



Gebruik en misbruik van de P-waarde

Inhoud

1. Kritiek op de P -waarde
2. Interpretatie van de P -waarde
3. Gebruik en nut van de P -waarde
4. Misinterpretaties van de P -waarde
5. Ongewenste consequenties
6. Ongewenste prikkels
7. Advies

Inhoud

1. Kritiek op de P -waarde
2. Interpretatie van de P -waarde
3. Gebruik en nut van de P -waarde
4. Misinterpretaties van de P -waarde
5. Ongewenste consequenties
6. Ongewenste prikkels
7. Advies

STATISTICAL ERRORS

THE CULT

STATISTICAL

ACADEMIA AND CLINIC

Toward Evidence-Based Medical Statistics. 1: The *P* Value Fallacy

Steven N. Goodman, MD, PhD

Why Most Publications
Are False

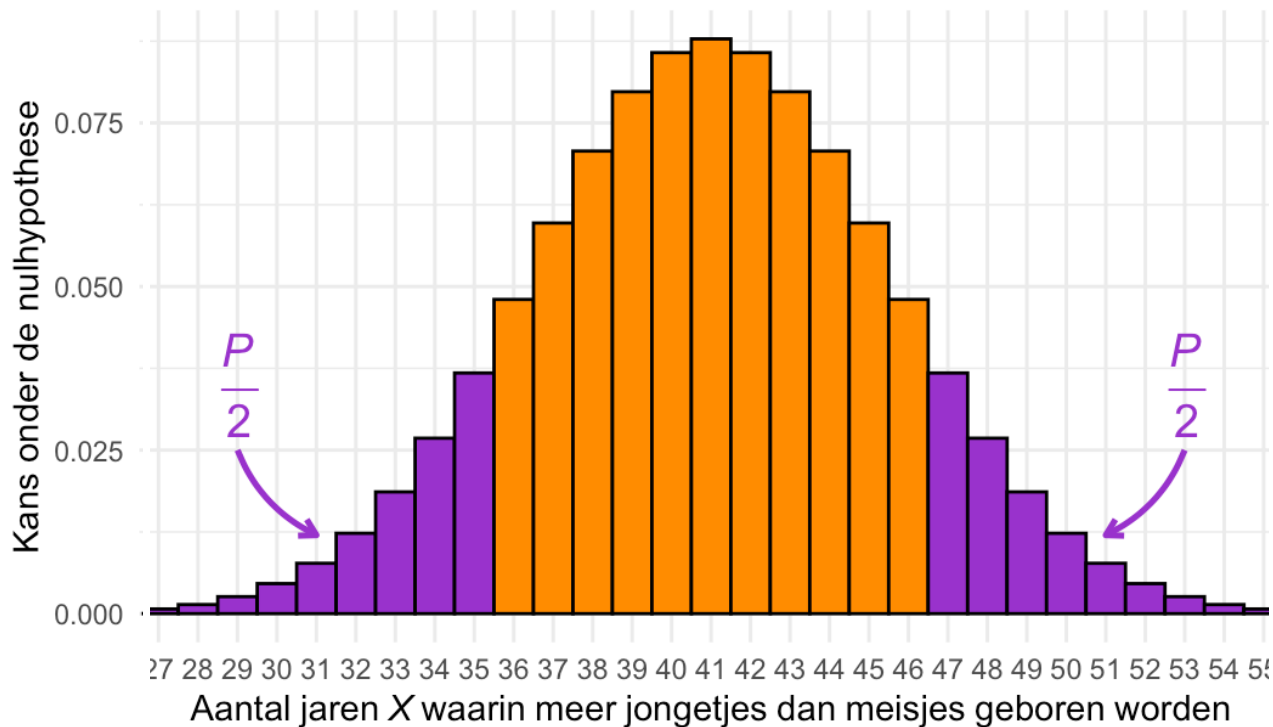
The Earth Is Round ($p < .05$)

Jacob Cohen

Inhoud

1. Kritiek op de P -waarde
2. Interpretatie van de P -waarde
3. Gebruik en nut van de P -waarde
4. Misinterpretaties van de P -waarde
5. Ongewenste consequenties
6. Ongewenste prikkels
7. Advies

Herhaling: de definitie van de P -waarde

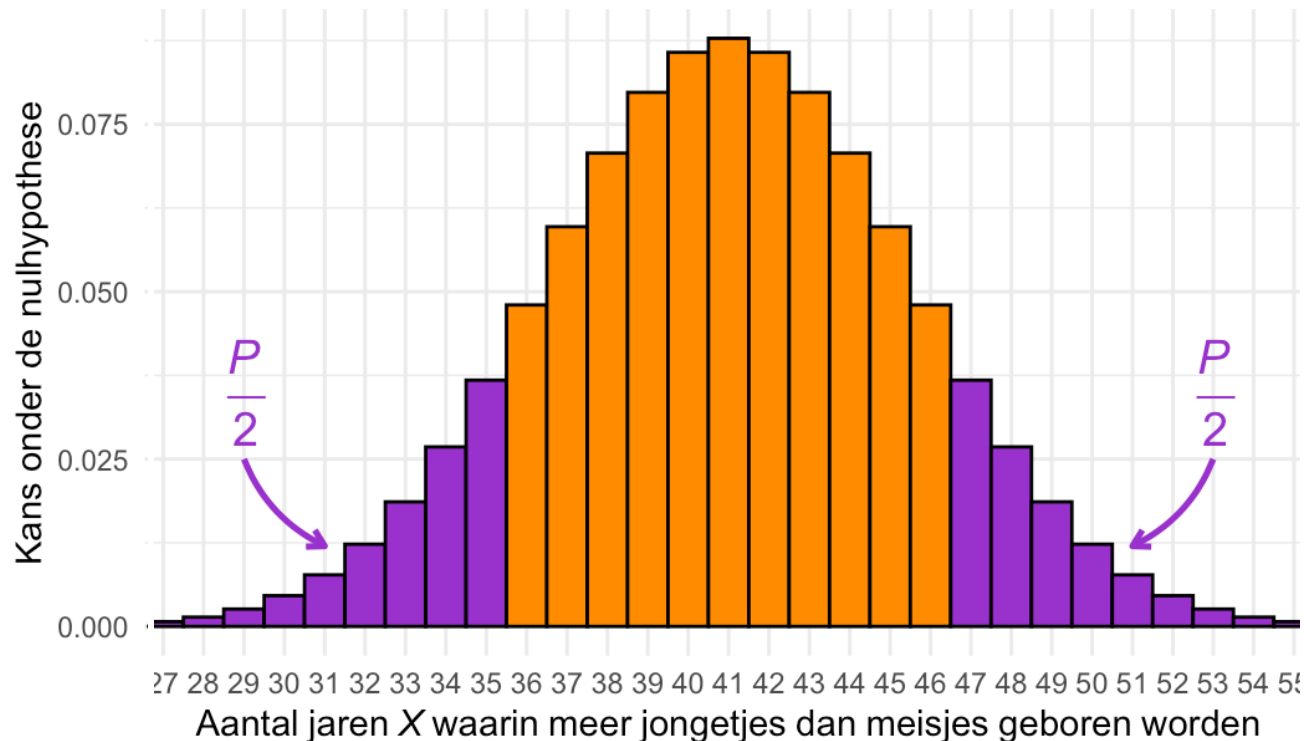


Neem aan dat

- ons statistische model correct is, en
- H_0 waar is,

wat is dan de kans op een toetsingsgrootte die minstens zo extreem is als de waarde die we gevonden hebben?

Interpretatie van de P -waarde

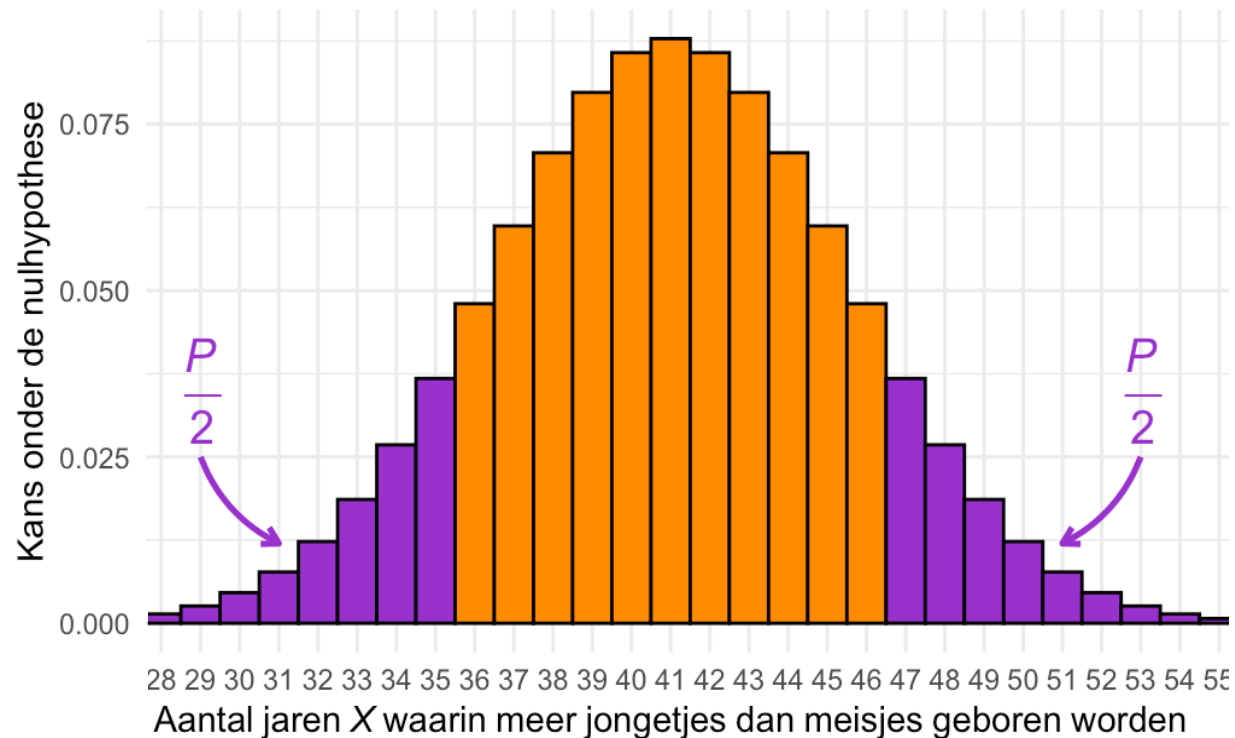


- Hoe opmerkelijk/extreem is de toetsingsgrootte onder H_0 ?
- Hoe vaak zou ik onder H_0 zo'n extreme toetsingsgrootte krijgen als ik hetzelfde onderzoek onafhankelijk meerdere keren zou herhalen?

Inhoud

1. Kritiek op de P -waarde
2. Interpretatie van de P -waarde
3. Gebruik en nut van de P -waarde
4. Misinterpretaties van de P -waarde
5. Ongewenste consequenties
6. Ongewenste prikkels
7. Advies

De P -waarde als mate van bewijs tegen H_0



Grote P -waarde:

- Toetsingsgrootte is niet opmerkelijk onder H_0 .
- Geen bewijs tegen H_0 .

Kleine P -waarde:

- Toetsingsgrootte opmerkelijk onder H_0 .
- Bewijs tegen H_0 .

De P -waarde als input voor besluitprocedures

Procedure:

- Kies α , meestal 0,05
- Verwerp H_0 als $P \leq \alpha$
- Kans om een type-1-fout te maken kleiner dan α .

Welke besluiten worden op deze basis genomen?

Binnen de wetenschap:

- Moet ik H_0 verwerpen en H_A aannemen?
- Is dit bewijs sterk genoeg om het te publiceren?
- Wetenschappelijke tijdschriften: Accepteren voor publicatie?

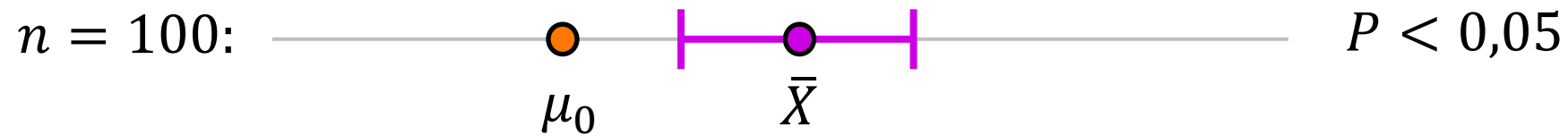
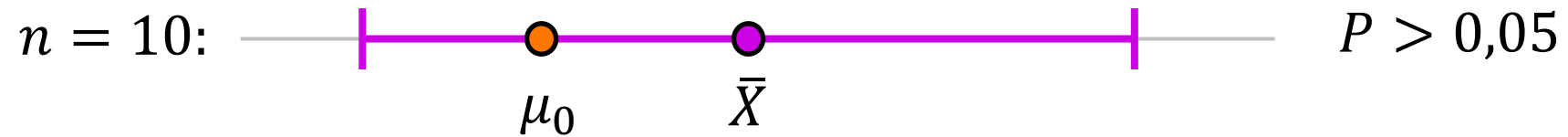
Buiten de wetenschap:

- Bijv.: Patient behandelen of niet?

Inhoud

1. Kritiek op de P -waarde
2. Interpretatie van de P -waarde
3. Gebruik en nut van de P -waarde
4. Misinterpretaties van de P -waarde
5. Ongewenste consequenties
6. Ongewenste prikkels
7. Advies

Een grote P -waarde zegt niet dat H_0 waar is

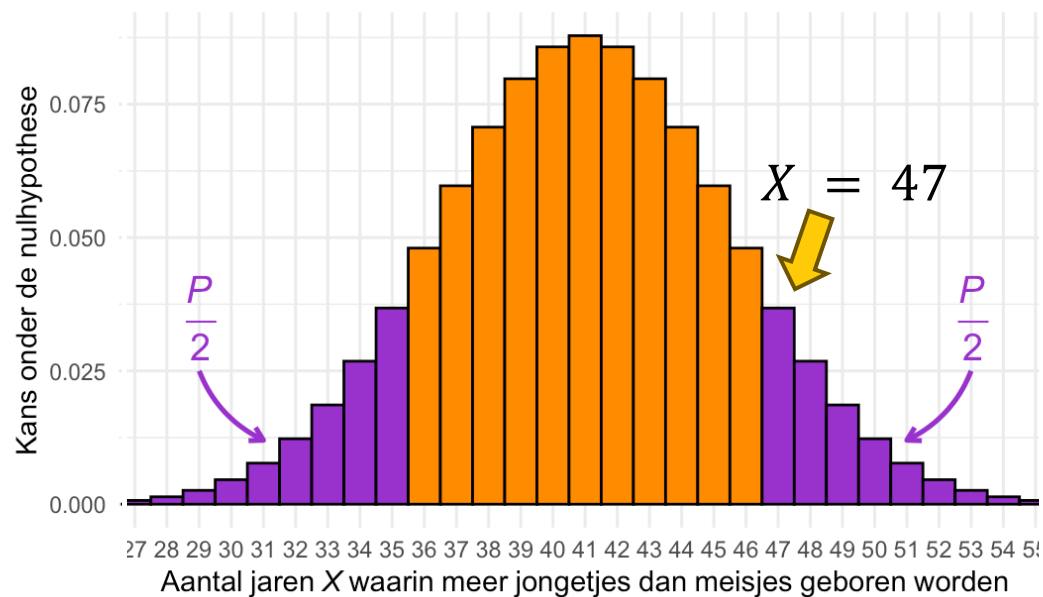


Afwezigheid van bewijs
is niet hetzelfde
als bewijs van afwezigheid.

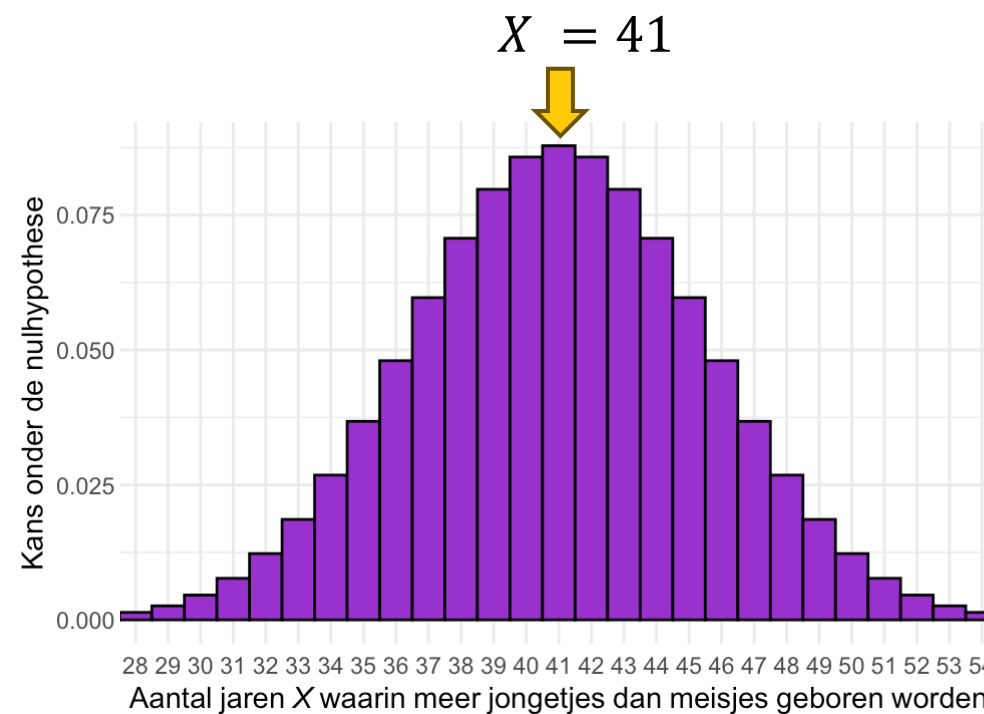
De P -waarde is niet de kans dat H_0 waar is

De kans op toetsingsgrootte (of extremer) onder H_0 is niet de kans dat H_0 waar is gegeven de toetsingsgrootte.

$$H_0: p_J = 0,5$$



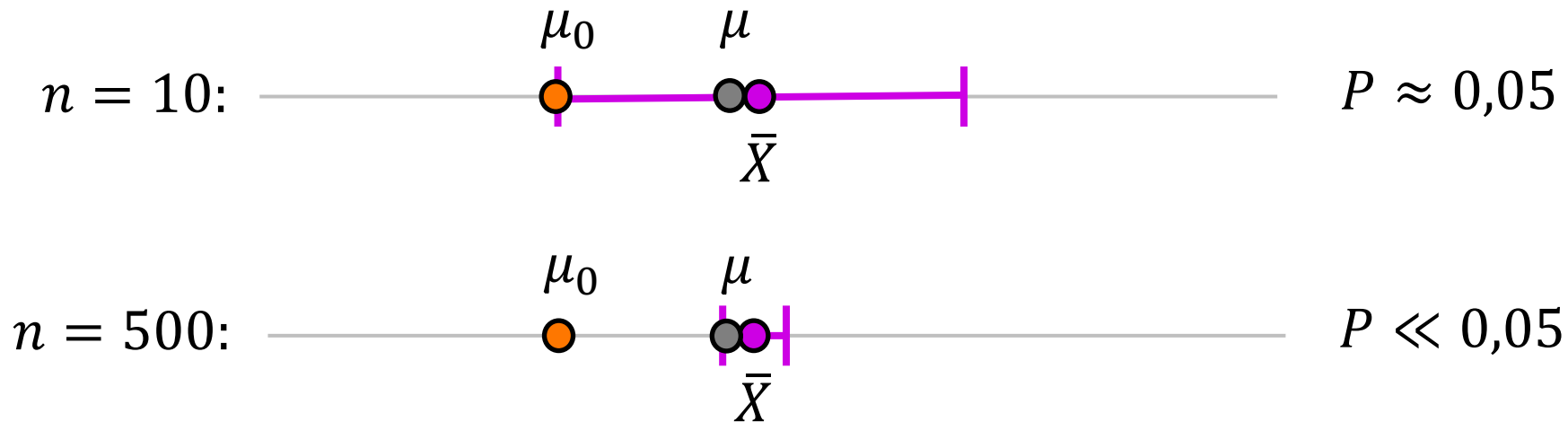
$$H_0: p_J = 0,5$$



De P -waarde is geen maat voor de “effectgrootte”

Voorbeeld hieronder:

Zelfde afwijking tussen μ_0 en μ , totaal andere P -waarde.



De P -waarde is geen maat voor de relevantie

Voorbeeld (geleend van Nuzzo et al (2014)):

“Last year, (...) a study of more than 19000 people showed that those who meet their spouses online are (...) more likely to have high marital satisfaction ($p < 0.001$) than those who meet offline.”

Gemiddelde *marital satisfaction* op een schaal van 1 tot 7:

Online daters: 5.64

Niet online: 5.48

Highly significant! Maar relevant?

Een kleine P -waarde kan ook het gevolg zijn van een ongeschikt statistisch model



- Lucia de Berk: verpleegkundige, in 2003 veroordeeld voor moord op 4 patienten en poging tot moord op 3.
- 7 jaar later alsnog vrijgesproken.
- Statistisch bewijs:
 P -waarde van 1 op 342 miljoen.
- Maar is het model (hypergeometrisch) wel redelijk?
Kans op overlijden gelijk voor alle shifts, verpleegkundigen, weekend/weekdagen, seizoenen, jaren?
- Model met variatie tussen shifts:

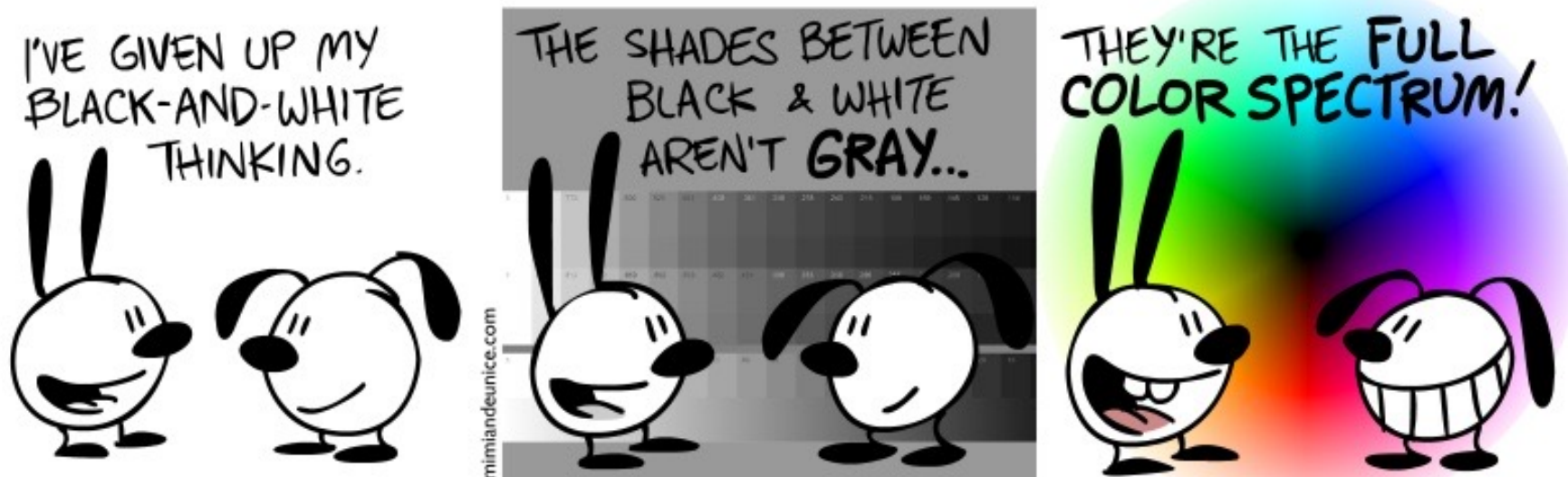
$$P = 0,14$$

Inhoud

1. Kritiek op de P -waarde
2. Interpretatie van de P -waarde
3. Gebruik en nut van de P -waarde
4. Misinterpretaties van de P -waarde
5. Ongewenste consequenties
6. Ongewenste prikkels
7. Advies

De drempelwaarde van $\alpha = 0,05$ werkt zwart-wit-denken in de hand

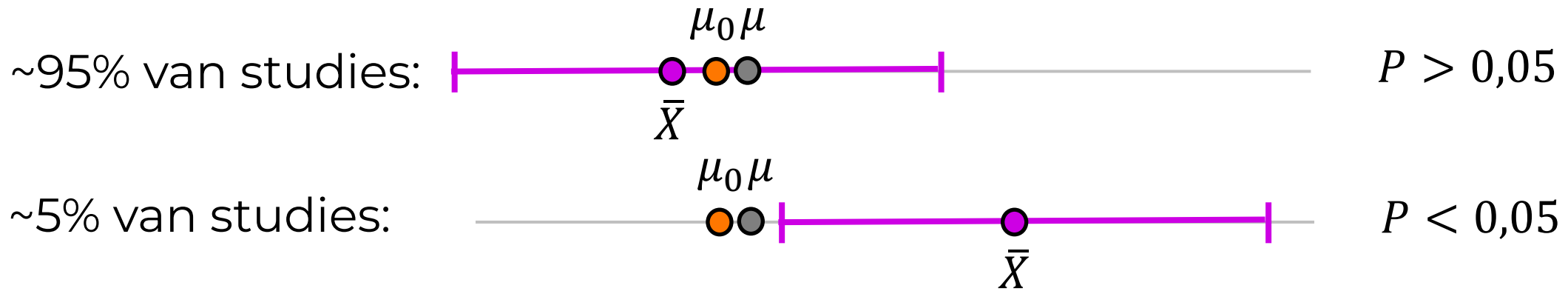
- Drempel is arbitrair: geen substantieel verschil tussen 0,04 en 0,06.
- Bovendien: $P = 0,05$ is maar matig bewijs; replicatie nodig!
- Daardoor krijg je veel fout-positieven.
- Inderdaad: Veel gepubliceerde resultaten worden bij replicatie niet weer gevonden.



<https://mimiandeunice.com>

Why Most Published Research Findings Are False

John P. A. Ioannidis (2005)



“Replication crisis”:

- Stel, er worden 1000 kleine onderzoeken uitgevoerd.
Te klein om eventuele effecten betrouwbaar te ontdekken.
- Dan zal alsnog 5% van de studies een significant effect rapporteren.
- Dat levert 50 publicaties op met effecten die bij replicatie waarschijnlijk *niet* weer gevonden worden.

Selecteren op P -waarde leidt tot publication bias

Voorbeeld: studie van Turner et al (2008).

- Onderzocht 74 klinische studies over antidepressiva, geregistreerd bij de Food and Drug Administration (FDA).
- Volgens FDA 51% van de studies positieve resultaten.
- Van de 38 studies met positieve resultaten werden er 37 gepubliceerd.
- Negatieve studies werden:
 - Niet gepubliceerd (22 van 36 gevallen), of
 - Gepubliceerd met een positieve draai.
- Gevolg: effectiviteit van antidepressiva sterk overschat.

Data dredging verergert het probleem

Data dredging:

In een dataset exploratief allerlei toetsen uitvoeren om te zien of je een positief resultaat kunt vinden.

Gevolg:

Grote kans op vals-positieven.

Veelgebruikte oplossing:

P -waarden verhogen:

$$P_{\text{gecorrigeerd}} = v P$$

aantal toetsen v .

“Bonferroni-correctie”

Chocolate hoax.

Besluitprocedure suggereert dat de overweging of H_0 verworpen moet worden genomen kan worden *puur op basis van de P -waarde van één experiment.*

P -waardes nemen niet mee:

- Voorkennis?
- Hypothese plausibel op biologische gronden?
- Eerder onderzoek?
- Coherentie met bestaande theorie?

Inhoud

1. Kritiek op de P -waarde
2. Interpretatie van de P -waarde
3. Gebruik en nut van de P -waarde
4. Misinterpretaties van de P -waarde
5. Ongewenste consequenties
6. Ongewenste prikkels
7. Advies

P-hacking

Probleem: Druk om te significante resultaten te publiceren.

Neiging:

Je analyse (of zelfs de gegevens) aanpassen met als doel een *P*-waarde significant te krijgen.

- Verschillende toetsen proberen
bijv.: t-toets voor 2 steekproeven, Welch t-toets, Mann-Whitney, ...
- Selectief uitbijters verwijderen.

Gevolg: Misleiding, wetenschapsfraude

<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	

Inhoud

1. Kritiek op de P -waarde
2. Interpretatie van de P -waarde
3. Gebruik en nut van de P -waarde
4. Misinterpretaties van de P -waarde
5. Ongewenste consequenties
6. Ongewenste prikkels
7. Advies

Advies!

De P -waarde is een nuttige maat van het bewijs tegen H_0 , verschaft door een specifieke dataset.

Maar: Baseer je eindoordeel nooit alleen op een P -waarde.

Een zorgvuldige afweging kan niet worden vervangen door één getal.

- Bereken ook **betrouwbaarheidsintervallen** van het “effect”. Die geven je een schatting mét de onzekerheid.
- *Vraag je altijd af:*
Is het aangetoonde effect **relevant**?
- *Vraag je altijd af:*
Hoe **plausibel** was H_0 bij voorbaat?
- *Vraag je altijd af:*
Wat **weet** ik nog meer over dit onderwerp?
Welke andere **informatie**, **studies**, en theorieën zijn er nog meer?
- Denk goed na of het **statistische model** van toepassing is.
(Denk aan Lucia.)

