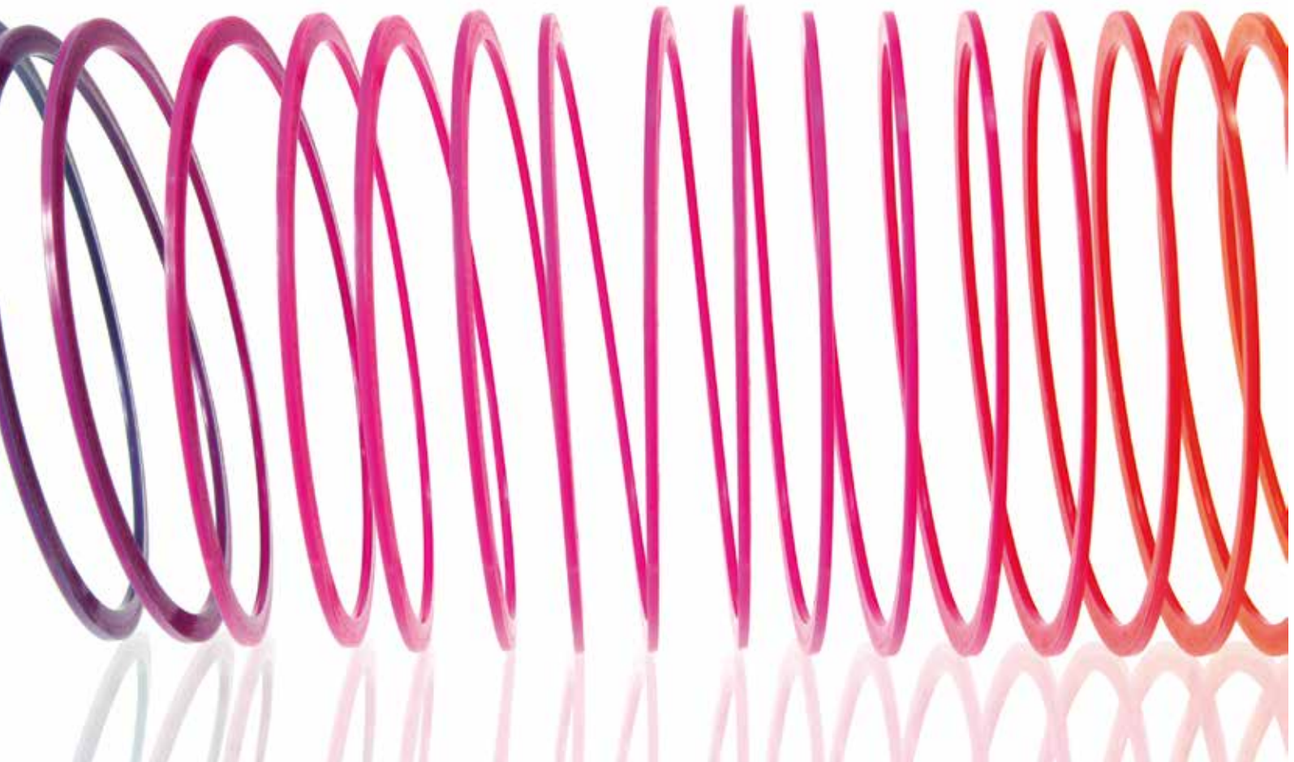


FLEXIBELE STATISTIEK

Van p-waarden naar e-waarden; Robuustere conclusies met minder data

De rigide natuur van p-waarden maakt ze steeds minder passend bij huidige mogelijkheden van dataverwerking: een groeiende dataset continu analyseren en data uitwisselen tussen verschillende instituten was nog nooit zo makkelijk. Een flexibeler alternatief zijn e-waarden, waarbij de e staat voor *evidence*. In dit artikel laten we zien hoe voor een breed scala aan scenario's van nulhypothesetoetsen in datastromen op simpele wijze e-waarden kunnen worden ontworpen, en hoe deze kunnen worden toegespitst op het zo snel mogelijk verzamelen van bewijs voor een alternatieve hypothese.

Rosanne J. Turner

De afgelopen jaren zijn er veel goede ontwikkelingen geweest op het gebied van data-infrastructuur voor wetenschappelijk onderzoek, zoals bijvoorbeeld de *Personal Health Train* en het FAIR-data initiatief (Van Soest et al., 2018; Wilkinson et al., 2016). Wetenschappers hebben nog nooit zoveel mogelijkheden gehad om veilig data uit te wisselen met collega's in andere instituten en om via dashboards data live te analyseren. Deze ontwikkelingen brengen interessante statistische uitdagingen met zich mee. Wat moet een onderzoeker bijvoorbeeld doen als hij na het behalen van zijn geplande steekproefgrootte een net niet significant resultaat behaalt, en een collega aanbiedt dat hij tien extra samples toe kan voegen? Of als de geplande steekproefgrootte nog niet behaald is, maar de onderzoeker tussendoor uit nieuwsgierigheid alvast een p-waarde uitrekent, die dan significant blijkt te zijn?

Sequentieel toetsen met klassieke p-waarden

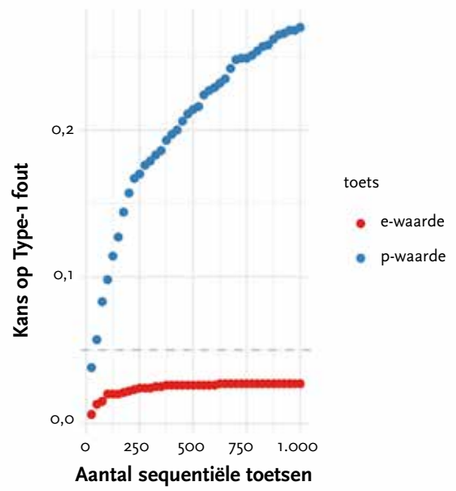
Peter Grünwald gaf eerder al in *STAtOR* een mooi overzicht van hoe het gebruik van klassieke nulhypothesetoetsen en p-waarden in (onder andere) dit soort situaties problematisch is (Grünwald, 2015). Klassieke toetsen zoals de gepaarde t-toets of Fishers exacte toets geven alleen een garantie op de kans om onterecht de nulhypothese te verwerpen, de Type-1 fout (*false positive rate*), als vooraf de steekproefgrootte exact wordt vastgesteld.

Stel dat we als statistiekconsultant een onderzoeker proberen te adviseren die de nulhypothese wil toetsen dat medicijn *a* een even grote kans op succes biedt als medicijn *b*. De onderzoeker heeft een dashboard voor hun studie opgezet en krijgt iedere keer als in beide groepen één patiënt het medicatietraject heeft afgerond een update. De data komen dus binnen in gebalanceerde blokken van groepsgrootte $n_a = 1$ in groep *a*, en $n_b = 1$ in groep *b*. Het liefst zou de onderzoeker op ieder moment dat het dashboard geüpdatet wordt een nulhypothesetoets doen, om zo snel mogelijk de studie af te kunnen ronden. In figuur 1 is geïllustreerd wat er zou gebeuren als we voor deze analyse Fishers exacte toets zouden gebruiken: onze kans om een Type-1 fout te maken blijft alsmaar stijgen naarmate we meer en meer datapunten verzamelen. Eigenlijk is onze hele statistische analyse oninterpreteerbaar geworden!

E-waarden

In figuur 1 is ook te zien dat met een analyse met zogenoemde e-waarden in plaats van p-waarden de Type-1 fout wel begrensd blijft. E-waarden zijn een alternatief voor p-waarden voor het doen van nulhypothesetoetsen, waarbij de e staat voor *evidence*. Zoals deze naam misschien al doet vermoeden zijn de e-waarden een maat voor bewijs voor de alternatieve hypothese in de data. Door de definitie van e-waarden blijven deze naar verwachting laag als data in werkelijkheid gegenereerd worden onder de nulhypothese: alle niet-negatieve *random variables* (kansvariabelen) met een verwachte waarde van hoogstens 1 onder alle verdelingen in de nulhypothese zijn e-waarden (Grünwald et al., 2019; Vovk & Wang, 2021).

Door deze definitie blijft het product van sequentieel verzamelde e-waarden zelf ook weer een e-waarde, ongeacht de methoden die de onderzoeker heeft gebruikt om te besluiten of de studie door moest lopen of stoppen (details in Grünwald et al., 2019). Dit betekent dat we onze onderzoeker zouden kunnen adviseren iedere keer als het dashboard geüpdatet is een e-waarde uit te reke-



Figuur 1. De geschatte kans op het maken van een Type-1 fout bij het toetsen van de nulhypothese dat de kans op succes in twee groepen hetzelfde is, onder sequentieel toetsen met Fishers exacte toets en e-waarden (aangepast uit Turner et al, 2021)

nen met deze nieuwe datapunten. En nu komt het mooie: uit de eigenschap dat e-waarden onder de nulhypothese een verwachte waarde van hoogstens 1 hebben volgt dat we ze kunnen gebruiken voor een nulhypothesetoets. We verwerpen de nulhypothese als de e-waarde groter is dan $1/\alpha$, wat ons een Type-1 fout garantie geeft van α, zelfs als we na iedere update opnieuw ons product van sequentiële e-waardes toetsen. Als onze onderzoeker bijvoorbeeld een significantieniveau van 0,05 aan wil houden, adviseren we de nulhypothese te verwerpen zodra het product van de sequentiële e-waarden groter is dan 20.[1]

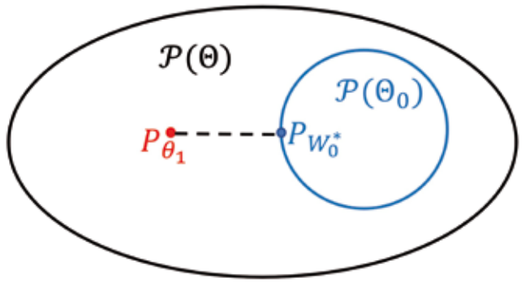
Zoveel mogelijk bewijs

E-waarden bieden dus die Type-1 fout garantie onder sequentieel toetsen, maar hoe gedragen ze zich als de alternatieve hypothese waar is? De lezer kon zich bij de brede definitie van e-waarden in de vorige paragraaf wellicht gelijk een aantal 'domme' keuzes voorstellen, bijvoorbeeld een *random variable* die altijd de waarde 1 aanneemt. In Grünwald et al. (2019) wordt voorgesteld e-waarden zo te kiezen, dat ze optimaal bewijs verzamelen voor de alternatieve hypothese in de sequentiële setting die we hierboven beschreven. De e-waarde die dit doet voor een puntalternatief $p_{\theta_1}^*$ kan gevonden worden door *reverse information projection*:

$$E^*(z^n) = \frac{p_{\theta_1^*}(z^n)}{p_{W_0^*}(z^n)}, \text{ zodat } D\left(P_{\theta_1^*} \parallel P_{W_0^*}\right) = \min_{W_0 \in W(\Theta_0)} D\left(P_{\theta_1^*} \parallel P_{W_0}\right). \quad (1)$$

Hierbij is D de Kullback-Leibler divergentie (of: relatieve entropie), een divergentiemaat uit de informatietheorie. De verzameling $\left\{P_{W_0} : W_0 \in W(\Theta_0)\right\}$, kan gezien worden als de verzameling van Bayesiaanse marginaalverdelingen voor de data z^n, ééntje voor elke prior W_0 op de nulhypothese Θ_0.[2] We kiezen dan de prior W_0^* zodat $P_{W_0^*}$ de marginaal over de nulverdelingen is die het 'minst verschilt van' of het 'dichtste bij' ons puntalternatief staat, zoals geïllustreerd (in Euclidische afstand) in figuur 2.

Het mooie aan vergelijking (1) is dat we een algemene uitdrukking hebben om voor elke combinatie van een alternatief en nulhypothese een goede e-waarde te definieren, maar het vinden van een oplossing lijkt wellicht niet triviaal. Voor een heel aantal toetsscenario's is ook nog geen analytische oplossing gevonden, en moet gebruik gemaakt worden van een numerieke benadering via optimalisatie, zoals geïllustreerd in Turner, 2019 en Lardy, 2021.



Figuur 2. De *reverse information projection* van een puntalternatief op de nulhypothese (figuur samengesteld analoog aan, als simplificatie van, Grünwald et al., 2019, figuur 1)

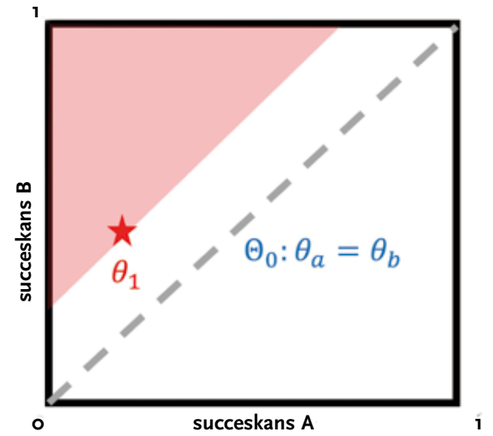
Een simpele oplossing voor data in twee groepen

Verrassend genoeg blijkt er voor een grote groep scenario's, namelijk waarbij we kansverdelingen in twee groepen vergelijken, een simpele generieke vorm van e-waarden te bestaan (Turner et al., 2021). Voor deze scenario's nemen we aan dat er een set aan kansverdelingen is, $P(\Theta) := \left\{P_\theta, \theta \in \Theta\right\}$, en dat datapunten $y_{i,a}$ en $y_{i,b}$ in groep a en b gegenereerd worden door een bepaalde verdeling uit deze set, P_{θ_a} en P_{θ_b} – dit is ons puntalternatief. Verder leggen we van tevoren vast hoeveel datapunten we in dit blok in elke groep gaan verzamelen: n_a in groep a, en n_b in groep b (tussen verschillende datablokken in mogen we deze aantallen wel veranderen). Indien we dan de nulhypothese willen toetsen dat de kansverdelingen in groep a en b gelijk zijn, versimpelt de e-waarde tot:

$$\frac{\prod_{i=1}^{n_a} p_{\theta_a}(y_{i,a})}{\prod_{i=1}^{n_a}\left(\frac{n_a}{n} p_{\theta_a}(y_{i,a}) + \frac{n_b}{n} p_{\theta_b}(y_{i,a})\right)} \frac{\prod_{i=1}^{n_b} p_{\theta_b}(y_{i,b})}{\prod_{i=1}^{n_b}\left(\frac{n_a}{n} p_{\theta_a}(y_{i,b}) + \frac{n_b}{n} p_{\theta_b}(y_{i,b})\right)}. \quad (2)$$

Deze formule kan zelfs verder versimpeld worden als de nulhypothese een bepaalde eigenschap bezit, convexiteit. Dan kunnen we de noemers in vergelijking (2) vervangen door de waarschijnlijkheid van de data onder een bepaalde kansverdeling uit onze set, $p_{\theta_0} = \frac{n_a}{n} p_{\theta_a} + \frac{n_b}{n} p_{\theta_b}$. In dit geval weten we ook dat deze e-waarde de beste e-waarde is om bewijs te verzamelen voor ons puntalternatief.

In het voorbeeld van onze onderzoeker is dit toevallig zo. Stel dat onze onderzoeker verwacht dat in groep a het medicijn succesvol is bij 20% van de patiënten, en dat het in groep b twee keer zo goed zal werken (de rode ster in figuur 3). De beste e-waarde voor dit alternatief ziet er dan als volgt uit: in de teller komt simpelweg de kans op de data in elke groep onder Bernoulli verdelingen met parameterwaardes 0,2 en 0,4, en in de noemer de kans op de gehele data onder een Bernoulli verdeling met parameterwaarde 0,3.



Figuur 3. De parameterruimte voor het nulhypothesetoets-scenario waarbij proporties in twee groepen vergeleken worden. De gestippelde lijn, Θ_0, geeft de parameterruimte van de nulhypothese aan

Leren van eerder geziene data

In de praktijk komt het niet vaak voor dat men genoeg voorkennis heeft om een de e-waarde toe te spitsen op een puntalternatief, maar bijvoorbeeld wel op een minimaal klinisch relevant verschil tussen de groepen (de roze driehoek in figuur 3), of een minimale odds ratio. In dat geval kan een prior op de parameterruimte geplaatst worden, en kan in combinatie met data gezien in eerdere blokken de beste e-waarde om bewijs voor de alternatieve hypothese van vorm (2) benaderd worden (details in Turner et al., 2021; en Turner et al., 2022). Het blijkt dat we met deze 'lerende' sequentiële methode net zo snel of soms zelfs sneller de nulhypothese kunnen verwerpen dan met klassieke hypothesetoetsen, waarbij we met de e-waarden altijd nog de mogelijkheid houden de dataset verder uit te breiden. Ook al gebruiken we hier priors, de aanpak is niet echt Bayesiaans. De Type-I foutgarantie geldt namelijk altijd, wat voor prior we ook kiezen – de keuze beinvloedt alleen hoe snel we de nulhypothese kunnen verwerpen als de nulhypothese niet waar is.

Kortom, e-waarden bieden mooie mogelijkheden data op een meer flexibele, maar nog steeds robuuste, manier te analyseren. Hopelijk kan het aanbieden van dit soort makkelijk aanpasbare statistische methoden bijdragen aan het verminderen van toekomstige *research waste* (Glasziou & Chalmers, 2018).

Deze tekst is gebaseerd op eerdere publicaties (Turner, 2019; Turner et al., 2021). Dank aan co-auteur Peter Grünwald.

Noten

1. Indien de onderzoeker heel graag een p-waarde wil rapporteren kan dat ook: 1 delen door de e-waarde geeft een conservatieve p-waarde.
2. Definitie van een Bayesiaanse marginaalverdeling: $P_{W_0}(Z^n) = \int w_0(\theta_0)\, P_{\theta_0}(Z^n) d\theta_0$.

Literatuur

Glasziou, P., & Chalmers, I. (2018). Research waste is still a scandal; An essay by Paul Glasziou and Iain Chalmers. *British Medical Journal, 363*.

Grünwald, P. D. (2015). Paranormale statistiek: Over de vele problemen met p-waarden, en een redelijk alternatief. *STAtOR, 4*.

Grünwald, P.D., De Heide, R. & Koolen, W. (2019). *Safe testing*. Preprint available on arXiv as arXiv:1906.07801.

Lardy, T. (2021). *E-values for hypothesis testing with covariates*. Master's thesis. Leiden University.

Turner, R. J. (2019). *Safe tests for 2 x 2 contingency tables and the Cochran-Mantel-Haenszel test*. Master's thesis, Leiden University.

Turner, R. J., Ly A., & Grünwald, P.D. (2021). *Two-sample tests that are safe under optional stopping, with an application to contingency tables*. Preprint available on arXiv as arXiv:2106.02693 [stat.ME].

Turner, R. J., Ly, A., Pérez, M. F., ter Schure, J. A., & Grünwald, P. D. (2022). Safestats: Safe anytime-valid inference. R package version 0.8.6, available at https://cran.r-project.org/web/packages/safestats/index.html..

Van Soest, J., Sun C., Mussmann O. et al. (2018). Using the personal health train for automated and privacy-preserving analytics on vertically partitioned data. In *Building continents of knowledge in Oceans of Data: The future of co-created eHealth,* 581–585.

Vovk, V., & Wang, R. (2021). E-values: Calibration, combination, and applications. *The Annals of Statistics, 49*(3), 1736–1754.

Wilkinson, M.D, Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*, 160018.

Rosanne Turner is PhD student statistiek bij Professor Peter Grünwald en Professor Floortje Scheepers aan het Centrum Wiskunde en Informatica en het UMC Utrecht. In 2018 behaalde zij een PhD in de Geneeskunde en in 2019 sloot zij haar Master Statistical Science for the Life and Behavioural Sciences af aan de Universiteit Leiden, waarbij zij voor haar scriptie de Jan Hemelrijk Award van de VVSOR ontving.
E-mail: Rosanne.Turner@cwi.nl