Statistiek in de rechtszaal

Kracht en grenzen van statistiek bij aanklacht en proces

Prof. dr. Herman Callaert

1. Juridische terminologie

In België is het de onderzoeksrechter die bij rechtszaken gelast is met het onderzoek. Hij kan een verdachte aanhouden voor een bepaalde tijd. Daarna moet de verdachte naar de raadkamer waar over het verdere lot wordt beslist. De raadkamer beslist of er al dan niet een vervolging voor een effectieve rechtbank zal gebeuren. De kamer van inbeschuldigingstelling dient als een beroepsinstelling van de raadkamer. Het is de raadkamer of de kamer van inbeschuldigingstelling die een verdachte naar de correctionele rechtbank stuurt of naar het assisenhof (enkel voor doodslag, moord of politieke misdrijven).

Het verhaal dat hieronder wordt verteld is gebaseerd op waar gebeurde feiten in de Verenigde Staten, waar het juridische systeem enigszins anders is. Om de zaak eenvoudig te houden hebben we de twee stappen in de procedure herleid tot gemakkelijk herkenbare namen. We gebruiken “kamer van inbeschuldigingstelling” voor de procedure waarbij onderzocht wordt of een verdachte al dan niet naar een rechtbank wordt doorverwezen. Als de verdachte wordt doorverwezen dan zal het hier naar het “hof van assisen” zijn omdat het over een aanklacht voor moord gaat.

2. Een verpleegster beschuldigd

Kristen Gilbert werkte in het midden van de jaren ’90 als verpleegster in het “Veteran’s Administration Hospital”, een ziekenhuis in Northampton, een stad in de staat Massachusetts van de Verenigde Staten.

Gilbert was erg gerespecteerd omwille van haar inzet en ervaring. Zij had de reputatie dat zij een crisis zeer goed kon inschatten. Wanneer een patiënt een hartstilstand kreeg dan was zij er als eerste bij en dan gaf ze onmiddellijk de ‘code blauw’, het signaal om de hulp in te roepen van het reanimatieteam. Zij bleef kalm en ze wist hoe ze epinephrine (een synthetische vorm van adrenaline) moest toedienen om het hart van de patiënt opnieuw op gang te brengen. Vaak leidde haar interventie tot een goede afloop voor de patiënt.

Na enige tijd kregen haar medeverpleegsters het vermoeden dat er iets niet klopte. Ze hadden de indruk dat er te veel ‘code blauw’ werd gegeven tijdens de dienst van Gilbert. De verdenkingen werden sterker naarmate er ook meer patiënten stierven. Harde feiten waren er echter niet, want niemand had haar ooit een dodelijke injectie zien toedienen. Globaal was er ook geen groot verschil tussen het aantal overlijdens in haar ziekenhuis en het aantal overlijdens in andere analoge ziekenhuizen.

De Raad van Bestuur van het “Veteran’s Administration Hospital” vertrouwde de zaak niet meer en gaf de verdachte situatie door aan onderzoeksrechter William Welch.

Onderzoeksrechter Welch verzamelde argumenten om de kamer van inbeschuldigingstelling te overtuigen dat Gilbert voor het hof van assisen moest gebracht worden. Welch zocht hierbij zowel naar een motief als naar getuigenissen en feiten die bezwarend konden zijn voor Gilbert.

Uit onderzoek naar haar privé-leven bleek dat Gilbert gescheiden was en moeder van twee jonge kinderen. Zij had een vaste job die goed betaalde en haar kwaliteiten als verpleegster waren overal gekend. Welk motief kon zij hebben om patiënten te vermoorden? Vermoorden inderdaad, want de fatale dosissen stimulerende medicijnen voor het hart die zij toediende waren zeker geen daden van euthanasie. Haar patiënten waren niet terminaal ziek, zij waren niet oud en zij hadden een relatief goede gezondheid. Daarom juist was de dood van die patiënten zo merkwaardig. Onderzoeksrechter Welch kwam te weten dat Gilbert hield van de actie tijdens een crisis. Zij stond er stralend bij, elke keer dat men bij een crisis moest erkennen “hoe snel zij er weer bij was” en “hoe professioneel zij het weer had aangepakt”. Bovendien kwam Welch te weten dat Gilbert een nieuwe vriend had die ook in dat ziekenhuis werkte, en dat zij op die manier indruk op hem wilde maken.

Naast dit mogelijke motief verzamelde Welch ook getuigenissen en feiten. Van collega’s die met Gilbert samenwerkten kwam hij te weten dat ze zonder moeite aan de nodige medicijnen kon geraken. Een dokter gaf meer uitleg over de symptomen die de overleden patiënten vertoonden. Verder onderzocht Welch gegevens over de dienstregeling in het ziekenhuis. Om dit cijfermateriaal goed te interpreteren kreeg hij hulp van de statisticus Stephen Gehlbach.

Waren het mogelijke motief, de getuigenissen van de dokter en van de collega’s, en het cijfermateriaal voldoende argumenten om de kamer van inbeschuldigingstelling te overtuigen? Het belangrijkste argument waarmee Welch de kamer van inbeschuldigingstelling overtuigde om de verpleegster Gilbert naar assisen door te verwijzen kwam van de statistiek.

3. Toetsen van hypothesen

De kamer van inbeschuldigingstelling moest beslissen of er een assisenproces tegen Gilbert zou komen. De belangrijkste vraag was dan ook: zijn er tijdens de dienst van Gilbert meer overlijdens dan gewoonlijk? Niet één of twee extra doden – want dat kon gemakkelijk te wijten zijn aan louter toeval – maar genoeg om verdacht te lijken? Als men dit niet kon aantonen dan was er niet genoeg bewijsmateriaal om haar aan te klagen.

Onderzoeksrechter Welch erkende dat de vraag “hoeveel extra doden zijn er nodig om verdacht te lijken?” alleen kon beantwoord worden met behulp van statistiek. Hij vroeg aan de statisticus Gehlbach om een analyse te maken van de ziekenhuisgegevens en om zijn bevindingen uit te leggen aan de rechters van de kamer van inbeschuldigingstelling.

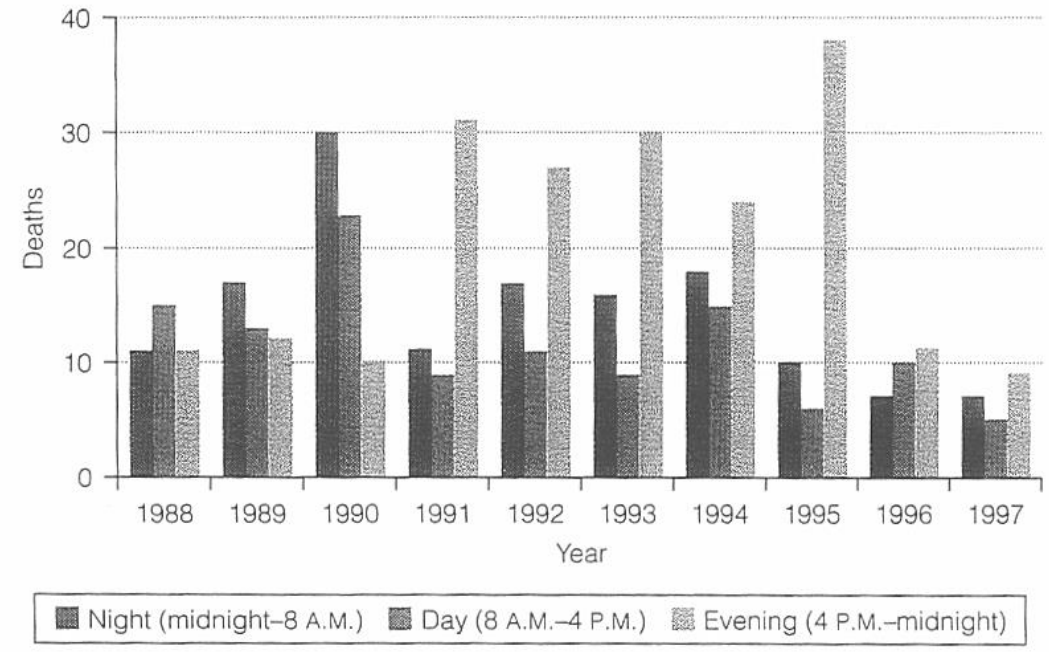
Hieronder vind je een overzicht van Gehlbach’s argumenten. Als je deze argumenten leest, beeld je dan in dat je één van de leden bent van de kamer van inbeschuldigingstelling. Vind je de argumenten sterk genoeg om Gilbert door te verwijzen naar de rechtbank?

De statistische methode die Gehlbach gebruikt heet ‘toetsen van hypothesen’. Het is een methode die gebruik maakt van kansrekening om te bepalen of een gebeurtenis al dan niet als “uitzonderlijk” kan worden beschouwd. Toetsen van hypothesen gebruikt een eigen manier

van redeneren die niet eenvoudig is, en er komen ook technische berekeningen bij kijken. Het was een uitdaging voor Gehlbach om aan mensen die weinig van statistiek kenden een duidelijke uitleg te geven. Hij deed die uitleg volledig mondeling, zodat hij zichzelf verplichtte om niet met details van berekeningen op de proppen te komen.

4. Het betoog van Gehlbach

De volgende paragrafen zijn een samenvatting van de drie delen van Gehlbach's betoog.
Het eerste deel gaat over het patroon van het aantal doden op de afdeling waar Gilbert werkte. Het tweede deel gaat over variabiliteit en p-waarden. Het laatste deel gaat over een statistische toets om te onderzoeken of het abnormaal aantal doden op Gilbert's afdeling op een duidelijke manier in verband kan gebracht worden met haar aanwezigheid.



Figuur 1: het patroon van dodenaantal per jaar en per shift

4.1. Het patroon van het dodenaantal

De rechters van de kamer van inbeschuldigingstelling kregen de bovenstaande figuur te zien. Gelbach gaf daarbij de volgende uitleg.

De grafiek die jullie hier zien, zijn de gegevens van het ziekenhuis waar Kristen Gilbert werkte. Elk blok van drie staven stelt 1 jaar voor, te starten bij 1988 tot en met 1997. Elke staaf stelt 1 shift voor. De hoogte van een staaf komt overeen met het aantal doden tijdens die shift, in dat jaar.

Bekijk nu de verandering van de staven van jaar tot jaar. De staven van de eerste twee jaren zijn kort, ruw geschat zo'n 10 doden per jaar per shift. Maar vanaf 1990 tot 1995 is er een drastische stijging: ieder jaar is er wel 1 shift met 25 tot 35 doden per jaar. Voor de volgende twee jaren zijn de staven alweer kort, net onder de 10 doden per jaar per shift.
Hoe staat dit verloop in verband met Kristen Gilbert's werktijd in het ziekenhuis? Gilbert begon in het ziekenhuis te werken vanaf maart 1990 en in februari 1996 is ze daar gestopt. Het dodenaantal per jaar volgt het patroon van haar werkgeschiedenis: kleine aantallen wanneer ze er niet werkte en grote aantallen wanneer ze er wel werkte.

Als je de staven afzonderlijk per shift bekijkt dan valt je nog wat anders op. Je ziet dat er altijd in 1 van de 3 shiften veel meer doden zijn dan in de andere twee. Voor 5 van de 6 jaar, van 1991 tot 1995, was dat de avondshift. Tijdens die 5 jaren deed Kristen de avondshift.

En wat kan je zeggen over de uitzondering 1990? Het was toen de nachtshift, die het grootste aantal overlijdens telde. In dat jaar had Kristen de nachtshift.

Op het eerste zicht blijkt er een duidelijk verband te zijn tussen het grote aantal doden en de werktijden van Gilbert. In principe zou dit nochtans het resultaat kunnen zijn van louter toeval. Je verwacht immers dat het aantal overlijdens in een ziekenhuis niet constant is en dat er schommelingen zijn van shift tot shift en van jaar tot jaar. Om aan te tonen dat het aantal sterfgevallen zo abnormaal is dat het eigenlijk niet meer aan louter toeval kan worden toegeschreven heb je een statistische toets nodig. Om de rechters voor te bereiden op dit soort statistisch redeneren, legde Gehlbach eerst de basisregels ervan uit aan de hand van een eenvoudig voorbeeld dat iedereen zich kon voorstellen.

4.2. Variabiliteit en p-waarde

Om het basisidee van een statistische toets te begrijpen kan je denken aan een eenvoudig kansexperiment waarbij je lukraak een kaartje uit een vaas trekt. In die vaas zitten 1000 kaartjes, witte en zwarte. Danny beweert dat het er 500 witte en 500 zwarte zijn, maar je weet niet of je de bewering van Danny wel mag geloven. Je mag die kaartjes in de vaas niet gaan natellen, maar je mag wel iets anders doen. Alle kaartjes zijn even groot en zij worden perfect door elkaar gemengd. Dan mag jij geblinddoekt een kaartje uit die vaas trekken, de kleur noteren en dan het kaartje terug in de vaas leggen. Daarna worden de kaartjes weer goed gemengd, je trekt terug lukraak een kaartje,noteert de kleur, enz.. Je herhaalt dit 10 keer zodat je op het einde 10 kleuren hebt opgeschreven. Hoe kan je nu achterhalen of Danny de waarheid heeft gesproken? Je begint met Danny te geloven en je kijkt naar wat er zou gebeuren als er echt 500 witte en 500 zwarte kaartjes in die vaas zouden zitten. Dit is je hypothese, het startpunt van je redenering.

Als de hypothese waar is dan heb je bij elke trekking 50% kans op wit en 50% kans op zwart. Als je 10 kaartjes trekt dan verwacht je dat er ongeveer 5 zwarte tussenzitten. Maar het kunnen er ook 6, 7 of zelfs 8 zijn. Dat is gewoon te wijten aan de variabiliteit bij elk experiment waarbij het toeval een rol speelt. Maar onderstel nu eens dat die 10 kaartjes allemaal zwart zijn. Dat is toch wel een extreme uitkomst die je eigenlijk niet had verwacht als de vaas evenveel witte als zwarte kaartjes heeft. Is dit resultaat extreem genoeg om verdacht te zijn? Of omgekeerd, als Danny nu voorstelt om die vaas te gebruiken om tegen hem een spel te spelen waarbij jij verliest als je zwart trekt, zou je dat dan doen? Hoe extreem moet een uitkomst zijn vooraleer je Danny niet meer vertrouwt?

Om deze vraag te beantwoorden berekenen statistici een p-waarde. Je vertrekt van de hypothese dat de vaas met kans 50% wit oplevert en met kans 50% zwart. Bij 10 trekkingen zou dat dan 5 zwarte moeten opleveren, maar je weet dat er variabiliteit op zit en dat je dikwijls niet exact 5 zwarte kaartjes zal hebben . Hoe groot mag de afwijking zijn zodat jij ze nog aanvaardbaar vindt? Je kan je hierbij laten leiden door kansen uit te rekenen.

Wat is bijvoorbeeld de kans om minstens één zwarte kaart meer te hebben dan de verwachte 5? Wat is dus de kans dat je minstens 6 zwarte kaarten hebt? De kansrekening leert dat die kans 38 % is. Dat betekent het volgende. Als je 100 keer 10 kaartjes trekt en telkens noteert hoeveel zwarte er tussen zitten, dan zal je zien dat er ongeveer 38 keer ten minste 6 zwarte kaartjes tussenzaten. Dat is dus iets wat gemiddeld 38 keer op honderd gebeurt. Dat is helemaal niet extreem of verdacht. Je hebt hier een p-waarde van 0.38.

Je kan ook uitrekenen wat de kans is dat er minstens 7 kaartjes van de 10 zwart zijn. Dat is 17 % zodat de p-waarde hier gelijk is aan 0.17. Als je dus 100 keer 10 kaartjes trekt dan zal het ongeveer 17 keer gebeuren dat er tussen die 10 kaartjes tenminste 7 zwarte zitten. Dat komt dus al wat minder voor maar het is zeker nog niet uitzonderlijk te noemen.

De kans dat je tenminste 8 zwarte kaartjes hebt bij 10 trekkingen is 5.5%. De p-waarde is hier 0.055 wat behoorlijk klein is. Zij wijst op iets wat zelden gebeurt, slechts een zestal keer op 100.

Om minstens 9 zwarte kaartjes te hebben bij 10 trekkingen is de kans 1 %. Een p-waarde van 0.01 is zeer klein. Als de vaas er echt zou uitzien zoals Danny beweerde dan gebeurt het maar 1 keer op 100 dat je minstens 9 zwarte kaartjes hebt. En dat zou dan toevallig nu bij jou gebeurd zijn?

Bij 10 zwarte kaartjes is het helemaal erg. De kans om zoiets uit te komen is 1 op 1000. Een p-waarde van 0.001 vertelt je dat je beter Danny niet meer gelooft en zeker niet met die vaas dat spel tegen hem speelt. Als de vaas er echt zou uitzien zoals hij gezegd heeft, dan zou je bij 1000 keer 10 kaartjes trekken maar één keer 10 zwarte vinden. De ene keer op 1000 waarbij zoiets gebeurt zou dan nu juist bij jou gebeurd zijn zeker? Geloof je dat? Of is het verstandiger om te besluiten dat de vaas er helemaal niet uitziet zoals Danny zei? In de statistiek verwerpt men de beginhypothese (hier is dat de bewering van Danny) van zodra de p-waarde klein is (meestal neemt men 0.05 als grens, of soms ook 0.01).

Zo worden p-waarden gebruikt door statistici. Als je een resultaat ziet dat een kleine p-waarde heeft, dan heb je ofwel te maken met een hoogst uitzonderlijke gebeurtenis (en dan kan je misschien je naam laten opnemen in het Guinness Book of Records), ofwel was de hypothese waarmee je gestart was verkeerd (en dat is de meest voor de hand liggende verklaring natuurlijk).

	Sterfgevallen tijdens de shift		
Gilbert aanwezig	Ja	Neen	Totaal
Ja	40	217	257
Neen	34	1350	1384
Totaal	74	1567	1641

Tabel 1: De data waarop de statistische toets is gebaseerd

Opmerking: De tabel is gebaseerd op de volgende gegevens:

- Aantal dagen 547
- Aantal shiften 1641
- Aantal overlijdens 74
- Overlijdens per shift 0.045
- Shifts waar KG aanwezig was 257
- Verwachte aantal overlijdens 11.59
- Echte aantal overlijdens 40

In wat volgt wordt slechts één onderdeel van een reeks toetsen die Gehlbach heeft uitgevoerd, besproken.

4.3. Een statistische toets.

Op dit moment laat Gehlbach aan de rechters de gegevens zien uit bovenstaande tabel. Hij geeft daarbij de volgende commentaar.

In februari 1996 hebben de collega's van mevrouw Gilbert aan hun diensthoofd hun bedenkingen geuit. Snel daarna heeft Gilbert ontslag genomen en is ze opgestapt. De tabel vat alle gegevens samen die zijn verzameld tijdens de 18 maanden die vooraf gingen aan die fameuze februari van 1996. Die periode telt 547 dagen. Met 3 shiften per dag maakt dat 1641 shiften. Daarvan zijn er 74 waarbij er minstens één sterfgeval was. De kans dat er tijdens een shift minstens één sterfgeval voorkomt is dus $\frac{74}{1641}$ of 0.045, wat 45 per 1000 is.

Beeld je nu in dat elke shift gesymboliseerd wordt door het trekken van een kaartje uit een vaas en dat zwart gelijk staat met minstens één sterfgeval in die shift. In dit geval moet je dus trekken uit een vaas waarin 1000 kaartjes zitten waarvan er 45 zwart zijn en 955 wit.

Onderstel nu eens dat een willekeurige verpleegster in die voorbije periode 200 shiften zou gewerkt hebben en dat het aantal shiften met minstens één sterfgeval helemaal niets met haar aanwezigheid zou te maken hebben. We kunnen dan gewoon de wetten van de kansrekening toepassen. Het is dan alsof die verpleegster 200 keer een kaartje uit die vaas zou getrokken hebben. Elke keer had ze een kans van 45 op duizend om een zwart kaartje te trekken en een kans van 955 op duizend om een wit kaartje te trekken. Op die 200 keer verwacht je dan dat zij ongeveer 9 keer een zwart kaartje zou getrokken hebben want $200 \times \frac{45}{1000} = 9$. Ook hier heb je terug enige variabiliteit en misschien kom je bij 200 trekkingen slechts 7 of 8 zwarte kaartjes tegen, of misschien wel 10, 11 of 12. Zoiets verwacht je en dat is helemaal niet uitzonderlijk. Maar als je echt te veel zwarte kaartjes hebt, dan is dat wel verdacht. Men zal vermoeden dat je aanwezigheid op een of andere manier samenhangt met de sterfgevallen tijdens uw dienst. Hoeveel sterfgevallen je moet hebben om niet meer te kunnen volhouden dat dit allemaal puur toeval is kom je te weten door de p-waarde te berekenen. Dit doen we nu voor het geval van Gilbert.

In totaal waren er 257 shiften waarbij Gilbert van dienst was. Als haar aanwezigheid niets met die sterfgevallen te maken had, dan verwacht je dat er ongeveer 12 van haar shiften waren met minstens één overlijden. Dat is immers zoals 257 keer een kaartje trekken uit die vaas met 45

zwarte en 955 witte kaartjes. Je verwacht dan dat je 11 of 12 keer de kleur zwart hebt genoteerd bij die 257 trekkingen want $257 \times \frac{45}{1000} = 11.6$.

Wat laten de gegevens zien? Als je terug naar de tabel kijkt dan zijn er bij Gilbert niet 11 of 12 shiften met minstens één sterfgeval, maar 40. Hoe extreem is 40? Zou je uit die vaas 40 keer op 257 een zwart kaartje kunnen trekken door louter toeval of is 40 echt verdacht? Om dit te beantwoorden berekenen je de p-waarde. En die p-waarde is hier minder dan 1 op 100 miljoen. Dat betekent dat het vrijwel onmogelijk is om tijdens de shiften van Gilbert zoveel sterfgevallen te hebben door louter toeval. Er was dus duidelijk een samenhang tussen de aanwezigheid van Gilbert en het hoge aantal sterfgevallen tijdens haar shiften. De uiterst kleine p-waarde was daarvoor het wetenschappelijke bewijs.

5. Gilbert op het proces

De rechters van de kamer van inbeschuldigingstelling vonden het betoog van Gehlbach overtuigend genoeg om Gilbert in staat van beschuldiging te stellen. Gilbert werd 4 keer voor moord en 3 keer voor poging tot moord aangeklaagd en zij werd naar het hof van assisen doorverwezen. Het openbaar ministerie besliste om de doodstraf te eisen. Voor Kristen Gilbert werd dit het proces op leven of dood.

Rechter Michael A. Ponsor was de voorzitter van het assisenhof en nog vooraleer het proces begon moest hij beslissen of de assisenjury het statistisch bewijs, dat de kamer van inbeschuldigingstelling had overtuigd, ook mocht te zien krijgen.

Je zou kunnen denken dat het toch geen probleem was om aan de jury de statistische evidentie die Gehlbach had gevonden te tonen. Maar zo eenvoudig werkt dat niet. Als de aanklager met de statisticus Gehlbach op de proppen komt dan zal de verdediging zeker ook op zoek gaan naar een goede statisticus die de argumenten van Gehlbach zal proberen te ontkrachten. De kans bestaat dan dat de jury bestookt wordt met een massa ingewikkelde statistische argumenten en dat zij uren moet luisteren naar twee experten die elkaar tegenspreken. Zo'n situaties zijn niet zeldzaam bij een rechtspraak en het gebeurt dan wel eens dat de jury er de brui aan geeft en van die hele statistische argumentatie niets meer wil weten. Zoiets helpt het proces helemaal niet vooruit, integendeel.

Om dit te voorkomen en om geen 'duelerende experten' op de jury los te laten, had de verdediging van Gilbert een andere expert, George Cobb, ingeschakeld en aan hem gevraagd om vooraf een geschreven rapport te maken voor rechter Ponsor. In dat rapport argumenteerde Cobb dat het niet goed was om aan de jury de statistische bewijzen van Gehlbach te tonen. Hij bracht daarvoor verschillende argumenten aan.

In de volgende paragrafen staat een samenvatting van de belangrijkste punten van het rapport van Cobb. Zet jezelf deze keer in de plaats van rechter Ponsor. Vind je dat Cobb goede argumenten heeft? Zou je nu nog toestaan dat de assisenjury de statistische bewijzen te zien krijgt?

6. Context van de opmetingen

Tijdens Gehlbach's betoog werd duidelijk wat de p-waarde bij 'toetsen van hypothesen' betekent. Een heel kleine p-waarde zegt dat het resultaat te extreem afwijkt van wat je verwacht om nog toegeschreven te kunnen worden aan louter toevallige variabiliteit. Voor de kamer van inbeschuldigingstelling was de cruciale vraag: zijn er te veel sterfgevallen tijdens de dienst van Gilbert, zodat het in de ogen van de wetenschap als verdacht kan worden beschouwd? Met een p-waarde die kleiner was dan 1 op 100 miljoen was het antwoord duidelijk: "ja".

In het rapport van Cobb werd de nadruk gelegd op wat de p-waarde niet betekent.

Cobb legde uit dat de interpretatie van een p-waarde ook te maken heeft met de manier waarop de gegevens zijn opgemeten en dus met de hele context van de situatie. Als je daar geen rekening mee houdt dan trek je snel een verkeerde conclusie. En dat gebeurt veel meer dan je denkt, want intuïtief wijst alles hier op de schuld van Gilbert. Maar met statistiek zal je dat niet kunnen aantonen beweert Cobb. Daarom pleit hij ervoor om de statistische argumenten niet aan de jury te tonen, zodat ze niet op het verkeerde been worden gezet.

7. Cobb's rapport voor rechter Ponsor

De technische details laten we buiten beschouwing en we beperken ons tot de drie belangrijke punten in het rapport van Cobb. Eén daarvan is dat hij akkoord gaat met de basisconclusie van Gehlbach. De andere twee leggen uit wat je uit een kleine p-waarde niet mag besluiten.

7.1. De statistici van de aanklager en van de verdediging gaan akkoord

Zoals reeds gezegd gebeurt het regelmatig dat twee experten niet akkoord gaan wanneer zij een getuigenis afleggen en daarbij wetenschappelijke argumenten gebruiken. Toch was dit niet het geval in de zaak Gilbert. Cobb ging akkoord met Gehlbach's argumenten. Zowel Cobb als Gehlbach zijn ervan overtuigd dat het verband tussen Gilberts aanwezigheid op de afdeling en het ongewoon aantal doden te sterk is om het aan louter toeval te kunnen toeschrijven.

Ze denken ook dat, omdat er geen andere logische verklaring is gevonden, er genoeg aanwijzingen zijn om Gilbert in staat van beschuldiging te stellen. Waarom zou de assisenjury Gelbach's getuigenis dan toch best niet horen? Om op deze vraag te antwoorden moet je de twee andere punten van Cobb verder bestuderen.

7.2. Associatie is niet noodzakelijk oorzaak

Bij de twee stappen in de juridische procedure van dit verhaal moeten beslissingen genomen worden die steunen op een verschillende bewijslast.

De kamer van inbeschuldigingstelling beslist alleen maar of Gilbert al dan niet voor het assisenhof moest komen. Wogen de verdenkingen zwaar genoeg om de vele kosten voor de het gerecht en de psychologische druk op Gilbert te rechtvaardigen? Hier moet dus niet beslist worden of Gilbert al dan niet schuldig is.

De drempel van bewijslast is veel lager voor een kamer van inbeschuldigingstelling dan voor een assisenhof. De kamer van inbeschuldigingstelling moet alleen maar nagaan of de graad van verdenking groot genoeg is. Om een antwoord te zoeken op dit soort vragen is de logica van 'toetsen van hypothesen' ontwikkeld.

In de statistiek wil men redelijk zeker zijn vooraleer men gaat spreken over abnormaal gedrag. Meestal neemt men een p-waarde die kleiner is dan 0.05 (of 0.01). Een gebeurtenis met een lage p-waarde lijkt dan onmiddellijk verdacht omdat "een zo grote afwijking die louter toevallig zou gebeurd zijn" bijna niet meer als verklaring kan aangenomen worden. Maar een lage p-waarde geeft geen uitleg over wat er gebeurd is. Ze zegt niet: "dit is de reden voor al die sterfgevallen." Ze zegt alleen maar: "Wat de oorzaak ook mag zijn, je kan er redelijk zeker van zijn dat het hier niet gaat om louter toevallige variabiliteit, daarvoor is de afwijking echt te groot".

Op het ogenblik van het proces ligt de zaak helemaal anders. De jury heeft er niet veel aan om te weten dat die 40 sterfgevallen een uitzonderlijk groot aantal was dat niet zomaar door toevallige variabiliteit kan verklaard worden. Wat de jury wil weten is of die sterfgevallen veroorzaakt werden door Gilbert omdat ze dodelijke injecties toediende. Op deze vraag kan een lage p-waarde niet antwoorden, omdat je hier te maken hebt met een observatiestudie en niet met een toevalsgecontroleerd experiment. Dit argument haalde Cobb aan om te zeggen dat het statistisch bewijs niet geschikt was voor deze jury.

Nota
Als je niet vertrouwd bent met de begrippen "observatiestudie" en "toevalsgecontroleerd experiment" dan lees je best de afzonderlijke tekst "Statistische studies naar het soort van samenhang". De volgende zinnen bouwen op die begrippen voort. Je kan die eventueel nu overslaan en rechtstreeks naar punt 3 gaan.

De gegevens van Gilbert's proces waren verkregen door een observatiestudie. Om er een toevalsgecontroleerd experiment van te maken had de aanwezigheid van Gilbert moeten bepaald worden met een toevalsgenerator (om haar lukraak toe te wijzen aan één of andere shift waarop ze zou moeten werken).

Omdat het geen statistisch experiment was kon de kleine p-waarde andere mogelijke oorzaken niet uitsluiten. Misschien is er wel iets speciaals met de avondshift en sterven patiënten bij voorkeur op dat ogenblik van de dag. Als dat zo zou zijn dan zou dat al heel wat verklaren want Gilbert had voor het grootste deel van haar carrière tijdens de avondshift gewerkt.

Rechter Ponsor gaf een ander voorbeeld, volledig hypothetisch, maar het zou dezelfde statistische cijfers kunnen opleveren. "Stel dat er een warmwaterketel was ontploft tijdens de dienst van Gilbert en dat er verschillende slachtoffers zouden geweest zijn. Dan zouden er

tijdens de shiften van Gilbert ook verdacht veel overlijdens zijn geweest. Zo'n ongeluk zou een kleine p-waarde opleveren en de aanwezigheid van Gilbert onmiddellijk associëren met het hoge dodencijfer. Maar dat zou geen bewijs zijn dat Gilbert daar enige schuld aan had".

In de statistiek zegt men: "een verband tussen twee dingen is nog geen bewijs dat het ene de oorzaak is van het andere". Het is verleidelijk om op basis van een toets te besluiten dat je de oorzaak gevonden hebt van zodra je een duidelijk verband ziet. Met gegevens van een observatiestudie kan dit gemakkelijk leiden tot een verkeerde conclusie.

7.3. De valkuil van de aanklager

Het rapport van Cobb gaf nog een tweede, verwante 'valkuil' waarin men dikwijls trapt bij het toetsen van hypothesen. De p-waarde is een kans die men berekent in de onderstelling dat de starthypothese waar is. In dit voorbeeld betekent dit dat je ervan uitgaat dat Gilbert niet schuldig is en dat haar aanwezigheid op die shiften niets te maken heeft met het aantal doden. Dat aantal shiften waarbij minstens één dode valt is dan louter aan het toeval te wijten.

De redenering gaat dan als volgt verder. Als het aantal shiften met minstens één dode louter aan het toeval te wijten is, dan verwacht ik bij Gilbert 11 of 12 shiften met minstens één dode, en geen 40. Als ik nu toch het extreme resultaat van 40 shiften zie, dan is het niet redelijk om te denken dat dit nog louter aan het toeval te wijten is. Merk op dat deze logica niets zegt over wat er dan wel de oorzaak zou kunnen zijn.

Kijk nu hoe gevaarlijk deze redenering wordt als je niet oppast.
"Stel dat Gilbert niet schuldig is, en dat de sterftecijfers toe te schrijven zijn aan louter toeval, zoals bij het trekken van een kaartje uit een vaas. De kans is minder dan 1 op 100 miljoen dat er zoveel shiften met sterfgevallen zijn tijdens haar werktijd." (correct)
Het is een snelle sprong naar de kortere versie: "Als Gilbert onschuldig is, dan is het bijna onmogelijk om zoveel meer shiften met sterfgevallen te hebben dan wat je normaal zou verwachten." (ook correct)
En dan: "Met dit grote aantal shiften met doden is de kans minder dan 1 op 100 miljoen dat Gilbert onschuldig is." (niet correct!)

Deze vorm van fout redeneren is zo verleidelijk en komt zoveel voor, dat ze bij statistici gekend staat als "de valkuil van de aanklager" (*the prosecutor's fallacy*).

Omdat deze foutieve redenering zo verleidelijk is, was het te verwachten dat het statistisch bewijs fout zou geïnterpreteerd worden door de jury. De jury zou onmiddellijk kunnen denken dat hier statistisch bewezen werd dat Gilbert schuldig was. Dit zou de aanklager bevoordelen en een eerlijk proces in de weg kunnen staan.

8. Conclusie

Rechter Ponsor besliste dat de resultaten van het statistisch onderzoek niet toegelaten werden tijdens het assisenproces van Gilbert. Toch was er genoeg ander, niet-statistisch bewijs gevonden om de jury te overtuigen. Na veel dagen van beraadslaging werd Gilbert veroordeeld voor 3 moorden met voorbedachten rade, 1 moord zonder voorbedachten rade, en

2 moordpogingen. De jury stemde 8 tegen 4 voor de doodstraf. Omdat er niet unaniem is gestemd werd de doodstraf niet voltrokken. Gilbert zit nu een levenslange gevangenisstraf uit, zonder enige kans op vervroegde vrijlating.

De statistische analyse die Gilbert's aanwezigheid in verband bracht met de overmaat aan sterfgevallen was een belangrijk argument om haar in verdenking te stellen. De twee stappen in de rechtsprocedure tonen mooi aan hoe belangrijk het is om 'toetsen van hypothesen' op de juiste manier te interpreteren. Eerst en vooral zorgt een kleine p-waarde ervoor dat je de louter toevallige variabiliteit kan uitsluiten als een aanvaardbare verklaring voor een extreem fenomeen dat je hebt waargenomen. Ze zegt ons dat de waarneming zo extreem is dat we ze kunnen beschouwen als een verrassing voor de wetenschap. Maar verder vertelt een kleine p-waarde je niet wat de oorzaak is van die verrassing wanneer je te maken hebt met een observatiestudie.

9. Zelfevaluatie

1. Als je weet tijdens welke shiften Kirsten Gilbert werkte dan kan je uit figuur 1 reeds vermoeden dat er iets verdacht aan de hand is. Wat? Verklaar je antwoord.

2. Waarom zijn de bezwarende feiten die je uit figuur 1 kan aflezen geen bewijs dat Gilbert schuldig is? Verklaar de redenering je hier moet gebruiken.

3. Wat is "de valkuil van de aanklager"? Leg duidelijk uit wat het probleem hier is

4. Wat is een p-waarde? Illustreer met een eenvoudig zelfgekozen voorbeeld

5. Waarom kan je in deze observatiestudie geen schuld bewijzen als je alleen maar naar het cijfermateriaal mag kijken? Wat zou een verstrengelende factor kunnen zijn? En welk probleem zou dat kunnen geven? Verklaar je antwoorden.