Ronald Meester
Afdeling Wiskunde
Vrije Universiteit Amsterdam
r.w.j.meester@vu.nl

Onderzoek

Waarom p-waardes niet gebruikt mogen worden als statistisch bewijs

Het toetsen van hypotheses is een van de belangrijkste onderdelen van elke inleidende cursus in de statistiek. Soms word je geconfronteerd met twee elkaar beconcurrerende hypotheses H_1 en H_2 en word je op de een of andere manier geacht een keuze te maken op basis van de data die tot je beschikking staat. Maar soms is niet het maken van een keuze het uiteindelijke doel, maar om te bepalen of de data op de een of andere manier bewijs geeft dat een specifieke hypothese waar is. Het onderscheid tussen enerzijds keuzes maken en anderzijds bewijs vinden wordt vaak vergeten en niet herkend, maar het is een wezenlijk verschil. In dit artikel dat gebaseerd is op zijn voordracht op het KWG Wintersymposium van 12 januari belicht Ronald Meester het begrip 'statistisch bewijs'.

In deze bijdrage gaat het mij om statistisch bewijs. Wat maakt dat bepaalde gegevens bewijs geven voor het waar zijn van een bepaalde hypothese? En wat is dat eigenlijk, statistisch bewijs? De meeste lezers zullen zich wel de colleges of cursussen in de statistiek herinneren die ze in het verleden hebben gevolgd, en ik vermoed ook dat de meesten een onmiddellijke associatie met significantie-toetsen en p-waardes hebben. Natuurlijk, sluitend bewijs dat een bepaalde hypothese waar is zal de statistiek niet vaak kunnen geven, maar we kunnen, zo is de algemene gedachte, wel procedures beschrijven die de kans op foutieve claims klein maken. Door het kiezen van een geschikt *significantieniveau* α (meestal $0{,}01$ of $0{,}05$, maar dat is in wezen volkomen arbitrair) en het uitvoeren van een significantietest, kunnen we ervoor zorgen dat we in hooguit een klein en gecontroleerd aantal gevallen een foutieve beslissing nemen door bewijs te claimen terwijl de hypothese eigenlijk niet waar is. Een dergelijke procedure zou dan de bewijskracht voor een bepaalde hypothese moeten kwantificeren, middels de van tevoren vastgestelde α.

Het is precies die laatste uitspraak die ik in deze bijdrage wil onderzoeken, en zal verwerpen. Ik zal dat langs verschillende lijnen doen. Ik begin met het precies omschrijven van de klassieke statistische procedure, samen met de bijbehorende logica, rationale, en intuïtie. Vervolgens zal ik aan de hand van een geruchtmakende rechtszaak laten zien dat deze procedure ervoor heeft gezorgd dat mensen op ondeugdelijke gronden schuldig zijn bevonden aan moord. Op zichzelf geldt in de wiskunde natuurlijk het adagium dat zodra er één geval bekend is waarbij een procedure faalt, de hele procedure als ondeugdelijk moet worden beschouwd, en kennelijk niet in algemeenheid kan worden geaccepteerd. Ik zou het dan ook kunnen laten bij dit ene, schokkende, tegenvoorbeeld. Maar om meer gevoel te krijgen voor wat er aan de hand zou kunnen zijn, zal ik het falen van de procedure illustreren met een aantal gevallen waarin het volgen ervan tot absurditeiten leidt. De conclusie zal zijn, onontkoombaar, dat deze manier van statistisch bewijs leveren niet deugt. Een veelgeciteerd artikel van Ioannidis [1] claimt dat ongeveer 80% van de statistisch ondersteunde wetenschappelijke resultaten onjuist is. Deze conclusie, zeg ik er dan maar onmiddellijk bij, is het directe gevolg van falende statistische methodes. Dat is een tamelijk schokkende conclusie als je bedenkt dat onze leerboeken ermee vol staan, dat wij onze studenten en leerlingen dit ook leren, en dat in de wetenschappelijke praktijk deze methode schering en inslag is. Ik denk dat de *reproductiecrisis* waar de wetenschap volgens velen middenin zit, het rechtstreekse gevolg is van collectief onbegrip van wat statistisch bewijs eigenlijk is.

Maar om niet alleen maar kritiek te leveren, zal ik in dit artikel ook kort pleiten voor een totaal andere visie op bewijs, waarbij bewijs slechts relatief van aard kan zijn. Ik zal betogen dat er niet zoiets bestaat als bewijs voor hypothese H, maar dat je in plaats daarvan altijd moet spreken over bewijs voor hypothese H_1 *ten opzichte van* hypothese H_2, waarbij weliswaar H_1 en H_2

elkaar dienen uit te sluiten, maar waarbij het niet nodig is dat ze elkaars complement zijn. Deze visie is niet nieuw, en wordt al decennialang gepropageerd door verschillende van mijn collega's, in soms zeer overtuigend proza, en vanuit verschillende perspectieven bezien. En toch lukt het maar niet om de wetenschappelijke wereld ervan te overtuigen dat het zo niet verder kan met de nog steeds dominante visie dat p-waardes iets te maken hebben met statistisch bewijs. Over het waarom daarvan kan ik alleen maar gissen, maar ik heb daar wel enkele ideeën over die ik ook met de lezer zal delen.

Klassieke procedure en eerste kritiek erop

Stel je bent wetenschapper en je wilt graag een bepaalde hypothese verwerpen. Dat 'graag' is natuurlijk een tikje vilein, want horen wij wetenschappers niet neutraal te zijn? In theorie wel, maar zo werkt het meestal niet. Een farmaceut wil bijvoorbeeld dat het nieuwe medicijn beter is dan het oude, dus zal hij of zij graag de hypothese verwerpen dat dat niet zo is. Goed, er is dus een hypothese H, en er is data E. We zien die data als een stochastische grootheid en schrijven e voor een realisatie ervan. Vervolgens is er een toetsingsgrootheid T. Deze T is een functie van E, en dus zelf ook weer een stochastische grootheid. Bij elke realisatie e van E hoort een realisatie t van T. We kiezen T zo dat we de kansverdeling van T op zijn minst bij benadering kennen als we aannemen dat H waar is. Een onschuldig voorbeeld maakt hopelijk veel duidelijk: Stel H is de hypothese dat een bepaalde munt zuiver is, en E is het resultaat van 100 worpen. Een geschikte T is nu bijvoorbeeld het aantal keer dat we kop hebben gegooid. Onder de aanname dat H waar is, heeft T een bekende verdeling, namelijk binomiaal met parameters 100 en $\frac{1}{2}$.

Men wil nu een uitspraak doen over het al dan niet waar zijn van hypothese H, en de procedure gaat als volgt. We definiëren een *kritiek gebied* K, een deelverzameling van de mogelijke uitkomsten van T, met de eigenschap dat de kans dat T onder H in K terecht komt klein is. Hoe klein? Dat bepalen we van tevoren, en meestal wordt hiervoor $0{,}01$ of $0{,}05$ genomen. De keuze wordt doorgaans aangegeven met α. De uiteindelijke instructie luidt nu als volgt: "Verwerp H als de toetsingsgrootheid T in K terecht is gekomen." In het voorbeeld van zojuist wordt K doorgaans gedefinieerd als het grootste gebied van de vorm $K_R := \{n : |n - 50| \geq R\}$, waarvoor geldt dat de kans onder H dat $T \in K_R$ hooguit α is. Met andere woorden, je verwerpt H als het aantal keren kop te groot of te klein is, en dat is natuurlijk heel intuïtief.

Een equivalente variant op deze procedure is dat men geen kritiek gebied kiest, maar simpelweg de kans uitrekent dat onder H de uitkomst van T meer afwijkt van 50 (in bovengenoemd voorbeeld) dan de daadwerkelijk waargenomen afstand. Als deze kans, een *p-waarde* genoemd, kleiner is dan α, dan wordt dat als bewijs voor het onwaar zijn van H opgevat, en opnieuw wordt de bewijskracht dan gekwantificeerd met α. Deze versie is zoals gezegd equivalent aan de eerste, maar heeft volgens velen het voordeel dat het noemen van de p-waarde extra informatie geeft. Een p-waarde van $0{,}0001$ lijkt sterker bewijs te geven voor het complement H^c van H dan een p-waarde van $0{,}02$. Ik zal me verder vooral op p-waardes richten.

De logica achter deze procedures is bedrieglijk eenvoudig: als de kans onder H op de waarneming die we hebben gedaan *of op een nog extremere waarneming* erg klein is, dan geloven we niet langer dat H waar is. Meestal wordt het verwerpen van H nu opgevat als *bewijs* dat H niet waar is, en de grootte van de bewijskracht wordt dan doorgaans op α gesteld. Men verwerpt H dan op *significantieniveau* α.

Op dit punt aangekomen is er al onmiddellijk een interessant punt van kritiek te formuleren op deze gang van zaken. Statistisch bewijs voor het al dan niet waar zijn van H zou gebaseerd moeten zijn op wat we waarnemen, op de data E dus. Neem nu twee hypotheses H_1 en H_2, en stel dat de data zodanig is dat deze twee hypotheses precies dezelfde kansen toekennen aan alle waargenomen data. Een redelijke notie van statistisch bewijs zou geen onderscheid kunnen maken tussen H_1 en H_2. Immers, als een wetenschapper beweert dat de data een bepaalde bewijskracht heeft voor H_1 (wat dat ook precies moge betekenen) dan kan een andere wetenschapper zeggen dat deze bewijskracht ook voor H_2 moet gelden, want de kansverdelingen van H_1 en H_2 zijn identiek op de waargenomen data. Alleen de waargenomen gegevens zouden mogen worden gebruikt om een uitspraak te doen over de bewijskracht van de data voor een specifieke hypothese. Zelf-evident als dit principe klinkt, het wordt in een p-waardeprocedure wel degelijk geschonden. Immers, als we even bij het voorbeeld van H_1 en H_2 blijven, dan zien we dat de p-waardes die bij H_1 en H_2 horen helemaal niet hetzelfde hoeven te zijn, omdat ze afhangen van de kansen op niet-waargenomen uitkomsten, en die kansen kunnen verschillend zijn onder H_1 en H_2. In de evaluatie van forensisch DNA-bewijs komt deze situatie echt voor, maar het gaat te ver om die voorbeelden hier te bespreken.

Dus als we p-waardes op de een of andere manier willen zien als kwantificering van statistisch bewijs, dan moeten we accepteren dat deze kwantificering afhangt van niet waargenomen data. Dat lijkt mij niet acceptabel.

Laten we nog eens verder kijken naar de keuze van het kritieke gebied, want het verhaal is nog niet helemaal af. In plaats van H verwerpen als de uitkomst van T te veel afwijkt van wat je verwacht, zou je wiskundig net zo goed kunnen afspreken dat je H verwerpt als T te *dicht* bij die verwachting ligt. Als je bijvoorbeeld 10.000 keer gooit met een zuivere munt, dan is de kans dat het aantal koppen tussen 4.998 en 5.002 (inclusief) ligt gelijk aan $0{,}03988$. Deze kans is kleiner dan $0{,}05$, dus we kunnen ook afspreken om H te verwerpen als het aantal keren kop *te dicht* ligt bij wat je verwacht. Dat zal niemand willen doen, omdat het natuurlijk vreemd klinkt om H te verwerpen als T doet wat je verwacht onder H. Het punt hierbij is dat er zeker verschillende kritieke gebieden te bedenken zijn die onder H kans ten hoogste α hebben, maar dat die verschillende gebieden zich niet allemaal hetzelfde gedragen als H *niet* waar is. Als H niet waar is wil je juist met zo groot mogelijke kans in het kritieke gebied terechtkomen, en men probeert dan ook om deze zogenaamde *power* bij *gegeven* α zo groot mogelijk te maken.

Terug naar de logica van p-waardes. Fisher, een van de grondleggers van de moderne statistiek, formuleerde het als volgt [2]:

> "Belief in the [null] hypothesis as an accurate representation of the population sampled is confronted by the logical disjunction: Either the hypothesis is untrue, or the value of [the test statistic] has attained by chance an exceptionally high value."

Met andere woorden: hetzij H is onjuist of er heeft zich iets heel bijzonders voorgedaan. Hoewel deze disjunctie behoorlijk overtuigend klinkt, is het in tegenstelling tot wat Fisher beweert helemaal geen logische disjunctie. Het zou namelijk best zo kunnen zijn dat de waargenomen waarde van de toetsingsgrootheid T onder H^c *ook zeer onwaarschijnlijk is*. (Het hoeft natuurlijk helemaal niet zo te zijn dat de verdeling van T onder H^c zomaar bepaald kan worden, maar het gaat me hier om het principe.) Het feit dat T een extreme waarde heeft aangenomen onder H kunnen we voorlopig voor kennisgeving aannemen, want zonder verdere kennis over het gedrag van T onder een alternatief is er helemaal niets te concluderen. Als de uitkomst van T onder H^c *ook* extreem is, dan kunnen we alleen maar zeggen dat hetzij (1) H is waar en er heeft zich iets bijzonder voorgedaan, hetzij (2) H^c is waar en er heeft zich iets bijzonders voorgedaan. Uit deze disjunctie is, uiteraard, geen enkele conclusie te trekken want ze is feitelijk niet meer dan een tautologie. We kunnen concluderen dat de rationale achter de hele procedure gemankeerd is. Fisher zat er eigenlijk dus gewoon naast.

Een schokkend voorbeeld

De disjunctie van Fisher leidde twee decennia geleden tot een schokkende uitspraak in een geruchtmakend proces in Engeland. Sally Clark was op een gegeven moment alleen thuis met haar eerste zoon van drie maanden, toen het kind in de nacht overleed, zonder enige aanwijzing over de oorzaak van het overlijden. Dit sterfgeval werd toegeschreven aan wiegendood. Twee jaar later echter gebeurde precies hetzelfde met haar tweede zoon, op dat moment ook ongeveer drie maanden oud. Na dit tweede sterfgeval werd Sally Clark aangeklaagd en veroordeeld voor dubbele moord. Deze veroordeling hield stand in hoger beroep, en was vooral (of eigenlijk uitsluitend) gebaseerd op de berekening van de medicus (!) Sir Roy Meadoy die had aangevoerd dat de kans op een dubbele wiegendood, onder de aanname van onschuld, gelijk was aan 1 gedeeld door 72 miljoen. Op basis van dit getal werd Sally Clark veroordeeld. Immers, volgens de disjunctie van Fisher is ze of schuldig of er heeft zich iets zeer bijzonders voor gedaan, namelijk een gebeurtenis met de astronomisch kleine kans van 1 op 72 miljoen.

Het getal van 1 gedeeld door 72 miljoen is overigens zeer discutabel, omdat het uitgaat van onafhankelijkheid tussen de twee gevallen van wiegendood. Echter, er is ruime eensgezindheid onder artsen dat er ook een genetische component moet zijn die samenhangt met wiegendood, en dit feit maakt de onafhankelijkheidshypothese aanvechtbaar.

Mijn punt van kritiek heeft echter niet zozeer met dat getal te maken als wel met de logica. Laten we deze situatie eens iets wiskundiger beschrijven. Laat M de gebeurtenis zijn dat Sally Clark een dubbele moord heeft begaan, en laat W de gebeurtenis zijn van een dubbele wiegendood. We schrijven E voor de gebeurtenis dat er twee gestorven kinderen zijn, en we gaan er voor het gemak even van uit dat er geen andere acceptabele verklaringen zijn dan moord en wiegendood. Deze aanname betekent dat $E = W \cup M$, en mag gezien de omstandigheden realistisch genoemd worden. Natuurlijk geldt ook dat $W \cap M = \emptyset$, zie Figuur 1.

Welnu, de instructie van Fisher luidt dat we ons moeten concentreren op $P(E \mid M^c)$, de kans dat E optreedt onder de aanname dat Sally Clark onschuldig is. Als deze kans te klein wordt, aldus Fisher, dan hebben we geen keus en moeten we M^c, de onschuldhypothese, verwerpen. Deze kans is gelijk aan

$$P(E \mid M^c) = \frac{P(E \cap M^c)}{P(M^c)} = \frac{P(W)}{P(W) + P(E^c)}. \qquad (1)$$

Dat deze conditionele kans klein is wekt geen verbazing, aangezien $P(W)$ ongetwijfeld een stuk kleiner is dan $P(E^c)$, in Figuur 1 enigszins geïllustreerd door de verhoudingen.

Maar het vreemde is dat we uiteindelijk niet in het minst geïnteresseerd zijn in $P(E \mid M^c)$. We zijn geïnteresseerd in de kans dat Sally Clark een dubbele moord heeft begaan, gegeven het bewijs E. Deze kans is een hele andere, namelijk

$$P(M \mid E) = \frac{P(M \cap E)}{P(E)} = \frac{P(M)}{P(E)}.$$

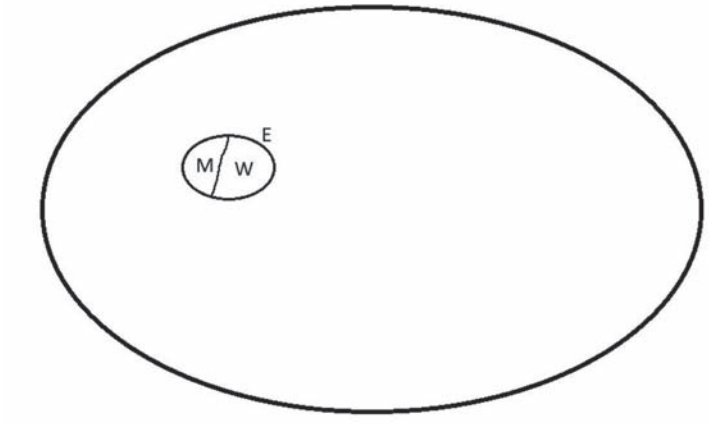
De kans dat Sally Clark onschuldig is gegeven het bewijs E is dan

$$P(M^c \mid E) = \frac{P(M^c \cap E)}{P(E)} = \frac{P(W)}{P(E)},$$

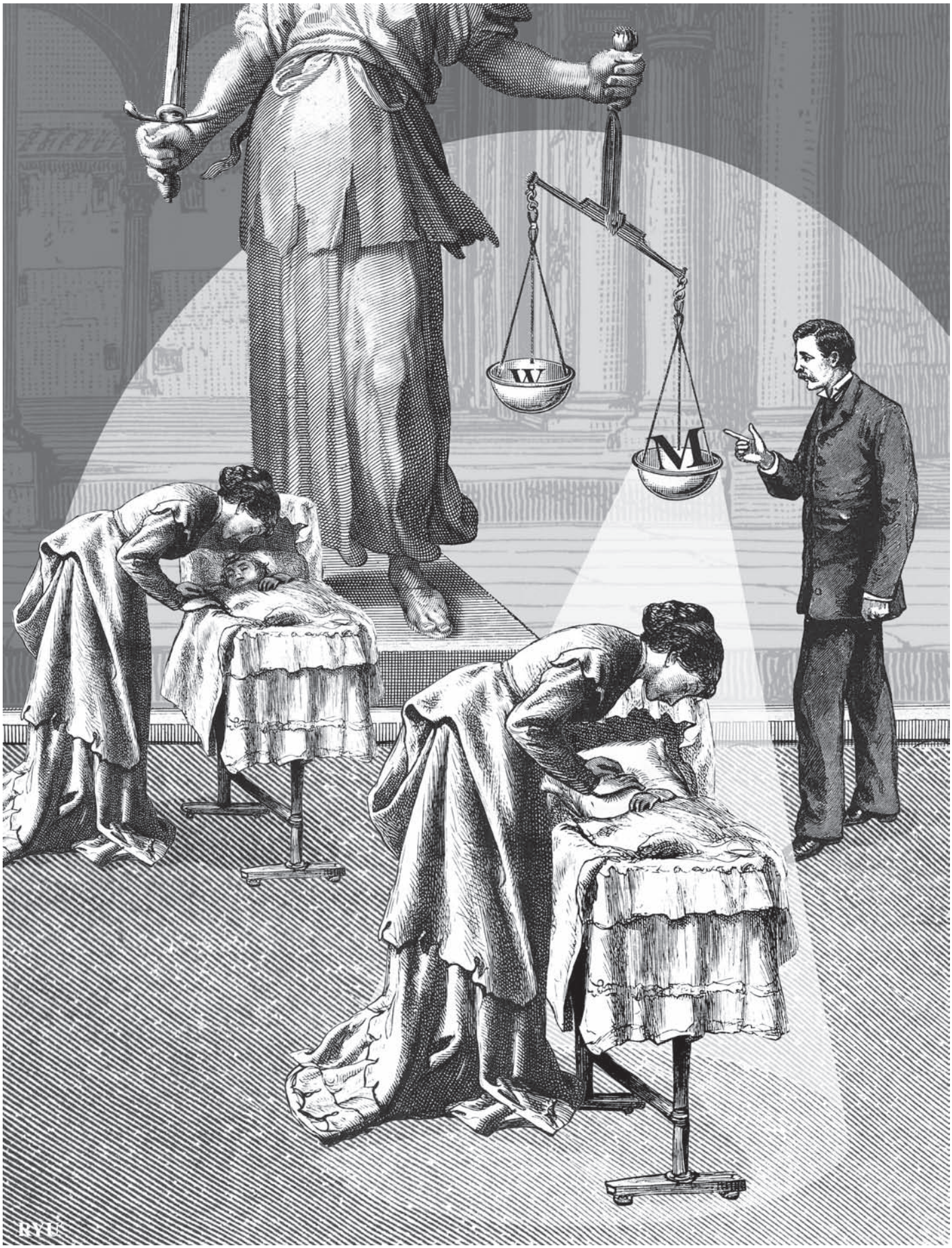
dus de kansverhouding tussen schuld en onschuld wordt volledig bepaald door de verhouding tussen $P(M)$ en $P(W)$. De berekening van Sir Roy Meadow had op deze verhouding geen betrekking, en was voor de schuldvraag dus simpelweg niet relevant. Ondanks het feit dat de kans in (1) extreem klein is, is met de beste wil van de wereld dat niet op te vatten als bewijs voor M.

Misschien dat de lezer bovenstaand argument niet helemaal overtuigend vindt vanwege het feit dat we spreken over de kans op W en de kans op M. Maar ook als je niet bereid bent om aan die kansen een betekenis toe te kennen gaat het mis. Immers, het bewijs E heeft zowel onder aanname van W als M conditionele kans 1: $P(E \mid W) = P(E \mid M) = 1$. Ook in die zin zien we dus dat het bewijs E *geen enkel onderscheid* kan maken tussen M en W. Het waargenomen bewijs is immers onder beide hypotheses even waarschijnlijk, in dit geval zelfs met een kans gelijk aan 1.

We kunnen op dit punt aangekomen dus rustig concluderen dat de procedure als voorgesteld door Fisher geen kwantificering van statistisch bewijs oplevert. Het



Figuur 1 Een Venn-diagram met alle relevante gebeurtenissen in de zaak Sally Clark.



Illustratie: Ryu Tajiri

heeft daar in feite eigenlijk weinig mee te maken. Het feit dat zich iets bijzonders heeft voorgedaan is in zichzelf nooit reden tot veel zorg. Als ik morgen de Staatsloterij win is dat ook een bijzondere gebeurtenis, en de kans dat ik win gegeven dat ik niet vals speel is erg klein. Maar dat feit impliceert natuurlijk niet dat ik waarschijnlijk gefraudeerd zal hebben.

Wat zegt een p-waarde dan wel?

Je kunt je met recht afvragen hoe het mogelijk is dat p-waardes toch als kwantificering van statistisch bewijs gezien worden. Hoe kunnen voorbeelden als zojuist gegeven aan de aandacht van al die wetenschappers zijn ontsnapt? Deze vraag heeft denk ik geen enkelvoudig antwoord, maar misschien komen we iets verder door na te gaan wat een p-waarde dan *wel* zegt.

Een p-waarde vertelt de onderzoeker wat de kans is dat onder aanname van hypothese H de uitkomst ten minste zo extreem is als wat is waargenomen. Als de p-waarde onder een van tevoren afgesproken drempel blijft, het significantieniveau α, dan vindt men de uitkomst dermate extreem onder H dat H niet langer geloofd wordt, en dus moet worden verworpen.

Ik merkte al eerder op dat er nog een ander type fout bestaat, namelijk H ten onrechte niet verwerpen. Er is geen wiskundige reden waarom het ene type fout anders behandeld zou moeten worden dan het andere type, en de asymmetrie is een keuze van 'de' wetenschap, of van verdere ethische of morele overwegingen. Het feit dat de kans op onterechte verwerping van H gecontroleerd moet worden, terwijl de kans op onterecht accepteren alleen maar zo klein mogelijk moet worden gemaakt binnen de ruimte die de eerste eis nog toelaat, duidt er op dat men bijzondere betekenis toekent aan het eerste type fout. Deze bijzondere betekenis is dat men simpelweg denkt dat α een maat voor het *bewijs* voor het complement van H is. Met die interpretatie is het logisch dat men die fout wil kunnen controleren omdat wetenschappelijk bewijs hoge standaarden vereist. Een eventueel juiste conclusie niet kunnen concluderen vanwege *gebrek aan bewijs* is minder erg dan een onterechte conclusie vanwege misleidend bewijs. In een juridische context staat H vaak voor de onschuldhypothese van de verdachte. Deze zal alleen verworpen worden als er echt bewijs is, en ook in deze situatie wordt onterecht verwerpen van H erger gevonden dan het niet verwerpen van H terwijl H niet waar is. Dat is maatschappelijk heel begrijpelijk, en is misschien een van de redenen van de populariteit van p-waardes.

Als een onderzoeker de geschetste procedure volgt, dan weet hij of zij dat in hooguit een fractie α van de keren dat H waar was, deze ondanks dat toch wordt verworpen. Dat is een frequentistische uitspraak, die iets zegt over de kwaliteit van de procedure als geheel. Wat echter cruciaal is, is dat als het experiment eenmaal uitgevoerd is en de data verkregen, er niet zomaar een uitspraak gedaan kan worden over de kans dat je in *dat geval* de juiste beslissing hebt genomen. Een p-waarde is gericht op de procedure, terwijl statistisch bewijs zich juist op de specifiek verkregen data moet richten van een enkel experiment. Op die manier bezien richt een p-waarde zich gewoon op een andere vraag dan de vraag naar statistisch bewijs.

Een kansuitspraak over de juistheid van de beslissing kan alleen gegeven worden als je bereid en in staat bent om van tevoren vast te stellen wat de kans is dat H optreedt. In het voorbeeld van Sally Clark zoals ik dat boven beschreef lijkt dat zeker niet onredelijk. Je kunt je immers afvragen hoe groot de kans is dat een moeder haar twee kinderen ombrengt door, bijvoorbeeld, naar statistieken te gaan kijken. Maar bij veel situaties lijkt een uitspraak over de kans op H niet zinvol. Als H bijvoorbeeld de hypothese is dat een nieuw medicijn niet beter werkt dan een oud, hoe zou je dan de kans daarop moeten inschatten? In feite is het formalisme van Fisher en de hele p-waarde-technologie juist ook ontworpen om een statistische uitspraak te kunnen doen zonder daar rekening mee te moeten houden, maar we zien dat dit dus niet zo goed werkt.

Een p-waarde is dus primair een uitspraak over de procedure, en geen uitspraak die in een specifiek geval veel zegt. In het voorbeeld van Sally Clark zagen we dat om een kansuitspraak te doen over de schuldvraag gegeven het bewijs, het simpelweg onvoldoende is om de p-waarde te berekenen. Bij een gegeven realisatie van het experiment kunnen we met een p-waarde niet uitdrukken wat de kans is dat we het goed hebben gedaan, en als kwantificering van bewijswaarde zijn p-waardes dus totaal ongeschikt. In de wetenschap is dat natuurlijk een belangrijke conclusie, maar in het recht zijn de catastrofale gevolgen van statistisch onbegrip misschien nog wel ernstiger. Als een kleine p-waarde als bewijs tegen de onschuldhypothese wordt opgevat, dan wordt de p-waarde voor bewijsdoeleinden gebruikt waar ze niet geschikt voor is.

Er is nog een gevaar bij het gebruik van p-waardes. Ik schreef al dat onderzoekers graag de p-waarde vermelden van hun resultaten. Wat men dan vaak doet is gewoon melden dat, bijvoorbeeld, de p-waarde $0{,}004$ is. Deze uitspraak suggereert dat het significantieniveau op $0{,}004$ gezet kan worden, maar dat is onjuist. Het significantieniveau kies je van tevoren en is een getal dat iets zegt over de kwaliteit van de gehele procedure, en als je die eenmaal hebt gekozen, dan is dat wat het is. Als je de gerapporteerde p-waarde per experiment verandert, dan pleeg je dus eigenlijk wetenschappelijke fraude omdat je een bewijswaarde claimt die niet klopt.

Paradoxen

Ik heb tot nu toe enkele argumenten gegeven tegen het gebruik van p-waardes als kwantificering van sterkte van bewijs. Het is instructief en onderhoudend om te zien wat er zou gebeuren wanneer je een dergelijke interpretatie wel toe zou staan. Ik zal met drie voorbeelden laten zien dat dit tot absurde situaties leidt. Dat is niet verwonderlijk, want als je uitgangspunt onjuist is, dan kun je alles verwachten.

Stiekem kijken

Stel een wetenschapper wil hypothese H ontkrachten. Laten we voor het gemak aannemen dat hij wil aantonen dat de succeskans van een bepaald experiment niet gelijk is aan $\frac{1}{2}$. Stel hij doet 20 experimenten en stel dat dit tot 14 successen leidt. Laten we $\alpha = 0{,}05$ nemen. Wat is de kans dat, onder de aanname H dat de succeskans $\frac{1}{2}$ is, de afwijking van wat we verwachten (in dit geval 10) minstens 4 is? Een kleine berekening met de binomiale verdeling laat zien dat deze kans gelijk is aan $0{,}115$, dus te groot om H te verwerpen. Echter, de wetenschapper merkt op dat als hij 15 successen zou hebben gezien, de bijbehorende p-waarde $0{,}041$ geweest zou zijn, en dat is wel klein genoeg om H te verwerpen. Het verwerpen van H is dus eigenlijk net niet gelukt.

De wetenschapper besluit hierop om 20 extra experimenten te doen. Stel nu eens dat die tweede serie voor hem een stuk beter verloopt, en dat hij maar liefst 19 successen ziet. Zijn gezamenlijke score is nu 33 successen in 40 experimenten, en de bijbehorende p-waarde is $0{,}0000423$, zoals weer eenvoudig is uit te rekenen. De wetenschapper concludeert nu dat de gezamenlijke score van de twee series zodanig is dat H (ruimschoots) kan worden verworpen.

Is dit een correcte gang van zaken? Nee. Al zou de wetenschapper een tweede serie van 100.000 experimenten hebben gedaan met daarin 100.000 successen, dan nog zou hij volgens het p-waardeparadigma niet hebben mogen concluderen dat H verworpen kan worden. Waarom? Wel, de wetenschapper dient voorafgaand aan het experiment het kritieke gebied te bepalen. Dit kritieke gebied moet zodanig zijn dat de kans om er onder H toch in terecht te komen, hooguit α is. Maar aangezien de wetenschapper H al verworpen zou hebben als er na 20 experimenten aanleiding toe was geweest, was de kans om dat na 20 experimenten ten onrechte te doen al $0{,}05$. Na de tweede serie kan die kans alleen maar groter worden, en zal dus uiteindelijk groter zijn dan de toegestane α.

Nu begrijpt natuurlijk iedereen dat een serie van 100.000 worpen met alleen maar successen bewijs *moet* zijn tegen de hypothese H dat de succeskans $\frac{1}{2}$ is. Het feit dat deze procedure dan toch niet toestaat om H te verwerpen is dan ook geen gevolg van een slechte intuïtie, maar van een ondeugdelijke procedure. Elke redelijke procedure zou dit als overweldigend bewijs tegen H moeten zien, maar de procedure met p-waardes doet dat niet. De ondeugdelijkheid van p-waardes als bewijswaarde is hiermee opnieuw geïllustreerd.

Meerdere hypotheses

Als $q \leq \alpha$ de kans voorstelt op onterecht verwerpen van H, dan zal zo'n een op de $1/q$ correcte hypotheses toch verworpen worden. Stel je wilt aantonen dat iemands favoriete kleur van invloed is op zijn of haar kans om kop te gooien met een gegeven munt. Het is eenvoudig uit te rekenen dat onder de hypothese dat de kans op kop $\frac{1}{2}$ is, de kans om 0, 1, 9 of 10 keer kop te gooien ongeveer $0{,}0215$ is, en dus vormt $\{0,1,9,10\}$ een geschikt kritiek gebied voor iedere afzonderlijke kleur. Stel we kiezen $1/0{,}0215 \approx 47$ mensen uit, die elk een andere favoriete kleur hebben, en iedereen gooit 10 keer met dezelfde munt. Er zal dan naar verwachting één persoon zijn die in dit kritiek gebied terecht komt. Als dat toevallig de persoon is die van groen houdt, dan kan de wetenschapper publiceren dat groen als favoriete kleur de kans op kop bij het gooien van een munt beïnvloedt. De hypothese H dat iemand die van groen houdt kans $\frac{1}{2}$ heeft op het gooien van kop wordt dan verworpen.

Natuurlijk is dit een anekdotisch voorbeeld, maar onderschat het niet. De neiging om alleen te publiceren als het een keertje lukt om een hypothese te verwerpen leidt tot *publication bias*, waarmee de geloofwaardigheid van de wetenschap in het algemeen in het geding komt. De problemen zijn eigenlijk zelfs nog groter dan dat. We kunnen ook de hypothese H' onderzoeken dat de succeskans $\frac{1}{2}$ is bij *alle* kleuren. Het gebied waarin iemand 0, 1, 9 of 10 keer kop ziet is nu veel te groot: de kans dat iemand van de twintig personen daarin terecht komt is ongeveer $\frac{1}{3}$ zoals je makkelijk kunt uitrekenen. Een correct en redelijk kritiek gebied wordt nu gevormd door die uitkomsten waarbij er *iemand* is die alleen maar kop of alleen maar munt gooit. Inderdaad is de kans daarop ongeveer $0{,}04$, en daarmee dus kleiner dan het significantieniveau $\alpha = 0{,}05$. Maar stel nu eens dat de persoon die groen mooi vindt 9 keer kop ziet en alle anderen nooit alleen maar kop of alleen maar munt. Als we alleen maar kijken naar wat groen heeft gegooid, dan wordt de hypothese H verworpen, zoals we net zagen. Maar als we nu alle andere kleuren ook in beschouwing nemen, dan zien we dat de hypothese H' *niet* verworpen wordt. Dat is op zijn zachtst gezegd vreemd: we kunnen H' niet verwerpen, dus we verwerpen niet dat *iemand* een afwijkende succeskans heeft, maar tegelijkertijd leidt concentreren op groen ertoe dat we wel verwerpen dat groen een afwijkende succeskans heeft. Vooral voor de groene persoon is dat nogal merkwaardig: zijn eigen data is onveranderd, maar puur en alleen omdat anderen ook aan het gooien zijn geslagen moet de conclusie over groen ook aangepast worden.

Deze merkwaardige situatie illustreert andermaal dat p-waardes eigenschappen hebben die op gespannen voet staan met enkele elementaire principes waaraan statistisch bewijs zou moeten voldoen.

Eenzijdig versus tweezijdig

Stel opnieuw dat we geïnteresseerd zijn in de onbekende succeskans p van een bepaald experiment. Als we het experiment 100 keer uitvoeren en we nemen significantieniveau $\alpha = 0{,}05$ dan leert een eenvoudige berekening dat we de hypothese H dat $p = \frac{1}{2}$ verwerpen als het aantal keer succes minstens 61 of hoogstens 39 is. Soms echter is er reden om *eenzijdig* te toetsen. We nemen dan de hypothese H' dat $p \leq \frac{1}{2}$ en we verwerpen H' als het aantal successen te hoog is. Een korte berekening laat zien dat een geschikt kritiek gebied nu gegeven wordt door $\{59,60,\ldots,100\}$, want de kans om onder H' in dit gebied terecht te komen is maximaal als $p = \frac{1}{2}$ en dan kleiner dan $0{,}05$.

Welnu, wat gebeurt er nu als het aantal waargenomen successen gelijk is aan 60? In dat geval verwerpen we $p = \frac{1}{2}$ niet maar we verwerpen *wel* dat $p \leq \frac{1}{2}$. Dat is onbegrijpelijk, want $p = \frac{1}{2}$ is een veel sterkere hypothese dan $p \leq \frac{1}{2}$. Maar het is wel het gevolg van een procedure die nog steeds door het wetenschappelijk establishment wordt geaccepteerd.

Een alternatief

Is er een alternatief voor de p-waarde als maat voor statistisch bewijs? Het antwoord is ja, maar dit antwoord komt wel met een prijs. De notie van statistisch bewijs die ik hier kort wil introduceren veronderstelt namelijk dat statistisch bewijs *uitsluitend relatief* is. In plaats van te zeggen dat data E al dan niet bewijs voor H oplevert, stelt deze benadering dat je alleen kunt zeggen dat de data E bewijs voor of tegen H_1 op kan leveren ten opzichte van een andere hypothese H_2.

Wat een *likelihood* benadering van statistisch bewijs concreet behelst, is het uitrekenen van het quotiënt

$$LR_{H_1,H_2}(E) := \frac{P(E \mid H_1)}{P(E \mid H_2)},$$

oftewel de verhouding van de kans op de geobserveerde data E onder H_1 en de kans hiervan onder H_2. Als dit quotiënt groter is dan 1, dan ondersteunt de data H_1 ten opzichte van H_2, en als het kleiner is dan 1, dan ondersteunt de data H_2 ten opzichte van H_1. Als het quotiënt gelijk is aan 1 is de data neutraal en geeft de data geen manier om onderscheid te maken tussen H_1 en H_2.

We hebben in deze bijdrage al een voorbeeld gezien van een dergelijke *likelihood*

ratio. In het voorbeeld van Sally Clark zagen we dat $P(E \mid W) = P(E \mid M) = 1$, zodat de likelihood ratio gelijk is aan 1, en de data E dus geen enkele informatie geeft over W versus M. Deze conclusie hadden we al eerder getrokken.

Waarom insisteert deze benadering op het relatief zijn van statistisch bewijs? Een eenvoudig voorbeeld helpt om dit te begrijpen. Stel we hebben een vaas met 100 ballen, allemaal wit of zwart. Hypothese H_1 zegt dat er 90 witte en 10 zwarte ballen zijn, hypothese H_2 zegt dat er 50 witte en 50 zwarte ballen zijn, en hypothese H_3 zegt dat alle ballen wit zijn. Stel we trekken 10 ballen met teruglegging, en deze 10 ballen zijn allemaal wit. Geeft deze data E bewijs voor bijvoorbeeld H_1? Om hier iets over te zeggen berekenen we $P(E \mid H_1) = (9/10)^{10}$, $P(E \mid H_2) = (\frac{1}{2})^{10}$ en $P(E \mid H_3) = 1$. We vinden dan

$$LR_{H_1,H_2}(E) \approx 357$$

en

$$LR_{H_1,H_3}(E) \approx 0{,}349.$$

De data ondersteunt H_1 meer dan H_2, maar H_3 meer dan H_1. De vraag of de data bewijs voor H_1 oplevert is dus simpelweg niet te beantwoorden, omdat het ervan afhangt waar je H_1 mee wilt vergelijken.

Ook in een juridische context is het ontzettend belangrijk om het bewijs in een bepaalde zaak te beschouwen vanuit de schuldhypothese, maar ook vanuit de onschuldhypothese, dit ter voorkoming van een tunnelvisie zoals we in het voorbeeld van Sally Clark al hebben gezien. Het gebruik van een likelihood ratio in die context is inmiddels gelukkig standaard.

Is een likelihood ratio nu een goede afspiegeling van het bewijs voor H_1 ten opzichte van H_2? We kunnen hier enig inzicht in krijgen door na te gaan of de bezwaren tegen p-waardes die ik heb aangedragen bij likelihood ratio's niet bestaan. Allereerst merk ik op dat een likelihood ratio alleen afhangt van de waargenomen data E, dus het eerste bezwaar tegen p-waardes dat ik aanvoerde geldt hier niet. Het gaat immers precies om de kans op E gegeven de twee hypotheses.

Wat betreft het 'stiekem kijken': dat bezwaar valt weg bij het gebruik van likelihood ratio's. De onderzoeker kan gerust zijn of haar gehele data gebruiken, en er is geen reden om bijzondere maatregelen te nemen omdat hij of zij halverwege het proces al eens eerder naar de data heeft gekeken. Natuurlijk kan de onderzoeker proberen om net zo lang door te gaan met experimenten tot de likelihood ratio de gewenste kant op wijst. Maar dat is wezenlijk anders dan het stiekem kijken wat ik eerder beschreef. Bij het stiekem kijken speelt de onderzoeker vals doordat hij geen volledige openheid geeft over de statistische procedure zoals deze is uitgevoerd. De bewijskracht (in termen van de p-waarde) hangt af van eerdere afspraken die niet te controleren zijn, en waarbij ook te goeder trouw makkelijk fouten worden gemaakt. Als een onderzoeker stopt als de likelihood ratio de gewenste kant op wijst, dan is het simpelweg zo dat de bewijswaarde is wat hij is. Dat is dus niet vals spelen, want het is echt waar wat de likelihood ratio zegt en iedereen kan dat controleren. Het aardige van likelihood ratio's is overigens dat als een onderzoeker H_1 met H_2 wil vergelijken, de kans dat de likelihood ratio *ooit*, na hoeveel pogingen ook, meer dan een factor k de verkeerde kant op zal wijzen, kleiner is dan $1/k$.

Het probleem dat bij de beschouwing van meerdere hypotheses ontstond, bestaat simpelweg niet bij likelihood ratio's. Natuurlijk kan een likelihood ratio de verkeerde kant op wijzen, en bijvoorbeeld kleiner dan 1 zijn terwijl toch H_1 waar is. Dat is geen fout van de methodiek maar een logisch gevolg van het feit dat het om kanstheoretische zaken gaat waarin bewijs gewoon 'toevallig' de verkeerde kant kan op wijzen. Dat is de natuur der dingen en kan en hoeft niet ondervangen te worden.

De paradox die we tegenkwamen bij het verschil tussen één- versus tweezijdig toetsen ten slotte, bestaat bij likelihood ratio's ook niet. Wat we wel zien, en dat zou als zwakte maar net zo goed als kracht gezien kunnen worden, is dat we bij het bepalen van een likelihood ratio geen samengestelde hypothese kunnen nemen zoals $H' : p \leq \frac{1}{2}$. We moeten $H : p = \frac{1}{2}$ vergelijken met een specifieke keuze voor het alternatief.

In de praktijk wordt het werken met een likelihood ratio vaak ingebed in de zogenaamde *odds*-vorm van de regel van Bayes:

$$\frac{P(H_1 \mid E)}{P(H_2 \mid E)} = \frac{P(E \mid H_1)}{P(E \mid H_2)} \times \frac{P(H_1)}{P(H_2)},$$

waarin de *posterior odds* $P(H_1 \mid E)/P(H_2 \mid E)$ uitgedrukt worden als het product van de likelihood ratio en de *prior odds* $P(H_1)/P(H_2)$. Om deze reden wordt het werken met de likelihood ratio ook wel 'Bayesiaanse kansrekening' genoemd, maar dat is feitelijk nogal vreemd, want kansrekening impliceert de regel sowieso. Deze naamgeving is bovendien des te ongelukkiger omdat ze verwarring met Bayesiaanse statistiek in de hand werkt. Wel is het zo dat het gebruik van de regel doorgaans impliceert dat kansen subjectief geïnterpreteerd dienen te worden en niet frequentistisch. Statistisch bewijs is in deze visie een kwestie van informatie: als er meer informatie beschikbaar is, verandert de waarde van het bewijs, en zal de likelihood ratio dus anders worden.

Over deze interpretatie, de eigenschappen van de likelihood ratio en het gebruik ervan is ongelooflijk veel te vertellen. In 2020 publiceer ik hier samen met Klaas Slooten een boek over [3]. Hopelijk helpen al deze inspanningen om het gebruik van p-waardes als bewijs terug te dringen, en om met andere ogen naar statistisch bewijs te gaan kijken. Het zou heel goed zijn als de p-waarde uit het curriculum van de middelbare school, hogeschool en universiteit gehaald zou worden, of op zijn minst zou worden genuanceerd. Ik begrijp natuurlijk dat dit niet op heel korte termijn zal gebeuren, en dat het voorlopig nog zo zal zijn dat we leerlingen en studenten methodes meegeven die niet geschikt zijn voor de geclaimde doeleinden. Mijn advies is om dan zo goed en zo kwaad als het gaat het belang van deze p-waardes te nuanceren. Zolang de procedure op het examen gevraagd kan worden moeten we uitleggen *hoe* het werkt. Maar het lijkt me dat we best kritisch mogen zijn over *wat* de procedure doet. De enige route die ik niet acceptabel vind, is net doen alsof er niets aan de hand is.

Referenties

1 J.P.A. Ioannidis, Why Most Published Research Findings are False, *PlosMed* 2(8) (2005), e124.

2 R.A. Fisher, *Statistical Methods for Research Workers*, Vol. 13, Oliver and Boyd, 1925.

3 R. Meester en K. Slooten, *Theory and Philosophy of Statistical Evidence in Forensic Science*, Cambridge University Press, verwacht in 2020.