

HET STATISTISCH INSTRUMENTARIUM

REDE

Uitgesproken bij de aanvaarding van het ambt van
gewoon hoogleraar in de Medische Statistiek
aan de Rijksuniversiteit te Leiden
op vrijdag 15 mei 1987

door

DR. J. C. van HOUWELINGEN



1987

RIJKSUNIVERSITEIT TE LEIDEN

Mijnheer de rector magnificus,
zeer gewaardeerde toehoorders,

Op de financieel-economische pagina van de Volkskrant zijn tegenwoordig grafieken aan te treffen, die het verloop van bepaalde economische indices in de tijd weergeven. Een daarvan is de grafiek van de Dow Jones-index over de laatste drie weken. Een nauwkeurige beschouwer moet het opvallen dat de indeling langs de verticale as van dag tot dag kan veranderen, afhankelijk van de mate waarin de index de afgelopen tijd aan verandering onderhevig is geweest. De schaling van de verticale as wordt altijd zo gekozen dat de beschikbare ruimte geheel gebruikt wordt. De consequentie daarvan is dat een continue stijging van 10 punten per dag hetzelfde plaatje oplevert als een stijging van 100 punten per dag. Zonder goed naar de schaal van de verticale as te kijken, kan men uit de grafiek niet opmaken of het de afgelopen weken rustig of onrustig is geweest op Wall Street.

De indruk die een grafiek op de lezer maakt is sterk beïnvloedbaar door de degene die de grafiek vervaardigt. Bewust of onbewust kunnen daar allerlei fouten bij gemaakt worden. Een uitvoerige beschrijving daarvan is te vinden in Edward R. Tufte's fraaie boek "The Visual Display of Quantitative Information"¹⁾. Dat de wereld statistici wantrouwt vanwege de wijze waarop gegevens worden gepresenteerd, blijkt wel uit het feit dat de boektitel "How to Lie with Statistics"²⁾ welhaast tot een gevleugelde uitdrukking geworden is.

Fraaie gevallen van statistische manipulatie zijn te vinden in het werk van Florence Nightingale. Voorbeelden daarvan worden gegeven in een aan haar gewijd artikel in de Scientific American van maart 1984. Om goed duidelijk te maken wat het effect was van eerst schromelijke verwaarlozing en daarna haar persoonlijk ingrijpen op de sterftcijfers in het ziekenhuis van Scutari tijdens de Krimoorlog in de winter van 1854/1855 vervaardigde zij een diagram dat werd opgenomen in het eindrapport van de op haar instigatie

ingestelde Royal Commission on the Health of the Army. De gegevens die zij weergaf waren de maandelijkse sterftcijfers als percentage van de patiëntenpopulatie op jaarbasis. Dat op zich geeft natuurlijk al indrukwekkende cijfers. In februari 1855 was het getal 415%. De gegevens werden grafisch weergegeven op een cirkel door stralen vanuit het middelpunt met een lengte evenredig aan het percentage. De uiteinden hiervan werden verbonden en het gebied binnen de grafiek donker gekleurd. Op deze wijze ontstond een zeer dramatisch effect, omdat de stralen op zich wel evenredig zijn met het percentage, maar het gekleurde oppervlak dat begrensd wordt door een deel van de grafiek, evenredig is met het kwadraat van het percentage. Zo kon het niemand ontgaan dat na de piek van 415% in februari de sterftcijfers door het ingrijpen van Florence Nightingale zeer snel waren teruggelopen.

Men kan stellen dat Florence Nightingale de statistiek als vergrootglas hanteerde om de door haar relevant geachte feiten onder de aandacht van anderen te brengen. Dat zij daarbij, bewust of onbewust, een extra vergroting aanbracht door de getallen visueel te kwadrateren, is maar een bijkomstigheid. Als zij de percentages had weergegeven als functie van de tijd in een nu meer gebruikelijke grafiek, was het effect niet veel minder dramatisch geweest. Overigens, had William Playfair³⁾ dergelijke grafieken al enige decennia eerder geïntroduceerd, maar zijn aanpak was in onbekendheid geraakt.

Dat de statistische wereld haar dankbaar was voor haar bijdrage aan de ontwikkeling van het vakgebied, moge blijken uit haar verkiezing tot Fellow of the Royal Statistical Society in 1858 en de toekenning van het erelidmaatschap van de American Statistical Association in 1874. De dichtregels van Longfellow

"Lo! in that house of misery

A lady with a lamp I see."

zouden in overdrachtelijke zin zo geïnterpreteerd kunnen worden dat zij als lichtend voorbeeld kan dienen

voor zowel medisch statistici als voor werkenden in de intramurale gezondheidszorg.

Uit het voorbeeld van de diagrammen van Florence Nightingale kunnen wij leren dat de kunst van het overdrijven een onderdeel is van de statistiek. Hoewel er voor gewaakt moet worden om niet de verdenking op zich te laden een statistische leugenaar of een liegende statisticus te zijn, ben ik toch van mening dat overdrijving inherent is aan statistiek. Men dient zich daarvan natuurlijk wel bewust te zijn.

Bekijken wij als voorbeeld de volgende situatie. Ter vergelijking van twee behandelingstherapieën worden patiënten gerandomiseerd over twee groepen. In beide groepen wordt voor elke patiënt nagegaan of de behandeling gewerkt heeft. Om te toetsen of het verschil statistisch significant is, wordt de bekende chikwadraat-toets voor de 2×2 -tabel gebruikt. De toets levert een chikwadraat-waarde en een overschrijdingskans of P-waarde op. Veronderstel dat in beide groepen 100 patiënten zitten en dat de effectiviteit van de ene therapie 10% en van de andere 20%. Men vindt dan een chikwadraat van 3.92 en een P-waarde iets kleiner dan 5%.

Als de groepen twee maal zo groot zijn, maar de effectiviteiten ongewijzigd, wordt de chikwadraat twee maal zo groot, dus 7.84 en de P-waarde kleiner dan 1%. Een argeloze toepasser zal concluderen dat in het tweede geval de P-waarde veel kleiner is, het verschil dus veel significanter, en dus ook groter. In deze valkuil vallen helaas nog velen. Vermoedelijk niet in dit eenvoudige versimpelde geval, maar wel in gevallen waarbij ingewikkelder technieken gebruikt worden.

Effect van een behandeling en corresponderende P-waarde zijn weliswaar gekoppeld, maar de P-waarde is vertekend door het vergrootglas van de statistische toets. Immers, een statistische toets beoordeelt het gevonden verschil niet in absolute termen, maar in termen van de toevalsvariatie waarmee het gevonden verschil behept kan zijn. Bij een constant verschil en een toenemende steekproefgrootte neemt de mogelijke toevalsvariatie af

en daarom wordt het virtuele verschil onder het statistisch vergrootglas groter.

Hoewel oorspronkelijk bedoeld als verkleinglas om te voorkomen dat verschillen die de toevalsvariatie niet te boven gaan, als werkelijke verschillen worden gepresenteerd, loopt de statistische toets door het ongewenste blindstaren op de P-waarde nu constant het risico om nauwelijks relevante verschillen op te blazen en op die wijze tot de verkeerde conclusie te leiden.

Een andere vertekening is die, welke veroorzaakt wordt door de proefopzet. Als wij in hetzelfde voorbeeld als hierboven, de patiënten zo over de groepen verdelen dat in de ene groep 50 en in de andere 150 patiënten terecht komen en nog steeds uitgaan van effectiviteiten van 10, respectievelijk 20 %, dan wordt de chikwadraat 2.60 en de P-waarde 10%. Omdat de totale steekproefomvang niet verschilt, die is immers nog steeds 200, zou men ten onrechte kunnen concluderen dat het verschil kleiner is dan bij de eerstgenoemde situatie met 100 patiënten in elke groep. Natuurlijk zal men bij het daadwerkelijk opzetten van het experiment er naar streven om groepen van gelijke grootte te krijgen. In allerlei observationeel onderzoek valt er echter niets te kiezen en kan deze vertekening vaak ongemerkt optreden. In datzelfde kader is ook een waarschuwing op zijn plaats om voorzichtig te zijn met het gebruik van de correlatiecoëfficiënt omdat die inherent met deze vertekening behept is. De correlatie tussen de variabelen y en x is alleen te vergelijken met de correlatie tussen de variabelen y en z als x en z dezelfde verdeling hebben.

In het algemeen kan men stellen dat deze vertekeningen voorkomen bij het statistisch toetsen of er een effect bestaat of niet. Een toets dient naar mijn mening gelijk op te gaan met een schatting van het effect. Daartoe is dan soms bezinning nodig op de definitie van het effect. Om een voorbeeld te noemen, in allerlei onderzoek worden overlevingskrommes met elkaar vergeleken. Toetsen op geen verschil worden uitgevoerd zoals de logrank-toets of de generaliseerde Wilcoxon-toets,

en de gevonden P-waardes worden aangedragen als indicatie voor een mogelijk verschil. Daarbij wordt dan verzuimd aan te geven in hoeverre de overlevingen zelf daadwerkelijk verschillen. Als maat daarvoor kan dienen het verschil in mediane overleving of in, bij voorbeeld, 5-jaars overleving. Dat op zich is veel belangrijker dan de P-waarde die alleen dient als statistisch hulpmiddel.

Aangesproken door de klacht over statistisch significante, maar praktisch irrelevante verschillen zullen ongetwijfeld de diagnose-stellers zijn. Iedereen weet dat mannen gemiddeld langer zijn dan vrouwen, daar is nauwelijks een statistische loupe voor nodig. Om echter eenduidig vast te stellen of iemand een man of een vrouw is, is de lengte alleen als gegeven niet voldoende. Daarvoor zit er te veel overlap in de lengteverdeling van mannen en vrouwen. Nu is het vaststellen van iemands sexe behalve bij de beoefenaars van sommige sporten een opgelost probleem en zal ieder van u zonder veel moeite in staat zijn van alle hier aanwezigen puur door middel van visuele inspectie de sexe vast te stellen. Dat doet u dan door het combineren van verschillende kenmerken, die worden waargenomen, en met gebruik van het in het menselijke brein aanwezige expert-systeem dat door middel van enige routine-checks binnen een fractie van een seconde de juiste classificatie uitvoert.

In het algemeen zijn diagnose- of classificatie-problemen verre van eenvoudig op te lossen. Niet voor niets is het één van de twee erkende onderzoeksprojecten van de sectie medische statistiek. Praktisch zijn zij van het grootste belang. Zonder een goede diagnose kan de medicus niet aan de slag. Om een diagnostisch apparaat te creëren met behulp van statistische technieken verzamelt men eerst van twee groepen met bekende geverifieerde diagnose, noem ze voor het gemak zieken en gezonden, allerlei informatie die van belang kan zijn voor het stellen van de diagnose. Op basis van deze gegevens hoopt men dan een systeem te ontwikkelen om het gewenste onderscheid te kunnen maken. De hoop is

dat door het verzamelen van genoeg gegevens eenduidige classificatie mogelijk is.

Het is in dit kader nuttig om even stil te staan bij het verschil tussen het classificatieprobleem en het toetsingsprobleem. Bij het toetsen of twee populaties verschillen op een bepaald kenmerk verzamelt men informatie door meer leden van elke populatie te bekijken om tenslotte met behulp van het statistisch vergrootglas het verschil zo uitvergroot te hebben dat geconcludeerd kan worden dat het statistisch significant is. Bij het classificatieprobleem verzamelt men voor elk individu uit een vaste, niet te kleine, steekproef, zoveel informatie door allerlei kenmerken te meten dat op grond daarvan beide populaties zo goed als volledig te scheiden zijn. Het instrument dat hiertoe nodig is zou men kunnen betitelen als het statistisch ontleedmes. De meer gebruikelijke naam is discriminant-analyse.

De eerste bijdragen aan de ontwikkeling van deze analysetechniek werden geleverd door Sir Ronald A. Fisher die de lineaire discriminantfunctie introduceerde. Door lineaire combinatie van de grootheden die aan een individu worden gemeten, kan een nieuwe grootheid gedefinieerd worden die de groepen veel beter scheidt dan elke variabele op zich. Het geeft altijd wat problemen bij de gebruiker dat op deze wijze appels en peren bij elkaar worden opgeteld, bij voorbeeld lengte en bloeddruk, maar het succes van de methode, blijkend uit lage misclassificatiekansen, heeft deze bezwaren meestal overstemd.

Een andere methodiek is om van beide groepen de simultane of multivariate verdeling van de gemeten kenmerken te bepalen en dan een classificatieregels te vinden gebaseerd op de ratio van de kansdichtheden. Een natuurlijke manier om dat te doen is via de eeuwen-oude regels van Bayes⁴⁾. Uitgaande van een a priori kans wordt de a posteriori kans bepaald en deze wordt gebruikt als hulpmiddel bij de classificatie, eventueel in een besliskundige context waarbij de effecten van verkeerde beslissingen gekwantificeerd worden in een

verliesmatrix. Onder de veronderstelling van multivariaat normale verdelingen leidt dit tot dezelfde toewijzingsregel als Fisher's lineaire discriminantfunctie. Een fraai voorbeeld van deze Bayesiaanse aanpak met a posteriori kansen is te vinden in het werk van Veltkamp, Drion en Loeliger⁵⁾ op het gebied van Haemofilie-draagsterschap uit 1968. Zij publiceerden fraaie curves die de kans op draagsterschap relateren aan de bepaling van het gehalte factor VIII of factor IX.

Een andere aanpak is om het type curve dat op deze wijze ontstaat, de zogenaamde logistische kromme, als uitgangspunt te nemen voor de analyse. Deze methode, logistische regressie genaamd, wordt vaak aan Cornfield⁶⁾ toegeschreven, maar heeft zijn wortels in een eerder verleden. De logit-transformatie is al in het werk van Finney⁷⁾ in 1952 te vinden en ook het werk van Cox⁸⁾ op dit terrein mag niet onvermeld blijven. In tegenstelling tot de Bayesiaanse aanpak die via a priori kansen komt tot a posteriori kansen, gaat de logistische regressie direct uit van de a posteriori kans, dat wil zeggen dat het de kans modelleert dat iemand tot een bepaalde populatie behoort, bijvoorbeeld de populatie der zieken, gegeven alle informatie die over het individu beschikbaar is. Het voordeel is dat op deze wijze ook experimenteel instelbare variabelen in het model kunnen worden opgenomen. In de Bayesiaanse aanpak is dat onmogelijk, omdat dergelijke variabelen niet als toevallige, of stochastische, variabelen te interpreteren zijn. Genoemd voordeel maakt het ook mogelijk om het model te gebruiken bij de analyse van het effect van bepaalde therapieën op bij voorbeeld de vijfjaars-overleving. De laatste jaren heeft deze techniek een grote vlucht genomen, mede dankzij de opname in statistische programmapakketten. In de epidemiologie is het heel populair.

Een zeker gevoel van onbehagen is bij de gebruikers echter blijven bestaan om diverse redenen. Ten eerste worden de onbekende parameters in dergelijke modellen geschat via numerieke algoritmen die nogal

ondoorgegrondelijk zijn. Omdat er geen eenvoudig rekenvoorschrift is, is het de statisticus zelf vaak onduidelijk hoe een bepaalde computeroutput tot stand komt, laat staan dat hij het de onderzoeker voor wie hij deze techniek aanwendt, kan uitleggen wat er precies gebeurt. Ten tweede is het model op zich moeilijker te doorgronden dan de vertrouwde lineaire regressie. Het effect van het weglaten van variabelen is moeilijk te voorzien en ook de "goodness of fit", dat is de mate waarin het model klopt, is niet of slecht gedefinieerd. Ten derde wordt het logistische model zelf als modelering van de wijze waarop variabelen kansen beïnvloeden door velen aangevochten. De keuze voor het logistische model impliceert het gebruik van de "odds ratio" als effect-maat, een keuze die niet door iedereen als natuurlijk wordt ervaren. Hier en daar gaan geluiden op om de logistische regressie slechts spaarzaam te gebruiken en in eerste instantie te volstaan met eenvoudiger technieken gebaseerd op kruistabellen.

De bezorgheid over het gebruik van niet-volledig begrepen technieken deel ik, maar ik denk dat het de taak van de beoefenaren van de statistiek is om door middel van nader onderzoek te komen tot een beter verstaan van deze technieken ten einde de bezwaren tegen het gebruik ervan weg te nemen. Een onderdeel van dergelijk onderzoek kan zijn na te gaan of er geen algemenere modellen voor dichotome variabelen formuleerbaar zijn, vergelijkbaar met en mogelijk nauw verwant aan de wijze waarop Habbema en Hermans⁹⁾ in de jaren '70 de discriminant-analyse bevrijdden van de knellende banden van de veronderstelde normale verdeling.

Zolang het echter gaat om diagnostiek sec op basis van de beschikbare informatie, en niet om de interpretatie van de parameters in het model, zijn alle technieken redelijk equivalent en een uitstekend hulpmiddel bij diagnostiek. Als ontleedmes zijn zij alle bruikbaar. De één snijdt langs een rechte lijn, de andere langs een iets andere rechte en de derde laat een zigzag-route toe, maar alle scheiden zij bekwaam gezonden van

zieken, als er tenminste voldoende informatie aanwezig om een goede scheiding mogelijk te maken.

Zoals al uit de hiervoor gebezigde terminologie al is op te maken, is de classificatie-problematiek te zien als een speciaal geval van regressieanalyse. Men zou de regressieanalyse kunnen betitelen als het manusje van alles van de statistiek, of in meer medische terminologie als het statistisch panacee, of om in de instrumentarium-beeldspraak te blijven als de statistische combinatietang.

Regressieanalyse probeert een antwoord te geven op de vraag of en in hoeverre een stel verklarende of onafhankelijke variabelen, meestal aangeduid met x , invloed heeft op één of meer afhankelijke of te verklaren variabelen, meestal aangeduid met y . Een voorbeeld is de invloed van zwangerschapsduur als onafhankelijke variabele op geboortegewicht als afhankelijke variabele. De relatie kan gevisualiseerd worden door eerst de waarnemingen aan de in het onderzoek betrokken zwangerschappen weer te geven in een zogeheten scatterdiagram, met op de x -as zwangerschapsduur en op de y -as geboortegewicht, en daarin de regressielijn te tekenen die het verband versimpeld weergeeft. De interpretatie van de gevonden relatie is in het algemeen een moeilijke en hangt sterk af van de onderzoeksopzet. Duidelijk onderscheid is daarbij nodig tussen observationeel onderzoek, waarbij alleen een samenhang wordt geconstateerd, en experimenteel onderzoek, waarbij de x -variabele kan worden ingesteld en hopelijk tot een causaal verband geconcludeerd kan worden. Het standaard voorbeeld van zo'n ingestelde variabele is de therapiekeuze in een gerandomiseerd klinisch onderzoek. In het algemeen moet grote omzichtigheid betracht worden met de interpretatie. Dat bovendien door toevalsfluctuatie ten gevolge van een beperkte steekproefomvang de relatie niet exact kan worden vastgesteld is een complicatie die wij even buiten beschouwing laten.

In multiple regressie, waarbij meer dan één onafhankelijke x optreedt is de interpretatie nog veel

moeilijker. Laat ik dat proberen te demonstreren aan een versimpeld voorbeeld. Aan een groep personen is lengte in cm, gewicht in kg en sexe bepaald. Men vindt voor de mannen een gemiddelde lengte van 180 en een gemiddeld gewicht van 80 en voor de vrouwen een gemiddelde lengte van 170 en een gemiddeld gewicht van 70. Nogmaals, de getallen zijn gefingeerd en dienen alleen om het probleem duidelijk te maken. De gegevens worden nu onderworpen aan regressieanalyse met gewicht als te verklaren variabele en lengte en sexe als verklarende variabelen. Als regressievergelijking komt uit de bus

$$\text{gewicht} = \text{lengte} - 100 .$$

De sexe komt in die vergelijking niet voor, dat wil zeggen de regressiecoëfficiënt voor sexe is gelijk aan nul. Moeten wij nu concluderen dat geslacht geen invloed heeft op gewicht? Mij dunkt van niet, want iedereen weet dat sexe voor een groot deel de lengte bepaalt en zo op impliciete wijze wel degelijk van invloed is. Wat wij wel kunnen concluderen is dat in dit voorbeeld lengte volstaat om gewicht te voorspellen. Willen wij echter tot een goede interpretatie komen, dan ligt het hier voor de hand om lengte te corrigeren voor sexe en te kijken of deze gecorrigeerde lengte invloed heeft. Voor de mannen is de gecorrigeerde lengte gelijk aan lengte min 180 (de gemiddelde lengte voor mannen) en de regressievergelijking kan zo gelezen worden dat het gewicht gelijk is aan 80 (het gemiddelde voor de mannen) + de gecorrigeerde lengte. Voor de vrouwen geldt iets analoogs.

De standaard computeroutput leert ons niet dat het model ook zo geïnterpreteerd kan worden. Het regressiemodel sec geeft de goede regressiecoëfficiënt voor het effect van de voor sexe gecorrigeerde lengte, maar geeft een verkeerd beeld van de invloed van de sexe. Om de invloed van de variabele lengte te bepalen moet nagegaan worden in hoeverre de lengte binnen een sexe varieert. Is binnen een sexe de lengte vrijwel constant, dan kan de lengte gemist worden in een model voor het gewicht en is sexe alleen voldoende. Is de

variatie in lengte binnen een sexe groot dan is lengte een belangrijke variabele in het model voor gewicht. Als de verklarende variabelen in het model niet zo simpel zijn als in dit versimpelde voorbeeld, dan is het veel moeilijker om te bepalen welke verklarende variabelen welke andere verklarende variabelen beïnvloeden en is het noodzakelijk om een zekere hiërarchie binnen de verklarende variabelen aan te brengen. Als die structuur niet wordt aangebracht, loopt men het risico om, in de termen van het voorbeeld, te verklaren dat sexe geen invloed heeft op gewicht. Om inzicht te krijgen in de structuur binnen de groep verklarende variabelen is overleg nodig tussen de statisticus en de betrokken onderzoeker. Maar zelfs na zo'n overleg is prudentie geboden bij de interpretatie van de coëfficiënten in een regressiemodel.

Een heel andere vraag is of zo'n regressiemodel gebruikt kan gaan worden om uitspraken te doen over één enkel individu. Afgezien van de vraag of het überhaupt gewenst is om dergelijke uitspraken te doen -daar zitten allerlei ethische aspecten aan-, is het nuttig om na te gaan of en onder welke voorwaarden het doen van uitspraken mogelijk is. Een eerste aspect is natuurlijk de beperktheid van de informatie waarop het model gebaseerd is. Dat heeft tot consequentie dat de regressiecoëfficiënten niet exact bekend zijn, maar dat is een secundair probleem. Laten wij veronderstellen dat ons model gebaseerd is op een grote steekproef en dat dus de regressiecoëfficiënten nagenoeg foutloos zijn. De moeilijkheid is dan dat een regressiemodel alleen maar een uitspraak doet over de gemiddelde of te verwachten waarde van de te verklaren y-variabele, gegeven de waargenomen waarde van de verklarende x-variabelen. In ons voorbeeld kunnen wij bij een bepaalde lengte aangeven wat het gemiddeld gewicht is. Hoe het werkelijke gewicht van een individu om deze modelwaarde spreidt, wordt niet door de vergelijking gegeven. Wat nog wel als bijproduct van alle berekeningen gegeven wordt is een maat voor de spreiding om de modelwaarde in de vorm van een residuele

standaardafwijking. In termen van ons voorbeeld weten wij wat de gemiddelde fout is als wij gewicht uit lengte voorspellen met de formule "gewicht = lengte - 100". Om die informatie te kunnen benutten moeten wij veronderstellen dat de spreiding om het model niet afhangt van de verklarende variabele x . Dat is de eis van gelijke spreiding, in het vakjargon homoskedasticiteit genoemd. Het is niet eenvoudig om te controleren of aan deze eis voldaan is, en zekerheid of aan deze eis inderdaad is voldaan, is moeilijk te geven.

Dat brengt ons nog niet waar wij willen zijn. Wij kunnen nu in termen van het voorbeeld, bij een gegeven lengte het gemiddelde gewicht en de standaardafwijking in het gewicht aangeven. Om echter kansuitspaken te doen, zoals een uitspraak over de kans dat een individu van 1 meter 70 meer dan 90 kg weegt, moeten er ook veronderstellingen gemaakt worden over de kansverdeling. De neiging zal dan groot zijn om maar te veronderstellen dat men met de normale of Gauss-verdeling te doen heeft. Velen leven nog in het misverstand dat de term normaal als normatief geïnterpreteerd moet worden, in andere woorden dat afwijkingen ten opzichte van het model normaal verdeeld behoren te zijn. Dat is echter vaak niet het geval. Scheve verdelingen kunnen evengoed optreden. De juistheid van de veronderstelling over de verdeling van de afwijkingen ten opzichte van het model is op dat moment van wezenlijk belang en heeft directe consequenties voor de uitspraken die men wil doen over één individu. Zolang het gaat om uitspraken over de populatie, is de veronderstelling van normaliteit niet zo belangrijk, omdat meestal een vorm van middeling optreedt die benaderde normaliteit van de interessante grootheden, zoals de regressiecoëfficiënten, met zich meebrengt, maar zodra de uitspraken over één individu gaan, zijn die zeer gevoelig voor de gemaakte veronderstellingen.

Wij zien dat prognostiek, het hanteren van statistiek als glazen bol, veel meer omzichtigheid vereist dan het

opstellen van een regressievergelijking. Daarbij komt nog dat de nieuwere statistische modellen die de laatste jaren ontwikkeld zijn, veel ingewikkelder zijn dan gewone regressie.

Een voorbeeld daarvan is het Cox-regressiemodel voor overlevingskrommen. De onderzoeker die deze techniek op zijn getallenmateriaal wil toepassen, is nog wel in staat om te begrijpen dat uit de statistische analyse blijkt dat sommige variabelen een positieve invloed hebben op de overleving, andere een negatieve, en sommige helemaal geen invloed, maar hoe hij de resultaten van deze statistische analyse ook kan gebruiken om voor een patient een prognose op te stellen, zal hem veel minder tot helemaal niet duidelijk zijn.

Dat verklaart dat veel wordt teruggegrepen op een dichotome indeling in patiënten met een goede prognose en patiënten met een slechte prognose, waarbij voor beide groepen de Kaplan-Meier-overlevingskromme wordt getekend. Als eerste aanschouwelijke presentatie is dit zeker een acceptabele gang van zaken, maar in wezen is het toch een vorm van versimpeling die niet altijd recht doet aan de werkelijkheid. Net als bij de eerder genoemde logistische regressie, ligt hier voor de beoefenaren van de statistiek de taak om er verder op te studeren en het model inzichtelijk en toegankelijk te maken voor de toepasser. Het Cox-model is in; volgens velen, onder wie tijdschriftredacteurs, moet het gebruikt worden bij de analyse van overlevingsgegevens; dan moet het ook begrepen worden !

Dat begrijpen is van een andere orde dan het begrijpen waarom bij de statistische toetsen die voor dit model zijn ontwikkeld normale benaderingen gebruikt mogen worden. Over dat laatste heeft Richard Gill¹⁰⁾ onlangs gesproken bij zijn ambtsaanvaarding als bijzonder hoogleraar in de toepassingen van de wiskunde. Zonder het belang daarvan te kleineren, wil ik toch benadrukken dat een ander begrijpen daaraan vooraf dient te gaan, namelijk het begrijpen van wat het veronderstelde proportional hazards model precies inhoudt, of die veronderstellingen redelijk zijn en of

praktijk-situaties inderdaad door dit model te beschrijven zijn. Vaak blijkt dat nieuwe therapieën wel beter werken- in de termen van dit model: een kleinere hazard hebben- in de eerste maanden van de behandeling, maar dat na verloop van tijd dat verschil verdwijnt of in het ergste geval zelfs van teken verandert. Vertrouwen op het model om extrapolaties in de tijd te maken is dan ook zeer gevaarlijk. Als toegepaste statistici zich over die problemen niet druk maken, wordt er weer een kans gemist.

Kortom, de statistiek kan aan de vraag om een glazen bol voldoen, maar dat wordt dan geen eenvoudige bol, die alleen van tijd tot tijd afgestoft dient te worden. Er blijken allerlei regelknoppen aan te zitten, die met kennis van zaken moeten worden gehanteerd. Zonder die kennis zouden wij wel eens een beeld van de toekomst van één enkele patiënt kunnen oproepen dat absoluut niet realistisch is.

Het laatste instrument uit het statistisch instrumentarium, dat ik vandaag voor u wil presenteren, is de statistische hark van de meta-analyse. Gelijk bladeren in de herfst die bijeen geharkt worden tot één grote bladerhoop, harkt men onderzoeken, meestal aangeduid met de engelse term "trial", uit verschillende streken en tijden bij elkaar tot een grote hoop bestaande uit veel patiënten en creëert daarmee een statistisch significant resultaat. Op die manier is de meta-analyse in de medische literatuur geïntroduceerd door Peto¹¹⁾ in 1980. Zo simpel liggen de zaken echter meestal niet. Het bijeenvoegen van diverse trials gaat uit van de veronderstelling dat in elke trial dezelfde resultaten te verwachten zijn, onafhankelijk van waar en wanneer de trial gehouden wordt. Nu is dit minder absurd dan het lijkt omdat in de uitkomst van met name gerandomiseerde clinical trials al een correctie is aangebracht voor mogelijke locale factoren, doordat de uitkomst gepresenteerd wordt als een relatief risico of een verwante maat die aangeeft in hoeverre behandeling A beter is dan behandeling B. Als echter met verschillende patiënten-populaties, onder verschillende

omstandigheden met mogelijk andere technieken gewerkt wordt, is er geen reden om a priori te veronderstellen dat het relatief risico van therapie A ten opzichte van therapie B een universele natuurconstante is. Toetsen op homogeniteit versus heterogeniteit hebben dan ook een geringe waarde omdat de nulhypothese van homogeniteit niet erg plausibel is.

Realistischer is om, zoals onder andere bepleit in een recent artikel van DerSimonian en Laird¹²⁾, uit te gaan van de heterogeniteit van de trials. Fraai zou zijn als die heterogeniteit verklaard zou kunnen worden met behulp van bepaalde kenmerken van de trials in zoiets als een regressieanalyse. Vaak echter is er verder nauwelijks nadere informatie over de trials aanwezig en kan er het beste uitgegaan worden van een toevallig optredend verschil in effect (bijvoorbeeld relatief risico) van trial tot trial. Dit is het klassieke variantiecomponentenmodel dat in de statistiek al lang bekend is en stamt uit de landbouwkundige veredelings-techniek. Onderscheiden worden de primaire eenheden, in een meta-analyse de trials, en de secundaire eenheden, in de meta-analyse de patiënten. Gebaseerd op deze klassieke technieken kan geschat worden wat het gemiddelde effect is, gemiddeld over de trials, en wat de spreiding in het werkelijke effect van trial tot trial is. Dat kan dan weer gebruikt worden om na te gaan wat, bij voorbeeld, de kans is dat bij een willekeurige trial een positief werkelijk effect optreedt. Het verschil tussen deze aanpak en de Peto-aanpak is dat bij Peto in zekere zin de patiënten als primaire onderzoekseenheden optreden, terwijl bij de laatste aanpak de trials zelf als primaire onderzoeks eenheid optreedt.

Deze aanpak is overigens niet alleen beperkt tot de meta-analyse. Ook bij multi-centre trials dient men na te gaan in hoeverre het effect van centrum tot centrum varieert. In technische termen is de interactie tussen centrum en behandeling van belang. In de grote bypass-study¹³⁾ van de EC/IC Bypass Study Group wordt geconcludeerd dat gemiddeld over alle centra de bypass geen

effect heeft, zonder na te gaan in hoeverre het effect van centrum tot centrum varieert. Er zullen natuurlijk centra blijven, die beweren dat het bij hen wel werkt. Deze discussie, die nu wel gevoerd wordt¹⁴⁾, had vermeden kunnen worden door de interactie tussen centrum en therapie beter te bestuderen.

Een ander voordeel van de bovengeschetste methode van meta-analyse is de mogelijkheid om een gecorrigeerde effect-schatting per trial te geven. Deze gecorrigeerde schatting zal in het algemeen dicht bij het gemiddelde of "overall" effect liggen. Op die manier wordt een compromis bereikt tussen het rücksichtlos combineren van de resultaten à la Peto en het volledig geïsoleerd bekijken van elke trial apart. De op deze wijze verkregen gecorrigeerde effect-schattingen liggen dicht bij de werkelijkheid dan de oorspronkelijke ongecorrigeerde schattingen. De constructie van deze verbeterde schattingen is gebaseerd op de zogenaamde Bayesiaanse modellering dat het effect van trial tot trial op toevallige wijze varieert en de regel van Bayes om tot een betere schatting te komen.

Omdat en passant de onderliggende verdeling van het effect over de trials empirisch wordt bepaald, spreekt men in dit verband wel van de empirische Bayes-methodologie. Deze methodologie is in de jaren vijftig in een algemener en meer theoretisch kader geïntroduceerd door Robbins¹⁵⁾ en toen door sommigen als een doorbraak in statistische methodologie beschouwd¹⁶⁾. Zelf heb ik jaren theoretisch onderzoek op dit terrein gedaan en het heeft mij dan ook aangenaam getroffen dat deze methodologie nu na zoveel jaren haar praktische toepassing gaat vinden onder andere in de medische wereld.

Op de geschetste manier verijnd is de meta-analyse niet alleen in staat om de trials als bladeren bij elkaar te harken, maar om ze ook ondertussen naar effect-grootte te sorteren. Ook mijn sectie probeert aan de ontwikkeling van deze techniek haar steentje bij te dragen.

Aan het eind van mijn rede wil ik allereerst Hare Majesteit de Koningin mijn dank betuigen voor het besluit mij op deze leerstoel te benoemen. De faculteit Geneeskunde dank ik voor het in mij gestelde vertrouwen en de welwillende wijze waarop men mij in de eerste periode is tegemoet getreden. Veel van wat ik hier ter berde heb ik gebracht, is de vrucht van de vele gesprekken die ik de laatste maanden met medici gevoerd heb. Ik hoop het vertrouwen niet te beschamen en verder voort te gaan op de weg van consolidatie van de zeer nuttige rol die medische statistiek in deze faculteit speelt en het onderzoeken van nieuwe technieken die voor medisch onderzoek van groot belang kunnen zijn. Voor een vruchtbare samenwerking is wederzijds contact nodig. Ik zal dat zeker zoeken en ik ben ervan overtuigd dat u van uw kant voor uw statistische en methodologische problemen de weg naar mijn sectie zult weten te vinden.

Lieve ouders,

u wil ik bedanken voor het feit dat u mij op het pad gezet hebt dat hierheen geleid heeft, ondanks het feit dat de academische wereld u volledig vreemd was en u enkel af kon gaan op de adviezen van mijn onderwijzers en leraren. Uw morele steun heb ik altijd hogelijk gewaardeerd en het verschaft mij bijzonder veel vreugde dat u vandaag hier aanwezig bent.

Lieve Hermien en alle kinderen,

onze gecompliceerde gezinssituatie is door mijn benoeming hier in Leiden er niet eenvoudiger op geworden. Onze onderlinge band heeft op mij altijd zeer inspirerend gewerkt en ik ben blij dat jullie allemaal in de vreugde van deze dag kunt delen.

Hooggeleerde Leppink, beste Gerard,

op jouw instituut heb ik mij kunnen ontwikkelen tot een volwaardig statisticus en heb ik mij de praktische en theoretische vaardigheden kunnen eigen maken die voor de beoefening van het vak onontbeerlijk zijn. Jij hebt

mij de ruimte gelaten om mijn eigen onderzoek, onderwijs en consultatie te doen, maar was altijd bereid om over alle zaken van gedachten te wisselen en advies te geven. De gelegenheid om in alle rust over de problemen na te denken mis ik nu soms node.

Mijne heren medewerkers en oud-medewerkers van de Mathematische Statistiek in Utrecht, vele jaren hebben wij lief en leed gedeeld en een nostalgisch gevoel bekruipt mij telkenmale als ik aan die tijd terug denk. Met name Joop Faber en André de Meijer wil ik bedanken voor de goede tijd in Utrecht.

Mijne dames en heren van de sectie Medische Statistiek, u wil ik danken voor de wijze waarop u mij, toch een volledige buitenstaander, in uw midden hebt opgevangen. In deze woelige tijden van krimp en groei rust op onze sectie de verantwoordelijkheid om niet alleen anderen bij te staan in de statistische en methodologische aspecten van hun onderzoek, maar om ook het eigen vakgebied te blijven verkennen en waar mogelijk te vernieuwen. De toenemende druk van buiten zullen wij gezamenlijk moeten weerstaan om op die wijze zowel de kwaliteit van ons werk als ons goede humeur te bewaren. Ik heb er het volste vertrouwen in dat dat zal lukken.

Dames en heren studenten van de faculteit Geneeskunde, ik heb getracht u iets van het statistisch instrumentarium te schetsen. Met behulp van soms vergezochte metaforen heb ik getracht aan te tonen dat statistiek meer is dan een verzameling ijverig gehanteerde, doch vaak onbegrepen formules. Wilt u die formules zinvol hanteren, dan zult u op zijn minst de achterliggende gedachte moeten begrijpen. Van mijn kant zal ik mij alle inspanning getroosten om u dat begrip bij te brengen.

Ik heb gezegd.

Verwijzingen.

1. E. R. Tufte, The Visual Display of Quantitative Information, Graphics Press, 1983.
2. D. Huff, How to Lie with Statistics, Penguin, 1973.
3. W. Playfair, The Commercial and Political Atlas, 1786.
4. T. Bayes, Essay Towards Solving a Problem in the Doctrine of Chances, 1763.
5. J. J. Veltkamp, E. F. Drion, E. A. Loeliger, Thrombos. Diathes. haemorrh., XIX, 403-422, 1968.
6. J. Cornfield, Proc. 3rd Berkeley Symp., 4, 135-148, 1956.
7. D. J. Finney, Probit Analysis, Cambridge University Press, 1952.
8. D. R. Cox, The Analysis of Binary Data, Methuen, 1970.
9. J. D. F. Habbema, J. M. H. Hermans, Statistical Methods for Clinical Decision Making, Dissertatie Leiden, 1978.
10. R. D. Gill, Gemiste Kansen, Oratie Leiden, 1987.
11. R. Peto, Lancet i, 1172-1173, 1980.
12. R. DerSimonian, N. Laird, Contr. Clinic. Trials, 7, 177-188, 1986
13. EC/IC Bypass Study Group, New England J. Med., 313, 1191-2000, 1985.
14. H. Robbins, Proc. 3rd Berkeley Symp., 1, 157-164, 1956.
15. J. Neyman, Rev. Inst. Intern. Statist., 30, 11-27, 1962,

Ontvangen 04-06-1987
Geaccepteerd: 04-06-1987