Piranha's. Foto: Today I Found Out

DE BAYESIAANSE LEERCYCLUS

De regel van Bayes lijkt op het eerste gezicht weinig inspirerend. De regel (niet eens een wet) volgt direct uit de definitie van conditionele kans, en geeft aan hoe de kans op A gegeven B (bv. de kans dat een willekeurige Ajax-fan ook een Nederlander is) zich verhoudt tot de kans op B gegeven A (de kans dat een willekeurige Nederlander ook een Ajax-fan is). Hoe bedrieglijk, hoe oppervlakkig is deze eerste indruk! In werkelijkheid laat de regel van Bayes zien hoe organismen middels het toetsen van voorspellingen hun kennis optimaal kunnen bijwerken. Dit universele leerproces kan behalve door dieren en software programma's ook gebruikt worden om wetenschappelijke vragen te beantwoorden.

Eric-Jan Wagenmakers & Quentin F. Gronau

'Zelfs een ezel stoot zich in het gemeen niet tweemaal aan dezelfde steen.' Ezels leren namelijk van hun ervaring, en dit aanpassingsvermogen deelt de ezel met alle ons bekende diersoorten – katten, salamanders, spinnen, ja zelfs eencellige slijkzwammen zijn in staat om te leren. Dat kan ook moeilijk anders, want de evolutie boetseert de natuur op een meedogenloze manier: organismen die zich niet kunnen aanpassen aan hun omgeving wacht de vergetelheid.

Maar hoe kunnen organismen leren van hun omgeving? In algemene zin kan er alleen geleerd worden als er meerdere concurrerende hypotheses bestaan. Als er maar een enkele hypothese bestaat dan spreekt men ook wel van een religieuze overtuiging, een rotsvaste mening die door geen enkel feit aan het wankelen kan worden gebracht. Om te leren moeten we dus beginnen met meerdere concurrerende hypotheses, ieder met een zekere mate van plausibiliteit of waarschijnlijkheid. Een jonge piranha in de Amazone voelt van veraf een beweging in het water; de ene hypothese is dat het gaat om een mogelijke prooi, de andere dat het gaat om een mede-piranha. Om meer te weten te komen zwemt onze piranha dichterbij. Zo verzamelt de piranha nieuwe gegevens, en deze gegevens zouden aanleiding moeten geven tot leren, dat wil zeggen tot een aanpassing van de relatieve waarschijnlijkheid van de hypotheses. Het is intuïtief dat hypotheses toenemen en afnemen in waarschijnlijkheid al naar gelang hun voorspellende verdienste: de 'prooi-hypothese' voorspelt een heftig spartelen, en de 'mede-piranha hypothese' voorspelt een meer gelijkmatige beweging. Wanneer de nieuwe gegevens nu wijzen op een heftig spartelen dan neemt de plausibiliteit van de prooi-hypothese toe en die van de mede-piranhahypothese af.

Op basis van dergelijke algemene overwegingen komen we tot de volgende kwalitatieve wetmatigheid:

bijgewerkte kennis over de wereld	=	bestaande kennis over de wereld	×	relatieve voorspellende verdienste

Dit zegt dat het leerproces – het bijwerken van kennis op basis van geobserveerde data – gestuurd wordt door de voorspellende verdienste van de concurrerende hypotheses. Deze redenering van het gezonde verstand wordt precies gemaakt door de regel van Bayes:

$$\underbrace{p(\theta \mid \text{data})}_{\text{Posterior overtuiging voor } \theta} = \underbrace{p(\theta)}_{\text{Prior overtuiging voor } \theta} \times \underbrace{\frac{p(\text{data} \mid \theta)}{p(\text{data})}}_{\text{Relatieve voorspellende verdienste voor } \theta}$$

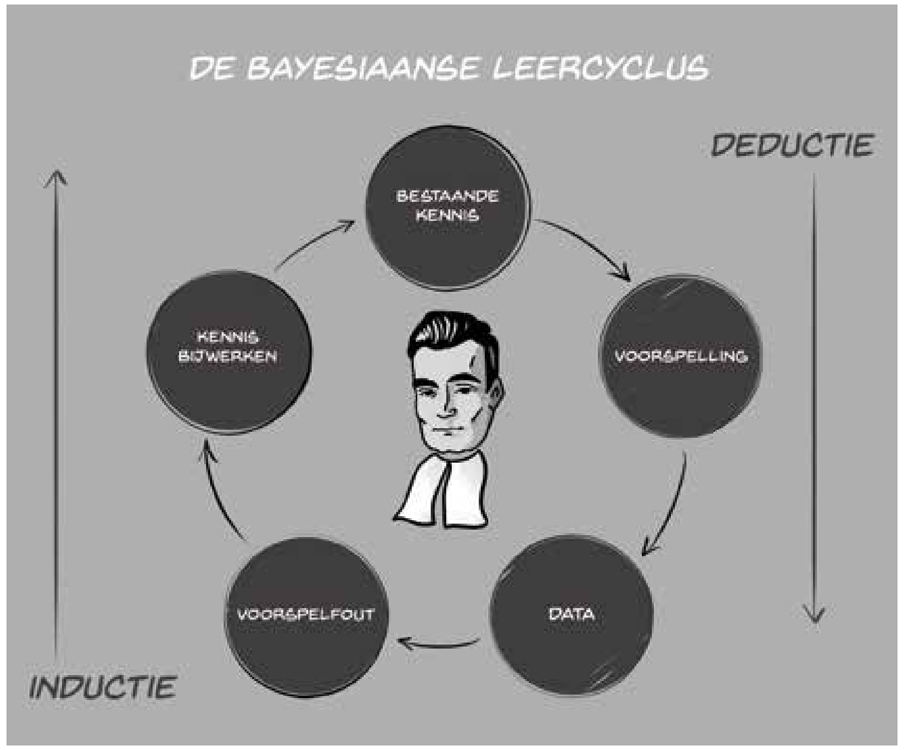
De term 'prior' staat voor de bestaande overtuiging voordat de data zijn geobserveerd, 'posterior' staat voor de bijgewerkte overtuiging, dus nadat de data zijn geobserveerd, en de Griekse letter θ (theta) representeert een van meerdere hypotheses of proposities waarover we iets willen leren. De overtuiging voor een bepaalde hypothese θ neemt toe naarmate zij de data *beter* voorspelt dan gemiddeld, en neemt af naarmate zij de data *slechter* voorspelt dan gemiddeld (zie ook Rouder & Morey, 2017).

Het is tenslotte nog van groot belang dat het leerproces nooit hoeft te stoppen; de posterior-overtuiging voor θ, de bijgewerkte kennis dus, dient als prior-kennis voor de voorspelling van de volgende reeks gegevens. Dit is niet alleen theoretisch elegant, maar voor een organisme zoals onze piranha, die een leven lang geconfronteerd wordt met een continue stroom gegevens, is dit ook praktisch relevant: nadat de kennis is bijgewerkt hebben de oude data hun werk gedaan en kunnen ze veilig worden vergeten – het enige wat de piranha hoeft te doen is de nieuwe data gebruiken om zijn al bestaande kennis bij te schaven. Figuur 1 geeft een schematisch overzicht van de Bayesiaanse leercyclus waarin voorspellen (een deductief proces) en kennis bijwerken (een inductief proces) elkaar telkens afwisselen.

Op middelbare scholen en universiteiten wordt de regel van Bayes vaak anders onderwezen, namelijk als truc om van p(A|B) te gaan naar p(B|A). Stel we weten dat er in 2010 ongeveer 16,6 miljoen Nederlanders waren en 737 miljoen Europeanen, dus p(NL) = 16,6/737 ≈ 0,023. Stel we weten ook dat er in Europa 7,1 miljoen Ajax-fans zijn, dus p(Ajax-fan) = 7,1/737 ≈ 0,010. Ten slotte weten we dat van die 7,1 miljoen fans er 4,1 miljoen Nederlanders zijn, dus p(NL | Ajax-fan) = 4,1/7,1 ≈ 0,577. Figuur 2 beeldt deze situatie uit als een Venn-diagram. Dan kunnen we via de bovenstaande regel van Bayes uitrekenen wat de kans is dat een willekeurige Nederlander ook fan is van Ajax: p(Ajax-fan | NL) = p(Ajax-fan) x p(NL | Ajax-fan) / p(NL) = 0,010 x 0,577/0,023 ≈ 0,251. De regel van Bayes is natuurlijk onmisbaar voor de kansrekening, maar het is jammer om de daarmee verbonden Bayesiaanse leerfilosofie vervolgens over te slaan.

Een voorbeeld van Bayesiaans leren: Het getal π

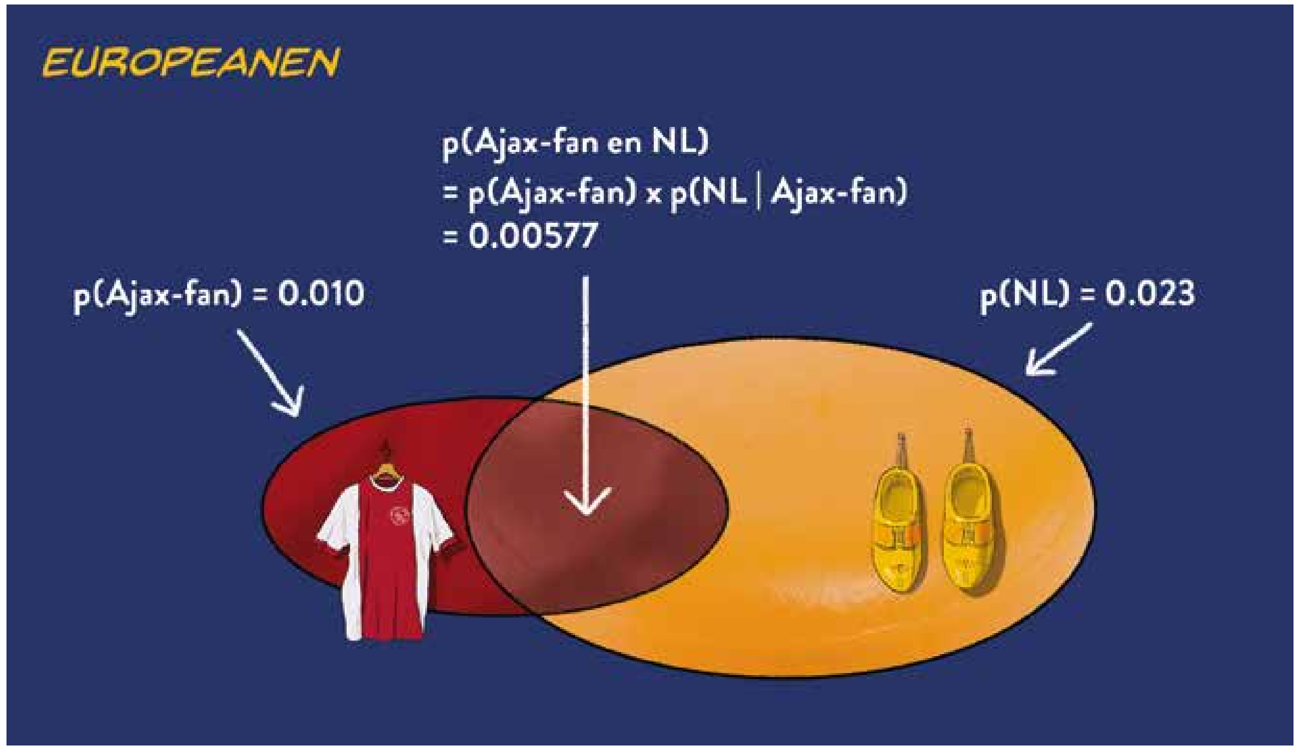
Om de eigenschappen van Bayesiaans leren te verduidelijken behandelen we nu een concreet probleem rondom



Figuur 1. Een conceptuele representatie van de Bayesiaanse leercyclus: vanuit bestaande kennis wordt een deductieve voorspelling gedaan; confrontatie met geobserveerde data leidt tot een voorspelfout, en deze geeft aanleiding tot het inductief bijwerken van de bestaande kennis; hierna kan de leercyclus worden hervat, met dien verstande dat de bijgewerkte kennis de plaats inneemt van de eerdere bestaande kennis (zie ook Jevons, 1874). Illustratie: Viktor Beekman instagram.com/viktordepiktor

het irrationale getal π, de verhouding tussen de omtrek en de diameter van een cirkel. Het getal π is 3,1415926535… en zo voort tot in de oneindigheid. Een prangende vraag die wiskundigen al langere tijd uit hun slaap houdt is of π 'normaal' is, dat wil zeggen, dat iedere specifieke reeks decimalen van een bepaalde lengte even vaak voorkomt. Er is een sterk vermoeden dat π inderdaad normaal is, maar niemand heeft het nog wiskundig kunnen bewijzen. Wij kunnen dat ook niet, maar wel kunnen we kwantificeren in hoeverre de data ondersteuning bieden voor de concurrerende hypotheses. Voor de eenvoud zullen we hier enkel onderzoeken of de even decimalen van π net zo vaak voorkomen als de oneven decimalen, en nemen we alleen de eerste honderd decimalen in beschouwing (voor een uitgebreider benadering zie Gronau & Wagenmakers, 2018).

Voordat we het leerproces kunnen starten dienen we eerst onze concurrerende hypotheses te formuleren, en wel zo specifiek dat ze ieder concrete voorspellingen kunnen maken. Onze eerste hypothese, H0, is dat π normaal is, en dit houdt in dat de even decimalen net zo vaak voorkomen als de oneven decimalen. Deze hypothese voorspelt dat de kans op een even decimaal telkens precies 0,5 is. Op dit moment schrikt de klassiek geschoolde statisticus misschien wakker: 'Hoe kun je nu spreken van *de kans op een even decimaal*? Een gegeven decimaal is immers even of oneven, en dat kun je zo opzoeken of uitrekenen!' Dat is natuurlijk waar. De decimalen van π zijn deterministisch bepaald, en het lijkt misschien vreemd om hier een statistische methode toe te passen. Maar deze schijn bedriegt. De decimalen van π mogen dan deterministisch bepaald zijn, de uitkomst van de worp van een eerlijke munt is dat evenzeer! Met perfecte kennis van de onderliggende fysische processen (met name de verticale kracht, de draaiimpuls, en wellicht nog de luchtweerstand) kan de uitkomst feilloos worden voorspeld. Jevons (1874, pp. 197-198) verwoordde het deterministisch perspectief als volgt: 'Kans bestaat niet (…) De bliksem twijfelt niet over waar hij zal inslaan; in de zwaarste storm is geen willekeur; geen korreltje zand ligt op het strand, of oneindige kennis zou kunnen verklaren waarom het zich daar bevindt; en de vlucht van ieder vallend blad wordt bepaald door de mechanische principes die regeren over de bewegingen van de hemellichamen. Kans bestaat derhalve niet in de natuur, en valt niet te verenigen met kennis; zij is enkel een uitdrukking, zoals Laplace al opmerkte, van onze onwetendheid over de betreffende oorzaken, en ons daaruit voortvloeiend onvermogen om het resultaat te voorspellen, of om het feilloos voort te brengen. In de natuur is het geschieden van een gebeurtenis voorbestemd vanaf de eerste constructie van het universum. *Waarschijnlijkheid behoort volledig tot de*

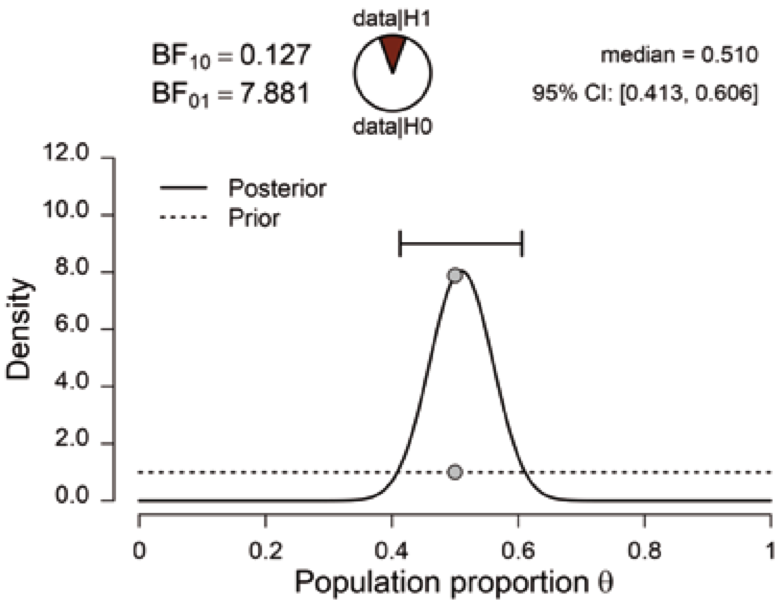


Figuur 2. In dit Venn-diagram kan het overlappende gebied (Nederlandse Ajax-fans) worden uitgedrukt hetzij als proportie van Ajax-fans (gegeven dat iemand Ajax-fan is, hoe groot is de kans dat hij of zij Nederlander is?), hetzij als proportie van Nederlanders (geven dat iemand Nederlander is, hoe groot is de land dat hij of zij Ajax-fan is?); dit ligt ten grondslag aan de regel van Bayes. Illustratie: Viktor Beekman instagram.com/viktordepiktor

gedachtenwereld.' (onze vertaling; de laatste zin in het origineel luidt: *Probability belongs wholly to the mind*). Met andere woorden, als we aannemen dat π normaal is, en u wordt gevraagd wat de waarschijnlijkheid is dat de 99ste decimaal van π even is, dan kunt u vanuit uw onwetendheid – even aangenomen dat u het antwoord niet uit uw hoofd kent – niet anders doen dan deze waarschijnlijkheid inschatten op 0,5.

De 'π is normaal' hypothese, H0, is star en kent de onbekende proportie θ van even decimalen een enkele waarde toe. De tweede hypothese, H1, is flexibeler en staat toe dat de proportie θ alle mogelijke waardes aanneemt tussen 0 en 1. Om H1-voorspellingen te kunnen laten maken moet het duidelijk zijn wat de a priori waarschijnlijkheid is van ieder van die waardes voor θ. Er zijn nu plotseling twee soorten prior waarschijnlijkheid, en het is van belang deze scherp te onderscheiden. De eerste 'prior' is die voor de discrete hypotheses H0 en H1. Welke hypothese is waarschijnlijker? Komen even decimalen van π net zo vaak voor als oneven decimalen, of is de verhouding toch scheef? We analyseren hier de eerste 100 decimalen van π, dus we reizen mentaal terug in de tijd tot de periode van Archimedes. Het lijkt dan niet gemakkelijk om aan H0 en H1 zinvolle, niet arbitraire prior kansen toe te kennen. We worden daarom iets minder ambitieus en stellen ons als doel om de relatieve voorspellende verdienste van H0 versus H1 te kwantificeren, dat wil zeggen de evidentie die de data aandragen – de mate waarin de kennis bijgewerkt moet worden. De relatieve voorspellende verdienste van H0 versus H1 wordt ook wel de 'Bayes factor' genoemd, en wij korten hem af als BF_{01}. De tweede prior is die voor de continue set van deelhypotheses θ die vallen onder de overkoepelende hypothese H1. Deze prior is een verdeling over de waardes van θ die aangeeft 'onder de aanname dat even decimalen van π *niet* net zo vaak voorkomen als oneven decimalen, wat kunnen we dan verwachten over de mate waarin dit het geval is?' Een populaire keuze is de uniforme prior, die deze vraag beantwoordt met 'geen idee; iedere waarde voor de proportie θ is even waarschijnlijk'.

We vergelijken nu dus twee concurrerende hypotheses die ieder concrete voorspellingen maken. De starre H0 stelt dat $\theta = 0{,}5$, en de flexibele H1 stelt dat iedere waarde van θ tussen 0 en 1 even waarschijnlijk is. De Bayesiaanse leercyclus passen we nu toe op de eerste honderd decimalen van π; hiervan zijn er 51 even. Een gemakkelijke manier om deze Bayesiaanse analyse uit te voeren is met het gratis statistisch softwareprogramma JASP (jasp-stats.org) dat in onze groep aan de Universiteit van Amsterdam wordt ontwikkeld. Figuur 3 geeft een overzicht van het resultaat (zie ook



Figuur 3. Een Bayesiaanse analyse van de vraag of het getal π net zoveel even als oneven decimalen kent, gebaseerd op de eerste 100 decimalen (waarvan 51 even). De figuur is gegenereerd door JASP, jasp-stats.org

https://osf.io/xbuqc/). De twee belangrijkste conclusies zijn als volgt:

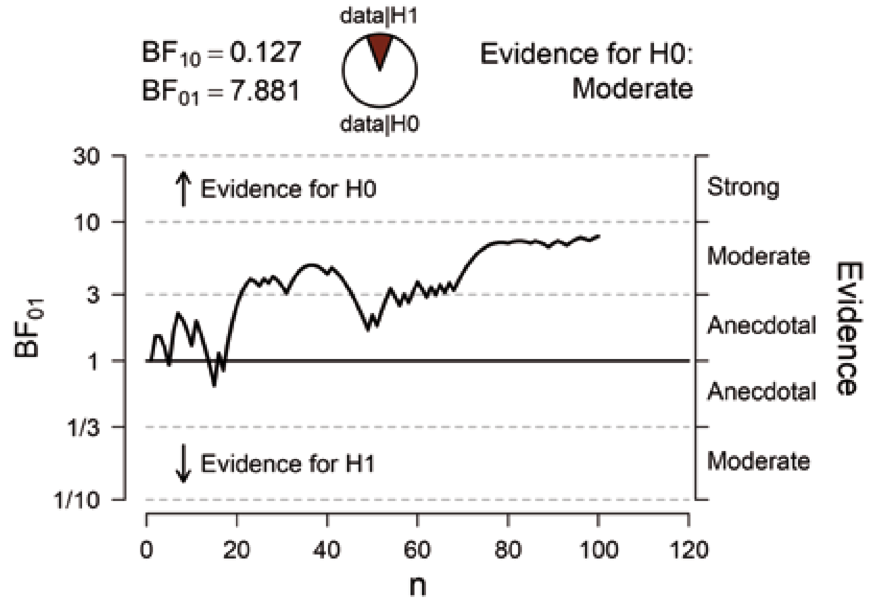
1. De data bieden steun voor H0; BF_{01} = 7,881, hetgeen betekent dat H0 de data bijna 8 keer zo goed voorspelde als H1. De 'pizza plot' poogt de sterkte van de evidentie intuïtief te maken. Het witte 'mozzarella' gedeelte van de pizza representeert de posterior-model kans voor H0 (onder een 50-50 prior-model kans) en het rode 'pepperoni' gedeelte representeert de posterior-model kans voor H1. Stel je prikt met je wijsvinger blindelings in de pizza, en hij landt in het niet-dominante beleg (in dit geval de pepperoni). Hoe verrast ben je? De intensiteit van je ingebeelde verrassing is een maat voor de sterkte van de evidentie.
2. Indien we aannemen dat H1 waar is (een aanname die door de data dus enigszins wordt weersproken) dan kunnen we de continue prior- en posterior-verdeling voor θ interpreteren. Waardes van θ lager dan 0,4 en hoger dan 0,6 zijn minder waarschijnlijk geworden door de data, terwijl waardes tussen 0,4 en 0,6 juist waarschijnlijker zijn geworden. De waarde van θ = 0,5 is 7,881 keer zo waarschijnlijk geworden, en dit bevestigt dat H0 (die immers gedefinieerd is als θ = 0,5) de data bijna 8 keer zo goed voorspelde als H1. Verder valt 95% van de posterior verdeling tussen θ = 0,413 en θ = 0,606, wat betekent dat we 95% zeker zijn dat de ware waarde zich in dit interval bevindt.

De evidentie voor H0 versus H1 kan ook worden bijgehouden terwijl de decimalen een voor een aan de analyse worden toegevoegd. De leercyclus wordt dus 100 keer doorlopen, en na iedere decimaal wordt opgeslagen hoe goed H0 en H1 de data tot dat punt voorspelden. Het resultaat van deze sequentiële analyse staat in figuur 4. Het is duidelijk dat de evidentie voor H0 over het algemeen toeneemt naarmate er meer decimalen worden meegenomen in de analyse. De categorieën van evidentie op de tweede y-as zijn afkomstig van de Bayesiaanse pionier Harold Jeffreys; ze zijn bedoeld als grove heuristiek, als hulpmiddel bij interpretatie, maar niet als harde drempel voor alles of niets beslissingen.

De voordelen van Bayesiaans leren voor de wetenschappelijke praktijk

De Bayesiaanse analyse van π hierboven is bijzonder relevant voor de praktische onderzoeker. Ten eerste laat het voorbeeld zien dat het mogelijk is om evidentie te verzamelen in het voordeel van de nulhypothese. Als de nulhypothese de geobserveerde data beter voorspelt dan de alternatieve hypothese, dan is dit evidentie voor de afwezigheid van een effect; dit contrasteert met het scenario waar beide hypotheses de geobserveerde data ongeveer even goed voorspellen, zodat er sprake is van afwezigheid van evidentie. In de wetenschap is het vaak cruciaal om evidentie te kunnen verzamelen voor de afwezigheid van een effect: 'veroorzaakt vaccinatie autisme of doet het dat *niet*?'; 'repliceert het effect van *power posing* of repliceert het *niet*?'; 'ondersteunen de data de hypothese dat lijstlengte *niet* van invloed is op de geheugenprestatie?'

Ten tweede laat het voorbeeld zien dat evidentie bijgehouden kan worden terwijl de data geleidelijk binnenkomen. Als we niet tevreden zijn met de evidentie na 100 decimalen kunnen we eenvoudigweg meer decimalen



Figuur 4. Naarmate er meer decimalen van π worden toegevoegd aan de analyse stijgt the evidentie voor de nulhypothese dat even en oneven getallen net zo vaak voorkomen. De figuur is gegenereerd door JASP, jasp-stats.org

toevoegen tot we wel overtuigd zijn. Conceptueel is dit hetzelfde als de piranha die, nog onzeker over de bron van de beweging die hij heeft waargenomen, dichterbij gaat zwemmen om meer informatie te verzamelen en zo zijn onzekerheid te verkleinen. Ter illustratie hebben de eerste 100 miljoen decimalen van π geanalyseerd (hiervan zijn er 50.006.474 even); de relatieve voorspellende verdienste van H0 was aanzienlijk groter dan die van H1 – de Bayes factor BF_{01} is ongeveer 3451. Voor de pragmatische onderzoeker betekent dit een toename in efficiëntie: wanneer een experiment in een vroeg stadium al doorslaggevende evidentie levert kan de datacollectie eerder worden stopgezet dan oorspronkelijk gepland.

De kern van het Bayesiaanse gedachtegoed is eenvoudig: verklaringen worden waarschijnlijker als ze de data relatief goed voorspellen, en minder waarschijnlijk als ze de data relatief slecht voorspellen. De uitdaging voor Bayesiaanse onderzoekers is om concurrerende modellen zo te specificeren dat ze zinnige voorspellingen maken, voorspellingen die de moeite van het toetsen waard zijn. De aard van de voorspellingen wordt gedeeltelijk bepaald door de prior verdeling, en dit vormt zowel de achilleshiel als de unieke kracht van het Bayesiaanse paradigma. Hoe zorgvuldiger de prior verdeling wordt gekozen, hoe zinniger de voorspellingen, en hoe betekenisvoller de statistische conclusies.

In ons voorbeeld van π hebben we een uniforme verdeling aangenomen voor de proportie θ van even decimalen. Andere keuzes zijn mogelijk en leiden onherroepelijk tot andere resultaten. Is dit een probleem? Wij denken van niet. Stel er is een aantal verschillende prior verdelingen, die allemaal verdedigbaar zijn; als de resultaten kwalitatief overeen komen dan hebben we ontdekt dat onze conclusies robuust zijn tegen redelijke veranderingen van prior aannames; als daarentegen de resultaten kwalitatief sterk verschillen dan moeten we erkennen dat de data onvoldoende doorslaggevend zijn, zodat plausibele, verdedigbare veranderingen in prior kennis leiden tot grote verschillen in de uiteindelijke conclusie. Dergelijke onzekerheid dient te worden benoemd en gerespecteerd, en niet onder het tapijt geveegd.

We hopen dat er in de toekomst meer aandacht wordt besteed aan de Bayesiaanse leercyclus, zowel in onderzoek als in onderwijs. In de woorden van James Clerk Maxwell: 'de ware Logica van deze wereld is de Kansrekening.'

Literatuur

Gronau, Q. F., & Wagenmakers, E.-J. (2018). Bayesian evidence accumulation in experimental mathematics: A case study of four irrational numbers. *Experimental Mathematics, 27*, 277-286.

JASP Team (2018). JASP(Version 0.9) [Computer software].

Jevons, W. S. (1874). *The principles of science: A treatise on logic and scientific method*. London: MacMillan.

Rouder, J. N., & Morey, R. D. (2017). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, DOI: 10.1080/00031305.2017.1341334.

Eric-Jan Wagenmakers is hoogleraar Neurocognitieve Modellering aan de Faculteit der Maatschappij- en Gedragswetenschappen (programmagroep Psychologische Methodenleer) van de Universiteit van Amsterdam.
E-mail: E.M.Wagenmakers@uva.nl

Quentin F. Gronau is PhD-student aan de Faculteit der Maatschappij- en Gedragswetenschappen (programmagroep Psychologische Methodenleer) van de UvA.
E-mail: quentin.f.gronau@gmail.com