KONINKL. NEDERL. AKADEMIE VAN WETENSCHAPPEN – AMSTERDAM
Overgenomen uit Verslag Dl. LXI, No. 1, 1952

VOORDRACHT

WISKUNDE

WISKUNDIGE CONSULTATIE TEN BEHOEVE VAN MEDISCH, BIOLOGISCH EN ANDER ONDERZOEK

DOOR

D. VAN DANTZIG

In de beginperiode van statistisch onderzoek werd meestal stilzwijgend ondersteld, dat een statistische collectie onder gelijkblijvende omstandigheden door één enkel getal (b.v. een gemiddelde) gekarakteriseerd kon worden (b.v. JOHN GRAUNT, 1662, die een schatting van het sterftequotient te Londen van 1 : 32 of 1 : 30 maakte, welke door WILLIAM PETTY (1687) gebruikt werd om de grootte der bevolking van Parijs, Amsterdam, Rome e.d. uit sterftewaarnemingen te schatten). [1])

In de tweede fase ($\pm$ 1800) werd men zich van spreidingsverschijnselen bewust. Men onderstelde, dat elke statistische verdeling door gemiddelde en spreiding gekarakteriseerd was en met de zgn. *normale verdeling* overeenstemde. Deze is ten onrechte naar GAUSS (1809) en soms ook naar LAPLACE (1778) genoemd, daar reeds DE MOIVRE (1733) haar kende. Zij werd vooral door A. QUETELET (omstreeks 1830)en FRANCIS GALTON (omstreeks 1870) toegepast.

In de derde fase ($\pm$ 1900) werd men zich van scheefheid en andere afwijkingen der verdelingen van de normale bewust en trachtte men daarmede door het invoeren van meer parameters rekening te houden (KARL PEARSON, W. C. KAPTEYN, M. VAN UVEN, C. V. L. CHARLIER e.a.).

Kenmerkend voor de tweede en de derde fase zijn methoden, strekkende tot het zo goed mogelijk schatten van de „ware waarden" der in de verdelingen voorkomende parameters (b.v. methode der kleinste quadraten van A. M. LEGENDRE en C. F. GAUSS; aannemelijkste („maximum likelihood")-schattingen volgens R. A. FISHER.

De (huidige) vierde fase is gekenmerkt door een kritische houding zowel ten aanzien van de grondslagen der waarschijnlijkheidsrekening (begin: R. VON MISES, 1919), als van de uit statistisch waarnemingsmateriaal te trekken conclusies. De voor de vorige perioden karakteristieke *schattings*problemen worden meer en meer vervangen door een exacte *toetsings*theorie (J. NEYMAN, E. S. PEARSON, e.a.) van uit statistisch materiaal te trekken conclusies, waardoor o.a. ook beperking tot korte

[1]) D. VAN DANTZIG, Enkele historische betrekkingen tussen mathematische en verzamelende statistiek, Statistica **4**, 233–247 (1950) en de daar vermelde literatuur.

waarnemingsreeksen mogelijk wordt (R. A. FISHER). In het bijzonder worden meer en meer de vóór dien gebruikelijke onderstellingen omtrent een bijzondere vorm van de verdelingsfunctie (b.v. normaliteit) vermeden, daar zij veelal niet toetsbaar en, indien wel, vaak niet vervuld zijn. Deze algemene methoden worden *parametervrij* of *rang-invariant* genoemd, daar zij overwegend slechts gebruik maken van de rangschikking naar grootte der waargenomen grootheden, niet van hun getallenwaarde (A. KOLMOGOROFF, 1934; M. G. KENDALL, 1938 en vele anderen). In de laatste jaren hebben zich de toetsingstheorie en de daarmede samenhangende theorie van betrouwbaarheidsintervallen (Engels: confidence-intervals) en tolerantie-grenzen nog verwijd, enerzijds tot een algemene theorie van rationeel doelgericht handelen [1]) (A. WALD's „Statistical decision functions"), anderzijds tot een wiskundige analyse van het begrip „informatie" (N. WIENER, C. SHANNON), die vooral ook voor de logika belangrijk belooft te worden. Deze laatstgenoemde theorieën zullen heden echter niet ter sprake komen.

De grondgedachte van de moderne statistische theorie is de volgende. Men kiest vooraf willekeurig een getal α, de *onbetrouwbaarheidsdrempel* (Engels: level of significance), waarvoor conventionaliter vaak 0,05 (5 %) wordt genomen en men laat uitsluitend conclusies toe, die *behoudens een waarschijnlijkheid* α (afkorting: spr α, d.w.z. salva probabilitate α) gelden. Uitdrukkelijk toegestaan wordt dus dat gemiddeld (b.v.) ten hoogste 1 op 20 conclusies fout mag zijn. Natuurlijk kan desgewenst voor α een kleinere waarde gekozen worden. Ook kan (volgens A. WALD's theorie der decisiefuncties) met de *ernst* van verschillende fouten rekening worden gehouden. Belangrijker echter is, dat iedere conclusie geldt behoudens een bepaalde waarschijnlijkheid, die weliswaar hoogstens 0,05 is, maar in werkelijkheid meestal zéér veel kleiner, zodat de werkelijke onbetrouwbaarheid, de zgn. *overschrijdingskans* (dat is de kleinste onbetrouwbaarheidsdrempel, waarbij deze conclusie nog getrokken zou zijn) doorgaans ver beneden de oorspronkelijk gekozen drempel blijft. Bovendien kan door doelbewuste combinatie van methoden en dienovereenkomstige opzet der experimenten de onbetrouwbaarheid der eindconclusies willekeurig klein worden gemaakt. Hoofdzaak is de beperking tot uitspraken, die nauwkeurig getoetst (anglicistisch: getest) kunnen worden.

In het bijzonder wordt deze gedachte toegepast bij het *toetsen van hypothesen*. Men kan namelijk waarschijnlijkheden alleen berekenen op grond van een bepaald *waarschijnlijkheidstheoretisch model*, dat bepaalde onderstellingen omtrent de te gebruiken waarschijnlijkheidsverdelingen inhoudt. Het feit dat dit model, zoals bij alle op de ervaring toegepaste wiskunde, de werkelijkheid slechts op vereenvoudigde wijze weergeeft, kan natuurlijk, als de afwijkingen tussen model en werkelijkheid te groot

---

[1]) Zie mijn voor de Statistische Dag 1952 te houden voordracht over „De natuur als tegenspeler".

worden, de bovengenoemde onbetrouwbaarheidsdrempel min of meer illusoir maken. Een gedeelte van dit model (dat dan gewoonlijk de „nulhypothese" of de te toetsen hypothese wordt genoemd) wordt dan getoetst op grond van waargenomen verschijnselen. Indien deze verschijnselen op grond van het model te onwaarschijnlijk zijn, wordt de hypothese verworpen. Vaak wordt zulk een hypothese juist opgesteld met de bedoeling en in de hoop, haar te kunnen verwerpen.

Voorbeeld. Men heeft van een aan onregelmatige continue fluctuaties onderhevige fysiologische grootheid tien paar metingen verricht, telkens vóór, zowel als na het toedienen van een bepaald middel, waarvan men vermoedt, dat het de bedoelde grootheid verhoogt. In 9 van de 10 gevallen vindt men na de toediening een grotere waarde dan er voor, in één geval een kleinere. Men hoopt nu te kunnen concluderen dat het bedoelde middel verhogend op de fysiologische grootheid werkt. Een eenvoudige, maar vrij ruwe, methode daartoe is de volgende „*tekentoets*" van R. A. FISHER.

Men stelt als nulhypothese, dat het middel in het geheel geen uitwerking heeft. De kansen op hogere en op lagere waarde na de toediening zijn dan gelijk ($= \frac{1}{2}$). De kans, dat men minstens 9 maal een verhoging vindt, is $11/2^{10}$. Indien men vooraf reeds met zekerheid kan zeggen, dat verlaging onmogelijk kan optreden, kan men hiermede volstaan (éénzijdige toetsing). Indien men echter (zoals meestal het geval is) verlaging niet a priori kan uitsluiten, moet men ook rekening houden met de proefreeksen, waarbij men minstens 9 maal een *verlaging* gevonden zou hebben. De kans daarop is eveneens $11/2^{10}$, tezamen $11/2^9$. Men kan dus de nulhypothese inderdaad spr 0,05 verwerpen. De werkelijke onbetrouwbaarheid van deze verwerping is slechts $11/2^9 \approx 0{,}022$.

Het doel van een statistische toetsing kan echter niet uitsluitend zijn, het trekken van onvoldoend betrouwbare conclusies te voorkomen. Immers al te voorzichtige toetsingsmethoden zouden er toe leiden, dat vrijwel nooit een hypothese verworpen zou kunnen worden. De toets dient daarom een voldoend groot *onderscheidingsvermogen* te hebben, d.w.z. ook de kans op ongerechtvaardigd niet-verwerpen moet voldoend klein gehouden worden. Het niet-verwerpen van een hypothese houdt nog niet het aanvaarden daarvan in. Immers talrijke andere hypothesen zouden wellicht, indien getoetst, evenmin verworpen worden. Men dient dus verdere conclusies niet te baseren op een niet-verworpen nulhypothese, maar op een gehele klasse van hypothesen, die alle bij toetsing niet-verworpen zouden worden, een zgn. *betrouwbaarheidsgebied.* Ten gevolge van wiskundige moeilijkheden is dit echter vaak niet mogelijk.

Bij dit alles is ondersteld, dat het waarnemingsmateriaal beschouwd kan worden als een „aselecte" (Engels: random) steekproef uit de collectie („populatie") waaromtrent men een conclusie wenst te trekken. De daardoor noodzakelijk gemaakte, soms zéér ingrijpende, beperking van het

geldigheidsgebied der conclusies kan vaak aanzienlijk worden verzacht door een doeltreffend aan het experiment voorafgegaan aselecteringsproces („randomization").

Deze algemene beschouwingen mogen worden toegelicht aan de hand van enkele zeer summiere opmerkingen, ontleend aan een aantal in het Mathematisch Centrum verrichte onderzoekingen. Daar de daarbij vermelde medische en biologische onderzoekingen zelve geheel buiten mijn competentiegebied vallen, is de bedoeling dat uitsluitend de *statistische* methodiek ter beoordeling staat.

1. Ter verwerking van waarnemingsresultaten van Dr J. F. VISSER, [1]) gepubliceerd in diens Utrechtse dissertatie, werd een statistisch onderzoek [2]) ingesteld, o.a. naar de invloed van ergotamine op de aciditeit van het maagsap bij ulcuslijders.

De reproduceerbaarheid der dagcurven, t.w. het overwegend voorkomen van stijgingen dan wel dalingen tussen opeenvolgende maaghevelingen, kon nog met behulp van de genoemde tekentoets worden vastgesteld. Ter vaststelling van de uitwerking der ergotamine bleek deze toets echter een te gering onderscheidingsvermogen te hebben, hetgeen niet te verwonderen is, daar zij alleen berust op het positief dan wel negatief zijn van een verschil, zonder de grootte van dit verschil in aanmerking te nemen. Ten einde hieraan tegemoet te komen, werd door J. HEMELRIJK een symmetrietoets ontworpen, die in sommige gevallen en in het bijzonder bij de dagelijkse gemiddelden, wèl tot een resultaat leidde, en sindsdien in vele andere gevallen kon worden toegepast.

2. Een methode ter diagnose van een moeilijk herkenbare kinderziekte (coeliaki) berust daarop, dat de onregelmatig fluctuerende vetresorptie bij toediening van een ongunstig (t.w. tarwehoudend) dieet een sterke daling ondergaat. Gevraagd werd, gedurende hoe lange tijd een gunstig en een ongunstig dieet moest worden gegeven, en hoe groot de daling moest zijn, om (spr α) de diagnose „coeliaki" te kunnen stellen, en dit in een voor artsen gemakkelijk toepasbaar rekenvoorschrift weer te geven. Op verzoek van de afdeling Bewerking Waarnemingsuitkomsten van de organisatie T.N.O. werd dit probleem ter hand genomen. Door combinatie van twee recente in Amerika ontstane methoden (de toets van WILCOXON en de sequente analyse van A. WALD) werd door spr. en J. HEMELRIJK een methode ontworpen, door J. H. B. KEMPERMAN uitgewerkt, en door de Rekenafdeling van het M.C. berekend. Er werd een schema opgesteld, waarmede de duur van toediening van het *ongunstige* dieet van aanvankelijk 14 dagen tot ten hoogste 7 dagen kon worden bekort, en wel kon

[1]) Onderzoek naar het effect van enkele sympathicusremmende stoffen bij patiënten met maag- of duodenumzweren, diss. Utrecht 1950.

[2]) Mej. A. M. J. A. VERBEEK, J. HEMELRIJK en H. NIEUWENHUIS, Rapport S 40 van het M.C., afgedrukt als aanhangsel van het genoemde proefschrift.

deze des te korter worden genomen, naarmate het geval (voor zover dit zich in de verminderde vetresorptie uit) ernstiger is.

Aan de hand van enkele verdere voorbeelden moge aangetoond worden van hoe groot belang een systematische opzet der contrôleproeven en zorgvuldige statistische bewerking daarvan is, en in het bijzonder een zorgvuldige bepaling van de nauwkeurigheid, waarmede de metingen *zonder* wijziging van omstandigheden reproduceerbaar zijn en welke verrassingen zich daarbij kunnen voordoen.

3. Bij een onderzoek door Dr J. GROEN c.s. naar de invloed van het cholesterolgehalte van de voeding op dat van het bloed bleek, dat de waarnemingsuitkomsten van betrekkelijk grote groepen proefpersonen, zelfs met verschillende diëten, gelijktijdig sterke stijgingen of dalingen vertoonden. [1]) Op grond daarvan werd vermoed, en door statistische toetsing bevestigd, dat ongewilde afhankelijkheden in de waarnemingsresultaten geslopen waren. En wel was dit het gevolg van het feit, dat telkens een groep van 8 bloedsera met eenzelfde standaardoplossing vergeleken was, en dat deze laatste somtijds enkele andere bewerkingen onderging dan de 8 bloedsera. Doordat de groepen van 8 op één standaard betrokken sera niet aselect over de 6 groepen van proefpersonen waren verdeeld, ontstonden grote moeilijkheden bij de statistische bewerking. In het bijzonder was het gevaar voor schijnconclusies bijzonder groot, doordat een effect aan dieetverandering kon worden toegeschreven dat in werkelijkheid door een verandering in de standaardoplossing kon zijn veroorzaakt. Hoewel het bij dit groot opgezette en zorgvuldig uitgevoerde experiment, mede dank zij het te hulp nemen van een groot deel van het moderne statistische arsenaal (tekentoets van FISHER, dubbele dichotomie van E. S. PEARSON, toets voor twee steekproeven van WILCOXON, symmetrietoets van J. HEMELRIJK, toets voor m rangschikkingen en rangcorrelatie, beide van M. G. KENDALL) gelukte, ondanks de afhankelijkheden de belangrijkste der verwachte verschijnselen vast te stellen, zijn toch vermoedelijk enkele resultaten verloren gegaan, die door doelmatige aselectering wellicht verkregen hadden kunnen worden. Dit wijst wel op de wenselijkheid de statistische bewerking van het door een onderzoek te verzamelen materiaal even zorgvuldig voor te bereiden als het medisch onderzoek zelf.

4. Bij een recent onderzoek van CHR. L. RÜMKE moesten een groot aantal tellingen van eosinophiele cellen verricht worden. De enigszins primitieve waarnemingsomstandigheden bij militairen te velde maakten een eenvoudige ijkingsmethode der verdunningspipetten noodzakelijk, die daartoe door RÜMKE werd bedacht (luchtbelijking). Ten einde de betrouw-

[1]) Zie Rapport S 71 door J. VAN KLINKEN en Dr J. HEMELRIJK Jr., Math. Centrum 1951.

baarheid van deze methode te onderzoeken werd een aantal pipetten ter ijking volgens een standaardmethode (kwikijking) aan een laboratorium gegeven. Bovendien werden bloedtellingen ter ijking gebruikt. Met de luchtbel- en kwikijking bleek, dat afwijkingen van ongeveer 3—5 % van de door de fabriek opgegeven waarde normaal waren en dat de verdunningsfactor voor sommige pipetten zelfs wel 9 of 11 bedroeg, in plaats van de opgegeven waarde 10. Bovendien bleken deze ijkingen nòch onderling nòch met de derde methode overeen te stemmen. Ten slotte toonden de beide eerstgenoemde methoden, dat er tussen pipetten die één en pipetten die twee mengkraaltjes bevatten een systematisch verschil in verdunningsfactor bestaat. Deze resultaten leiden dus tot de conclusie, dat bij een dergelijk onderzoek een nauwkeurige voorafgaande ijking van de verdunningspipetten door middel van bloedtellingen onontbeerlijk is, en niet door kwik- en andere ijkingsmethoden kan worden vervangen.