Statistische Zonden

Een overzicht van twijfelachtige onderzoekspraktijken (en hoe deze te herkennen)

Casper Albers
Twitter: @caal

KWG Wintersymposium 2019

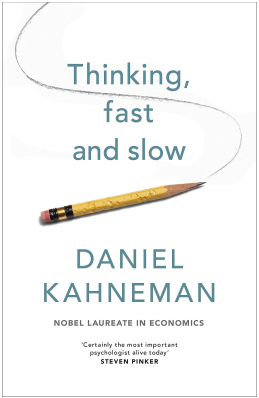
university of groningen

Introductie

Wat is statistiek?

Verklaringen zoeken voor verschillen tussen metingen.

Waarom is statistiek moeilijk?

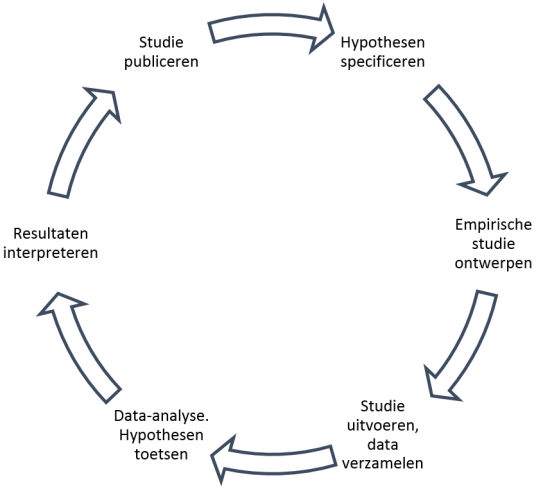


Daniel Kahneman

Thinking fast and slow (*ons feilbare denken*).

- Systeem 1 – Thinking Fast.
  Moeiteloos, intuïtief, heuristisch, onbewust
- Systeem 2 – Thinking Slow.
  Analytisch, kritisch denken.

De empirische wetenschappelijke cyclus



Statistische ‘zonden’

- Situaties waar de intuïtie je in de steek laat (paradoxen).
- Situaties door gebrek aan statistische kennis/vaardigheden.
- (Quasi-)opzettelijk dingen fout doen.

Intuïtieve fouten

Steekproef en populatie

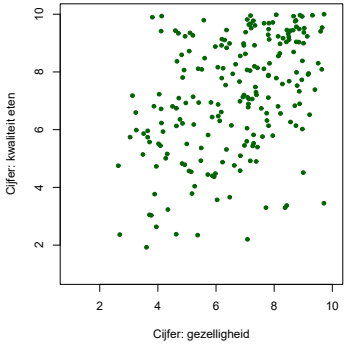
Je gebruikt een steekproef om iets over de populatie te zeggen.
Maar wat is de populatie? *Reichenbachs referentieklassenprobleem.*

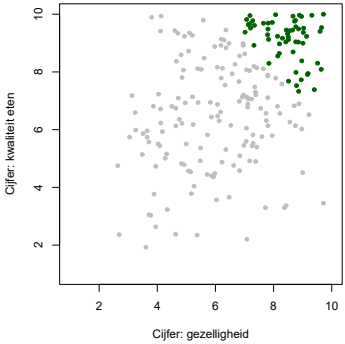
Twee bronnen van verstoring:

1. Steekproefvariatie
2. Systematische afwijking (representativiteit)

'n =veel' is zinloos als onbekend is met welk kansmechanisme n uit N gehaald is.

Steekproef en populatie: Berkson's paradox





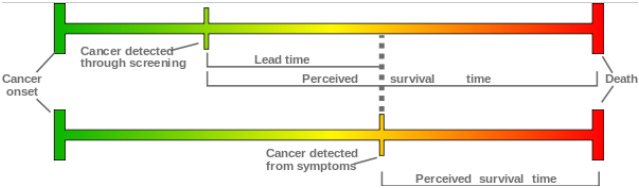
210 restaurants. Correlatie: 0.42.

58 restaurants met totaalscore boven 16. Correlatie: -0.33

Berkson's paradox: door te controleren op 'verkeerde' variabelen krijg je schijnverbanden.

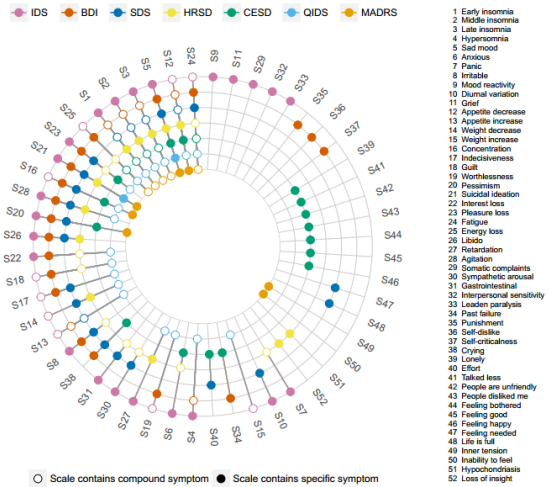
Houdt rekening met 'lead time bias'

While running for president of the United States the former New York mayor Rudy Giuliani announced in a 2007 campaign advertisement, "I had prostate cancer, 5, 6 years ago. My chance of surviving prostate cancer—and thank God, I was cured of it—in the United States? Eighty-two percent. My chance of surviving prostate cancer in England? Only 44 percent under socialized medicine."[1]



Bron: Gigerenzer (BMJ, 2013) / Wikipedia

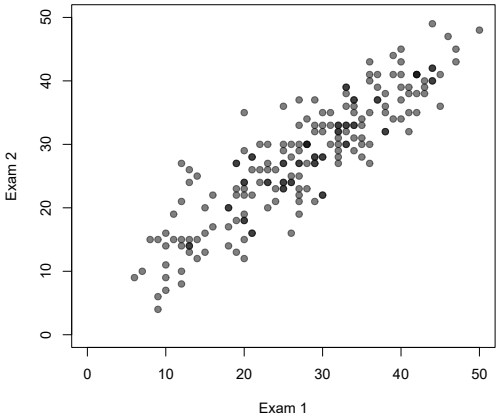
Weet wat je meet (en speel niet vals!)

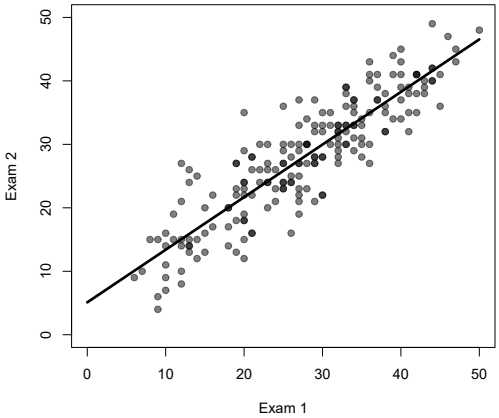


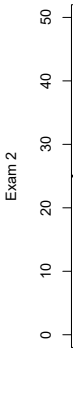
Bron: Fried, E. (2017)

Weet wat je meet (en speel niet vals!)

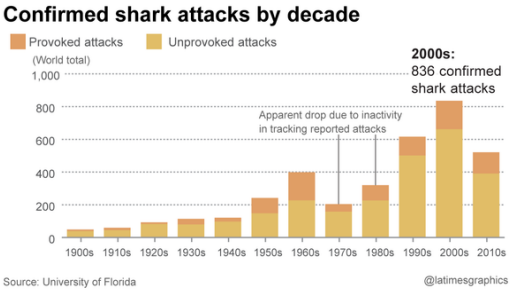
Regressie en subgroep-selectie







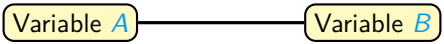
Absoluut vs. relatief risico



- In de jaren '00 waren haaien **4**× **dodelijker** dan in de jaren '50.
- Krantenkoppen: **HORRORHAAIEN**
- Ongeveer een half miljard dooien in de jaren '00.
- Risico op haaigerelateerde dood: $<$ 1 op 500 miljoen.

Correlatie & causatie

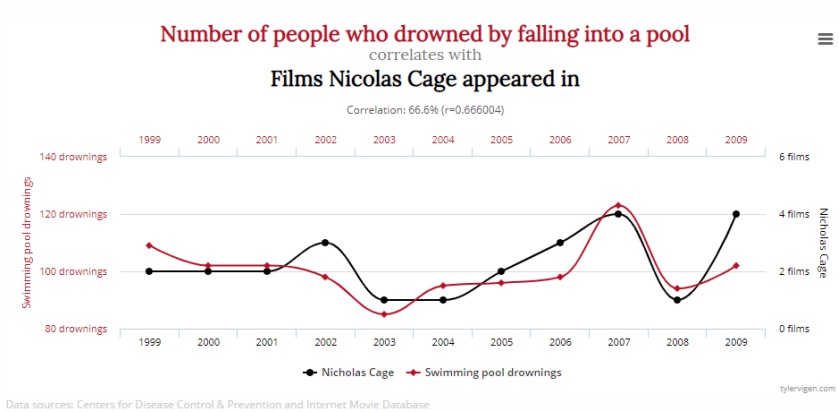
Er zijn verschillende manieren waarop twee variabelen kunnen samenhangen (d.w.z. correleren). Sommige zijn interessanter dan anderen.



Post Hoc Ergo Propter Hoc

- Correlatie impliceert niet causatie, maar
- Causatie impliceert wel correlatie.

Correlatie $\neq$ causatie



tylervigen.com

Voorbeeld: *hidden moderators*

Studie (Van der Lee & Ellemers, PNAS, 2015) in opdracht van NWO:

Geslacht	Aanvragen	Honoreringspercentage
Man	1635	17.7%
Vrouw	1188	14.9%

De statistische analyse gaf $p = .045$, nèt onder $\alpha = .05$.

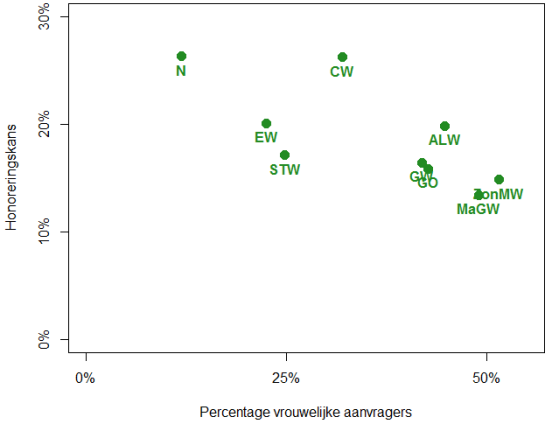
deVolkskrant

NWO discrimineert vrouwelijke wetenschappers

Naomi Ellemers met Koningin Maxima in 2010, toen de professor de KNAW Mortanprijs in ontvangst nam voor haar inzet voor gelijke kansen in de wetenschap.

Voorbeeld: *hidden moderators*

Post-peer anayse (Albers, PNAS, 2015): geen discriminatie maar een hidden moderator.

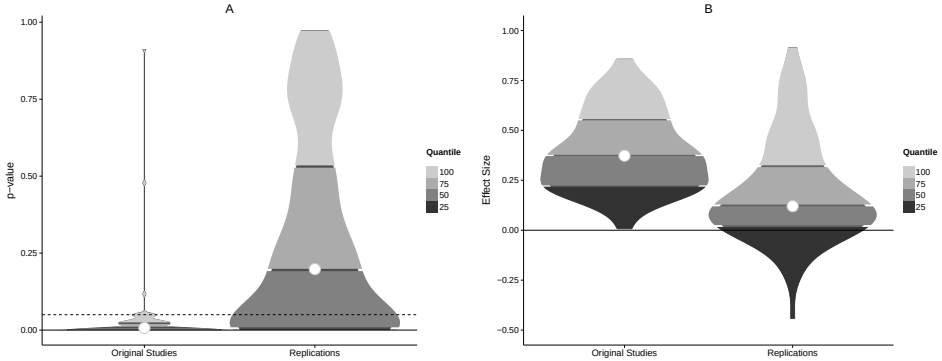


De reproducibiliteitscrisis

Het Reproducibiliteitsproject

- Open Science Collaboration (2015),
  Estimating the reproducibility of psychological science, Science.
- 100 empirische studies uit drie toonaangevende psychologische tijdschriften zijn geselecteerd.
- Studies zijn zo nauwkeurig mogelijk gerepliceerd.
- Doel: vergelijk effectgrootte en p-waarde van originele studie en replicatie.

Het Reproducibiliteitsproject



Sterk bewijs voor **file drawer effect**: alleen significante resultaten worden gepubliceerd.

Questionable Research Practices

Verklaringen voor samenhang

'Eerlijke' verklaringen:

1. A oorzaak van B
2. A veroorzaakt door B
3. A en B hebben gemeenschappelijke oorzaak
4. A beïnvloedt B indirect via C
5. Toeval.

Nog een belangrijke oorzaak:

Je niet aan de statistische spelregels houden.

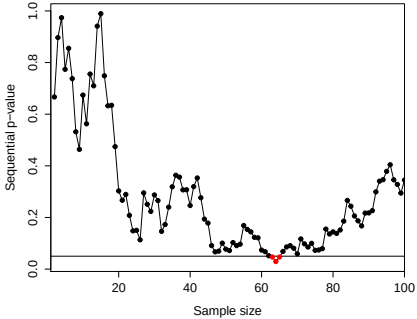
Questionable Research Practices

Een QRP is het (vaak onbewust) volgen van een foutieve statistische/methodologische procedure, met als gevolg dat de resultaten onbetrouwbaar en mogelijk vertekend zijn. Veelgemaakte QRPs:

- Niet alle variabelen van een studie opnoemen
- Bepalen of je nog aanvullende metingen doet nadat je al naar resultaten gekeken hebt
- HARKing
- Uitbijters of andere datapunten weghalen omdat je dan mooiere resultaten krijgt
- Een alternatieve statistische methode gebruiken omdat je dan mooiere resultaten krijgt
- (Als editor:) een paper niet accepteren omdat de p-waarde niet $< \alpha$ is.

Sequentieel toetsen

- De p-waarde is ontworpen voor het uitvoeren van **een enkele** toets
- Sequentieel toetsen: meervoudige vergelijkingen $\rightarrow$ kanskapitalisatie $\rightarrow$ aanpassingen nodig



HARKing

Idee achter NHST: als H_0 waar is kans van α (.05) op Type I fout.

Drie groepen vergelijken. Meerdere mogelijkheden:

- Groep 1 en 2 verschillen
- Groep 1 en 3 verschillen
- Groep 2 en 3 verschillen

Elke keer afzonderlijk toetsen? Opstapeling Type I fouten: 'Overall error rate' $\approx$ 14%.

Kanskapitalisatie. Twee mogelijkheden:

1. Hiervoor corrigeren.
   Bv. Bonferroni-correctie.
2. Stiekem doen alsof je alleen in de vergelijking 2–3 geïnteresseerd bent, omdat die toevallig significant is.
   HARKing, of *p*-hacking, een *questionable research practice*.

QRPs komen veel voor

John, Loewenstein, Prelec, 2012:

QRP	Percentage
Niet alle gemeten variabelen vermelden	63.4
Sequentieel toetsen (zonder correctie)	55.9
Meerdere studies doen. Alleen de 'succesvolle' vermelden	45.8
Uitbijter-selectie gebaseerd op p-waarden	38.2
Niet alle condities van een studie vermelden	27.7
Een p-waarde van bv. 0.054 'handig afronden' naar $< .05$	20.0
Data fabriceren of falsificeren	0.6

QRP komen niet alleen binnen de psychologie voor.

Oplossingen

Oplossingen

Trust but verify

Betere opleiding in statistisch denken;

Betere controlemechanismen;

Open Science: delen van *alle* onderzoeksmaterialen, incl. (geanonimiseerde) data.

Preregistratie: vooropgesteld en vooraf gedeeld compleet onderzoeks- en analyseplan.

Registered reports: artikel geschreven *voordat* de data verzameld is. Editor bepaald *vooraf* of het geaccepteerd wordt; daarna pas data-verzameling en analyse.

Perverse prikkels wegnemen.

Forensische statistiek

Gepubliceerd werk bevat incomplete gegevens.

Bijvoorbeeld wel gemiddelde, sd, n, maar niet elke losse meting.

Forensische statistiek gaat na of gepubliceerde gegevens kunnen kloppen.

Drie veelgebruikte 'tools':

1. De GRIM test
2. Sprite
3. Statcheck

De GRIM test

- 'granularity-related inconsistency of means (GRIM) test'; Brown & Heathers, 2016
- Veel data op een Likertschaal met integer-scores
- Gerapporteerd: steekproefgrootte en gemiddelde

> Participants ($N = 52$) responded to the manipulation check question, "To what extent did you believe our story about the dog having eaten our homework?" on a 1–7 Likert-type scale. Results showed that they found our story convincing ($M = 6.28$, $SD = 1.22$).

$$6.28 \times 52 = 326.56$$

$$326/52 \approx 6.269 \quad 327/52 \approx 6.288$$

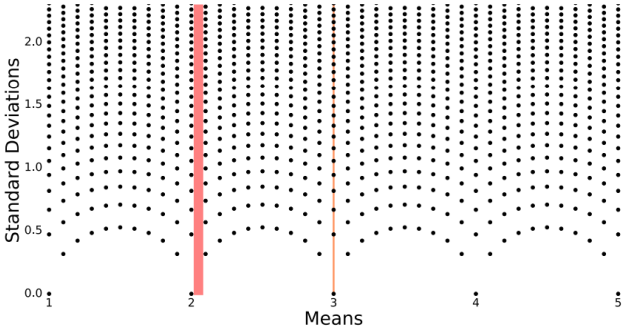
$\rightarrow$ onmogelijk resultaat

Steekproef van 71 recente publicaties: 36 hadden fouten.

Sprite

- ‘Sample Parameter Reconstruction via Iterative TEchniques (SPRITE)’
- Heathers, Anaya, Van der Zee & Brown, 2018
- Uitbreiding van het GRIM-idee, kijkt nu ook naar de SD $= \sqrt{\sum (x_i - \bar{x})^2}$.

$n = 10$ integer data:



Sprite & Brian Wansink

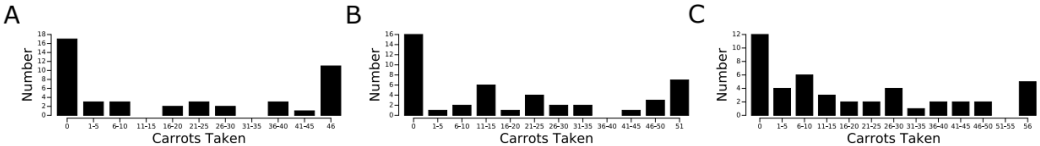


Figure 6. Possible data distributions. Possible distributions the carrot data can take for the control condition, taken (N=45, mean=19.4, SD=19.9).

Omstreden voedingsprof Brian Wansink stapt op na intern onderzoek

De reputatie van de ooit vooraanstaande Amerikaanse voedingspsycholoog Brian Wansink ligt aan diggelen. Donderdagavond maakte zijn werkgever, de Cornell University, bekend dat Wansink aan het eind van het academisch jaar zijn hoogleraarschap neerlegt. Een interne onderzoekscommissie van de universiteit verweet Wansink 'wetenschappelijk wangedrag', 'onjuiste rapportage van onderzoeksgegevens' en 'problematische statistische technieken'.

Sinds 2017: 17 papers retracted (een 2×), 15 corrections

Statcheck

Statistische 'spellingscontrole'; Nuijten & Epskamp, 2016.

Voorbeeld:

Steekproef, $n = 25$ uit $N(\mu, 1)$. $\bar{x} = .3687$.

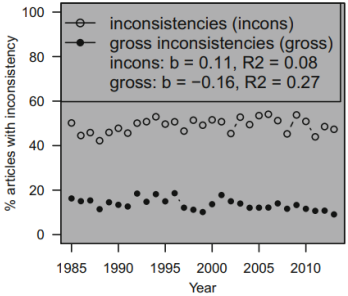
H_0: $\mu = 0$, H_A: $\mu > 0$.

$Z = \bar{x}/(1/\sqrt{25}) = 1.843 \rightarrow p = .033$

- Statcheck controleert of gerapporteerde toetsgrootheden, vrijheidsgraden en p-waarden overeenkomen.
- Checkt $t-$, $F-$, χ^2-, $Z-$verdeling en Pearson's correlaties, mits gerapporteerd in APA-style (default binnen sociale wetenschappen).

Statcheck

- 8 toonaangevende psychologische tijdschriften onderzocht (Nuijten et al., 2015)
- *Alle* artikelen 1985 – 2013 (automatisch) gescand. In 54.4% minstens één p-waarde
- 258,105 gerapporteerde p-waarden in 16,695 artikelen.
- In 49.6% van de artikelen ≥ 1 inconsistentie ($p_{\text{Gerapporteerd}} \neq p_{\text{Berekend}}$).
- In 12.9% van de artikelen ≥ 1 grove inconsistentie ($p_{\text{G.}} \neq p_{\text{B.}}$ en 1 vd 2 was significant).



Conclusie

Conclusie

Ja.

Vaak, maar niet altijd.