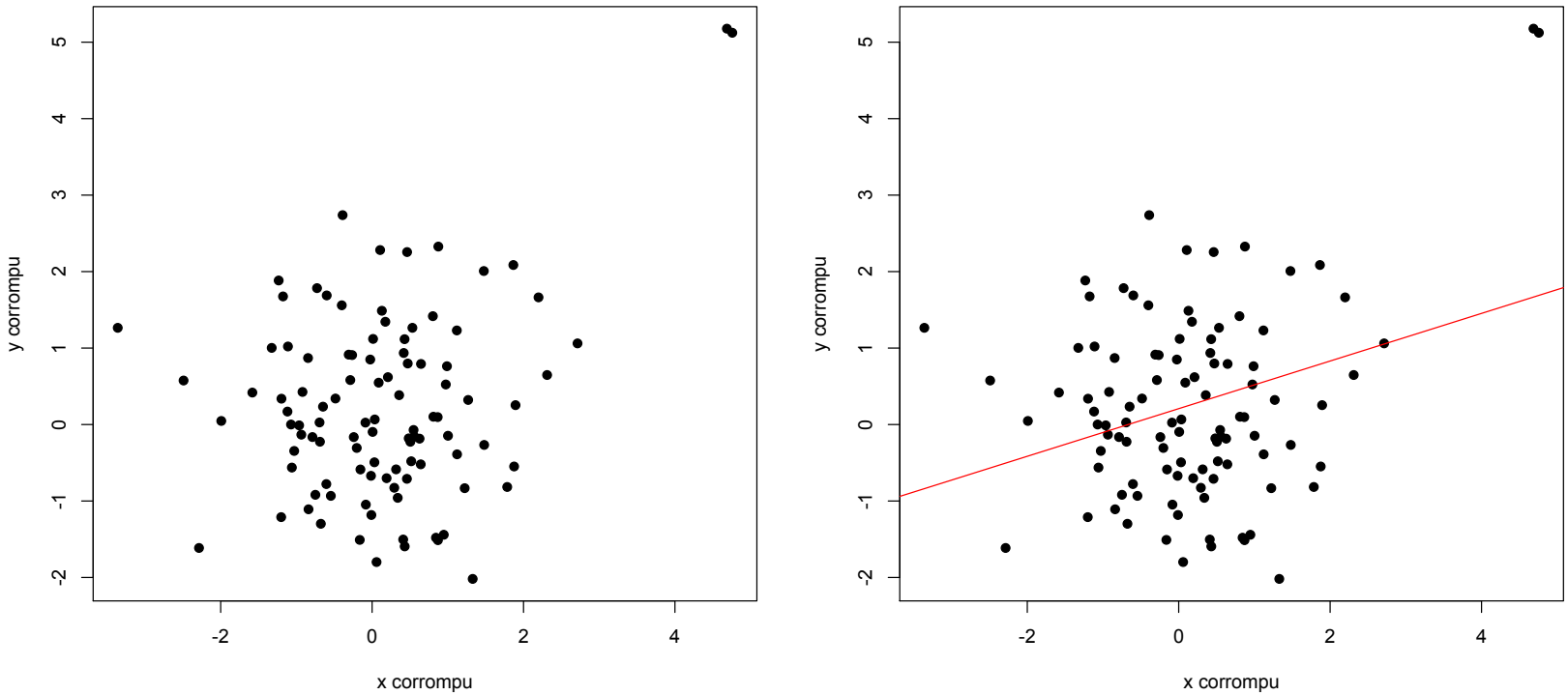
Soient deux variables aléatoires indépendantes de distribution $N(0, 1)$ et un échantillon de taille $n = 100$:



Comme attendu, la droite des moindres carrés est horizontale, aucune tendance ni association n'est détectée (cor=0.011, p-valeur=0.92).

Avec deux observations atypiques

Supposons que les données soient corrompues par 2 observations atypiques:



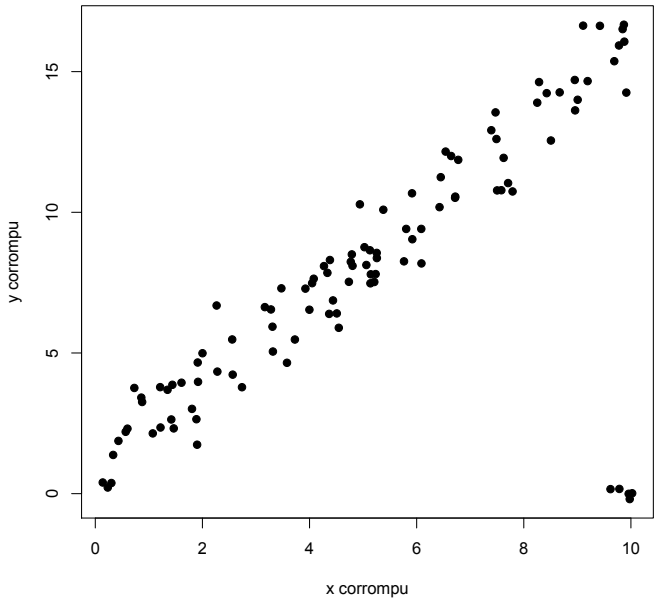
Dans ce cas, la technique des moindres carrés détecte une tendance et l'absence de corrélation est rejetée (cor$=0.30$, p-valeur$=0.002$).

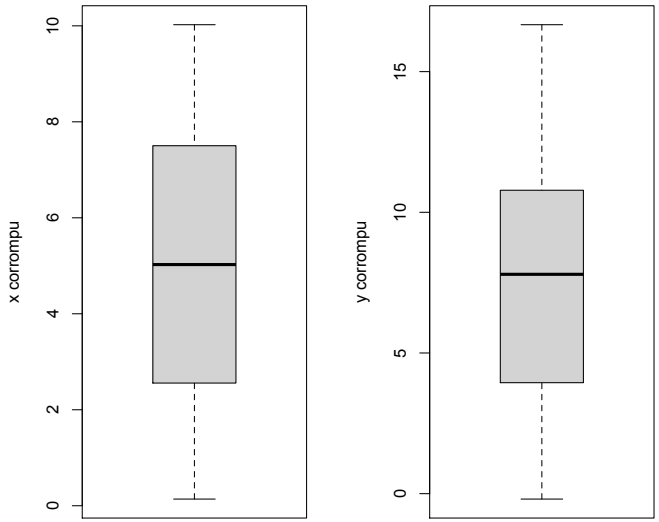
... comme attendu mathématiquement vu le critère imposé.

NB: la normalité des résidus n'est pas rejetée...

Observations atypiques/influentes en dimension p

L'exemple précédent peut paraitre peu convaincant car les observations atypiques sont "visibles". Cependant, ces observations atypiques ne sont pas toujours faciles à épingler:

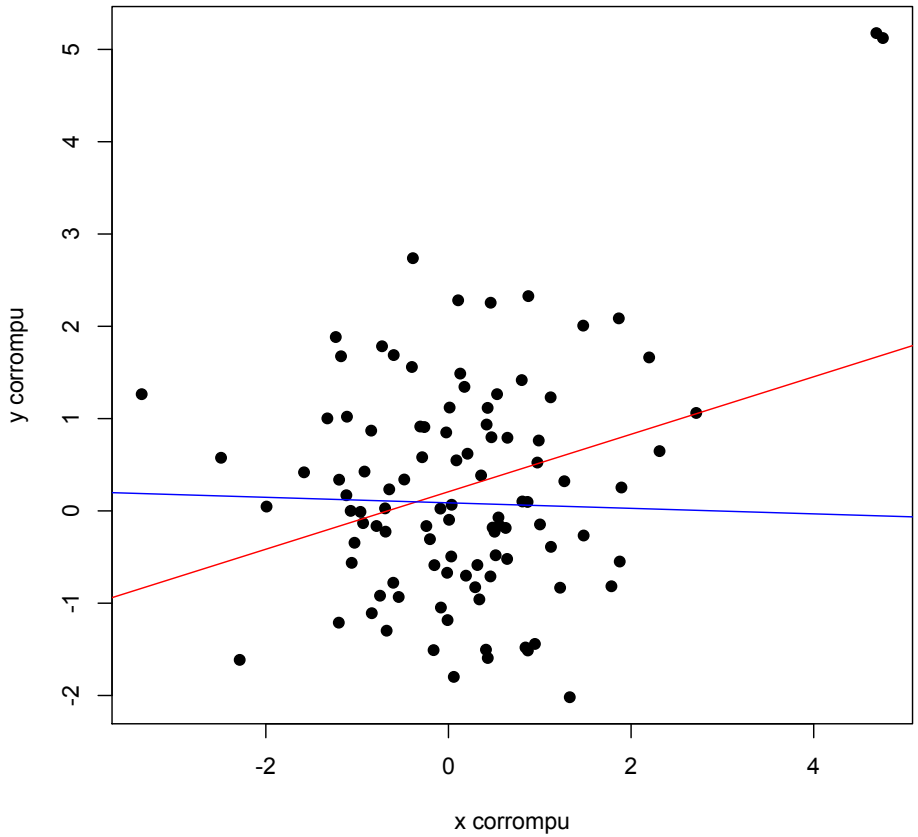




Les observations atypiques en dimension p ne sont pas nécessairement repérables en dimension $p-1$.

Utilisation de techniques “robustes”?

Droite de régression LTS (*least trimmed squares*) de P.J. Rousseeuw:



Corrélation de rang de Kendall: 0.023.

Taille d'échantillon et dimension

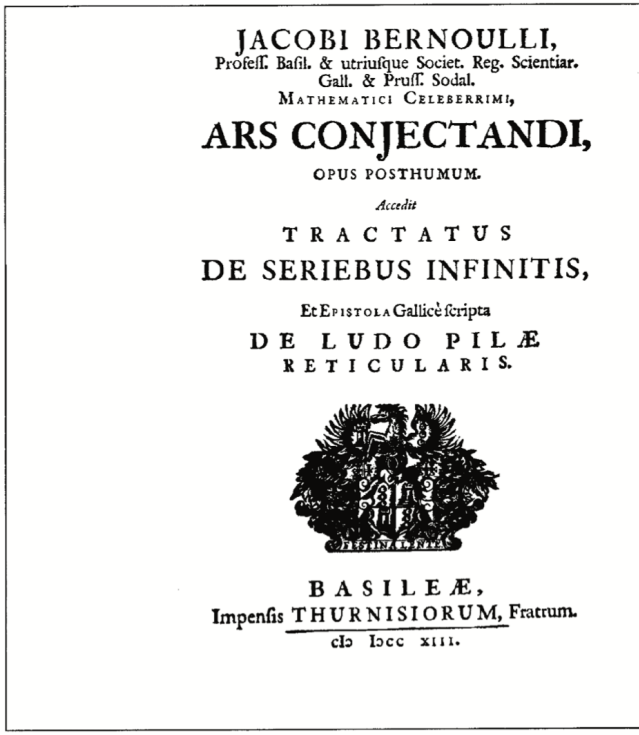
La taille n d'un échantillon est liée à de nombreux aspects de l'expérimentation et de l'analyse statistique:

- Coût de l'expérimentation
- Protocole de l'expérimentation
  - Données répétées ?
  - Individus discernables?
  - ...
- Nombre p de facteurs considérés :
  Rule of thumb: $n/p > 5$

  NB: on parle de données *plates* si $n < p$ et dans un tel contexte, certaines techniques usuelles ne sont plus appropriées.
- ...

n trop petit

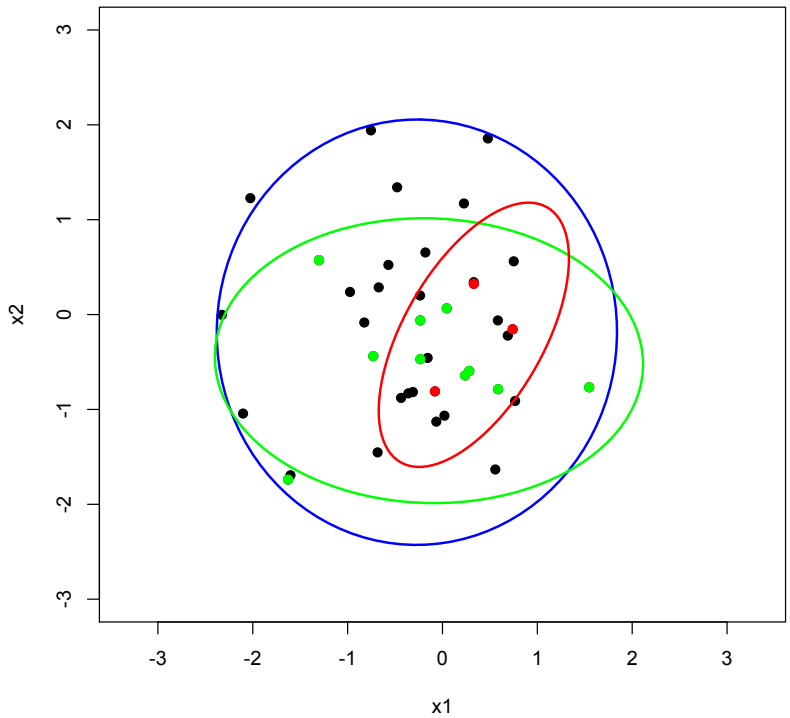
Il ne peut échapper à personne que, pour juger par ce moyen de quelque événement, il ne suffirait pas d'avoir fait choix d'une ou de deux expériences, mais qu'il serait requis une grande quantité d'expériences : tout être des plus stupides, par je ne sais quel instinct naturel, par lui-même et sans le guide d'aucun enseignement (chose absolument admirable) tient pour évident que, plus on aura recueilli de nombreuses observations de ce genre, moins grand sera le danger de s'écarter du but.


JACOBI BERNOULLI,
Profess. Basil. & utriusque Societ. Reg. Scientiar.
Gall. & Pruss. Sodal.
Mathematici Celeberrimi,
ARS CONJECTANDI,
OPUS POSTHUMUM.
Accedit
TRACTATUS
DE SERIEBUS INFINITIS,
Et Epistola Gallicè scripta
DE LUDO PILÆ
RETICULARIS.
BASILEÆ,
Impensis THURNISIORUM, Fratrum.
clɔ Iɔcc xiii.

Ars Conjectandi, Jacques Bernoulli (rédaction entre 1684 et 1689), publication par son neveu, Nicolas Bernoulli, en 1713.

Problèmes avec n petit

- Perte de puissance
- Risque de détecter des effets non présents:
  Ex: détection d'une corrélation entre deux variables indépendantes



n "trop grand"

Tout écart infinitésimal devient statistiquement significatif mais probablement scientifiquement inintéressant!

Exemple: test d'indépendance entre deux variables binaires:
Supposons que la distribution jointe soit la suivante:

	Variable Y	
Var X	Succès	Echec
Succès	0.27	0.23
Echec	0.23	0.27

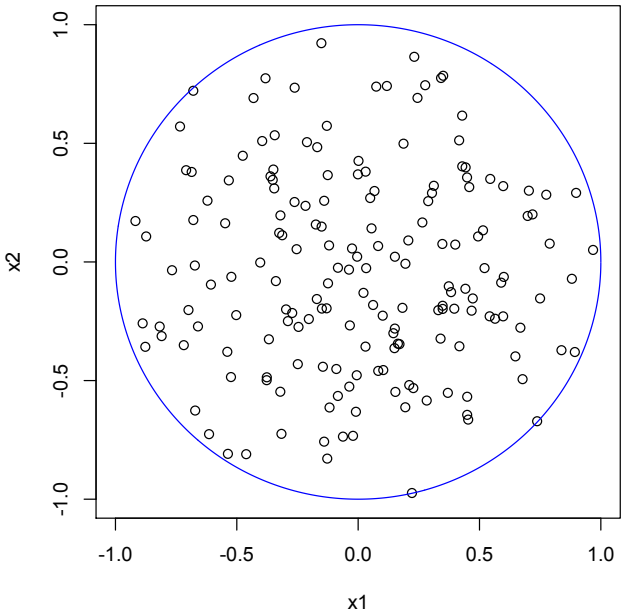
En supposant que l'on obtienne des échantillons respectant parfaitement la distribution jointe, le test d'indépendance du χ^2 donne:

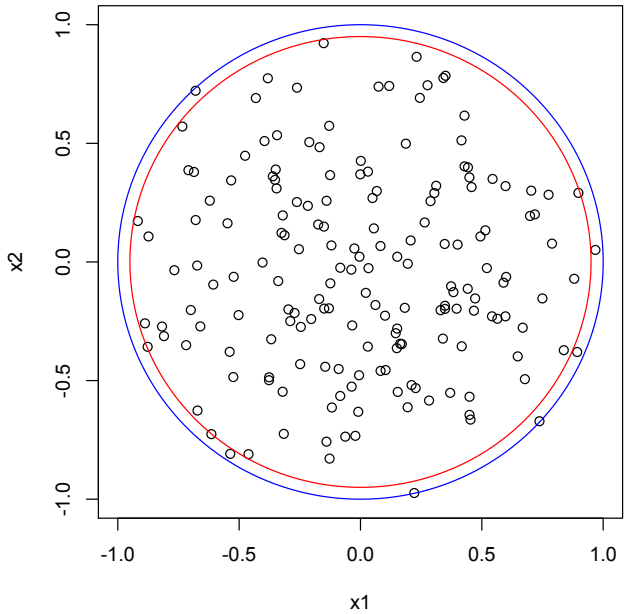
n	50	100	500	1000
p-valeur	0.7773	0.5485	0.08924	0.01364

et c'est évident mathématiquement...

p grand et le "Curse of dimensionality"

Lorsque l'on dispose de p variables/facteurs explicatifs, il n'est pas possible de visualiser les données globalement. On a tendance à généraliser en dimension p ce que l'on connait en dimension 2 ou 3.
Exemple: données distribuées de manière uniforme à l'intérieur du cercle bleu (centre 0 et rayon 1):





La probabilité d'observer une valeur dans la "couronne" (partie située entre le cercle bleu et le cercle rouge (de rayon 0.95) vaut 0.0975.

Et en p dimensions?

p	2	3	5	10	25	100	250	500
Probabilité	0.0975	0.143	0.226	0.401	0.723	0.994	0.999	1

En grande dimension, toutes les observations se trouvent, sous l'uniformité, dans la couronne...

Sources: On the curse of dimensionality, Ch. Giraud, 2021.

Conclusion

La statistique est une science et non un livre de recettes...