Paranormale Statistiek

CWI Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam
Mathematisch Instituut Universiteit Leiden

Paranormale Statistiek

CWI Peter Grünwald

Dia 35-37 zijn, met toestemming, overgenomen van het onvolprezen **xkcd.org** **Dank!**

Dia 5-10, 31 en 61 zijn, met toestemming, deels overgenomen van een voordracht van **E.J. Wagenmakers.** **Dank!**

P-waardes... deugen niet!

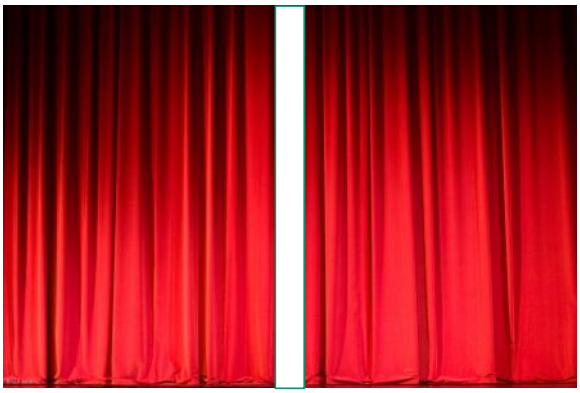
CWI Peter Grünwald

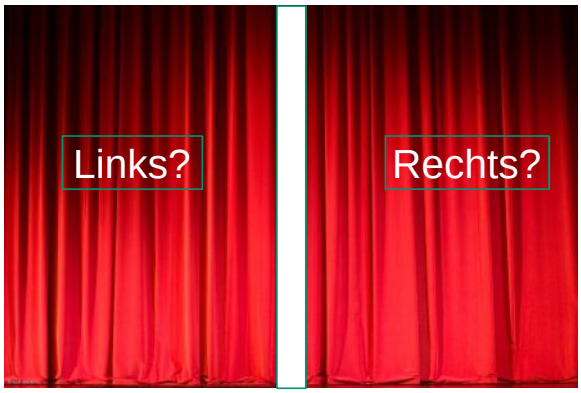
P-waardes... deugen niet!

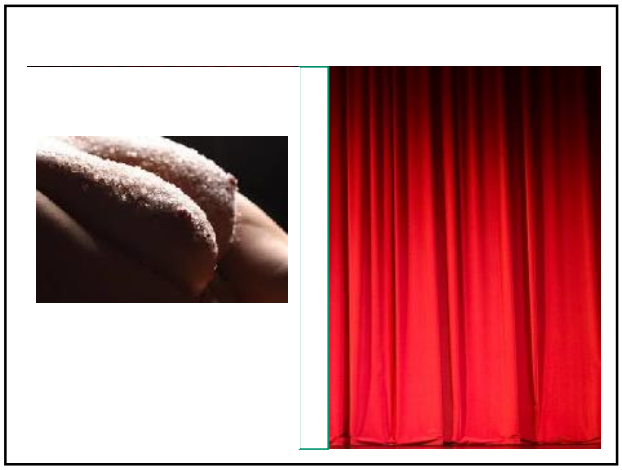
...en enige andere aspecten van

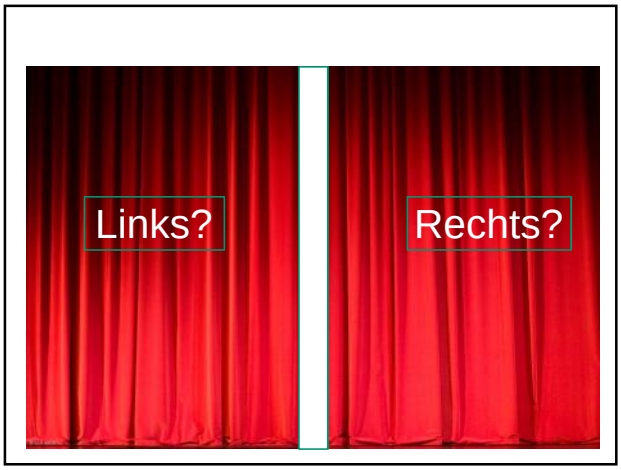
nulhypothese significantietoetsen

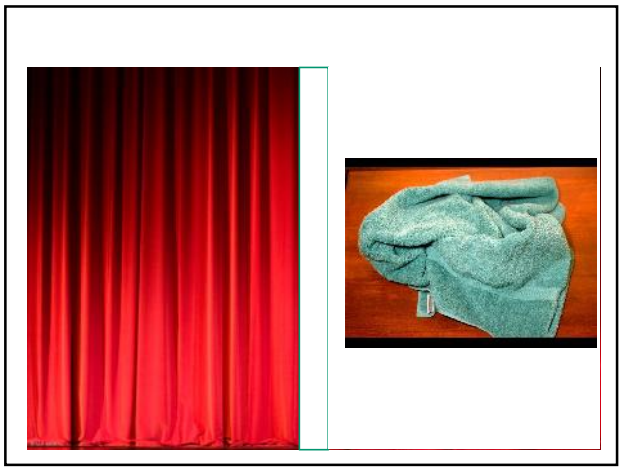
(de gangbare methode voor hypothesetoetsen in geneeskunde, psychologie, biologie...) deugen trouwens ook niet!

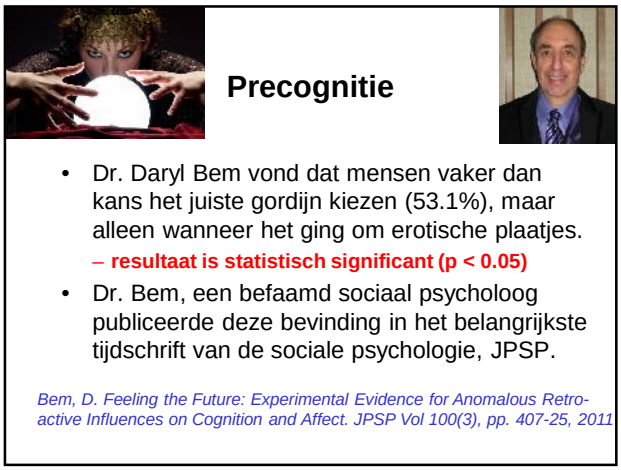












Precognitie

- Dr. Daryl Bem vond dat mensen vaker dan kans het juiste gordijn kiezen (53.1%), maar alleen wanneer het ging om erotische plaatjes.
  - **resultaat is statistisch significant ($p < 0.05$)**
- Dr. Bem, een befaamd sociaal psycholoog publiceerde deze bevinding in het belangrijkste tijdschrift van de sociale psychologie, JPSP.

Bem, D. Feeling the Future: Experimental Evidence for Anomalous Retro-active Influences on Cognition and Affect. JPSP Vol 100(3), pp. 407-25, 2011

Commotie Alom!

- Bem's onderzoek haalt de New York Times, Oprah, etc., maar wordt natuurlijk ook van alle kanten bekritiseerd
- Belangrijkste kritiek komt van de groep van 'onze eigen' Prof. Dr. Eric-Jan Wagenmakers* (UvA), in het artikel

Wagenmakers et al. *Why Psychologists must change the way they analyze their data – the case of Psi.* Comment on Bem (2011). *JPSP 100*, 2011

Menu

Er is het een en ander mis met gangbare praktijk van nulhypothesetoetsen / **p-waardes**:

1. publicatiebias
2. interpretatiemoeilijkheden
3. zeer beperkte toepasbaarheid
   (zodat veel toepassingen eigenlijk incorrect zijn)

Menu

Er is het een en ander mis met gangbare praktijk van nulhypothesetoetsen / **p-waardes**:

1. publicatiebias
2. interpretatiemoeilijkheden
3. zeer beperkte toepasbaarheid
   (zodat veel toepassingen eigenlijk incorrect zijn)

Bayesiaanse methode

– voorkomt probleem 2&3 geheel, 1 deels...maar heeft andere problemen

Test Martingaal methode

–'almost the best of both worlds'

Menu

Er is het een en ander mis met gangbare praktijk van nulhypothesetoetsen / **p-waardes**:

1. publicatiebias
2. interpretatiemoeilijkheden
3. zeer beperkte toepasbaarheid
   (zodat veel toepassingen eigenlijk incorrect zijn)

Bayesiaanse methode

– voorkomt probleem 2&3 geheel, 1 deels...maar heeft andere problemen

Test Martingaal methode ← **Leuke Wiskunde!**

–'almost the best of both worlds'

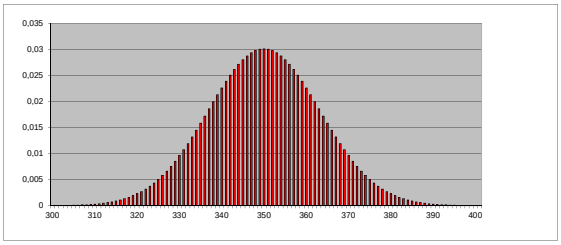
Hoe werkt nulhypothese toetsen?

- Stel we laten 700 mensen raden of het plaatje links of rechts zit
- Nulhypothese H0 (de 'status quo') gerepresenteerd als **kansverdeling** over *Test Statistic T*
- *Hier: T* = #mensen dat goed kiest
- Volgens H0 is $T \sim \mathrm{Bin}(0.5, 700)$verdeeld
  - i.e. verdeling van het aantal keren kop in 700 onafhankelijke worpen met een eerlijke munt

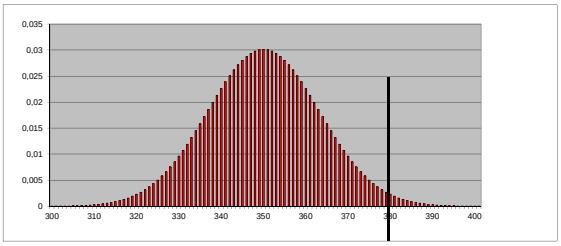
Hoe werkt nulhypothese toetsen?

- Volgens H0 is $T \sim \text{Bin}(0.5, 700)$verdeeld
- Alternatieve Hypothese H1: *munt niet eerlijk*
- Volgens H1 is $T \sim \text{Bin}(p, 700)$ voor $p > 0.5$
- We identificeren H0 dus met een enkele, en H1 met een verzameling kansverdelingen

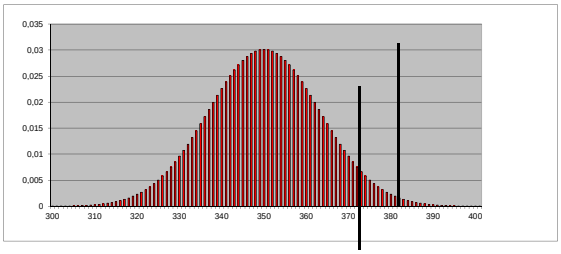
Verdeling van T onder H0



- We doen nu het experiment en zien dat *T=380*. **De p-waarde is de kans dat we deze *of een nog extremere* waarde zouden krijgen**,
  ≈ de oppervlakte onder de grafiek rechts van de lijn. We vinden, voor *T = 380,* dat *p = 0.02*



- We spreken van te voren een ***significance level*** ~ af en noemen het resultaat 'significant' als $p \leq \alpha$



Percentage nodig voor p < 0.05

n	afwijking	%
600	22	0,54
500	19	0,54
400	17	0,54
300	15	0,55
200	13	0,57
150	11	0,57
100	59	0,59

Menu

Er is het een en ander mis met gangbare praktijk van nulhypothesetoetsen / **p-waardes**:

1. **Interpretatiemoeilijkheden I**
2. Publicatiebias
3. Interpretatiemoeilijkheden II
4. zeer beperkte toepasbaarheid
   (zodat veel toepassingen eigenlijk incorrect zijn)

What Do Doctors Know About Statistics? Wulff et al.,*1987*

What Do Doctors Know About Statistics? Wulff et al.,*1987*

Een dubbelblinde gerandomiseerde toets van een nieuw geneesmiddel leidt tot de conclusie dat het **'significant beter' is dan de placebo (p < 0.05).**

What Do Doctors Know About Statistics? Wulff et al.,*1987*

Een dubbelblinde gerandomiseerde toets van een nieuw geneesmiddel leidt tot de conclusie dat het **'significant beter' is dan de placebo (p < 0.05).**
Welke uitspraak klopt het best?

1. Het is wetenschappelijk bewezen dat het geneesmiddel beter werkt dan de placebo
2. Als het geneesmiddel niet werkt, is er minder dan 5% kans op zo'n soort resultaat
3. Er is minder dan 5% kans dat het geneesmiddel niet beter werkt dan de placebo
4. Geen idee

Math teachers! *What Do Doctors Know About Statistics?*

Een dubbelblinde gerandomiseerde toets van een nieuw geneesmiddel leidt tot de conclusie dat het **'significant beter' is dan de placebo (p < 0.05).**
Welke uitspraak klopt het best?

1. Het is wetenschappelijk bewezen dat het geneesmiddel beter werkt dan de placebo
2. Als het geneesmiddel niet werkt, is er minder dan 5% kans op zo'n soort resultaat
3. Er is minder dan 5% kans dat het geneesmiddel niet beter werkt dan de placebo
4. Geen idee

Math teachers! *What Do Doctors Know About Statistics?*

Een dubbelblinde gerandomiseerde toets van een nieuw geneesmiddel leidt tot de conclusie dat het 'significant beter' is dan de placebo (p < 0.05).
Welke uitspraak klopt het best?

1. Het is wetenschappelijk bewezen dat het geneesmiddel beter is dan de placebo
2. **Als het geneesmiddel niet werkt, is er minder dan 5% kans op zo'n soort resultaat**
3. Er is minder dan 5% kans dat het geneesmiddel niet beter werkt dan de placebo
4. Geen idee

Prosecutor's Fallacy: standaard-verkeerde interpretatie

Een dubbelblinde gerandomiseerde toets van een nieuw geneesmiddel leidt tot de conclusie dat het 'significant beter' is dan de placebo (p < 0.05).
Welke uitspraak klopt het best?

"Er is minder dan 5% kans dat het geneesmiddel niet beter werkt dan de placebo"

Dit is de zgn. Prosecutor's Fallacy! Vrijwel alle mensen, ook wiskundigen, hebben de neiging zo te rederenen, maar het klopt niet!

Prosecutor's Fallacy

- p-waarde zegt iets over de geobserveerde data (of meer extreme gevallen) **gegeven** dat de nul hypothese waar is.
- De p-waarde zegt dus niet direct iets over de kans dat de *nul hypothese* waar is!

  Pr(D | H0) is niet gelijk aan Pr(H0 | D)

  Deze twee kansen kunnen enorm verschillen!

Prosecutor's Fallacy

- Pr(D | H0) heel anders dan Pr(H0 | D)
  Mocht u twijfelen:

Wat is Pr(lengte > 1.90 | speler in de NBA)?
Wat is Pr(speler in de NBA | lengte > 1.90)?

- Op zich wordt dit studenten wel vaak verteld, maar het blijft een bron van ellende…

Prosecutor's Fallacy

- Pr(D | H0) heel anders dan Pr(H0 | D)
- Deskundige in zaak Lucia de Berk: "de kans dat een verpleegkundige bij toeval bij zoveel of meer incidenten aanwezig is, is 1 op 342 miljoen (een p-waarde!)
- Rechter: deskundige is gevraagd te bepalen *wat de kans op toeval is*

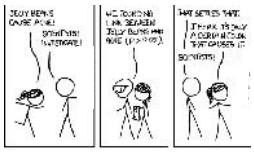
Juiste Interpretatie (Neyman-Pearson, 1937)

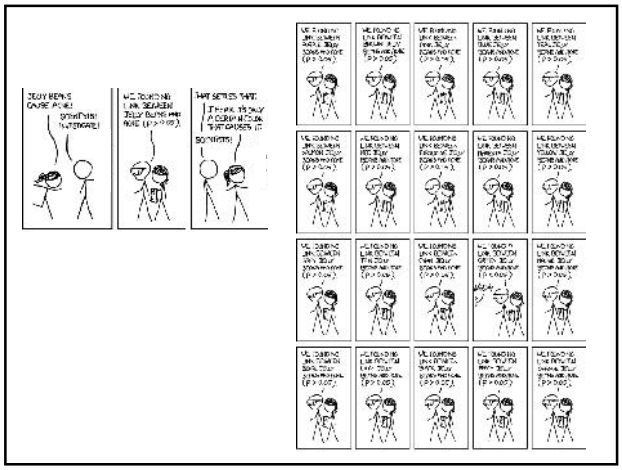
- We zetten *van te voren* een significantielevel (bijv. $\alpha = 0.05$) vast.
- Wanneer we nu waarnemen $p \leq \alpha$ zeggen we 'we verwerpen H0'. Anders 'accepteren we H0'.
- Wanneer we nu **herhaaldelijk** hypothesetoetsen (over verschillende onderwerpen) uitvoeren **zullen we gemiddeld genomen in hoogstens een fractie ~ van alle keren H0 verwerpen terwijl hij waar is**

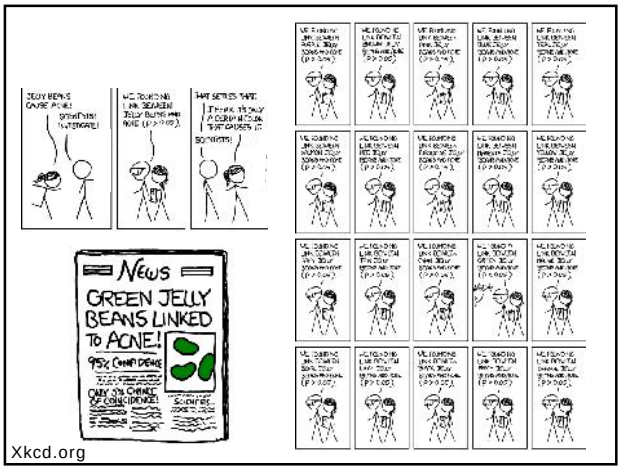
Menu

Er is het een en ander mis met gangbare praktijk van nulhypothesetoetsen / **p-waardes**:

1. Interpretatiemoeilijkheden I
2. **Publicatiebias**
3. Interpretatiemoeilijkheden II
4. zeer beperkte toepasbaarheid
   (zodat veel toepassingen eigenlijk incorrect zijn)







Menu

Er is het een en ander mis met gangbare praktijk van nulhypothesetoetsen / **p-waardes**:

1. publicatiebias

2. interpretatiemoeilijkheden II

(dit vertellen ze je *niet* op de universiteit)

1. zeer beperkte toepasbaarheid
   (zodat veel toepassingen eigenlijk incorrect zijn)

Juiste Interpretatie (Neyman-Pearson, 1937)

- We zetten *van te voren* een significantielevel (bijv. $\alpha = 0.05$) vast.
- Wanneer we nu waarnemen $p \leq \alpha$ zeggen we 'we verwerpen H0'. Anders 'accepteren we H0'.
- Wanneer we nu herhaaldelijk hypothesetoetsen (over verschillende onderwerpen) uitvoeren **zullen we gemiddeld genomen in hoogstens een fractie ~ van alle keren H0 verwerpen terwijl hij waar is**

- Neyman-Pearson zeggen eigenlijk: als je $p \leq \alpha$ waarneemt, dan moet je alleen maar 'reject' rapporteren, en niet de grootte van *p* – *die zegt niet zoveel!*

- Neyman-Pearson zeggen eigenlijk: als je $p \leq \alpha$ waarneemt, dan moet je alleen maar 'reject' rapporteren, en niet de grootte van *p* – *die zegt niet zoveel!*
- Maar dat vinden mensen begrijpelijkerwijs heel raar: als je *p* = 0.00001 hebt, heb je toch veel meer 'bewijs' dan bij *p* = 0.05. **Dus ze noemen de p-waarde wel!**
- ***...en dan begint het gedonder pas echt!***

Interpretatie-Probleem II

Wat betekent "een hele kleine p-waarde?"

Zou dit het zijn?

- Stel ik voer een reeks van n, zeg $n = 10^{10}$ toetsen achter elkaar uit, steeds weer in andere context. Laat p_j de p-waarde zijn die de *j*-de toets oplevert. Dan geldt, met grote kans, voor alle $0 < q < 1$:

$$\frac{\#\{i \in \{1,2,\ldots,n\} : p_i \leq q, H_0 \text{ is true}\}}{\#\{i \in \{1,2,\ldots,n\} : p_i \leq q\}} \lessapprox q$$

(dus binnen de deelverzameling van alle toetsen met uitkomst $p \leq 0.05$, is H0 HOOGSTENS in ongeveer in 5% van de gevallen waar; binnen de verzameling van toetsen met $p \leq$ 0.04%, ongeveer in 4%; etc.)

Zou dit het zijn?

- Stel ik voer een reeks van n, zeg $n = 10^{10}$ toetsen achter elkaar uit, steeds weer in andere context. Laat p_j de p-waarde zijn die de *j*-de toets oplevert. Dan geldt, met grote kans, voor alle $0 < q < 1$:

$$\frac{\#\{i \in \{1,2,\ldots,n\} : p_i \leq q, H_0 \text{ is true}\}}{\#\{i \in \{1,2,\ldots,n\} : p_i \leq q\}} \lessapprox q$$

- **Niet goed: dit is wederom de prosecutor's fallacy!**

Zou dit het zijn?

- Stel ik voer een reeks van n, zeg $n = 10^{10}$ toetsen achter elkaar uit, steeds weer in andere context. Laat p_j de p-waarde zijn die de *j*-de toets oplevert. Dan geldt, met grote kans, voor alle $0 < q < 1$:

$$\frac{\#\{i \in \{1,2,\ldots,n\} : p_i \leq q, H_0 \text{ is true}\}}{\#\{i \in \{1,2,\ldots,n\} : p_i \leq q\}} \lessapprox q$$

- **Niet goed!**
- **Probeer het slimmer te doen via (correct) alternatief voor Neyman-Pearson interpretatie**

Juiste Basis-Interpretatie II (Besliskundig, Wald, 1940)

- Steeds als ik p ≤ 0.05 observeer, doe ik een investering van € 20.
  - Als H0 toch correct was ben ik dat geld kwijt.
  - Als H0 inderdaad fout was, dan win ik iets (de precieze waarde doet er voor ons niet toe)
- Ik zou natuurlijk pech kunnen hebben, maar mijn verlies op de lange termijn is vrijwel zeker begrensd, want met zeer grote kans geldt:

$$\frac{1}{n}\sum_{i=1}^{n} \text{Verlies}_i \leq 1$$

- Steeds als ik p ≤ 0.01 observeer, doe ik een investering van € 100.
  - Als H0 toch correct was ben ik dat geld kwijt.
  - Als H0 inderdaad fout was, dan win ik iets (de precieze waarde doet er voor ons niet toe)
- Ik zou natuurlijk pech kunnen hebben, maar mijn verlies op de lange termijn is vrijwel zeker begrensd, want met zeer grote kans geldt:

$$\frac{1}{n}\sum_{i=1}^{n} \text{Verlies}_i \leq 1$$

- Steeds als ik p <0.1 observeer, investeer ik € 10. Als H0 toch correct was ben ik dat geld kwijt. Als H0 inderdaad fout was, dan win ik iets (onbepaalds)
- Steeds als ik p <0.01 observeer, investeer ik € 100. Als H0 toch correct was ben ik dat geld kwijt. Als H0 inderdaad fout was, dan win ik iets (onbepaalds)
- Steeds als ik p <0.001 observeer, investeer ik €1000. Als H0 toch correct was ben ik dat geld kwijt. Als H0 inderdaad fout was, dan win ik iets (onbepaalds)
- **Je zou nu hopen dat nog steeds met grote kans geldt:** $\frac{1}{n}\sum_{i=1}^{n} \text{Verlies}_i < \ldots$

- **HELAAS: onder H0 geldt...**

$$\mathbf{E}[\text{Verlies}] = \frac{1}{10}\cdot 10 + \frac{1}{100}\cdot 100 + \frac{1}{1000} + \ldots = \infty$$

....dus als H0 steeds maar weer waar is, dan geldt met kans 1:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \text{Verlies}_i = \infty$$

Kleine p-waardes

- Interpretatie 'p-waarde veel kleiner dan significantielevel' volstrekt onduidelijk
- Het hele bouwwerk is alleen te interpreteren als je 'reject' (als $p \leq \alpha$) of 'accept' rapporteert, en niet p zelf!
- **Daarom kun je ook niet zonder dat rare significance level!**

Menu

Er is het een en ander mis met gangbare praktijk van nulhypothesetoetsen / **p-waardes**:

1. publicatiebias
2. interpretatiemoeilijkheden
3. **zeer beperkte toepasbaarheid** (zodat veel toepassingen eigenlijk incorrect zijn)

Menu

Er is het een en ander mis met gangbare praktijk van nulhypothesetoetsen / **p-waardes**:

1. publicatiebias
2. interpretatiemoeilijkheden
3. **zeer beperkte toepasbaarheid** - maar men past het toch toe, ook als het niet kan (je moet wat) - zodat veel toepassingen eigenlijk incorrect zijn

Beperking van de p-waarde

- p-waardes zijn slechts gedefinieerd als we van te voren weten wat de mogelijke uitkomsten van het experiment zijn, en wat voor kansen ze hebben onder de nul/alternatieve hypothesen...
- **Dit lijkt alleszins redelijk – maar is het niet!**

Beperking van de p-waarde

- **De Weerman/Vrouw:**
  We kunnen p-waardes niet gebruiken om te bepalen wie beter is:
  Marjon de Hond (NOS) of Peter Timofeeff (RTL)


Beperking van de p-waarde

- p-waardes zijn slechts gedefinieerd als we van te voren weten wat de mogelijke uitkomsten van het experiment zijn, en wat voor kansen ze hebben onder de nul/alternatieve hypothesen...
- dit impliceert dat *voordat* het experiment begint een **protocol ("sampling plan")** opgesteld moet zijn...
  - Bekijk precies 100 patienten
  - Bekijk steeds weer nieuwe patienten totdat er een patient > 39 graden koorts krijgt
  - ...etc.

Beperking van de p-waarde

- p-waardes zijn slechts gedefinieerd als er *voordat* het experiment begint een protocol ("sampling plan") opgesteld is
- Dit lijkt ook weer een redelijke eis: als we door mogen 'sampelen' totdat de resultaten er toevallig even goed uitzien en op dat moment stoppen (**optional stopping**) dan lijkt het alsof we de boel bedotten
- Maar is het wel zo redelijk!?!?

57

Het kan beter...

- We zouden graag willen werken met een methode die ook 'achteraf' gebruikt kan worden, als we het protocol niet weten!
  - zulke methoden bestaan!
  - handig bijproduct: **onderzoeker mag lekker meer data vergaren als hij een 'veelbelovend maar nog niet heel overtuigend' onderzoeksresultaat ziet**

Menu

Problemen met p-waardes:

1. publicatiebias
2. interpretatiemoeilijkheden
3. zeer beperkte toepasbaarheid
   (zodat veel toepassingen eigenlijk incorrect zijn)

Bayesiaanse methode

– voorkomt probleem 2&3 geheel, 1 deels...maar heeft andere problemen

Test Martingaal methode

–'almost the best of both worlds'

De Stelling van Bayes

Posterior odds = likelihood ratio * prior odds

$$\frac{\Pr(H_0 \mid D)}{\Pr(H_1 \mid D)} = \frac{\Pr(D \mid H_0)}{\Pr(D \mid H_1)} \cdot \frac{\Pr(H_0)}{\Pr(H_1)}$$

- Als je bereid bent om kansen aan H0 en H1 en binnen H1 toe te kennen...

De Stelling van Bayes, Bem

H0: mensen kunnnen niet in toekomst kijken

D: 383 'hits'

Vrij klein

bijna 1

$$\frac{\Pr(H_0 \mid D)}{\Pr(H_1 \mid D)} = \frac{\Pr(D \mid H_0)}{\Pr(D \mid H_1)} \cdot \frac{\Pr(H_0)}{\Pr(H_1)}$$

vrij groot

heel klein

$$\Pr(D \mid H_1) = \int p^{n_0}(1-p)^{n_1} w(p)\,dp$$

BEM revisited

"De bewijskracht voor een extreme bewering moet worden geschaald naar haar buitenissigheid"

Pierre-Simon Laplace, 1749 – 1827,
Vader (met Bayes) van de Bayesiaanse statistiek

Bayes kan dus 'publication bias' voorkomen
...maar dan moet je wel 'goede priors' hebben!

Menu

Problemen met p-waardes:

1. publicatiebias
2. interpretatiemoeilijkheden
3. zeer beperkte toepasbaarheid
   (zodat veel toepassingen eigenlijk incorrect zijn)

Bayesiaanse methode

– voorkomt probleem 3 geheel, 1 deels...maar heeft andere problemen

Test Martingaal methode

–'almost the best of both worlds'

Test Martingaal

Ville (1939), Levin (1973), Vovk (1993-nu), G. (2012)

- Laat $X_1, X_2, X_3, \ldots \sim P_{H0}$ en voor alle i, M_i een deterministische functie van $X_1, \ldots, X_i$ zdd

$$M_i \geq 0$$
$$\mathbf{E}\left[M_i \mid X_1, \ldots, X_{i-1}\right] \leq 1$$

Dan is $S_1, S_2, S_3, \ldots$ met $S_n = \prod_{i=1}^{n} M_i$ een **test-martingaal onder H0**

Heldere Interpretatie!

- Neem voor het gemak weer als H0:
  $X_1, X_2, \ldots \sim$ i.i.d. Bernoulli$(1/2)$
- Stel er worden op elk tijdstip *i* twee loterij-tickets aangeboden.
- Beide tickets kosten €1. Ticket *j* betaalt €2 uit als uitkomst *j* is, met $j \in \{0, 1\}$
  - **Je mag je geld over beide tickets verdelen, en ook bijv. 1/3 of 7 tickets kopen**
  - **Als H0 waar is, is dit een eerlijk spel!**

Heldere Interpretatie!

- Beide tickets kosten €1. Ticket *j* betaalt €2 uit als uitkomst *j* is, met $j \in \{0, 1\}$
- We beginnen nu met €1 startkapitaal en verdelen dat over beide tickets voor X_1 Vervolgens herverdelen we ons eindkapitaal weer over beide tickets voor X_2 Vervolgens voor X_3 en .. en ...
- Voor elke sequentiele herverdelingsstrategie is er een test martingaal zodat S_i je kapitaal op tijdstip *i* is, en v.v.

Interpretatie is universeel

- S_i geeft aan hoeveel geld je op tijdstip *i* hebt gewonnen door sequentieel te gokken op uitkomsten met bepaalde gokstrategie onder contracten die eerlijk zouden zijn als H0 waar was, en je kapitaal steeds te herinvesteren
- **Hoe meer geld, hoe meer 'bewijs' tegen H0 (buitengewoon logisch!)**
- Verschillende martingalen (gok-strategieen) corresponderen met verschillende alternatieve hypothesen

Resultaten: best of both worlds

- S_n volledig bepaald door
  $\Pr(X_1, \ldots, X_n \mid H_0)$
  de kans op *de daadwerkelijk geobserveerde (en niet 'counterfactual')* data
- Je kunt bepalen wie de beste weersvoorspeller is!

Resultaten: best of both worlds

- **Stelling:** Voor elke alternatieve hypothese H1 en elke verdeling over de kans-verdelingen in H1 geldt dat de **Bayes factor**

$$S_n = \Pr(X_1, \ldots, X_n \mid H_1) / \Pr(X_1, \ldots, X_n \mid H_0)$$

een test martingaal is (Savage, 1961)

Resultaten: best of both worlds

- **Stelling:** Voor elke alternatieve hypothese H1 en elke verdeling over de kans-verdelingen in H1 geldt dat de **Bayes factor**

$$S_n = \Pr(X_1, \ldots, X_n \mid H_1) / \Pr(X_1, \ldots, X_n \mid H_0)$$

een test martingaal is (Savage, 1961)

- **Maar niet omgekeerd!**

(G. De Rooij, Van Erven, Journal of the Royal Statistical Society Series B, 2012)

Resultaten: best of both worlds

- **Standaard p-waarde:** for all $0 \leq \alpha \leq 1$:

$$\Pr(p_{\text{standard}}(T) \leq \alpha) = \alpha$$

- **Stelling**: elke supermartingaal kan gezien worden als een **robuuste p-waarde** (Doob (1950s), Vovk, G. (2000s))

$$\Pr\left(\exists n : \frac{1}{S_i} \leq \alpha\right) \leq \alpha$$

Resultaten: best of both worlds

- **Stelling**: elke supermartingaal kan gezien worden als een **robuuste p-waarde** (Doob (1950s), Vovk, G. (2000s))

$$\Pr\left(\exists n : \frac{1}{S_i} \leq \alpha\right) \leq \alpha$$

- **Gevolg**: *als* je per se een significance level wilt gebruiken, geldt de Neyman-Pearson interpretatie ook al is het sampling plan onbekend/doe je aan optional stopping

Take Home Message

- Standaard p-waardes hebben geen heldere interpretatie en zijn zeer beperkt toepasbaar
- Test martingalen hebben heldere interpretatie **(geld!)**, zijn breed toepasbaar, en zijn te relateren aan Bayesiaanse methoden *en* gerobustificeerde p-waardes