MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 101

Lezingen over Statistiek

I. Beschouwingen van algemene aard.

Door

Prof. Dr J.Hemelrijk

1953.

§ 1. De natuurwetenschappen houden zich bezig met verschijnselen, die zich onder gelijke omstandigheden steeds op dezelfde wijze voordoen. Althans, dat zegt men wel eens en in sommige gevallen is het ook wel zo: als men een zwaar voorwerp los laat, valt het naar beneden; als men in een U-vormige buis kwik giet en vervolgens in één van beide benen nog wat water,komt de waterspiegel hoger dan de kwikspiegel in het andere been; als men zoutzuur en zilvernitraat bijeenvoegt krijgt men een neerslag van zilverchloride, etc. In zulke eenvoudige gevallen kan men bovenstaande bewering wel handhaven. Bij kwantitatieve bepalingen echter gaat dit minder goed. Bij herhaling van een bepaling vindt men vaak een andere uitkomst dan de eerste maal en men behelpt zich dan vaak door het gemiddelde van een aantal bepalingen als uitkomst te nemen. In een dergelijk geval is men gewend van "waarnemingsfouten" te spreken en in gedachten de gevonden schommelingen niet te beschouwen als inhaerent aan het verschijnsel of de grootheid, die men onderzocht, maar als een gevolg van een onvolmaakte meettechniek. De grondslag van deze voorstelling van zaken is gelegen in de causale denkwijze, die door de eerste zin van deze paragraaf gekarakteriseerd kan worden.

Hoewel dit schema -een causaal verschijnsel en een met waarnemingsfouten behepte onderzoekingstechniek- vaak goede diensten kan bewijzen, is het lang niet altijd voldoende, om een bepaald onderzoekingsgebied te beschrijven. De statistiek, waarvan men zich ook binnen dit schema kan bedienen (er is een onderdeel van de statistiek, dat de "theorie der waarnemingsfouten" wordt genoemd), stelt zich dan ook in de regel op een ruimer standpunt, dat als grensgeval het causale standpunt bevat. Dit ruimere standpunt is gemakkelijker duidelijk te maken door voorbeelden op een minder exact gebied dan de natuur- of scheikunde en wij zullen, om het ons voorlopig gemakkelijk te maken, in dit eerste hoofdstuk enkele principes van de statistische denkwijze bespreken aan de hand van voorbeelden op medisch gebied. Wij komen later terug op het gebied der exacte natuurwetenschappen en der techniek en zullen later zien, dat dezelfde statistische beschouwingswijze ook daar van veel nut kan zijn en zelfs onontbeerlijk moet worden genoemd.

§ 2. Op het gebied der medicijnen en aanverwante wetenschappen is het vaak moeilijk een overzicht over een bepaald vraagstuk te krijgen, omdat er zoveel verschillende, en vaak deels onbekende, factoren in het spel zijn, die invloed op het resultaat uitoefenen. Indien men, zoals bij de natuurwetenschappen vaak het geval is, alle omstandigheden op één na geheel of nagenoeg constant kan houden, of althans de wijzigingen kent en tot op zekere hoogte beheerst, kan men de invloed van de bepalende factoren afzonderlijk onderzoeken. Een dergelijke beheersing van alle invloeden is bij experimenten, die op levende wezens betrekking hebben, in de regel niet mogelijk, zelfs in de meest eenvoudige gevallen niet. Terwijl men bij natuurwetenschappelijk geschoolden soms een zekere weerstand tegen de statistiek aantreft op grond van het feit, dat men meent met een streng causaal schema te kunnen volstaan, heeft men in medische kringen vaak de neiging het gebruik van wiskundige methoden in het algemeen te weren, daar men geen vertrouwen heeft in de toepasbaarheid daarvan, juist op grond van de strikt logische opbouw. Hoe zou men, met behulp van die zo exacte wiskunde iets kunnen bereiken op een gebied, waar steeds weer vaagheden en onzekerheden optreden en op zullen blijven treden! Het antwoord luidt: door de onzekerheid in de wiskundige verwerking toe te laten in de vorm van het begrip "waarschijnlijkheid" [1]). Men kan een wiskundig apparaat opbouwen (de mathematische statistiek), waarmee onzekere en aan fluctuaties onderhevige gegevens tot niet geheel zekere conclusies worden verwerkt. Dat kan men zonder wiskunde uiteraard ook wel. Maar de wiskundige behandeling heeft het voordeel, dat de mate van zekerheid der conclusies gespecificeerd wordt. De onzekerheid ervan wordt nl. in een getal, dat de "onbetrouwbaarheid" van de gebruikte methode genoemd wordt, uitgedrukt en dit getal heeft een scherp omschreven betekenis: het is de kans, om een verkeerde conclusie te trekken en geeft dus aan hoeveel foute conclusies bij veelvuldige toepassing van deze methode ongeveer verwacht moeten worden. Men kan in vele gevallen met behulp van deze wiskundige methoden conclusies bereiken, die men zo op het oog niet zou durven trekken, of die diep in het waarnemingsmateriaal verborgen liggen.

---

[1]) Wij zullen in dit hoofdstuk nog geen definitie van het begrip "waarschijnlijkheid" geven, maar aannemen, dat de lezer zich wel ongeveer kan voorstellen, wat daarmee bedoeld wordt. Verderop hopen wij exacter te werk te gaan.

Anderzijds wordt men vaak voor het trekken van overhaaste conclusies behoed. Men kan de mathematische statistiek dan ook zien als een apparaat ter verscherping van het gezonde verstand, een apparaat in de hand van de onderzoeker, zoals het lancet in de hand van de chirurg. Met behulp van dit apparaat kan men trachten de effecten van de omstandigheid, die men wenst te onderzoeken, te scheiden van de gevolgen van bijkomstige omstandigheden. Of, anders uitgedrukt: men kan de systematische effecten te voorschijn halen uit de poel van "toevallige" effecten, waarmee men op deze gebieden steeds te maken heeft.

§ 3. Een voorbeeld moge dit verduidelijken. Bij een onderzoek van het histaminegehalte van het bloed bij normale kinderen en kinderen met t.b.c., werden in twee kleine groepjes van 5 resp. 6 kinderen de volgende getallen gevonden:

Histaminegehalte van

I	II
normale kinderen	kinderen met t.b.c.
5,4	4,2
7,2	6,0
10,8	6,4
6,2	5,4
6,6	7,8
	7,6

De eenheid, waarin deze gehalten zijn uitgedrukt, is voor beide groepen van gegevens dezelfde en is voor de statisticus niet interessant. Het gehele onderzoek was natuurlijk veel uitgebreider, maar voor ons voorbeeld zullen wij alleen deze twee groepjes waarnemingen beschouwen.

De vraag is nu, of het histaminegehalte bij normale kinderen systematisch verschilt van dat bij kinderen met t.b.c. Daarbij wordt ondersteld, dat de twee groepjes kinderen onder overigens zoveel mogelijk gelijkwaardige omstandigheden zijn onderzocht. Is er nl. naast het verschil "normaal-t.b.c." nog een ander systematisch verschil tussen de groepjes, dan kan ook dat een verschil tussen de histaminegehalten veroorzaken.

Statistisch gezien kan dan ook alleen de vraag worden onderzocht, of de beide groepjes getallen op een systematisch verschil wijzen of niet. Het aanwijzen van de oorzaak van een, eventueel gevonden, systematisch verschil is niet meer de taak van de statisticus, maar van de medicus.

Indien de aantallen onderzochte kinderen in de beide groepen (die we verder met I en II zullen aangeven) groot zijn en indien b.v. alle waarnemingen in groep I kleiner zijn dan alle waarnemingen in groep II, dan zal niemand eraan twijfelen, dat er een systematisch verschil bestaat tussen de twee groepen van kinderen. Evenmin, indien de situatie juist andersom is. In een dergelijk extreem geval kan men de statistiek gevoegelijk missen. Indien daarentegen de waarnemingen der beide groepen zo door elkaar heen liggen, dat de groepen ongeveer samenvallen, zal men het er in het algemeen over eens zijn, dat er geen aanwijzingen bestaan voor een systematisch verschil. Tussen deze uitersten in liggen de twijfelgevallen (en deze zijn eer regel dan uitzondering). Dan kan men de mathematische statistiek te hulp roepen, om tot een beslissing te komen.

De statisticus kan zich bij dit probleem van verschillende technieken bedienen. Eén der eenvoudigste daarvan, die in al dergelijke gevallen van toepassing is, zullen wij kort beschrijven. Men begint, met van ieder der waarnemingen van groep I te tellen, hoeveel der waarnemingen van groep II groter zijn. Gelijken tellen daarbij voor $\frac{1}{2}$. De eerste waarneming van groep I (5,4) wordt in grootte overtroffen door 4 waarnemingen van groep II en is gelijk aan één waarneming van groep II; dit telt dus voor $4\frac{1}{2}$; de tweede (7,2) wordt overtroffen door 2 waarnemingen van groep II, de derde (10,8) door geen enkele (bijdrage 0), de vierde (6,2) door 3, en de vijfde (6,6) door 2. Alle tezamen geven zij

$$4\tfrac{1}{2}+2+0+3+2=11\tfrac{1}{2}.$$

Dit getal (de z.g. U van WILCOXON; zie [1] en [2] [2])) is een maat voor de onderlinge ligging van de groepen I en II. Indien de groep I geheel onder II ligt (d.w.z. indien de grootste waarneming van groep I kleiner is dan de kleinste van II), wordt U maximaal; in ons voorbeeld zou $5\times6=30$ de maximale waarde zijn; indien I geheel boven II ligt, wordt U minimaal en wel gelijk aan 0. De statisticus baseert nu zijn beoordeling van het gevonden resultaat op de waarschijnlijkheidsverdeling, die U zou bezitten, indien er geen systematisch verschil tussen de histaminegehalten van kinderen met en zonder t.b.c. zou bestaan,

---

[2]) Cijfers tussen vierkante haken verwijzen naar de litteratuurlijst aan het eind van dit hoofdstuk.

d.w.z. indien de verschillen tussen de groepen louter door toevallige omstandigheden ontstaan. Die onderstelling stelt hem nl. in staat uit te rekenen, hoe groot de kans is, dat u de waarden 0, 1, 2, etc. zal aannemen. Daarbij blijkt dan (zie fig. 1), dat de hoge en lage waarden van u kleine waarschijnlijkheden bezitten en de middenwaarden (bij ons voorbeeld de waarden in de buurt van $\frac{1}{2}\times 30=15$) grotere. De middelste waarde heeft de grootste waarschijnlijkheid en deze neemt naar beide zijden monotoon af.

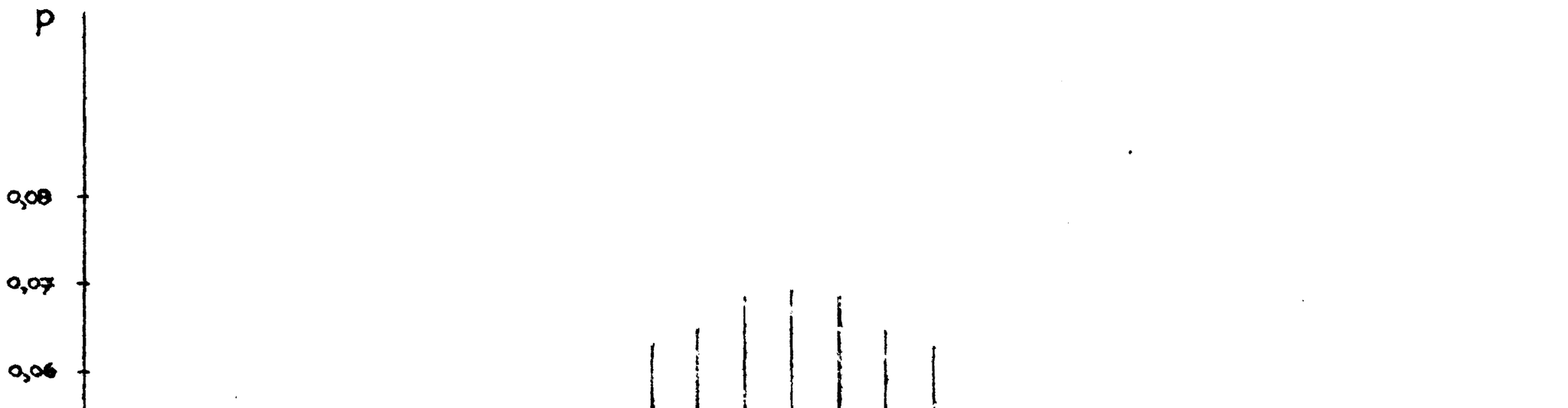


Fig. 1. De waarschijnlijkheidsverdeling van u voor twee niet systematisch verschillende groepen van 5 resp. 6 waarnemingen. De verticale lijntjes stellen de kans voor, dat u de op de abscis aangegeven waarde aanneemt. Zie ook de opmerking aan het eind van dit hoofdstuk.

Men berekent nu de som der waarschijnlijkheden van alle waarden van u, die even ver of nog verder van de middelste waarde verwijderd zijn dan de bij het experiment gevonden waarde, die in ons voorbeeld $11\frac{1}{2}$ bedroeg. Deze som van waarschijnlijkheden stelt de kans voor, om, als er geen systematisch verschil is tussen de beide onderzochte groepen, bij toeval een dergelijk resultaat te vinden als bij de werkelijk verrichte metingen, of een nog minder waarschijnlijk resultaat. De grootte van deze kans, die de overschrijdingskans wordt genoemd, geldt dan als een maat voor de aannemelijkheid van de onderstelling, dat er geen systematisch verschil is tussen beide groepen. Is de overschrijdingskans klein, dus bevindt men zich in één der "staarten" van

de verdeling van u, dan verwerpt men de gemaakte onderstelling en besluit tot het bestaan van een systematisch verschil. Gewoonlijk stelt men een bepaalde grens vast. Indien de overschrijdingskans daaronder ligt, besluit men tot het bestaan van een systematisch verschil en anders niet. Deze grens wordt de onbetrouwbaarheidsdrempel genoemd en men neemt daarvoor vaak op traditionele gronden de waarde 1/20. Dit betekent dan, dat men een kans van 1/20 loopt om, indien er geen systematisch verschil bestaat, toch daartoe te besluiten. Of ook: ongeveer één op de twintig keer, dat men een dergelijke proef doet, terwijl er geen systematisch verschil is, zal men toch concluderen, dat dit er wel is. Vindt men één op de twintig keer te veel, dan kan men zich van een kleinere drempel, b.v. 1/100 of 1/1000 bedienen. Hoe kleiner men echter deze breuk kiest, hoe groter een wèl bestaand systematisch verschil moet zijn, wil men een goede kans hebben, om het te ontdekken.

Berekenen wij in ons voorbeeld de overschrijdingskans, dan blijkt deze 0,6 te zijn. Deze waarnemingen geven dus geen aanleiding, om te vermoeden dat er een systematisch verschil in het histamingehalte bestaat tussen gezonde kinderen en kinderen met t.b.c. Er zij met nadruk op gewezen, dat een dergelijke negatieve conclusie essentieel negatief is; d.w.z. dat men daaruit niet de conclusie moet trekken, dat er nu ook werkelijk geen systematisch verschil is. Er is er alleen geen ontdekt. Misschien is er wèl een, maar is het te klein, om met zo weinig waarnemingen al ontdekt te worden. Met deze mogelijkheid moet men rekening houden. Zou de uitslag anders zijn geweest, dus zou er wèl een systematisch verschil zijn ontdekt, dan behoeft men een dergelijk vaag voorbehoud niet te maken. Het gehele voorbehoud ligt dan opgesloten in het opgeven van de gebruikte onbetrouwbaarheidsdrempel.

Ter oriëntering zij nog vermeld, dat bij twee groepen van 5 en 6 waarnemingen, zoals in ons voorbeeld, de kans op geheel gescheiden liggen van de groepen (dus $u=0$ of $u=30$) slechts 0,004 bedraagt, terwijl de kans op één der waarden 0, 1, 2, 3, 27, 28, 29 of 30 voor u gelijk is aan 0,03. Neemt men er de waarden 4 en 26 nog bij, dan stijgt deze kans tot 0,052. Dit geldt in de onderstelling, dat er geen systematisch verschil tussen de groepen is, de onderstelling die men wenst te "toetsen" en eventueel te verwerpen. Men ziet dus, dat zelfs bij dergelijke geringe aantallen reeds een duidelijk resultaat verkregen kan worden, indien het verschil tussen de onderzochte groepen voldoende groot is.

de verdeling van u, dan verwerpt men de gemaakte onderstelling en besluit tot het bestaan van een systematisch verschil. Gewoonlijk stelt men een bepaalde grens vast. Indien de overschrijdingskans daaronder ligt, besluit men tot het bestaan van een systematisch verschil en anders niet. Deze grens wordt de onbetrouwbaarheidsdrempel genoemd en men neemt daarvoor vaak op traditionele gronden de waarde 1/20. Dit betekent dan, dat men een kans van 1/20 loopt om, indien er geen systematisch verschil bestaat, toch daartoe te besluiten. Of ook: ongeveer één op de twintig keer, dat men een dergelijke proef doet, terwijl er geen systematisch verschil is, zal men toch concluderen, dat dit er wel is. Vindt men één op de twintig keer te veel, dan kan men zich van een kleinere drempel, b.v. 1/100 of 1/1000 bedienen. Hoe kleiner men echter deze breuk kiest, hoe groter een wèl bestaand systematisch verschil moet zijn, wil men een goede kans hebben, om het te ontdekken.

Berekenen wij in ons voorbeeld de overschrijdingskans, dan blijkt deze 0,6 te zijn. Deze waarnemingen geven dus geen aanleiding, om te vermoeden dat er een systematisch verschil in het histaminegehalte bestaat tussen gezonde kinderen en kinderen met t.b.c. Er zij met nadruk op gewezen, dat een dergelijke negatieve conclusie essentieel negatief is; d.w.z. dat men daaruit niet de conclusie moet trekken, dat er nu ook werkelijk geen systematisch verschil is. Er is er alleen geen ontdekt. Misschien is er wèl een, maar is het te klein, om met zo weinig waarnemingen al ontdekt te worden. Met deze mogelijkheid moet men rekening houden. Zou de uitslag anders zijn geweest, dus zou er wèl een systematisch verschil zijn ontdekt, dan behoeft men een dergelijk vaag voorbehoud niet te maken. Het gehele voorbehoud ligt dan opgesloten in het opgeven van de gebruikte onbetrouwbaarheidsdrempel.

Ter oriëntering zij nog vermeld, dat bij twee groepen van 5 en 6 waarnemingen, zoals in ons voorbeeld, de kans op geheel gescheiden liggen van de groepen (dus $u=0$ of $u=30$) slechts 0,004 bedraagt, terwijl de kans op één der waarden 0, 1, 2, 3, 27, 28, 29 of 30 voor u gelijk is aan 0,03. Neemt men er de waarden 4 en 26 nog bij, dan stijgt deze kans tot 0,052. Dit geldt in de onderstelling, dat er geen systematisch verschil tussen de groepen is, de onderstelling die men wenst te "toetsen" en eventueel te verwerpen. Men ziet dus, dat zelfs bij dergelijke geringe aantallen reeds een duidelijk resultaat verkregen kan worden, indien het verschil tussen de onderzochte groepen voldoende groot is.

§ 4. In het bovenstaande is één der moderne toetsingsmethoden der mathematische statistiek kort beschreven. Details over de berekening der overschrijdingskans zijn daarbij niet gegeven. Deze kan men vinden in [2]. Er is nog een ander aspect van de mathematische statistiek, dat wij hier kort willen bespreken. Zoals reeds is gezegd, is de statistiek er niet alleen op uit, om de mogelijkheden voor het trekken van conclusies te vergroten, maar ook, om deze conclusies op een hechte basis te plaatsen en schijnresultaten te vermijden.

Schijnresultaten kunnen ontstaan, doordat ongecontroleerde of zelfs onvermoede nevenomstandigheden een systematisch verschil tussen twee groepen waarnemingen (of in het algemeen een systematisch effect) veroorzaken, dat dan ten onrechte aan een bekende omstandigheid of oorzaak wordt geweten. Zo zou b.v., indien de boven beschreven waarnemingen van groep I in een andere tijd van het jaar zijn verricht dan die van groep II, een seizoeneffect tot een systematisch verschil kunnen leiden, dat men dan ten onrechte aan de t.b.c. van groep II zou kunnen wijten. Omgekeerd is het om dezelfde reden mogelijk dat een bestaand effect, dat bij een betere proefopstelling wel ontdekt zou zijn, nu niet aan het licht komt. Beide gevaren zijn verre van denkbeeldig. Want behalve de zich veelvuldig voordoende seizoeneffecten zijn er nog vele andere mogelijke oorzaken voor schijneffecten, zoals verschil van sociaal milieu of voeding der beide groepen kinderen, verschil in lichaamsbeweging, enz., of ook: een verloop in de meettechniek tijdens het onderzoek.

Een typisch voorbeeld treft men ook aan op een geheel ander gebied, dat der prophylactische inenting tegen infectieziekten. Inenting kan gewoonlijk slechts op basis van vrijwilligheid geschieden en men gaat er dan ook vaak toe over, bij het beproeven van een prophylacticum,dat door inenting wordt toegediend, van een bepaalde groep mensen allen, die zich vrijwillig daartoe aanmelden, in te enten. Een dergelijke proef kan echter op zichzelf niet tot een conclusie omtrent de al of niet werkzaamheid van het prophylacticum leiden. Want als men achteraf de veelvuldigheid van het optreden van de ziekte, waar het om gaat, beschouwt in de groep der vrijwillig ingeënten en deze vergelijkt met die bij de overige mensen van de onderzochte groep, dan zijn er twee mogelijke oorzaken voor een systematisch verschil: in de eerste plaats kan het prophylacticum gunstig (of misschien ook ongunstig) hebben gewerkt en in de tweede plaats kan de selectie, die in de vrijwilligheid der aanmelding voor de inenting schuilt, eveneens in beide richtingen een systematisch verschil veroorzaken. Niet zelden zullen juist die mensen,

die ook overigens beter op hun gezondheid letten, zich wel aanmelden en de zorgelozen niet. Daardoor kan dan een verschil tussen de beide groepen ontstaan, zonder dat het prophylacticum daar ook maar iets mee te maken heeft. Achteraf kan men de twee mogelijke oorzaken niet meer van elkaar scheiden, zodat er op deze wijze geen conclusie te bereiken valt. Ook niet, indien er geen systematisch verschil gevonden wordt; want de twee mogelijke oorzaken zouden elkaar ook kunnen tegenwerken. Een dergelijk experiment is dus in de regel waardeloos.

§ 5. Het middel, dat de statistiek biedt, ter vermijding van dergelijke foute conclusies bestaat in principe daarin, dat men de bijkomstige omstandigheden, ook de onbekende, door een systeem van lotingen tot " toevallige omstandigheden " maakt. Zij kunnen dan nog slechts bij toeval een schijnresultaat veroorzaken en niet meer systematisch. De kans, dat zij per toeval tot een foute conclusie aanleiding zullen geven, is dan echter bekend: deze is nl. verwerkt in de onbetrouwbaarheidsdrempel van de toegepaste statistische methode. Het gevaar voor het systematische optreden van schijneffecten is nu verdwenen.

Wij kunnen dit toelichten aan het voorbeeld der inenting. De moeilijkheid zou bij dat probleem geheel verdwijnen, indien men de groep van vrijwilligers (dus de groep personen die ingeënt zijn) door loting uit de gehele groep van personen zou kunnen afzonderen. De selectie op grond van vrijwilligheid zou men dan vervangen door een "aselecte" keuze [3]), die niet door enige systematische factor bepaald wordt. Het is echter duidelijk, dat men niet door loting kan bepalen, of iemand vrijwilliger is of niet. Men kan echter wel de groep der vrijwilligers bij de aanmelding door loting in twee groepen verdelen en de ene daarvan met het prophylacticum inenten, terwijl de andere (zonder het te weten) slechts het zeker niet werkzame oplosmiddel ingespoten krijgt. Op die wijze verkrijgt men twee groepen, die vergeleken kunnen worden zonder gevaar voor schijnresultaten. Bovendien kan de tweede groep (die dus een schijninenting heeft gehad) nog met de groep der niet-vrijwilligers vergeleken worden, om het effect van de selectie door vrijwilligheid en de eventuele suggestieve werking van het geven van een schijninenting te beoordelen. Tegen deze proefopzet bestaan wellicht be-

---

[3]) De term "aselect" (Engels: random) is ingevoerd door Prof. Dr D.VAN DANTZIG, De natuur als tegenspeler, Statistica 5 (1952), pp. 149-160.

zwaren van ethische aard. De medische ethiek is een terrein, waar de statisticus zich niet mee behoort te bemoeien. Dit ontneemt hem echter niet het recht de medicus erop te wijzen, dat slechts een proefopzet in de trant van de bovenstaande, waarbij het selectieproces door het één of andere aselecteringsprocédé geneutraliseerd wordt, de mogelijkheid biedt iets te weten te komen over de werking van het gebruikte prophylacticum.

§ 6. Een ander voorbeeld van het nut van een aselecteringsprocedure is het vermijden van foute conclusies, die tengevolge van een verloop in de waarnemingstechniek optreden. Dit verschijnsel is minder zeldzaam, dan men wellicht zou denken. Bij de meeste reeksen van gelijksoortige bepalingen, b.v. ook bij histaminebepalingen, treden in één of meer der phasen van de bepaling veranderingen op, al of niet van systematische aard, die maken, dat de bepalingen, die op verschillende dagen of zelfs op verschillende tijden van één dag verricht zijn, niet zonder meer als gelijkwaardig kunnen worden beschouwd. De schijnresultaten, die daardoor kunnen ontstaan, kan men vermijden door het treffen van twee maatregelen. De eerste daarvan is, dat men de proef, die men wenst te verrichten, in een aantal kleine gelijksoortige proeven splitst en ieder daarvan op één dag uitvoert en de tweede is, dat men de volgorde der bepalingen op ieder van de proefdagen opnieuw door loting vaststelt. De meetresultaten van iedere dag apart zijn dan, wat de bijkomstige omstandigheden betreft, aselect gemaakt en dus voor een statistische analyse geschikt. De resultaten van deze analyses voor de verschillende proefdagen kunnen tot een geheel worden gecombineerd, waarbij dan ook de "dag-verschillen" zijn uitgeschakeld.

§ 7. De techniek der aselecte keuze, die uitgevoerd kan worden met behulp van tabellen van aselecte getallenrijen [4]), maakt het dus mogelijk schijnresultaten te vermijden. Men kan echter op deze wijze niet verhinderen, dat nevenomstandigheden, die een grote invloed op het meetresultaat uitoefenen, een eventueel wèl bestaand systematisch verschil, dat men juist wenst op te sporen, "onzichtbaar" maken. Indien het effect, waarnaar men zoekt, slechts gering is in vergelijking met de werking van niet relevante invloeden, dan is de kans groot, dat dit effect door die invloeden verdoezeld zal worden. Dit kan slechts voorkomen

---

[4]) Deze techniek heet in het Engels "randomization", de aselecte getallenrijen worden "random numbers" genoemd [3].

worden door het aantal waarnemingen te vergroten of door, op de boven voor de "dag-verschillen" beschreven methode, de niet relevante invloeden zoveel mogelijk uit te schakelen. Dit laatste heeft het voordeel, dat men het aantal waarnemingen niet zo groot behoeft te nemen. Het is van belang deze kwesties bij ieder onderzoek van tevoren grondig te overwegen en er bij de proefopzet rekening mee te houden. Daarbij behoort men tevens van tevoren zijn gedachten te laten gaan over de wijze, waarop men zich voorstelt de verkregen resultaten te analyseren.

Litteratuur.

[1] F.Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945), pp. 80-82.

[2] H.R.van der Vaart, Gebruiksaanwijzing voor de toets van Wilcoxon, Rapport S 32 (M 4) van het Mathematisch Centrum, Amsterdam 1950.

[3] M.G.Kendall and B.Babington Smith, Tables of random sampling numbers, Cambridge Un. Press 1939.

Opmerking. In figuur 1 en die delen van de tekst, die op deze figuur betrekking hebben, is geen rekening gehouden met de mogelijkheid van het optreden van gelijke waarnemingen. Indien dit verschijnsel zich wel voordoet, zoals bij ons numerieke voorbeeld, verandert de waarschijnlijkheidsverdeling enigszins.

MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING
Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 101

Lezingen over Statistiek

II. Waarschijnlijkheidsrekening als basis voor de statistiek.

Door

Prof. Dr J.Hemelrijk

1953.

1. De experimentele wet der grote getallen.

Statistiek is toegepaste waarschijnlijkheidsrekening [1]) en wij zullen daar dus ook iets over moeten vertellen. De whr is een onderdeel der zuivere wiskunde en berust op een aantal axioma's, die betrekking hebben op het begrip waarschijnlijkheid [1]). Dit begrip wordt niet nader gedefinieerd; de eigenschappen van whn worden door axioma's vastegelegd en met behulp daarvan kunnen wij uit gegeven whn andere afleiden.

Om het contact met de experimentele werkelijkheid, waarop men de whr wil gaan toepassen, niet te verliezen, kan men zich een wh ongeveer voorstellen als een frequentiequotiënt [1]) in een zeer lange reeks experimenten [2]). De toepassing van de whr op experimenten berust nl. op het volgende merkwaardige verschijnsel, dat de experimentele wet der grote getallen wordt genoemd.

Als men een experiment E, dat verschillende uitkomsten kan hebben, een aantal (n) malen op dezelfde wijze herhaalt en het aantal malen (x), dat een bepaald resultaat S optreedt, deelt door n , dan verkrijgt men een tussen 0 en 1 gelegen getal

$$fq(S) \stackrel{def}{=} x/n ,$$

dat men het frequentiequotiënt van S noemt. Verricht men nu een aantal dergelijke reeksen van experimenten, dan blijkt, dat de verschillen der voor fq(S) gevonden waarden bij toenemende n geringer worden.

Toepassing van de whr op een bepaald probleem veronderstelt in feite, dat dit verschijnsel zich ook bij dat probleem zou voordoen als men steeds langere reeksen experimenten verrichtte.

Als voorbeeld van een dergelijke reeks van experimenten zullen wij, wegens de eenvoud en de practische uitvoerbaarheid, worpen met een dobbelsteen beschouwen. Deze dobbelsteen behoeft niet "zuiver" te zijn.

---

[1]) Wij gebruiken de volgende afkortingen:

wh (meervoud whn)= waarschijnlijkheid,
whr= waarschijnlijkheidsrekening,
fq (meervoud fqn)= frequentiequotiënt.

[2]) Wij gebruiken het woord "experiment" ook in de zin van: "het verrichten van een waarneming".

2. De keuze der axioma's en de interpretatie van whn.

Zoals boven reeds gezegd is kunnen wij ons een wh vagelijk voorstellen als een fq in een lange reeks experimenten. Een wh is, scherper gezegd, niet anders dan het mathematische analogon van een dergelijk fq en de whr is, o.a., een mathematisch model voor zulke reeksen experimenten. Daarom worden de axioma's der whr analoog aan enkele der eenvoudigste eigenschappen van fqn genomen. Uitgaande van bepaalde, gegeven of gegeven onderstelde, whn kunnen dan met behulp van de axioma's andere whn worden berekend en stellingen worden afgeleid en wij hopen dan, dat ook andere eigenschappen van fqn een analogon in whr zullen blijken te bezitten, zodat wij whn met vrucht zullen kunnen interpreteren als fqn in lange reeksen experimenten. Want daarop zullen wij dan experimenteel controleerbare voorspellingen van practische aard kunnen baseren en daar is het om te doen. Inderdaad zal blijken, dat dit mogelijk is. Het voordeel van deze opzet is, dat het axiomatische systeem zeer eenvoudig van structuur wordt en dus goed handelbaar is.

3. Eigenschappen van fqn.

De eenvoudige eigenschappen van fqn, die wij als axioma's voor whn zullen invoeren, zullen wij niet streng afleiden, maar slechts illustreren aan het voorbeeld van een reeks worpen met een dobbelsteen. Iedere worp geeft als uitkomst één der getallen 1, 2, 3, 4, 5 of 6. De gehele reeks, R, ziet er dus b.v. als volgt uit:

(1) $\qquad R: 2,5,3,6,2, \ldots ,4,1,6,3. \qquad$ (n worpen).

Het fq van de uitkomst "2" b.v. is gedefinieerd als het quotiënt van het aantal tweeën in R en n en is dus een getal tussen 0 en 1. Dit geldt uiteraard algemeen en voor een willekeurige uitkomst A bij een willekeurige reeks van experimenten geldt dus

(2) $\qquad 0 \leqq fq(A) \leqq 1.$

Verder is in R b.v. $fq(0)=0$, d.w.z.

(3) $\qquad fq(A)=0 \qquad$ als A onmogelijk is.

Beschouw nu bij R de uitkomst A;"één der getallen 1, 2, 3, 4, 5 of 6 wordt verkregen". Deze uitkomst is zeker en bezit $fq=1$. Dus

(4) $\qquad fq(A)=1 \qquad$ als A zeker optreedt.

Tenslotte vatten wij nog de volgende twee uitkomsten in het oog:

A ≡ "er komt een even getal uit"
B ≡ "er komt een getal < 5 uit".

Dan geldt in R, zoals gemakkelijk na te gaan is

(5) $$fq(A \text{ of } B) = fq(A) + fq(B) - fq(A \text{ en } B) \quad ,$$

waarin " A of B " betekent, dat minstens één van deze beide gebeurtenissen [3]) A of B optreedt.

Al deze eigenschappen gelden voor iedere reeks R.

4. De axioma's der whr.

De wh van een uitkomst A van een experiment E geven wij aan met P[A] (P ="probability"). De axioma's zijn nu geheel analoog aan (2),...,(5). Zij luiden:

Ax. 1: $0 \leqq P[A] \leqq 1$ voor iedere A bij iedere E.

Ax. 2: $P[A]=0$ als A bij E onmogelijk is.

Ax. 3: $P[A]=1$ als A bij E zeker is.

Ax. 4: $P[A \text{ of } B] = P[A] + P[B] - P[A \text{ en } B]$.

Opmerkingen. Als de uitkomsten A en B elkaar uitsluiten, d.w.z. niet tegelijk kunnen optreden, is de laatste term in Ax. 4 gelijk aan 0 (nl. volgens Ax.2). Ax. 4 gaat dan over in

(6) $P[A \text{ of } B] = P[A] + P[B]$ als A en B elkaar bij exp.E uitsluiten.

Deze eigenschap geldt ook analoog bij fqn en laat zich bovendien gemakkelijk tot meer dan twee uitkomsten generaliseren. Echter niet tot oneindig veel en dan ontbreekt ook het analogon bij fqn, daar bij een werkelijk uitgevoerde reeks experimenten (die altijd eindig is) niet oneindig veel uitkomsten op kunnen treden. Deze leemte wordt opgevuld door een vijfde axioma, dat wij buiten beschouwing zullen laten. Wij zullen vooral van (6) en van de generalisatie daarvan voor meer dan 2 gebeurtenissen vaak gebruik maken.

5. Stochastische onafhankelijkheid.

Behalve gewone fqn en whn voeren wij nu nog voorwaardelijke in. Wij beschouwen daartoe opnieuw R en twee gebeurtenissen:

A ≡ "er komt een even getal uit"

B ≡ "er komt een getal < 5 uit".

Wij definiëren nu het voorwaardelijke fq van A onder de voorwaarde B door

(7) $$fq(A|B) \overset{def}{=} \frac{fq(A \text{ en } B)}{fq(B)} .$$

---

[3]) De woorden "gebeurtenis", "uitkomst" en "waarnemingsuitkomst" worden door elkaar in ongeveer gelijke betekenis gebruikt.

Anders gezegd: in plaats van R beschouwen wij de deelreeks R' van R, bestaande uit alle uitkomsten < 5; het gewone fq van A op deze deelreeks is het voorwaardelijke fq op R zelf van A, onder voorwaarde B.

Verder noemen wij de uitkomsten A en B <u>onafhankelijk op R</u>, indien op R geldt

(8) $$fq(A|B) = fq(A).$$

Voor de speciale gebeurtenissen A en B, die wij boven genoemd hebben, geldt

$$fq(A \text{ en } B) = fq(2 \text{ of } 4) = fq(2) + fq(4),$$

$$fq(B) = fq(1,2,3 \text{ of } 4) = fq(1) + fq(2) + fq(3) + fq(4)$$

en

$$fq(A) = fq(2) + fq(4) + fq(6).$$

Indien dus b.v. in R alle getallen 1,2,...,6 even vaak voorkomen, zijn linker- en rechterlid van (8) beide gelijk aan $\frac{1}{2}$, zodat A en B dan onafhankelijk zijn op R, daar dan alle fqn= $\frac{1}{6}$ zijn (deze voorwaarde is echter geen noodzakelijke voorwaarde).

Voor whn voeren wij nu analoge definities in. De <u>voorwaardelijke wh van A onder de voorwaarde B</u> definiëren wij als

(9) $$P[A|B] = \frac{P[A \text{ en } B]}{P[B]}$$

en wij noemen A en B <u>stochastisch onafhankelijk</u> bij experiment E, als

(10) $$P[A|B] = P[A]$$

is.

Een voorbeeld van twee stochastisch onafhankelijke gebeurtenissen vormen weer de genoemde uitkomsten A en B, als de kans [4]) op ieder der 6 mogelijke uitkomsten gelijk is (volgens de axioma's en (6) is ieder van deze kansen dan gelijk aan $\frac{1}{6}$) [5]).

Voor twee stochastisch onafhankelijke gebeurtenissen A en B geldt volgens (9) en (10):

(11) $$P[A \text{ en } B] = P[A]P[B],$$

hetgeen, evenals (6), tot meer dan twee gebeurtenissen kan worden uitgebreid. De analoge eigenschap geldt voor fqn als A en B op R onafhankelijk zijn.

---

[4]) Kans is synoniem met wh.

[5]) Wij noemen de dobbelsteen dan "zuiver". In dit stadium is "zuiver" dus een wiskundig begrip, waarvan de practische betekenis later zal blijken te zijn, dat in een lange reeks worpen alle fqn ongeveer gelijk aan $\frac{1}{6}$ worden.

Opmerking. Ook (11) zullen wij herhaaldelijk gebruiken. Wij wijzen er echter met nadruk op, dat dit geen axioma is en ook niet een eigenschap, die uit de axioma's volgt; (11) volgt uit de definities (9) en (10) en geldt alleen voor uitkomsten, waarvan de stochastische onafhankelijkheid binnen het mathematische model gegeven is of gepostuleerd wordt. Wij behandelen in de volgende paragraaf een belangrijk voorbeeld van onafhankelijkheid.

6. Onafhankelijkheid van experimenten.

Tot nu toe hebben wij geen onderstelling gemaakt over de onderlinge onafhankelijkheid van de reeks experimenten R. Alles, wat in de vorige paragrafen is gezegd, geldt onafhankelijk van de volgorde der uitkomsten in R en blijft geldig, als wij reeksen van het type

$$1,1,1,1,\ldots$$

of

$$1,2,3,4,5,6,1,2,\ldots$$

of iets dergelijks beschouwen, waarbij iedere term op de één of andere wijze uit de vorige volgt. Deze gevallen interesseren ons echter zeer weinig en wij zullen speciaal reeksen van experimenten beschouwen, waarbij een dergelijke regelmaat niet optreedt. Wij beperken ons nl. tot reeksen experimenten, waarbij de uitkomst van ieder experiment niet van de reeds verkregen uitkomsten afhangt. Geven wij het nummer van het experiment aan door een index bij de uitkomst, zodat b.v. A_1 betekent: uitkomst A bij de eerste worp, dan houdt deze eis dus, in wh-theoretische terminologie, in, dat

$$P[A_2] = P[A_2|A_1] = P[A_2|B_1] = P[A_2|C_1] = \ldots$$

$$P[B_3] = P[B_3|A_1 \text{ en } B_2] = P[B_3|A_1 \text{ en } A_2] = \ldots$$

enz., zodat wij steeds (11) mogen toepassen. Zo wordt dan b.v. de wh, om bij het eerste experiment A en bij het tweede B te verkrijgen:

$$P[A_1 \text{ en } B_2] = P[A_1].P[B_2].$$

Is het experiment bovendien iedere keer gelijk, dan wordt dit in het mathematische model uitgedrukt door de relaties

$$P[A_1] = P[A_2] = \ldots$$

$$P[B_1] = P[B_2] = \ldots$$

etc.,

en dan kunnen wij dus kortweg schrijven

(11') $$P[A_1 \text{ en } B_2] = P[A].P[B]$$

met weglating van de indices in het tweede lid, daar deze er toch niet toe doen. In het eerste lid kunnen wij ze niet missen, daar A en B b.v. uitkomsten kunnen zijn, die bij één experiment niet tegelijk op kunnen treden, maar wel bij opeenvolgende experimenten.

Deze stochastische onafhankelijkheid van de uitkomsten van de verschillende experimenten van een reeks is een wiskundige eigenschap, opgelegd aan de whn in het mathematische model en wij moeten ons dus nog trachten te realiseren, wanneer wij deze eigenschap redelijkerwijze vervuld kunnen achten.

Het analogon in termen van fqn verkrijgen wij, als wij een reeks R van experimenten veranderen in een reeks R* van paren experimenten, b.v. het eerste en tweede experiment van R, het derde en vierde, enz. Geven wij nu aan de eerste uitkomst van ieder paar de index 1 en aan de tweede de index 2, dan kunnen wij op de zo verkregen reeks R* van paren van experimenten fqn van de vorm

$$fq(A_1),\ fq(A_2|B_1),\ fq(C_2|A_1),\ \text{etc.}$$

invoeren en de onafhankelijkheid van de eerste en tweede uitkomsten van de paren wordt dan gegeven door

$$fq(A_1) = fq(A_1|A_2) = fq(A_1|B_2) = fq(A_1|C_2) = \ldots$$

enz. Hetzelfde kunnen wij doen voor drietallen, enz. en indien wij nu verwachten, of door experimentele verificatie weten, dat in reeksen van experimenten aan deze voorwaarden steeds beter voldaan wordt naarmate deze reeksen langer genomen worden, dan kunnen wij de stochastische onafhankelijkheid der verschillende experimenten binnen het model zonder bezwaar postuleren.

Beschouwen wij b.v. een reeks worpen met een dobbelsteen, op de gebruikelijke wijze uitgevoerd, dan kunnen wij ons ternauwernood voorstellen, dat de resultaten van vroegere worpen de uitkomst van een nieuwe worp zullen beïnvloeden (dit heeft met zuiverheid van de dobbelsteen niets te maken). Dit betekent echter o.a. juist, dat wij in een lange reeks van worpen-paren ongeveer hetzelfde fq van een bepaalde uitkomst A bij de tweede worpen van ieder paar verwachten als bij de tweede worpen van die paren, die bij de eerste worp een bepaalde uitkomst B hebben gegeven. Men kan deze verwachting natuurlijk experimenteel verifiëren.

7. Mathematisch model van een reeks onafhankelijke experimenten.

Samenvattend kunnen wij dus zeggen,dat wij als mathematisch model voor een reeks gelijksoortige experimenten E_1, E_2, etc. (zoals worpen met een dobbelsteen), die geen invloed uitoefenen op elkaars uitkomsten, en die ieder de uitkomsten A, B, C, etc., kunnen hebben, nemen: de wh van ieder der uitkomsten A, B, C, etc. is bij ieder experiment dezelfde en de uitkomsten zijn stochastisch onafhankelijk, terwijl de axioma's der whr gelden.

Zoals reeds eerder gezegd kunnen wij dan binnen dit model allerlei nieuwe whn berekenen en stellingen afleiden en zodoende trachten te verifiëren, of whn, afgezien van de door de axioma's eraan gegeven eigenschappen, ook overigens een goede overeenkomst vertonen met fqn in lange reeksen waarnemingen.

8. De binomiale verdeling.

Hulpstelling: Het aantal verschillende verdelingen van k objecten in twee groepen van h resp. k-h objecten bedraagt

(12) $$\binom{k}{h} = \frac{k!}{h!(k-h)!} .$$

Bewijs. Wij leggen de k objecten op een rij en nemen de eerste h als de ene groep. Er zijn k! rangschikkingen, maar daar de h! permutaties van de eerste h objecten in de rij en eveneens de (k-h)! van de overige aan de verdeling in twee groepen niets veranderen, moet het aantal rangschikkingen door h!(k-h)! gedeeld worden, om het aantal verdelingen in twee groepen te verkrijgen. H.u.v. (12).

Opmerking: is $h = k-h = \frac{1}{2}k$, dan geldt het lemma slechts, indien verwisseling van de twee groepen geacht wordt tot een nieuwe verdeling te leiden. M.a.w. men moet dan een eerste en een tweede groep onderscheiden.

Wij beschouwen nu een reeks onafhankelijke experimenten, zoals in § 7 beschreven en vatten speciaal één bepaalde uitkomst in het oog, die wij met S aangeven. Om de terminologie een beetje levendig te maken, duiden wij deze uitkomst aan als een "succes" en de overige mogelijke uitkomsten als "mislukkingen" (aangegeven door $\bar{S}$). Het aantal successen bij n experimenten geven wij aan met $\underline{x}$. $\underline{x}$ kan dus de waarden 0,1,2,...,n aannemen en wel ieder met een bepaalde wh; wij zeggen dan, dat $\underline{x}$ een stochastische grootheid is (≡ een grootheid met een wh-verdeling) en geven dit aan door onderstreping van het symbool x; x zonder onderstreping gebruiken wij, om een bepaalde waarde, die $\underline{x}$ aan kan nemen, aan te geven. Dan geldt:

Stelling 1. Bij een reeks van n onafhankelijke experimenten, die ieder een wh p op succes hebben, bezit het aantal successen $\underline{x}$ een wh-verdeling van de vorm

(13) $$P[\underline{x}=x]=\binom{n}{x}p^x q^{n-x} \quad (q=1-p).$$

Opmerkingen: formule (13) geldt voor iedere gehele waarde van x, indien wij $\binom{n}{x}=0$ stellen voor $x<0$ en $x>n$. De door (13) aangegeven wh-verdeling wordt de binomiale verdeling of de verdeling van Bernoulli genoemd.

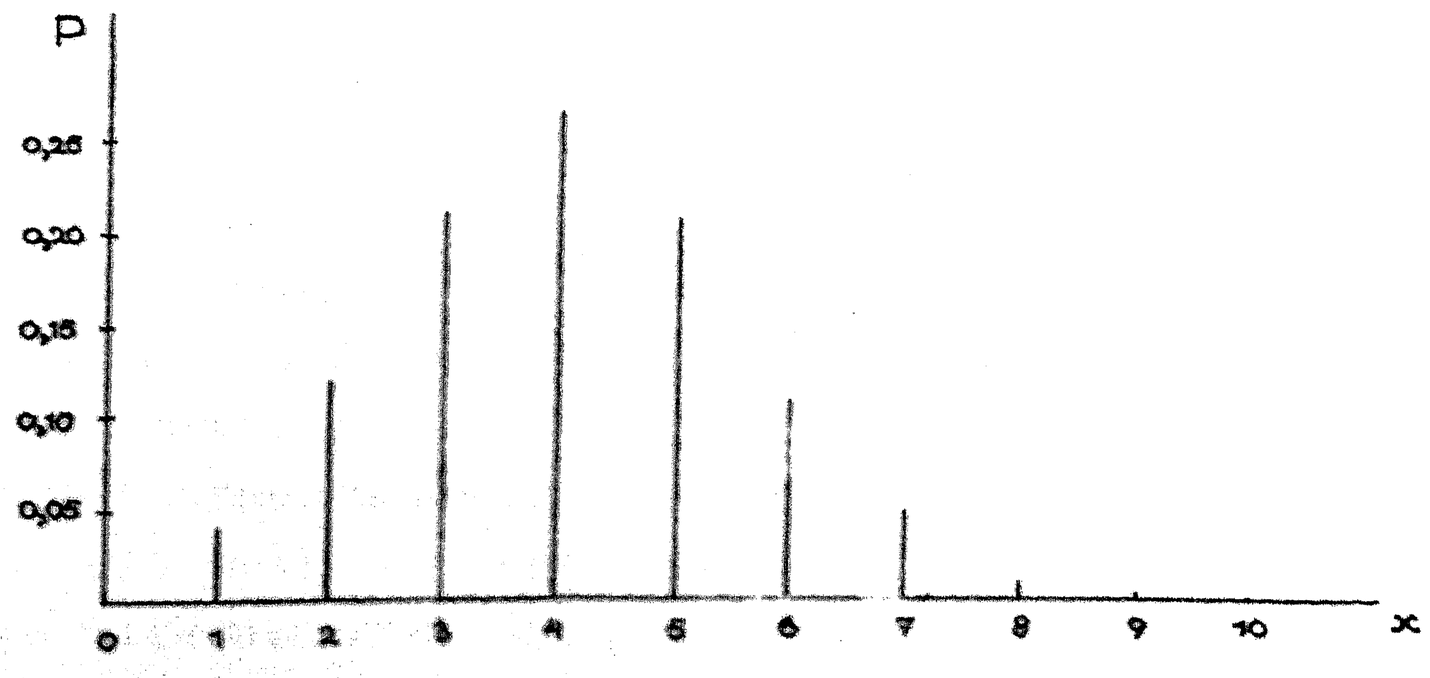
Bewijs van de stelling.

De kans, dat er een bepaalde reeks uitkomsten van de vorm $S\bar{S}\bar{S}S...S$ verkregen wordt, waaronder x successen, is gelijk aan $pqqp...p = p^x q^{n-x}$. Dit volgt direct uit de vermenigvuldigingsregel (11') voor onafhankelijke gebeurtenissen. Er zijn, volgens de hulpstelling van § 8, $\binom{n}{x}$ verschillende dergelijke reeksen. Het optreden van twee verschillende reeksen tegelijk is onmogelijk, zodat wij formule (6) kunnen toepassen. Dan blijkt, dat de kans op het verkrijgen van één van deze reeksen gelijk is aan het rechterlid van (13), maar het is tevens de kans om x successen te verkrijgen, dus gelijk aan het linkerlid van (13).

Voorbeeld. Voor $n=10$ en $p=0{,}4$ is de binomiale verdeling in onderstaande tabel en grafiek gegeven.

Binomiale verdeling voor $n=10$, $p=0{,}4$.

x	P[x=x]	x	P[x=x]
0	0,006	5	0,201
1	0,040	6	0,111
2	0,121	7	0,042
3	0,215	8	0,011
4	0,251	9	0,002
		10	0,0001



9. De theoretische wet der grote getallen.

Wij hebben de volledige afleiding van de binomiale verdeling gegeven, omdat wij deze later nog zullen gebruiken. Uit deze verdeling volgt nu een stelling, die het theoretische analogon van de experimentele wet der grote getallen is en waaruit volgt, dat wij een wh inderdaad, op de basis van het gekozen model, kunnen interpreteren als een fq in een lange reeks waarnemingen. Deze stelling zullen wij, om tijd te sparen, zonder bewijs vermelden.

Stelling 2. Beschouwen wij een reeks van n onafhankelijke experimenten, die ieder een kans p op succes bieden, en noemen wij x het aantal successen en geven wij het fq van het aantal successen aan door y:

(14) $$\underline{y} \stackrel{\text{def}}{=} \underline{x}/n,$$

dan bezit y een wh-verdeling, waarvoor geldt:

(15) $$\lim_{n\to\infty} P[|\underline{y}-p|>c] = 0$$

voor iedere c>0.

Deze stelling wordt de theoretische wet der grote getallen genoemd en beschrijft juist de experimentele wet der grote getallen in wh-theoretische terminologie, indien wij aannemen, dat gebeurtenissen, die een zeer kleine wh bezitten, niet of slechts zeer zelden optreden. Dit resultaat is, gezien de wijze, waarop de axioma's der whr aan de experimentele wet der grote getallen zijn aangepast, allerminst verwonderlijk, maar het is wel verheugend, want het geeft te kennen, dat de gekozen wijze van mathematisch beschrijven van de bedoelde verschijnselen adaequaat is.

10. Toepassingsprincipe; samenvatting.

Op grond van het bovenstaande kunnen wij nu dus een statistisch verschijnsel, d.i. een verschijnsel, dat aan de experimentele wet der grote getallen voldoet, onderzoeken met behulp van de whr. Daarbij kan een afgeleide wh geïnterpreteerd worden als een fq in een lange reeks van experimenten op grond van het toepassingsprincipe, inhoudende, dat gebeurtenissen met zeer geringe wh zich slechts zeer zelden zullen voordoen, zodat men deze gevoeglijk buiten beschouwing kan laten.

Een groot voordeel van deze opzet is, dat men de toepasbaarheid ervan experimenteel kan toetsen door statistische experimenten op te zetten, waarbij volgens het wh-theoretische model een bepaalde uitkomst een zeer kleine wh bezit en dan te verifiëren, dat deze uitkomst niet optreedt. De interpretatie van een wh als