Kan dat geen toeval zijn?

Kan dat geen toeval zijn?

Een kritische blik op statistische bewijsvoering

Ronald Meester en Klaas Slooten

Amsterdam University Press

Ontwerp omslag: Gijs Mathijs Ontwerpers
Ontwerp binnenwerk: Crius Group, Hulshout

ISBN 978 94 6372 508 8
e-ISBN 978 90 4855 658 8
DOI 10.5117/9789463725088
NUR 916

Inhoudsopgave

Voorwoord

Dit boek is bedoeld voor iedereen die op de een of andere manier te maken heeft met, of geïnteresseerd is in, statistisch bewijs: wetenschappelijke onderzoekers, studenten, docenten, wiskundigen, scholieren, filosofen, juristen, managers, en vast nog veel meer mensen. Het is geen gewoon boek over statistiek, en het is zeker geen receptenboek dat de lezer gaat vertellen welke test bij welk probleem hoort, of welke software voor welk probleem gebruikt kan worden. Wat is het dan wel voor boek? Dat zullen we nu uitleggen.

In dit boek leggen we uit hoe statistische bewijsvoering in zijn werk gaat. We laten zien wat voor vragen er worden gesteld, en wat de achterliggende logica is achter de manier waarop we die vragen proberen te beantwoorden. In grote lijnen zijn er twee manieren om statistische bewijsvoering te doen, die beide oude wortels hebben. De ene manier is om een onderzoekshypothese als bewezen te beschouwen als de data die je verkregen hebt, heel slecht passen bij het niet waar zijn van de onderzoekshypothese, en de andere bestaat uit het zoeken naar wat de beste verklaring is voor de data die je wilt duiden. Deze vormen van statistiek staan bekend als respectievelijk de klassieke (of frequentistische) aanpak, en de Bayesiaanse aanpak. In tegenstelling tot wat deze naamgeving suggereert, heeft de Bayesiaanse aanpak de oudste fundamenten: deze gaat terug tot Thomas Bayes (1702-1761), terwijl de frequentistische aanpak is ontwikkeld vanaf het begin van de twintigste eeuw.

Er bestaan veel leerboeken in de statistiek die je vertellen hoe je, met de klassieke aanpak, data zou moeten interpreteren, hoe je hypotheses moet opstellen, hoe je een test kunt ontwerpen, en hoe je kunt besluiten of je een hypothese al dan niet verwerpt. Echter, in dit boek willen we de lezer laten zien dat bepaalde aspecten van deze klassieke manier om bewijs te interpreteren problematisch zijn. Het gevolg hiervan is dat gerapporteerde claims ongefundeerd kunnen zijn. Deze problemen rond statistisch bewijs zijn mede de oorzaak van wat de 'replicatiecrisis' wordt genoemd: de kennelijke onmogelijkheid om onderzoeksresultaten te reproduceren. Het lukt dan dus niet om met een tweede onderzoek de conclusie te bevestigen die uit het eerste onderzoek is getrokken. Wat als wetenschappelijk 'bewezen' werd beschouwd, blijkt dat bij nader inzien misschien toch niet te zijn.

Dit is een serieus probleem. De betrouwbaarheid en status van statistisch verkregen bewijs staan op het spel. Dit heeft mogelijk grote gevolgen, niet alleen voor de wetenschap, maar ook voor de samenleving als geheel,

bijvoorbeeld voor de gezondheidszorg of de rechtspraak. Data worden immers niet alleen vergaard om te onderzoeken of een bepaalde hypothese correct is, maar om aan de correctheid (of onjuistheid) ervan conclusies te verbinden. En dit kan op allerlei terreinen gebeuren, want de data die statistisch worden geëvalueerd kunnen van alles betreffen. Telkens als deze data mede door het toeval worden bepaald, heb je een statistische interpretatie nodig. Dat kan gaan over de effectiviteit van een medicijn, data over het verband tussen voeding en gezondheid, opiniepeilingen, of data over botsende elementaire deeltjes in een deeltjesversneller: telkens worden de vergaarde data mede door het toeval bepaald.

In dit boek zullen we op een leesbare en begrijpelijke manier uitleggen hoe het mogelijk is dat het bedrijven van statistiek met zo veel problemen gepaard gaat. We zullen hierbij over veel dingen na moeten denken. Wat is dat eigenlijk, statistisch bewijs? Dat lijkt misschien een eenvoudige vraag, maar we zullen zien dat hij een stuk lastiger te beantwoorden is dan je misschien op het eerste gezicht denkt. We zullen daarom teruggaan tot de essentie, namelijk de manier van denken en de redeneringen aan de basis van de statistiek, omdat deze ons duidelijk maken wat we van statistiek wel, en wat we juist niet mogen verwachten. Op die manier zullen we uitleggen dat de klassieke manier om hypotheses te toetsen géén antwoord geeft op de vraag of de data bewijs leveren voor of tegen een hypothese, en dat die daarom ontoereikend is om daar conclusies over te trekken. Vervolgens zullen we laten zien hoe nadenken over de aard van statistisch bewijs ons op het juiste pad kan brengen om op een andere manier te redeneren en tot conclusies te komen die wél logisch gerechtvaardigd zijn.

Uiteraard zijn we niet de eersten die kritiek hebben op de huidige statistische praktijk, en een aantal van de problemen die we in dit boek signaleren zijn door collega's al onderkend. We noemen er een paar. Biostatisticus Richard Royall uitte in 1996 al stevige kritiek op de klassieke statistiek [27]. In 2005 publiceerde de epidemioloog John P.A. Ioannidis een artikel waarin hij beweerde dat de meeste gepubliceerde wetenschappelijke onderzoeksresultaten niet waar zijn [18]. Recentelijk verwees statisticus William Briggs het gebruik van de klassieke *p*-waardes naar de prullenbak [6]. Er zijn ook tijdschriften die het gebruik van de klassieke *p*-waardes helemaal in de ban hebben gedaan, zoals bijvoorbeeld *Basic and Applied Psychology*, *Epidemiology* en *Political Analysis*, zie [10], [15] en [30]. Dat er ondanks deze – en veel vergelijkbare – publicaties tot nu toe weinig is veranderd, is opmerkelijk, zeker als je in ogenschouw neemt dat er ook oplossingen zijn voorgesteld, waarover later meer.

Onze bijdrage is in zoverre afwijkend van de eerdergenoemde publicaties, dat wij niet in de eerste plaats op de wiskunde focussen, maar vooral uitleggen hoe de klassieke statistiek werkt, waarom en wanneer dit soms niet goed gaat, en wat je in de plaats daarvan zou kunnen doen. We willen graag laten zien dat de statistiek zoals wij die voorstaan heel goed overeenkomt met onze intuïtie over wat we eigenlijk kunnen verwachten van statistisch onderzoek. We leggen uit wat we wel, maar ook wat we juist niet van statistisch bewijs kunnen of mogen verwachten; in welke mate er objectiviteit te vinden is of na te streven is, waar objectiviteit eindigt, en of het nu eigenlijk een voordeel of een nadeel is dat een methode objectief is. Telkens nemen we de lezer mee in het waarom van onze benadering, en leggen de vinger exact op de zere plekken. Dit boek is daarom zowel filosofisch als toegepast van aard. We hebben in de literatuur geen andere tekst gevonden met deze benadering, ondanks dat dit volgens ons de beste manier is om de aard van statistische redeneringen te begrijpen. We hopen dan ook dat dit boek iets bij zal dragen aan een meer realistische kijk op statistisch bewijs.

Het boek is als volgt opgebouwd. In Deel I bespreken we de klassieke statistiek. In het eerste hoofdstuk bespreken we significantietoetsen. We concluderen allereerst dat veel gerapporteerde claims onjuist zijn, en geven een huiveringwekkend voorbeeld van wat er mis kan gaan, uit de juridische praktijk. In Hoofdstuk 2 komt de *p*-waarde uitgebreid aan bod. We leggen uit wat een *p*-waarde eigenlijk is, en behandelen waarom deze *p*-waarde eigenlijk niet goed gebruikt kan worden om sterkte van statistisch bewijs te meten. Als *p*-waardes, zoals vaak gebeurt, wel als zodanig opgevat worden, kun je dus problemen verwachten. Hoofdstuk 3 gaat over betrouwbaarheidsintervallen. Deze zijn nauw verbonden met *p*-waardes en kennen gelijksoortige interpretatieproblemen, waar we ook uitgebreid op in zullen gaan.

In Deel II bespreken we een meer Bayesiaanse benadering. We beginnen in Hoofdstuk 4 met uiteenzetten wat statistisch bewijs dan eigenlijk wél is, en we introduceren daarbij een instrument om de sterkte van het bewijs mee uit te drukken, de zogeheten likelihood ratio. We concluderen dat statistisch bewijs bijna nooit absoluut kan zijn, maar alleen relatief ten opzichte van een alternatieve verklaring. In Hoofdstuk 5 geven we een aantal concrete voorbeelden van likelihood ratio's in actie, en komen we ook toe aan het belangrijke verschil tussen bewijs en overtuiging. Daarna besteden we in Hoofdstuk 6 aandacht aan de vraag in hoeverre data misleidend kunnen zijn. Immers, het is onredelijk te verwachten dat elke snipper data je tot de waarheid zal leiden omtrent hoe die data tot stand zijn gekomen, maar je kunt je wel afvragen hoe vaak het hoe erg mis kan gaan. Daar blijken grenzen voor te bestaan, gelukkig. Ook zullen we zien hoe de likelihood

ratio ingezet kan worden om goede onderzoeksopzetten te ontwerpen. Dit besluit Deel II.

Deel III is gewijd aan de statistische praktijk. In Hoofdstuk 7 bespreken we een uitgebreide statistische analyse van twee concrete situaties. In Hoofdstuk 8 leggen we uit dat er situaties zijn waarin *p*-waardes, ondanks alle methodologische problemen die ermee samenhangen, soms pragmatisch toch gebruikt kunnen worden.

Het boek wordt wiskundig gezien nergens moeilijker dan een eerste cursus in toegepaste statistiek; de meest ingewikkelde wiskunde bestaat uit het uitrekenen van wat integralen, het rekenen met conditionele kansen, en het toepassen van de *abc*-formule voor tweedegraads vergelijkingen.

Echter, de wiskunde zelf is bij het begrijpen van statistiek vaak niet het grootste probleem. Statistiek is meer dan wiskunde. De moeilijkheid zit hem vooral in de betekenis van wat we doen. Dus ondanks het feit dat we de materie zo toegankelijk mogelijk presenteren, wordt van de lezer wel enige inspanning vereist. Wellicht zijn er passages die je een paar keer moet lezen, en is het soms ook nodig om met pen en papier even na te rekenen wat wij hebben gedaan. We denken dat de beloning hier ruimschoots tegen opweegt.

De lezer die bekend is met de mathematische statistiek zal wellicht verbaasd opmerken dat de concepten die we in dit boek propageren binnen die mathematische statistiek allang geaccepteerd zijn. Wat wij echter vooral hopen te bereiken is dat een overdenking van wat statistiek al dan niet vermag, welhaast automatisch naar een min of meer Bayesiaanse aanpak leidt, zonder daarbij overigens de nuttige aspecten van de klassiekere 'frequentistische' statistiek uit het oog te verliezen.

Sommige passages in dit boek zijn iets technischer van aard. Deze hebben we met * gelabeld – dat betekent dat de inhoud van die passage net even dieper gaat dan de rest van het boek, en dat je verder kunt gaan zonder dat stuk te lezen. Voor de liefhebbers raden we die passages natuurlijk wel aan. Wij wensen de lezer veel plezier, en we hopen dat het de inspanning waard zal zijn.

Verschillende mensen hebben ons erg geholpen met allerlei soorten commentaar op eerdere versies van dit manuscript. Onze dank gaat uit naar (in alfabetische volgorde) Marc Jacobs, Wouter Kager, Boukje Meester, Luit Jan Slooten, Robert van der Toorn, Wessel van Wieringen en Harry van Zanten.

Ronald Meester en Klaas Slooten