8655

S- WERK ARCHIEF

STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

LEIDING: PROF. DR D. VAN DANTZIG
ADVISEUR VOOR STATISTISCHE CONSULTATIE: PROF. DR J. HEMELRIJK

Rapport S 225 (V 17)

Statistische proefopzetten

door

Prof. Dr. J. Hemelrijk

November 1957

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 er zorgvuldig voor moet zorgen, dat men in bepaalde opzichten niet weet wat men doet.

Deze situatie is op aanschouwelijke wijze te illustreren aan de hand 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, b.v. 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 het zelfde getal voor nemen - althans als de ander eerst moet raden. Op deze wijze bereikt men, dat de winstkansen voor beide spelers gelijk worden, wàt 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 gevoegelijk buiten beschouwing blijven.

Dit soort voor de gek houderij heeft op zich zelf niets met medische experimenten te maken, maar toch is het één van de essentiële bijdragen van de statistiek tot het perfectionneren van experimenten van velerlei - ook van medische - aard. Men kan er, zoals in het bovengenoemde geval, ingewikkelde en onoverzichtelijke situaties in één slag mee vereenvoudigen tot theoretisch en statistisch gemakkelijk te beheersen problemen en men bereikt daarmee twee belangrijke resultaten. In situaties, waar onoverzichtelijke en niet geheel (of geheel niet) bekende invloeden op kunnen treden, kan men nl. enerzijds het optreden van "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 ène (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 verstrekt is. Vervolgens worden de door de proefpersoon verstrekte antwoorden met de werkelijkheid vergeleken om te beslissen of zijn bewering juist is of niet. De statistische techniek, die daarbij wordt gebruikt, behoeft niet hier te worden uitgelegd, daar het hier alleen om de proefopzet gaat.

Welke voorzorgen moeten nu genomen worden, om te zorgen, dat het experiment zich voor een goede statistische uitwerking leent? Ten eerste uiteraard de reeds genoemde, dat de proefpersoon niet weet welk middel hij op een bepaalde dag verstrekt krijgt. Ook indien hij hiervan niet op de hoogte is, kunnen zich echter nog complicaties voordoen. Immers de arts, die bepaalt in welke volgorde de middelen A en B worden verstrekt, moet deze volgorde op de één of andere wijze bepalen en indien hij zich daarbij door zijn eigen intuïtie of willekeur laat leiden zal er in deze volgorde zonder twijfel het één of andere patroon verschijnen. Zo zal b.v. zeker de neiging aanwezig zijn om te vermijden, dat gedurende een relatief groot aantal dagen achtereen hetzelfde middel wordt gegeven. De proefpersoon zal zijnerzijds, daar hij weet, dat hij nu eens het éne, dan weer het andere middel krijgt, bij het vaststellen van zijn antwoorden een soortgelijke neiging vertonen. Indien de psychologie van de arts en van de proefpersoon enigzins 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 zich zelf reëel genoeg, maar het heeft met de te onderzoeken vraag niets te maken. 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 beïnvloeding van 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 aan de wat vergezochte mogelijkheid van telepathie te denken, toch zijn wijze van overhandigen daarvan misschien beïnvloeden, waardoor opnieuw moeilijkheden van dezelfde aard als boven geschetst kunnen ontstaan. (Tussen twee haakjes zij opgemerkt, 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 schijn-effect, een ander voorbeeld moge duidelijk maken, dat even goed een reëel effect verdoezeld kan worden. Veronderstel, dat een arts twee geneesmiddelen of -methoden tot zijn beschikking heeft voor een niet gevaarlijke ziekte, maar die zich toch in duidelijk verschillende graden van ernst voor kan doen. Het éne 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 ernstige patiënten. De ernstigere 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 beide 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, omdat eventuele verbeteringen bij patiënten, die de ziekte in heftige mate vertonen wellicht opvallender zijn dan bij minder ernstige. 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 zoveel hardnekkiger is dan bij de minder ernstige, dat middel Y in een voordelige positie verkeert, omdat juist de gemakkelijker te genezen patiënten dit krijgen. Het is in de 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, zijn diagnose omtrent eventuele vooruitgang der patiënten zal kunnen beïnvloeden. Daardoor komt het middel, waarin hij het meeste vertrouwen stelt, in het voordeel tegenover het andere. Ook dit dient vermeden te worden.

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 ernstige patiënten nog niet met middel Y wil behandelen, dan blijven deze (voorlopig) buiten de proef. De arts bepaalt dus eerst of een patiënt mee zal doen - 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 dragende zelf voorlopig niet te weten 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 te voren 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 opgeborgen wordt, 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, al lang vergeten welk middel in het uitgereikte doosje zit, zodat enerzijds zijn keuze van het middel inderdaad toevallig is en anderzijds zijn diagnose niet meer door de - nu immers niet aanwezige - kennis (X of Y) beïnvloed kan worden.

Natuurlijk kan men bij dergelijke proeven veel gemak ondervinden van de hulp van een assistent(e), vooral in moeilijker te organiseren gevallen, daar deze wel 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 dit 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 evident is, vaak als vrijwel onmogelijk beschouwd worden.

Om dit te demonstreren beschouwen wij een getallenvoorbeeld. Wij veronderstellen daartoe, dat de twee middelen X en Y achter elkaar bij dezelfde patiënt beproefd kunnen worden en dat de volgorde, waarin dit geschiedt, van geen invloed is op de uitwerking en de diagnose (wellicht wat vergaande veronderstellingen, maar die voor dit voorbeeld gewenst zijn). Bij iedere patiënt kan nu nagegaan worden op welk der twee middelen hij het gunstigst reageert (zonder dat uit die waarneming de conclusie getrokken dient te worden, dat dit verschil zich, bij iedere patiënt afzonderlijk, steeds op deze wijze zal voordoen; van twee waarnemingen zal, indien zij - om welke reden dan ook - ongelijk zijn, altijd één van beide de grootste zijn). Indien nu een arts bij 20 patiënten verschillende reacties op de middelen gevonden heeft en bij 14 daarvan was de uitwerking van X gunstiger, 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 handelswijze 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% van zijn toepassingen tot de verkeerde conclusie leidt, dat één der middelen beter is dan het andere. Is dit werkelijk het geval, is b.v. Y beter dan X, dan is de kans dat men toch X als het beste aan zal wijzen 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 zeer groot is, zodat 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 ontdekt te worden (in statistische jargon: het <u>onderscheidingsvermogen</u> van een proef neemt sterk toe met het aantal waarnemingen). Hij breidt de reeks b.v. uit tot 100 proeven. De statisticus zal hem nu vertellen, dat hij - indien hij zijn kans op een foute conclusie tot hoogstens 1% wil beperken - slechts dan tot een verschil tussen de middelen mag concluderen als één van beide middelen in minstens 64 van deze 100 gevallen een betere uitwerking heeft dan het andere. Tot zoverre is alles goed. Maar indien de arts er niet voor gezorgd heeft onwetend te zijn omtrent de gebruikte middelen en zich, zij het in geringe mate, bij zijn diagnose zou laten beïnvloeden door zijn persoonlijke verwachting omtrent de deugdelijkheid der middelen, zodat hij b.v. in 5 twijfelgevallen (op 100 proeven) zijn oordeel ten onrechte ten gunste zou doen uitvallen van dat middel, waarvan hij het meest verwacht, dan wordt de kans, dat dit middel ten onrechte als het beste aangewezen zal worden alweer tot ongeveer 5% verhoogd. Laat hij zich in 7 twijfelgevallen door zijn mening verleiden tot een gunstige uitspraak, dan wordt deze kans zelfs 10%. Maakt men dus enerzijds de proef aanzienlijk onderscheidender door het aantal waarnemingen te vergroten, anderzijds wordt hij daardoor ook veel gevoeliger voor fouten (ook kleine) in de proefopzet en de beoordeling der afzonderlijke waarnemingen. Hetgeen onderstreept, dat alleen een zorgvuldige proefopzet zowel betrouwbaar als onderscheidend kan zijn. Het is verspilling van arbeid en kennis te werken met niet zorgvuldig opgezette kleine proeven, daar deze geen objectieve beoordeling toestaan; het is nog meer verspilling grote 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.

Het is uiteraard niet nodig, dat de dokter zo ver gaat, dat hij dobbelstenen hanteert bij het behandelen van zijn patiënten. De patiënten zouden de indruk kunnen krijgen, dat er met hen gesold wordt; ook al zou deze indruk misplaatst zijn, toch is het wenselijk, dat hij in het geheel niet gewekt wordt. Er zijn echter uitgebreide tabellen beschikbaar van z.g. <u>aselecte getallen</u> (Engels: "random numbers"), waarvan men zich kan bedienen voor alle doeleinden van loting. Deze zou men, als dat nodig is, zelfs in aanwezigheid van de patiënt kunnen raadplegen zonder zich zijn toorn of wanvertrouwen op de hals te halen. Bij dit alles vergete men niet, dat het juist in het belang van de patiënten is, dat betere geneesmiddelen zo snel mogelijk ontdekt worden en niet door onvolmaakte proefopzetten over het hoofd worden gezien.

Naast het boven nogal uitvoerig besproken elementaire, maar fundamentele, doch niet vaak consequent toegepaste, lotingsprincipe heeft de statistiek nog vele andere bijdragen tot de goede proefopzet te leveren. Het zou te ver voeren daarop in te gaan, wij noemen slechts enkele middelen bij name: de sequente analyse, die erop uit is het aantal waarnemingen dat nodig is om tot een conclusie te komen zo veel mogelijk te beperken; het verdelen van de patiënten in zoveel-mogelijk homogene 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 gesuggereerd wordt, een statisticus raadpleegt, doet men er goed aan dit te doen vóór men aan een experiment begint. Hierop kan nauwelijks genoeg de nadruk gelegd worden. 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 eerder aan Haarlemmer Olie denken: het helpt niet werkelijk. Talloos zijn de onderzoekingen, die gestrand zijn op een onzorgvuldige opzet en waarbij de statisticus alleen nog maar de fouten in deze opzet en hun fatale gevolgen aan kan wijzen, zonder de experimentator uit de daaruit voortvloeiende moeilijkheden te kunnen helpen.

Wendt men zich echter van tevoren tot een statisticus, dan zal men vrijwel steeds al direct voor de vraag gesteld worden: "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 boven betoogd, dat de experimentator in sommige opzichten zorgvuldig moet vermijden te weten wat hij doet, hij dient deze onwetendheid ook zorgvuldig te beperken tot die punten, die de statisticus hem aangeeft; overal elders werkt onwetendheid negatief en dat geldt uiteraard niet alleen voor de arts, maar voor iedere experimentator.

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, beschouwd te worden als essentiële onderdelen van het onderzoek als geheel. Te vaak nog worden deze aspecten (ook bij de begroting in geld en tijd) niet als onderdelen van de proef beschouwd, hetgeen er niet zelden toe leidt, dat wel waarnemingsmateriaal verzameld wordt, maar dat dit achteraf niet grondig kan worden geanalyseerd, ofwel omdat tijd en geld, of zelfs een sta tisticus, ontbreken, ofwel omdat het materiaal achteraf niet geschikt blijkt te zijn om het gestelde doel mee te bereiken, de gestelde vraag te beantwoorden. Dit lijkt misschien absurd, maar dat het voorkomt zal ieder onderzoeker bekend zijn. Dit kan zelfs zover gaan, dat de statistiek (N.B!) er een slechte naam van krijgt. Heeft men echter van tevoren met de statistische aspecten voldoende en op de juiste wijze rekening gehouden, dan kan men vaak in verrassende korte tijd resultaten bereiken, waarvoor men anders een veel langere tijd nodig zou hebben, of die zelfs vrijwel onbereikbaar zouden zijn gebleven. Waarmee niet gezegd wil zijn, dat een gesprek met een statisticus vooraf niet tot teleurstellingen zou kunnen leiden. Zo kan het b.v. blijken, dat het doel, dat men zich stelt langs de voorgestelde weg niet bereikbaar geacht moet worden, b.v. omdat de statistiek (nog) niet beschikt over een geschikte analysemethode voor het gestelde probleem, 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 bereikt zou hebben. Wat de tijdsduur van de statistische verwerking van uitgebreid waarnemingsmateriaal betreft, deze is in het algemeen van dezelfde orde van grootte 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 hierop vaak aanzienlijk worden bespaard.

Men kan de statistiek bij medische en physiologische (en ook bij andere) proefnemingen op twee wijzen gebruiken: als <u>detectiemiddel</u> en als bewijsmiddel. Het uitermate belangrijke onderscheid tussen deze twee gebruiksmethoden wordt nog wel eens uit het oog verloren. Zoekt men b.v. naar nieuwe geneesmiddelen voor een bepaalde ziekte en probeert men vele mogelijk werkzame middelen, daarbij iedere proef goed opzettende en bij de verwerking statistische toetsen toepassende, dan kan men nog lelijk bedrogen uitkomen als men deze onderscheiding niet duidelijk maakt. 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. Practische zekerheid wordt alleen verkregen door vaak herhaalde bevestiging, absolute zekerheid nooit. Daarom is dupliceren van door anderen verrichte proeven zo nuttig, 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 uitgedrukt kan worden, zodat men weet, hoe vaak men een foute conclusie trekt. Daardoor is men dan anderzijds weer wat scheutiger met die fouten: veelal wordt een kans van 1/20 op een foute conclusie toegelaten. Dit betekent, dat bij het onderzoeken van 100 volkomen onwerkzame middelen, 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 en op de markt brengt. Als men nu maar kon aanwijzen, niet alleen hoeveel fouten er in een reeks statistische analyses ongeveer gemaakt worden, maar ook welke conclusies fout zijn, dan was men van deze moeilijkheid verlost. En dit kan als men van de statistiek 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 analysemethoden: "Zie maar eens wat er uit te halen is") overgaat op de statistiek als bewijsmiddel. Men kan een statistisch verwerkte proef alleen dan als een overtuigende aanwijzing voor een bepaald verschijnsel, b.v. de werkzaamheid van een geneesmiddel, beschouwen, indien aan de volgende eisen voldaan is.

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

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

C. De statistische analysemethoden worden van tevoren gekozen en van deze keuze wordt niet afgeweken.

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

E. De mogelijkheid van schijneffecten (en verdoezelingen) wordt van tevoren te niet gedaan door in die stadia van de proeven, waarin onbekende of onberekenbare oorzaken systematische invloed zouden kunnen uitoefenen, statistisch 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 veel afschrikwekkendere lijst uitgebreid kunnen worden. Men houdt er zich dan ook vrijwel nooit volledig aan. De verleiding is te groot om, als zich tijdens de proef een onverwacht effect (schijnt) voor te doen, de proef onderweg te wijzigen of de statistische analyse uit te breiden tot dit nieuwe verschijnsel. In vele gevallen zou het ook onjuist zijn niet aan deze verleiding toe te geven. Maar men beseffe wel, dat men van het bewijsmiddel teruggaat naar het 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 b.v. soms vrij lange rijen van gelijke getallen achter elkaar.) Toetsing van zulke bijzondere constellaties en pogingen tot verklaring op medische gronden zijn zeker zinvol, indien men er maar voldoende van doordrongen is, dat een nadere bevestiging door een experiment, dat wel aan de boven geformuleerde eisen voldoet, onontbeerlijk 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 gebruik ervan en een waarschuwing tegen verkeerd gebruik.*) Laat mij daar dan nog een

*)Men leze in dit verband ook de voortreffelijke voordracht, die Prof. Dr.D.K.de JONGH op 11 december 1956 gehouden heeft voor de Medische Biologische sectie van de Vereniging voor Statistiek ("Design for decision in het clinische experiment", Mededelingen van de Medisch Biologische Sectie van de V.v.S.,1957 nr.2).

aansporing aan toevoegen: houdt Uw experimenten zo eenvoudig als maar enigzins 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. Vermeldt bovendien in Uw publicaties in hoeverre U Uw oorspronkelijke doelstellingen en voorgenomen methoden hebt gehandhaafd en in hoeverre deze gewijzigd zijn tijdens Uw onderzoek. Geeft 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 waardering van de bereikte resultaten mogelijk. En dat is voor een publicatie een eerste vereiste.

En ten slotte nog enkele opmerkingen over het gebruik van placebos en controlegroepen. Tegen het gebruik daarvan rijzen niet zelden bezwaren van medisch-ethische aard, die buiten de competentie van de statisticus vallen. Deze heeft de plicht er steeds weer op 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 een vergrote kans op beide mogelijke fouten: de 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 ten ene male onbekend. Het is verkeerd gebruik van de statistiek indien men een andere suggestie wekt. De medisch-ethische consequenties hiervan vallen opnieuw buiten de competentie van de statisticus, maar deze dient er steeds op te wijzen, dat het gebruik van goede controles in wezen onontbeerlijk is. De directe 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 ingewikkeldere en technisch moeilijkere aard. Samenvattend luidt de moraal van dit verhaal: gebruikt 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 juist gebruik een aanzienlijke vergroting van Uw onderscheidingsvermogen geeft - brengt nodeloze gevaren mee voor Uw patiënten.