OPTIONAL STOPPING MET BAYES FACTORS
mogelijkheden en beperkingen

Optional stopping **betekent 'naar de resultaten kijken om te beslissen of je nog wat data aan het experiment toe wilt voegen'. Vaak wordt er beweerd dat de Bayes factor voor hypothesetoetsen geldig blijft tijdens optional stopping. Echter, dat hangt af van de specifieke situatie.**

Rianne de Heide

Er zijn drie verschillende wiskundige definities te geven voor het fenomeen *optional stopping*. Of we in de praktijk kunnen zeggen 'de Bayes factor methode voor hypothesetoetsen blijft geldig tijdens optional stopping' is een subtiele kwestie, die afhangt van de specifieke kenmerken van de gegeven situatie: welke modellen en welke priors worden gebruikt, en wat is het doel van de analyse.

De afgelopen jaren bleken de conclusies van papers verrassend vaak onjuist wanneer de experimenten werden gerepliceerd. Een deel van deze *replicability crisis* wordt veroorzaakt doordat in praktijk noodzakelijke voorwaarden voor het gebruik van de klassieke (frequentistische) methodes worden geschonden (John, Loewenstein & Prelec, 2012). Een van die voorwaarden is dat het experimentele protocol van te voren volledig moet zijn vastgelegd. In de praktijk passen onderzoekers het juist vaak aan, door onvoorziene omstandigheden, of ze verzamelen gegevens tot ze een bevredigend resultaat zien. Dit laatste wordt *optional stopping* genoemd, en kan er voor zorgen dat hypotheses veel vaker ten onrechte worden verworpen dan de foutgaranties van de statistische methodes beloven. Hypothesetoetsen met *Bayes factors* wordt al lang bepleit als een alternatief voor traditionele hypothesetoetsmethodes dat verschillende problemen kan oplossen, en in het bijzonder werd al vroeg beweerd dat Bayesiaanse methodes geldig blijven tijdens optional stopping (Lindley, 1957; Raiffa & Schlaifer, 1961; Edwards, Lindman, & Savage, 1963). In het licht van de replicability crisis stonden deze claims kortgeleden opnieuw in de belangstelling (Wagenmakers, 2007; Rouder, 2014; Schönbrodt et al., 2017; Yu et al., 2014; Sanborn & Hills, 2014). Maar wat betekenen ze wiskundig? Het blijkt dat verschillende auteurs verschillende dingen bedoelen met 'Bayesiaanse methodes kunnen omgaan met optional stopping'; en bovendien blijken dergelijke claims vaak alleen geldig te zijn in informele zin of in beperkte situaties. In het paper (Hendriksen, De Heide, and Grünwald 2020) geven we een systematisch overzicht en een formalisering van dergelijke claims, en leggen hun relevantie voor de praktijk uit: kunnen we er op vertrouwen dat Bayes factor hypothesetoetsen geldig blijven tijdens optional stopping, of niet? We zullen zien dat het antwoord genuanceerd ligt. Ten tweede breiden we het bereik van die claims uit tot meer algemene situaties, waarin ze nooit formeel zijn geverifieerd en waarvoor dat niet triviaal is. In het paper (De Heide & Grünwald, 2020) lichten we die claims vervolgens toe voor een publiek van methodologen en toegepaste statistici met behulp van computersimulaties.

Bayesiaans leren

Bayesianisme gaat over een bepaalde interpretatie van het begrip *waarschijnlijkheid*: als *geloofsgraden*. Een Bayesiaan begint met het uitdrukken van dit geloof als een waarschijnlijkheidsfunctie. We noemen dit de *prior*, en we duiden het aan met $\mathbb{P}(\theta)$, waarbij θ de parameter (of meerdere parameters) van het model is. Na specificatie van de prior, verkrijgen we de data D en likelihood $\mathbb{P}(D|\theta)$. Nu kunnen we de posterior $\mathbb{P}(\theta|D)$ berekenen met behulp van de stelling van Bayes:

$$\mathbb{P}(\theta|D) = \frac{\mathbb{P}(D|\theta)\mathbb{P}(\theta)}{\mathbb{P}(D)}.$$

Stel dat we een nulhypothese $\mathcal{H}_0$ willen toetsen tegenover een alternatieve hypothese $\mathcal{H}_1$. We kunnen dit op een Bayesiaanse manier doen met *Bayes factors*: we beginnen met de *prior odds* $\mathbb{P}(\mathcal{H}_1)/\mathbb{P}(\mathcal{H}_0)$, ons geloof voordat we de data zien. Vaak geloven we dat beide hypotheses even waarschijnlijk zijn, en dan zijn onze prior odds 1 staat tot 1. Daarna verzamelen we de data D, en we werken onze odds bij met onze nieuwe kennis, met behulp van de stelling van Bayes:

$$\text{posterior odds } (\mathcal{H}_1 \text{ vs. } \mathcal{H}_0) = \frac{\mathbb{P}(\mathcal{H}_1|D)}{\mathbb{P}(\mathcal{H}_0|D)} = \frac{\mathbb{P}(\mathcal{H}_1)}{\mathbb{P}(\mathcal{H}_0)}\frac{\mathbb{P}(D|\mathcal{H}_1)}{\mathbb{P}(D|\mathcal{H}_0)}.$$

De posterior odds is onze bijgewerkte overtuiging over welke hypothese waarschijnlijker is.

Optional stopping: drie begrippen

Een wenselijke eigenschap van hypothesetoetsen is geldigheid tijdens optional stopping: we verzamelen wat gegevens, we kijken naar de tussenresultaten, en besluiten op basis daarvan of we stoppen of nog even doorgaan en nog wat meer data verzamelen. Informeel noemen we het 'gluren naar de data tot dusver om te beslissen of we nog meer data gaan verzamelen of niet' *optional stopping*, maar als we het wat preciezer willen maken wat het nou echt betekent als een toets geldig blijft tijdens optional stopping, blijkt het dat verschillende benaderingen (frequentistisch, subjectief en objectief Bayesiaans) leiden tot verschillende interpretaties en definities. Het blijkt dat we drie verschillende wiskundige concepten kunnen onderscheiden, die we identificieren en formeel definiëren in het paper (Hendriksen, De Heide & Grünwald, 2020).

Subjectief Bayesiaans optional stopping

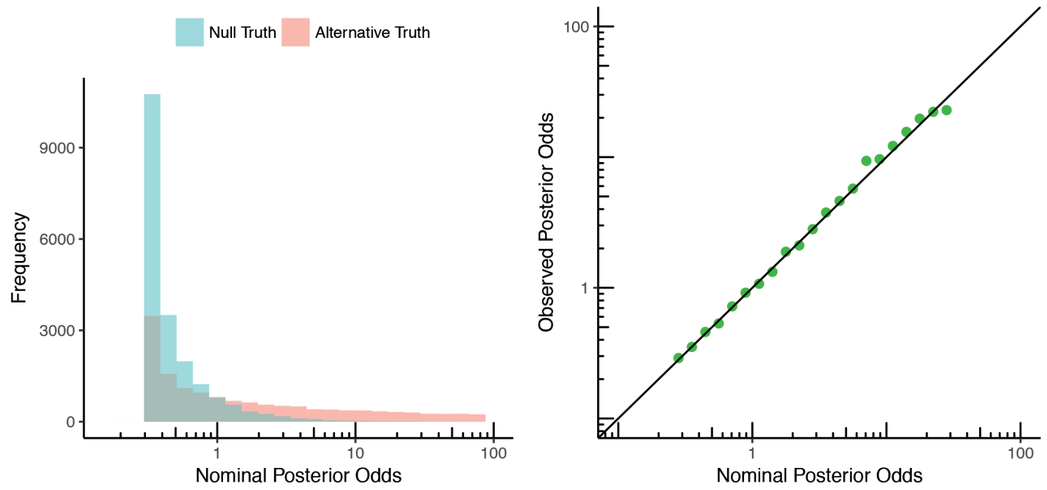
Het eerste begrip noemen we *subjectief Bayesiaans optional stopping* of π-*onafhankelijkheid*. Als je het puur subjectief Bayesiaans beschouwt — alleen als je werkelijk in je prior gelooft — dan wordt Bayesiaans updaten van prior naar posterior niet beïnvloed door de gebruikte stopregel: je eindigt met dezelfde posterior als je de steekproefgrootte n van te voren had vastgelegd, of als hij bijvoorbeeld was bepaald omdat je het resultaat na n datapunten goed genoeg vond. In die zin is een subjectief Bayesiaanse procedure niet afhankelijk van de stopregel.

Kalibratie

De tweede vorm van optional stopping noemen we *kalibratie*. Zoals (Rouder, 2014) schrijft: 'Als een herhalingsexperiment een posteriorkans van 3,5 staat tot 1 oplevert in het voordeel van de nulhypothese, dan verwachten we dat de data 3,5 keer zo waarschijnlijk door de nulhypothese zijn geproduceerd, als door de alternatieve hypothese.' In meer wiskundige taal kan dit worden uitgedrukt als:

$$\text{post-odds } (\mathcal{H}_1 \text{vs.} \mathcal{H}_0 | \text{“post-odds } (\mathcal{H}_1 \text{vs.} \mathcal{H}_0 | D) = a\text{”}) = a.$$

We zeggen dat deze vergelijking *kalibratie van de posterior odds* uitdrukt. Het blijkt dat deze kalibratie niet standhoudt als je er geen puur subjectieve visie van Bayesianisme op nahoudt, en in het bijzonder geldt dit niet voor



Figuur 1. Posterior odds in een experiment waarin we toetsen of de verwachtingswaarde van een normale verdeling 0 is (H_0), versus niet-nul (H_1), van 20:000 experimenten. (a) de empirische steekproevenverdeling als histogram onder H_0 (blauw) en H_1 (roze). (b) Kalibratieplot: de geobserveerde posterior odds als een functie van de nominal posterior odds

de *default* priors die enkele Bayesianen in de psychologie bepleiten (Wagenmakers, 2007; Rouder et al., 2012). Om een eerste idee van de problemen hieromtrent te krijgen: default priors zijn soms afhankelijk van de data. In dat geval is het onduidelijk wat *optional stopping* nou eigenlijk betekent, omdat als je een prior $P_1(\theta)$ gebruikt die gebaseerd is op een steekproefgrootte van n, maar je bent gestopt bij $n'<n$, dan had je eigenlijk prior $P'_1(\theta)$ gebaseerd op n' moeten gebruiken... maar dan zou je kunnen stoppen bij $n''<n'$, etc. Zie ons paper (De Heide & Grünwald, 2020) voor een uitgebreide discussie en veel voorbeelden.

Frequentistisch optional stopping

De derde vorm is een frequentistische interpretatie van het omgaan met optional stopping, die gaat over het controleren van de Type-I-fout van een experiment. Een Type-I-fout treedt op als we de nulhypothese verwerpen wanneer deze waar is, ook wel een *vals positief* genoemd. De frequentistische interpretatie van geldigheid tijdens optional stopping is dat de Type-I-foutgarantie nog steeds geldt als we de onderzoeksopzet — en dus de stopregel — niet van tevoren vastleggen, maar we mogen stoppen als we een significant resultaat zien. In het geval dat $\mathcal{H}_0$ enkelvoudig is (bestaat uit slechts één hypothese), is er een bekend verband tussen Bayes factors en Type-I-foutwaarschijnlijkheden: als we $\mathcal{H}_0$ verwerpen dan en slechts dan als de posterior odds in het voordeel van $\mathcal{H}_0$ kleiner zijn dan een vast niveau α, dan hebben we een garantie dat onze Type-I-fout ten hoogste α is. En interessant genoeg geldt dit niet alleen voor een vaste steekproefgrootte, maar ook tijdens optional stopping. Echter, voor *meervoudige* $\mathcal{H}_0$, geldt dit verband niet. Behalve in het speciale geval waar *alle* vrije parameters in $\mathcal{H}_0$ zogenoemde *nuisance parameters* zijn, die een groepsstructuur bezitten en die zijn uitgerust met de bijbehorende right-Haar prior, en worden gedeeld met $\mathcal{H}_1$, zoals we bewijzen in (Hendriksen, Heide, and Grünwald 2020). Maar voor algemene priors en voor meervoudige $\mathcal{H}_0$ is dit meestal niet het geval.

Conclusie

Er zijn drie verschillende wiskundige definities te geven aan het fenomeen *optional stopping*. Of we in de praktijk kunnen zeggen 'de Bayes factor methode voor hypothesetoetsen blijft geldig tijdens optional stopping' is een subtiele kwestie, die afhangt van de specifieke kenmerken van de gegeven situatie: welke modellen en welke priors worden gebruikt, en wat is het doel van de analyse.

Literatuur

Edwards, W., Lindman, H., & Savage. L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193–242. https://doi.org/10.1037/h0044139.

De Heide, R., &Grünwald, P.D. 2021. Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review, 28*(3), 795–812.

Hendriksen, A., de Heide, R., & Grünwald, P. 2021. Optional stopping with Bayes factors: a categorization and extension of folklore results, with an application to invariant situations. *Bayesian Analysis, 16*(3), 961–989.

John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*.

Lindley, D.V. (1957). A statistical paradox. *Biometrika, 44*(1/2), 187–92. https://doi.org/10.2307/2333251.

Raiffa, H., & Schlaifer. R. (1961). *Applied statistical decision theory*. Harvard University Press.

Rouder, J.N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4.

Rouder, J.N, Morey, R.D., Speckman, P.L., & Province, J.M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356–374.

Sanborn, A.N., & Hills, T.T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review, 21*(2), 283–300. https://doi.org/10.3758/s13423-013-0518-9.

Schönbrodt, F.D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*(2), 322–339. https://doi.org/10.1037/met0000061.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of P values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/bf03194105.

Yu, E.C., Sprenger, A.M., Thomas, R.P., & Dougherty, M.R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review, 21*(2), 268–282. https://doi.org/10.3758/s13423-013-0495-z.

Rianne de Heide werkt als universitair docent bij de afdeling wiskunde van de Vrije Universiteit Amsterdam. Ze is geïnteresseerd in de wiskundige en filosofische grondslagen van de statistiek en machine learning. Haar onderzoek betreft in het bijzonder (sequentieel) hypothesetoetsen, Bayesiaans leren, banditproblemen, en sinds kort ook explainable AI. Ze won in 2022 de Willem R. Van Zwet Award voor haar proefschrift *Bayesian learning: challenges, limitations and pragmatics*.
Website: https://riannedeheide.github.io
E-mail: r.de.heide@vu.nl

Gerrit Stemerdink

column

DEC-10 systeem

De Hoofdige Boer

'Elk weet, waar 't Almensch Kerkje staat, en kent de laan, die derwaart gaat.' Zo begint het gedicht van Staring over de Hoofdige Boer. De weg naar dat kerkje werd doorsneden door een modderige ondiepe voorde die men moest oversteken om ter kerke te kunnen gaan. Dat gaf vaak schade aan kleding en schoeisel, daarom werd op aandringen van de predikant een brug gebouwd. Iedereen blij, behalve Scholte Stuggink die het niet op die nieuwlichterij had voorzien. Hij bleef, voorzien van lieslaarzen, de oude doorsteek gebruiken. Hij was koppig, hoofdig in het Nederlands van de 19e eeuw, en bleef doen zoals zijn voorouders ook altijd hadden gedaan.

Aan die Hoofdige Boer moest ik denken toen ik problemen had met een softwareleverancier en koppig voet bij stuk hield. Tenslotte ben ik ook een Achterhoeker, net als Starings Scholte Stuggink.

Ik loop al heel lang mee in dit vak, en dan met name in de gegevensverwerking. Daar heeft zich in de loop der jaren een heilvolle standaardisatie voltrokken. Niet alleen heilvol omdat daarmee het gemak voor de gebruiker toeneemt, maar vooral omdat er fouten worden vermeden.

De rekencentra van Nederlandse universiteiten hadden begin jaren tachtig vorige eeuw verschillende merken en types computer in gebruik. Toen ik in 1982 bij het Rekencentrum van de toenmalige Landbouwhogeschool ging werken was daar een DEC-10 systeem in gebruik. Dat was in zoverre een afwijkend concept omdat men niet de gebruikelijke 32-bits woordlengte hanteerde, maar 36-bits. Dat had voordelen voor de rekennauwkeurigheid, van een getal konden er iets meer decimalen worden opgeslagen. Onder andere Tilburg, Twente en Rotterdam hadden dit ook, of de iets grotere versie DEC-20. Verder waren er 32-bits IBM-systemen in Delft, Nijmegen en Leiden en 60-bits CDC-systemen in Amsterdam, Utrecht en Groningen en nog wat minder gangbare merken. PC's waren er nog niet, die kwamen pas enkele jaren later.

Een programma dat ik graag in Wageningen wilde hebben was LISREL, voor Linear Structural Relations, geschreven door Karl Jöreskog. De distributie daarvan werd geregeld via een Amerikaans bedrijf. Kleinere organisaties hadden niet zelf de mogelijkheid een product over te zetten naar andere systemen, dat besteedden ze uit aan bijvoorbeeld een bevriende universiteit die over het gewenste computersysteem beschikte. Zo kwam de DEC-10 versie van LISREL via een Australische universiteit. Wat er precies is misgegaan met die conversie is me nooit duidelijk geworden, maar ik kreeg in Wageningen afwijkende uitkomsten van de bijgeleverde set testen. Op een dag probeerde ik het nog eens en plotseling was het wél goed. Na enige tijd van alles te hebben gecontroleerd en vergeleken ontdekte ik dat de juiste uitkomsten alleen werden geproduceerd als ik de geheugencapaciteit voor mijn job klein hield. Om sneller te kunnen werken met zo'n gecompliceerd programma had ik de eerste keer gekozen voor meer geheugen. Ik vond dat ik het programma op deze manier niet aan de gebruikers beschikbaar mocht stellen, het moest gewoon onder alle geheugenpartities de juiste uitkomsten geven. De Amerikaanse distributeur wist geen raad met het probleem, en de Australische universiteit was van mening dat het correct was, ik moest de gebruikers maar verplichten met weinig geheugen te werken. Ik heb het probleem toen voorgelegd aan Jöreskog die mij steunde in mijn besluit het programma in deze staat niet te accepteren. Gelukkig hadden we de licentie nog niet betaald, ik heb het contract *on hold* gezet totdat er een oplossing kwam.

Helaas bleef de distributeur me bestoken met herinneringsnota's en dreigde met advocaten. Op een gegeven ogenblik werd ik gebeld door juristen van het Ministerie van Landbouw, zij hadden van de Nederlandse ambassade in de VS het verzoek gekregen die vervelende man bij de Agricultural University tot enige spoed te manen. Ik voelde me in mijn koppigheid zo langzamerhand een echte Hoofdige Boer en heb ze uitgelegd dat ik niet van plan was ook maar iets te betalen voor een niet correct werkend product. Dat de oorspronkelijke maker van dat product me hierin steunde gaf de doorslag. Ik heb alle ontvangen materiaal zoals tapes en manuals naar het ministerie gestuurd die het in een *diplomatic pouch* naar Amerika verzond alwaar de Landbouwattaché het overhandigde aan de distributeur.

Weken commotie, ergernis en veel vergeefs werk, alleen omdat wij op een ander computersysteem werkten dan de oorspronkelijke maker. Gelukkig liggen die ruige pioniersjaren inmiddels achter ons.

Gerrit Stemerdink is eindredacteur van *STAtOR*.
E-mail: gjstemerdink@hotmail.com