Peter D. Grünwald
Centrum Wiskunde & Informatica, Amsterdam, en Mathematisch Insituut, Universiteit Leiden
peter.grunwald@cwi.nl

Onderzoek

Toetsen als gokken: een redelijk alternatief voor de p-waarde

De laatste tijd is er in de media veel aandacht geweest voor het feit dat veel wetenschappelijke resultaten niet reproduceerbaar zijn. Behalve bij de betreffende vakgebieden zit er volgens onderzoeker Peter Grünwald van het CWI en Universiteit Leiden ook een substantieel probleem bij de gebruike wiskunde: p-waarden doen maar zeer ten dele wat ze horen te doen. In dit artikel gaat hij in op bezwaren die kleven aan de standaardmethode voor statistisch onderzoek, en laat zien dat er een veelbelovend alternatief bestaat.

Er kleven een aantal fundamentele bezwaren aan de standaardmethode voor statistisch onderzoek, het p-waarde-gebaseerde nulhypothesetoetsen (pHT). Sommige van deze bezwaren zijn al sinds ongeveer 1960 bekend [5, 14] maar tegen die tijd was pHT al zo wijdverbreid geraakt in de toegepaste wetenschappen dat alle — herhaalde — pogingen tot verandering van paradigma op niets uitliepen. (Ik sta zelf keer op keer weer te kijken hoe weinig wetenschappers nu feitelijk van deze bezwaren afweten, en hoevelen ze afdoen als 'niet essentieel' of 'er bestaat nu eenmaal geen perfecte methode'.) Ik zal hier een aantal van deze problemen bespreken, en laten zien dat er wel degelijk een veelbelovend alternatief bestaat — de *toets-martingaalmethode*. Een toets-martingaal meet feitelijk de opbrengst van een *gokstrategie*. In standaard pHT staat een *kleine p-waarde* voor veel evidentie tegen de nulhypothese; in de toets-martingaalmethode staat een *grote (virtuele) financiële winst* voor evidentie tegen de nulhypothese. Dit idee werd geintroduceerd door Volodya Vovk [18] (toevallig — of niet — Kolmogorovs laatste promovendus), en bouwt zelf weer verder op het werk van, onder anderen, de vroeg gestorven J. Kiefer (1924–1981) [3], het werk van Wald en vooral Robbins op het gebied van sequentieel toetsen (Lai [12] geeft een mooi overzicht) en de klassieker van Ville [17], een van de eerste en meest invloedrijke artikelen over het gebruik van 'martingalen' in de kansrekening.

Mede door de zeer theoretische insteek van Vovk heeft de toets-martingaalmethode tot voor kort een slapend bestaan geleid. Om de methode praktisch toe te passen moesten nog veel details uitgewerkt worden. Sinds 2010 is hier een begin mee gemaakt, onder anderen door Vovk zelf [16] en ook — samen met co-auteurs — door mijzelf, als onderdeel van mijn Vici-project 'Safe Statistics' [6, 13]. De methode is nu klaar om in statistische software verwerkt te worden — iets waar ik mij de komende jaren op zal richten. De toets-martingaalmethode is verwant aan de Bayesiaanse statistiek, maar verschilt op een aantal cruciale punten, waar ik ook kort op in zal gaan.

Dit stuk is een wiskundige uitwerking en uitbreiding van een eerder stuk dat verschenen is in *StatOr* [9], en een deel van de tekst is overgenomen van het *StatOr*-artikel. Het *StatOr*-artikel was op zijn beurt gebaseerd op de lezing 'Paranormale Statistiek', die ik in 2015 gaf op de jaarlijkse Nederlandse Wiskunde Dagen.

Motivatie: de reproducibility crisis

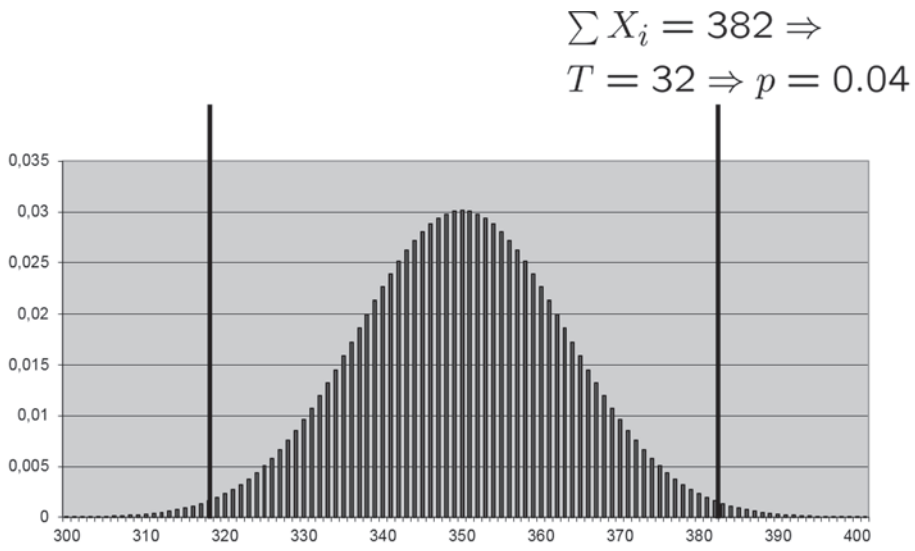
Het tijdschrift *Science* heeft er een speciale thema-uitgave aan gewijd en ook *The Economist* vond het een hoofdartikel waard: een schrikbarend hoog percenta-

ge wetenschappelijke resultaten is niet reproduceerbaar. Met name in de geneeskunde en psychologie wordt onderkend dat er sprake is van een *reproducibility crisis*. In de geneeskunde is dit al een aantal jaren duidelijk, met name door het werk van Ioannidis [11] ('Why most published research findings are false'). Voor de psychologie werd het afgelopen augustus nog eens bevestigd, alweer in *Science*: een grote groep onderzoekers probeerde een groot aantal psychologische studies zo zuiver mogelijk te reproduceren; zoals in vrijwel alle Nederlandse kranten stond te lezen bleek minder dan de helft van de onderzoeken reproduceerbaar [1].

De twee belangrijkste redenen voor de reproducibility crisis zijn wellicht publicatiebias en het feit dat verschillende populaties en experimentele condities vaak niet met elkaar vergelijkbaar zijn. (In plaats van uit te leggen wat publicatiebias is, verwijs ik liever naar de cartoon https://xkcd.com/882 van het onvolprezen xkcd.com: een plaatje zegt hier meer dan duizend woorden.) Zo zijn de proefpersonen in psychologische experimenten meestal studenten psychologie, die zelfs al het een en ander weten over hoe psychologische experimenten in elkaar zitten — wat voor die studenten geldt, hoeft zeker niet voor de populatie als geheel te gelden. Maar er zit ook een substantieel (en vaak onderschat) probleem bij de wiskunde zelf — en daar gaat het in dit artikel over. p-waarden doen namelijk maar zeer ten dele wat ze horen te doen. Ik zal nu eerst uitleggen wat p-waarden zijn, alvorens de problemen te bespreken.

Definitie van de p-waarde

Stel we observeren data $X^N \equiv X_1,\ldots,X_N$. We hebben twee mogelijke verklaringen voor deze data: de nulhypothese H_0, die de status quo representeert, en daarnaast de alternatieve hypothese H_1. In medische toetsen staat H_0 vaak voor 'nieuw geneesmiddel werkt niet', en H_1 voor 'geneesmiddel is werkzaam'. Zowel H_0 als H_1 worden gerepresenteerd als verzamelingen van kansverdelingen. Voor het gemak beperken we ons in deze opfrisser tot het geval dat $H_0 = \{P_0\}$ *enkelvoudig* is, dat wil zeggen slechts een enkele verdeling P_0 bevat. Een p-waarde is een functie p die de verzameling van mogelijke uitkomsten afbeeldt op $[0,1]$, en waarvoor geldt dat



De p-waarde voor het Bernoulli-voorbeeld, bij $N = 700$ met toetsingsgrootheid $T = |\sum X_i - N/2|$. De grafiek laat voor elke waarde van $\sum X_i$ zijn hoe groot de kans erop is onder H_0. We observeren 382 enen, hetgeen overeenkomt met een p-waarde van $0{,}04$: dit is de kans dat het aantal enen $T = 32$ of nog méér afwijkt van het verwachte aantal 350. Deze kans is gelijk aan de 'oppervlakte' van de grafiek buiten de meest linker- en rechterstreep. Als we precies 350 $(T = 0)$ enen zouden observeren is de p-waarde 1; de twee strepen staan bij 350 en vallen dan over elkaar, de oppervlakte buiten de strepen is dan 1. Hoe groter de geobserveerde T, hoe verder de strepen van het midden verhuizen. We verwerpen de nulhypothese als T zo groot is dat de bijbehorende p-waarde kleiner of gelijk is aan het van te voren gekozen significance level α. Bijvoorbeeld, als we $\alpha = 0{,}1$ hadden gekozen, verwerpen we de nulhypothese zodra we $|T| \geq 23$ observeren, omdat $|T| > 23$ equivalent is met $\mathrm{p} \leq 0{,}1$

Voor elke $0 \leq \alpha \leq 1$: $P_0\big(\mathrm{p}(X^N) \leq \alpha\big) \leq \alpha$. (1)

De meer gangbare indirecte definitie (bijvoorbeeld op Wikipedia) komt neer op de striktere eis dat (1) met gelijkheid geldt, dus $P_0\big(\mathrm{p}(X^N) \leq \alpha\big) = \alpha$, maar dit is eigenlijk niet noodzakelijk voor de klassieke toetsingstheorie. Om de definitie verder toe te lichten gebruiken we het *Bernoulli-voorbeeld* waarbij de $X_i \in \{0,1\}$ binair zijn, de nulhypothese H_0 zegt dat X_i onafhankelijk Bernoulli$(\frac{1}{2})$-verdeeld zijn, dus uitkomsten van een eerlijke munt, en de alternatieve hypothese $H_1 = \{P_{1,\theta} \mid \theta \neq 1/2\}$ zegt dat de munt niet eerlijk is.

p-waarden kunnen gedefinieerd worden aan de hand van verschillende statistieken van de data, de zogenaamde toetsingsgrootheden. Laten we eerst even aannemen dat het aantal datapunten N vast ligt. In het Bernoulli-voorbeeld kunnen we bijvoorbeeld als toetsingsgrootheid $T(X^N) = |\sum_{i=1}^{N} X_i - N/2|$, de discrepantie tussen het daadwerkelijke aantal enen en het verwachte aantal enen, nemen. We zetten $\alpha_t := P_0(T(X^N) \geq t)$, en we definiëren vervolgens $\mathrm{p}(X^N) := \alpha_{T(X^N)}$ als de kans op een uitkomst die minstens zo extreem is als de uitkomst die we daadwerkelijk hebben waargenomen. Aangezien voor elke t geldt $\{X^N : T(X^N) \geq t\} = \{X^N : \alpha_{T(X^N)} \leq \alpha_t\}$ volgt dan ook

$$P_0(\mathrm{p}(X^N) \leq \alpha_t) = P_0(\alpha_{T(X^N)} \leq \alpha_t) = P_0(T(X^N) \geq t) = \alpha_t,$$

zodat (1) geldt met gelijkheid voor alle α die gelijk zijn aan α_t voor een t in het bereik van T; dan is eenvoudig in te zien dat (1) zelf (met ongelijkheid) geldt voor alle $0 \leq \alpha \leq 1$, zodat $\mathrm{p}(X^N)$ een valide p-waarde is. Met behulp van een toetsingsgrootheid T kunnen we dus een p-waarde definiëren, en om de toets compleet te maken hebben we nu nog een *onbetrouwbaarheidsdrempel* oftewel *significance level* α nodig — vaak wordt deze op $0{,}05$ gezet. We 'verwerpen' dan de nulhypothese als de waargenomen p-waarde $\mathrm{p}(X^N) \leq \alpha$. De rationale hierachter is dat met deze procedure de kans op een *Type-I fout* kleiner of gelijk is aan α:

$$P_0(\text{'Ik verwerp } H_0\text{'}) \leq \alpha. \qquad (2)$$

Dit wil zeggen dat, als we een leven lang hypothesen toetsen en steeds dezelfde drempel α aanhouden, we *zelfs* in het geval dat de nulhypothese *altijd* waar is, op de lange termijn maar een fractie α van de keren de nulhypothese verwerpen. Mochten — zoals hopelijk het geval is — de nulhypothesen in sommige experimenten niet waar zijn, dan is de fractie van de toetsen waarin de nulhypothese wel waar is, maar we die toch verwerpen, zelfs kleiner dan α.

Na deze opfrisser kunnen we nu een aantal van de problemen met p-waarden de revue laten passeren.

Probleem 1: *Beperkte toepasbaarheid, wegens onbekende kansen en stopregel.* Om p-waarden te gebruiken moet je een toetsingsgrootheid T hebben zodat je de kans $P_0(T(X^N) \geq t)$ kunt uitrekenen. Je kunt pHT niet gebruiken in vele eenvoudige scenario's waarbij toetsen intuïtief heel wel mogelijk is, maar zo'n toetsingsgrootheid niet valt te definiëren. Een simpel voorbeeld is weersvoorspelling: tot een aantal jaren geleden gaven de weerman op RTL 4 en de weervrouw op NOS elke dag een 'kans dat het morgen regent'. Je zou willen toetsen wie van de twee dit beter doet. Als de RTL-man consequent hoge regenkans geeft voor dagen waarop de zon blijkt te schijnen en de NOS-vrouw dat niet doet, zou je willen concluderen dat de NOS-vrouw beter is. Het is echter onmogelijk om een p-waarde te berekenen onder de nulhypothese 'ze zijn even goed'. Daarvoor moet je namelijk niet alleen weten hoe goed de voorspellers het doen op de data die echt hebben plaatsgevonden, *maar ook hoe goed ze het zouden hebben gedaan in situaties die zich niet hebben voorgedaan!* Je moet dus weten wat de RTL-man voor kans zou voorspellen voor vandaag als het eergisteren geregend zou hebben, ook al scheen in werkelijkheid eergisteren de zon. Dit soort 'counterfactual' voorspellingen zijn intuïtief irrelevant (en niet benodigd in de toets-martingaalmethode), maar ze zijn nodig om p-waarden te berekenen [4, 8]. Een gerelateerd probleem is dat, om valide p-waarden uit te kunnen rekenen, het experimentele protocol ofwel de *stopregel* volledig bekend moet zijn bij begin van het experiment. Voorbeelden van zo'n protocol in het Bernoulli-voorbeeld zijn 'zet N vast op 100' of 'ga door met data verzamelen totdat je voor het eerst achterelkaar drie enen hebt gezien' (dan is N zelf een stochast). In de praktijk weten we vaak niet van tevoren wat de stopregel is, of willen we het protocol aanpassen aan de hand van onvoorziene omstandigheden; met standaard pHT mag dat niet — terwijl het in de praktijk voortdurend gebeurt. Het mooie van de toets-martingaalmethode is dat ze resultaten oplevert die robuust zijn — ze blijven geldig, ook al wordt het protocol 'onderweg' aangepast (dit punt, en nog enkele kritiekpunten waar ik hier niet eens op inga, worden in detail uitgewerkt in [9]).

Probleem 2: *Interpretatie.* Stel een nieuw geneesmiddel is in twee ziekenhuizen getoetst — de nulhypothese is, zoals altijd, dat het geneesmiddel niet beter werkt dan het placebo. Ziekenhuis A rapporteert p-waarde p_A en ziekenhuis B rapporteert p-waarde p_B. Natuurlijk zijn de twee patiëntenpopulaties verschillend, maar we zouden toch graag een allereerste indicatie willen krijgen van de gezamenlijke bewijskracht van de twee experimenten. Hoe doen we dit? Het meest voor de hand ligt de p-waarden te vermenigvuldigen, maar zoals Fisher [7] al aangaf is dit niet goed: omdat p-waarden kleiner dan 1 zijn wordt het resultaat altijd kleiner, wat de oorspronkelijke p-waarden ook waren. In werkelijkheid moet je een correctie toepassen. Er zijn meerdere 'correcte' correcties, die echter allemaal een ander antwoord geven. Welke correctie moet je gebruiken? Zou het niet fijner zijn om een methode te hebben waarbij er een uniek-optimale manier is om resultaten te combineren? De toets-martingaal biedt zo'n manier — je kunt gewoon vermenigvuldigen.

Ook is het niet duidelijk wat het precies betekent als je een p-waarde observeert die veel kleiner is dan de van tevoren gekozen drempel α, zeg $\alpha = 0{,}05$. Het significantieconcept is bedacht door Neyman en Pearson, en als je hun werk letterlijk neemt, impliceert het dat, als je een p-waarde observeert die veel kleiner is dan α, je die informatie eigenlijk weg kunt gooien — je dient alleen maar te rapporteren 'nulhypothese verworpen', dus '$p \leq \alpha$', ook al is de waargenomen p kleiner dan $0{,}001$. Het probleem is namelijk dat je zo'n kleinere p niet kunt vertalen naar een kleinere Type-I-fout. Die is namelijk alleen gedefinieerd als we de verzameling uitkomsten in precies twee delen splitsen: die met $p \leq \alpha$ (significant) en die met $p > \alpha$. Maar nu willen we de verzameling uitkomsten bijvoorbeeld in minstens drie verzamelingen opsplitsen ($p > 0{,}05$, $0{,}001 < p \leq 0{,}05$, $p \leq 0{,}001$). Dan is strikt genomen de kans op een Type-I-fout (2) niet eens meer gedefinieerd; en als we hem wel formeel zouden willen definieren ontstaan er grote problemen [2, 20]. We zouden bijvoorbeeld een conditionele uitspraak willen doen als '*gegeven* dat ik een toets-resultaat krijg met $p \leq 0{,}001$ en daarom besluit de nulhypothese te verwerpen, is mijn Type-I-fout op zijn hoogst $0{,}001$'. Zo'n uitspraak is echter onjuist: als we het extreme geval onder (2) bekijken waarin we ons hele leven lang hypothesen toetsen en de nulhypothesen altijd waar zijn, zullen we toch 1 op de 1000 keer $p \leq 0{,}001$ waarnemen en de nulhypothese verwerpen, en *gegeven* dat we $p \leq 0{,}001$ waarnemen, doen we dan met zekerheid een foute uitspraak, en is onze Type-I-fout dus 1!

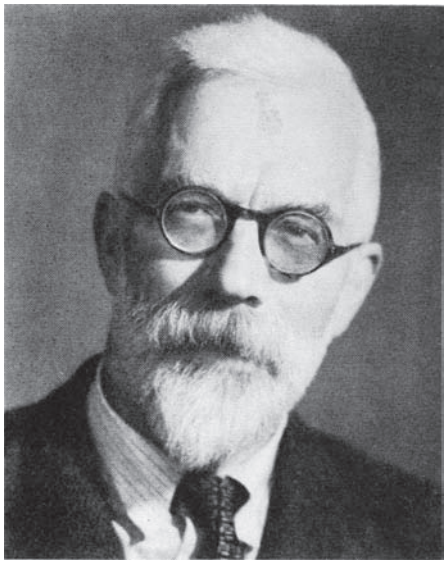


Foto: flickr.com, W.I.B. Beveridge

Sir Ronald Fisher (links) en Jerzy Neyman (rechts), twee van de grootste statistici die ooit geleefd hebben, in de jaren dertig beiden werkzaam aan University College London. Er wordt vaak gesproken over de controverse tussen de twee grote scholen binnen de statistiek, de Bayesiaanse en de frequentistische. Maar hoewel Neyman en Fisher beiden 'frequentistisch' waren, waren ze het volstrekt oneens over hypothesetoetsen — hetgeen het des te merkwaardiger maakt dat de huidige praktijk feitelijk een amalgaam is van de door hen voorgestelde methodes.

Toch is het duidelijk dat een kleinere p wel *iets* zegt over extra bewijsmateriaal tegen de nulhypothese, en dus is het begrijpelijk dat de gewoonte is ontstaan om die kleinere p toch te rapporteren. Aldus is de standaard hypothesetoets feitelijk een merkwaardig amalgaam van twee methodes: Fisher, wellicht de grootste statisticus die ooit geleefd heeft, stelde voor om p-waardes te rapporteren als mate van bewijskracht, maar moest niets hebben van significance levels en Type-I-fouten. Neyman, ook een van de grootste statistici, ging het om garanties op Type-I-fouten, en zag de daadwerkelijke p-waarde als minder belangrijk — zie Figuur 2. De huidige praktijk suggereert dat een kleinere p-waarde een kleinere Type-I fout impliceert, maar we hebben al gezien dat dat niet zo is. Maar wat zegt die kleinere p dan? Dat is nog niet zo eenvoudig!

Probleem 3: *nog meer interpretatie — de p-waarde versus de posterior.* In bovenstaand kader staat een eenvoudige vraag over p-waardes die aan Amerikaanse artsen werd voorgelegd. Er waren 397 respondenten. Van hen koos 15 procent antwoord 1, 19 procent antwoord 2 (correct), 52 procent antwoord 3 en 15 procent antwoord 4. Dat betekent dat meer dan de helft van de artsen de zogenaamde *prosecutor's fallacy* (antwoord 3) begaat: zij verwarren de kans op de (nul-)hypothese gegeven de data met de kans op de data gegeven de hypothese. (De prosecutor's fallacy heet niet voor niets zo — zij kwam en komt nog steeds ook voor in de rechtszaal, onder andere in de geruchtmakende zaak tegen Lucia de Berk.) Een p-waarde $< 0{,}05$ betekent dat die laatste kans kleiner is dan $0{,}05$; over de eerste kans kun je binnen het pHT-paradigma niets zeggen. (Preciezer: er is voordat het experiment een verzameling $\mathcal{R}$ van uitkomsten vastgesteld met kans $< 0{,}05$ onder de nulhypothese, en $p < 0{,}05$ betekent dat de daadwerkelijke uitkomst in $\mathcal{R}$ viel.) Binnen het pHT-paradigma wordt de 'echte' toestand van de wereld namelijk als een onbekend maar vast, niet-random gegeven gezien: ofwel H_0 is waar, ofwel H_1 is waar, maar je kunt geen kansverdeling op H_0 en H_1 leggen. Een Bayesiaans statisticus is wel bereid dit te doen; het ligt dan voor de hand om bijvoorbeeld te stellen dat *a priori* $p(H_0) = p(H_1) = \frac{1}{2}$ (hier en in het vervolg gebruiken we kleine letters voor kansfuncties, en hoofdletters voor kansmaten). Zij kan dan met de stelling van Bayes de *a posteriori kans* $p(H_0 \mid X^N)$ op H_0 gegeven de data bepalen, als

$$p(H_0|X^N) = \frac{p(X^N|H_0)p(H_0)}{p(X^N|H_0)p(H_0)+p(X^N|H_1)p(H_1)}. \quad (3)$$

waarbij $p(X^N \mid H_0) := p_0(X^N)$ de kans is op data X^N onder de nulhypothese. In ons Bernoulli-voorbeeld (eerlijke muntjes) hebben we dus $p(X^n \mid H_0) = 2^{-N}$. De kans $p(X^N|H_1) := \int_{\theta\in\Theta} p_\theta(X^n|H_1)\,\pi(\theta)\,d\theta$ wordt gedefinieerd als de *marginale* kans op X^N volgens P_θ, waarbij θ zelf ook weer beschouwd wordt als verdeeld volgens een *a priori* kansdichtheidsfunctie $\pi(\theta)$. Bayes zelf en kort daarna Laplace namen voor onze Bernoulli H_1 de homogene verdeling $\pi(\theta) \equiv 1$, zodat

$$p(X^N \mid H_1) = \int_0^1 p_\theta(X^N)\,d\theta = \int_0^1 \theta^{\Sigma_{i=1}^n X_i}(1-\theta)^{N-\Sigma_{i=1}^N X_i} d\theta.$$

Er zijn binnen de statistiek hele veldslagen gevoerd over de merites van de Bayesiaanse aanpak; Fisher en Neyman wezen hem categoraal af. Maar zelfs als je sympathie hebt voor hun kritiek, is het toch zinvol om eens te kijken wat er zou gebeuren in een *geïdealiseerde* situatie waarin de natuur H_0 wel degelijk kiest door een eerlijk muntje te gooien. Je vindt dan dat bij data die een p-waarde van $0{,}05$ opleveren, de posterior kans van H_0 in de meeste gevallen vele malen *groter* is dan $\frac{1}{20}$ — hij kan in feite willekeurig dicht bij 1 liggen. (Dit probleem zou veel minder erg zijn als er een vaste omrekenfactor van p-waardes naar posterior kansen zou bestaan. Maar die bestaat helaas niet — de omrekenfactor hangt af van de modellen, de toetsingsgrootheid, de hoeveelheid geobserveerde data N en de (stop-) regel waarmee N werd bepaald.) Dit geldt ook als we een andere gladde prior op p leggen die nergens 0 is — de precieze keuze van zo'n verdeling doet er heel weinig toe. Aangezien de meeste artsen (en wetenschappers!) p-waarden toch eerder op de Bayesiaanse manier interpreteren, is dit erg zorgelijk: $p < 0{,}05$ zegt, in termen van de kans op de nulhypothese gegeven de data, veel minder dan de meeste mensen (althans artsen) denken.

Dit interpretatieprobleem wordt vaak weggewuifd — statistiek is nou eenmaal moeilijk, we kunnen niet verwachten dat artsen (of rechters) het begrijpen. Probleem is dat het in ieder geval bij artsen wel gaat om mensen die geacht worden vakliteratuur te lezen, die bol staat met resultaten die significant zijn op '$p < 0{,}05$' of '$p < 0{,}01$'. Bovendien: onder het motto *What do Math Teachers know about Statistics* heb ik dezelfde vraag als hierboven ook aan mijn (700-koppige) publiek gesteld tijdens een voordracht op de *Nederlandse Wiskunde Dagen*. Hoewel daar slechts 20 procent antwoord 3 gaf, gaf een schrikbarend percentage van 40 procent antwoord 4. Blijkbaar vinden wiskundedocenten het ook nog heel moeilijk. Samenvattend hebben we dus een methode, ontwikkeld in termen van Type-I-fouten, die geen goede Type-I-fout-interpretatie heeft (Probleem 2 hierboven), en ook geen goede Bayesiaanse interpretatie (Probleem 3).

Valt een veel algemener toepasbare methode die een veel concretere interpretatie heeft dan niet toch te prefereren? Ik

Toets uw eigen interpretatie van p-waarden!

Het geruchtmakende artikel 'What Do Doctors Know about Statistics?' [21] beschrijft de resultaten van een enquête onder Amerikaanse artsen, waarbij een aantal basale vragen over statistiek werd gesteld. Een van de vragen was: *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 (dus er is minstens 95% kans dat het beter werkt).
4 Geen idee.

Zie de hoofdtekst voor het juiste antwoord, en de ontluisterende resultaten onder artsen en wiskundeleraren.

denk het wel — zowel Bayesiaans toetsen als de toets-martingaal zijn zo'n methode. Er is dan ook — bijvoorbeeld in de psychologie [19] — een soort Bayesiaanse revolutie gaande. Ik denk dat dit een stap in de goede richting is, maar dat de toets-martingaalmethode uiteindelijk meer mogelijkheden en minder interpretatieproblemen biedt — ik kom hierop terug aan het eind van dit artikel.

Een redelijk alternatief: de toets-martingaal

Voordat we de algemene definitie geven lichten we het idee van de toets-martingaal toe aan de hand van het Bernoulli-voorbeeld: stel dat we mogen *gokken* op de uitkomsten $X_1, X_2, \ldots$ op de volgende manier: we hebben een bepaald beginkapitaal K_0 en mogen dat vrijelijk verdelen over uitkomst '$X_1 = 0$' en '$X_1 = 1$'. Nadat X_1 gerealiseerd is, zeg $X_1 = x$, krijgen we twee keer onze inzet op x uitbetaald; de inzet op de andere uitkomst zijn we kwijt. Dus als we bijv. een fractie $q(1)$ van ons kapitaal op $X_1 = 1$ inzetten en een fractie $q(0) := 1 - q(1)$ op $X_1 = 0$, dan is ons nieuwe kapitaal K_1 na ronde 1 gelijk aan $K_1 = 2q(X_1)K_0$. We mogen nu het nieuwe kapitaal K_1 vrijelijk verdelen over de twee mogelijke uitkomsten van X_2 en krijgen opnieuw twee keer onze inzet op de echte uitkomst uitbetaald, terwijl we de inzet op de andere uitkomst verliezen. Als we bijvoorbeeld een fractie $q'(1)$ inzetten op $X_2 = 1$ en $q'(0) := 1 - q'(1)$ op $X_2 = 0$, dan is ons eindkapitaal na ronde 2 gelijk aan $K_2 := 2q'(X_2)K_1 = 4q'(X_2)q(X_1)K_0$. Zo gaat het spel door: we kunnen steeds K_j herverdelen over uitkomst X_{j+1}, en krijgen steeds twee keer onze inzet op de daadwerkelijke uitkomst terug, en dit gaat zo door tot aan ronde N als er geen data meer is, en we eindigen met eindkapitaal K_N. Deze manier van sequentieel-gokken-met-herinzet heet *Kelly gambling* in de economische literatuur. Een *gokstrategie* is een functie die elk initieel segment $X^t \equiv X_1, \ldots, X_t$ van elke lengte $t \geq 0$ afbeeldt op een reëel getal $q(1 \mid X^t) \in [0,1]$ dat aangeeft, gegeven verleden $X_1, \ldots, X_t$, welke proportie van ons tot nog toe verzamelde kapitaal K_t we inzetten op uitkomst $X_{t+1} = 1$. We schrijven dus voortaan $q(X_2 \mid X^1)$ in plaats van $q'(X_2)$, om duidelijk te maken dat de strategie die we hanteren af mag hangen van het verleden. Ons eindkapitaal bij data X^N wordt dan gegeven door $K_N = M(X^N) \cdot K_0$, waarbij

$$M(X^N) := 2^N \prod_{t=1}^{N} q(X_t \mid X^{t-1}) \qquad (4)$$

We noemen de functie M, gedefinieerd op datasequenties van willekeurige lengte, de *toets-martingaal behorend bij gokstrategie r*.

We merken eerst op dat de uitbetaling (een factor 2 van de inzet) neerkomt op een *eerlijke* gok *als* de nulhypothese waar is. Immers, bij deze uitbetaling geldt dat, wat voor gokstrategie we ook hanteren, ons *verwacht* eindkapitaal onder de nulhypothese nooit groter is dan ons beginkapitaal. Als we met een bepaalde gokstrategie juist *wel* veel geld winnen, is dat dus een indicatie dat de nulhypothese onjuist is. Dit is in het kort waar de martingaalmethode op neerkomt: bij klassiek toetsen leggen we een *test statistic* (toetsingsgrootheid) vast, die een p-waarde bepaalt; hoe kleiner de p-waarde, hoe groter de indicatie dat de nulhypothese onjuist is. In de martingaalmethode leggen we een *gokstrategie* vast; hoe meer geld we daarmee winnen, hoe groter de indicatie dat de nulhypothese onjuist is.

In het Bernoulli-voorbeeld zouden we bijvoorbeeld op tijdstip t kunnen kijken naar de waargenomen frequentie van enen tot nu toe, $\hat{\theta}_t := \sum_{i=1}^{t} X_i / t$. Als dit sterk afwijkt van $\frac{1}{2}$ is dit, intuïtief, een indicatie dat de nulhypothese onjuist is en dat er meer winst te behalen valt als we een proportie van ongeveer $\hat{\theta}_t$ van ons geld op 1 zouden zetten. Nu is het gevaarlijk om *precies* $\hat{\theta}_t$ in te zetten: als $\hat{\theta}_t$ gelijk is aan 0 (alleen nullen gezien) of 1 (alleen enen gezien) zouden we *al* ons geld op 0 respectievelijk 1 zetten, en dus met een bepaalde kans ook al ons geld verliezen. We kunnen ons indekken voor dit risico door, net iets minder agressief, op tijdstip t een proportie van $\breve{\theta}_t = \sum_{i=1}^{t} (X_i + 1) / (t + 2)$ op uitkomst 1 in te zetten. Dit blijkt een zeer effectieve gokstrategie te zijn onder het alternatief H_1: als deze alternatieve hypothese juist is, winnen we met bovenstaande gokstrategie exponentieel veel geld. Meer specifiek, is het eenvoudig om aan te tonen dat ons verzamelde kapitaal K_N op tijdstip N gegarandeerd tenminste

$$M(X^N) = (N+1)^{-1} \exp(2N(\hat{\theta}_N - \tfrac{1}{2})^2) \qquad (5)$$

keer K_0 bedraagt [8]. Volgens de wet van de grote aantallen zal $\hat{\theta}_N$ naar θ convergeren, zodat (5) exponentieel stijgt als $\theta \neq \frac{1}{2}$. De ongelijkheid van Hoeffding impliceert zelfs dat voor elke vaste $\epsilon > 0$, de kans dat $P_\theta(|\hat{\theta}_N - \theta| > \epsilon)$ exponentieel klein is in N; als $\theta \neq \frac{1}{2}$ maken we dus exponentieel veel winst met kans vrijwel 1.

Opmerkingen

We mogen dus gokstrategieën gebruiken waarbij onze inzet op tijdstip t van het verleden $X_1, \ldots, X_{t-1}$ afhangt; dit is zelfs cruciaal om winst te kunnen maken als de nulhypothese niet klopt. Maar de gokstrategie *zelf* mag niet afhangen van de data; deze moet feitelijk vaststaan voordat we de data gezien hebben. Als we de gokstrategie namelijk achteraf bepalen kunnen we altijd de triviale strategie gebruiken die al het geld op tijdstip t inzet op de daadwerkelijke uitkomst X_{t+1}; dat zou natuurlijk op bedrog neerkomen. *Wel* kunnen we vrijelijk gokstrategieën combineren. In het bovenstaande geval zou bijvoorbeeld scepticus 1 kunnen denken dat de data weliswaar onafhankelijk Bernoulli θ zijn, maar dat $\theta \neq \frac{1}{2}$; scepticus 2 zou kunnen denken dat elke X_t weliswaar marginale kans $\frac{1}{2}$ heeft, maar niet onafhankelijk is van X_{t-1}; dit zou getest kunnen worden door een gokstrategie die de hoeveelheid geld ingezet op 1 afhankelijk maakt van wat er op tijdstip $t-1$ is gebeurd. Als we nu denken dat scepticus 1 of scepticus 2 weleens gelijk zou kunnen hebben, maar we weten niet wie, dan kunnen we een *nieuwe* gokstrategie maken, waarbij we de gokstrategie van scepticus 1 en 2 combineren: we investeren 50 procent van ons beginkapitaal K_0 in de strategie van scepticus 1, en 50 procent in de strategie van scepticus 2. Vervolgens laten we beide sceptici hun strategieën spelen. Als een van de twee eindkapitaal K_N behaalt, behalen wij minstens eindkapitaal $K_N/2$. Formeel: als M en M' twee toets-martingalen zijn, dan is $\frac{1}{2}M + \frac{1}{2}M'$ er ook een. Als een van de twee sceptici 'gelijk' heeft, zal zijn kapitaal exponentieel groeien, en is die factor twee al snel verwaarloosbaar: we doen het dus 'bijna' zo goed als de beste van de twee. Het is lang niet zo eenvoudig om twee p-waardetoetsen te combineren.

Verder is van belang dat, als we deze methode gebruiken voor een daadwerkelijke hypothesetoets, we niet *echt* hoeven te gokken en we dus ook niet *echt* iemand hoeven te vinden die bereid is onze inzetten te accepteren en eventuele winsten uit te betalen:

Het gaat om een puur virtueel spel, waarbij we kijken *hoeveel winst we* zouden *maken als we volgens een bepaalde gokstrategie* zouden *gokken, onder uitbetalingen die eerlijk* zouden *zijn als de nulhypothese waar was*.

Formele definitie

Stel H_0 en H_1 zijn allebei verzamelingen van kansverdelingen over X^∞. De nulhypothese is dat de data $X_1,\ldots,X_N$ verdeeld zijn volgens een $P \in H_0$. Voor het gemak nemen we even aan dat alle uitkomsten X_t in een eindige verzameling $\mathcal{X} = \{0,1,\ldots,k\}$ vallen; uitbreiding naar het aftelbaar oneindige en continue geval is echter eenvoudig. Een toets-martingaal wordt gedefinieerd door twee componenten: een gokstrategie en een eerlijk uitbetalingsproces.

Een *gokstrategie* is een functie die, voor elk tijdstip t, gegeven elk *verleden* $X^{t-1} \equiv (X_1,\ldots,X_{t-1})$ een vector $v_0, v_1, \ldots, v_{k-1}$ bepaalt, waarbij v_j de fractie van het kapitaal, verzameld tot aan tijdstip $t-1$ is, dat ingezet zal worden op de uitkomst $X_t = j$. Als de daadwerkelijke uitkomst $X_t = j'$ is, dan schrijven we $v_{j'}$ als $q(X_t \mid X^{t-1})$. Elke functie $q : \bigcup_{t \geq 0} \mathcal{X}^t \to \mathbb{R}^k$ waarbij alle componenten van $q(\cdot \mid X^t)$ niet-negatief zijn en voor alle t, alle mogelijke realisaties van X^{t-1}, geldt dat $\sum_{x \in \mathcal{X}} q(x \mid X^{t-1}) = 1$, is een geldige gokstrategie. In ons Bernoulli-voorbeeld gebruikten we bijvoorbeeld $q(1 \mid X^t) = (\sum_{i=1}^{t} X_i + 1)/(t+2)$.

Een *uitbetalingsproces* r is een functie die, voor elk tijdstip t, gegeven elk verleden $X^{t-1} \equiv (X_1,\ldots,X_{t-1})$ een vector $w_0, w_1, \ldots, w_{k-1}$ bepaalt, waarbij w_j de factor weergeeft waarmee de inzet op uitkomst j vermenigvuldigd wordt als $X_t = j$ daadwerkelijk de uitkomst is. Als de daadwerkelijke uitkomst j' is, dan schrijven we $w_{j'}$ als $r(X_t \mid X^{t-1})$. In ons Bernoulli-voorbeeld was $r(1 \mid X^{t-1}) = r(0 \mid X^{t-1})$ voor alle mogelijke realisaties van X^{t-1}. In het algemeen geldt dat als een fractie $q(X_t \mid X^{t-1})$ van het verzamelde kapitaal op tijdstip $t-1$ ingezet wordt op uitkomst X_t, dan is het kapitaal na tijdstip t dus gelijk aan $q(X_t \mid X^{t-1}) r(X_t \mid X^{t-1})$.

Eerlijke uitbetalingsprocessen: we komen nu bij het cruciale punt, namelijk dat we willen kijken naar weddenschappen die eerlijk (geen verwachte winst of verlies) zouden zijn onder H_0. We zullen dit alleen definiëren voor het geval dat onder alle verdelingen in H_0 de data onafhankelijk zijn; uitbreiding naar het algemene geval is zeer eenvoudig. In het geval dat de data onder H_0 onafhankelijk zijn, zeggen we dat een uitbetalingsproces *eerlijk* is *onder* H_0 *voor gegeven gokstrategie* q als voor elk tijdstip t, elk verleden $X_1,\ldots,X^{t-1}$ geldt:

Voor alle $P \in H_0$:

$$\mathbf{E}_{X_t \sim P}\big(q(X_t \mid X^{t-1}) \cdot r(X_t \mid X^{t-1})\big) = 1. \quad (6)$$

Dit betekent dat het *verwachte* eindkapitaal $\mathbf{E}[K_t \mid X^{t-1}]$ na uitkomst X_t, gegeven uitkomsten X^{t-1}, altijd gelijk is aan het gerealiseerde kapitaal K_{t-1} na uitkomsten X^{t-1}. 'Eerlijk' betekent dus dat je in verwachting geen winst maakt. Verdere intuïtie achter deze definitie is het eenvoudigst in het geval dat H_0 enkelvoudig is, $H_0 = \{P_0\}$. In dat geval bestaat er een uitbetalingsproces r' dat eerlijk is onder H_0 voor *willekeurige* q. Het is daarmee ook eerlijk voor q die zodanig zijn dat er, voor elke realisatie X^{t-1}, een $x \in \mathcal{X}$ bestaat met $q(x \mid X^{t-1}) = 1$ en voor $x' \neq x$, $q(x' \mid X^{t-1}) = 0$, i.e. op tijdstip t wordt al het geldt op x gezet. Dan kan (6) herschreven worden als:

$$\begin{aligned} &\text{Voor alle } x \in \mathcal{X} \text{ met } P_0(X_t = x) > 0: \\ &P_0(X_t = x) \cdot r'(x \mid X^{t-1}) = 1, \\ &\text{i.e. } r'(x \mid X^{t-1}) = \frac{1}{p_0(x)}. \end{aligned} \quad (7)$$

De uitbetaling bij een uitkomst moet dus omgekeerd evenredig zijn met de kans op die uitkomst. In ons Bernoulli-voorbeeld betekent dit dus dat, bij beide uitkomsten, de inzet op de gerealiseerde uitkomst wordt verdubbeld.

De *toets-martingaal* behorende bij gokstrategie q en eerlijk (voor q) opbrengstproces r, is nu het proces $M^1, M^2, \ldots$ waarbij $M^n \equiv M(X^n)$ gedefinieerd is als

$$M(X^n) := \prod_{t=1}^{n} q(X_t \mid X^{t-1}) \cdot r(X_t \mid X^{t-1}). \quad (8)$$

We hebben dus dat $K_n = M(X^n) \cdot K_0$, en wanneer je begint met beginkapitaal $K_0 = 1$, dan $K_n = M(X^n)$: *de toets-martingaal op tijdstip* n *meet het eindkapitaal dat je zou hebben verworven met beginkapitaal* 1.

We kunnen nu een hoge waarde van een toets-martingaal direct interpreteren als 'sterke indicatie dat nulhypothese onjuist is' — immers, als H_0 waar is dan is ons verwachte eindkapitaal niet groter dan ons beginkapitaal. Maar als we, zoals in klassiek pHT, graag grenzen op onze Type-I-fout willen, dan kunnen we een gelddrempel A instellen, en de nulhypothese verwerpen als $M(X^N) \geq A$. Het mooie is nu dat deze procedure toch weer gerelateerd kan worden aan klassiek pHT: de type-I-fout van zo'n procedure is op zijn hoogst $\alpha := 1/A$, maar in dit geval geldt dit *onafhankelijk van de stopregel die* N *bepaalt*:

Stelling: Martingaal groot ⇒ Robuuste p-waarde klein. *Laat* M *een willekeurige toets-martingaal zijn* (*zodat er dus een gokstrategie bestaat zodanig dat*, *voor alle* n, $M(X^n)$ *het verzamelde kapitaal is op tijdstip* n *bij beginkapitaal* 1 *en* '*eerlijke*' *uitbetalingen*). *Dan geldt* (*vergelijk met* (1)):

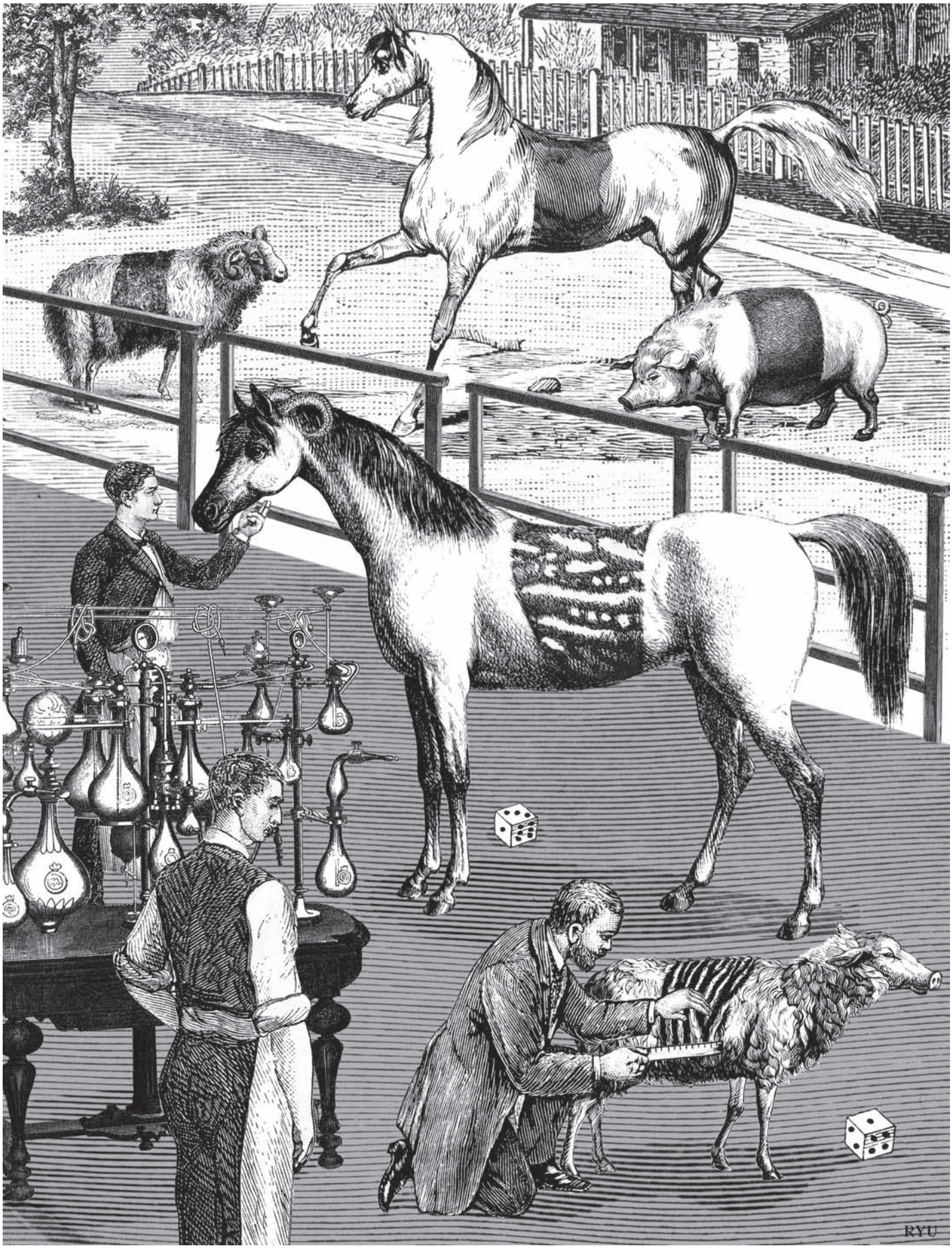
Voor elke $0 \leq \alpha \leq 1$:

$$P_0\Big(\text{Er bestaat een } n \text{ met} \frac{1}{M(X^n)} \leq \alpha\Big) \leq \alpha. \quad (9)$$

Deze stelling geldt voor algemene toets-martingalen en willekeurige H_1; hij kan worden uitgebreid naar meervoudige, willekeurige H_0. Het bewijs volgt eenvoudig uit het 'optional stopping theorem' in martingalentheorie; voor een expliciet bewijs zie [16]. Ze impliceert dat de kans dat er *überhaupt* een n bestaat waarbij de martingaal over A heengaat kleiner is dan $1/A$. We kunnen $1/M(X^n)$ dus als een *robuuste* p-waarde zien: verwerpen als $1/M(X^n)$ kleiner is dan $1/A$ leidt tot een Type-I-foutkans $\leq 1/A$ onafhankelijk van de gebruikte stopregel! Daar komt bij dat $M(X^n)$ berekend kan worden zonder 'counterfactual kansen' te hoeven weten — het is zonder meer toepasbaar op de weersvoorspellingstoets van Probleem 1. Ook kunnen toets-martingalen van onafhankelijke experimenten — als eerste provisorische indicatie van de gecombineerde evidentie — zonder meer vermenigvuldigd worden [9].

Toets-martingalen versus Bayes

In vele praktische situaties, het Bernoulli-voorbeeld incluis, is $H_0 = \{P_0\}$ enkelvoudig. In dat geval definieert elke Bayesiaanse hypothesetoets met uniforme a priori kansen op H_0 en H_1 automatisch een toets-martingaal, zoals we nu laten zien. Neem voor het gemak weer aan dat de uitkomstenruimte $\mathcal{X}$ eindig is en dat data volgens P_0 onafhankelijk is, en beschouw eerst een willekeurige kansverdeling Q over X^∞, waarbij we de conditionele kans die Q toekent aan de gerealiseerde X_t gegeven X^{t-1} schrijven als $q(X_t \mid X^{t-1})$. Omdat voor elk verleden X^{t-1} geldt dat $\sum_{x \in \mathcal{X}} q(x \mid X^{t-1}) = 1$ en $q(x \mid X^{t-1}) \geq 0$ kunnen we q óók zien als gokstrategie, die



Illustratie: Ryu Tajiri

ons vertelt om een fractie $q(x \mid X^{t-1})$ van ons kapitaal op tijdstip t te zetten op uitkomst $X_t = x$. Uit (7) volgt dat voor enkelvoudige H_0 geldt dat $r(x \mid X^{t-1}) = 1/p_0(x)$ een eerlijk uitbetalingsproces is, en dan geeft (8) dat

$$M(X^N) = \prod_{t=1}^{N} \frac{q(X_t \mid X^{t-1})}{p_0(X_t)} = \frac{q(X_1,\ldots,X_N)}{p_0(X_1,\ldots,X_N)} \qquad (10)$$

een toets-martingaal is. De laatste gelijkheid volgt hier door *telescoping*, dat wil zeggen het herhaaldelijk toepassen van de definitie van conditionele waarschijnlijkheden en uitvermenigvuldigen. Dit betekent dus dat voor *elke* kansverdeling Q met kansfunctie q, de likelihood ratio $q(X^N)/p_0(X^N)$ een toets-martingaal definieert. Dit geldt dus ook als we Q gelijk zetten aan $P(\cdot \mid H_1)$, zodat $q(X^N) := \int p_\theta(X^N)\,\pi(\theta)\,d\theta$, de Bayesiaanse kansverdeling van de data gegeven model H_1. We krijgen dan via de stelling van Bayes (3) dat, met uniforme priors $P(H_0) = P(H_1) = \frac{1}{2}$, de a posteriori odds (kansverhouding) oftewel de *Bayes factor* $B^N := p(H_1 \mid X^N)/p(H_0 \mid X^N)$ gezien kan worden als een toets-martingaal — er moet volgens de regel van Bayes immers gelden dat $B^N = M(X^N)$ met M als in (10).

Een mooi voorbeeld hiervan is de toets-martingaal die we definieerden boven (5): deze is inderdaad precies gelijk aan $p(H_1 \mid X^N)/p(H_0 \mid X^N)$ waarbij $P(\cdot \mid H_j)$ gedefinieerd is met een homogene π zoals onder (3) — feitelijk laat (5) dus een (monotone transformatie van een) Bayesiaanse posterior kans op H_1 zien! Dezelfde constructie werkt voor willekeurige (bijvoorbeeld met oneindige $\mathcal{X}$, afhankelijke $X_t, \ldots$) H_0 en H_1, en willekeurige priors π binnen H_1, zolang H_0 maar enkelvoudig is.

We kunnen dus elke Bayesiaanse toets met prior kans $\frac{1}{2}$ op enkelvoudige nulhypothese als een martingaaltoets zien en de posterior odds tegen H_0 als de virtuele winst van de bijbehorende gokstrategie. Dit feit is op zich al bekend sinds het werk van Doob in de jaren veertig. Een veel recenter inzicht is echter, dat er ook martingaaltoetsen bestaan die geen Bayesiaanse interpretatie hebben en toch zinvol (en soms zelfs te prefereren) zijn — met uitzondering van het klassieke werk van Robbins en anderen op het gebied van sequentieel toetsen (dit gaan weer terug tot de jaren vijftig — zie [12]) werden nuttige niet-Bayesiaanse martingaaltoetsen tot voor kort eigenlijk nauwelijks gebruikt. In de laatste fase van mijn Vici-project heb ik gewerkt aan het ontwerpen van dit soort toetsen. Ik sluit dit artikel af met drie korte voorbeelden:

Voorbeeld 1. Het redelijker alternatief
Beschouw een martingaaltoets waarbij we de nulhypothese verwerpen zodra $M(X^N) \geq A$, zodat de kans op een Type-I-fout kleiner is dan $1/A$. Stel nu dat H_0 niet waar is, en dat de data verdeeld zijn volgens $P_\theta \in H_1 \backslash H_0$. Er treedt een Type-II-fout op als we toch de nulhypothese accepteren, dus als, bij steekproefgrootte N, $P_\theta(M(X^N) < A)$. We willen graag dat een goede toets ook een kleine kans op een Type-II-fout geeft. Beschouw nu ons Bernoulli-voorbeeld. Voor elke redelijke martingaaltoets, en voor elke $\theta \neq \frac{1}{2}$, gaat de Type-II-foutkans uiteindelijk naar 0 naarmate we meer data observeren die verdeeld zijn volgens deze P_θ. Het is ook duidelijk dat naarmate θ dichter bij $\frac{1}{2}$ zit, we meer data nodig zullen hebben, het probleem wordt dan immers moeilijker. We kunnen nu een kleine $0 < \epsilon \ll 1$ vastleggen, en ons afvragen *hoe ver* θ van $\frac{1}{2}$ moet af zitten om te garanderen dat de kans op een Type-II-fout bij steekproefgrootte N kleiner is dan ϵ. Dit is een maat voor de efficiëntie van een martingaaltoets. In het Bernoulligeval geldt, voor elke Bayesiaanse toets met voldoende 'gladde' a priori verdelingen, dat we, met $\theta_0 = \frac{1}{2}$,

$$(\theta - \theta_0)^2 \geq C \cdot \frac{\log N}{N}$$

moet gelden om Type-II-foutkans ϵ te kunnen garanderen bij N uitkomsten, voor een constante C die afhangt van ϵ. Iets dergelijks geldt voor Bayesiaanse toetsen met enkelvoudige H_0 zodra H_1 een voldoende regulier 'parametrisch model' is, zoals bijvoorbeeld Poisson, (multivariaat) normaal, Gamma enzovoort.

Nu blijkt echter dat er ook martingaaltoetsen zijn waarbij we al Type-II-foutkans ϵ halen zodra we θ kiezen met

$$(\theta - \theta_0)^2 \geq C \cdot \frac{\log\log N}{N}.$$

We kunnen met zo'n toets dus, met eenzelfde Type-I-fout, een kleinere Type-II-fout halen: de toets is gevoeliger voor afwijkingen van de nulhypothese. [13] beschrijft martingaaltoetsen die deze efficiëntie halen, gebaseerd op een keuze voor q die de *switch distribution* werd genoemd door Van Erven e.a. [6]. Let wel: deze toets heeft dus nog steeds alle martingaal-voordelen zoals geldigheid en implementeerbaarheid zonder kennis van de stopregel. De klassieke Neyman–Pearson p-waarde gebaseerde toets haalt, nog net iets beter, C/N, maar de prijs is dat deze garantie alleen geldt en de toets alleen implementeerbaar is als die N van tevoren vastgelegd is. Asymptotisch gezien is de switch-martingaaltoets dus bijna zo krachtig als de klassieke toets.

Voorbeeld 2. Het ontbrekende alternatief
Soms wil men toetsen of een nulhypothese (althans tot op zekere hoogte) correct is, zonder dat men een heel specifieke alternatieve hypothese in gedachten heeft. Er zijn tegenwoordig bijvoorbeeld machientjes te koop die een reeks enen en nullen produceren die via quantummechanische effecten tot stand zijn gekomen. De producenten claimen dat deze reeksen 'echt' random (onafhankelijk Bernoulli($\frac{1}{2}$) zijn. Wanneer we de reeksen nu proberen te comprimeren met een standaard data-compressor als `rar` of `zip` (aanwezig op uw laptop) en we vinden dat we substantieel kunnen comprimeren, dan is dit een duidelijke aanwijzing dat de data in werkelijkheid *niet* volledig random zijn. Zo'n soort datacompressietoets is door Ryabko en Monarev [15] ook daadwerkelijk uitgevoerd op (niet-quantum) pseudo-toevalsgeneratoren, en zij vonden dat ze niet bepaald goed waren: substantiële compressie door middel van `rar` bleek mogelijk. Ryabko's compressietoets is geen standaard Neyman–Pearson-nulhypothesetoets, omdat het niet duidelijk is wat het alternatief precies is: bij een standaard toets zouden we een precies alternatief moeten formuleren (zoals 'de data komen van een 1ste-orde Markov-keten'). Om dezelfde reden is Ryabko's toets ook niet Bayesiaans. Ryabko's toets — die mij buitengewoon overtuigend lijkt — kan echter wel degelijk als een supermartingaaltoets geïnterpreteerd worden — nog een extra reden om ons niet te beperken tot Bayesiaans toetsen.

We zien hier een twistpunt waarbij Neyman en de Bayesianen op één lijn stonden tegenover Fisher. Voor Neyman en Bayes moet er altijd een precies geformuleerd alternatief H_1 zijn: minder vertrouwen in H_0 betekent automatisch meer vertrouwen in H_1. Ryabko's randomness-toets is daarentegen Fisheriaans: compressie leidt ertoe dat men H_0 verwerpt, maar niet dat men

een bepaalde H_1 accepteert — zie Figuur 2 en het voortreffelijke overzicht 'Could Fisher, Jeffreys and Neyman have agreed on testing?' [2].

Voorbeeld 3. De meervoudige nulhypothese Het grootste, en interessantste, verschil tussen Bayesiaanse en martingaaltoetsen ontstaat echter wanneer de nulhypothese meervoudig is. De definitie van toets-martingaal vereist dan dat we een uitbetalingsproces q en opbrengstproces r construeren zodanig dat (6) geldt voor *alle* $P \in H_0$. Vovk [18] stelt deze eis maar noch Vovk noch enig ander auteur werkt verder uit hoe zo'n q en r geconstrueerd zouden moeten worden — dit zou dan op zo'n manier moeten gebeuren dat ook de kans op Type-II-fouten snel naar nul gaat. In de eerste vier jaar van mijn Vici-onderzoek ben ook ik hier geen stap verder mee gekomen, maar in het laatste jaar is het dan uiteindelijk gelukt om voor een willekeurige H_0 en gokproces q een bijbehorend eerlijk opbrengstproces r te construeren met de benodigde eigenschappen. De resulterende toetsen bieden beduidend sterkere garanties dan Bayesiaanse toetsen, waarvoor (6) voor meervoudige H_0 in het algemeen niet geldt. Zo blijven Bayesiaanse toetsen in het meervoudig-H_0-geval alleen geldig onder 'optioneel stoppen' als men de a priori verdeling π_0 binnen H_0 daadwerkelijk gelooft. In de praktijk worden meestal π_0 gekozen die handig rekenen — met 'echt geloof' heeft dat niet zoveel te maken. De martingaaltoetsen blijven geldig onder optioneel stoppen onder *elke* $P \in H_0$.

Biografie

Peter Grünwald is senior onderzoeker aan het Centrum Wiskunde & Informatica en hoogleraar statistisch leren aan de Universiteit Leiden. In 2010 ontving hij, samen met Harry van Zanten, de Van Dantzigprijs van de VvS+Or, de hoogste Nederlandse onderscheiding op het gebied van statistiek en OR. Hij is auteur van *The Minimum Description Length Principle* (MIT Press, 2007), een boek over data-analyse met behulp van datacompressie, gerelateerd aan de supermartingaalmethode. Dank aan Stéphanie van der Pas, die mij attendeerde op [21], en aan Mark de Rooij en Eric-Jan Wagenmakers voor meerdere nuttige discussies.

Referenties

1 270 auteurs, Estimating the reproducibility of psychological science, *Science* 349 (6251), 2015.

2 J.O. Berger, Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science* 18(1) (2003), 1–12.

3 C. Brownie en J. Kiefer, The ideas of conditional confidence in the simplest setting, *Comm. Statistical Theory and Methods* 6 (69) (1977), 691–751.

4 A.P. Dawid, Present position and potential developments: Some personal views, statistical theory, the prequential approach, *Journal of the Royal Statistical Society, Series A* 147(2) (1984), 278–292.

5 W. Edwards, H. Lindman en L.J. Savage, Bayesian statistical inference for psychological research, *Psychological Review* 70 (1963), 193–242.

6 T. van Erven, P. Grünwald en S. de Rooij, Catching up faster by switching sooner: A predictive approach to adaptive estimation with an application to the AIC-BIC dilemma, *Journal of the Royal Statistical Society, Series B* 74(3) (2011), 361–397; with discussion, 397–417.

7 R. Fisher, *Statistical Methods for Research Workers*, Genesis Publishing, 1925.

8 P. Grünwald, *The Minimum Description Length Principle*, MIT Press, Cambridge, MA, 2007.

9 P. Grünwald, Paranormale statistiek: over de vele problemen met p-waarden, en een redelijk alternatief, *STAtOR* 16(3) (2015), 9–16.

10 P. Grünwald, Safe probability, Technical report, 2016, arxiv.org/abs/1604.01785.

11 J. Ioannidis, Why most published research findings are false, *PLoS Medicine* 2(8) (2005), doi:10.1371/journal.pmed.0020124.

12 T.L. Lai, Martingales in sequential analysis and time series, 1945–1985, *Electronic Journal for History of Probability and Statistics* 5(1) (2009).

13 S. van der Pas en P.D. Grünwald, Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in single parameter model selection, Preprint, 2014, arXiv:1408.5724.

14 J.W. Pratt, On the foundations of statistical inference: Discussion of Birnbaum's paper, *Journal of the American Statistical Association* 57 (1962), 314–315.

15 B. Ya. Ryabko en V.A. Monarev, Using information theory approach to randomness testing, *Journal of Statistical Planning and Inference* 133(1) (2005), 95–110.

16 G. Shafer, A. Shen, N. Vereshchagin en V. Vovk, Test martingales, Bayes factors and p-values, *Statistical Science* 26(1) (2011), 84–101.

17 J. Ville, Etude critique de la notion de collectif, *Monographies des Probabilités* 3 (1939).

18 V.G. Vovk, A logic of probability, with application to the foundations of statistics, *Journal of the Royal Statistical Society, Series B* 55 (1993), 317–351; with discussion.

19 E.J. Wagenmakers, A practical solution to the pervasive problems of p-values, *Psychonomic Bulletin and Review* 14(5) (2007), 779–804.

20 R.L. Wolpert, Testing simple hypotheses, in H.H. Bock en W. Polasek, eds., *Data Analysis and Information Systems: Statistical and Conceptual Approaches*, Springer, 1996, pp. 289–297.

21 H.R. Wulff, B. Andersen, P. Brandenhoff en F. Guttler, What do doctors know about statistics? *Statistics in Medicine* 6(1) (1987), 3–10.