VERENIGING „NEDERLANDSCH TIJDSCHRIFT VOOR GENEESKUNDE"

WETENSCHAPPELIJKE VERGADERING OP 7 DECEMBER 1957 IN DE ZETEL DER VERENIGING, JAN LUYKENSTRAAT 5, AMSTERDAM

VOORZITTER: PROF. DR. J. R. PRAKKEN

Prof. Dr. J. HEMELRIJK*, *Statistische proefopzetten*

Bij ieder experiment, van welke aard dan ook, is het van groot belang, dat men „weet wat men doet". Merkwaardigerwijze houdt dit bij experimenten van statistische aard vaak in, dat men zorgvuldig ervoor moet zorgen, dat men in bepaalde op.. .nten *niet* weet wat men doet.

Deze situatie is op aanschouwelijke wijze voor te stellen met behulp van de zg. speltheorie. Een eenvoudig spel van psychologische aard is het volgende: twee spelers nemen ieder naar eigen keuze 0, 1 of 2 geldstukken in de gesloten hand en leggen deze hand op tafel. Ieder mag nu raden, hoeveel geldstukken de ander in de hand heeft en indien de een goed raadt en de ander niet, wint de eerste een vaste inzet. Een speler die de psychologie van zijn tegenstander doorziet, kan met dit spel de ander gewoonlijk gemakkelijk overtroeven. Als verweer hiertegen kan nu de zwakkere speler trachten, zijn blijkbaar al te doorzichtige tactiek gecompliceerder, of althans minder gemakkelijk voorspelbaar, te maken. Het is niet zo gemakkelijk, dit doeltreffend te doen, terwijl de tegenstander bovendien een dergelijke verandering in de tactiek ook weer kan doorzien. De mogelijkheid bestaat echter om iedere psychologische tactiek uit te schakelen door te zorgen, dat men zelf niet weet wat men zal gaan doen, zodat ook de tegenstander dat niet kan raden. Daartoe behoeft men het aantal munten dat men in de hand neemt, slechts te laten bepalen door een toevalsmechanisme, bv. door te werpen met een dobbelsteen (waarbij men de 5 en de 6 voor 0 telt, de 1 en de 3 voor 1, en de 2 en de 4 voor 2). Het essentiële van een dergelijk toevalsmechanisme is nl. juist de *onvoorspelbaarheid* van de uitkomst, een onvoorspelbaarheid, die niet alleen voor de speler zelf geldt, maar ook voor zijn tegenstander. Op dezelfde wijze kan men bovendien bepalen, welk aantal men zal raden bij zijn tegenspeler; men kan daar zelfs hetzelfde getal voor nemen — althans als de ander het eerst moet raden. Op deze wijze bereikt men, dat de winstkansen voor beide spelers gelijk worden, wat ook de tactiek van de tegenspeler moge zijn. Het spel is gereduceerd tot een „eerlijk" toevalsspel. De complicatie, dat de tegenspeler misschien bezwaar zal maken tegen het werpen met een dobbelsteen, kan hier gevoeglijk buiten beschouwing blijven.

Dit soort voor-de-gek-houderij heeft op zichzelf niets met medische experimenten te maken, maar toch is het een van de essentiële bijdragen van de statistiek tot het perfectioneren van experimenten van velerlei — ook van medische — aard. Men kan, zoals in het bovengenoemde geval, ingewikkelde en onoverzichtelijke situaties in één slag ermee vereenvoudigen tot theoretisch en statistisch gemakkelijk te beheersen problemen en men bereikt daarmee twee belangrijke resultaten. In situaties waarin zich onoverzichtelijke en niet geheel (of in het geheel niet) bekende invloeden kunnen voordoen, kan men nl. enerzijds „schijneffecten" verwachten, terwijl anderzijds effecten die men graag zou ontdekken, door die onbekende invloeden „verdoezeld" en daardoor aan de waarneming onttrokken kunnen worden. Wij zullen trachten, dit aan enkele eenvoudige — en daarom wellicht wat gekunstelde — voorbeelden te demonstreren.

Een proefpersoon beweert, dat hij twee slaapmiddelen naar aanleiding van hun uitwerking van elkaar kan onderscheiden. Het ene (A) werkt snel, maar kort; het andere (B) werkt langzamer, maar langduriger. Het is zeer wel mogelijk, dat hij gelijk heeft, maar men wenst dit niet zonder bewijs aan te nemen. Daarom geeft men de proefpersoon een aantal dagen lang een capsule, die hij voor het slapen inneemt, en die één van beide slaapmiddelen bevat. De proefpersoon mag natuurlijk niet weten, welk middel dat is; hem wordt gevraagd, de volgende dag naar aanleiding van zijn bevindingen zijn mening hierover mede te delen. Deze proef wordt een aantal malen herhaald, tot van beide middelen een niet te klein aantal capsules is verstrekt. Vervolgens worden de door de proefpersoon gegeven antwoorden met de werkelijkheid vergeleken, opdat men kan beslissen of zijn bewering juist is of niet. De statistische techniek die daarbij wordt gebruikt, behoeft hier niet te worden uitgelegd, daar het hier alleen om de proefopzet gaat.

Welke voorzorgen moeten nu worden genomen om te zorgen, dat het experiment zich voor een goede statistische uitwerking leent? Ten eerste uiteraard de reeds genoemde, dat de proefpersoon niet mag weten, welk middel hij op een bepaalde dag toegediend krijgt. Ook indien hij hiervan niet op de hoogte is, kunnen zich echter nog complicaties voordoen. Immers de arts die vaststelt, in welke volgorde de middelen A en B worden verstrekt, moet deze volgorde op de een of andere wijze bepalen; indien hij zich daarbij door zijn eigen intuïtie of willekeur laat leiden, valt het niet te vermijden, dat in deze volgorde tot op zekere hoogte een patroon verschijnt. Zo zal bv. ongetwijfeld de neiging aanwezig zijn om te vermijden, dat gedurende een relatief groot aantal dagen achtereen hetzelfde middel wordt gegeven. De proefpersoon zal zijnerzijds bij het vaststellen van zijn antwoorden een soortgelijke neiging tonen, daar hij weet, dat hij nu eens het ene dan weer het andere middel krijgt. Indien de psychologie van de arts en van de proefpersoon enigszins parallel lopen in dit opzicht, is de kans niet gering, dat de proefpersoon een vrij groot aantal malen het juiste antwoord zal geven, ook als hij er in feite alleen maar naar raadt.

Op deze wijze kan dus een verkeerde conclusie uit de bus komen, in de vorm van een „schijneffect". Deze term is eigenlijk misplaatst. Het effect is op zichzelf reëel genoeg, maar het heeft met de te onderzoeken vraag niets te maken.

Rapport S 225 (V 17) van de Statische Afdeling van het Mathematisch Centrum te Amsterdam.

*Hoogleraar Technische Hogeschool Delft; Adviseur voor Statistische Consultatie van het Mathematisch Centrum, Amsterdam.

De parallel met het spelvoorbeeld is duidelijk: de arts en de proefpersoon spelen als het ware tegen elkaar, en ook als de proefpersoon niet bewust zal trachten, het schema van de arts te doorzien, dan zal hij dit toch onbewust doen. Geen enkel schema, behalve een toevallige volgorde, kan aan dit bezwaar tegemoet komen. Hoe ingenieus men een schema ook zou maken, altijd bestaat de mogelijkheid, dat men de één of andere systematische overeenkomst met de psychologie of het „slaappatroon" van de proefpersoon over het hoofd ziet. Indien men echter de volgorde door loting bepaalt, zijn deze zorgen opgeheven, daar dan iedere systematiek ontbreekt. En met de grilligheden van het toeval wordt door de statisticus bij zijn verwerking rekening gehouden. Daar is nl. de gehele statistische techniek juist op gericht en zelfs op gebaseerd.

Overigens zijn zelfs dan nog niet àlle zorgen van de baan, daar de mogelijkheid van psychologische invloed op de proefpersoon door de arts die hem het middel dagelijks overhandigt, niet uitgesloten is. Indien de arts weet, welk middel hij de patiënt geeft, kan dit — ook zonder dat men aan de wat vergezochte mogelijkheid van telepathie denkt — toch misschien invloed hebben op zijn wijze van overhandigen daarvan, waardoor opnieuw moeilijkheden van dezelfde aard als boven geschetst, kunnen ontstaan. (Men bedenke, dat een soortgelijk gevaar in hoge mate aanwezig is, als niet de proefpersoon zijn oordeel over de uitwerking van een middel geeft, maar indien de arts deze moet bepalen). Het is dus van belang, dat de capsules aan de proefpersoon worden overhandigd door iemand die niet weet, wat zij bevatten. Dit is, bij een proef als de hier beschrevene, gemakkelijk te realiseren. En het valt aan te bevelen, deze proefopzet aan de proefpersoon mede te delen; ook het feit, dat het uit te reiken middel dagelijks door loting wordt bepaald, daar dit hem duidelijk de vruchteloosheid van raden op grond van psychologische overwegingen voor ogen stelt. Hij kan zich dan ongestoord concentreren op het vaststellen van de uitwerking van het middel.

Hadden wij hier een voorbeeld van een mogelijk schijneffect, een ander voorbeeld moge duidelijk maken, dat evengoed een reëel effect kan worden verdoezeld. Veronderstel, dat een arts twee geneesmiddelen of -methodes tot zijn beschikking heeft voor een ziekte die niet gevaarlijk is, maar zich toch in duidelijk verschillende graden van ernst kan voordoen. Het ene middel (X) kent hij reeds lang, het andere (Y) is nieuw, en de werking is hem nog niet uit eigen ervaring bekend. Wel neemt hij aan — op grond van gepubliceerde ervaringen van anderen — dat middel Y niet schadelijk is. Hij besluit nu „middel Y ook eens te proberen", maar beperkt zich daarbij in eerste instantie tot zijn niet-ernstig zieke patiënten. De ernstig zieke patiënten geeft hij het beproefde middel X.

Deze situatie zal zich ongetwijfeld in de praktijk voordoen en nu valt in het geheel niet te voorspellen, wat er zal gebeuren. De twee groepen van met verschillende middelen behandelde patiënten zijn nl. ten gevolge van de toegepaste selectie in het geheel niet vergelijkbaar meer. Als het nieuwe middel Y in feite beter is dan het oude, is het toch niet onmogelijk, dat dit, bij vergelijking van de met beide middelen behaalde resultaten, niet wordt opgemerkt; eventuele verbeteringen bij ernstig zieke patiënten zullen immers wellicht duidelijker zijn dan bij de minder zieke. Zo kan een betere werking van Y zich aan de waarneming onttrekken. Anderzijds is het mogelijk, dat middel X beter is dan Y, doch dat de ziekte bij de ernstigere patiënten veel hardnekkiger is dan bij de minder ernstige; middel Y verkeert dan in een gunstige positie, omdat juist de gemakkelijker te genezen patiënten dit krijgen. Het is in den regel zeer moeilijk, zo niet onmogelijk, een dergelijke warwinkel te ontwarren en een goed gefundeerd oordeel te vormen.

Daar komt nog bij, dat de graad van vertrouwen, die de arts in de beide middelen stelt, invloed op zijn diagnose zal kunnen uitoefenen omtrent eventuele vooruitgang der patiënten. Daardoor komt het middel waarin hij het meeste vertrouwen stelt, in het voordeel tegenover het andere. Ook dit dient te worden vermeden.

Een „waterdichte" proefopzet krijgt men weer alleen als de arts ervoor zorgt, dat hij tot op zekere hoogte niet weet, wat hij doet. Indien hij ernstig zieke patiënten nog niet met middel Y wil behandelen, blijven deze (voorlopig) buiten de proef. De arts bepaalt dus eerst, of een patiënt zal meedoen — naar aanleiding van de ernst van de ziekte — en als dit niet zo is, behandelt hij hem gewoon met X. Het resultaat van deze behandeling wordt echter niet gebruikt bij de vergelijking der twee middelen. Wil hij de patiënt wel in de proef betrekken, dan loot hij (op de één of andere wijze; daarover later) of hij middel X of Y zal geven. Vervolgens geeft hij hem dit middel, daarbij zorg dragend, dat hijzelf voorlopig niet weet, welk het is. Dit lijkt ingewikkeld en het is ook niet altijd realiseerbaar. Vaak kan men echter wel degelijk aan deze eisen voldoen. Om de situatie te vereenvoudigen: indien het gaat om de toediening van pillen kan de arts van tevoren een aantal doosjes klaarmaken, de helft met X, de andere helft met Y. De doosjes worden door loting genummerd; op een lijst, die daarna tot het einde van de proef wordt opgeborgen, wordt aangetekend welke nummers X en welke Y bevatten en de doosjes worden verder in willekeurige volgorde (maar het mag ook op nummer) aan de niet ernstige patiënten uitgereikt. Als de arts geen fenomenaal geheugen heeft, is hij bij het uitreiken allang vergeten, welk middel in het uitgereikte doosje zit. Op deze wijze is enerzijds zijn keuze van het middel inderdaad toevallig en anderzijds staat zijn beoordeling niet meer onder invloed van de — nu immers niet aanwezige — kennis.

Natuurlijk kan men bij dergelijke proeven veel gemak ondervinden van de hulp van een assistent(e), vooral in moeilijker te organiseren gevallen, daar deze wèl mag weten, wat de arts voorlopig niet dient te weten.

Vermoedelijk worden zo ver gaande voorzorgen vrij zelden genomen, terwijl er wellicht ook een zekere weerstand bestaat tegen zulke hocus-pocus. Niettemin zou het uiterst belangrijk zijn, indien men irrelevante gevoelsweerstanden overwon en er een gewoonte van maakte, deze soort, in wezen vrij eenvoudige voorzorgen te nemen. Het vormen van een oordeel is, zelfs bij een goede proefopzet, al moeilijk genoeg. Bij een slechte proefopzet moet het, behalve als het verschil tussen de vergeleken middelen onmiskenbaar is, vaak als vrijwel onmogelijk worden beschouwd.

Om dit aan te tonen beschouwen wij een getallenvoorbeeld. Wij veronderstellen daartoe, dat de twee middelen X en Y achter elkaar bij dezelfde patiënt kunnen worden beproefd en dat de volgorde waarin dit geschiedt, van geen invloed is op de uitwerking en de beoordeling (wellicht wat ver gaande veronderstellingen, màar die voor dit voorbeeld gewenst zijn). Bij iedere patiënt kan nu worden nagegaan op welk der twee middelen hij het gunstigst reageert (zonder dat uit die waarneming de conclusie moet worden getrokken, dat dit verschil zich bij iedere patiënt afzonderlijk, steeds op deze wijze zal voordoen; indien twee waarnemingen — om welke reden dan ook — ongelijk zijn, zal altijd één van beide de grootste zijn). Indien nu een arts bij 20 patiënten verschillende reacties op de middelen heeft gevonden en bij 14 van hen de uitwerking van X gunstiger was, doch slechts

bij 6 die van Y, wat zal hij daaruit dan concluderen? De kans lijkt mij niet gering, dat hij middel Y maar weer zal laten vallen en voortaan weer steeds met X zal werken. Deze handelwijze zou echter overhaast zijn. De kans op een dergelijke of een nog extremere uitkomst is nl., als X en Y even goed zijn, nog ongeveer 1 : 9. Dit betekent, dat de geschetste methode, bij vergelijking van twee gelijkwaardige middelen, in meer dan 10 pct van zijn toepassingen tot de verkeerde conclusie leidt, dat één der middelen beter is dan het andere. Is dit wèrkelijk het geval, is bv. Y beter dan X, dan is de kans, dat men toch X als het beste zal aanwijzen, uiteraard kleiner, maar nog geenszins te verwaarlozen. Indien men zijn oordeel op intuïtieve wijze bepaalt, zal men vrij vaak het slechtste middel voor het beste verslijten, tenzij het verschil zo groot is, dat er geen twijfel over kan bestaan.

Laten wij nu veronderstellen, dat een arts, zich bewust van de verscherpte blik die statistische verwerking van zijn waarnemingen hem kan verschaffen, naar middelen uitziet om tot een verantwoorde conclusie te komen. In de eerste plaats zal hij dan trachten, de lengte van de proefreeks minder klein te nemen, daar bij een kleinere proef alleen grove verschillen een redelijke kans bezitten om te worden ontdekt (in statistisch jargon: het *onderscheidingsvermogen* van een proef neemt sterk toe met het aantal waarnemingen). Hij breidt de reeks bv. uit tot 100 proeven. Indien hij zijn kans op een verkeerde conclusie tot ten hoogste 1 pct wil beperken, mag hij — volgens de statistici — pas tot een verschil tussen de middelen concluderen, als één van beide middelen in ten minste 64 gevallen van de 100 een betere uitwerking heeft dan het andere. Tot zoverre is alles goed. Maar indien de arts er niet voor heeft gezorgd, onwetend te zijn omtrent de gebruikte middelen, en zijn beoordeling — zij het in geringe mate — onder invloed zou staan van zijn persoonlijke verwachting omtrent de deugdelijkheid der middelen, zodat hij bv. in 5 twijfelgevallen (op 100 proeven) ten onrechte dàt middel gunstig zou beoordelen, waarvan hij het meest verwacht, dan wordt de kans, dat dit middel ten onrechte als het beste zal worden aangewezen, alweer tot ongeveer 5 pct verhoogd. Laat hij zich in 7 twijfelgevallen door zijn mening verleiden tot een gunstige uitspraak, dan wordt deze kans zelfs 10 pct. De proef die men aldus aanzienlijk beter onderscheidend maakt door het aantal waarnemingen te vergroten, wordt daardoor ook veel gevoeliger voor (ook kleine) fouten in de proefopzet en in de beoordeling der afzonderlijke waarnemingen, hetgeen onderstreept, dat alleen een zorgvuldige proefopzet zowel betrouwbaar als onderscheidend kan werken. Het uitvoeren van niet zorgvuldig opgezette, kleine proeven is verspilling van arbeid en kennis, daar deze geen objectieve beoordeling toestaan; het is nog meer verspilling, uitgebreide proeven te doen om het onderscheidingsvermogen te vergroten, en daarbij de betrouwbaarheid prijs te geven door na te laten, de proef zorgvuldig voor te bereiden en uit te voeren.

De dokter behoeft uiteraard niet zo ver te gaan, dat hij dobbelstenen hanteert bij het behandelen van zijn patiënten. De patiënten zouden de indruk kunnen krijgen, dat er met hen wordt gesold; ook al zou deze indruk misplaatst zijn, toch is het wenselijk, dat hij in het geheel niet wordt gewekt. Er zijn uitgebreide tabellen beschikbaar van zg. aselecte getallen (Engels: „random numbers"), waarvan men zich kan bedienen voor alle doeleinden van loting. Deze zou men — zo nodig — zelfs in aanwezigheid van de patiënt kunnen raadplegen, zonder diens toorn of wantrouwen op te wekken. Bij dit alles vergete men niet, dat het juist in het belang van de patiënten is, dat betere geneesmiddelen zo spoedig mogelijk worden ontdekt, en niet door onvolmaakte proefopzetten onopgemerkt blijven.

Naast het boven nogal uitvoerig besproken elementaire maar fundamentele, en desondanks niet vaak consequent toegepaste lotingsprincipe kan de statistiek nog vele andere bijdragen tot de goede proefopzet leveren. Het zou te ver voeren, daarop in te gaan; wij noemen slechts enkele middelen: de sequente analyse, die erop uit is, het aantal waarnemingen, dat nodig is om tot een conclusie te komen, zoveel mogelijk te beperken; het verdelen van de patiënten in zo homogeen mogelijke groepen om met behulp van een dan iets ingewikkelder analyse een groter onderscheidingsvermogen te verkrijgen (in dit verband is het van belang, verschillende details die in verband met ziekte, behandeling, diagnose en genees-snelheid kunnen staan, van alle patiënten te noteren; liever te veel dan te weinig); uitgebalanceerde proefopzetten voor ingewikkelde experimenten (waarbij men het niet zonder voorafgaand advies van een statisticus zal kunnen stellen; het inwinnen van een dergelijk advies is trouwens altijd aan te raden).

Indien men, zoals in de voorafgaande zin wordt gesuggereerd, een statisticus raadpleegt, moet men dit doen, voor men aan een experiment begint. Hierop kan nauwelijks genoeg nadruk worden gelegd. Men zou zelfs kunnen zeggen, dat deze stap tot de proefopzet behoort. Want de statistiek is helaas geen allesomvattende methode, waarmee men ieder waarnemingsmateriaal met goed gevolg kan analyseren. Verre van een wondermiddel te zijn, doet statistiek bij niet zorgvuldig opgezette proeven aan Haarlemmer olie denken: het helpt niet werkelijk. Talloze onderzoekingen zijn gestrand op een onzorgvuldige opzet; de statisticus kan dan alleen nog maar de fouten in deze opzet en hun noodlottige gevolgen aanwijzen, zonder de onderzoeker uit de daaruit voortvloeiende moeilijkheden te kunnen helpen.

Wendt men zich echter tevoren tot een statisticus, dan zal men vrijwel steeds al dadelijk voor de vraag worden gesteld: „Wat is precies het doel van uw onderzoek?". De beantwoording van deze vraag kost vaak meer moeite dan men zou verwachten, en juist daaruit blijkt het grote nut van de vraag. Want al hebben wij betoogd, dat de onderzoeker in sommige opzichten zorgvuldig moet vermijden, te weten wat hij doet, hij dient deze onwetendheid ook zorgvuldig te beperken tot die punten, welke de statisticus hem aangeeft; overal elders werkt onwetendheid negatief; dat geldt uiteraard niet alleen voor de arts, maar voor iedere onderzoeker.

Is de vraag naar het doel van het onderzoek eenmaal bevredigend beantwoord, dan luidt de tweede vraag: „En hoe denkt U dat doel te bereiken?". Bij het beantwoorden van deze vraag kan de statisticus reeds actief behulpzaam zijn. Want al is hij gewoonlijk leek op het gebied van de geneeskunde, hij is het niet op het gebied van proefopzetten, en de statistische aspecten van de proefopzet dienen — evenals trouwens de statistische analyse achteraf — te worden beschouwd als essentiële onderdelen van het onderzoek als geheel. Te vaak nog worden deze aspecten (ook bij de begroting in tijd en geld) niet als onderdelen van de proef beschouwd, hetgeen er niet zelden toe leidt, dat wel waarnemingsmateriaal wordt verzameld, maar dat dit achteraf niet grondig kan worden geanalyseerd, òfwel omdat tijd en geld, of zelfs een statisticus ontbreken, òfwel omdat het materiaal achteraf niet geschikt blijkt te zijn voor het gestelde doel. Dit lijkt misschien absurd, maar dat het

vóórkomt, zal iedere onderzoeker bekend zijn. Dit kan zelfs zo ver gaan, dat de statistiek er een slechte naam door krijgt. Heeft men echter tevoren voldoende n op de juiste wijze rekening met de statistische aspecten gehouden, dan kan men vaak in verrassend korte tijd resultaten bereiken, waarvoor men anders veel langer nodig zou hebben, of die zelfs vrijwel onbereikbaar zouden zijn gebleven. Waarmee niet wil zijn gezegd, dat een gesprek met een statisticus vooraf niet tot teleurstellingen zou kunnen leiden. Zo kan bv. blijken, dat het doel dat men zich stelt, langs de voorgestelde weg niet bereikbaar moet worden geacht, bv. omdat de statistiek (nog) niet beschikt over een analysemethode die bij het gestelde probleem past, ofwel daarvoor onuitvoerbare proeven voorschrijft. Ook deze teleurstelling kan nuttig zijn, omdat er gewoonlijk weinig kans is, dat men dan toch met een experiment iets zou hebben bereikt. Wat de tijdsduur van de statistische verwerking van uitgebreid waarnemingsmateriaal betreft, deze is in het algemeen ongeveer even groot als de tijd, nodig voor het verrichten van de waarnemingen. Stippelt men echter de proef uit met een bepaalde statistische analysetechniek in gedachten, dan kan er vaak aanzienlijke tijd worden bespaard.

Men kan de statistiek bij medische en fysiologische (en ook bij andere) proefnemingen op twee wijzen toepassen: als detectiemiddel (toepassing van statistische toetsing op reeksen van proeven of op uitgebreid waarnemingsmateriaal zonder van tevoren scherp gestelde vragen en van tevoren bepaalde statistische analysemethodes: „Zie maar eens, wat eruit te halen is") en als bewijsmiddel. Het uitermate belangrijke onderscheid tussen deze twee gebruiksmethodes[3] wordt nog wel eens uit het oog verloren. Zoekt men bv. naar nieuwe geneesmiddelen voor een bepaalde ziekte en probeert men vele middelen op hun werkzaamheid, daarbij iedere proef goed opzettend en bij de verwerking statistische toetsen toepassend, dan kan men nog lelijk bedrogen uitkomen, als men geen scherp onderscheid maakt tussen deze twee methodes. De oorzaak daarvan is, dat iedere statistische analyse — zoals ook iedere andere methode — de mogelijkheid tot verkeerde conclusies openlaat. Dit is misschien betreurenswaardig, maar het is onvermijdelijk. Alleen door vaak herhaalde bevestigingen wordt praktische zekerheid verkregen, absolute zekerheid nooit. Daarom is het zo nuttig, proeven die door anderen zijn verricht, te herhalen, ook al lijkt het op het eerste gezicht wellicht verspilling van energie (wat het in bijzondere gevallen ook wel is). Het verschil tussen de statistiek en de meeste andere wetenschappen is, dat men zich niet alleen van de mogelijkheid van verkeerde conclusies scherp bewust is, maar dat de mate van onbetrouwbaarheid — bij een goede proefopzet althans — ook exact in een getal kan worden uitgedrukt, zodat men weet, hoe vaak men een verkeerde conclusie trekt. Daardoor is men dan anderzijds weer wat scheutiger met die fouten: veelal wordt een kans van 1 : 20 op een verkeerde conclusie toegelaten. Dit betekent, dat bij het onderzoeken van 100 volkomen onwerkzame middelen, er ongeveer 5 toch werkzaam zullen schijnen door toevallige omstandigheden. En houdt men daarmee geen of onvoldoende rekening, dan valt het niet te vermijden, dat men — zelfs bij een overigens juiste proefopzet — onwerkzame middelen als werkzame aanprijst. Als men nu maar kon aanwijzen, niet alleen hoeveel fouten er in een reeks statistische analyses ongeveer worden gemaakt, maar ook welke conclusies fout zijn, dan was men van deze moeilijkheid verlost. En dit kan, als men van de statistiek als detectiemiddel overgaat op de statistiek als bewijsmiddel. Men kan een statistisch verwerkte proef alleen dan beschouwen als een overtuigende aanwijzing voor een bepaald verschijnsel, bv. de werkzaamheid van een geneesmiddel, indien aan de volgende eisen is voldaan:

A. Het doel van het onderzoek moet tevoren duidelijk zijn geformuleerd en daarvan wordt niet afgeweken.

B. De te verrichten proeven worden tevoren zorgvuldig beschreven en daarvan wordt niet afgeweken.

C. De statischische analysemethodes worden tevoren gekozen en van deze keuze wordt niet afgeweken.

D. Uit een reeks proeven, uitgevoerd volgens plan, worden alleen conclusies getrokken omtrent de tevoren gestelde vragen en met hun aantal wordt rekening gehouden bij de keuze van de onbetrouwbaarheid (kans op een onjuiste conclusie), die voor iedere toetsing afzonderlijk wordt toegelaten.

E. De mogelijkheid van schijneffecten (en verdoezelingen) wordt tevoren tenietgedaan, door in die stadia van de proeven, waarin onbekende of onberekenbare oorzaken systematische invloed zouden kunnen uitoefenen, statistische aselectering (loting) toe te passen.

Dit lijstje van lang niet lichte voorwaarden is vermoedelijk nog niet eens volledig; in ieder geval zou het gemakkelijk tot een nog meer afschrikwekkende lijst kunnen worden uitgebreid. Men houdt er zich dan ook vrijwel nooit volledig aan. Als zich tijdens de proef een onverwacht effect voordoet (schijnt voor te doen), is de verleiding om de proef onderweg te wijzigen of de statistische analyse uit te breiden tot dit nieuwe verschijnsel, te groot. In vele gevallen zou het ook onjuist zijn, niet aan deze verleiding toe te geven. Maar men beseffe wel, dat men dan van bewijsmiddel teruggaat naar detectiemiddel.

In ieder uitgebreid waarnemingsmateriaal doen zich wel bijzondere constellaties van cijfers voor. (Dit is ook het geval, als men in een tabel van aselecte getallen bladert; men vindt daarin bv. soms vrij lange rijen van gelijke getallen achterelkaar). Toetsing van zulke bijzondere constellaties en pogingen tot verklaring op medische gronden zijn ongetwijfeld zinvol, indien men er maar voldoende van doordrongen is, dat een nadere bevestiging door een proef die wèl aan de boven geformuleerde eisen voldoet, onontbeerlijk is.

Ter illustratie bespreken wij een passage uit het proefschrift van K. H. Brandt (1957) *Over de plaats van vorming der urobilinogenen in het menselijk organisme.* (De hiervolgende beschouwingen zijn ontwikkeld in samenwerking met Prof. Dr. C. G. G. van Herk, medewerker van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam).

De bedoelde passage betreft het verschil tussen de zg. B-gal die wordt uitgescheiden na de galblaas te zijn gepasseerd, en de C-gal die rechtstreeks uit de lever komt. In de galblaas wordt de gal door resorptie van water ingedikt; bij eenzelfde persoon is de bilirubineconcentratie in de B-gal dus aanmerkelijk hoger dan die in de C-gal. Bij 21 patiënten van Brandt, die bijna allen een onbelaste anamnese hadden

TABEL I. Concentraties van urobilinogenen (in mg/100 ml) en bilirubine (in e/100 ml) in B- en C-gal

Nr. patiënt	$[ur]_B$	$[ur]_C$	$[bil]_B$	$[bil]_C$
1	0,37	0,06	96,96	15,70
2	0,108	0,005	99,26	2,82
9	0,312	0,208	98,7	8,7
12	1,38	0,57	46	9,74
19	0,01	0,005	20,6	13,1

wat lever- en galwegen betreft (één patiënt had een steen; één had hepatitis infectiosa doorgemaakt, maar had tijdens het onderzoek een goede leverfunctie), varieerde de verhouding van beide concentraties van ruim 1,5 tot 50. Misschien zijn deze zeer uiteenlopende getallen ten dele toe te schrijven aan een meer of minder sterke vermenging van gal met vocht in het duodenum. Zolang dit vocht zelf geen galkleurstoffen bevat, is een dergelijke verdunning voor de verdere discussie zonder belang.

Ook voor de urobilinogenen vond BRANDT bij al deze personen, op één uitzondering na, de hoogste concentratie in de B-gal.

Van de waarnemingen, bij deze 21 personen verricht (BRANDT, bl. 116) zijn enkele gegevens in tabel I overgenomen. Daarbij zijn concentraties, zoals gebruikelijk, door vierkante haken aangegeven. (Om ons onbekende redenen zijn de gegevens over patiënt 21 hier niet vermeld. Deze zijn ons door de schrijver verstrekt en ook in deze beschouwingen opgenomen. Weglating zou echter geen essentiële wijziging tot gevolg hebben. De tabel bevat enkele onbetekenende rekenfouten).

Van bela· 's verder, dat er geen bilirubine in de galblaas wordt gevormd of afgescheiden. Wel kan deze stof daar verdwijnen, althans onder pathologische omstandigheden, maar voor ons betoog is dit (gelukkig) niet van belang.

Voor iedere patiënt beschouwt BRANDT nu de beide quotiënten:

$$q_{bil} = [bil]_B / [bil]_C,\ q_{ur} = [ur]_B / [ur]_C.$$

Indien in de galblaas alleen indikking plaatsvindt, is er geen reden, waarom deze quotiënten onderling zouden verschillen, behalve dan door onvermijdelijke onnauwkeurigheden bij de chemische analyse*. Zou echter blijken, dat q_{ur} systematisch kleiner is dan q_{bil}, dan zou dit wijzen op resorptie of afbraak van urobilinognene in de galblaas.

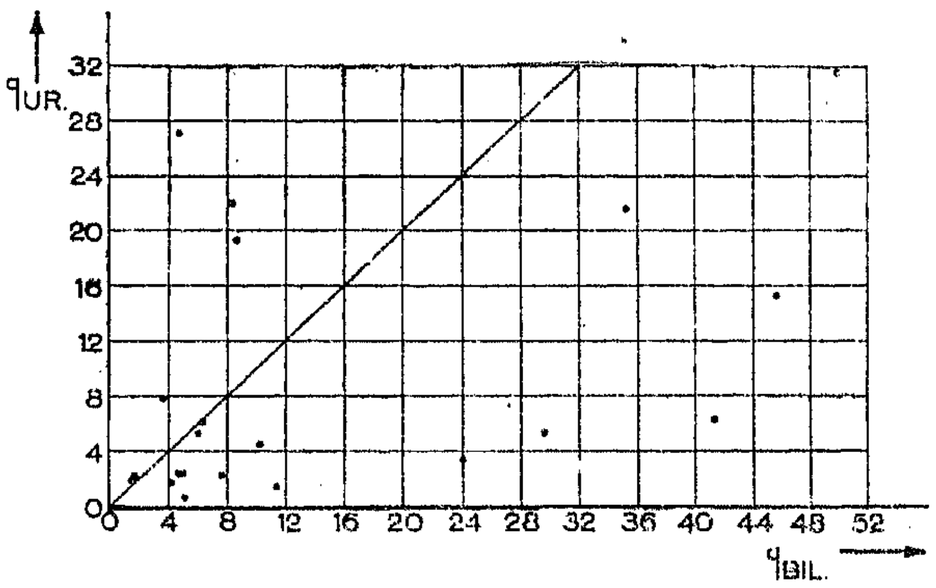

Fig. 1. Concentratiequotiënten van 21 patiënten met vrijwel normale lever en galwegen.

*Met een mogelijk storende invloed van het leverritme op de samenstelling van de gal is geen rekening gehouden. Hierover hebben wij nl. geen gegevens kunnen verkrijgen. JORES (1937) geeft een dagkromme voor het bilirubinegehalte van het bloed (maximum om ongeveer 0 uur, minimum om ongeveer 4 uur) en voor de urobilinogeenuitscheiding in de urine (maximum om ongeveer 12 uur, minimum om ongeveer 16 uur), maar spreidingen worden daarbij niet vermeld. Bij het gebruikelijke onderzoek worden de B- en C-gal op verschillende delen van een etmaal gesecerneerd. Voor de statistische aspecten van deze discussie is dit overigens van geen belang.

Omgekeerd zou door uitscheiding of aanmaak van urobilinogenen (al dan niet uit bilirubine) in de galblaas, q_{ur} systematisch groter worden dan q_{bil}. Dit zou ook het geval zijn bij afbraak of selectieve resorptie van bilirubine aldaar.

Statistisch bezien gaat het er dus om, of van de beide quotiënten het ene bij een voldoend aantal patiënten voldoende kleiner is dan het andere. Dit houdt nl. een aanwijzing in voor een systematisch verschil, en daarmee voor ten minste één der verschijnselen: resorptie, aanmaak enz. van hetzij urobilinogenen, hetzij bilirubine in de galblaas.

In figuur 1 zijn de beide concentratiequotiënten voor deze 21 patiënten grafisch tegen elkaar uitgezet (BRANDT bl. 131).

Ligt een punt in deze figuur boven de bissectrice door de oorsprong, dan is $q_{ur} > q_{bil}$, terwijl voor punten eronder het omgekeerde geldt. Geen der punten ligt precies op deze lijn, wat trouwens ook niet kon worden verwacht.

Van de 21 punten liggen er nu 6 boven en 15 onder de bissectrice. Evenals bij een eerder genoemd voorbeeld verkeren wij daarom in een situatie waarin geen voldoende aanwijzigingen voorhanden zijn voor een systematisch verschil, hoewel figuur 1 toch wel een vermoeden in die richting kan doen ontstaan.

Nu wijken een aantal punten, zowel boven als onder de bissectrice, ver van deze lijn af. Dit doet vermoeden, dat de nauwkeurigheid der quotiënten gering is. Immers, als in de galblaas alleen indikking plaatsvindt, zijn alle afwijkingen van de bissectrice aan onnauwkeurigheden van de analyse toe te schrijven. Maar ook als dit niet het geval is en er bv. ook afbraak van urobilinogenen plaatsvindt, zijn deze punten onverenigbaar met een redelijk uniforme galblaasfunctie van de onderzochte personen. De verhouding q_{ur} tot q_{bil} varieert van 0,132 tot 5,74, zodat de grootste waarde hiervan ruim 43 maal de kleinste bedraagt. Voor personen met ogenschijnlijk normale lever en galwegen — zij het ook met mogelijke andere afwijkingen — lopen deze getallen toch wel zo ver uiteen, dat men ze niet zonder meer aan functieverschillen kan toeschrijven. Er is trouwens een veel meer voor de hand liggende verklaring.

Over de nauwkeurigheid der bepalingen waren wij niet ingelicht; vermoedelijk was deze onbekend. Deze onbekendheid — een lacune in de opzet van het onderzoek — had tot gevolg, dat de statistiek niet als bewijsmiddel kon worden gehanteerd, zodat het waarnemingsmateriaal geen duidelijke conclusie toeliet.

Maar het is niet onaannemelijk — zoals nader zal blijken — dat dit anders zou zijn geweest, indien ook de nauwkeurigheid van de analyse behoorlijk was onderzocht, o.a. door proeven met afgewogen hoeveelheden urobilinogeen en door het verrichten van een voldoend aantal duplowaarnemingen. Bij ontbreken hiervan kan de statistiek altijd nog fungeren als detectiemiddel.

Redelijke speculaties over de analysenauwkeurigheid kunnen nl. het effect dat in figuur 1 in zwakke mate aanwezig schijnt te zijn, duidelijker naar voren brengen. In den regel is immers de relatieve nauwkeurigheid van een kwantitatieve bepaling (bij eenzelfde methode) geringer, wanneer het kleinere hoeveelheden van de onderzochte stof betreft. De relatieve verliezen zijn dan in het algemeen groter, de meeste aflezingen naar verhouding onnauwkeuriger.

Hieruit volgt, dat ook de verhouding van twee hoeveelheden van eenzelfde stof des te onnauwkeuriger wordt bepaald, naarmate deze hoeveelheden kleiner zijn. Omgekeerd zal van twee ongeveer gelijke, aldus verkregen quotiënten datgene met de grootste teller en noemer het betrouwbaarst zijn.

Bovendien zijn de door verliezen veroorzaakte fouten éénzijdig. Bij de zo juist beschouwde quotiënten zullen daarom, in het algemeen gesproken, zowel de teller als de noemer te klein uitvallen. Is één van deze grootheden aanmerkelijk kleiner dan de andere, dan zal de relatieve invloed van de verliezen op het kleinste getal het sterkst zijn. Is dit getal de noemer, dan zal voor het quotiënt een te grote, en soms zelfs een buitensporig grote waarde worden gevonden.

Wij vermoeden, dat een en ander zich bij de boven beschouwde grootheden voordoet, speciaal bij de quotiënten q_{ur}. Immers, de concentratie van bilirubine in de gal is groter dan die van urobilinogeen, en de voor q_{bil} benodigde aflezingen waren blijkbaar steeds in ten minste twee cijfers mogelijk; vermoedelijk is daarom q_{bil} nauwkeuriger bepaald dan q_{ur}.

Daarentegen is in het bijzonder de noemer $[ur]_C$ van q_{ur} soms zeer klein (en in niet meer dan één cijfer aangegeven), dus vermoedelijk zeer onnauwkeurig. Dat de drie kleinste van $[ur]_C$ in de (volledige) tabel alle gelijk zijn aan 0,005, ondersteunt dit vermoeden in niet geringe mate. Wellicht liet de chemische analyse een verder gaande specificatie van het begrip „zeer klein" niet toe.

Het ligt voor de hand dat ecarteren van de minst goede waarnemingen een onderzoek als het nu besprokene ten goede moet komen. De onnauwkeurige waarnemingen kunnen een eventueel aanwezig effect gemakkelijk verdoezelen.

De meest voor de hand liggende wijze van ecarteren bestaat in het weglaten van die gevallen waarbij $[ur]_C$ klein, en q_{ur} dus onbetrouwbaar is. Wel is het gewenst, dat daarbij genoeg punten uit figuur 1 overblijven om over de structuur van de resterende puntenwolk een behoorlijke indruk te kunnen krijgen. Nu is bij 13 van de 21 beschouwde patiënten $[ur]_C \leqslant 0{,}104$; bij de 8 overige varieerde deze grootheid van 0,208 tot 1,31. (De lacune tussen 0,104 en 0,208 in de waarden van $[ur]_C$ verschafte een ongedwongen criterium voor het ecarteren). De met deze laatsten corresponderende betrouwbaardere punten zijn in figuur 2 weergegeven.

Deze liggen nu alle duidelijk beneden de bissectrice, hetgeen een sterke aanwijzing voor een systematisch verschil tussen q_{bil} en q_{ur} zou inhouden, als bovenstaande procedure verantwoord kon worden geacht. Was een uitkomst van de aard van figuur 2 verkregen na ecartering op grond van nauwkeurigheidsbepalingen, dan was een ondubbelzinnig resultaat bereikt.

Bij de gevolgde methode verdwijnen evenwel juist die punten waarvoor $[ur]_C$ klein, dus q_{ur} in het algemeen groot is, dus overwegend hooggelegen punten in figuur 1. En inderdaad zijn ook de hooggelegen punten onder de bissectrice verdwenen. Toch was het niet vanzelfsprekend, dat alle overgebleven punten onder de bissectrice zouden liggen; q_{ur} kan nl. ook groot zijn door een grote teller (waarover later). Maar het trekken van een conclusie uit het verkregen resultaat alléén is zeker niet verantwoord.

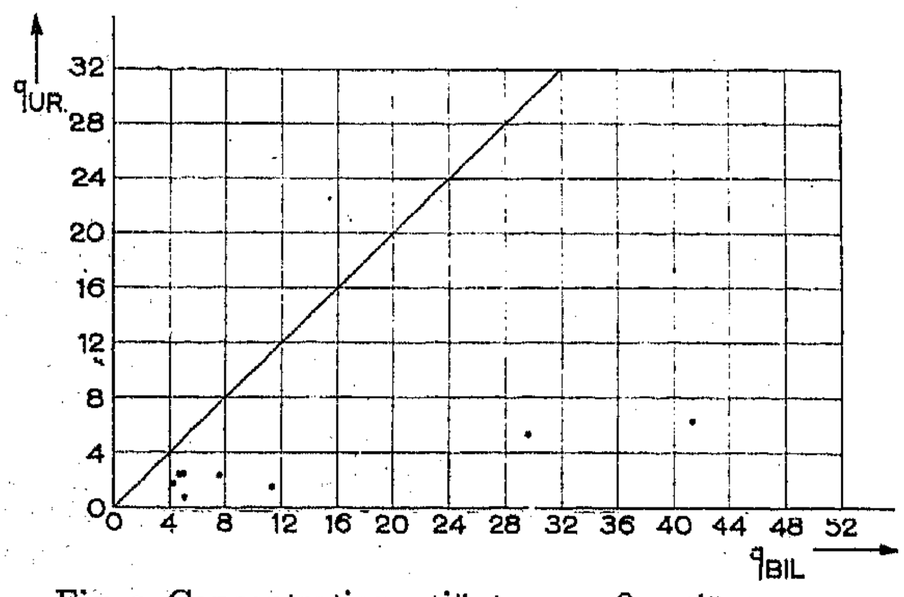

Fig. 2. Concentratiequotiënten van 8 patiënten met de grootste waarden van $[ur]_C$.

Men kan echter trachten, op grond van andere criteria te ecarteren, en zien of men aldus aanwijzingen verkrijgt die in dezelfde richting gaan. Dit blijkt inderdaad het geval. Daarbij moet men bedenken, dat iedere methode van ecarteren willekeurig is (zoals de grens voor $[ur]_C$, die aan fig. 2 ten grondslag lag, vrij arbitrair was). Er zijn immers geen voldoende gegevens over de nauwkeurigheid om tot niet-arbitraire richtlijnen te komen.

Om hierin een beter inzicht te krijgen kan men de tellers $[ur]_B$ en de noemers $[ur]_C$ tegen elkaar uitzetten, zoals in figuur 3 is gedaan (terwille van de duidelijkheid met verschillende schalen op de horizontale en verticale as).

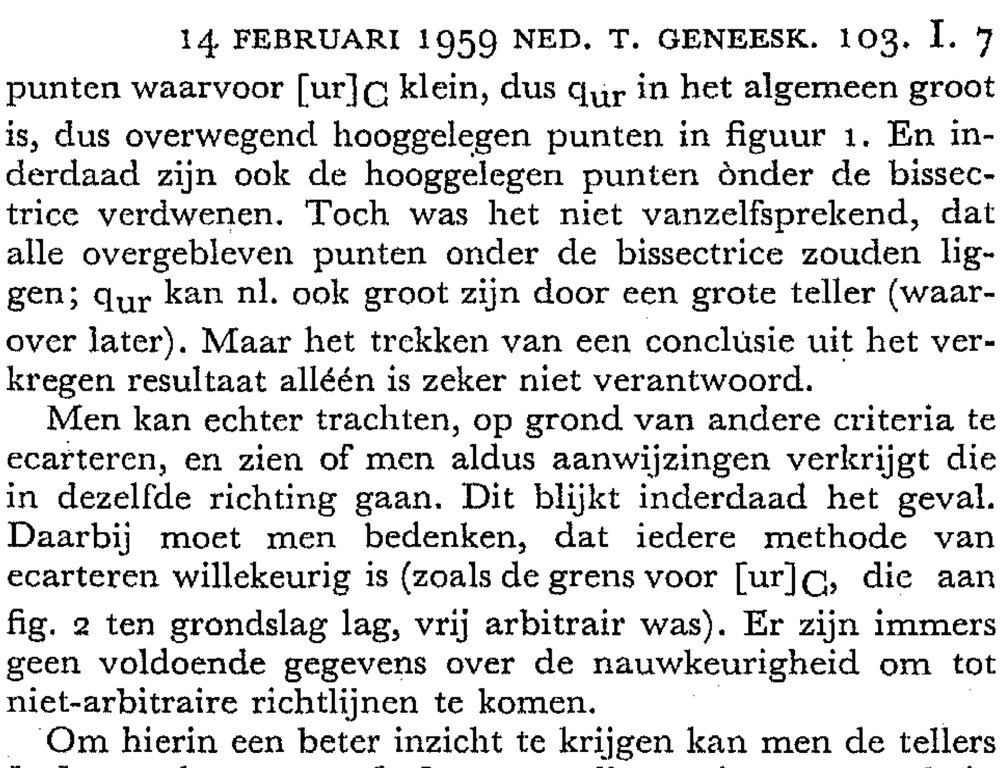

Fig. 3. Tellers en noemers van q_{ur}.

De punten die in deze figuur onder de met I aangegeven stippellijn liggen, beantwoorden dan aan de hierboven geecarteerde waarnemingen. Daarbij werd alleen rekening gehouden met de grootheden $[ur]_C$. Wij zagen evenwel, dat ook kleine waarden van $[ur]_3$ tot vermoedelijk onbetrouwbare uitkomsten leiden. In het bijzonder is er reden om een quotiënt q_{ur} te wantrouwen, indien zowel de teller als de noemer klein zijn, wat voor die punten van figuur 3 geldt, die dichtbij de oorsprong liggen. Een hierop gebaseerde ecartering doet figuur 4 ontstaan.

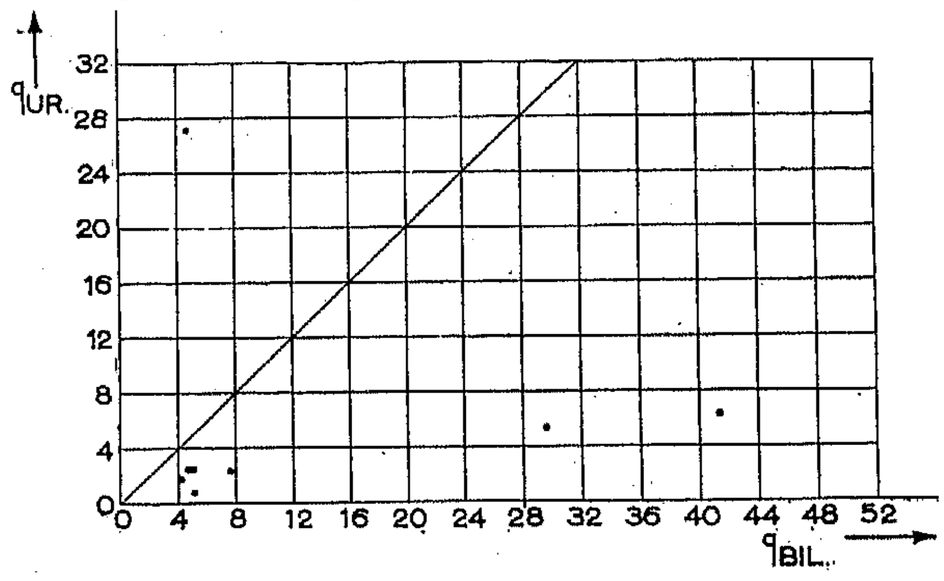

Fig. 4. Concentratiequotiënten van 8 patiënten met de grootste waarden van $[ur]_B + 2[ur]_C$, resp. van $[ur]_B$.

Ditmaal beantwoorden de weggelaten waarnemingen aan die punten in figuur 3, die onder de (alweer tamelijk willekeurige) stippellijn II liggen.

Voor het mèrendeel zijn de overgebleven punten in figuur 4 dezelfde als in figuur 2. Alle op één na liggen weer onder de bissectrice. Ook nu is de aanwijzing voor een systematisch verschil aanzienlijk duidelijker dan in figuur 1, zij het zwakker dan in figuur 2, wegens het ene punt ver boven de bissectrice.

Een derde methode komt tegemoet aan het bezwaar, dat hooggelegen punten in figuur 2 nogal opzettelijk waren geëlimineerd. Men bevordert het tegendeel, door juist die waarnemingen met de grootste tellers $[ur]_B$ uit te zoeken. Evenals bij figuur 2 zullen ook ditmaal de in het algemeen betrouwbaarder punten uit figuur 1 overblijven. Kiest men weer 8 punten — nu de 8 met de grootste tellers $[ur]_B$ — dan zijn dit (min of meer toevallig) precies dezelfde als in figuur 4 aangegeven, zodat ook nu weer 7 van de 8 punten onder de bissectrice liggen. Met name liggen de punten, beantwoordend aan de 3 hoogste waarden van de teller, eronder. Onze poging om bij voorkeur punten boven de bissectrice over te houden, heeft dus een averechtse uitwerking.

Alle „detectie"-argumenten wijzen dus in dezelfde richting: dat q_{ur} systematisch kleiner is dan q_{bil}, dus dat er afbraak of resorptie van urobilinogenen plaatsvindt in de galblaas.

Reeds eerder werd uiteengezet, dat dit niet als een bewijs mag worden opgevat. Men kan slechts de tweeledige conclusie trekken, dat er misschien wèl een duidelijke aanwijzing zou zijn verkregen, indien er meer aandacht aan de nauwkeurigheid was geschonken, en dat een verder onderzoek — indien deze kwestie medisch voldoende belangrijk wordt geacht — aanbevelenswaardig is.

Zoals vaak bij lezingen over de statistiek — een nog vrij jonge wetenschap, die nog niet voldoende gemeengoed is geworden — is deze lezing min of meer geworden tot een aansporing tot het gebruik ervan en een waarschuwing tegen verkeerd gebruik. Men leze in dit verband ook de voortreffelijke voordracht „Design for decision in het klinische experiment", die Prof. Dr. D. K. DE JONGH op 11 december 1956 gehouden heeft voor de Medisch Biologische sectie van de Vereniging voor Statistiek.

Laat mij daaraan dan een aansporing toevoegen: houd uw proeven zo eenvoudig als maar enigszins mogelijk is. Tracht niet in één onderzoek te veel vragen tegelijk te beantwoorden. Dit leidt er maar al te vaak toe, dat géén der gestelde vragen goed kan worden beantwoord. Vermeld bovendien in uw publikaties, in hoeverre U uw oorspronkelijke doelstellingen en voorgenomen methodes hebt gehandhaafd en in hoeverre deze gewijzigd zijn tijdens uw onderzoek. Geef de lezer inzicht in de wijze waarop U de statistiek hebt gebruikt: als bewijsmiddel of als detectiemiddel. De overtuigingskracht zal er in het eerste geval aan winnen, en de waarde in verband met verder onderzoek in het tweede. Alleen als ook deze details van het onderzoek vermeld zijn, is een juiste beoordeling van de bereikte resultaten mogelijk. En dat is voor een publikatie een eerste vereiste.

En tenslotte nog enkele opmerkingen over het gebruik van placebo's en controlegroepen. Tegen het gebruik daarvan rijzen niet zelden bezwaren van medisch-ethische aard, die buiten de competentie van de statisticus vallen. Deze heeft echter toch de plicht, steeds weer erop te wijzen, dat een proef zonder controlegroep in de meeste gevallen niet tot een verantwoorde conclusie kan leiden en dus vergeefse moeite is. Niet alleen dat, maar indien men met gegevens uit het verleden of gegevens van anderen als controle werkt, vloeien daaruit vaak gevaren voort voor (toekomstige) patiënten van de onderzoeker. Daar men dan nl. zijn conclusie baseert op de vergelijking van de eigen proef-uitkomsten met de bij andere, niet zorgvuldig equivalent gemaakte groepen van patiënten verkregen resultaten, bestaat er een vergrote kans op de beide mogelijke fouten: een minder goed middel als beter aanprijzen en een beter middel over het hoofd zien. De grootte van de kans op schijneffecten en verdoezelingen is in zulke gevallen niet gelijk aan de (bekende) onbetrouwbaarheid van de gebruikte statistische analysemethode, doch te enen male onbekend. Het is een verkeerd gebruik van de statistiek, indien men een andere suggestie wekt. De rechtstreekse hulp, die de statisticus kan geven, bestaat uit een beperking tot zo klein mogelijke controlegroepen door het aangeven van de meest doeltreffende proefopzet.

Al deze beschouwingen, die hier voor problemen van uiterst eenvoudige aard zijn gegeven, gelden analoog — doch in nog sterkere mate — voor problemen van ingewikkelder en technisch moeilijker aard.

Samenvattend luidt de moraal van dit verhaal: gebruik de statistiek bij het opzetten en uitwerken van uw proeven en bij het verzamelen en analyseren van uw gegevens, maar doe dit zorgvuldig; zowel het verwaarlozen als het verkeerd gebruiken van de statistiek — die bij goed gebruik een aanzienlijke vergroting van uw onderscheidingsvermogen geeft — brengen nodeloze gevaren mee voor uw patiënten.

Literatuur: BRANDT, K. H. (1957) Proefschrift Utrecht. — DE JONGH, D. K. (1956) Mededelingen van de Medisch Biologische Sectie van V. v. S., 1957 nr. 2. — JORES (1937) *Tab. Biol.* **14**, 95.

Discussie:

De heer JONGKEES merkt op, dat vergelijking van eigen resultaten omtrent geneesmiddelen, met gegevens uit de literatuur, toch wel tot betrouwbare conclusies aanleiding kan geven, als de verschillen zeer groot zijn. Hij vraagt, hoe de medicus, als leek-statisticus, weet, wanneer een uitkomst onmiskenbaar is en wanneer niet.

De heer HEMELRIJK antwoordt: Het vergelijken van eigen resultaten met gegevens uit de literatuur kan altijd als nuttig worden beschouwd. Grote verschillen kunnen echter ook ontstaan door niet relevante verschillen in de proefomstandigheden; voorzichtigheid bij de interpretatie van gevonden verschillen is daarom altijd geboden. Een bevestigend onderzoek dient, indien mogelijk, steeds als gewenst te worden beschouwd.

Op de tweede vraag, wanneer een uitkomst onmiskenbaar is en wanneer niet, kan geen algemeen geldend antwoord worden gegeven. Voor iemand die in het geheel geen ervaring heeft met statistische verwerking van waarnemingsmateriaal, is het op het oog beoordelen van verschillen steeds een gevaarlijke bezigheid. De vaardigheid in de beoordeling van de betekenis van verschillen hangt nauw samen met routine in deze bezigheid.

De heer GOSLINGS merkt op: U geeft als schema voor een vergelijkend onderzoek van de werking van twee geneesmiddelen X en Y op een ziekte die gedurende langere tijd een constant beloop heeft, wat de activiteit betreft, de raad om een groep te behandelen in de volgorde Y–X en een andere groep X–Y. Maar de reactie van de ziekte op het geneesmiddel kan zelfs bij constante, gelijke activiteit van het

ziekteproces ook nog invloed ondergaan, doordat de reactie op het middel in een zeker tijdsverloop verandert, bv. vermindert. Is het daarom niet beter, een groep te nemen in de volgorde X–Y–X en de tweede groep Y–X–Y? Slechts als X in het eerste geval en Y in het tweede geval beide keren nagenoeg gelijk blijven werken, kan een vergelijking met het andere middel wellicht beter geschieden. Of is dit toch niet nodig?

De heer HEMELRIJK antwoordt: Bij het behandelde voorbeeld is afgezien van complicaties van de in deze vraag geschetste aard. Indien het vermoeden bestaat, dat deze wel eens aanwezig zouden kunnen zijn, is het inderdaad gewenst, de proefopzet op zodanige wijze te veranderen, dat deze mogelijke storende invloeden ofwel geëlimineerd, ofwel in de analyse betrokken worden. De hier voorgestelde volgorde der proeven kan daarvoor inderdaad dienen.

Mevrouw RUYS vraagt: Is het wel geoorloofd, de werkzaamheid van middelen te beoordelen met behulp van gegevens uit de literatuur? Men heeft bij de behandeling van difterie met serum de overtuiging gekregen, dat dit een bijzonder werkzaam middel is. In de jaren dertig heeft FRIEDBERGER gemeend te kunnen aantonen, dat gewoon paardeserum even goed werkte als antidifterieserum. Is het dan niet mogelijk, dat de ernst van de ziekte minder is geworden en de serumtoediening geen effect heeft? Men heeft bij roodvonk in dezelfde periode ook een belangrijke vermindering in de ernst van de ziekte gezien. Men moet zich dus afvragen, of men niet steeds bij difterie zonder noodzaak antiserum geeft. Niemand heeft echter nu de moed, dit op grote schaal na te gaan.

De heer HEMELRIJK antwoordt: Het door mevrouw RUYS genoemde geval is een voorbeeld van de hierboven (in het antwoord op de vragen van de heer JONGKEES) gesignaleerde gevaren. Indien het antiserum niet schadelijk werkt, is het eventueel overbodig toedienen ervan uiteraard geen zeer groot bezwaar. Niettemin zou het wenselijk zijn, zekerheid te verkrijgen over de werking ervan. Dit is nu echter onmogelijk, althans zonder nieuwe proefnemingen, waartegen echter bezwaren van andere aard bestaan.

De heer VAN DEN BROEK vraagt, of het, gezien het feit dat nooit twee mensen aan elkaar gelijk zijn (genetisch, anatomisch, fysiologisch, psychisch), wel toelaatbaar is, statistische methoden op zulk een ongelijkwaardig materiaal toe te passen.

De heer HEMELRIJK antwoordt: De inhomogeniteit van een groep mensen die men aan een onderzoek onderwerpt, is één van de factoren, waarmede bij een goede proefopzet, volledig rekening wordt gehouden. De bovenbesproken procedure van loting bij het bepalen van de behandeling dient juist om de storende invloed van de tussen individuen bestaande verschillen op de betrouwbaarheid van de statistische toetsing op te heffen. Niet alleen is daarom toepassing van statistische methodes toelaatbaar, men kan zelfs stellen, dat het niet toepassen daarvan ontoelaatbaar is. Uiteraard is het juist in gevallen waarin deze verschillen tussen de individuen groot zijn, gewenst, de proef zo verantwoord mogelijk op te zetten, daar anders ook statistische verwerking niet tot betrouwbare conclusies kan leiden.

Niettemin is het duidelijk, dat met homogeen waarnemingsmateriaal betere resultaten zijn te verkrijgen dan met inhomogeen materiaal. Bij een goede proefopzet zullen in beide gevallen even vaak (of zo men wil even zelden) schijneffecten voorkomen, doch anderzijds zullen werkelijk bestaande effecten gemakkelijker — en daardoor ook vaker — worden ontdekt naarmate het proefmateriaal homogener van karakter is. Men drukt dit uit door te zeggen, dat het onderscheidingsvermogen van de toegepaste statistische toetsen groter is, indien het waarnemingsmateriaal homogeen is. Bij proeven op dieren zal men er dan ook naar streven, gebruik te maken van dieren die genetisch en ook in andere opzichten zo weinig mogelijk van elkaar verschillen. Indien de mogelijkheid van selectie van homogene groepen proefindividuen niet aanwezig is, kan het noteren van gegevens over de aard der tussen de individuen bestaande verschillen soms de mogelijkheid openen tot een ingewikkelder, maar meer onderscheidende statistische analyse, die berust op splitsing van het materiaal in homogene groepen. Binnen deze groepen moet dan weer aan strenge voorwaarden zijn voldaan. Dit alles dient dus te worden overwogen, voor men aan de proef begint, daar er anders niet goed verwerkbare waarnemingen kunnen worden verkregen.

M. M. HILFMAN, *secretaris*