DERDE GEDEELTE

STOCHASTISCHE STATISTIEK

I. Eendimensionale frequentieverdelingen

§ 15. *Toevallige afwijkingen*

Tot hier toe hebben wij ons uitsluitend bezig-gehouden met wat wij noemden systematische samenhang tussen verschillende grootheden. Wanneer een grootheid, nemen wij als voorbeeld het verbruik van suiker, alleen door dergelijke systematische samenhang met andere grootheden wordt bepaald, zeggen wij ook wel dat alleen systematische oorzaken of krachten [1]) daarop hebben gewerkt. Dat kunnen in het genoemde voorbeeld zijn de stand der inkomens en de prijs van suiker. Daarnaast zijn er echter gewoonlijk andere oorzaken (of krachten) die de omvang van een grootheid bepalen, die men *toevallige* oorzaken (krachten) noemt. Deze veroorzaken afwijkingen tussen de waarde van de beschouwde grootheid en de waarde die men

---

[1]) Het woord „krachten" wordt hier niet bedoeld in de physische betekenis.

op grond van de systematische oorzaken zou verwachten. Zulke afwijkingen noemt men *toevallige afwijkingen*; deze zullen het onderwerp van dit hoofdstuk vormen.

De grens die men legt tussen systematische en toevallige krachten is tot op zekere hoogte door subjectief inzicht bepaald; d. w. z. bepaald door de verhouding van den onderzoeker tot het object. Een toevallige kracht is er nl. een waarvan de uitwerking aan den onderzoeker van geval tot geval niet precies bekend is. Deze definitie is echter in vergelijking tot de in statistische vakkringen gebruikelijke terminologie wat wijd en we zullen, als dat nodig is, de krachten die aldus gedefiniëerd worden „toevallig in de wijdere zin" noemen. Gewoonlijk zullen we in het vervolg met toevallige krachten een iets enger gedefiniëerde groep bedoelen, die we voorzover nodig nog kunnen noemen „toevallig in de engere zin". Deze enger gedefiniëerde groep toevallige krachten voldoet dan nog aan de volgende voorwaarden:

a) ze zijn talrijk;

b) de uitwerking van elk afzonderlijk is klein in vergelijking tot die van alle tezamen;

c) de uitwerking van elk afzonderlijk is onafhankelijk van die der andere toevallige krachten [1]).

[1]) Voor de analyses in de sociale wetenschappen zal deze veronderstelling gewoonlijk als voldoende in overeenstemming

Op het eerste gezicht zal het wellicht lijken alsof dit belangrijke beperkingen zijn. In de werkelijkheid zal men echter krachten die niet aan deze voorwaarden voldoen en waarvan de uitwerking niettemin aan den onderzoeker onbekend is, weinig aantreffen. Wanneer het aantal der krachten waarvan de uitwerking onbekend is, in tegenstelling tot wat onder *a*) werd aangenomen, klein is, zal men ze waarschijnlijk met name kunnen noemen en hun uitwerking kunnen bestuderen. Evenzo, wanneer hun uitwerking, in contrast met het bij *b*) veronderstelde, belangrijk is.

Hetzelfde kan tenslotte worden gezegd van krachten die niet onafhankelijk zouden zijn van

---

met de waarnemingen worden gemaakt. Voor de fijnere analyses in de natuurwetenschappen moet men hier nog onderscheid maken tussen gevallen van strikte onafhankelijkheid en gevallen waarbij deze strikte onafhankelijkheid afwezig is. Als b.v. het voorkomen van een positieve uitwerking van één der krachten in kwestie medebrengt dat ook enkele andere positieve uitwerking hebben of gewoonlijk positieve uitwerking hebben, dan zal de totale uitwerking dezer krachten te zamen de neiging hebben om vaak òf vrij groot positief òf vrij groot negatief te zijn. Daardoor zal, zoals we nog nader zullen zien, de frekwentieverdeling lichtelijk andere vormen gaan aannemen dan in het geval van zuivere onafhankelijkheid. Toch zou men, wanneer het verschijnsel in kwestie binnen zekere perken blijft, nog wel van toevallige krachten kunnen spreken. Dat is dan een kwestie van liefhebberij.

de andere. Deze zouden òf, tezamen met deze andere, een wèl belangrijke uitwerking kunnen hebben òf, als dat niet het geval is, bij de andere kunnen worden ondergebracht, waardoor ze als afzonderlijke kracht niet behoeven te worden genoemd. In 't algemeen zal men daarom inderdaad kunnen beweren dat de krachten waarvan de uitwerking aan de onderzoeker niet bekend is, ook voldoen aan de drie gestelde criteria, waarbij men niet vergete dat het bij alle practische onderzoekingen om benaderingen gaat en de woorden groot en klein, belangrijk en onbelangrijk, veel speling toelaten. Het behoeft echter niet altijd zo te zijn dat de krachten met onbekende uitwerking tevens aan de drie gestelde criteria voldoen. Daarnaast volgt uit het subjectieve karakter van het begrip „onbekend" ook dat verdergaand onderzoek er herhaaldelijk toe kan leiden dat een aanvankelijk onder de toevallige krachten gerangschikte kracht later tot de systematische kan worden gerekend. In het genoemde voorbeeld van het verbruik van suiker zal men als toevallige krachten de volgende oorzaken van afwijkingen kunnen noemen: afwijkingen in de aanvoer van vruchten waarbij suiker wordt gebruikt, afwijkingen in de voorraden van tussenhandelaren, invloed van feestdagen, enz.

Tenslotte worde nog even opgemerkt, dat het feit dat de uitwerking van een toevallige

kracht aan den onderzoeker onbekend is, vaak een gevolg is van de omstandigheid dat men de richting van de kracht zelve niet kent (niet dat men, als die richting bekend ware, niet zou weten in welke richting de daardoor veroorzaakte afwijking zou liggen). Bij het voorbeeld van de suiker: men weet niet of in een gegeven maand of jaar er een extra grote of extra kleine aanvoer van vruchten is geweest.

Het werken met het begrip toevallige krachten is vaak, vooral in de sociale wetenschappen, te beschouwen als een *tijdelijk* hulpmiddel bij het onderzoek; tijdelijk in deze zin dat men als uiteindelijk ideaal van het onderzoek toch ziet de verklaring van alle bewegingen van de beschouwde grootheid uit systematische krachten [1]). Tevens is het een zekere *stylering*, in

[1]) Dit is niet het geval wanneer de „micro-structuur" en de „macro-structuur" een principiëel verschil vertonen, of de „micro-structuur" ontoegankelijk is. Dit is vrijwel het geval bij de meeste toevalsspelen, zoals b.v. het werpen met dobbelstenen, waar het vrijwel ondoenlijk is om van elk geval van werpen na te gaan hoe de nauwkeurige stand van hand, dobbelstenen, enz. is geweest aan het begin (waaruit in theorie de uitslag van de worp zou zijn te vinden).

We raken hier aan problemen die voor de natuurwetenschappen en dus voor de wereldbeschouwing van groot belang zijn, doch buiten het raam van dit werkje vallen, nl. of het hierboven gezegde over het hulpkarakter van het begrip toevallige kracht ook voor de natuurwetenschappen geldt. Het schijnt dat de nieuwste ontwikkeling dezer wetenschappen tot het tegengesteld inzicht leidt. Voor een zeer duidelijke elemen-

zoverre dat de eisen waaraan toevallige krachten in de engere zin moeten voldoen zelden geheel en al vervuld zijn. Deze stylering komt het scherpst en meest exact uit bij het geval dat men als de basis van bijna alle vraagstukken van theoretische statistiek (zie § 1) kan beschouwen, nl. dat van het *toevalsspel*. Aan dit toevalsspel zullen we daarom, zoals gebruikelijk is, enige nadere aandacht schenken.

§ 16. *Toevalsspelen*

Bij toevalsspelen treedt nl. zeer duidelijk de hoofdeigenschap der toevallige krachten naar voren, t.w. dat de toeschouwer van te voren niet weet wat het resultaat zal zijn van een „worp" of een „trekking".

Denkt men zich, om de gedachten te bepalen, het geliefkoosde speelgoed der theoretisch-statistici, nl. een loterijtrommel (in de taal van veel statistiekleerboeken en boeken over waarschijnlijkheidstheorie heet dat een „urn" of een „vaas"!) waarin zich 10 rode en 10 witte papiertjes bevinden. Daaruit mag eenmaal getrokken worden; het aantal malen wit dat men trekt, kan dan zijn 0 of 1. Wat men nu nog weet is alleen, *welke mogelijkheden er zijn* en,

---

taire uiteenzetting zij verwezen naar Prof. Dr. Ph. Kohnstamm, Het ongeloof en de natuurwetenschap (No. 6 uit de serie geschriften Psychologie van het Ongeloof, Amsterdam 1935); speciaal § 3—7.

in een aantal gevallen die ons het meest interesseren, *hoe vaak elk dezer gevallen zich kan voordoen.* Wanneer men nl. alle achtereenvolgende trekkingen verricht welke mogelijk zijn zonder terugleggen van de getrokken papiertjes, is het duidelijk dat men in het geheel tien maal een rood en tien maal een wit papiertje moet trekken.

Niet in alle soorten van dergelijke spelen die men kan construeren is het zo gesteld dat men a priori weet hoe vaak elke mogelijkheid kan voorkomen. Wanneer men nl. trekkingen verricht „met terugleggen", d. w. z. na elke trekking het getrokken briefje weer in de trommel werpt en opnieuw trekt („na schudden"), is er a priori geen zekerheid daaromtrent. Wel leert *het experiment* dat als men zeer vaak een trekking verricht, er een bepaalde verhouding tussen rode en witte papiertjes als resultaat steeds meer benaderd wordt. En deze verhouding is dezelfde als bij het spel zonder terugleggen, nl. één. Men zou kunnen beweren dat er *geen enkele reden* is om aan te nemen dat in dit geval de rode trekkingen talrijker zouden zijn dan de witte of omgekeerd, waardoor sommige schrijvers de stelling ook bewezen achten. Over de kracht van deze argumentatie lopen de meningen uiteen. In 't algemeen zal men echter kunnen beweren dat het aangeven a priori van welke gebeurtenissen even waarschijnlijk zijn de taak is van een *theorie* betreffende

het object in kwestie, welke theorie als *gegeven* optreedt in hetgeen nu volgt: de waarschijnlijkheidsrekening.

In de *waarschijnlijkheidsleer* (of -rekening) spreekt men nu in 't algemeen van *universa* (populaties, massa's) (de 10 witte en 10 rode papiertjes in ons eerste geval; de gezamenlijke uitkomsten van alle trekkingen in ons tweede geval), waarin men zekere *kenmerken* onderscheidt (wit en rood in onze gevallen), van welke kenmerken men de *frequentieverdeling* bestudeert. Dat bestuderen komt, naar de opvatting van de bekende waarschijnlijkheidstheoreticus R. v. Mises, neer op het volgende: in bepaalde eenvoudige gevallen is deze verdeling gegeven — zoals in de beide door ons behandelde voorbeelden —; de problemen der waarschijnlijkheidsleer bestaan nu in het berekenen van dergelijke frequentieverdelingen in ingewikkelder gevallen, die uit de eenvoudige op een gegeven manier worden afgeleid.

Het belangrijkste voorbeeld dat voor ons betoog hiervan zal worden besproken is het volgende. Als er in de trommel een zeer groot aantal N witte en N rode briefjes zijn en we trekken daaruit telkens opnieuw zonder terugleggen n briefjes, hoeveel witte zullen daarbij dan kunnen zijn en hoe vaak zal elk dezer mogelijkheden voorkomen (dit laatste relatief genomen t.o.v. elkaar)?

De eerste vraag is gemakkelijk te beantwoorden. Het aantal witte briefjes (kortweg te noemen: *het resultaat* van de trekking) zal elk geheel getal van 0 tot en met n kunnen zijn. Om de tweede vraag te kunnen beantwoorden, moeten we een overzicht maken van alle denkbare groepen van n briefjes die we uit de $2N$ aanwezige van de „totale populatie" kunnen formeren. Om dit te bereiken, vragen we ons eerst af, op hoeveel verschillende manieren we één wit papiertje uit de $2N$ kunnen trekken. Daartoe zijn blijkbaar N mogelijkheden: het 1e, het 2e, het 3e .... enz. tot en met het Nde witte briefje kan door ons worden genomen. Op hoeveel wijzen kunnen we er twee trekken? Bij elk der N mogelijkheden, zojuist genoemd, hebben we $N - 1$ mogelijkheden om een ander wit briefje te trekken, nl. een der nog overgebleven. Er zijn dus

$$N(N-1)$$

mogelijkheden om 2 witte briefjes te trekken; maar daarbij hebben we elk paar tweemaal geteld: b.v. hebben we het paar No. 17, No. 3, dan hebben we dat eenmaal bereikt door te beginnen met No. 17 en daar No. 3 bij te nemen; de andere maal door met No. 3 te beginnen en daar No. 17 bij te nemen. Er zijn dus slechts

$$\frac{N(N-1)}{2}$$

verschillende tweetallen te vormen.

Op een volkomen analoge wijze berekent men dat het aantal drietallen dat men kan trekken uit N exemplaren bedraagt

$$\frac{N(N-1)(N-2)}{1.2.3}$$

en het aantal k-tallen

$$\frac{N(N-1)\ldots\ldots(N-k+1)}{1.2.3\ldots\ldots k}.$$

Gebruiken we de bekende schrijfwijze

$k!$ („k faculteit") $= 1.2.3.\ldots\ldots k$,

dan laat zich de laatste uitdrukking schrijven

$$\frac{N!}{k!\,(N-k)!}$$

Na deze voorbereidende opmerkingen kunnen we de vraag beantwoorden, op hoeveel verschillende wijzen zich een combinatie van k witte en $n-k$ rode briefjes laat verkrijgen. De k witte briefjes kunnen, zoals we zagen, op

$$\frac{N!}{k!\,(N-k)!}$$

wijzen worden verkregen; bij elk dezer gevallen laten zich dan $n-k$ rode briefjes op

$\dfrac{N!}{(n-k)!\,(N-n+k)!}$ wijzen verkrijgen; in

totaal hebben we dan

$$\frac{(N!)^2}{k!\,(N-k)!\,(n-k)!\,(N-n+k)!} \text{ gevallen, of}$$

uitgeschreven:

$$\frac{N(N-1)\ldots(N-k+1)\,N(N-1)\ldots(N-n+k+1)}{k!\,(n-k)!}$$

Nu interesseren we ons alleen voor de relatieve frequentie van elk dezer gevallen; bovendien willen we, zoals reeds opgemerkt, aannemen, dat N zeer groot is.

Het aantal factoren in de teller is nu, onafhankelijk van k, gelijk aan n. De verhouding der frequenties voor de gevallen $k = 0 \ldots . n$ wordt niet gewijzigd door de frequenties te delen door N^n; daardoor wordt de teller $1\,(1-\frac{1}{N})\,(1-\frac{2}{N})\ldots.(1-\frac{k-1}{N})\,1\,(1-\frac{1}{N})\ldots.$ $(1-\frac{n-k-1}{N})$ hetgeen, voor N zeer groot, eenvoudig niet te onderscheiden is van 1. De relatieve frequenties worden daardoor evenredig met:

$$\frac{1}{k!\,(n-k)!}, \qquad k = 0 \ldots . n$$

wat we tot een reeks gehele getallen kunnen maken (als we dat wensen) door vermenigvuldiging met $n!$, waardoor alweer de verhouding der frequenties zich niet wijzigt. De

getallen die aldus worden verkregen, worden vaak met het afkortende symbool

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

aangeduid en *binomiale coëfficiënten* genoemd. Het zijn nl. de coëfficiënten die men verkrijgt door het uitvermenigvuldigen van de n-de macht van een tweeterm (een binomium). Dit levert nl. op:

$$(a+b)^n = a^n + \binom{n}{1} a^{n-1}b + \binom{n}{2} a^{n-2}b^2 + \binom{n}{3} a^{n-3}b^3 \ldots + \binom{n}{n-1} a\, b^{n-1} + \binom{n}{n} b^n.$$

(formule van Newton).

Er bestaat een gemakkelijke wijze om deze coëfficiënten in getallenwaarde te verkrijgen, nl. door de zgn. driehoek van Pascal op te schrijven, welke hieronder voor de gevallen tot en met $n = 8$ is afgedrukt:

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1

De eigenaardigheid van deze getallenfiguur is, dat elk getal de som is van de twee getallen

die er boven staan. In de zojuist ingevoerde schrijfwijze is de onderste reeks getallen

$\binom{8}{0}\ \binom{8}{1}\ \binom{8}{2}\ \binom{8}{3}\ \binom{8}{4}\ \binom{8}{5}\ \binom{8}{6}\ \binom{8}{7}\ \binom{8}{8}$, zoals men gemakkelijk narekent. In het algemeen is de n-de rij getallen de rij

$$\binom{n}{0}\ \binom{n}{1}\ \ldots\ldots\ \ldots\ldots\ \binom{n}{n-1}\ \binom{n}{n}$$

Daarbij is de definitie van $\binom{n}{0}$ gekozen in aansluiting met de regelmatigheden in de binomiale coëfficiënten voor $k \geqq 1$: $\binom{n}{0} = \binom{n}{n} = 1$.

Men kan gemakkelijk inzien, dat de binomiale coëfficiënten inderdaad op de aangeduide wijze kunnen worden berekend, d.w.z. als som van de twee er boven staande coëfficiënten. Ten eerste door het eenvoudig door narekenen te controleren voor het algemene geval:

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

Ten tweede echter ook door de ontstaanswijze bij het uitvermenigvuldigen van $(a + b)^n$ na te gaan. Denkt men zich dat men $(a + b)^{n-1}$ reeds berekend heeft en daarvoor de vorm

$$A_1 a^{n-1} + A_2 a^{n-2} b + \ldots\ldots + A_k a^{n-k} b^{k-1} + A_{k+1} a^{n-1-k} b^k \ldots\ldots + A_n b^{n-1}$$

heeft verkregen. Deze moet nu met $a + b$ worden vermenigvuldigd om $(a + b)^n$ te verkrijgen en men vraagt zich af, welke coëfficiënt in de laatste uitdrukking de term met $a^{n-k} b^k$ krijgt. Deze term kan slechts op twee manieren

ontstaan, n.l. door $A_k a^{n-k} b^{k-1}$ met b te vermenigvuldigen en door $A_{k+1} a^{n-1-k} b^k$ met a te vermenigvuldigen. Dat levert de coëfficiënt $A_k + A_{k+1}$ op. Nu is echter A_k niets anders

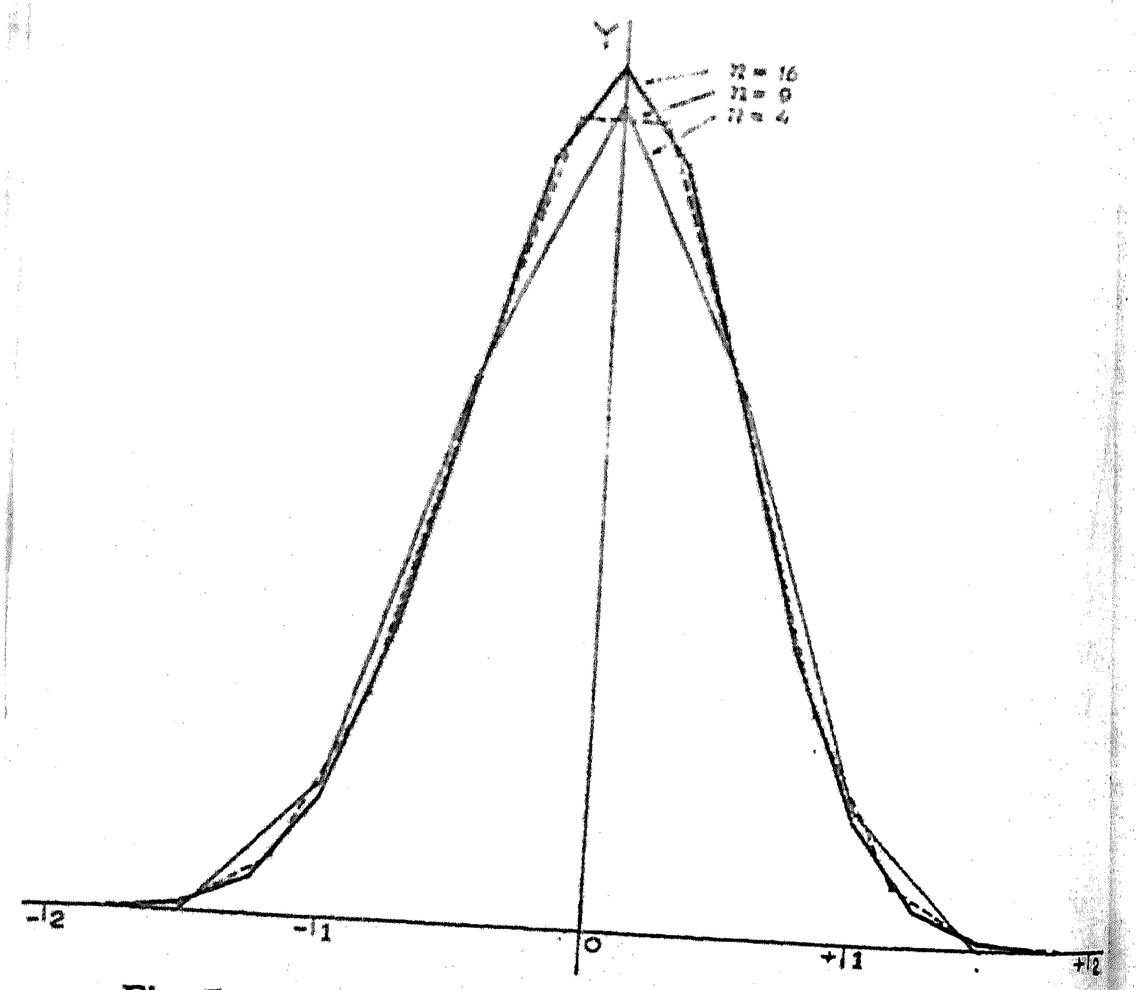

Fig. 5. Drie achtereenvolgende benaderingen van de curve van Gauss: binomische verdelingen voor $n = 4$, 9 en 16, bij constante standaarddeviatie.

dan $\binom{n-1}{k-1}$ en A_{k+1}: $\binom{n-1}{k}$, terwijl de nieuw verkregen coëfficiënt $\binom{n}{k}$ is. Inderdaad is dus

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

De binomiale coëfficiënten hebben nog vele

verdere belangwekkende eigenschappen, waarover echter in dit werkje verder niet kan worden gesproken. Slechts één eigenschap, reeds uit de getallenvoorbeelden gebleken, is van zodanig belang, dat we hem hier afzonderlijk uitspreken en dat is de symmetrie-eigenschap

$$\binom{n}{k} = \binom{n}{n-k}$$

welke direct volgt uit de definitie en dus betekent dat de coëfficiënten van één rij in de driehoek van Pascal in omgekeerde volgorde gelezen dezelfde zijn als in de natuurlijke volgorde. Dit betekent dat de *frequentieverdeling van onze trekkingsuitkomsten grafisch wordt voorgesteld door een symmetrische lijn* (zie fig. 5). Er volgt ook onmiddellijk uit, dat de mediaan, de modus en het rekenkundig gemiddelde van deze frequentieverdeling samenvallen.

§ 17. *De wet van Gauss; andere verdelingswetten*

Denken we ons nu het aantal trekkingen n bij onze loterij, dat in één steekproef wordt verricht, steeds groter wordend. Dan wordt dus ook het aantal $n + 1$ der mogelijke uitkomsten steeds groter en in onze frequentietabel dus het aantal klassen. Denken we ons daarbij tevens de „breedte" der klassen afnemend en wel evenredig met $\frac{1}{\sqrt{n}}$. Dat zou betekenen dat de „waarde" van één wit briefje in die ver-

houding zou worden verkleind. In de sfeer der reële toepassingen van deze theorie betekent het, dat de afwijking, die één der toevallige krachten op zich zelf kan veroorzaken, kleiner wordt in diezelfde verhouding. Denken we ons nu verder ook de *frequenties berekend per eenheidsinterval*. Dat wil zeggen, dat we een frequentie behorende bij een klassebreedte of klasseïnterval van $^1/_4$ met 4 vermenigvuldigen; een frequentie berekend bij een klasseïnterval van $^1/_3$ met 3 vermenigvuldigen, enz. Wanneer we nu van deze „genormaliseerde frequenties" of „frequenties per eenheidsinterval" een grafiek maken voor verschillende waarden van n, dan zullen we bemerken, dat deze grafieken meer en meer naderen tot één bepaalde vloeiende kromme lijn (Vgl. fig. 5). Men noemt deze lijn de *curve van Gauss*, of van *Laplace*, ook wel naar een bekend statisticus, die haar veel gebruikte, de curve van *Quetelet*. De verdeling, die er mee correspondeert, noemt men de *verdeling van Gauss* of de *normale verdeling* of ook wel de verdeling volgens de *foutenwet*. De vergelijking voor de *relatieve frequentie y per eenheidsinterval* luidt:

$$y = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}} \qquad (17.1)$$

waarbij x de *afwijking van het gemiddelde resultaat* betekent in de eenheden die men heeft aan-

genomen. Dit betekent dan, dat in een interval dat niet, in de aangenomen eenheden, **één** bedraagt, maar dx, de totale relatieve frequentie bedraagt:

$$ydx = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}}\, dx$$

De grootheid σ is, zoals men gemakkelijk kan narekenen, de *standaarddeviatie* van deze verdeling; immers, daar

$$\int_{-\infty}^{+\infty} e^{-\xi^2}\, d\xi = \sqrt{\frac{\pi}{2}}$$

hebben wij voor het kwadraat van de standaarddeviatie van de verdeling (17.1):

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2\, e^{-\frac{x^2}{2\sigma^2}}\, dx = \sigma^2 \qquad (17.2)$$

Het zal duidelijk zijn dat, als men de eenheden waarin men x meet verandert, ook σ verandert en wel omgekeerd evenredig, terwijl het getal x dat een vaste afstand weergeeft evenredig met σ verandert. Men ziet aan de de formule, dat y dan in dezelfde verhouding verandert als die eenheid, hetgeen klopt met de definitie van y als frequentie over een eenheidsinterval.

De afleiding van de formule voor y, als grens geval voor $n \rightarrow \infty$ kan worden gevonden b.v. bij Davis & Nelson, Elements of Statistics.

Een tabel aangevende de waarden

$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} t^2}$, d.w.z. voor het geval dus dat de standaarddeviatie 1 is, vindt men b.v. bij Davis & Nelson, Elements of Statistics, blz. 397. Volgens de hier geschetste ontstaanswijze treedt dus een verdeling van Gauss op, wanneer de waarde van een bepaalde grootheid behalve door systematische krachten (die voor alle beschouwde gevallen als identiek worden aangenomen) wordt bepaald *door een groot aantal krachten die elk op zich zelf een kleine afwijking, even vaak naar links of naar rechts, kunnen veroorzaken, zonder dat de uitwerking van de ene invloed heeft op die van de andere.* Zoals men ziet, is dit een bijzonder geval van wat wij in § 15 toevallige krachten hebben genoemd.

Het laat zich bewijzen dat de verdelingswet van Gauss ook optreedt, wanneer de krachten die de kleine afwijkingen veroorzaken anders zijn. Nl. wanneer niet slechts één bepaalde positieve of één bepaalde negatieve afwijking kan voorkomen, doch een gehele frequentieverdeling van mogelijkheden bestaat. Ook mag wel één kracht in vergelijking tot de overige

grote afwijkingen veroorzaken, mits de afwijkingen, door deze ene kracht veroorzaakt, dan zelf optreden volgens de frequentieverdeling van Gauss. Een en ander vindt men exact behandeld o.a. bij Darmois, Statistique mathématique, speciaal blz. 234/236 en wat daaraan voorafgaat.

Essentiëel voor de genese van de verdeling van Gauss is echter, dat de uitwerking van elk der krachten niet afhangt van de uitwerking der overige krachten of van de volgorde waarin ze optreden. In zulke gevallen komen min of meer belangrijke afwijkingen van de wet van Gauss voor. Met de ruwe benadering die voorlopig in de sociale wetenschappen nog noodzakelijk is, kan men zeggen dat er een aantal verschijnselen zijn waarbij de verdeling van Gauss optreedt.

Voor een discussie van de kwestie in hoeverre in de biologie de normale verdeling een rol speelt vgl. men Prof. Dr. L. G. M. Baas Becking en E. F. Drion, On the origin of frequency distributions in biology, Acta Biotheoretica, Leiden 1936, pag. 133.

Een aantal economische en sociale verschijnselen vertoont een enigszins *scheve frequentieverdeling*. Een hypothese waarmee men deze verdelingen verklaren kan is, dat de uitwerking van k in de als „gunstig" gekozen richting werkende krachten niet evenredig is met k,

maar van de vorm ma^k is, waarbij a een voor elk object andere constante kan zijn en m de ongestoorde waarde is. Men kan dan zeggen dat de logarithmen van de uitwerkingen met k evenredig zijn en de frequentieverdeling wordt wederom normaal, als men de logarithmen neemt van de afwijkingen. *Gibrat* noemt de wet die deze verdelingen tot stand brengt „la loi de l'effet proportionnel" [1]); *Van der Wijk* paste deze gedachte onafhankelijk toe op de inkomensverdeling [2]).

Scheve verdelingen kunnen echter natuurlijk ook op andere, meer algemene, wijze ontstaan. Een eenvoudig voorbeeld krijgt men reeds bij de binomiale verdeling, zodra daarbij het aantal witte en rode briefjes (§ 16) in het universum ongelijk wordt aangenomen. Dan wordt de frequentieverdeling niet voorgesteld, zoals in § 16, door de termen van $(a + b)^n$ voor $a = b$, maar voor $a \neq b$, hetgeen tot een scheve curve aanleiding geeft, zoals uit onderstaande tabel blijkt.

Termen van de ontwikkeling van $(a + b)^n$ voor $a = 1$, $b = 2$ en $n = 11$:

Termnumm.	Term	Termnumm.	Term
1	1	12	1 × 2048 = 2048
2	11 × 2 = 22	11	11 × 1024 = 11264
3	55 × 4 = 220	10	55 × 512 = 28160
4	165 × 8 = 1320	9	165 × 256 = 42240
5	330 × 16 = 5280	8	330 × 128 = 42240
6	462 × 32 = 14784	7	462 × 64 = 29568

[1]) Vgl. R. Gibrat, Les inégalités économiques, Parijs 1931.
[2]) T. a. p.

Zoals men ziet, ligt de top niet in het midden, doch tussen de 8e en 9e term. Bovendien vindt de daling naar de zijde van de lagere termnummers langzamer plaats dan naar de zijde van de hogere termnummers.

Behalve afwijkingen van de wet van Gauss door scheefheid bestaan er eveneens afwijkingen in de vorm van een *onder-* of een *bovennormale spreiding*.

Een zeer eenvoudig voorbeeld van de laatste is de frequentieverdeling van de uitkomsten van een toevalsspel met één munt. Als men dan vraagt naar het aantal malen dat b.v. „kruis" wordt geworpen, is het antwoord: er zijn slechts 2 uitkomsten mogelijk, 0 of 1, en elk van beide is even waarschijnlijk. De frequentieverdeling heeft dan een zeer eenvoudige gedaante: de frequentie is bij 0 en 1 een even groot, van nul verschillend, getal en bij alle andere getallen: nul. In dit geval is de frequentieverdeling discontinu; men kan echter ook zeer goed gevallen construeren, waarbij zij continu is en dezelfde eigenschap vertoont. Daartoe is reeds voldoende dat de vele kleine krachten die wij reeds eerder beschouwden niet een van elkaars uitwerking onafhankelijke afwijking vertonen, doch waarbij enkele dier krachten een positieve afhankelijkheid vertonen, d.w.z. dat een gegeven afwijking, door de ene veroorzaakt, vaak een in dezelfde richting vallende afwijking

door de andere ten gevolge heeft. Zo vertoont b.v. het sterftecijfer een bovennormale spreiding als gevolg van het optreden van besmettelijke ziekten. De dood van sommigen sleept dan de dood van anderen mede.

Een *ondernormale spreiding* zal optreden wanneer een negatieve afhankelijkheid optreedt, dus wanneer een afwijking, door een der krachten veroorzaakt, vaak meebrengt een tegengesteld gerichte afwijking door een andere. Een extreem voorbeeld daarvan is aanwezig, wanneer men met twee munten werpt waarvan de ene aan beide zijden een muntteken en de andere aan beide zijden een kruisteken heeft. Dan zal noodzakelijk bij elke keer dat er een kruis boven komt ook een munt bovenkomen. Er is in dit geval in 't geheel geen spreiding zelfs. Men kan ook tussengevallen construeren door te werpen met 3 munten waarvan er 2 als zojuist aangeduid zijn en één normaal is.

Voor een formeel-systematische behandeling der van de curve van Gauss afwijkende curven zij nog verwezen naar Jones, A First Course in Statistics, Ch. XV, of naar Kelley, Statistical Methods (de zgn. *curven van Pearson*).

§ 18. *Toepassing van waarschijnlijkheidstheoretische frequentieverdelingen*

Ons betoog vertoonde tot hier toe deze

gang, dat wij een universum van trekkingen beschouwden, toegelicht aan een speciaal geval, en daarvan de frequentieverdeling over verschillende mogelijkheden voor het resultaat beschouwden. Nu komt dan evenwel de stap die de overgang tot de praktische toepassingen vormt en over de toelaatbaarheid waarvan in vakkringen afwijkende meningen naast elkaar bestaan. We gaan nu aan het voorspellen. Wanneer iemand een willekeurig aantal Z keren een trekking van n briefjes verricht uit de boven ingevoerde trommel, dan voorspellen we, dat de frequentieverdeling van de verschillende uitkomsten die hij daarbij verkrijgt, benaderd wordt door de frequentieverdeling die we vroeger voor *het gehele universum* van alle denkbare steekproeven van n objecten verkregen. Laten we het aantal steekproeven dat zich in dat universum bevindt met M aangeven, en de frequenties van elk der mogelijke uitkomsten $F_0, F_1, F_2 \ldots . F_n$, dan voorspellen we dat de frequentieverdeling die gevonden wordt bij Z proeven zal zijn

$$\frac{F_0}{M} Z, \quad \frac{F_1}{M} Z \ldots . \frac{F_n}{M} Z$$

of, als we weer het begrip waarschijnlijkheid invoeren en met de letter w aangeven (w_0, $w_1, \ldots . w_n$):

$$w_0 Z \quad w_1 Z \ldots . w_n Z.$$

Daarbij moet dan natuurlijk op gehele getallen worden afgerond.

Deze stap: het overgaan tot voorspellingen bij reële gebeurtenissen wordt door sommige waarschijnlijkheidstheoretici reeds vroeger in het betoog gedaan. Ik heb er de voorkeur aan gegeven de berekeningen eerst te geven en de kritische stap het laatst te doen. De uitkomsten zijn natuurlijk dezelfde, doch de formulering der praemissen is wat anders.

Het experiment vertoont nu veelal wel is waar een ruwe benadering van de aangegeven frequenties, doch gewoonlijk geen exacte. De vraag is, of daarmee de uitgesproken voorspelling en de daarin belichaamde theorie weerlegd moet worden geacht. Om twee redenen is dat naar de gangbare opvatting niet zo. In de eerste plaats omdat stilzwijgend is aangenomen — wat moeilijk te controleren is — dat er alvorens de trekkingen verricht worden een zekere symmetrie-eis vervuld is. B.v. dat de rode en de witte briefjes goed dooreen in de trommel liggen. Komt in een concreet geval de voorspelling niet uit, dan kan men zich natuurlijk steeds redden met de opmerking dat aan deze eis niet volkomen was voldaan. Dat kan men intussen door verfijning der experimenten („goed schudden", enz.) bereiken.

De tweede reden waarom een niet-uitkomende voorspelling geen ongeluk voor de aangegeven

theorie is, is gelegen in de omstandigheid dat ook de theorie die afwijkingen als mogelijk ziet en van het optreden daarvan ook weer de waarschijnlijkheid kan berekenen. Daartoe gaat zij over tot het beschouwen van een ingewikkelder universum, nl. tot het universum van alle denkbare combinaties van Z steekproeven die uit het oorspronkelijke universum kunnen worden getrokken. Het principe van de behandeling zal den lezer duidelijk zijn; de behandeling in concreto moet worden uitgesteld (§ 25). De theorie kan zich zodoende steeds formeel redden. En de beoordeling van het al dan niet bruikbaar zijn van de theorie wordt enigszins subjectief. Het objectief waardevolle ligt m. i. in de berekeningen over de universa van ingewikkelder soort, die, zoals hierboven ook gedaan is, kunnen worden gegeven voordat men ook maar enig voorspellingselement heeft geïntroduceerd. Overigens is het een communis opinio dat ook voor de bovenbedoelde voorspellingsdoeleinden de theorie practisch zeer goed bruikbaar is.

Bij de wijze van voorstellen die wij hebben gekozen staat één ding vast: wanneer men trekkingen zonder terugleggen doet en daarbij het universum eenvoudig uitput, dan geldt voor die serie trekkingen de voorspelling exact. *Wanneer men dus een toenemend aantal steekproeven doet, zal men* (in dit laatste geval) *een*

nadering tot de voorspelde uitkomst zien optreden. Deze uitspraak duidt men wel eens aan met de *wet der grote getallen.* Men neemt ze ook gewoonlijk aan voor gevallen van trekkingen met terugleggen [1]).

Met de wet van de grote getallen geeft men ook vaak een andere stelling aan, die hieronder bij de verdere bespreking der steekproefvraagstukken nog nader ter sprake zal komen.

Een eerste, betrekkelijk elementaire, toepassing van deze wet bestaat hierin dat men een zo groot mogelijk materiaal bezigt bij zijn onderzoekingen, hetgeen inderdaad, binnen de grenzen die later zullen worden aangegeven, goed is. Niet zo goed is echter, wanneer men vergroting van het materiaal tracht te verkrijgen ten koste van de kwaliteit er van. Dit kan zeer gevaarlijk zijn.

Is eenmaal een bepaalde frequentieverdeling voor de mogelijkheden vastgesteld, dan kan men daarmee ook een verdere, vaak optredende vraag, beantwoorden. In het door ons gekozen voorbeeld kan men nl. vragen: wat is de waarschijnlijkheid dat de uitkomst ligt boven een gegeven getal m? We hebben dan eenvoudig een sommatie — of bij het werken met een continue curve een integratie — uit te voeren.

---

[1]) Vgl. R. v. Mises, Wahrscheinlichkeit, Statistik und Wahrheit, blz. 79.

In het discontinue geval van de binominale verdeling luidt het antwoord:

$$\frac{1}{2^n}\left\{\binom{n}{m+1}+\binom{n}{m+2}+\cdots\binom{n}{n-1}+\binom{n}{n}\right\}$$

Daarbij is de vorm { } de totale frequentie van alle gevallen waarin de uitkomst hoger dan m is, en waarbij de totale frequentie van alle mogelijkheden is 2^n. Daarom is door 2^n gedeeld.

In het continue geval is de uitkomst

$$\frac{1}{\sigma\sqrt{2\pi}}\int_m^{+\infty} e^{-\frac{x^2}{2\sigma^2}}\,dx$$

Men bedenke dat het getal m nu een andere betekenis heeft dan bij het discontinue geval.

Tabellen van het (halve) oppervlak als functie van t $\left(=\frac{x}{\sigma}\right)$ vindt men o.a. bij Davis & Nelson, Elements of Statistics, blz. 398; voor het gehele oppervlak buiten een bepaald interval $-t$ tot $+t$ (bij dezen schrijver door x aangegeven) bij Fisher, Statistical Methods for Research Workers, blz. 79.

De praktische toepassingen van de wet van Gauss bestaan in de eerste plaats hierin, dat men van een gegeven verdeling vaststellen kan, of deze toegeschreven kan worden aan toevallige

krachten dan wel niet. Daartoe moet deze verdeling dan met die van Gauss voldoende overeenstemmen. De vraag of de overeenstemming voldoende is kan men op een betrekkelijk primitieve wijze beantwoorden door te trachten een curve van Gauss zo goed mogelijk aan te sluiten bij de gegevens, b.v. door het gemiddelde en de standaarddeviatie van de gegeven verdeling te becijferen en vervolgens een curve van Gauss te tekenen die als gemiddelde en standaarddeviatie dezelfde getallen vertoont. De beoordeling van de mate van overeenstemming kan dan „op het gezicht" geschieden en heeft uiteraard een subjectief karakter. Als een eerste verkenning van het terrein kan dit echter vaak reeds voldoende opleveren.

Een meer objectieve maatstaf verkrijgt men natuurlijk, wanneer men de waarschijnlijkheid berekent dat de waargenomen verdeling door een toevalsspel van gegeven karakter ontstaan is. Dit is echter een vraagstuk dat wat zijn gecompliceerdheid betreft, de grenzen van dit boekje overschrijdt. Verwezen worde b.v. naar de behandeling er van bij Derksen, Inleiding tot de Correlatierekening, § 27.

De gegeven frequentieverdeling kan b.v. zijn de frequentieverdeling van de afwijkingen die een historische reeks vertoont van een zgn. voortschrijdend gemiddelde. Men gebruikt nl. het voortschrijdend gemiddelde vaak als een

methode om de toevallige afwijkingen uit te schakelen [1]). Strikt genomen zou een dergelijke contrôle echter moeten worden ingesteld, d.w.z. zou men moeten nagaan of de aldus uitgeschakelde bedragen inderdaad verdeeld zijn volgens de wet van Gauss. Men zou o. m. op grond van dergelijke analyses kunnen besluiten welke soort van voortschrijdend gemiddelde voor de gegeven reeks de beste is. Voor een zeer uitvoerige behandeling van deze vraagstukken zie men Anderson, Die Korrelationsrechnung in der Konjunkturforschung, Bonn 1929.

Vaak wordt gebruik gemaakt van de eigenschap der normale verdeling dat praktisch geen gevallen liggen op een afstand van meer dan 3σ ter weerszijden van het gemiddelde. Inderdaad is de waarschijnlijkheid voor een afwijking op grotere afstand slechts ongeveer 0.1%. Een voorbeeld van een dergelijke toepassing is te vinden bij de periodieke keuring van grondstoffen of producten in industriële bedrijven. Als de kwaliteit van een stof in een getal is uit te drukken (b.v. een vetgehalte, een hardheid, een brandduur), dan kan men voor elke keuring een punt in een grafiek tekenen (fig. 6) waarvan de ordinaat bedoeld getal is. Na enige tijd laat

[1]) Zie b.v. M. J. de Bosch Kemper, Inleiding tot de analyse van conjunctuurlijnen, Ec. en Soc. Kroniek 3e kwartaal 1929.

zich van de frequentieverdeling het gemiddelde en de standaarddeviatie berekenen. Door nu op een afstand van $\pm$ 3σ van het gemiddelde horizontale lijnen te trekken verkrijgt men een strook waar binnen de volgende waarnemingspunten moeten liggen, wil alleen door toevallige krachten variatie intreden. Zodra een of meer punten buiten deze strook komen te liggen, is de waarschijnlijkheid groot dat een systematische

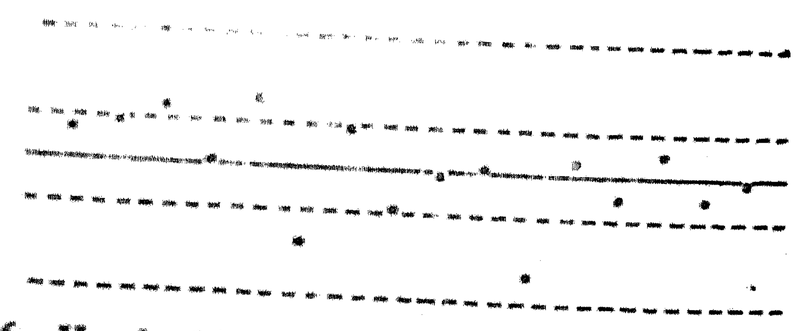

Fig. 6. Voorbeeld van een praktische toepassing van de wet van Gauss. Een serie waarnemingen van b.v. de kwaliteit van grondstoffen. De stippellijnen zijn getekend op afstanden $\pm$ σ en $\pm$ 3σ van het gemiddelde. Wanneer een nieuwe waarneming buiten de buitenste lijnen valt, is er een sterke aanwijzing van een systematische afwijking.

verandering in de kwaliteit der producten is opgetreden. Dit kan dus een waarschuwingsteken zijn voor de betreffende afdeling.

Voor een nadere behandeling van een aantal soortgelijke en meer ingewikkelde gevallen worde verwezen naar het sinds 1934 verschijnende Supplement to the Journal of the Royal Statistical Society (Industrial and Agricultural Research Section).

Een contrôle op het eerste gezicht is mogelijk, wanneer de frequentieverdeling sterk van die van Gauss afwijkt, b.v. tweetoppig is. Een goed voorbeeld van een tweetoppige frequentieverdeling wordt gegeven in fig. 7 [1]), waar de frequentie van verschillende uitbreidingspercentages in de katoenindustrie o. a. van 1930 op 1931 is voorgesteld. We zien daarbij dat een belangrijk deel van de beschouwde ondernemingen een inkrimping van 10 à 15% van de productie te zien heeft gegeven en een ander belangrijk deel een inkrimping van 25 à 30%. Daartussen in liggen echter weinig gevallen. Waarschijnlijk moet men hier denken aan het bestaan van twee groepen van ondernemingen, nl. ondernemingen die hoofdzakelijk voor de export en andere die hoofdzakelijk voor de binnenlandse markt werkten. De laatstgenoemde vertoonden een geringere achteruitgang van de productie dan de eerstgenoemde.

II. Tweedimensionale frequentieverdelingen

§ 19. *De correlatietabel*

In dit hoofdstuk houden we ons bezig met universa en de frequentieverdelingen van gelijktijdig optredende of op andere wijze bijeen-

[1]) Ontleend aan De Ned. Conjunctuur, Juni 1934.

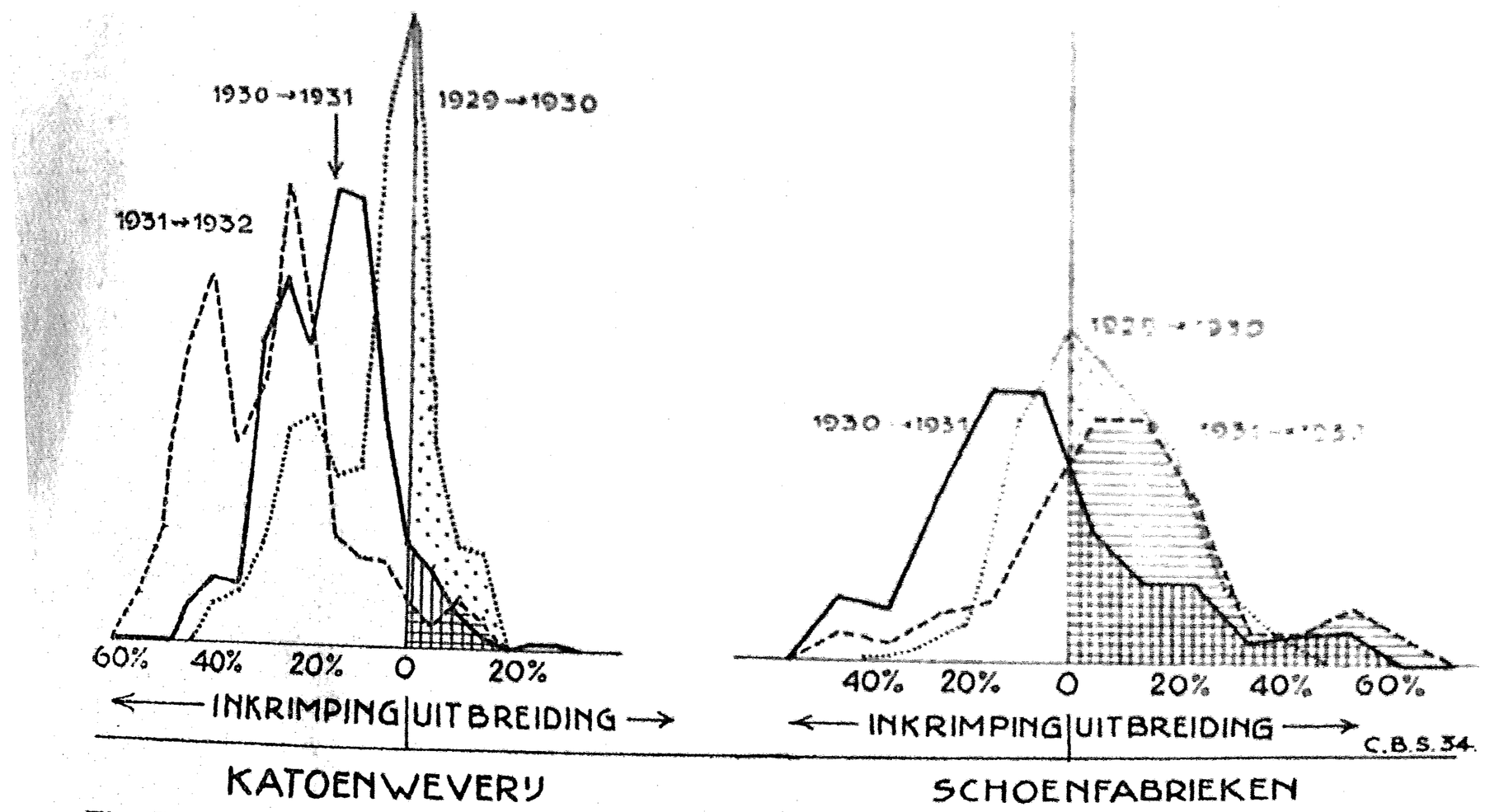

Fig. 7. Voorbeeld van een tweetoppige verdelingscurve: percentage van de ondernemingen in de katoenindustrie resp. de schoenindustrie hetwelk een gegeven achter- of vooruitgang van productie vertoonde van 1929 op 1930, enz. Vooral de curven voor de katoenindustrie zijn duidelijk tweetoppig.

behorende paren van getallen. Een concreet voorbeeld kan zijn de prijs van varkensvlees en het verbruik daarvan. Laten we de eerste grootheid aangeven met X, de tweede met Y. Er zal dan een serie bijeenbehorende waarnemingen zijn, die voorlopig de rol van het universum uit onze voorgaande beschouwingen kunnen spelen. Het universum bestaat nu uit paren. Het is overzichtelijk te beschrijven in een frequentieverdeling die nu echter tweedimensionaal moet zijn. We kunnen nl. beide grootheden X en Y in een aantal klassen onderbrengen, die we voor X aanduiden met X_i, voor Y met Y_j. Zowel de i als de j kan verschillende waarden aannemen; zeggen we b.v. i van 1 tot n en j van 1 tot m. Onze frequentietabel geeft nu voor elke willekeurige combinatie van i en j aan het aantal gevallen f_{ij} dat die combinatie vertoont. We kunnen ook van een grafische voorstelling gebruik maken, die overigens bijna identiek is met deze tabel, nl. een spreidingsdiagram, waarin elke waarneming wordt aangegeven door een punt met de „coördinaten" X_i en Y_j, d.w.z. in het hokje dat tegelijkertijd in de strook van klasse X_i en in die van klasse Y_j ligt. Zou men de klassen zeer klein nemen, dan zou praktisch elke waarneming in een eigen klasseinterval liggen en zou men de zuivere spreidingsfiguur verkrijgen. Daarin hebben de frequenties dan steeds de waarde nul of één.

Het zal duidelijk zijn, dat wanneer de waargenomen punten volkomen op één curve liggen we te maken hebben met een geval van functioneel verband tussen X en Y, ook wel genoemd een geval van volkomen correlatie. Dit is echter in de praktijk zo goed als nooit het geval en dan spreken we van onvolkomen correlatie. De genoemde tabel is een *correlatietabel.*

Men spreekt daarbij nu van *normale correlatie,* wanneer de frequentieverdeling een tweedimensionale generalisatie van de wet van Gauss vormt en wel wanneer de frequentie van de wiskundige vorm

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}\, e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2}+2r\frac{xy}{\sigma_1\sigma_2}\right)}\, dx\, dy$$

is. De frequentie per eenheidsgebied is dan natuurlijk gelijk aan bovenstaande uitdrukking, gedeeld door $dx\, dy$. Men ziet gemakkelijk in dat, wanneer de frequentie per eenheidsgebied als z-coördinaat wordt beschouwd, er een oppervlak ontstaat (het *normale-correlatieoppervlak*) dat als verticale doorsneden steeds curven van Gauss en als horizontale doorsneden ellipsen vertoont.

Voorlopig willen we aannemen dat we met een willekeurige tweedimensionale frequentieverdeling te maken hebben; we willen daarbij veronderstellen dat we tot oneindig kleine klasse-

intervallen zijn overgegaan, waardoor we dus in plaats van de frequentieverdeling eenvoudig de opsomming van de waarnemingspunten als gegeven kunnen beschouwen. We geven nu met de index k een rangnummer van de waarnemingen aan (b.v. de tijd) en hebben dan dus de waarnemingen

$$X_k, Y_k \quad k = 1 \ldots . N$$

waarin N het aantal waarnemingen voorstelt.

§ 20. *De regressielijnen*

We stellen ons nu allereerst de vraag, hoe we een rechte lijn kunnen bepalen die „zo goed mogelijk aansluit" bij de waarnemingen. Een dergelijke lijn kan als een samenvatting van de waarnemingen beschouwd worden en als zodanig praktische waarde hebben.

Het is duidelijk dat deze vraag thans met minder stelligheid te beantwoorden is dan in het geval van een *volkomen* rechtlijnige correlatie het geval zou zijn geweest. Inderdaad zal blijken dat er meer dan één redelijke oplossing is. Daartoe moet echter de vage term „zo goed mogelijk aansluiten bij" mathematisch worden gedefiniëerd. Dit kan geschieden met behulp van de methode der kleinste kwadraten; d.w.z. dat dus gevraagd wordt naar de lijn die zodanige afwijkingen vertoont van de waargenomen punten dat de som der kwadraten dezer afwijkingen zo klein mogelijk is. Men kan de afwijkingen

tussen een punt en een lijn echter op verschillende wijze definiëren; en daarom zijn er hier ook verschillende oplossingen.

De eerste wijze is, dat men de afwijkingen meet in de y-richting. Geeft men de gevraagde rechte lijn aan door $Y'_k = a + bX_k$

dan moet dus

$$\Phi = \sum_1^N{}^k (Y'_k - Y_k)^2 = \sum_1^N{}^k (a + bX_k - Y_k)^2 \text{ min.}$$

zijn. Dit levert voor a en b de vergelijkingen:

$$\frac{\partial \Phi}{\partial a} = 0 \qquad \frac{\partial \Phi}{\partial b} = 0$$

of $\sum_1^N{}^k (a + bX_k - Y_k) = 0$ en

$$\sum_1^N{}^k X_k (a + bX_k - Y_k) = 0$$

of

$$Na + b \sum_1^N{}^k X_k = \sum_1^N{}^k Y_k \text{ resp. } a \sum_1^N{}^k X_k + b \sum_1^N{}^k X_k^2 = \sum_1^N{}^k X_k Y_k.$$

We kunnen hieruit a en b en dus de bedoelde lijn vinden. De formules worden nog iets overzichtelijker, wanneer we de X_k en Y_k voortaan meten vanaf hun gemiddelde en die nieuwe getallen aanduiden met x_k en y_k. Dan is dus $\sum_1^N{}^k x_k = \sum_1^N{}^k y_k = 0$ en we houden over:

$$a' = 0 \qquad b' = \frac{\sum_1^N{}^k x_k y_k}{\sum_1^N{}^k x_k^2} \text{ [1]} \tag{20.1}$$

De tweede wijze, waarop men de afwijkingen kan meten is in de richting van de x-as. We hebben dan de rollen van x en y eenvoudig te verwisselen en krijgen voor de coëfficiënten a'' en b'' in dit geval

$$a'' = 0 \qquad b'' = \frac{\sum_1^N{}^k x_k y_k}{\sum_1^N{}^k y_k^2}$$

waarbij te bedenken valt, dat het nu een lijn is met de vergelijking

$$x_k' = a'' + b'' y_k$$

zodat b'' nu de richtingscoëfficiënt t.o.v. de y-as is. Zou men de richtingscoëfficiënt van deze tweede lijn ook t.o.v. de x-as willen kennen, dan zou men dus moeten nemen

$$\frac{1}{b''} = \frac{\Sigma y^2}{\Sigma xy}$$

(waarbij kortheidshalve de indices zijn weggelaten). In 't algemeen zullen deze lijnen niet

[1]) Voor het bijzondere geval dat men de x en de y in zodanige eenheden meet, dat $\sigma_x = \sigma_y = 1$ is (d.w.z. x en y ,op standaarddeviatie zet") is deze coëfficiënt b' even groot 'als de hierna te bespreken correlatiecoëfficiënt r. Soms wordt b' dan ook in zulke gevallen door het symbool r weergegeven; dit komt mij verwarrend voor.

samenvallen. Wel gaan ze beide — wegens het nul zijn van a' en a'' — door de oorsprong van het nieuwe assenstelsel, d.w.z. door het zwaartepunt van de waarnemingen (het punt met de coördinaten

$$X = \frac{\Sigma X_k}{N} \quad \text{en} \quad Y = \frac{\Sigma Y_k}{N}).$$

Dat de twee lijnen niet samenvallen, is te illustreren aan het volgende eenvoudige geval. Laten we 4 waarnemingen hebben waarvoor geldt:

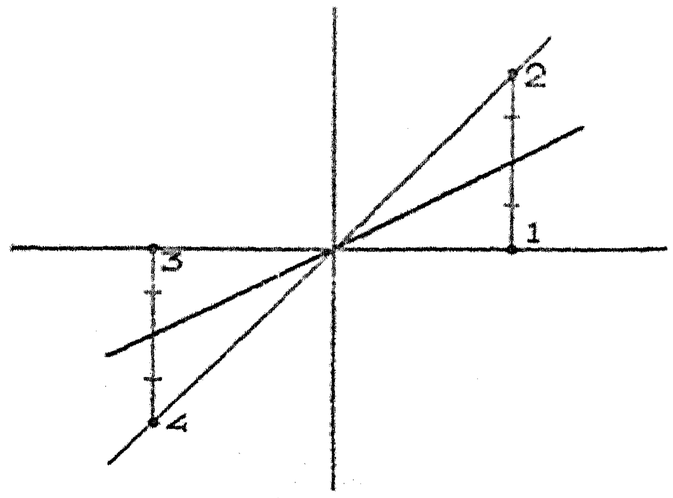

Fig. 8. De twee regressielijnen in een zeer eenvoudig geval: spreidingsdiagram voor de punten (0, ± 1), (± 1, ± 1).

$$x_1 = x_2 = +1 \qquad x_3 = x_4 = -1$$
$$y_1 = y_3 = 0 \qquad y_2 = 1;\ y_4 = -1$$

Men ziet gemakkelijk in dat de regressielijn waarbij de afwijkingen in de y-richting zijn gemeten verloopt als door de zware lijn

aangegeven, met $b' = \frac{1}{2}$; terwijl de andere volgens de dunne lijn verloopt, met $\frac{1}{b''} = 1$.

Naast de hier besproken regressielijnen, die men wel de elementaire noemt, zijn er nog meer te onderscheiden, al naarmate men de definitie van „beste aansluiting" verschillend kiest. Men kan nog denken aan een regressielijn waarbij de loodrecht op die lijn gemeten afstanden een zo klein mogelijke kwadraatsom hebben, enz. Een andere praktisch goed bruikbare regressielijn verkrijgt men nog door te nemen

$$b^* = \frac{\sigma_y}{\sigma_x} = \sqrt{\frac{\sum_1^N{}^k\, y_k^2}{\sum_1^N{}^k\, x_k^2}} \qquad (20.2)$$

waarin, zoals men ziet, de x en y symmetrisch zijn behandeld in die zin, dat wanneer men de regressielijn t.o.v. de andere as berekent men dezelfde verkrijgt.

Daar

$$b' \times \frac{1}{b''} = b^{*2}$$

zien we dat de hellingscoëfficiënt b^* het meetkundig gemiddelde is van de hellingscoëfficiënten der twee elementaire regressielijnen. De laatst behandelde regressielijn ligt dan ook steeds tussen de beide andere in.

Tenslotte kunnen we vaststellen dat de twee elementaire regressielijnen dan en slechts dan samenvallen — en dan ook met de laatst besproken regressielijn coïncideren — wanneer

$$b' = \frac{1}{b''} \quad \text{of} \quad b'b'' = 1, \text{ dus wanneer}$$

$$\frac{(\Sigma\, x\, y)^2}{\Sigma\, x^2 . \Sigma\, y^2} = 1.$$

Zoals in de volgende paragraaf zal worden besproken, is dit het geval wanneer er lineair functioneel verband is tussen x en y.

§ 21. *De correlatiecoëfficiënt*

Deze coëfficiënt definiëert men gewoonlijk als

$$r = \frac{\Sigma\, xy}{\sqrt{\Sigma\, x^2\, \Sigma\, y^2}} \quad \text{of} \quad \frac{\Sigma\, xy}{n\, \sigma_x\, \sigma_y}$$

waarbij de sommaties zich over alle waarnemingen uitstrekken. Van deze coëfficiënt laat zich eenvoudig bewijzen dat hij steeds tussen — 1 en + 1 ligt. Daartoe voeren we eerst de nieuwe grootheden ξ en η in volgens:

$$\xi_k = \frac{x_k}{\sigma_x} \qquad \eta_k = \frac{y_k}{\sigma_y}$$

welke wel genormaliseerde coördinaten genoemd worden. Men zegt ook wel dat ξ en η „op standaarddeviatie staan". Hun standaarddeviaties zijn één, zoals men gemakkelijk inziet en reeds werd opgemerkt. In vergelijkingen uitgedrukt:

$$\sum_1^n{}^k\, \xi_k^2 = n \qquad \sum_1^n{}^k\, \eta_k^2 = n \qquad (21.1)$$

Door deze transformatie wordt:

$$n\,r = \Sigma\,\xi\,\eta \qquad (21.2)$$

We gaan nu een tweede transformatie uitvoeren en wel voeren we in: $\zeta = \dfrac{\xi + \eta}{2}$

$$\alpha = \frac{\xi - \eta}{2}$$

Daardoor wordt $\xi = \zeta + \alpha$ en $\eta = \zeta - \alpha$ en de vergelijkingen (21.1) en (21.2) worden dan:

$$\Sigma\,\zeta^2 + 2\,\Sigma\,\zeta\,\alpha + \Sigma\,\alpha^2 = n \qquad (21.3)$$

$$\Sigma\,\zeta^2 - 2\,\Sigma\,\zeta\,\alpha + \Sigma\,\alpha^2 = n \qquad (21.4)$$

$$nr = \Sigma\,\zeta^2 - \Sigma\,\alpha^2 \qquad (21.5)$$

Uit de twee eerste dezer vergelijkingen leidt men af dat

$$\Sigma\,\zeta\,\alpha = 0.$$

en dus $\qquad \Sigma\,\zeta^2 + \Sigma\,\alpha^2 = n$

Uit een combinatie van de laatste vergelijking en (21.5) volgt nu: $\quad nr \leqq n$

waarbij het gelijkteken alleen geldt wanneer *alle α's gelijk aan nul zijn.* Bovendien ziet men gemakkelijk in dat

$$nr \geqq -\,n \qquad (21.6)$$

en dat het gelijkteken hier alleen geldt wanneer *alle ζ's gelijk aan nul zijn.*

Samenvattend hebben we dus

$$-\,1 \leqq r \leqq +\,1$$

waarbij $r = -\,1$ geldt wanneer elke $\xi_k = -\,\eta_k$ en waarbij $r = +\,1$ geldt wanneer elke $\xi_k = \eta_k$.

Omgekeerd volgt ook uit $r = -\,1$ dat elke

$\xi_k = -\eta_k$ moet zijn en uit $r = +1$, dat elke $\xi_k = \eta_k$ moet zijn. Deze stellingen zijn dus omkeerbaar. Men noemt deze gevallen de gevallen van *volkomen rechtlijnige correlatie* of *rechtlijnig functioneel verband*. Inderdaad is voor de oorspronkelijke grootheden X en Y thans het verband aanwezig

$$Y_k - Y = \frac{\pm \sigma_y}{\sigma_x} (X_k - X)$$

hetgeen een algemeen lineair verband is. Wanneer omgekeerd is $Y_k = a + b X_k$
dan volgt daaruit $\eta_k = \pm \xi_k$
zoals men gemakkelijk narekent. In dit geval staan de zaken dus zó (vgl. § 12) dat de ξ en η *gelijk*, de x en y *proportioneel* en de X en Y *lineair verbonden* zijn.

Van de correlatie-coëfficiënt laat zich ook nog gemakkelijk bewijzen dat hij een maatstaf is voor de spreiding van de afwijkingen tussen de werkelijke waarden van y_k en de met behulp van de regressievergelijking $y'_k = b'x_k$ „berekende" waarden y'_k van y_k.

Inderdaad vinden we voor deze spreiding

$$\frac{1}{n} \sum_1^n{}^k (y_k - y'_k)^2 = \frac{1}{n} \sum_1^n{}^k \left(y_k - \frac{\Sigma xy}{\Sigma x^2} x_k\right)^2 =$$

$$= \frac{1}{n} \left\{ \Sigma y^2 - 2 \frac{\Sigma xy}{\Sigma x^2} \Sigma xy + \frac{(\Sigma xy)^2}{(\Sigma x^2)^2} \Sigma x^2 \right\} =$$

$$= \frac{1}{n} \left\{ \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} \right\}$$

waarbij kortheidshalve de indices verder zijn weggelaten.

Noemen wij de genoemde spreiding $\sigma_{2\cdot 1}$, dan is dus

$$\frac{\sigma_{2\cdot 1}^2}{\sigma_y^2} = 1 - r^2$$

of

$$r^2 = 1 - \frac{\sigma_{2\cdot 1}^2}{\sigma_y^2} \qquad (21.7)$$

Hieruit is ook nog duidelijk af te lezen dat wanneer $\sigma_{2\cdot 1}^2 = 0$, $r = 1$ en omgekeerd, en voorts dat de relatieve spreiding $\frac{\sigma_{2\cdot 1}^2}{\sigma_y^2}$ de waarde van r bepaalt.

§ 22. *Andere maatstaven voor de mate van samenhang*

Er bestaan nu nog enige andere maatstaven van de mate van samenhang die er tussen twee variabelen x en y kan bestaan; deze maatstaven zijn het duidelijkst te definiëren op een met (21.7) analoge wijze. Vervangt men nl. in (21.7) de spreiding $\sigma_{2\cdot 1}^2$ t.o.v. een rechte regressielijn, door $\sigma'^2_{2\cdot 1}$ t.o.v. een kromme lijn, dan gaat de formule over in $\varrho^2 = 1 - \frac{\sigma'^2_{2\cdot 1}}{\sigma_y^2}$

waarbij men ϱ de *correlatie-index* noemt.

Vervangt men verder $\sigma_{2\cdot 1}^2$ door de spreiding $\sigma''^2_{2\cdot 1}$ t.o.v. de gemiddelde y's die men in een aantal klasse-intervallen kan berekenen, dan

verkrijgt men de formule voor de *correlatieratio*:

$$\eta_{yx}^2 = 1 - \frac{\sigma''^2_{2\cdot 1}}{\sigma_y^2}$$

Het zal duidelijk zijn, dat men door de klasse-intervallen maar klein genoeg te kiezen, tenslotte een toestand kan bereiken (althans voor elk praktisch geval), waarbij elke waarneming een klasse vormt. Er is dan in het geheel geen spreiding meer en dan is $\eta_{yx}^2 = 1$. Men kan zodoende steeds een zo fijne klasseverdeling aannemen dat

$$\eta \geq \varrho,$$

terwijl men eveneens steeds een zo „geschikte" kromme regressielijn kan kiezen dat

$$\varrho \geq r.$$

Een vierde maatstaf is de zgn. *mean square contingency*; geeft men het aantal gevallen waarvan de x ligt in de klasse i en de y in de klasse j aan met n_{ij}, dan is de definitie van deze grootheid φ^2:

$$\varphi^2 = \Sigma_{i,j} \frac{\left(n_{ij} - \frac{\Sigma_i n_i \Sigma_j n_j}{N}\right)^2}{\sqrt{(N_x - 1)(N_y - 1)} \cdot \Sigma_i n_i \Sigma_j n_j}$$

waarbij N_x het aantal klassen voor x, N_y dat voor y en N het totaal aantal waarnemingen betekent. Voor een vergelijkende studie betr. het gedrag van de grootheden φ, η en r zij verwezen naar Derksen, Inleiding tot de Correlatierekening, speciaal § 10.

Zoals men ziet, heeft op de waarde van φ^2 slechts het aantal der waarnemingen binnen bepaalde klassen invloed, niet de absolute waarde dezer waarnemingen voor y en x. Men kan φ^2 dus ook berekenen wanneer een van beide of elk der grootheden niet meetbaar doch slechts „rangschikbaar" is. Voor dit laatste geval kan men ook nog als maatstaf bezigen de *coëfficiënt van Spearman*. Denkt men zich de beide reeksen gerangschikt en genummerd van 1 tot n, en vervolgens de rangnummers v_i van de elementen der tweede reeks vastgesteld die bij No. $i = 1 \ldots . n$ van de elementen der eerste behoren, dan is de coëfficiënt van Spearman

$$r_s = 1 - 6 \frac{\Sigma (i - v_i)^2}{n (n^2 - 1)}$$

Deze is echter niets anders dan een andere schrijfwijze voor de uitdrukking voor de correlatiecoëfficiënt tussen de overeenkomende rangnummers:

$$r_s = \frac{\Sigma i \, v_i}{\sqrt{\Sigma i^2 \, \Sigma v_i^2}},$$

hetgeen men gemakkelijk narekent.

Transformatie van Fréchet

Fréchet [1]) heeft bewezen dat men voor de correlatiecoëfficiënt r mag schrijven:

$$r = \varrho \, \eta$$

[1]) M. Fréchet, Sur le coefficient dit de corrélation et sur la corrélation en général, Revue de l'Institut International de Statistique, vol. IV, blz. 1 (1933).

waarbij η de correlatieratio is, en

ϱ de correlatiecoëfficiënt van de verdeling der kolomgemiddelden:

$$\varrho = \frac{\Sigma\, x\, \bar{y}(x)}{\sqrt{\Sigma\, x^2\, \Sigma\, y^2}}$$

waarbij de sommatie nu alleen in de x-richting is uit te voeren. We zien nu hoe slechts dan $r = 1$ kan zijn, wanneer ϱ èn $\eta = 1$ zijn, d.w.z. wanneer de kolomgemiddelden op een rechte lijn liggen ($\varrho = 1$) en tevens geen spreiding om die kolomgemiddelden aanwezig is ($\eta = 1$).

§ 23. *Berekening van y uit x op grond van de regressievergelijking; trendbepaling*

Het bij een correlatieonderzoek gevonden verband tussen twee grootheden y en x kan worden gebezigd om uit een gegeven waarde van de ene variabele x de waarde van de andere y te berekenen. Wanneer de correlatie volkomen en rechtlijnig is, levert deze vraag geen moeilijkheden op. De regressievergelijking geeft het antwoord.

Wanneer echter de correlatie niet volkomen is, behoort bij één waarde van x niet één waarde van y, doch een gehele frequentieverdeling van y. De vraag is, of we van deze frequentieverdeling nog iets te weten kunnen komen. Dit is het volgende.

De regressievergelijking

$$y'_k = b'x_k$$

geeft ons een soort gemiddelde berekende waarde voor y_k. Gemiddelde in die zin dat over alle waarnemingen genomen inderdaad $\overline{y'_k} = \overline{y_k}$ (beide $= 0$).

Voor een bepaald klasse-interval van x behoeft echter y'_k geenszins het gemiddelde te zijn; slechts wanneer de klassegemiddelden van de y's alle op een rechte lijn liggen zou dat het geval zijn. Dit is o.a. zo wanneer de correlatie een normale is (Zie § 19).

Evenzo levert ons vgl. (21.7) enig inzicht in de gemiddelde spreiding van de afwijkingen tussen berekende en werkelijke waarden. We kunnen deze vergelijking nl. schrijven in de vorm

$$\sigma_{2\cdot 1}^2 = \sigma_y^2 \, (1 - r^2) \qquad (23.1)$$

waaruit we zien dat de spreiding $\sigma_{2\cdot 1}^2$ t.o.v. σ_y^2 des te kleiner is naarmate de correlatie hoger is. Ook dit geldt echter weer voor het waarnemingencomplex als een geheel en niet noodzakelijk voor elk afzonderlijk klasseïnterval van x. Ook hier is echter bij normale correlatie dit laatste wel het geval.

Praktisch voorbeeld

In fig. 9 is een praktisch voorbeeld van enkelvoudige correlatie afgebeeld. De bovenste helft van de figuur geeft het spreidingsdiagram; de

bestudeerde grootheden zijn de index van de productie van productiemiddelen *P* en de index van de productie van consumptiegoederen

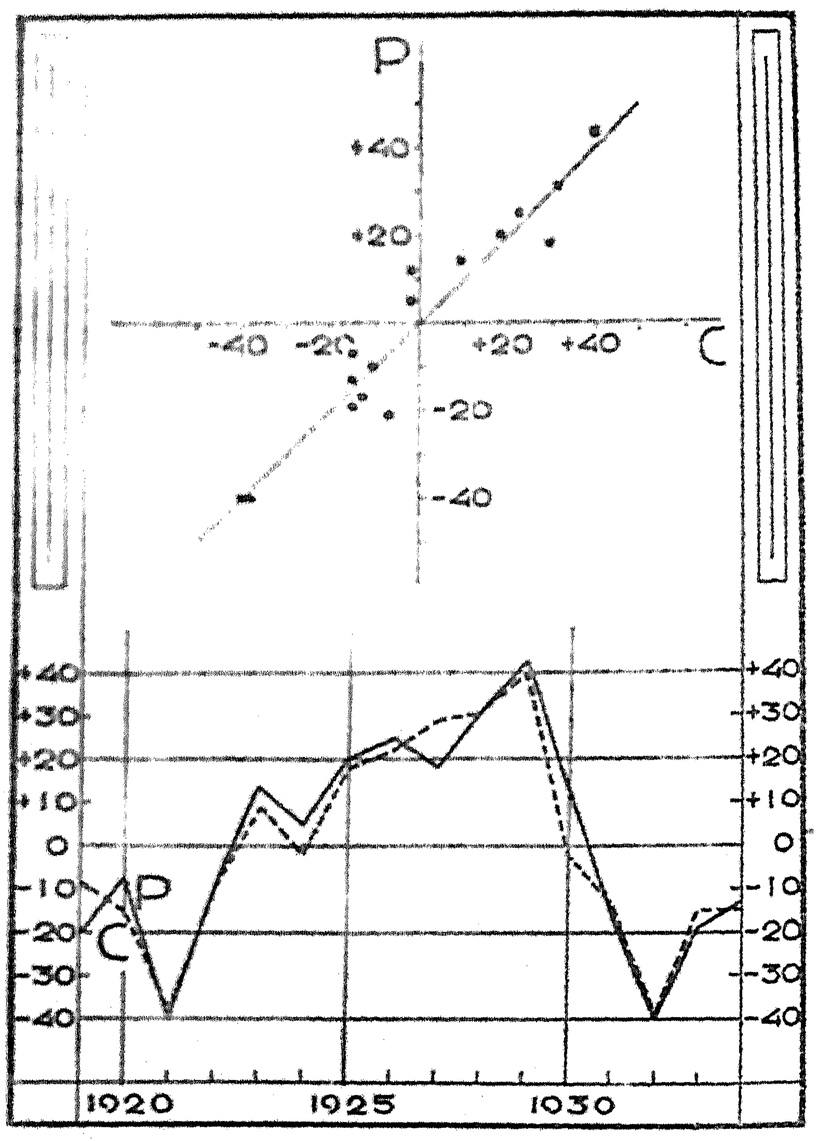

Fig. 9. Enkelvoudige correlatie tussen twee grootheden: productie van productiemiddelen *P* en productie van consumptiegoederen *C* (Ver. Staten).

C in de Verenigde Staten 1919—1934. Van deze laatste zijn de schommelingen 2.2 × vergroot, teneinde eenzelfde amplitude te verkrijgen.

De onderste helft geeft een historische grafiek waarbij P benaderd is door berekening uit C op grond van formule (20.2).

Bepaling van de trend

Een speciaal geval van het bepalen van een regressievergelijking, dat zeer veel toepassing vindt in de economische statistiek, is het bepalen van de zgn. trend of voortschrijdende beweging van een historische reeks. Hierbij is de ene variabele — en wel de als onafhankelijk beschouwde — de tijd. Al naarmate men een rechte of een kromme regressielijn aanneemt, is er sprake van een recht- of kromlijnige trend, welke laatste weer in vele verschillende typen kan worden onderscheiden. In vergelijking met het algemene geval treden twee vereenvoudigingen op in het geval van de trendbepaling. Ten eerste liggen de achtereenvolgende tijdstippen of -perioden, waarvoor men gegevens heeft, gewoonlijk op onderling gelijke afstanden: zij vormen een rekenkundige reeks van getallen. En ten tweede is er voor elke waarde van de tijd slechts één waarneming. De eerste omstandigheid maakt nog een vereenvoudiging mogelijk bij de berekening van de uitdrukking in de noemer van de regressiecoëfficiënt bij een rechtlijnige trend. Geven we de tijd aan met t (in plaats dus van met x in het algemene geval) en

de waarde van het gemeten verschijnsel met y, dan is de regressiecoëfficiënt

$$b = \frac{\Sigma yt}{\Sigma t^2}$$

waarin nu $\Sigma t^2 = 2(1^2 + 2^2 + 3^2 + \ldots + m^2)$, wanneer althans het aantal waarnemingen oneven is en gelijk aan $2m + 1$. De som van de kwadraten van alle gehele getallen tot en met m laat zich door een formule van eenvoudige bouw berekenen:

$$\sum_{1}^{m} i^2 = \frac{1}{6} m (m + 1) (2m + 1)$$

waardoor

$$\Sigma t^2 = \frac{1}{3} m (m + 1) (2m + 1).$$

Bij een even aantal waarnemingen wordt het iets ingewikkelder.

Of men een recht- dan wel een kromlijnige trend moet bepalen, is een kwestie die de wetenschap waartoe de bestudeerde vraagstukken behoren moet oplossen. Slechts indien deze keuze is geschied, kan het statistische werk een aanvang nemen. Voor een voorbeeld van een rechtlijnige trend zie men fig. 10.

Bepaling van de gemiddelde vertraging tussen twee reeksen

Een toepassing op een ander speciaal probleem is de bepaling van de gemiddelde vertraging tussen twee gegeven historische reeksen. Deze kan men bepalen door naar de vertraging te zoeken die

de grootste correlatiecoëfficiënt oplevert. Het is nl. gemakkelijk in te zien, om een geïdealiseerd geval te nemen, dat de correlatiecoëfficiënt tussen sin x en sin $(x - \alpha)$ het grootst is wanneer $\alpha = 0, 2\pi$, enz. is; het kleinst (d.w.z. — 1) als $\alpha = \pi, 3\pi$, enz., en nul als $\alpha = \frac{\pi}{2}, \frac{3\pi}{2}, \frac{5\pi}{2}$ enz.

Dit geïdealiseerde geval is in fig. 11 afgebeeld.

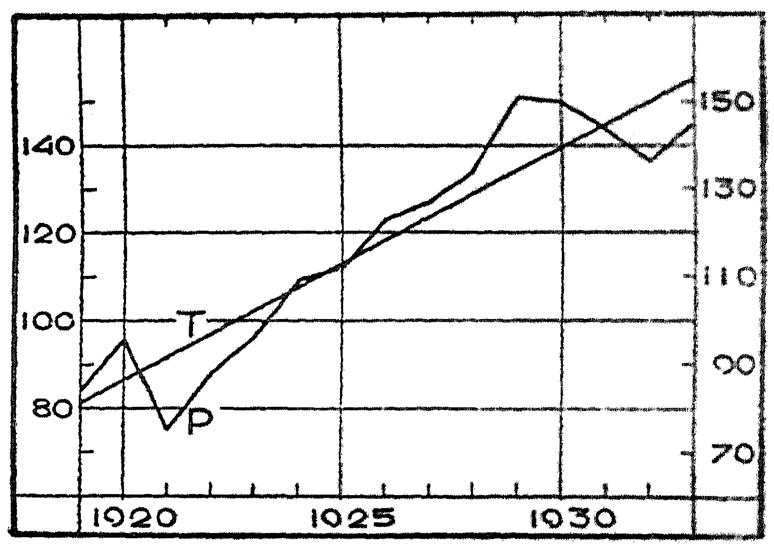

Fig. 10. Bepaling van den trend T van de industriële productie P in Zweden.

III. Meerdimensionale frequentieverdelingen

§ 24. *Meervoudige lineaire correlatie*

De in het voorafgaande hoofdstuk besproken beschouwingen laten zich ook op meer dan twee variabelen uitbreiden. Voor onderzoekingen die met gecompliceerde verschijnselen te maken hebben is zulks ook zeer wenselijk (vgl. Deel I).

Met name is het vraagstuk van betekenis om

een gegeven variabele zo goed mogelijk te benaderen door een combinatie van een aantal andere variabelen — een vraagstuk analoog met

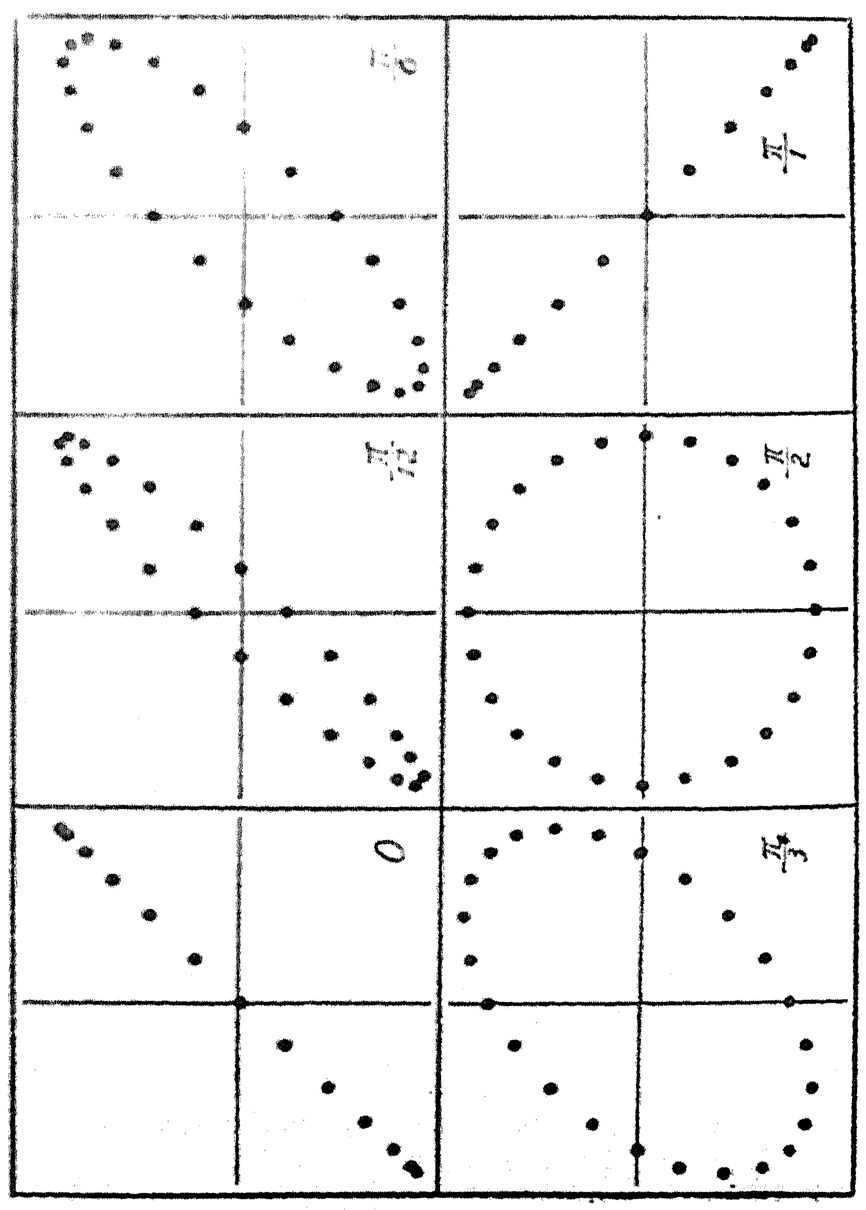

Fig. 11. Correlatie tussen sin x en sin $(x + \alpha)$ voor $\alpha = 0, \frac{\pi}{12}, \frac{\pi}{6}, \frac{\pi}{3}, \frac{\pi}{2}$ en $\frac{\pi}{1}$.

het in § 23 behandelde geval, waar y berekend wordt uit x. Men beperkt zich veelal — zij het niet uitsluitend — tot het geval dat een lineaire combinatie gevraagd wordt en spreekt van

multipele of meervoudige regressie en multipele correlatie. Bij de hier als uitgangspunt gekozen probleemstelling — de probleemstelling zoals ze zich gewoonlijk vanuit de praktijk aandient — wordt dus gevraagd naar de multipele regressievergelijking tussen b.v. x_1 en x_2, $x_3 \ldots . x_n$.

Overeenkomstig de twee elementaire regressielijnen die in het geval van slechts twee variabelen te onderscheiden waren, treedt in het geval met n variabelen een n-tal verschillende regressievergelijkingen op — nu te denken als $n - 1$ dimensionale ruimten in een n-dimensionale — die elk één der variabelen uitdrukken in de $n - 1$ andere. De gehele problematiek die hiermee samenhangt kan dienovereenkomstig op zeer symmetrische en elegante wijze behandeld worden wanneer men gebruik maakt van de leer der determinanten en der matrices. Daartoe zij echter in hoofdzaak verwezen naar b.v. Derksen, Inleiding tot de Correlatierekening en Frisch, Correlation and Scatter in Statistical Variables, Nordic Statistical Journal, 1928, blz. 36.

In dit werkje zal, overeenkomstig de eenvoudige opzet, slechts het hoofdprobleem worden behandeld op elementaire wijze. De notatie is echter uit bovengenoemde werken overgenomen ten einde geen nodeloze verwarring te stichten. Dienovereenkomstig worden de coëfficiënten (regressie- zowel als correlatie-)

aangegeven met n indices, waarvan er 2, de primaire, vooropstaan, gescheiden door een punt van de andere, b.v. $b_{13 \cdot 245 \ldots n}$.

Deze stelt voor de regressiecoëfficiënt van x_1 t.o.v. x_3, d.w.z. de coëfficiënt die in de vergelijking ter benadering van x_1 staat voor de x_3; de secundaire indices $2, 4, 5 \ldots n$ geven aan welke andere variabelen in de vergelijking voorkomen. Op dezelfde wijze bestaat ook $r_{13 \cdot 245 \ldots n}$, waarvan echter de betekenis eerst later zal worden vermeld.

Het hoofdvraagstuk is nu om te bepalen de coëfficiënten in de vergelijking

$$x_1' = b_{12 \cdot 34 \ldots n}\, x_2 + b_{13 \cdot 24 \ldots n}\, x_3 + \ldots + b_{1n \cdot 23 \ldots n-1}\, x_n$$

zodanig dat tussen x_1' en x_1 een „zo nauw mogelijke aansluiting bestaat", hetgeen we weer vertalen door te eisen dat

$$\Sigma\, x_{1 \cdot 2}{}^2 \ldots n = \Sigma\, (x_1 - x_1')^2 \text{ min.}$$

is. Hier worden dus eveneens symbolen met n indices, doch slechts met 1 primaire index, ingevoerd, welke de afwijkingen of *residuen* zijn tussen een variabele en zijn met behulp van de in de secundaire indices aangegeven variabelen benaderde waarde. Volkomen analoog met wat vroeger werd gevonden, vinden we nu de zgn. „normaalvergelijkingen":

$$\Sigma\, x_{1 \cdot 2 \ldots n}\, x_2 = 0 \qquad \Sigma\, x_{1 \cdot 2 \ldots n}\, x_3 = 0 \ldots \qquad \Sigma\, x_{1 \cdot 2 \ldots n}\, x_n = 0 \tag{24.1}$$

of, uitgeschreven:

$$\Sigma x_1x_2 = b_{12\cdot34\ldots n}\ \Sigma x_2^2 + b_{13\cdot24\ldots n}\ \Sigma x_2x_3 + \ldots\ldots b_{1n\cdot23\ldots n-1}\ \Sigma x_2x_n$$

$$\Sigma x_1x_3 = b_{12\cdot34\ldots n}\ \Sigma x_2x_3 + b_{13\cdot24\ldots n}\ \Sigma x_3^2 + \ldots\ldots b_{1n\cdot23\ldots n-1}\ \Sigma x_3x_n$$

. . . . . .

$$\Sigma x_1x_n = b_{12\cdot34\ldots n}\ \Sigma x_2x_n + b_{13\cdot24\ldots n}\ \Sigma x_3x_n + \ldots\ldots b_{1n\cdot23\ldots n-1}\ \Sigma x^2_n$$

Deze $n - 1$ vergelijkingen van de eerste graad zijn voldoende om de $n - 1$ onbekenden b te bepalen. Zoals men ziet, zijn deze vergelijkingen zeer symmetrisch opgebouwd. Een der factoren in de productsommen die overal als coëfficiënten optreden is steeds dezelfde in een (horizontale) rij, de andere factor steeds in een (vertikale) kolom. In praktische gevallen zal men alle productsommen

$$\begin{array}{ccccc} \Sigma x_1x_2 & \Sigma x_2^2 & \Sigma x_2x_3 & \ldots\ldots & \Sigma x_2x_n \\ \vdots & \vdots & \vdots & & \vdots \\ \Sigma x_1x_n & \Sigma x_2x_n & \Sigma x_3x_n & \ldots\ldots & \Sigma x^2_n \end{array}$$

uit de waarnemingen berekenen — waarbij te bedenken is dat alle x-waarden hier als afwijkingen van hun gemiddelden bedoeld zijn, omdat anders de aangenomen regressievergelijking nog een bekende term had moeten bezitten.

Op eenvoudige wijze laat zich voorts ook de correlatiecoëfficiënt $r_{1\cdot23\ldots n}$ tussen x_1 en x_1'

berekenen. Deze is krachtens definitie (§ 21)

$$r = \frac{\Sigma x_1 x_1'}{\sqrt{\Sigma x_1^2 \; \Sigma x_1'^2}} \tag{24.2}$$

Van de in deze uitdrukking voorkomende sommen zou ons $\Sigma x_1'^2$ nogal enig werk kosten. Deze kunnen we echter door een kunstgreep gemakkelijk berekenen. Daartoe denken we ons de vgln. (24.1) resp. vermenigvuldigd met $b_{12\cdot 3\ldots n}$, $b_{13\cdot 24\ldots n}$ .... $b_{1n\cdot 23\ldots n-1}$ en opgeteld. Er komt dan:

$$\Sigma x_{1\cdot 2\ldots n}\,(b_{12\cdot 34\ldots n}\,x_2 + b_{13\cdot 24\ldots n}\,x_3 + \ldots + b_{1n\cdot 23\ldots n-1}\,x_n) =$$
$$\Sigma x_{1\cdot 2\ldots n}\,x_1' = \Sigma (x_1 - x_1')\,x_1' = 0$$

of:

$$\Sigma x_1 x_1' = \Sigma x_1'^2 \tag{24.3}$$

Als we dit gebruiken om $\Sigma x_1'^2$ te berekenen, wordt (24.2):

$$r = \sqrt{\frac{\Sigma x_1 x_1'}{\Sigma x_1^2}} \tag{24.4}$$

of

$$r = \sqrt{\frac{b_{12\cdot 34\ldots n}\,\Sigma x_1 x_2 + \ldots b_{1n\cdot 23\ldots n-1}\,\Sigma x_1 x_n}{\Sigma x_1^2}}$$

hetgeen na het voorgaande gemakkelijk te berekenen valt.

Voor praktische doeleinden zal het meestal aangenamer zijn, niet de hier gebezigde notatie van de coëfficiënten b te gebruiken, doch enkelvoudige letters. Gemakshalve volgen hier nog

de formules voor $n = 4$ bij deze vereenvoudigde notatie:

$$x_1' = a\,x_2 + b\,x_3 + c\,x_4$$

$$\Sigma\, x_1x_2 = a\,\Sigma\, x_2^2 + b\,\Sigma\, x_2x_3 + c\,\Sigma\, x_2x_4$$

$$\Sigma\, x_1x_3 = a\,\Sigma\, x_2x_3 + b\,\Sigma\, x_3^2 + c\,\Sigma\, x_3x_4$$

$$\Sigma\, x_1x_4 = a\,\Sigma\, x_2x_4 + b\,\Sigma\, x_3x_4 + c\,\Sigma\, x_4^2$$

$$r = \sqrt{\frac{a\,\Sigma\, x_1x_2 + b\,\Sigma\, x_1x_3 + c\,\Sigma\, x_1x_4}{\Sigma\, x_1^2}}$$

Een gemakkelijke formule voor de totale correlatiecoëfficiënt in het geval van 3 reeksen is nog:

$$r_{1\cdot 23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2\, r_{12}\, r_{13}\, r_{23}}{1 - r^2_{23}}}$$

Uit de vergelijkingen (24.3) en (24.4) laat zich nog een conclusie trekken omtrent de verhouding van de standaarddeviaties van x_1 en x_1':

$$\frac{\sigma^2_{x_1'}}{\sigma^2_{x_1}} = \frac{\Sigma\, x_1'^2}{\Sigma\, x_1^2} = \frac{\Sigma\, x_1x_1'}{\Sigma\, x_1^2} = r^2$$

Dit betekent dus dat, naarmate de correlatiecoëfficiënt lager is, men een berekende waarde x_1' verkrijgt met een lagere standaarddeviatie dan de x_1, een en ander weder in volkomen analogie met het geval met de enkelvoudige correlatie. Evenals daar zou men daarom een soort van genormaliseerde regressievergelijking verkrijgen, indien men berekende een grootheid:

$$x_1^* = \frac{a}{r}\, x_2 + \frac{b}{r}\, x_3 + \frac{c}{r}\, x_4 \qquad (24.5)$$

hetgeen in een aantal onderzoekingen van het Centraal Bureau voor de Statistiek gedaan is.

Partiële correlatie

Uit de normaalvergelijkingen (24.1) leiden wij af dat bij een willekeurige productsom van residuen

$$\Sigma\, x_{i,kl\ldots m}\, x_{j,kl\ldots n}$$

alle gelijke secundaire indices bij een van beide factoren mogen worden weggelaten. Laten deze gelijke indices k en l zijn, dan is dus

$$\Sigma\, x_{i,kl\ldots m}\, x_{i,kl\ldots n} = \Sigma\, x_{i,\ldots m}\, \Sigma\, x_{i,kl\ldots n} \quad (24.6)$$

Het bewijs volgt uit de definitie van de residugrootheden in verband met de normaalvergelijkingen:

$$x_{i,kl\ldots m} = x_i - b_{ik,l\ldots m}\, x_k - b_{il,k\ldots m}\, x_l - \ldots\ldots b_{im,kl\ldots}\, x_m \quad \text{en}$$

$$x_{i,\ldots m} = x_i - \ldots\ldots b_{im,\ldots}\, x_m$$

met welke laatste vergelijking moge worden aangeduid dat daarin geen termen met x_k en x_l voorkomen. In het ontbreken van de dienovereenkomstige termen aan de rechterzijde van (24.6) ligt dus het verschil tussen rechterzijde en linkerzijde van die vergelijking. Dat ontbreken is echter inderdaad irrelevant, aangezien die termen de productsommen

$$\Sigma\, x_k\, x_{i,kl\ldots n} \quad \text{en} \quad \Sigma\, x_l\, x_{i,kl\ldots n}$$

als factoren bevatten en deze laatste krachtens de normaalvergelijkingen nul zijn.

Uit dien hoofde mogen we de normaalvgl.

$$\Sigma\, x_2\, x_{1\cdot 23\,\cdots\, n} = 0$$

ook schrijven

$$\Sigma\, x_{2\cdot 3\,\cdots\, n}\, x_{1\cdot 23\,\cdots\, n} = 0$$

aangezien uit de laatste vlg. de voorlaatste volgt door weglaten bij een der factoren van de secundaire indices die ze gemeen hebben.

De laatste vergelijking schrijven we uit:

$$\Sigma\, x_{2\cdot 3\,\cdots\, n}\, (x_1 - b_{12\cdot 3\,\cdots\, n}\, x_2 - \ldots\ldots) = 0.$$

De termen met x_3, enz. mogen we echter krachtens de normaalvgln. alweer weglaten, omdat ze tot somtermen aanleiding geven die nul zijn. We mogen nu in de overblijvende vergelijking weer bepaalde secundaire indices toevoegen:

$$\Sigma\, x_{2\cdot 3\,\cdots\, n}\, x_{1\cdot 3\,\cdots\, n} - b_{12\cdot 3\,\cdots\, n}\, \Sigma\, x^2_{2\cdot 3\,\cdots\, n} = 0.$$

Uit deze vergelijking is $b_{12\cdot 3\,\cdots\, n}$ te berekenen:

$$b_{12\cdot 3\,\cdots\, n} = \frac{\Sigma\, x_{1\cdot 3\,\cdots\, n}\, x_{2\cdot 3\,\cdots\, n}}{\Sigma\, x^2_{2\cdot 3\,\cdots\, n}} \qquad (24.7)$$

waaruit volgt, dat $b_{12\cdot 3\,\cdots\, n}$ ook mag worden *geïnterpreteerd* als de regressiecoëfficiënt in een geval van enkelvoudige correlatie (vgl. § 20, vgl. (20.1)) en wel tussen de grootheden $x_{1\cdot 3\,\cdots\, n}$ en $x_{2\cdot 3\,\cdots\, n}$, of de residuen van resp. x_1 en x_2 t.o.v. de grootheden x_3 tot en met x_n.

Men vindt natuurlijk op analoge wijze:

$$b_{21\cdot 3\,\cdots\, n} = \frac{\Sigma\, x_{2\cdot 3\,\cdots\, n}\, x_{1\cdot 3\,\cdots\, n}}{\Sigma\, x^2_{1\cdot 3\,\cdots\, n}}$$

en, evenals in § 20, kan men nu definiëren

$$r^2_{12\cdot3\ldots n} = r^2_{21\cdot3\ldots n} = b_{12\cdot3\ldots n}\, b_{21\cdot3\ldots n}$$

de correlatiecoëfficiënt tussen de genoemde residuen. Men noemt deze grootheden *partiële correlatiecoëfficiënten.*

Uit de normaalvergelijkingen blijkt, dat residuen niet correleren met die variabelen waarvan de indices als secundaire indices in die residuen voorkomen. Men kan dat aldus uitdrukken dat men zegt: *de residuen zijn de niet met de in de secundaire indices tot uitdrukking komende variabelen correlerende gedeelten van een andere variabele:* $x_{1\cdot23\ldots n}$ is b.v. het niet met $x_2, x_3 \ldots x_n$ correlerende gedeelte van x_1.

De partiële correlatiecoëfficiënten meten de correlatie tussen dergelijke residuen, waarin dan dezelfde secundaire indices voorkomen.

Spreiding der residuen van hogere orde

Voor het geval dat $n = 3$ bewijzen we nog een formule die een goed inzicht geeft in de afneming van de spreiding bij vergroting van n.

Volgens definitie is:

$$x_{1\cdot23} = x_1 - b_{12\cdot3}\, x_2 - b_{13\cdot2}\, x_3.$$

Daaruit volgt:

$$\Sigma\, x^2_{1\cdot23} = \Sigma\, x_{1\cdot23}\, x_{1\cdot2} \qquad (24.8)$$

omdat nl. $\Sigma\, x_{1\cdot23}\, x_2 = \Sigma\, x_{1\cdot23}\, x_3 = 0$ zijn (vgl. (24.1)) en men dus termen die daarmee evenredig zijn naar believen mag weglaten of met

een constante factor vermenigvuldigen. In de uitdrukking ter linkerzijde van (24.6) is nu de term met $x_{1\cdot 23}\, x_3$ geheel weggelaten, terwijl de term met $x_{1\cdot 23}\, x_2$ een andere coëfficiënt heeft gekregen: links heeft hij de coëfficiënt $b_{13\cdot 2}$, rechts de coëfficiënt b_{12}.

We transformeren nog verder:

$$\Sigma\, x^2_{1\cdot 23} = \Sigma\, x_{1\cdot 2}\, x_1 - b_{13\cdot 2}\, \Sigma\, x_{1\cdot 2}\, x_3.$$

Daarbij is weggelaten de term $- b_{12\cdot 3}\, \Sigma\, x_{1\cdot 2}\, x_2$, die echter ook nul is krachtens (24.1) voor $n = 2$. Voorts mogen we om dezelfde reden voor $\Sigma\, x_{1\cdot 2}\, x_1$ ook schrijven $\Sigma\, x^2_{1\cdot 2}$.

We krijgen nu:

$$\Sigma\, x^2_{1\cdot 23} = \Sigma\, x^2_{1\cdot 2} \left(1 - b_{13\cdot 2} \frac{\Sigma\, x_{1\cdot 2}\, x_3}{\Sigma\, x^2_{1\cdot 2}}\right) =$$

$$= \Sigma\, x^2_{1\cdot 2} \left(1 - b_{13\cdot 2} \frac{\Sigma\, x_{1\cdot 2}\, x_{3\cdot 2}}{\Sigma\, x^2_{1\cdot 2}}\right) =$$

$$= \Sigma\, x^2_{1\cdot 2}\, (1 - b_{13\cdot 2}\, b_{31\cdot 2}) = \Sigma\, x^2_{1\cdot 2}\, (1 - r^2_{13\cdot 2}).$$

Delende door n verkrijgt men daaruit, wanneer de standaarddeviaties met σ worden aangegeven:

$$\sigma^2_{1\cdot 23} = \sigma^2_{1\cdot 2}\, (1 - r^2_{13\cdot 2}) \qquad (24.9).$$

Wanneer we nog gebruik maken van (23.1), waarbij de in deze paragraaf ingevoerde notatie wordt gebezigd, dan bereiken wij de formule:

$$\sigma^2_{1\cdot 23} = \sigma^2_1\, (1 - r^2_{12})\, (1 - r^2_{13\cdot 2}) \qquad (24.10).$$

Deze formule doet ons, bij vergelijking met (23.1), zien, hoe bij elke nieuwe variabele die voor de „verklaring" van x_1 wordt aangewend, de standaarddeviatie der residuen wordt ver-

kleind; deze verkleining is des te belangrijker naarmate in ons geval van (24.) $r^2_{13\cdot 2}$ dichter bij de 1 ligt, d.w.z. er een hogere correlatie bestaat tussen $x_{1\cdot 2}$ en $x_{3\cdot 2}$, d.w.z. tussen „het element in x_3 dat niet met x_2 correleert" en „het element in x_1 dat niet met x_2 correleert".

Men ziet er ook uit dat het insluiten van een nieuwe variabele *altijd* de spreiding verkleint, ook dus wanneer de nieuwe variabele niets te maken heeft met de te verklaren reeks. Men zal dus alleen gewicht kunnen hechten aan een verbetering van de correlatie, wanneer deze verbetering aanzienlijk is.

Voor n variabelen bewijst men:

$$\sigma^2_{1\cdot 23\ldots n} = \sigma_1^2\,(1 - r_{12}^2)\,(1 - r^2_{13\cdot 2})\,(1 - r^2_{14\cdot 23})\ldots(1 - r^2_{1n\cdot 23\ldots n-1}).$$

Vermelden wij nog dat deze en nog aanmerkelijk meer resultaten op elegante wijze kunnen worden afgeleid door gebruik te maken van de zgn. *determinant van Gram*:

$$R = \begin{vmatrix} 1 & r_{12} & r_{13} & \ldots\ldots & r_{1n} \\ r_{12} & 1 & r_{23} & \ldots\ldots & r_{2n} \\ r_{13} & & & & \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{1n} & \ldots & \ldots & \ldots & 1 \end{vmatrix}$$

Voor dergelijke beschouwingen zij verwezen naar Derksen, Inleiding tot de Correlatierekening, Delft 1935, blz. 33 e.v. en de daar genoemde literatuur.

Voorbeelden

Een eenvoudig economisch voorbeeld wordt geïllustreerd door fig. 12 en bijgaand tableau,

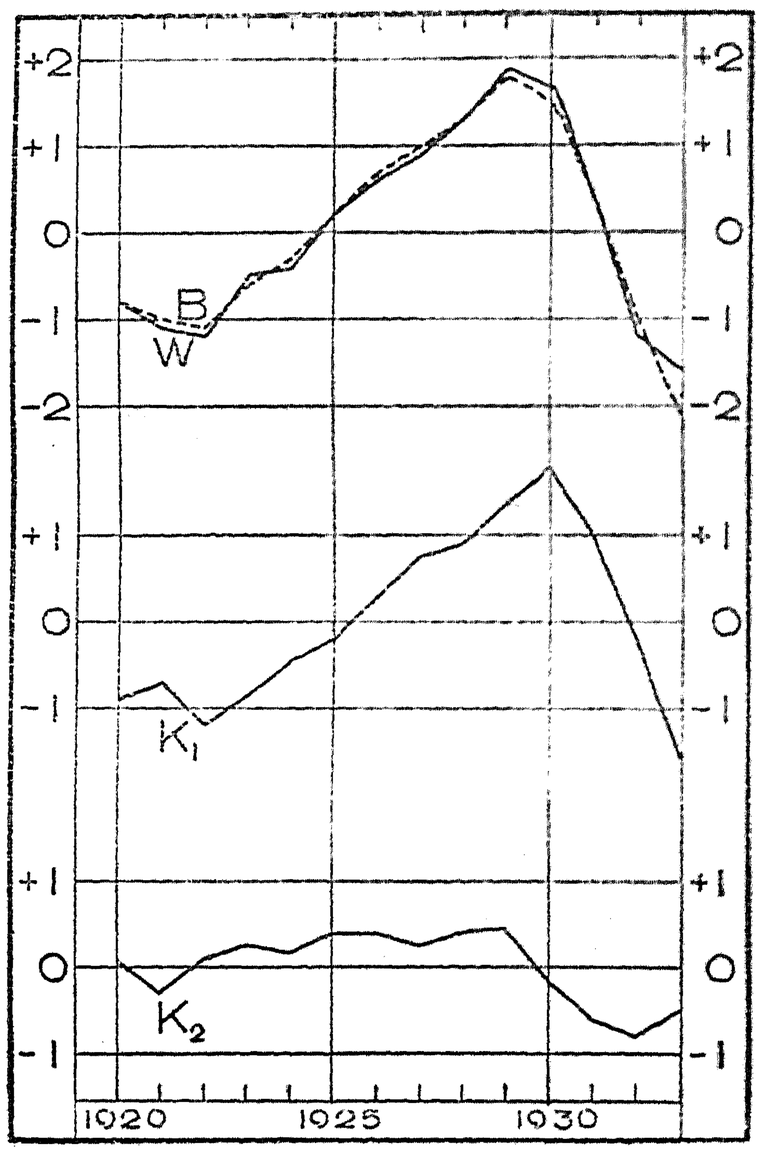

Fig. 12. Meervoudige correlatie tussen totale dividendsom W enerzijds en reservepositie en nettowinst anderzijds (Ver. Staten). B berekende dividendsom; K_1 invloed daarop van de reservepositie, K_2 invloed van de nettowinst. (Ontleend aan De Ned. Conj., Aug. 1935).

waarin men achtereenvolgens afgebeeld vindt, in de notatie van vgl. (24.5), de grootheden x, x^*

TABLEAU: Voorbeeld van een vereenvoudigde berekening van de regressievergelijking (2 verklarende reeksen). (Fig. 12).

	W	C	P	$W-\overline{W}=$ w	$C-\overline{C}=$ c	$P-\overline{P}=$ p	w^2	c^2	p^2
1920	2.90	4.3	4.28	—0.8	—5.5	+ 1.0	0.64	30.2	1.00
1921	2.69	5.7	—0.05	—1.1	—4.1	—3.3	1.21	16.8	10.89
1922	2.63	3.0	4.38	—1.2	—6.8	+ 1.1	1.44	46.2	1.21
1923	3.30	4.7	5.83	—0.5	—5.1	+ 2.6	0.25	26.0	6.76
1924	3.42	7.2	5.00	—0.4	—2.6	+ 1.8	0.16	6.8	3.84
1925	4.01	8.8	6.97	+ 0.2	—1.0	+ 3.7	0.04	1.0	13.69
1926	4.44	11.8	6.77	+ 0.6	+ 2.0	+ 3.5	0.36	4.0	12.25
1927	4.76	14.1	5.88	+ 0.9	+ 4.3	+ 2 6	0.81	18.5	6.76
1928	5.16	15.2	7.64	+ 1.3	+ 5.4	+ 4.4	1.69	29.2	19.36
1929	5.76	17.7	8.08	+ 1.9	+ 7.9	+ 4.8	3.61	62.4	23.04
1930	5.63	20.0	1.38	+ 1.8	+ 10.2	—1.9	3.24	104.0	3.61
1931	4.18	15.7	—3.14	+ 0.3	+ 5.9	—6.4	0.09	34.8	40.96
1932	2.63	8.4	—5.37	—1.2	—1.4	—8.6	1.44	2.0	73.96
1933	2,20	0.4	—2.36	—1.6	—9.4	—5.6	2.56	88.4	31.36
Totaal	53.71	137.0	45.29	+ 0.2	—0.2	—0.3	17.54	470.3	248.09
Gem.	3.84	9.8	3.24						

W = Dividendsom van alle naamloze vennootschappen in de V.S., mld. \$. C = Cumulatie van de niet uitgekeerde winst sinds 1 Jan. 1920, mld. \$. P = Nettowinst, mld. \$. Bronnen: Bull. Nat. Bureau of Economic Research No. 50; Nat. City Bank, Economic Conditions, April 1935, p. 61.

	wc		wp		cp		$a^*c + b^*p = B$ (berekende divid.) $0.17c + 0.09p = B$	
	+	—	+	—	+	—		
1920	4.40			0.80		5.50	—0.93 + 0.09	—0.8
1921	4.51		3.63		13.53		—0.70 —0.30	—1.0
1922	8.16			1.32		7.48	—1.16 + 0.10	—1.1
1923	2.55			1.30		13.26	—0.87 + 0.23	—0.6
1924	1.04			0.72		4.18	—0.44 + 0.16	—0.3
1925		0.20	0.74			3.70	—0.17 + 0.33	+ 0.2
1926	1.20		2.10		7.00		+ 0.34 + 0.32	+ 0.7
1927	3.87		2.34		11.18		+ 0.73 + 0.23	+ 1.0
1928	7.02		5.72		23.76		+ 0.92 + 0.40	+ 1.3
1929	15.01		9.12		37.92		+ 1.32 + 0.43	+ 1.8
1930	18.36			3.42		19.38	+ 1.77 —0.17	+ 1.6
1931	1.77			1.92		37.76	+ 1.00 —0.58	+ 0.4
1932	1.68		10.32		12.04		—0.24 —0.78	—1.0
1933	13.24		8.96		52.64		—1.60 —0.50	—2.1
Totaal	82.81	0.20	42.93	9.48	158.07	91.76	—0.03 —0.04	0.0

De normaalvgln.: $\Sigma wc = a\,\Sigma c^2 + b\,\Sigma cp$

$\Sigma wp = a\,\Sigma cp + b\,\Sigma p^2$

worden: $82.6 = 470\,a + 66.3\,b$

$33.4 = 66.3\,a + 248\,b$, waarvan de oplossingen zijn: $a = 0.163$, $b = 0.09$.

Daaruit volgt $r = \sqrt{\dfrac{a\,\Sigma wc + b\,\Sigma wp}{\Sigma w^2}} = 0.97$; zodat $a^* = \dfrac{a}{r} = 0.17$, $b^* = \dfrac{b}{r} = 0.09$.

(dooreengetekend om de mate van overeenstemming zo duidelijk mogelijk zichtbaar te maken), $\frac{a}{r}$ y en $\frac{b}{r}$ z. In dit voorbeeld is x de loop der dividendsom, uitgekeerd door de gezamenlijke naamloze vennootschappen in de Verenigde Staten over elk der jaren 1920—1933; y het bedrag der gecumuleerde niet-uitgekeerde winsten van vorige jaren (een maatstaf voor de reserves) en z het bedrag der nettowinst in het jaar in kwestie.

Men ziet uit de figuur dat de dividendpolitiek in de Verenigde Staten in belangrijke mate wordt bepaald door de reservepositie. Een vergelijking met andere landen, doorgevoerd in De Nederl. Conjunctuur van Aug. 1935, waaraan ook fig. 12 is ontleend, doet zien dat dit in die andere landen anders is.

Een voorbeeld van psychologische aard is te vinden in „Intelligentieverhoudingen in Nederland" [1]), waar, op een wat meer elementaire

[1]) Statistiek van Nederland; bewerkt door het Centraal Bureau voor de Statistiek; uitgave J. B. Wolters, Groningen —Batavia 1935, blz. 46 en 51.

Door Prof. Dr. Th. Goossens is (vgl. overdrachtsrede van den Rector van de R.K. Handelshogeschool te Tilburg op 16 Sept. 1935) kritiek uitgeoefend op de conclusie dat de intelligentie in Noord-Brabant het laagste was. Opgemerkt werd door hem dat de geestelijken buiten het onderzoek vielen. Voor de „voor onderwijs gecorrigeerde

wijze, getracht is de afzonderlijke invloeden van beroep en geografische plaats op de intelligentiegemiddelden voor verschillende soorten onderwijs genoten hebbende groepen recruten te bepalen, evenals de afzonderlijke invloeden van genoten onderwijs en geografische plaats op die intelligentiegemiddelden. De uitschakeling van de beroepsinvloed is b.v. aldus geschied, dat afgetrokken wordt de gemiddelde intelligentie die voor het beroep in kwestie is gevonden (men vergelijke de behandelingswijze in § 30).

Een tweede voorbeeld uit de psychologisch-pedagogische wereld is de reeds besproken methode van *Spearman* ter verklaring van de rapportcijfers voor verschillende vakken (zie § 13, einde); er werd reeds op gewezen, dat door Spearman een veelvuldig gebruik gemaakt wordt van de formules der multipele correlatie, doch dat anderzijds de probleemstelling een weinig afwijkt van de hier gevolgde.

IV. Steekproeven

§ 25. *Eendimensionale steekproeven uit discontinue universa*

Reeds even werd (in § 18) het vraagstuk

---

cijfers" geldt echter dit verwijt niet. Overigens bleek uit een antwoord van Dr. Ph. Idenburg, chef van de afd. Onderwijsstatistiek, dat het gemiddelde der ongecorrigeerde cijfers niet belangrijk zou stijgen, indien de geestelijken zouden worden meegeteld.

aangeroerd waarover in dit hoofdstuk meer uitvoerig zal worden gesproken: het centrale vraagstuk van de steekproevenleer. Het kan het best op deze plaats weer geïntroduceerd worden door de vraag: indien een meting wordt verricht aan een gedeelte van een universum, welke „waarde" heeft men dan aan de gevonden resultaten te hechten. De hierbij gebruikte vage term „waarde" laat zich verduidelijken door de volgende precizering van de probleemstelling: zal een overeenkomstige meting aan het gehele universum tot afwijkingen voeren van de gedeeltelijke meting en wat is de waarschijnlijkheid van een gegeven afwijking daarbij? In de uitdrukkingswijze van de voorafgaande bladzijden is deze vraag aldus te formuleren: gegeven de frequentieverdeling van een gedeelte van een universum (verder steeds te noemen: de steekproef); gevraagd: wat weet men van de frequentieverdeling van het universum zelf? [1])

---

[1]) Op deze plaats moge even genoemd worden de zgn. „method of maximum likelihood" van Fisher, voor het geval men de wiskundige vorm van de frequentieverdeling kent, doch de waarde der daarin voorkomende constanten niet. Als deze functie wordt voorgesteld door $\varphi\,(x, a, b) = 0$, en als gegeven zijn de waarnemingen $x_1, x_2 \ldots x_n$, dan bepaalt Fisher de waarden a en b door

$$\Pi\, \varphi\,(x_i, a, b) = \varphi\,(x_1, a, b)\ \varphi\,(x_2, a, b) \ldots \varphi\,(x_n, a, b)$$

tot een maximum te maken.

Voor de normale verdeling met constante standaarddeviatie wordt dit:

Een concreet getallenvoorbeeld wordt in § 26 gegeven.

Verder onderzoek wijst uit, dat strengere antwoorden kunnen worden gegeven door de probleemstelling iets anders te formuleren en wel in zekere zin om te keren: gegeven het universum, gevraagd de frequentieverdelingen van aan steekproeven te meten grootheden. B.v. gevraagd de frequentieverdeling van het gemiddelde van steekproeven met een bepaald aantal objecten, van de standaarddeviatie, enz.

Al naarmate het universum waaruit steekproeven genomen worden gecompliceerder is, zullen zich hierbij ook gecompliceerder vraagstukken voordoen. Slechts de eenvoudiger vraagstukken zullen hier worden behandeld. Daarbij zullen eerst de problemen worden besproken waarbij de steekproef door een discontinue eendimensionale frequentieverdeling kan worden beschreven, vervolgens de gevallen waarbij meerdimensionale frequentieverdelingen optreden. We beginnen met eendimensionale steekproeven uit discontinue universa.

Als eenvoudigste geval van de hier optredende problemen is wel het volgende te beschouwen:

$$e^{-\frac{1}{2\sigma^2}\left[(x_1-m_1)^2+(x_2-m_2)^2+\ldots+(x_n-m_n)^2\right]} \quad \text{max.}$$

of $(x_1-m_1)^2+(x_2-m_2)^2+\ldots+(x_n-m_n)^2$ min.

d.w.z. de methode der kleinste kwadraten.

er is een universum waarin slechts twee kenmerken te onderscheiden zijn, een positief en een negatief. De relatieve frequentie der positieve objecten is p, die van de negatieve $1 - p$, waarvoor we q schrijven. Er wordt een steekproef van n objecten genomen, waarvan het aantal positieve objecten wordt bepaald. Gevraagd wordt de frequentieverdeling van dit aantal, dat we kortweg het resultaat zullen noemen.

De structuur van het vraagstuk is deze, dat we ons hebben voor te stellen dat alle mogelijke steekproeven van n objecten worden beschouwd en deze gerangschikt naar hun resultaat. Van elk resultaat dient de waarschijnlijkheid te worden bepaald.

Dit vraagstuk hebben we, voor het geval dat $p = \frac{1}{2}$, reeds onder het oog gezien in § 16. Beschouwen we nu het geval dat p willekeurig is. Het aantal positieve objecten in het universum is nu pN, waarvan we gemakshalve aannemen dat het een geheel getal is [1]).

We denken ons het universum van alle steekproeven nu aldus verkregen dat eerst alle mogelijkheden beschouwd worden die zich bij trekking van één object voordoen, vervolgens de daaruit af te leiden mogelijkheden voor trekking van

---

[1]) Voor een behandeling van het geval dat pN niet geheel is zie men J. Westenberg, Diss. Amsterdam 1935, blz. 28, e.v.

twee, enz. Bij trekking van één object zijn N gevallen denkbaar, waarvan pN positief. Aan elk dezer N gevallen kunnen wederom N worden toegevoegd door er een tweede object aan toe te voegen. We hebben dus trekking *met* terugleggen aangenomen. Dit doen we ter vermijding van enige complicaties, die echter voor dit geval in § 26 onder het oog zullen worden gezien.

Van de dan verkregen N^2 gevallen zullen er p^2N^2 een resultaat 2 geven: elk der pN gevallen, waarvoor het resultaat van het eerste object reeds positief was, groeit uit tot pN gevallen waarin ook het tweede object positief uitvalt. Evenzo zullen er q^2N^2 gevallen zijn die een resultaat 0 (twee negatieve objecten) opleveren. De gevallen met resultaat 1 zullen $2pqN^2$ in getal zijn; want uit de qN negatieve eerstelingen groeien pqN^2 gevallen met een negatief en een positief object en evenzo uit de pN positieve eerstelingen, waarbij zich een negatief tweede object voegt. De frequentieverdeling is dus:

Resultaat	2	1	0
Frequentie	p^2N^2	$2pqN^2$	q^2N^2

Voegen we nu een derde object toe, dan wordt het totaal aantal mogelijkheden N^3. Daarvan zijn er nu met een resultaat 3 behept p^3N^3; die met een resultaat 2 ontstaan op twee wijzen: uit de gevallen met resultaat 2 bij in totaal 2 objecten door toevoeging van een

negatief object, wat bij $p^2N^2 \times qN$ gevallen optreedt. En voorts uit de gevallen met resultaat 1 bij in totaal 2 objecten, door toevoeging van een positief object. Dit is bij $2pqN^2 \times pN$ gevallen zo. In totaal $3p^2qN^3$. Op deze wijze doorredenerend vindt men voor de verdeling bij steekproeven met 3 objecten:

Resultaat	3	2	1	0
Frequentie	p^3N^3	$3p^2qN^3$	$3pq^2N^3$	q^3N^3.

Evenzo vindt men voor het geval van n objecten:

Resultaat	Frequentie	Waarschijnlijkheid
n	$\binom{n}{0}\ p^nN^n$	$\binom{n}{0}\ p^n$
$n-1$	$\binom{n}{1}\ p^{n-1}qN^n$	$\binom{n}{1}\ p^{n-1}q$
$n-2$	$\binom{n}{2}\ p^{n-2}q^2N^n$	.
$n-3$	$\binom{n}{3}\ p^{n-3}q^3N^n$	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
3	$\binom{n}{n-3}\ p^3q^{n-3}N^n$	.
2	$\binom{n}{n-2}\ p^2q^{n-2}N^n$	.
1	$\binom{n}{n-1}\ pq^{n-1}N^n$	$\binom{n}{n-1}\ pq^{n-1}$
0	$\binom{n}{n}\ q^nN^n$	$\binom{n}{n}\ q^n$

De derde kolom in deze tabel is zodanig, dat de som der getallen 1 bedraagt, zoals ook nodig is.

Hiermede is de frequentieverdeling die gevraagd werd aangegeven. Het is de moeite waard om er het gemiddelde en de standaardafwijking van te berekenen. Het gemiddelde kan uit de aard der zaak niet anders zijn dan pn, hetgeen men overigens gemakkelijk narekent door te bedenken dat

$$(n - k) \binom{n}{k} = n \binom{n-1}{k}.$$

Daardoor wordt het gemiddelde gelijk aan:

$$n \binom{n}{0} p^n + (n-1) \binom{n}{1} p^{n-1}q + \\ + (n-2) \binom{n}{2} p^{n-2}q^2 \ldots + 0. \binom{n}{n} q^n = \\ = n \left[\binom{n-1}{0} p^n + \binom{n-1}{1} p^{n-1}q + \binom{n-1}{2} p^{n-2}q^2 + \\ + \ldots \binom{n-1}{n-1} pq^{n-1} \right] = np\,(p+q)^{n-1} = np.$$

§ 26. *Directe berekening van de standaarddeviatie*

De berekening van de *standaarddeviatie* kan op de gebruikelijke wijze uit de frequentieverdeling geschieden. Overzichtelijker is wellicht een berekening waarbij niet eerst de frequentieverdeling wordt berekend. Noemt men het „resultaat" van elk object afzonderlijk x_{ij}, waarbij i het nummer van de steekproef en j het nummer van het object binnen de steekproef aangeeft, dan is, ten eerste, deze x altijd òf 0 (bij een negatief object), òf 1 (bij een positief) en, ten tweede, de uitkomst van steekproef No. i

$$\sum_{1}^{n}{}_j\, x_{ij}.$$

De afwijking die deze uitkomst vertoont van het algemeen gemiddelde pn, bedraagt $\sum_{1}^{n}{}_j\, x_{ij} - pn$ en de standaarddeviatie van al de uitkomsten σ_n is gedefinieerd door

$$M\,\sigma_n^2 = \sum_{1}^{M}{}_i \left(\sum_{1}^{n}{}_j\, x_{ij} - pn\right)^2$$

waarbij M het totaal aantal mogelijke steekproeven beduidt. Deze uitdrukking is als volgt te transformeren:

$$M\,\sigma_n^2 = \sum_{1}^{M}{}_i \left[\sum_{1}^{n}{}_j\, (x_{ij} - p)\right]^2 = \sum_{1}^{M}{}_i \left[\sum_{1}^{n}{}_j\, (x_{ij} - p)^2 + \right.$$

$$\left. + \sum_{1}^{n}{}_{j'} \sum_{1}^{n}{}_j\, (x_{ij} - p)\,(x_{ij'} - p)\right]. \qquad (26.1)$$

Hierbij betekent j' een ander getal dan j, zodat in de [ ]-vorm de kwadratische termen van de producttermen zijn gescheiden. Beschouwen we nu de sommatie over i en j tezamen, dan kunnen we opmerken dat in deze sommen alle objecten uit het oorspronkelijke universum gelijkelijk voorkomen. Het aantal der termen $(x_{ij} - p)^2$ bedraagt Mn. In de som over i, j en j' voeren we in plaats van de twee indices i en j (resp. i en j') die een bepaald object kenmerken, een nieuwe notatie in, nl. het rangnummer k resp. k', dat het object in kwestie kenmerkt binnen het *oorspronkelijke* universum (en dus niet binnen een steekproef). Het is duidelijk dat dan verschillende waarden van j en j' kunnen corresponderen met eenzelfde k,

al naarmate men zich in verschillende steekproeven bevindt. Tevens is duidelijk dat éénzelfde k nu in vereniging van een ander rangnummer k' niet éénmaal, maar meerdere malen voorkomt. Evenveel malen als twee bepaalde objecten in eenzelfde steekproef voorkomen. En dat is gelijk aan het aantal malen dat men uit de $N - 2$ overige objecten uit het universum een $n - 2$-tal kan kiezen. Dit aantal bedraagt, in analogie met wat vroeger werd gevonden:

$$Z = \frac{(N-2)!}{(n-2)!\,(N-n)!}$$

Onze tweede som kan daarmee worden geschreven als

$$\frac{(N-2)!}{(n-2)!\,(N-n)!} \sum_1^N{}_k \sum_1^N{}_{k'} (x_k - p)\,(x_{k'} - p).$$

Daarin is nu $\sum_1^N{}_{k'} (x_{k'} - p)$, waarmee de term $x_k - p$, corresponderend met één vaste waarde van k, vermenigvuldigd wordt, te vervangen door $p - x_k$, aangezien

$$\sum_1^N{}_{k'} (x_{k'} - p) + x_k - p = 0$$

en we krijgen dus als uitkomst van de tweede som in (26.1)

$$-\frac{(N-2)!}{(n-2)!\,(N-n)!} \sum_1^N{}_k (x_k - p)^2.$$

Aangezien het totaal aantal mogelijke steekproeven $M = \dfrac{N!}{n!\,(N-n)!}$ en in de eerste

som in (26.1) steeds alle objecten symmetrisch voorkomen, wordt de gehele uitkomst:

$$\sigma_n^2 = n\,\sigma^2 - \frac{(N-2)!}{(n-2)!\,(N-n)!}\,\frac{n!\,(N-n)!}{N!}\,N\sigma^2$$

$$\sigma_n^2 = n\,\sigma^2 - \frac{n\,(n-1)}{N-1}\,\sigma^2 = \frac{N-n}{N-1}\,n\,\sigma^2 \qquad (26.2)$$

Deze formule geldt algemeen, d.w.z. voor elke waarde van N; nemen we aan dat N zeer groot is in vergelijking tot n, dan is de uitkomst:

$$\sigma_n^2 = n\,\sigma^2 \qquad (26.3)$$

Daarbij geeft σ^2 de standaarddeviatie aan van de verdeling der afzonderlijke objecten. Het is gemakkelijk te berekenen dat $\sigma^2 = pq$. Immers, voor de afzonderlijke objecten zijn slechts twee mogelijkheden aanwezig: de resultaten 0 en 1, met een waarschijnlijkheid van q en p. Aangezien het gemiddelde resultaat p bedraagt, is dus:

$$\sigma^2 = q\,(0-p)^2 + p\,(1-p)^2 = qp^2 + pq^2 = pq.$$

De uitkomst is dus ook te schrijven:

$$\sigma_n^2 = npq \qquad (26.4)$$

Wanneer wij onder het „relatieve resultaat" m verstaan het resultaat, gedeeld door het aantal n der objecten in een steekproef, dan wordt daarvan het gemiddelde voor alle steekproeven p en de standaarddeviatie $s_n = \frac{\sigma_n}{n} = \sqrt{\frac{pq}{n}}$.

Deze uitkomst brengt tot uitdrukking, dat, hoe groter n wordt, des te kleiner deze relatieve standaarddeviatie wordt.

Dit betekent dat, hoe groter n genomen wordt, de werkelijke waarde p van het gemiddelde van het gehele universum des te beter benaderd wordt. De waarschijnlijkheid b.v. om een afwijking tussen steekproefgemiddelde m en algemeen („werkelijk") gemiddelde p te vinden van $3 \times s_n$ is ongeveer 0.001 [1]) en bij toenemende n nadert s_n tot nul. Men noemt deze laatste uitspraak ook wel de *wet van de grote getallen* (vgl. § 18), hoewel het gebruik van deze veel omstreden uitdrukking beter vermeden kan worden [2]).

Wegens het zeer grote praktische belang van de zaak moge een enkel concreet voorbeeld worden gegeven. Nemen we aan dat uit een voorraad producten van een fabriek een steekproef van 100 exemplaren wordt onderzocht op deugdelijkheid — waarvoor zekere van te voren gekozen normen worden genomen. Laten er van deze 100 90 exemplaren goed zijn; dan is dus $m = 0.9$. Wij wensen echter p te kennen. *We nemen nu aan,* dat m niet veel van p verschilt en berekenen nu de s_n volgens onze formule:

$$s_n = \sqrt{\frac{pq}{n}} \backsim \sqrt{\frac{0.9 \times 0.1}{100}} = 0.03 = s_n'.$$

[1]) De verdeling is nl. voor grotere n's ongeveer normaal.

[2]) Vgl. b.v. R. v. Mises, Wahrscheinlichkeit, Statistik u. Wahrheit, blz. 78 e. v.; O. Anderson, Einführung in die mathematische Statistik, blz. 86—88.

Was deze s_n' de juiste s_n, dan zou deze uitkomst betekenen dat de p met 99.9% waarschijnlijkheid ligt binnen de grenzen $0.90 \pm 3 \times 0.03$ dus 0.81 en 0.99. Wanneer nu s_n' niet precies juist is dan komt hierin toch slechts een geringe verandering. Het ongunstigst wanneer p in de buurt van 0.81 ligt. Dan wordt

$$s_n'' = \sqrt{\frac{0.81 \times 0.19}{100}} = 0.04$$

hetgeen niet veel verschil maakt.

Is het er om te doen om te weten dat minstens $^3/_4$ van de partij aan de gestelde eisen voldoet, dan is men in dit geval dus reeds praktisch zeker van zijn zaak.

Gewoonlijk zal echter een grotere nauwkeurigheid vereist worden. Onze formule levert de mogelijkheid tot een schatting van het aantal exemplaren dat men daarvoor minstens moet onderzoeken. Stel b.v. dat men een onzekerheidsmarge van hoogstens 0.02 wenst (met een waarschijnlijkheid van 99.9%). Dan moet dus blijkbaar

$$s_n = \tfrac{1}{6} \times 0.02 = 0.00333\,..$$

dus $$s_n^2 = \frac{pq}{n} = 0.0000111\,..$$

Aannemende dat de gevonden waarde voor p een goede eerste benadering is voor p, hebben we $pq = 0.9 \times 0.1 = 0.09$ en dus

$$0.0000111\,..\,n = 0.09 \quad \text{of } n = 8100.$$

Er moet dan dus een steekproef van 8100 ex. genomen worden. (Men ziet gemakkelijk in dat bij een kleinere q het getal n aanmerkelijk geringer kan zijn). Is men met de eerste benadering niet tevreden, dan kan op grond van de berekende standaarddeviaties een tweede, desgewenst een derde benadering berekend worden, waarmee men gewoonlijk het doel snel bereikt.

Als tweede voorbeeld kan er een worden aangehaald uit Johannsen, Elemente der exakten Erblichkeitslehre (Jena 1909), blz. 95, betr. het percentage der jongens onder pasgeborenen. In een kliniek waarin 19 jaar in 't geheel 32410 kinderen werden geboren, werd een jongenspercentage $p' = 52.09\%$ gevonden. Johannsen berekent nu daaruit de standaarddeviatie van het „universum" der afzonderlijke geboorten:

$$\sigma' = \sqrt{p'q'} = 49.96\%$$

$$s_n = \sqrt{\frac{pq}{n}} \backsim \frac{49.96}{\sqrt{32410}} = 0.28\%.$$

Het werkelijke cijfer der jongensgeboorten ligt dus met 99.9% waarschijnlijkheid tussen

$52.09 \pm 3 \times 0.28$ of 51.25 en 52.93%.

Voorts berekent schr. de standaarddeviatie voor het cijfer dat men gemiddeld per jaar — waarin gem. 1706 geboorten vielen — moet verwachten:

$$s_{1706} = \frac{49.96}{\sqrt{1706}} = 1.21.$$

Daaruit volgt dat het niet onmogelijk is dat

in afzonderlijke jaren eens een meisjesoverschot wordt gevonden; want $3 \times 1.21 = 3.63$ en $52.09 - 3.63 < 50$. Inderdaad wordt voor 1896 een jongenspercentage van 49.33 gevonden. Daartegenover in 1903: 54.81%.

Een enkel woord moge tenslotte nog worden gezegd over het geval waarbij de N niet groot in vergelijking tot n wordt verondersteld. Zoals formule (26.2) doet zien, worden dan de standaarddeviaties nog kleiner. Dit is ook begrijpelijk, omdat voor $n = N$ er van een standaarddeviatie geen sprake kan zijn.

§ 27. *Waargenomen en ware standaarddeviatie*

Het bereikte resultaat stelt ons voorts in staat om het onderscheid te bespreken tussen de *waargenomen standaarddeviatie* en de zgn. *ware standaarddeviatie* bij een serie van n waarnemingen $x_1 \ldots x_n$ van een zekere grootheid. Laten we het gemiddelde dezer grootheden x noemen en het gemiddelde van het universum waaruit ze zijn getrokken p. Onder de ware standaarddeviatie wordt dan verstaan de uitdrukking

$$\sigma_w = \sqrt{\frac{\Sigma (x_i - p)^2}{n}}. \qquad (27.1)$$

Deze moet zowel onderscheiden worden van de waargenomen standaarddeviatie

$\sigma' = \sqrt{\frac{\Sigma (x_i - x)^2}{n}}$ als van de onbekend ge-

dachte standaarddeviatie in het universum.

Om σ_w met het beschikbare materiaal zo goed mogelijk te benaderen, wordt van de volgende identiteit uitgegaan: $\Sigma(x_i - p)^2 = \Sigma(x_i - x + x - p)^2 = \Sigma(x_i - x)^2 + \Sigma(x - p)^2$, aangezien $\Sigma(x_i - x)(x - p) = (x - p)\,\Sigma(x_i - x) = 0$ is.

Delende door n krijgen we:

$$\frac{\Sigma(x_i - p)^2}{n} = \frac{\Sigma(x_i - x)^2}{n} + \frac{\Sigma(x - p)^2}{n}.$$

De laatste term beschouwen we als een benadering van σ_n^2, het kwadraat van de standaarddeviatie der relatieve steekproefresultaten, welke zelf weer benaderd kan worden door $\frac{1}{n}\sigma_w^2$, hoewel eigenlijk $\sigma_n^2 = \frac{1}{n}\sigma^2$. We krijgen zodoende

$$\sigma_w^2 = \sigma'^2 + \frac{1}{n}\sigma_w^2 \quad \text{of} \quad \sigma_w^2 = \frac{n}{n-1}\sigma'^2$$

hetgeen ook direct in de x_i kan worden uitgedrukt door:

$$\sigma_w^2 = \frac{\Sigma(x_i - x)^2}{n - 1} \tag{27.2}$$

Er wordt dus nu door $n - 1$ gedeeld in plaats van door n, zoals we tot hier toe deden. Verschillende schrijvers definiëren op deze wijze de standaarddeviatie van het begin af aan. Men vindt dus deze formule naast de vroeger gebruikte en men vindt niet altijd verschil gemaakt tussen de begrippen waargenomen en ware standaard-

deviatie. Voor grote waarden van n is het onderscheid natuurlijk gering.

We zullen de formule (27.2) verder als „de gecorrigeerde formule" aanhalen.

§ 28. *Tweede geval. Steekproeven uit niet-homogeen materiaal (geval van Lexis)*

Vervolgens willen we overgaan tot de behandeling van een ingewikkelder geval, nl. een geval waarbij de achtereenvolgende steekproeven worden genomen uit materiaal, waarvan de frequentieverdeling verschillend is, dus verschillende waarschijnlijkheden p_1, p_2 .... oplevert voor het hebben van een succes [1]). Daarbij willen we het ons in zoverre gemakkelijk maken, dat we nu aannemen dat elk dezer universa op zich zelf een groot aantal objecten bevat, zodat we de in de vorige paragraaf op grond daarvan aangebrachte vereenvoudigingen kunnen gebruiken.

Nemen we aan dat er in 't geheel M_1 steekproeven kunnen worden genomen uit een universum met de waarschijnlijkheid p_1 van een positief resultaat; M_2 steekproeven met p_2,

[1]) Men kan zich daarbij denken dat de afwijkingen tussen de resultaten en de bijbehorende p het gevolg zijn van meetfouten, terwijl de onderlinge afwijkingen tussen de p's reëel zijn. Men noemt de eerste dan wel „erreurs", de laatste „écarts". Deze interpretatie is natuurlijk niet de enige, doch slechts een uit meerdere.

enz,, waarbij M_f en p_f de laatste groep aanduiden en de index in 't algemeen geschreven wordt als h. Vgl. schema fig. 13. De gezochte

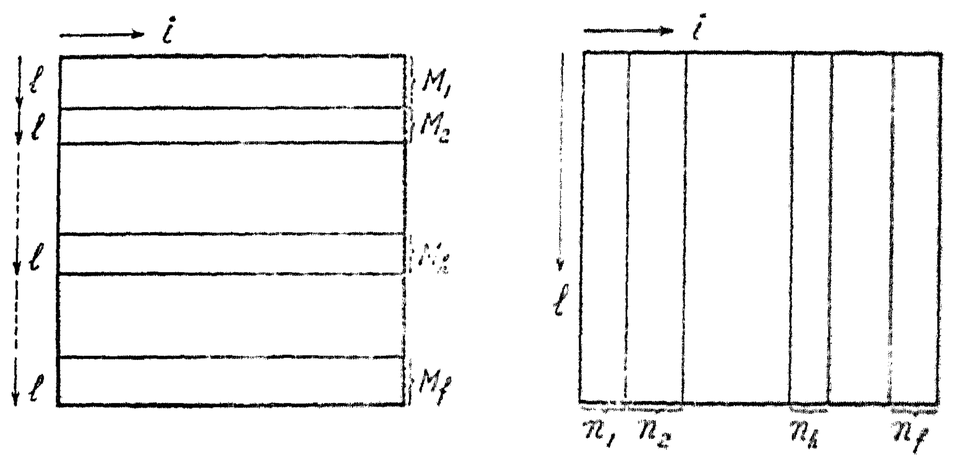

Fig. 13. Schematische voorstelling van de gevallen van Lexis en Poisson.

standaarddeviatie in dit geval σ_L wordt nu gevonden uit:

$$M\,\sigma_L^2 = \sum_1^f{}_h \sum_1^{M_h}{}_l \left[\sum_1^n{}_i (x_{ilh} - p)\right]^2 \qquad (28.1)$$

waarbij p de gemiddelde waarschijnlijkheid van een positief resultaat voorstelt, dus $p\,\Sigma\,M_h = \Sigma_h\,M_h\,p_h$; inderdaad stelt [ ] de afwijking voor tussen het resultaat van de l-de steekproef in de groep h en het algemeen gemiddelde der resultaten np. Het kwadraat van deze afwijking moet worden gesommeerd over alle denkbare steekproeven in alle groepen. Voor de uitwerking van (28.1) is het doelmatig om voor [ ] te schrijven:

$$\sum_1^n{}_i (x_{ilh} - p_h + p_h - p) = \sum_1^n{}_i (x_{ilh} - p_h) + (p_h - p)\,n,$$

waarvan het kwadraat is

$$[\sum_1^n{}_i\,(x_{ilh} - p_h)]^2 - 2\,n\,(p_h - p)\sum_1^n{}_i\,(x_{ilh} - p_h) +$$
$$+\ n^2\,(p_h - p)^2.$$

Bij sommeren van deze uitdrukking over l wordt de eerste term volkomen analoog met de eerste som in (26.1) bij het eenvoudigste geval, en kan *voor grote* M_h, gelijk gesteld worden aan $M_h n\,\sigma_h^2$; de tweede term wordt nul wegens de betekenis van p_h, terwijl de derde term voor alle steekproeven bij de groep constant is en dus na deze sommering oplevert

$$M_h n^2\,(p_h - p)^2.$$

Daarna sommerend over alle groepen krijgen we

$$M\,\sigma_L^2 = n\sum_1^f{}_h\,M_h\,\sigma_h^2 + n^2\sum_1^f{}_h\,M_h\,(p_h - p)^2 =$$
$$= n\sum_1^f{}_h\,M_h\,p_h\,q_h + n^2\sum_1^f{}_h\,M_h\,(p_h - p)^2 =$$
$$= n\sum_1^f{}_h\,M_h\,(p_h - p_h^2) + n^2\sum_1^f{}_h\,M_h\,(p_h - p)^2.$$

Voor de eerste som is te schrijven:

$$n\,Mp - n\sum_1^f{}_h\,M_h\,(p_h - p + p)^2 =$$
$$= n\,Mp - n\sum_1^f{}_h\,M_h\,[(p_h - p)^2 + 2p\,(p_h - p) +$$
$$+\ p^2] = n\,Mp - n\sum_1^f{}_h\,M_h\,(p_h - p)^2 - 0 - np^2\sum_1^f{}_h\,M_h.$$

Nu is $\sum_1^f{}_h\,M_h\,(p_h - p)^2$ niets anders dan M keer de standaarddeviatie σ_p van de getallen p,

terwijl $\sum_{1}^{f}{}_h M_h = M$; daardoor wordt de eind-uitkomst

$$M\sigma_L^2 = n\,Mp - n\,M\sigma_p^2 - n\,Mp^2 + n^2\,M\sigma_p^2$$

of

$$\sigma_L^2 = n\,pq + n\,(n-1)\,\sigma_p^2 \qquad (28.2)$$

Zouden we ook hier overgaan tot de relatieve resultaten en de standaarddeviatie s_L^2 daarvan, dan is dus:

$$\boxed{s_L^2 = \frac{pq}{n} + \frac{n-1}{n}\,\sigma_p^2} \qquad (28.3)$$

Bij deze uitkomst valt op te tekenen dat in de eerste plaats dus niet de gehele frequentieverdeling van de p's bekend behoeft te zijn, doch slechts de standaarddeviatie er van, en in de tweede plaats dat dus de standaarddeviatie σ_L groter is dan de σ_n bij een homogeen universum, en dat de afwijking tussen deze twee zijn maximum bereikt voor $n = \infty$, terwijl ze voor $n = 1$ verdwijnt. Dit laatste kan een middel zijn om door verschillende n's achter elkaar te nemen een splitsing te maken in de twee bestanddelen van de uitkomst. Dit betekent dus het op een andere wijze hanteren van de formule als voor het eerste geval werd aangegeven; nl. het nemen van een vrij groot aantal steekproeven ter contrôle van de formule van σ_L; eventueel dus voor twee uiteenloopende waarden

van n. Natuurlijk kan men de formule ook gebruiken voor het doel dat in § 26 beschreven werd, wanneer men langs andere weg een denkbeeld kan krijgen van de waarde van σ_p.

§ 29. *Gemengde steekproeven (geval van Poisson)*

Een juist tegenovergesteld resultaat verkrijgt men, wanneer in elke steekproef een vast aantal n_1 uit een eerste universum met waarschijnlijkheid p_1, een aantal n_2 met p_2, enz. wordt genomen (algemeen n_h met p_h, h lopend van 1 tot f). In dit geval wordt nl. de standaarddeviatie kleiner dan in het eenvoudige geval. We vinden nu nl.:

$$M\sigma_p^2 = \sum_1^M{}_l \left[\sum_1^f{}_h \sum_1^{n_h}{}_i (x_{ilh} - p)\right]^2 =$$

$$= \sum_1^M{}_l \left[\sum_1^f{}_h \sum_1^{n_h}{}_i (x_{ilh} - p_h)\right]^2 = \sum_1^M{}_l \left[\sum_1^f{}_h \sum_1^{n_h}{}_i (x_{ilh} - p_h)^2 + \right.$$

$$+ \sum_1^f{}_h \Sigma_i \Sigma_{i'} (x_{ilh} - p_h)(x_{i'lh} - p_h) +$$

$$\left. + \Sigma_h \Sigma_{h'} \Sigma_i \Sigma_{i'} (x_{ilh} - p_h)(x_{i'lh'} - p_{h'})\right]. \quad (29.1)$$

Hierin symboliseert de tweede som binnen de [ ]-vorm alle producten betrekking hebbend op objecten binnen de zelfde groep binnen één steekproef; daarin is $i' \neq i$. De derde som heeft betrekking op producten betr. objecten van verschillende groepen binnen dezelfde steekproef; daarbij is $h' \neq h$, doch doorlopen i en i' beide alle waarden. Deze beide sommen zullen in de einduitkomst verdwijnen; de laatste doet

dat reeds, wanneer althans M groot wordt aangenomen, binnen elke groep, na sommatie over l; bij de tweede kan men eerst een vast object uit de $h$de groep van het oorspronkelijke universum nemen en constateren dat na sommatie over l de cofactor nul wordt. Als einduitkomst krijgen we zodoende slechts

$$M\,(n_1\,\sigma_1^2 + n_2\,\sigma_2^2 + \ldots\ldots n_f\,\sigma_f^2).$$

Hierin is dan $\sigma_1^2 = p_1 q_1$, enz., zodat de uitkomst te schrijven is als

$$M \sum_1^f{}_h\, n_h\,(p_h - p_h^2) =$$

$$= M \sum_1^f{}_h\, n_h\, p_h - M \sum_1^f{}_h\, n_h\,(p_h - p + p)^2$$

en verder op analoge wijze als in § 28 werd aangegeven. Het resultaat is

$$\boxed{\sigma_p^2 = n\,pq - n\,\sigma_p^2} \qquad (29.2)$$

Als een eerste *voorbeeld* nemen we het geval dat we met 2 dobbelstenen werpen, waarbij van elk 3 kanten wit en 3 rood zijn. Blijkbaar is hier

$n = 2,\ n_1 = n_2 = 1,\ p_1 = p_2 = q_1 = q_2 = \frac{1}{2},\ \sigma_p = 0.$

De uitkomst is dan

$$\sigma_p^2 = \tfrac{1}{2}. \qquad (29.3)$$

Nu veranderen we de zaak aldus dat van de ene steen 5 zijden wit en 1 rood zijn en het bij de tweede steen net andersom is. We hebben nu

$n = 2,\ n_1 = n_2 = 1,\ p_1 = q_2 = \frac{5}{6},\ p_2 = q_1 = \frac{1}{6}$
en dus: $\sigma_p = \frac{1}{3},\ p = \frac{1}{2},\ q = \frac{1}{2}$.
We vinden zodoende voor

$$\sigma_p^2 = 2.\tfrac{1}{2}.\tfrac{1}{2} - 2\left(\tfrac{1}{3}\right)^2 = \tfrac{5}{18}. \qquad (29.4)$$

Een uiterste geval, zeer geschikt om de formule op de proef te stellen, krijgen wij als we nu aannemen dat de ene dobbelsteen uitsluitend witte en de andere uitsluitend rode kanten heeft. Dan is dus $p_1 = q_2 = 1,\ p_2 = q_1 = 0$, waardoor p en q wederom dezelfde waarde behouden: $p = q = \frac{1}{2}$. Nu is echter $\sigma_p = \frac{1}{2}$ geworden en daardoor wordt

$$\sigma_p^2 = 0. \qquad (29.5)$$

Dit klopt geheel met het feit dat er thans volkomen zekerheid heerst omtrent de uitkomsten; immers men kan niet anders werpen dan eenmaal rood en eenmaal wit.

§ 30. *Fisher's „Analysis of Variance"*

Met de zoëven behandelde steekproefproblemen zeer verwant zijn de problemen die een plaats vinden in de zgn. *analysis of variance*, genoemd naar de bekende Engelse statisticus R. A. Fisher [1]).

[1]) Voor een zeer duidelijke en praktische uiteenzetting in het Nederlands zij verwezen naar W. A. Bigot, De toepassing van Analysis of Variance bij proefvelden, Landbouwkundig Tijdschrift 1934, blz. 640 en 705. Voorts zij ook verwezen naar Prof. Dr. M. J. van Uven, Mathematical Treatment of the Results of Agricultural and Other Experiments, Groningen 1935 en J. Westenberg, Diss. Amsterdam 1935.

Om een denkbeeld te geven van de aard der hier behandelde vraagstukken moge een eenvoudig voorbeeld worden behandeld. Voor de details en de moeilijker problemen moet al weer verwezen worden naar de literatuur.

Gemeten moge worden een grootheid x die afhankelijk is van twee discontinu veranderend gedachte grootheden, die we daardoor als indices aan x kunnen toevoegen: x_{ik}. In de proefveldentechniek, waar deze methode veelvuldig toepassing vindt, kan b.v. i aangeven het nummer van het „blok" en k het nummer van het bemestingsrecept.

Voor den niet-landbouwkundigen lezer moge een kleine uitwijding gegeven worden ten einde de aard van deze begrippen duidelijk te maken. Men denke zich dat een aantal verschillende bemestingsrecepten op hun resultaten onderzocht moeten worden, d.w.z. dat nagegaan moet worden of de ene bemestingswijze een duidelijk beter resultaat geeft dan de andere, enz. De recepten worden genummerd met de index k, welke b.v. van 1 tot n loopt. Ten einde verschillen in natuurlijke vruchtbaarheid van de proefvelden geheel als toevallige afwijkingen te kunnen beschouwen, worden voor elke k een aantal percelen ingericht en wel zó dat telkens alle recepten k één perceel hebben binnen een „blok", een groter perceel, waarvan het gehele veld er weer

verscheidene bevat, zeggen we m, welke genummerd worden door de index i.

In het geheel zijn nu dus beschikbaar $m\,n$ waarden x_{ik}. Daaruit kunnen berekend worden:

1. alle receptgemiddelden, d. w. z. de gemiddelde uitkomst voor alle percelen waarin eenzelfde recept is toegepast:

$$x_{\cdot k} = \frac{1}{m} \sum_{1}^{m}{}_i \, x_{ik} \qquad (k = 1 \ldots . n) \qquad (30.1)$$

2. alle blokgemiddelden:

$$x_{i\cdot} = \frac{1}{n} \sum_{1}^{n}{}_k \, x_{ik} \qquad (i = 1 \ldots . m) \qquad (30.2)$$

3. het algemene gemiddelde:

$$x = \frac{1}{m\,n} \sum_{1}^{m}{}_i \sum_{1}^{n}{}_k \, x_{ik} \qquad (30.3)$$

Het is nu duidelijk wat de betekenis van de navolgende uitdrukkingen is:

$\bar{x}_{\cdot k} - \bar{x}$: de invloed van een bepaald recept k op de uitkomst;

$\bar{x}_{i\cdot} - \bar{x}$: de invloed van een bepaald blok i op de uitkomst;

$x_{ik} - x - (\bar{x}_{\cdot k} - \bar{x}) - (\bar{x}_{i\cdot} - \bar{x}) = x_{ik} - \bar{x}_{\cdot k} - \bar{x}_{i\cdot} + \bar{x}$, de afwijking van een individueel perceel van het algemeen gemiddelde, na correctie voor de „receptinvloed" en de „blokinvloed".

De kwadraten der overeenkomstige standaarddeviaties zullen zijn resp.:

$$\frac{\sum_{1}^{n}{}_k\ (\overline{x}_{.k} - \overline{x})^2}{n-1}, \quad \frac{\sum_{1}^{m}{}_i\ (\overline{x}_{i.} - \overline{x})^2}{m-1} \quad \text{en}$$

$$\frac{\sum_{1}^{m}{}_i \sum_{1}^{n}{}_k\ (x_{ik} - \overline{x}_{.k} - \overline{x}_{i.} + \overline{x})^2}{m\,n-1}$$

wanneer we de „gecorrigeerde formules" gebruiken (zie § 27).

Denken we ons nu een ogenblik dat de verdeling van de x_{ik}'s een volkomen Gaussiaanse zou zijn, zonder enige systematische invloed van de i en de k. Dan zou de waarde $\overline{x}_{.k}$ voor elke willekeurige k niet anders zijn dan de uitkomst van een steekproef van m gevallen. We weten uit de elementaire steekproeftheorie (§ 26), dat de standaarddeviatie van een dergelijke uitkomst tendeert naar een waarde die $\frac{1}{\sqrt{m}} \times$ zo groot is als de standaarddeviatie σ der x_{ik}'s in het oorspronkelijke universum. Evenzo zouden de $\overline{x}_{i.}$'s niets anders zijn dan de uitkomsten van een steekproef van n gevallen, en de standaarddeviatie daarvan zal bij een groot aantal waarnemingen $\frac{1}{\sqrt{n}} \times$ zo groot zijn als σ. Van de standaarddeviatie der uitdrukking

$$x_{ik} - \overline{x} - (\overline{x}_{.k} - \overline{x}) - (\overline{x}_{i.} - \overline{x})$$

laat zich door berekening gemakkelijk bewijzen, dat ze naar

$$\sqrt{\frac{(m-1)(n-1)}{mn-1}} \times \sigma$$

tendeert, bij een groeiend aantal waarnemingen.

In het gestelde geval van een normale verdeling der x_{ik}'s zouden we zodoende op drie wijzen de waarde van σ^2 kunnen benaderen en wel door te nemen resp.

$$\frac{m \sum_1^n{}_k \ (x_{.k} - \bar{x})^2}{n-1}, \quad \frac{n \sum_1^m{}_i \ (x_{i.} - \bar{x})^2}{m-1} \quad \text{en}$$

$$\frac{\sum_1^m{}_i \ \sum_1^n{}_k \ (x_{ik} - x_{i.} - x_{.k} + \bar{x})^2}{(m-1)\ (n-1)}$$

hetgeen ook geschreven kan worden als

$$\frac{\Sigma_i \ \Sigma_k \ (x_{.k} - \bar{x})^2}{n-1}, \quad \frac{\Sigma_i \ \Sigma_k \ (x_{i.} - \bar{x})^2}{m-1} \quad \text{en}$$

$$\frac{\Sigma_i \ \Sigma_k \ (x_{ik} - x_{i.} - x_{.k} + \bar{x})^2}{(m-1)\ (n-1)}. \qquad (30.4)$$

Zodra echter er systematische verschillen zouden optreden tussen de receptgemiddelden onderling en de blokgemiddelden onderling, zouden de eerste twee uitdrukkingen groter worden. De vraag is, of aan een van 1 afwijkende verhouding tussen die uitkomsten, zeggen we s_1^2 en s_2^2 resp. s_2^2 en s_3^2, betekenis als binnen de toevalsgrenzen liggend moet worden beschouwd. Hier kan een later te bespreken frequentieverdeling uitwijzen hoe groot de waarschijnlijkheid is dat een gegeven afwijking van

1 of een grotere optreedt. Fisher gebruikt, zoals in § 32 nog zal worden behandeld, de log. $\frac{s_1}{s_2} = z$ en de daarmee corresponderende waarschijnlijkheidsverdeling. Bij gegeven m en n kan men nu in tabellen (gegeven op de blz. 224 — 227 van de 4e druk van zijn Statistical Methods for Research Workers) de waarde van z aflezen, waarvan het overschrijden resp. 5% en 1% kans heeft. Is de in de werkelijkheid gevonden z kleiner dan de eerstgenoemde waarde, dan is, naar Fishers voorstel, s_1 slechts toevallig van s_2 verschillend; is daarentegen z groter dan de laatstgenoemde drempelwaarde, dan neemt hij een wezenlijk verschil aan en dus een systematische invloed van het gebezigde recept. Op dezelfde wijze kan men vaststellen of er een systematische invloed van het blok op de uitkomsten is, d.w.z. of er dus een „gang in de vruchtbaarheid" van de grond is waarover zich de proefvelden uitstrekken.

Er moge tenslotte nog op gewezen worden dat de drie uitdrukkingen (30.4), afgezien van de getallenfactoren, als termen beschouwd kunnen worden die optreden, wanneer men de standaarddeviatie der x_{ik}'s berekent. Men heeft n.l.:

$$\begin{aligned} \Sigma_i \Sigma_k (x_{ik} - \bar{x})^2 &= \Sigma_i \Sigma_k (\bar{x}_{i.} - \bar{x})^2 + \\ &+ \Sigma_i \Sigma_k (\bar{x}_{.k} - \bar{x})^2 + \\ &+ \Sigma_i \Sigma_k (x_{ik} - \bar{x}_{i.} - \bar{x}_{k.} + \bar{x})^2 \end{aligned} \qquad (30.5)$$

§ 31. *Meerdimensionale steekproeven*

De standaarddeviatie van de correlatiecoëfficiënt en van het verschil tussen twee grootheden

Als voorbeeld van een steekproefprobleem van een universum van twee dimensies behandelen we eerst het volgende vraagstuk. Gegeven een universum van bijeenbehorende paren van getallen, waarvan het ene getal steeds aangeeft de waarde van een verschijnsel X, het tweede de waarde van een verschijnsel Y. Het universum heeft dus het karakter van een tweedimensionale frequentieverdeling, of, als men wil, een correlatieoppervlak. Er zal in het universum een correlatiecoëfficiënt r bestaan tussen de reeks der getallen X en die der getallen Y. We nemen nu een steekproef van n paren uit dit universum, waardoor we een reeks X en een reeks Y ter lengte van n verkrijgen.

Dit verkrijgen we zoveel malen als een dergelijke steekproef op verschillende wijzen kan worden verkregen en we vragen naar de standaarddeviatie van de correlatiecoëfficiënt ϱ die daarbij telken male verkregen wordt. Deze standaarddeviatie σ_ϱ zal een beoordeelingsmaatstaf zijn voor de „fout" in de gevonden waarde ϱ_0 in een concreet geval, d.w.z. het verschil dat er tussen ϱ_0 en r kan bestaan.

We zullen dit vraagstuk echter niet in zijn algemene vorm oplossen, doch het zodanig vereenvoudigen dat het zo gemakkelijk mogelijk

te behandelen is zonder dat essentiële bestanddelen er van verloren zijn gegaan. Daartoe nemen we niet aan dat er getallen*paren* in het universum aanwezig zijn, doch getallen-viertallen, waarvan de 2 eerste steeds X-waarden voorstellen, de 2 laatste Y-waarden. Bovendien nemen we aan, dat er maar twee soorten van dergelijke viertallen aanwezig zijn, waarvan de waarschijnlijkheid resp. p en q bedraagt, nl.

X:	$+1 \quad -1$	$+1 \quad -1$
Y:	$+1 \quad -1$	$-1 \quad +1$
w:	p	q

Daaruit laat zich onmiddellijk berekenen welke waarde r heeft:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{2p - 2q}{2p + 2q} = \frac{p - q}{p + q} = p - q. \tag{31.1}$$

Door oplossen van p en q uit de vergelijkingen $p + q = 1$ en $p - q = r$ vinden we:

$$p = \frac{1 + r}{2} \quad \text{en} \quad q = \frac{1 - r}{2}. \tag{31.2}$$

Wanneer nu n trekkingen verricht worden waarvan k $\times$ het eerste viertal, dan is voor die steekproef de

$$\varrho = \frac{k - (n - k)}{n} = \frac{2k - n}{n} \tag{31.3}$$

Dit bedrag is nu te vergelijken met het „resultaat" van de steekproeven, waarvan steeds sprake was in § 26. In die § noemden we het relatieve

resultaat $\frac{k}{n}$; hier is het $2\frac{k}{n} - 1$. Dat brengt mede, dat de standaarddeviatie in ons geval 2 × zo groot is, omdat we alle resultaten in 2 × zo grote getallen uitdrukken. De standaarddeviatie was in § 26 gelijk aan $\sqrt{\frac{pq}{n}}$; hier wordt dat, nadat we p en q hebben uitgedrukt in r:

$$2\sqrt{\frac{pq}{n}} = 2\sqrt{\frac{(1+r)(1-r)}{4n}} = \sqrt{\frac{1-r^2}{n}}.$$

We hebben dus voor het beschouwde eenvoudige geval bewezen:

$$\sigma_r = \sqrt{\frac{1-r^2}{n}}. \qquad (31.3)$$

Deze formule wordt als eerste benadering, voor $\varrho - r$ klein, ook verkregen wanneer de reeksen X en Y alle willekeurige waarden kunnen aannemen, mits ze normaal verdeeld zijn [1]). Voor een afleiding van de formule onder deze voorwaarden zij verwezen naar Derksen, Inleiding tot de Correlatierekening, blz. 87, waar men tevens de formules vindt zowel voor de standaardfouten voor de correlatieratio en de partiële en multipele correlatiecoëfficiënten als ook voor de regressiecoëfficiënten.

[1]) Dan neemt men daarbij echter ook gewoonlijk aan dat elke waarneming van de andere onafhankelijk is, waardoor het aantal onafhankelijke waarnemingen n tevens het aantal termen van de reeks is.

Standaarddeviatie van het verschil van twee grootheden

Als een verdere toepassing behandelen we de vraag, de standaarddeviatie te bepalen van het verschil tussen twee grootheden waarvan de gemiddelden gelijk zijn. Laten we die grootheden aangeven met x_i en y_i; we hebben dan:

$$M\sigma_d^2 = \Sigma(x_i - y_i)^2 = \Sigma\{(x_i - x) - (y_i - y)\}^2 =$$
$$= \Sigma(x_i - x)^2 + \Sigma(y_i - y)^2 - 2\Sigma(x_i - x)(y_i - y) =$$
$$= M(\sigma_x^2 + \sigma_y^2 - 2\, r_{xy}\, \sigma_x\, \sigma_y)$$

of

$$\sigma_d^2 = \sigma_x^2 + \sigma_y^2 - 2\, r_{xy}\, \sigma_x\, \sigma_y.$$

Wanneer de twee grootheden bovendien als ongecorreleerd kunnen worden beschouwd, verkrijgt men de bekende formule:

$$\sigma_d = \sqrt{\sigma_x^2 + \sigma_y^2} \qquad (31.4)$$

ook in de foutentheorie veel gebruikt.

§ 32. *Steekproeven uit continue massa's. De verdelingen van Student en Fisher*

In principe brengen de vraagstukken van steekproeven uit massa's waarin niet enkele discontinue kenmerken, doch een gehele reeks van continu in elkaar overgaande mogelijkheden bestaan geen nieuwe vraagstukken. De mathematische behandeling is echter, speciaal voor de weinig wiskundig geschoolde lezers, minder doorzichtig en om die reden moge een enkele poging tot verheldering in deze materie hier worden gewaagd. Als voorbeeld zal hier 't een

en ander worden gezegd over de in de literatuur beroemd geworden „*verdeling van Student*" [1]).

Denken we ons een universum waarin als kenmerk optreedt een continu veranderlijke x met een normale verdeling, d.w.z. dat dus de waarschijnlijkheid van het optreden van een waarde tussen x en $x + dx$ wordt gegeven door

$$\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} dx$$

waarin m het gemiddelde en σ de standaarddeviatie van x is. Laat er een steekproef van n objecten genomen worden. De waarschijnlijkheid dat deze steekproef een resultaat oplevert dat voor de resp. objecten ligt tussen x_1 en $x_1 + dx_1$, x_2 en $x_2 + dx_2$ .... enz. zal dan — naar analogie van wat in § 25 behandeld werd — gelijk zijn aan het gedurig product van de waarschijnlijkheden voor het optreden der afzonderlijke waarden:

$$\frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\Sigma(x_i-m)^2} dx_1\, dx_2 \ldots . dx_n. \tag{32.1}$$

Voor de exponent kan worden geschreven:

$$\sum_{i=1}^{n} (x_i - m)^2 = \Sigma (x_i - \bar{x} + \bar{x} - m)^2 =$$
$$= \Sigma (x_i - \bar{x})^2 + 2\,\Sigma (x_i - \bar{x})(\bar{x} - m) +$$
$$+ \Sigma (\bar{x} - m)^2 = \Sigma (x_i - \bar{x})^2 + n\,(\bar{x} - m)^2,$$

[1]) Deze naam is een pseudoniem voor W. S. Gosset.

waarbij $\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots . x_n)$

het gemiddelde voor de steekproef is.
Als we de standaarddeviatie van de steekproef weergeven door s, dan is

$$ns^2 = \Sigma (x_i - \bar{x})^2$$

en wordt de waarschijnlijkheid (32.1):

$$\frac{1}{(\sigma \sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}[ns^2 + n(\bar{x} - m)^2]} dx_1 \, dx_2 \ldots dx_n. \tag{32.2}$$

Men kan nu een aantal berekeningen in dit gebied zeer vergemakkelijken door gebruik te maken van de stellingen van de zgn. meerdimensionale meetkunde. Men vat dan de combinatie van waarden $(x_1 \, x_2 \ldots x_n)$ op als de coördinaten van een punt in een n-dimensionale ruimte. Voor $n = 3$ is dat nog voorstelbaar; voor hogere n's niet; doch de formules voor de 3-dimensionale meetkunde kunnen worden gegeneraliseerd en het is daarbij gemakkelijk, om van de meetkundige termen gebruik te blijven maken. De niet met deze werkwijze vertrouwde lezer zal de volgende formules dus niet gemakkelijk kunnen doorzien.

Uit de laatste formule blijkt dat de waarschijnlijkheid van het optreden van steekproeven met eenzelfde $\bar{x}$ en s dezelfde is (wanneer ook de elementaire intervallen $dx_1 \ldots . dx_n$ gelijk worden gekozen). Al dergelijke steekproeven liggen op

eenzelfde $(n-1)$-dimensionaal „boloppervlak". Men bewijst [1]) dat voor al dergelijke gevallen de gezamenlijke waarschijnlijkheid (die dus gelijk is aan de inhoud van dat „boloppervlak" maal de integrand in de laatste formule) gelijk is aan

$$\frac{n^{\frac{n}{2}}}{2^{\frac{n-2}{2}}\,\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\sigma^n}\, e^{-\frac{ns^2}{2\sigma^2}}\, s^{n-2}\, ds\, e^{-\frac{n(\bar{x}-m)^2}{2\sigma^2}} \qquad (32$$

Hierin wordt nu een nieuwe veranderlijke ingevoerd en wel „*Student's ratio*" $z = \dfrac{\bar{x}-m}{S}$ (32.4)

en vervolgens wordt gevraagd naar de waarschijnlijkheid van het optreden van een bepaalde z, onverschillig welke waarde van s daarbij nu verder optreedt. Deze waarschijnlijkheid verkrijgt men door te integreren over alle waarden van s bij een gegeven waarde van z en het resultaat is:

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}\,(1+z^2)^{-\frac{n}{2}}\,dz \qquad (32.5)$$

welke verdelingswet de „*verdeling van Student*"

[1]) Zie Derksen, Inleiding tot de Correlatierekening, § 50, waaraan deze beschouwingen ontleend zijn.

aangeeft. Het is dus de waarschijnlijkheidsverdeling voor de verhouding van de afwijking tussen steekproefgemiddelde en algemeen gemiddelde tot de standaarddeviatie van de steekproef.

Men kan met behulp van deze formule berekenen hoe waarschijnlijk het optreden van een z, groter dan een in een concreet geval gevonden z-waarde, is. Daarbij is een voordeel tegenover de vroeger besproken methoden dat in z de s staat, d. i. de standaarddeviatie van de steekproef en niet die van het (onbekende) universum.

De z-verdeling van Fisher

Een zeer algemene verdeling is de zgn. z-verdeling van Fisher. De grootheid z die daarin een rol speelt, wordt als volgt gedefiniëerd. Uitgegaan wordt van $n_1 + n_2$ onderling onafhankelijke normaal verdeelde veranderlijken

$$x_i \quad (i = 1, 2, \ldots, n_1 + n_2).$$

Deze worden in twee groepen verdeeld, nl.

$$x_1, x_2 \ldots . x_{n_1} \quad \text{en} \quad x_{n_1+1}, \ldots . x_{n_1+n_2}.$$

Daaruit worden berekend:

$$n_1 s_1^2 = x_1^2 + \ldots . + x^2_{n_1} \tag{32.6}$$

$$n_2 s_2^2 = x^2_{n_1+1} + \ldots . + x^2_{n_1+n_2} \tag{32.7}$$

en vervolgens is $z = \log \frac{s_1}{s_2}$ en men vraagt naar de waarschijnlijkheidsverdeling van deze z. Er wordt voor gevonden

$$\frac{2 n_1^{\frac{1}{2} n_1} \, n_2^{\frac{1}{2} n_2} \, \Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right)} \frac{e^{n_1 z} \, dz}{(n_1 e^{2z} + n_2)^{\frac{1}{2}(n_1 + n_2)}} \tag{32.8}$$

Een toepassing van deze z-verdeling werd reeds besproken in § 30.

Zoals reeds werd opgemerkt, kunnen bij nieuwe steekproefproblemen vaak ook nieuwe, tot nog toe niet aangetroffen verdelingen hun intrede doen. Er zijn dus nog ettelijke andere verdelingen bestudeerd. De belangstellende lezer zij opnieuw verwezen naar Derksen, Inleiding tot de Correlatierekening, waar een omvangrijke literatuur wordt genoemd.

§ 33. *Een andere methode om de „significance" van een gevonden samenhang te beoordelen*

In een groot deel van de literatuur is het gebruikelijk om de betrouwbaarheid van een gevonden correlatiecoëfficiënt of regressiecoëfficiënt te meten door de berekening van de standaarddeviaties van deze grootheden. Daarbij gaat men dus uit van de voorstelling dat de waarnemingen waarover men heeft beschikt een steekproef vormen uit een groter universum en dat in dat grotere universum de „ware" correlatie- en de „ware" regressiecoëfficiënt heersen. Het onbevredigende bij deze wijze van voorstellen is wel dat ook deze „ware"

samenhang een stochastische zou zijn. Om deze reden is de andere opzet die door Frisch gekozen is om deze kwesties te belichten [1]) een vooruitgang te achten. Frisch gaat uit van de voorstelling dat in elke reeks een systematisch en een toevallig gedeelte moet worden onderscheiden en dat tussen de systematische gedeelten een totale correlatie moet worden aangenomen, terwijl tussen de toevallige geen correlatie wordt aangenomen. Deze laatste veronderstelling zou men natuurlijk ook anders kunnen nemen: er mag ook wel correlatie bestaan tussen de niet-systematische gedeelten. Nemen we echter de voorstellingswijze van Frisch over en beschouwen we het eenvoudigste geval dat zich hierbij kan voordoen, nl. dat van twee variabelen. Veronderstellen we dus dat deze twee variabelen x_1 en x_2 elk kunnen worden gesplitst:

$$x_1 = x_1' + x_1'' \qquad (33.1)$$

$$x_2 = x_2' + x_2'' \qquad (33.2)$$

en dat daarbij x_1' en x_2' de systematische gedeelten zijn en x_1'' en x_2'' de toevallige. We nemen nu met Frisch (pag. 52 e.v.) aan dat er geen correlatie bestaat tussen:

I. x_1'' en x_2'', dus $\Sigma x_1'' x_2'' = 0$.
II. x_1'' en x_2', „ $\Sigma x_1'' x_2' = 0$.
x_2'' en x_1', „ $\Sigma x_1' x_2'' = 0$.

[1]) R. Frisch, Statistical Confluence Analysis by Means of Complete Regression Systems, Oslo 1934, p. 52 e.v.

III. x_1' en x_1'', dus $\Sigma x_1' x_1'' = 0$.
x_2' en x_2'', „ $\Sigma x_2' x_2'' = 0$.

Daaruit volgt:

$$\Sigma x_1 x_2 = \Sigma x_1' x_2' + \Sigma x_1'' x_2' + \Sigma x_1' x_2'' + \Sigma x_1'' x_2'' = \Sigma x_1' x_2'$$

$$\Sigma x_1^2 = \Sigma x_1'^2 + 2 \Sigma x_1 x_1'' + \Sigma x_1''^2 = \Sigma x_1'^2 + \Sigma x_1''^2$$

$$\Sigma x_2^2 = \Sigma x_2'^2 + \Sigma x_2''^2$$

Als we nu nog invoeren de „storingsintensiteiten":

$$\lambda_1 = \frac{\Sigma x_1''^2}{\Sigma x_1^2} \quad \text{en} \quad \lambda_2 = \frac{\Sigma x_2''^2}{\Sigma x_2^2} \qquad (33.3)$$

dan is dus $(1 - \lambda_1) \Sigma x_1^2 = \Sigma x_1'^2$

$(1 - \lambda_2) \Sigma x_2^2 = \Sigma x_2'^2$

De gevonden correlatiecoëfficiënt tussen x_1 en x_2 zij r_{12}; dan is

$$r_{12} = \frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 \Sigma x_2^2}} \qquad (33.4)$$

De „ware" correlatiecoëfficiënt ϱ_{12} is die tussen x_1' en x_2':

$$\varrho_{12} = \frac{\Sigma x_1' x_2'}{\sqrt{\Sigma x_1'^2 \Sigma x_2'^2}} \qquad (33.5)$$

Daaruit volgt:

$$\varrho^2_{12} = \frac{r_{12}^2}{(1 - \lambda_1)(1 - \lambda_2)} \qquad (33.6)$$

Als we nu aannemen dat $\varrho_{12} = 1$, dan volgt daaruit:

$$(1 - \lambda_1)(1 - \lambda_2) = r^2_{12}$$

d.w.z. λ_1 en λ_2 moeten zodanig zijn, dat het punt (λ_1 λ_2) (bij een rechthoekig assenkruis voor λ_1 en λ_2) ligt op de gelijkzijdige hyperbool die in fig. 14 is aangegeven. We vinden dus zekere grenzen waarbinnen zich de λ_1 en de λ_2 hebben te bewegen. Dit brengt vanzelfsprekend ook

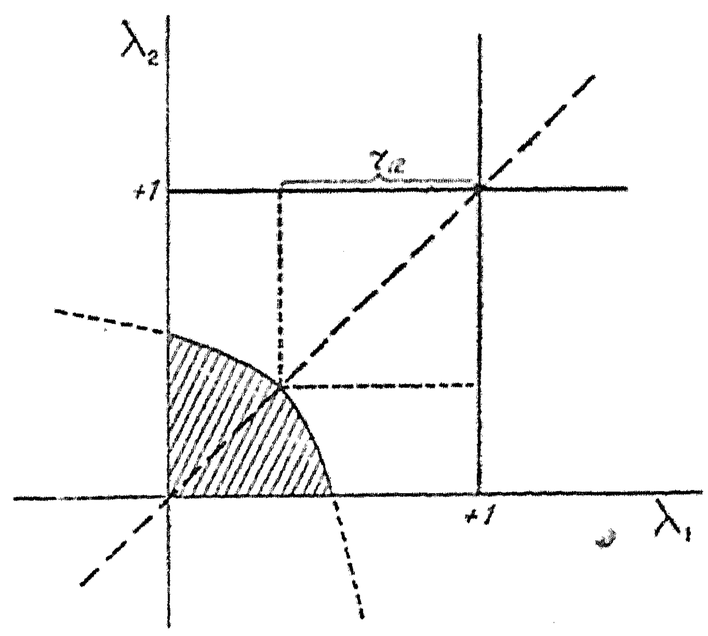

Fig. 14. Waarden van de storingsintensiteiten λ_1 en λ_2 bij onvolkomen correlatie tussen twee gestoorde grootheden, waarvan de systematische gedeelten volkomen correleren. (Ontleend aan Frisch, Confluence Analysis, Oslo 1934).

grenzen mede voor de regressiecoëfficiënt die tussen x_1' en x_2' gelden.

Immers

$$b_{12} = \frac{\Sigma x_1' \; x_2'}{\Sigma x_2'^2} = \frac{\Sigma x_1 \; x_2}{\Sigma x_2^2 \; (1 - \lambda_2)}$$

De minimumwaarde hiervan wordt bereikt als $\lambda_2 = 0$; dan is $b_{12} = \frac{\Sigma x_1 \; x_2}{\Sigma x_2^2}$; de maximum-

waarde als $\lambda_1 = 0$, en dan kunnen we er voor schrijven $\frac{1}{b_{21}} = \frac{\Sigma x_1^2 (1 - \lambda_1)}{\Sigma x_1 x_2} = \frac{\Sigma x_1^2}{\Sigma x_1 x_2}$.

Deze maximum- en minimumwaarde zijn juist de elementaire regressiecoëfficiënten van de oorspronkelijke gegevens: *de „ware" regressiecoëfficiënt ligt daar dus tussen in.* (Frisch, t.a.p., blz. 60).

Men ziet dat de soort van stellingen die men nu bereikt een geheel andere is. In 't geval dat men de steekproeftheorie toepast zijn theoretisch alle waarden van de „ware" regressiecoëfficiënt mogelijk; elke afwijking tussen „ware" en gevonden coëfficiënt heeft een zekere waarschijnlijkheid. Bij de opzet van Frisch zijn slechts de waarden binnen een zeker interval mogelijk en volgt uit elke „ware" waarde de storingsintensiteit van elk der variabelen.

Frisch past deze gedachtengang toe op veel ingewikkelder gevallen, die buiten het kader van dit boekje vallen. Het was de bedoeling van deze paragraaf slechts om de aard van de behandelingswijze duidelijk te maken. Met name combineert Frisch deze werkwijze met de veronderstelling van multicollineariteit. Daartoe worde echter verwezen naar de genoemde publicatie.