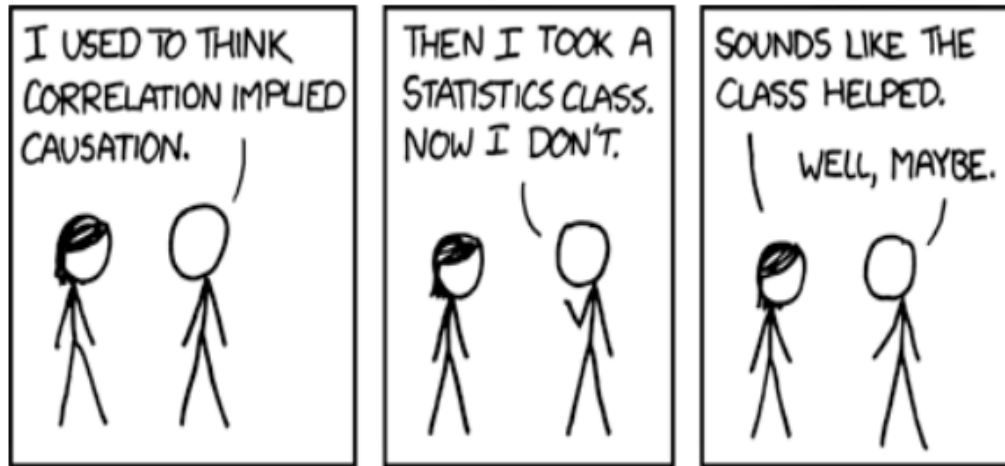


Statistisch bewijs – wat is dat eigenlijk?



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. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

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

It can be proven that most claimed research findings are false.

should be interpreted based only on p -values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. “Negative” research is also very useful. “Negative” is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2×2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of “true relationships” to “no relationships” among those tested in the field. R

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 be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α . Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2(8): e124.

Copyright: © 2005 John P.A. Ioannidis. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abbreviation: PPV, positive predictive value

John P.A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: joannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

Dit is onderzoek!

Handleiding voor kwantitatief en kwalitatief onderzoek

Ben Baarda

Tweede editie

**Tweede
geheel
herziene
editie**



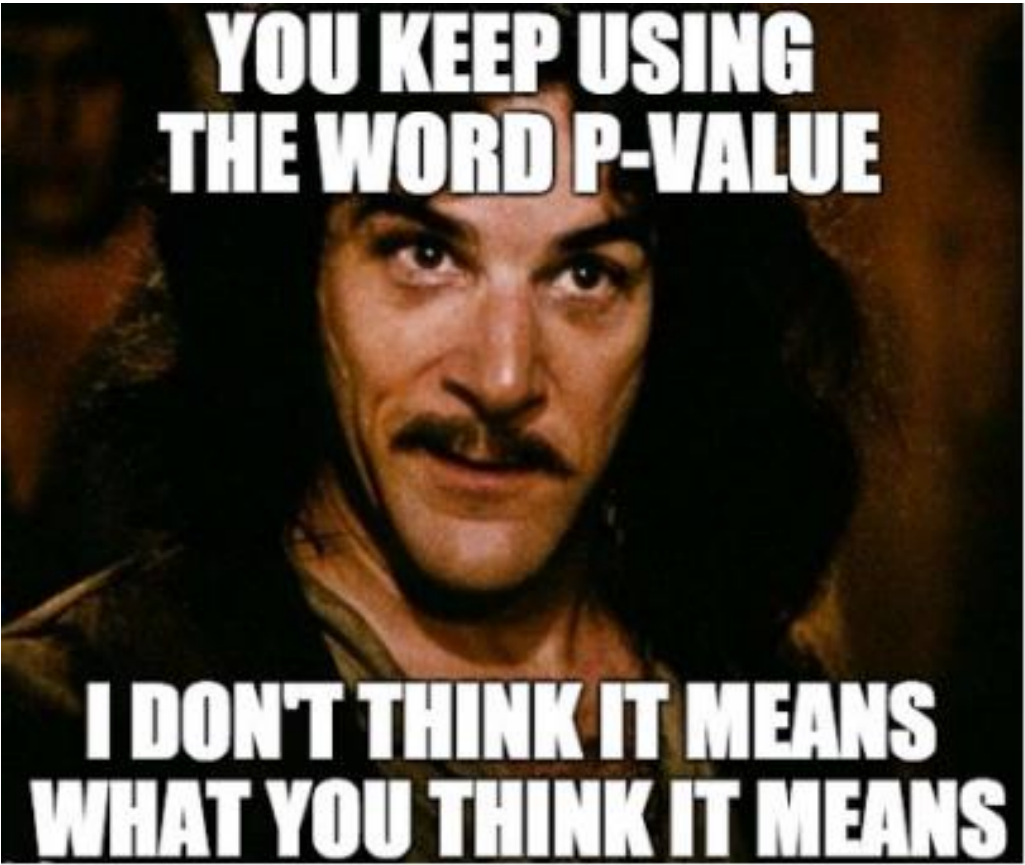
Noordhoff Uitgevers

Formuleer de hypothese die je wilt *ontkrachten*:

- Medicijn X werkt niet beter dan medicijn Y
- Verdachte S is niet schuldig aan het misdrijf
- Het Standaard model van de theoretische fysica is compleet: het Higgs deeltje bestaat niet
- Persoon Z kan de toekomst voorspellen
- Mondkapjes werken niet
- De zomers worden niet warmer
- Persoon U fraudeert niet met zijn belastingaangifte

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	]— HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	]— SIGNIFICANT
0.049	
0.050	]— OH CRAP. REDO CALCULATIONS.
0.051	]— ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	]— HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	]— HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥0.1	

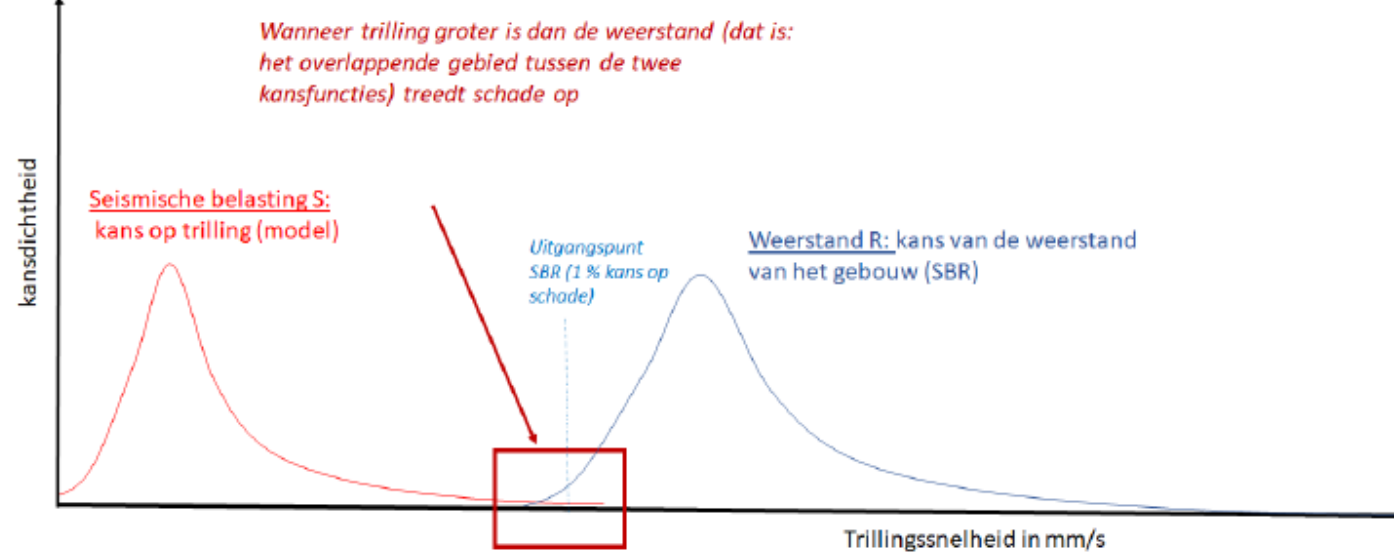
- P-waarde
- Significant
- Betrouwbaarheidsinterval
- Betrouwbaarheidsdrempel



**YOU KEEP USING
THE WORD P-VALUE**

**I DON'T THINK IT MEANS
WHAT YOU THINK IT MEANS**

- De juiste vraag is niet: wat is de kans op deze data onder deze hypothese?
- De juiste vraag is: nu ik deze data heb, wat is daarvoor de beste verklaring?
- Dit heeft grote consequenties: bewijs wordt relatief!



Figuur 5: Kansdichtheidsfunctie voor de weerstand van een gebouw, aangeduid met R (zie voetnoten 8 en 11 tot en met 13) en van de grootte van de topwaarde van de trillingssnelheid, aangeduid met S (volgens de modellen genoemd in voetnoten 1 tot en met 5).

De kans op schade kan worden bepaald uit de convolutie van de kansverdeling van de weerstand of kwetsbaarheid R en de kansdichtheid van de trillingsbelasting S. De kans op schade $P(R < S)$ volgt rekenkundig uit de convolutie-integraal:

$$P(R < S) = \int_0^{\infty} (1 - F_S(x)) \cdot f_R(x) dx$$

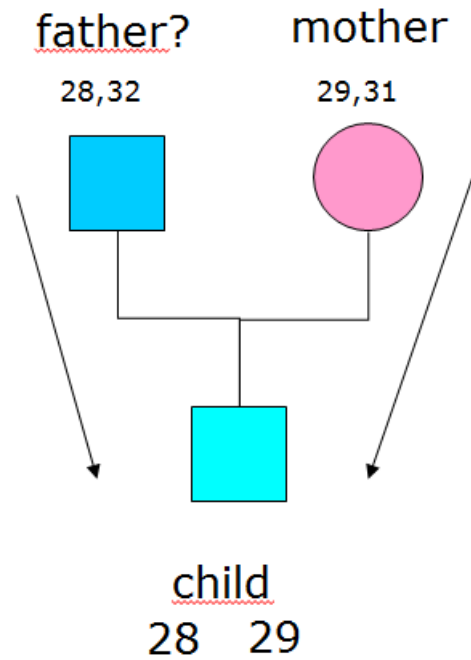
Waarin:

$F_S(x)$ is de kansverdeling van de belasting (trilling) S

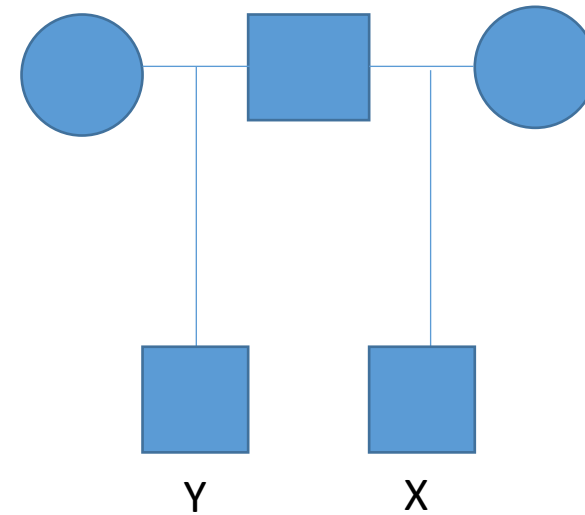
$f_R(x)$ is de kansdichtheid van de weerstand (kwetsbaarheid) van een gebouw R.

Twée forensische voorbeelden

Is dit de vader?



Is Y de halfbroer van X??



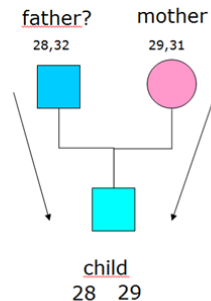
Hoeveel bewijs hebben we?



Het genotype van het kind is **100.000** keer waarschijnlijker als X de vader is dan wanneer een onbekend person de vader is

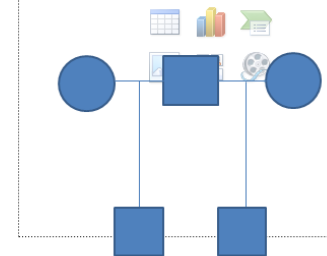
Het genotype van Y is **100.000** keer waarschijnlijker als hij de halfbroer van X is, dan wanneer hij ongerelateerd is aan X

Is he the father?



"Completely unacceptable. I don't know how any son of mine could... Oh. OK, I see you even failed the DNA test."

Is he a half-brother?



- Die factor 100.000 heet de likelihood ratio
- Deze vertelt me welke van de twee rivaliserende hypotheses de data beter verklaart, en hoeveel keer beter.
- Formeel: $P(\text{data} \mid \text{hypothese 1}) / P(\text{data} \mid \text{hypothese 2})$

Toekomst voorspellen...?



Daryl Bem

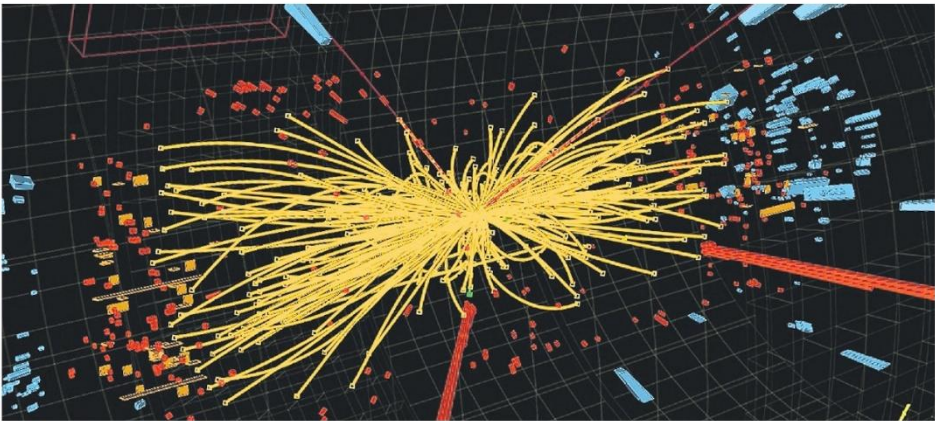


Aantal experimenten	Aantal keer succes	p-waarde voor $p=1/2$	Likelihood ratio $p=1/2$ versus $p \neq 1/2$
1528	828	0,02	1,64

Dat doen theoretisch fysici ook:

“vijf sigma”, komt overeen met een p-waarde van 0,0000000695

Aantal experimenten	Aantal keer kop		Likelihood ratio
10,000	5185		0,085
10,000,000,000	50,000,785,824		1,09



Grafische weergave van de sporen van een proton-protonbotsing in een van de deeltjesdetectoren van CERN, het deeltjesversnellerinstituut bij Genève. Foto AFP / CERN

Historische stap in het onderzoek naar de bouwstenen waaruit heelal is opgebouwd

Higgsdeeltje ‘vrijwel zeker’ ontdekt

Door BRUNO VAN WAYENBURG
AMSTERDAM. Na twee uur spanningsekkende praatjes komt CERN-directeur Rolf Heuer vanmorgen met de mededeling: „We hebben een ontdekking, de waarneming van een nieuw deeltje, met de eigenschappen van een Higgs-boson”.

Het Higgsdeeltje dus, het enige deeltje in het Standaardmodel van de natuurkunde waarvan het bestaan wel voorspeld was maar nog niet aangetoond.

Het is het deeltje dat andere deeltjes hun massa geeft.

Het is groot nieuws: de mededeling leidt tot een ontlading in de zaal bij het CERN, het deeltjesversnellerinstituut bij Genève. Het publiek van vooral natuurkundigen klappt en joelt.

Aankomende zaterdag zal de ontdekking ook de onderzoekers en pers die samengekomen zijn in Nijmegen, het instituut voor deeltjesfysica in Amsterdam.

Het lijkt er nu toch echt op dat ze, zij het virtueel, aanwezig zijn bij een historische aankondiging. Uit presentaties van Joe Incandela van de CMS-deeltjesdetector en van Gianotti van de ATLAS-detector, blijkt dat er maar een hele kleine kans op toeval is: minder dan 1 op de 3,5 miljoen.

„Er is absoluut een nieuw deeltje ontdekt, dat valt niet meer te ontkennen”, zegt ook Stan Benvise, projectleider van de Nederlandse inbreng in het ATLAS-experiment, waarbij 45 onderzoekers betrokken zijn.

Om het Higgsdeeltje aan te tonen, moet je nieuwe deeltjes maken. Dat is peperduur, maar eenvoudig. Je laat in een deeltjesversneller deeltjes met vrijwel de lichtsnelheid met elkaar botsen. Hoe harder de botsing, hoe meer energie er wordt omgezet in nieuwe deeltjes. Zoals het Higgs-boson, dat in de jaren zestig voorspeld werd door een theoretisch natuurkundigen. Het werd naar een van hen genoemd, Peter Higgs.

De Higgs is nodig om te verklaren hoe het komt dat alles massa heeft: doordat deeltjes worden afgeremd door het zogeheten Higgs-veld. Zo alomtegenwoordig als het uitge-smeerde Higgs is, zo ongraspbaar is het als deeltje.

Zo gauw het ontstaat uit de enorme energie die vrijkomt bij een botsing, zo snel valt het ook weer uit el-

kaar in verschillende elementaire deeltjes. Alleen die brokstokken zijn, meten na een botsing, goed te zien in de detectoren.

Maar veel vaker ontstaat er bij een botsing geen Higgs, maar een mix van al bekende deeltjes. Onderscheid maken tussen ‘Higgs- en niet-Higgs-botsingen’ is een kwestie van netjes meten en turven en zware statistiek.

Daarover gaat het allemaal, in de presentaties van Gianotti en Incandela.

In het deeltjesjagen gaat het om de sigma-waarde: een statistische maat voor de kans dat de gevonden botsingen, ook al lijken ze op een nieuw deeltje, toch een toevallige uitschieter zijn. De afspraak is dat je een deeltje pas mag claimen bij een sigma van 5: de kans dat het om een toevallige

uitschieter zou gaan is dan 1 op de 3,5 miljoen. Incandela eindigt met een tergende 4,9, net geen 5. Maar Gianotti komt na eindeloze details uit op 5,0.

„Nu moeten we onderzoeken of het ook de Higgs is”, zegt Benvise. Daar lijkt het wel op. „Al vervalt het wel iets vaker in twee fotonen dan je zou verwachten.”

Maar mocht het geen Higgs-boson zijn, of zelfs maar een licht afwijkend Higgs-boson, dan zou dat nog groter nieuws zijn. Want meer nog dan naar het Higgs-boson, snakken natuurkundigen naar ‘nieuwe natuurkunde’: metingen of deeltjes die eindelijk eens een keer niet overeenkomen met het Standaardmodel, is dus het devies. „Dit is pas het begin”, concludeert ook Gianotti.

Voor het onderwijs:

- Geen 'kookboeken'. Statistiek vereist wijsheid.
- Vanuit echte dossiers. Data is meestal niet zo mooi.
- Filosofie belangrijker dan formules.
- Hoeveel informatie kan data bevatten? Realisme.
- Beperkingen van statistiek zijn cruciaal.
- Dit vereist opleiding van de docenten.