EERSTE GEDEELTE

HET METEN

I. De natuur van het te meten object

§ 1. *Rangschikken en meten*

Niet alle verschijnselen waarover in getallen kan worden gesproken zijn toegankelijk voor eigenlijke *meting*. Het is nodig hier onderscheid te maken tussen verschijnselen die alleen rangschikking toelaten en zulke die bovendien werkelijke meting toelaten. Voorbeelden zijn hier wel het meest instructief.

Als men vaststelt dat de ene weg 10 km en de andere 2 km lang is, houdt dat in, dat men de lengte van de laatste weg 5 × kan afpassen op de eerste; dat die eerste 5 × zo lang is als de tweede. Dit is karakteristiek voor het *meten* in engere zin. Wanneer echter op school één leerling een 10 en de andere een 2 voor rekenen heeft, betekent dat niet dat de knappe jongen 5 × zo knap is of 5 × zo vlug een som kan maken als de minder fortuinlijke. Hier is slechts kwestie van een *volgorde*, evenals bij huisnum-

mers, enz. Hier kan niet gemeten worden, doch slechts gerangschikt. Natuurlijk kan men bij meetbare grootheden ook rangschikken: bij te rangschikken grootheden echter niet noodzakelijk altijd meten. Meetbare grootheden worden ook wel *extensiteiten*, de alleen rangschikbare grootheden *intensiteiten* genoemd [1]).

De voorbeelden van intensiteiten zijn niet weinig talrijk en hun belang voor de sociale wetenschappen is niet gering. Enkele voorbeelden mogen hier genoemd worden.

In de economie gebruikt men het begrip *bevredigingsgraad* (ofelimiteit, psychische welstand); de ene goederencombinatie (b.v. (1) een kleine woning en slecht eten) geeft een andere bevredigingsgraad dan andere combinaties (b.v.
(2) een kleine woning en goed eten;
(3) een grote woning en slecht eten;
(4) „ „ „ „ goed „) [2])
Het is niet mogelijk om te zeggen dat combinatie (4) b.v. 3 × zo veel bevrediging geeft als combinatie (1), enz., maar wel laat zich vaststellen, voor elk individu afzonderlijk, welke combinatie de grootste, welke de op een na grootste, welke de op een na kleinste en welke de kleinste be-

[1]) Het woord intensiteit heeft hier dan een andere betekenis dan in het dagelijks leven; er vallen b.v. ook onder verschijnselen die men gewoonlijk qualiteiten zal noemen.

[2]) Hierbij denken we ons de begrippen klein en groot, slecht en goed in dit verband precies omschreven.

vrediging schenkt. De volgorde is vast te stellen. En het is op deze omstandigheid dat de economie haar leerstellingen bouwt.

In de pedagogie maakt men gebruik van *cijfers voor rapporten,* voor allerhande werkstukken, enz. Zoals reeds gereleveerd, geven ook deze alleen (zo goed mogelijk) een volgorde aan voor de kwaliteiten der werkstukken van alle leerlingen in een klasse.

Voorbeelden van meetbare grootheden zijn het talrijkst in de physische wetenschappen, doch ook zeer talrijk in de economie. Goederenhoeveelheden, prijzen en geldsommen (inkomens, kredieten, enz.) zijn voorbeelden.

Een nadere beschouwing leert, dat het niet altijd zeker is of een grootheid „door zijn natuur" in de ene of de andere categorie behoort. Dat wordt b.v. gedemonstreerd door de opvattingen, welke met de relativiteitstheorie hun intrede hebben gedaan. Volgens deze theorie is de grootte van de lengtemaatstaven die wij bij het meten gebruiken niet zo absoluut als vroeger werd gemeend en hangt deze af van de beweging van die maatstaf. Het zou te ver voeren, hierop nader in te gaan. Het feit wordt daarom slechts gereleveerd.

§ 2. *De maatstaf*

Bij het meten van meetbare grootheden doen zich nog enkele technische kwesties voor, die

wegens hun groot belang bij sommige toepassingen hier afzonderlijk worden besproken.

Een eerste vraag die men bij het meten heeft te stellen is: waar leggen wij het beginpunt van onze schaal? Bij het meten van temperaturen b.v. zijn wij in het dagelijks leven gewend dit te leggen bij het vriespunt van water (als we met een Celsius-schaal werken). Alle temperaturen zijn dus gemeten vanaf dit vriespunt als nulpunt, als beginpunt. Men kan echter ook vanaf andere nulpunten meten. Zo meet de Fahrenheitschaal vanaf een lager gelegen temperatuur, zodanig dat het vriespunt bij 32° F. ligt. En de absolute of Kelvin-schaal meet vanaf het zgn. absolute nulpunt, ongeveer 273° C. onder het vriespunt van water. Het is een kwestie van doelmatigheid, waar men het nulpunt legt; slechts moet steeds duidelijk blijken hoe de keuze is gevallen. Hoewel dat niet zo gebruikelijk is, zou men ook de lengte van een mens, in plaats van op de gebruikelijke manier, kunnen meten vanaf een ander nulpunt, b.v. vanaf de hoogte van 1 meter. De hoogte van bomen meet men gewoonlijk vanaf de grond, die van bergen vanaf de zeespiegel. Zou men echter de wortels van de boom meerekenen, of de diepte van de diepste zee, dan zou men een soort „absolute hoogte" van bomen en bergen kunnen invoeren die vanaf een ander, lager gelegen punt wordt gerekend. Het is echter wel nuttig, de term absoluut, die al even werd

gebezigd, hier ter onderscheiding van andere, relatieve maatstaven te gebruiken. Alleen ten aanzien van de absolute lengte van b.v. een mens geldt dat een mens van 2 m. 2 × zo lang is als een van 1 m. Meet men de mensen in centimeters, vanaf een meter hoogte als nulpunt, dan is iemand van 60 cm. niet 6 × zo groot als iemand van 10 cm.; dat geldt alleen voor de absolute lengte. Wel kan men zeggen dat iemand van 60 cm. (in die maatstaf) 6 × zo veel uitsteekt boven een muurtje van 1 meter als iemand van 10 cm. (in die maatstaf). De lezer vergeve deze elementaire uiteenzettingen: ze worden soms, in gecompliceerder gevallen, vergeten.

Veelal neemt men, speciaal bij economisch-statistische onderzoekingen, als nulpunt van de schaal niet het absolute nulpunt (b.v. van prijzen, van productie) maar het een of ander gemiddelde. De getallen, die men dan voor het verloop van een prijs geeft, kan men dan meer precies aanduiden als „afwijkingen van het gemiddelde". Een voorbeeld volgt hieronder.

Prijzen: 12 8 9 11 10 Gemiddelde: 10
Afwijkingen van het gemiddelde resp.:
+2 —2 —1 +1 0

Een tweede kwestie, die bij het meten van belang is, is natuurlijk die van de *eenheden waarin men meet*. Men kan een lengte in meters, in centimeters, in kilometers, enz. uitdrukken. Men

kan een prijs in guldens per kg., of in shillings per quarter, enz. enz. uitdrukken. Heeft men een hele reeks prijsgegevens (b.v. de juistgenoemde), dan zullen de cijfers, wanneer men van de ene eenheid overgaat op de andere, alle in dezelfde verhouding worden gewijzigd. Nog ingewikkelder veranderingen zijn mogelijk, als men het hele systeem van meting verandert. Zo kan men bij overgang van het ene systeem op het andere twee reeksen cijfers krijgen, die niet evenredig zijn, doch waar de ene de logarithme is van de andere. Het is b.v. nuttig om, in tijden van zeer sterke prijswijzigingen niet te rekenen met de absolute waarde van het prijsniveau, doch met de logarithme daarvan. Wanneer het prijsniveau zich wijzigt van 100 op 1000 en van 1000 op 10.000, dan wijzigt de logarithme zich van 2.00 op 3.00 en van 3.00 op 4.00; aangezien in verschillende opzichten een verandering van 100 op 1000 dezelfde betekenis heeft voor het sociale leven als een verandering van 1000 op 10.000, is het gebruik van de logarithmen, waarbij de toeneming in beide gevallen door hetzelfde cijfer van 1.00 wordt weergegeven, voor de problemen waarom het dan gaat, aan te bevelen.

Een goed voorbeeld is ook de wijze waarop Van der Wijk [1]) de inkomens meet. In plaats

[1]) Vgl. De Economist, 80 (1931), blz. 112 e.v., 232 e.v., 305 e.v.

van dat te doen op de gebruikelijke wijze, in guldens per jaar (laten we dit getal i noemen), neemt deze auteur log $(i—i_0)$, waarbij i_0 een grootheid is die voor elk land en elk jaar afzonderlijk moet en kan worden vastgesteld en door hem „bestaansminimum" wordt genoemd.

Het spreekt verder wel van zelf, dat wanneer men met verschillende grootheden naast elkaar werkt, de keuze der eenheden niet voor elk daarvan vrij behoeft te zijn, doch integendeel vaak bepaald is door de keuze bij een der andere grootheden. Daar het van het geheel van de vraagstukken, die men beschouwt, afhangt welke vrijheden er hier al dan niet bestaan, wordt daarop hier nu niet verder ingegaan. Er worde in dit verband echter nog verwezen naar § 12.

§ 3. *Het aantal dimensies (vrijheidsgraden) van het te meten object*

Een zeer belangrijke vraag bij het meten van een verschijnsel is het aantal dimensies dat het heeft, d. w. z. *het aantal getallen dat minimaal nodig is, om het volledig te beschrijven.* Voorbeelden kunnen ook hier het meest verhelderend werken.

Wil men, om met een meetkundig voorbeeld te beginnen, een rechthoekig blok (of officiëel: een rechthoekig parallelopipedum) beschrijven, dan moet men drie getallen noemen: de lengte, de breedte en de hoogte. Laten we deze a, b en c noemen. Wil men daarentegen de afmetingen

van een lijn weergeven, dan is maar één getal nodig: de lengte. Men zou natuurlijk ook bij het blok wel één getal kunnen geven dat een globaal denkbeeld geeft van de grootte er van: b.v. de inhoud. Maar volledig beschrijven doet men het blok daarmee niet. Iemand die het blok moet dragen zal mogelijk reeds genoeg weten: uit de inhoud kan hij het gewicht berekenen (vooropgesteld dat de substantie gegeven is) en dat kan voor zijn doel genoeg zijn. Maar iemand die het blok moet bekleden, is daarmee niet geholpen. Die zou het totale oppervlak moeten kennen. Dat is ook maar één getal, maar op een andere manier verkregen uit getallen a, b en c. Terwijl de inhoud wordt voorgesteld door de uitdrukking abc, is het totale oppervlak 2 (ab + bc + ca). Iemand daarentegen die weten moet of hij het blok kan neerzetten in een bepaalde hoek van zijn kamer, moet alle drie de getallen kennen. Zo ook iemand die het blok zou moeten namaken. Daarbij zijn drie getallen nodig, niet minder, maar ook niet meer. Men zou daarbij nog andere getallen kunnen vermelden, zoals de inhoud, het oppervlak, de ruimtediagonaal b.v., maar dat zou overbodig zijn. Het bedoelde blok heeft in deze zin dus drie dimensies, een lijnsegment één.

Er moge de aandacht op gevestigd worden, dat het aantal dimensies op deze wijze gedefiniëerd, niet altijd overeenkomt met het aantal

dimensies van een lichaam in de zin van de meetkunde. Men zegt in de meetkunde b.v. dat een onregelmatig blok (b.v. een met scheve hoeken) ook slechts drie dimensies heeft. In onze zin heeft zo'n blok een veel groter aantal dimensies: men moet er veel meer getallen van kennen om het te kunnen construeren. In de natuurwetenschappen noemt men het aantal dimensies in onze zin veelal het *aantal vrijheidsgraden.*

Voor den mathematisch geschoolden lezer zij vastgesteld dat het aantal dimensies in onze zin is het aantal dimensies van de ruimte waarin men elk individueel blok door één punt zou kunnen karakteriseren; het aantal dimensies in de zin van de meetkunde is het aantal dimensies van de minstgedimensioneerde ruimte waarin het blok bestaanbaar is.

Het punt in de eerstgenoemde (hypothetische) ruimte bepaalt een voerstraal, lopende van de oorsprong van die ruimte naar het punt in kwestie. Zo'n voerstraal noemt men ook wel een *vector*. Deze is dus eveneens bepaald door meer dan één getal, nl. door zoveel getallen als de ruimte in kwestie dimensies heeft (in de gewone meetkundige zin). Een grootheid die men door slechts één getal kan vastleggen noemt men dan wel een *skalar* of een *skalaire grootheid*.

In verschillende der wetenschappen waarin

met veel statistische waarnemingen gewerkt wordt is voorts het woord „*complex*" voor een grootheid die door meerdere getallen wordt bepaald in gebruik.

Om tot ons blok terug te keren: essentiëel voor ons is thans dat één getal niet in staat is om het te beschrijven; men kan dus ook niet alle bestaanbare blokken rangschikken in één rij. Zo is het ook met veel andere begrippen. Een voorbeeld van geheel andere aard is b.v. iemands bekwaamheid in het schaakspel. Ook dat is geen grootheid die door één getal is weer te geven; bedoelde bekwaamheid heeft vele zijden. En zo kan het geschieden dat A beter schaakt dan B, en B beter dan C, terwijl toch A niet beter dan C schaakt. Hierdoor is het onmogelijk om naar deze bekwaamheid een rangschikking naar één reeks door te voeren.

In het bovenstaande zijn twee feiten van grote betekenis voor de meting van sociale en anthropologische verschijnselen naar voren gekomen; nl. ten eerste dat er *voor bepaalde verschijnselen méér dan een getal nodig is ter volledige beschrijving er van* en ten tweede, dat men *voor sommige doeleinden wel met één samenvattend getal kan volstaan,* doch dat *dit getal nooit alle eigenschappen van het object tegelijkertijd tot uitdrukking kan brengen* en dat *de wijze waarop*

dat samenvattend getal uit de volledige reeks cijfers moet worden berekend wordt bepaald door het doel dat men nastreeft.

Deze mogen thans nog eens uitvoerig worden toegelicht aan enkele voorbeelden uit de sociale en anthropologische wetenschappen.

§ 4. *Indexcijfers*

Een eerste zeer bekend voorbeeld is dat van de *indexcijfers* in de economische statistiek (overigens mutatis mutandis evenzeer bruikbaar voor andere objecten [1]).

Een indexcijfer is een samenvattend cijfer voor een meerdimensionaal (complex) verschijnsel, uitgedrukt in procenten van de waarde die dat samenvattend cijfer in een basisperiode of voor een ander vergelijkingsobject heeft. Als men b.v. vaststelt dat het indexcijfer van de groothandelsprijzen van afgewerkte producten in de Verenigde Staten van Noord-Amerika in het jaar 1934 bedroeg 70.5, dan wil dat zeggen dat „het prijsniveau" van bedoelde goederen in bedoeld land in het jaar 1934 70.5 % bedroeg van het overeenkomstige niveau in het basistijdperk, dat hier het jaar 1926 is. Met „het prijsniveau" is dan zo'n samenvattend begrip bedoeld als in ons meetkundige voorbeeld „de

[1]) Zo worden ook geregeld indexcijfers van de criminaliteit berekend.

inhoud" of „de oppervlakte", doch dan een samenvatting voor het stelsel der prijzen van allerlei afgewerkte producten. Zou men dat begrip prijsniveau volledig willen beschrijven, dan zou men de prijzen van alle afgewerkte producten afzonderlijk moeten kennen. Ook hier hebben we dus te doen met een meerdimensionaal verschijnsel. De vraag is nu: op welke wijze moet een samenvattend cijfer van al deze prijzen worden berekend? Nauw daarmee verwant is dan ook de vraag hoe een *indexcijfer* moet worden berekend (in welk cijfer bovendien nog het *relatieve* prijsniveau t. o. v. 1926 tot uitdrukking moet komen). Zoals reeds gezegd, is hier het doel waarvoor men het prijsniveau (resp. het indexcijfer) wenst te gebruiken beslissend.

Is dat doel b.v. om te weten hoeveel meer de gezamenlijke kopers van de afgewerkte producten moeten betalen, dan zal men rekening moeten houden met de goederen en de daarvan omgezette hoeveelheden, die Amerikaanse kopers hebben gekocht. Zou men daartegenover wensen te weten, hoeveel de verkopers van afgewerkte producten hebben ontvangen, dan moet men rekening houden met de goederen en de omgezette hoeveelheden die Amerikaanse verkopers hebben omgezet. In het laatste geval zal men dus b.v. de uitvoer meetellen, in het eerste geval de invoer. De doeleinden kunnen nog andere zijn. Men kan zich de vraag stellen, wat *bepaalde*

Amerikaanse kopers, b.v. alleen de consumenten als kopers hebben te betalen. Dan zal men van de afgewerkte producten alleen de consumptiegoederen nemen en daarvan de prijzen in de kleinhandel. Enzovoort.

Met het bovenstaande is intussen nog geen voldoend nauwkeurige omschrijving van de berekeningswijze gegeven: slechts de richtlijnen zijn vastgesteld. Volledigheidshalve zij de berekeningswijze nog medegedeeld voor het eerste der bovenomschreven mogelijke doeleinden.

Laten we de in het basisjaar (0) bestaande prijzen der goederen 1, 2, 3 . . . n aangeven met

$$p_{01} \quad p_{02} \quad p_{03} \cdots p_{0n}$$

en de hoeveelheden die in dat jaar zijn omgezet met

$$q_{01} \quad q_{02} \quad q_{03} \cdots q_{0n};$$

voor het jaar waarvoor we ons interesseren (1) resp.

$$p_{11} \quad p_{12} \quad p_{13} \cdots p_{1n} \text{ en}$$
$$q_{11} \quad q_{12} \quad q_{13} \cdots q_{1n}.$$

Het bedrag dat in het basisjaar is uitgegeven voor deze goederen is dus $p_{01}q_{01} + p_{02}q_{02} + \cdots p_{0n}q_{0n} = \Sigma p_0 q_0$

Het bedrag dat in het jaar 1 is uitgegeven is

$$p_{11}q_{11} + p_{12}q_{12} + \cdots p_{1n}q_{1n} = \Sigma p_1 q_1$$

Willen we de prijsniveau's vergelijken in deze twee jaren, dan kunnen we echter niet deze twee grootheden naast elkaar zetten; want het verschil $\Sigma p_0 q_0$ en $\Sigma p_1 q_1$ is niet alleen ver-

oorzaakt doordat er in jaar 1 andere prijzen heersten, doch evenzeer doordat er toen andere hoeveelheden zijn omgezet. We moeten de hoeveelheden daarom onveranderd denken. Er blijven verschillende mogelijkheden bestaan. Gebruiken we de hoeveelheden die in het basisjaar werden omgezet, dan is een goede maatstaf voor de verhouding der prijsniveau's in de jaren 1 en 0:

$$\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \tag{4.1}$$

gaan we daarentegen uit van de hoeveelheden die in het jaar 1 zijn omgezet, dan zouden we als maatstaf verkrijgen

$$\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \tag{4.2}$$

De uitkomsten (1) en (2) zullen niet verschillen wanneer $q_{01} : q_{11} = q_{02} : q_{12} = q_{03} : q_{13} = \ldots . q_{0n} : q_{1n}$; in 't algemeen zullen, wanneer de jaren 0 en 1 niet ver uiteenliggen de uitkomsten (1) en (2) ook niet veel verschillen. Het is ons niet mogelijk, op deze materie hier dieper in te gaan. Daartoe zij verwezen naar G. Haberler, Der Sinn der Indexzahlen, Tübingen 1927, en H. Staehle, A Development of the Economic Theory of Price Index Numbers, The Review of Economic Studies, II (1935), p. 163.

Vermeld worde nog, dat men formule (4.1) noemt de formule van *Laspeyre* en formule (4.2) die van *Paasche*. De getallen q vervullen de rol van *gewichten* of *wegingscoëfficiënten*.

Wanneer men niet uitgaat, zoals hierboven is geschied, van de absolute prijscijfers, doch, zoals ook vaak wordt gedaan (b.v. bij het indexcijfer van de groothandelsprijzen in Nederland, men vgl. Maandschrift van het Centraal Bureau voor de Statistiek, Augustus 1935, blz. 1217) van de relatieve cijfers t.o.v. een basisperiode, dan moet men als wegingscoëfficiënten ook andere getallen gebruiken, en wel de waarde van de omzetten in plaats van de hoeveelheden. Men ziet gemakkelijk in dat men zodoende dezelfde uitkomsten moet verkrijgen.

In de literatuur over dit onderwerp komt men nog velerlei andere formules tegen. Met name wordt vaak gebruik gemaakt van indexcijfers die niet, zoals de bovenstaande, een quotiënt zijn van sommen, doch een quotiënt van producten dus b.v.

$$I = \frac{p_{11}^{a_{01}} \quad p_{12}^{a_{02}} \quad p_{13}^{a_{03}} \cdots\cdots p_{1n}^{a_{0n}}}{p_{01}^{a_{01}} \quad p_{02}^{a_{02}} \quad p_{03}^{a_{03}} \cdots\cdots p_{0n}^{a_{0n}}}$$

en dergelijke. Bij de bespreking van verschillende soorten van gemiddelden zal daarvan nog nader sprake zijn. Thans zij er echter reeds op gewezen, dat deze formules zeer bezwaarlijk gefundeerd kunnen worden als uitvloeisel van een be-

paald met deze indexcijfers na te streven doel.

Voor buitengewoon uitvoerige beschouwingen over de verschillende mathematische mogelijkheden tot het berekenen van indexcijfers zij verwezen naar I. Fisher, The Making of Index Numbers, Boston & New York, 1922.

Hetgeen hier over prijsindexcijfers gezegd is, kan grotendeels ook op indexcijfers voor andere grootheden worden toegepast. Als voorbeeld moge nog gewezen worden op de in de laatste tijd nogal vaak gebezigde indexcijfers van de invoer van grondstoffen. Ook daar hangt het weer van het doel af welke *weging* men bezigt. Enkele denkbare doeleinden zijn:

1°. na te gaan wat de vervoersondernemingen van een verandering in de invoer van grondstoffen voor gevolgen ondervinden. In dat geval zal het er om gaan, hoe het totale gewicht (c.q. ook het volume) van de invoer van grondstoffen te beschouwen. Heeft men de invoer gegeven in kilogrammen, dan moet dus geen weging worden toegepast (c.q. een weging gelijk aan het „soortelijk volume", d. w. z. het omgekeerde van het soortelijk gewicht als weging worden gebruikt).

2°. na te gaan, welke hoeveelheid eindproducten daaruit zal worden geproduceerd, waarbij het weer er om gaat, te weten welke invloed de vervoerondernemingen voor het vervoer der

eindproducten zullen ondergaan van een verandering in de invoer van grondstoffen. Er zou dan een weging moeten worden toegepast die evenredig is met bepaalde „fabricatiecoëfficiënten" en wel de coëfficiënten die aangeven hoeveel eindproduct uit één kg. grondstof wordt vervaardigd.

3°. na te gaan, hoe de hoeveelheid werkgelegenheid zich wijzigt als gevolg van de wijzigingen in de hoeveelheden ingevoerde grondstoffen. Dan zal men als weging moeten bezigen de hoeveelheid arbeid die wordt toegepast bij het productieproces dat binnen het land in kwestie met behulp van de bedoelde grondstoffen plaats vindt.

Deze lijst zou te verlengen zijn. De bedoeling zal echter wel duidelijk zijn.

§ 5. *Andere voorbeelden*

Voorbeelden van samenvattende getallen, die gebezigd worden voor het karakteriseren van een in wezen meerdimensionale grootheid zijn ook de rapportcijfers die aan leerlingen worden toegekend en de rangcijfers die daaruit weer voor één leerling worden berekend, door b.v. van de rapportcijfers de som te nemen en aan de hoogste som het rangnummer één, aan de op een na hoogste het rangnummer 2, enz. toe te kennen. Inderdaad is één cijfer voor een enkel vak ook reeds te beschouwen als een samenvatting van

zeer verschillende kwaliteiten van een leerling. In dat cijfer toch zitten besloten de bekwaamheden in verschillende richting die voor de beoefening van een bepaald vak nodig zijn, b.v. bij talenstudie: het vertalen in en vanuit een andere taal, het stellen, de kennis van de grammatica, enz. Dat hetzelfde geldt voor de rangcijfers, is duidelijk. Dat men ook hier voor verschillende doeleinden op verschillende wijzen de samenvattende cijfers moet berekenen, is bekend genoeg. Ten aanzien van beoordelingen bij examens blijkt dat duidelijk hieruit, dat de zwaarte waarmee men de verschillende vakken laat meetellen voor verschillende inrichtingen van onderwijs uiteenloopt, evenals voor verschillende afdelingen van eenzelfde inrichting (gymnasium A en B b.v.).

Een ander voorbeeld uit de pedagogie is het begrip *intelligentieleeftijd*. Men geeft door een leeftijd aan het peil waarop zich een bepaalde leerling bevindt. Uiteraard is echter ook deze intelligentieleeftijd slechts dan een aanduiding voor het peil in alle opzichten, wanneer bedoelde leerling zich harmonisch in alle richtingen heeft ontwikkeld. Als hij echter in de ene richting merkbaar verder ontwikkeld is dan in de andere, dan kan de intelligentieleeftijd niet anders dan een gemiddelde zijn met alle nadelen hierboven bij de behandeling der indexcijfers genoemd. Om dan exacter geïnformeerd te zijn, moet men

de structuur van de intelligentie kennen (vgl. ook § 10 en fig. 2; eveneens § 13).

Het weergeven van een meerdimensionaal (complex) verschijnsel door een enkel cijfer is reeds lang in gebruik en talrijk zijn de ervaringen dan ook over de moeilijkheden die zulks medebrengt. Veel moeilijkheden spruiten voort uit de omstandigheid, dat men een voor een doel I geconstrueerd samenvattend cijfer ook wil gebruiken voor een ander doel II, waarvoor het eigenlijk niet kan worden gebruikt. Zo wordt vaak van een algemeen indexcijfer van groothandelsprijzen gebruik gemaakt om de invloed van prijsveranderingen uit de veranderingen in de in- en uitvoer te elimineren. Dat geschiedt dan in dier voege, dat men verhoudingscijfers vervaardigt voor de loop van de waarde van de uitvoer over een aantal jaren, b.v.

1926/'30	1929	1932	1934
100	106	45	38

en deze deelt door de overeenkomstige indexcijfers van de groothandelsprijzen, bv.

100	99.7	64.6	63.0	(P)

met als resultaat:

100	106	69	60

Wanneer nu echter het algemene prijsverloop, zoals dat in reeks (P) tot uitdrukking komt, sterk verschilt van het prijsverloop voor exportartikelen, dan is het resultaat kennelijk onbetrouwbaar.

Een zeer veel voorkomend geval van soortgelijke moeilijkheden is de meting van de productiefactor kapitaal. Deze productiefactor kan — nog geheel afgezien van de theoretisch-economische moeilijkheden die zich daarbij voordoen— in fysieke zin en in financiële zin worden gedefiniëerd. Wanneer we ons beperken tot de fysieke zin, dan nog hebben we te doen met een meerdimensionale grootheid. Eén aspect van de hoeveelheid kapitaal wordt gegeven door het vaak beschikbare getal van het vermogen, in pk, van de primaire krachtwerktuigen. Voor sommige doeleinden is dit ene getal inderdaad het juiste, zo b.v. voor de bestudering van de grenzen van het energieverbruik.

Voor andere doeleinden is dit getal echter ten enen male onvoldoende, b.v. voor het beoordelen van de productieve prestaties. Want daarbij kunnen ook zeer belangrijke diensten bewezen zijn door weinig „krachtige" maar zeer ingenieuze installaties. Zodoende probeert men de cijfers betr. de productiefactor kapitaal tegenwoordig aan te vullen met gegevens over het aantal spinspillen, het aantal papiermachines, het aantal schoenmachines, e.d. [1]).

Weer een belangrijk voorbeeld vormt het begrip *volkswelvaart*. Vaak probeert men de

---

[1]) Men vgl. de statistiek van voortbrenging en verbruik van het Centraal Bureau voor de Statistiek.

bewegingen in de volkswelvaart weer te geven door de beweging van het volksinkomen, desnoods gecorrigeerd voor prijsbewegingen en bewegingen in de grootte der bevolking. Ook dit cijfer is in velerlei opzichten onvoldoende omdat de volkswelvaart een grootheid is met een groter aantal dimensies. Bekend genoeg is de opmerking dat behalve de grootte van het volksinkomen ook de verdeling daarvan zeer belangrijk is. Daarnaast kan naast het geldinkomen een inkomen in natura aanwezig zijn (in de vorm van een gunstig klimaat, gunstige sociale zorg, e.d.) dat voor de welvaart van zeer grote betekenis is [1]).

II. Frequentieverdelingen als hulpmiddel bij de meting van meerdimensionale grootheden

§ 6. *Verdeling, gemiddelden en momenten*

In 't algemeen kan men zeggen dat de wijze, waarop meerdimensionale grootheden volledig kunnen worden beschreven bestaat in het vermelden van meer dan een cijfer.

Soms is het mogelijk deze cijfers op een systematische wijze te groeperen. Speciaal wanneer het aantal dimensies zeer hoog is, is dat van

---

[1]) O. Neurath, Was bedeutet rationale Wirtschaftsbetrachtung? Einheitswissenschaft, Schriften herausgegeben von Otto Neurath, in Verbindung mit Rudolf Carnap und Jørgen Jørgensen, No. 4, Wien 1935.

grote waarde. Zo b.v. bij een prijsniveau of een prijsindexcijfer waarin een groot aantal prijzen zijn gebezigd, of bij een volkswelvaartcijfer, waarbij men de gehele inkomensstructuur wenst te vermelden. Bovendien zal blijken dat deze rangschikking vaak de weg wijst om te komen tot betere samenvattende cijfers dan voorheen bekend waren.

Een zeer bekende en belangrijke wijze ter systematisering van grote hoeveelheden min of meer gelijksoortige cijfers is het gebruik van een *frequentieverdeling*. Dit begrip moge worden verduidelijkt met een aan de inkomensstatistiek ontleend voorbeeld. Ter karakterisering van de totaliteit der inkomens welke door de in de inkomstenbelasting aangeslagenen zijn opgegeven voor een bepaald jaar kan men natuurlijk alle inkomens der afzonderlijke aangeslagenen mededelen. Bij een aantal aangeslagenen van, voor ons land bv., rond 1.6 millioen zou dit echter een omvangrijk werk zijn. Eenvoudiger is, bepaalde inkomensklassen te onderscheiden, b.v.

klasse 1: inkomens *f* 801—1000
„ 2: „ *f* 1001—1200 enz.

en bij elke klasse het aantal aangeslagenen te vermelden dat daarin valt. Dit aantal noemt men de in elke klasse voorkomende frequentie en de aldus ontstaande tabel de *frequentietabel* of *frequentieverdeling*. Ze kan natuurlijk grafisch

worden voorgesteld en men spreekt dan van een *frequentiediagram.* Voor verschillende doeleinden is het nuttig om de grenzen van elke klasse evenver uiteen te nemen (een even grote klassebreedte of een even groot klasse-interval te nemen). Stelt men in het diagram de frequenties voor door de oppervlakte van de op het klasse-interval opgerichte rechthoekjes, dan stelt de hoogte dier rechthoeken voor de frequentie per eenheidsinterval. Door de klassen al kleiner en kleiner te nemen, nadert het frequentiediagram meer en meer tot een vloeiende lijn, de *frequentiecurve,* waarvan dus ook de hoogte in elk willekeurig punt voorstelt de *frequentie per eenheidsinterval* (en niet de frequentie zonder nadere aanduiding).

Soms laat zich de frequentiecurve met grote benadering door een mathematische formule weergeven.

Op het gebied van de inkomensverdeling is dat het geval met de bekende curve van Pareto

$$y = A\cancel{e^{-x^{\alpha}}} \quad (\alpha \sim 1.5) \quad x^{-\alpha}$$

waarbij intussen x voorstelt een bepaald inkomen en y, in afwijking van de boven gekozen definitie, het aantal personen, dat een inkomen *boven* x heeft.

Een andere, in vele gevallen betere, benadering is de curve van Van der Wijk, behandeld in De Economist 80 (1931), p. 335.

Een frequentieverdeling waarvan nog vaak sprake zal zijn, is die van Gauss (zie § 17).

Ter volledige karakterisering van een frequentieverdeling zijn, vooral wanneer een fijne klasse-indeling wordt gebezigd, dus ook nog een betrekkelijk groot aantal getallen nodig. Wil men ook hierop „bezuinigen", dan kan men enkele der belangrijkste eigenschappen van de volledige omschrijving bezigen. De meest bekende dezer samenvattende getallen zullen hier besproken worden.

Rekenkundig gemiddelde

Beschikt men over het oorspronkelijke materiaal (in het geval van de inkomens alle afzonderlijke inkomens) dan levert het berekenen van het rekenkundig gemiddelde geen probleem op: het is eenvoudig de som der inkomens gedeeld door het aantal.

Beschikt men alleen over een frequentieverdeling met klassengrenzen $x_0, x_1, x_2 \ldots x_n$ (dus n klassen) en frequenties $y_1, y_2 \ldots y_n$, dan pleegt men het rekenkundig gemiddelde te berekenen als

$$m = \frac{y_1 z_1 + y_2 z_2 + \ldots y_n z_n}{y_1 + y_2 + \ldots + y_n}$$

waarin $z_1 \ldots z_n$ de middens der klasse-intervallen zijn, dus $2z_1 = x_0 + x_1$, $2z_2 = x_1 + x_2$, $2z_3 = x_2 + x_3$ enz.

Deze uitkomst zal gewoonlijk van het ware rekenkundig gemiddelde enigszins afwijken; er bestaan methoden om een correctie te berekenen. Daarop zal hier echter niet worden ingegaan. Men zie de bekende leerboeken (b.v. Yule, An Introduction to the Theory of Statistics, Londen 1929.)

Meetkundig gemiddelde

Onder het meetkundig gemiddelde verstaat men, zoals bekend is, de n-de machtswortel uit het gedurig product der waarnemingen, waarbij n het aantal waarnemingen is. Ook hier kan men, wanneer een zeer groot aantal waarnemingen aanwezig is, een benaderingsformule gebruiken door van een klasse-indeling uit te gaan. Als de middens der klassen de waarden $z_1 \ldots . z_n$ hebben en de bezetting weer $y_1 \ldots y_n$ is, dan is het meetkundig gemiddelde te berekenen als

$$w = \sqrt[\Sigma y]{z_1^{y_1} z_2^{y_2} \ldots . z_n^{y_n}}$$

Men ziet gemakkelijk in, dat voor de logarithmen geldt:

$$\log w = \frac{y_1 \log z_1 + y_2 \log z_2 + \ldots . y_n \log z_n}{y_1 + y_2 + \ldots . y_n}$$

d.w.z. de logarithme van het meetkundig gemiddelde is het rekenkundig gemiddelde van de logarithmen der waarnemingen.

Modus

Onder de modus verstaat men het midden van het *dichtst bezette interval* (het interval met de grootste frequentie), vooropgesteld, dat alle intervallen even groot zijn genomen. In tegenstelling met de berekening van het rekenkundig gemiddelde en de nog te noemen mediaan is voor de berekening van de modus de verdeling in klassen noodzakelijk voor de uitvoering er van. Bij de hierboven gegeven definitie is de uitkomst ook nog afhankelijk van de wijze van indeling; doch gewoonlijk zal men kunnen zeggen, dat de uitkomst maar weinig meer verandert wanneer het materiaal omvangrijk is en men de indeling binnen zekere grenzen verder verfijnt. Een te grove klasse-indeling is daarom niet bruikbaar, evenmin als een al te fijne. De mathematisch geschoolde lezer, die zich over deze werkwijze verbaast, bedenke, dat het hier steeds gaat om het werken met cijfers waarin reeds onnauwkeurigheden en toevallige afwijkingen voorkomen.

Wanneer meer dan één klasse een maximale frequentie vertoont of zelfs ook maar een relatief maximale — d. w. z. een maximale in vergelijking tot de naaste omgeving naar beide zijden — dan is er dus niet één modus, doch kunnen er meer zijn. Wij zullen nog zien, hoe het optreden van twee- en meertoppige curven in de praktijk belangrijke aanwijzingen kan geven.

Mediaan

Onder de mediaan van een groep getallen verstaat men het middelste getal, na rangschikking dier getallen volgens grootte. Strikt genomen is er dus geen mediaan, wanneer er een even aantal getallen is. Men neemt dan een zgn. *gegeneraliseerde mediaan,* d. w. z. het rekenkundig gemiddelde van de middelste 2 (of 4, enz.) getallen. Ook in het geval van een oneven aantal getallen kan men een gegeneraliseerde mediaan nemen, doch dan b.v. van 3,5, enz. getallen. De berekening geschiedt het nauwkeurigste, wanneer men over het oorspronkelijke materiaal beschikt. Beschikt men alleen over de frequentieverdeling, dan kan men om te beginnen het klasse-interval aangeven, waarin de mediaan ligt. Van dit klasse-interval kan men dan eenvoudig het midden nemen; wil men nauwkeurig te werk gaan, dan kan men dat doen door aan te nemen dat alle oorspronkelijke getallen in dat klasse-interval gelijkelijk over dat interval verdeeld waren.

Het begrip mediaan heeft een zeer nuttige uitbreiding ondergaan doordat men zich de vraag heeft gesteld de getallen aan te geven die het materiaal niet in twee gelijke helften, doch in vier, tien of honderd gelijke delen verdelen. Die getallen noemt men resp. de (3) *kwartielen,* de (9) *decielen* en de (99) *percentielen.* Het zal

duidelijk zijn dat de middelste dezer getallen, dus de 2e kwartiel, de 5e deciel en de 50e percentiel aan de mediaan identiek zijn. Door het aangeven van b.v. de decielen kan men reeds een groot aantal bijzonderheden van een frequentieverdeling weergeven. Door het aangeven van de kwartielen kan men zich over de breedte zowel als over de graad van symmetrie reeds een zekere voorstelling maken: liggen de uiterste kwartielen ver uiteen, dan is de verdeling breed; ligt de linker kwartiel verder van de mediaan dan de rechter, dan is de verdeling asymmetrisch en wel breder aan de linkerzijde, enz.

Een duidelijk voorbeeld zijn de door Jhr. Ir. M. J. de Bosch Kemper berekende inkomensindices [1]). Deze zijn de percentielen van de inkomensverdeling voor Nederland in achtereenvolgende jaren, gebaseerd op de inkomstenbelasting, doch onder de veronderstelling dat de totaliteit der aangeslagenen wordt uitgebreid met de totaliteit der niet aangeslagenen, waarvan wordt aangenomen dat ze een inkomen onder de belastinggrens trekken. Niet alle percentielen zijn genomen, doch die bij 20.2, 13.01, 7.52, 3.39, 1.34, 0.668, 0.418, 0.146, 0,078, 0.036 en 0.019% [2]). In andere woorden geeft de

---

[1]) Zie b.v. De Ned. Conjunctuur, Mei 1935, blz. 11.

[2]) Zoals men ziet, is het begrip percentiel hier door ons nog wat wijder genomen dan volgens de hierboven aan-

laatste index steeds het inkomen van „de op 0.019% na rijkste Nederlander", waarbij het percentage is betrokken op de gehele Nederlandse bevolking.

Momenten

Men onderscheidt voorts aan elke frequentieverdeling de zgn. momenten: het 1e, 2e, 3e .... M-de moment. Daarbij kan men nog hebben „momenten t.o.v. het gemiddelde" en „momenten t.o.v. een willekeurig getal a". Zijn de waarden $x_1 \ldots x_n$, dan is het M-de moment t.o.v. de waarde a:

$$m'_M = \sum_1^n{}^i (x_i - a)^M$$

Wanneer voor a het gemiddelde wordt genomen, dan hebben we het M-de moment t.o.v. het gemiddelde of het M-de moment zonder nadere aanduiding:

$$m_M = \sum_1^n{}^i (x_i - \bar{x})^M$$

Men ziet daaruit dat het eerste moment eenvoudig de waarde nul heeft:

$$m_1 = \sum_1^n{}^i (x_i - \bar{x}) = 0$$

---

gegeven tekst precies het geval zou zijn, doordat nu ook gebroken rangnummers zijn aanvaard.

Spreiding (*strooiing, variance*); *standaarddeviatie* (*middelbare afwijking*). Onder de *spreiding* of *strooiing* (Eng. variance) van een serie waarnemingen verstaat men het gemiddelde kwadraat van de afwijkingen tussen de individuele waarden en het gemiddelde:

$$\frac{1}{n} \Sigma (x_i - x)^2 = \frac{m_2}{n}$$

waarin m_2 dus het tweede moment om het gemiddelde voorstelt.

De wortel uit deze uitdrukking wordt de standaarddeviatie of middelbare afwijking genoemd en veelal met σ aangegeven:

$$\sigma = \sqrt{\frac{\Sigma (x_i - x)^2}{n}} \quad \text{(vgl. ook § 12 en § 27)}$$

Scheefheid. Voor dit begrip zijn verscheidene definities in omloop. We noemen die van Pearson:

$$S = \frac{\bar{x} - \tilde{x}}{\sigma}, \text{ waarbij } \tilde{x} \text{ de modus is.}$$

Daarnaast wordt ook gebruikt (vgl. b.v. Davis & Nelson, Elements of Statistics, p. 89):

$$S'' = \frac{m_3}{2n\sigma^3}$$

Kurtosis. Een verdere eigenschap van fre-

quentieverdelingen die nogal eens ter sprake komt is de zgn. kurtosis (letterlijk: welving).

Een maatstaf hiervoor is (vgl. Davis & Nelson, t.a.p. blz. 313)

$$\frac{m_4}{n\sigma^4} - 3$$

Voor de normale curve is deze uitdrukking juist = 0; is ze positief, dan spreekt men wel van subnormale of Poisson'se verdelingen; is ze negatief, dan van hypernormale of Lexis-verdelingen.

TWEEDE GEDEELTE

SYSTEMATISCHE SAMENHANG TUSSEN TWEE EN MEER VERSCHIJNSELEN

I. QUALITATIEVE BESCHRIJVING VAN DE SAMENHANG

§ 7. *Doel van het meten*

Het meten van de omvang van bepaalde verschijnselen is in veel gevallen doel op zichzelf. Men houdt statistieken bij van prijzen, van lonen, van verkochte hoeveelheden, van het volksinkomen, van de betalingsbalans, omdat men zich voor de getallen betreffende deze verschijnselen op zich zelf interesseert.

In een toenemend aantal gevallen is echter het doel van de meting een verder liggend: men wenst de samenhang tussen meer dan een verschijnsel statistisch te onderzoeken. Dat wil zeggen: *men wenst statistisch te controleren of een op andere wijze aannemelijk gemaakte samenhang inderdaad aanwezig is.* Deze andere wijze van aannemelijk maken is gewoonlijk ontleend aan een der wetenschappen die zich met de bedoelde verschijnselen inlaat. Men zal b.v. op grond van de wetenschap der economie een

verband vermoeden tussen de prijs van een artikel en de productie daarvan. Of men zal op grond van de biologie een verband aanwezig achten tussen de temperatuur van het Zuiderzeewater en de ansjovisvangst.

Uitdrukkelijk wordt hier vastgesteld, dat de rol van de statistiek in deze slechts een controlerende kan zijn en het statistisch onderzoek moet worden voorafgegaan door het onderzoek binnen de betreffende tak van wetenschap. Men heeft wel gemeend, dat men langs statistische weg een verband kan bewijzen. Wat men kan bewijzen is echter nooit meer dan een overeenstemming in de bewegingen van twee of meer verschijnselen. In welke richting het causaal verband echter in zo'n geval loopt is statistisch bijna nooit met zekerheid uit te maken. Men kan alleen constateren of een bepaalde richting al dan niet met de feiten in overeenstemming is. Zo kan natuurlijk niet de oorzaak op een later tijdstip plaats hebben dan het gevolg, uit welke omstandigheid soms conclusies van betekenis zijn te trekken.

Een en ander zal nog duidelijker worden door een nadere analyse van de *logische structuurvormen* waarop men zoal kan stuiten. We zouden de hier volgende ontleding ook kunnen noemen de qualitatieve beschrijving van de samenhang. Deze moet, zoals reeds gezegd, steeds gebaseerd

worden op de wetenschap welke zich met de verschijnselen in kwestie bezighoudt.

§ 8. *Richting van het causale verband*

We hebben reeds even de kwestie aangeroerd, *in welke richting* het causale verband gedacht moet worden. Als twee grootheden samenhangen, kan het zijn dat A oorzaak en B gevolg is (symbolisch aan te geven door $A \rightarrow B$); het kan ook dat B oorzaak en A gevolg is ($B \rightarrow A$). Veelal liggen de zaken echter niet zo eenvoudig. Men pleegt dan nog al eens het woord wisselwerking te gebruiken. Daarmee wordt echter vaak het vraagstuk weggepraat. In vele gevallen kan men daar tot een scherper stellen van de logische structuur komen door een nadere onderscheiding in te voeren, speciaal door onderscheid te maken naar het tijdstip waarop men elk der verschijnselen A en B meet. We zullen dan met A_1 aangeven de omvang van A op tijdstip 1, met A_2 op tijdstip 2, en evenzo de symbolen B_1, B_2, enz. gebruiken. Als men met de symbolen A, B, enz. verschijnselen aanduidt, geven A_1, A_2, B_1, B_2, enz. *gebeurtenissen* aan (in de tijd gelocaliseerde verschijnselen). Het kan nu b.v. zo staan met de samenhang tussen A en B, dat A_1 bepalend is voor B_1 (A_1 oorzaak en B_1 gevolg), doch B_1 bepalend voor A_2 (B_1 oorzaak en A_2 gevolg), A_2 voor B_2, B_2 voor A_3 enz. Een en ander

is te verenigen in het volgende schema:

$$\begin{array}{ccccccccc} A_1 & & A_2 & & A_3 & & A_4 & & \\ \downarrow & \nearrow & \downarrow & \nearrow & \downarrow & \nearrow & \downarrow & \nearrow & \text{enz.} \\ B_1 & & B_2 & & B_3 & & B_4 & & \end{array}$$

Iets dergelijks doet zich voor bij de productie (A) en de prijzen (B) van sommige landbouw- en veeteeltproducten. De productie in het eerste jaar bepaalt door het „vraagverband" de prijshoogte in dat jaar (als er veel geproduceerd is, kan dit door de consumenten alleen opgenomen worden bij een lage prijs); de prijshoogte in dat jaar bepaalt (soms naast andere oorzaken) de omvang van de productie in het volgend jaar; enz. Noch prijs, noch productie is oorzaak zonder nadere aanduiding.

Een ander voorbeeld kan ontleend worden aan de psychologie of pedagogie als men de samenhang bespreekt tussen intelligentie en genoten onderwijs. De intelligentie van een bepaalde persoon kan zowel gevolg zijn van genoten onderwijs als ook oorzaak daarvan, zo wordt betoogd. Ook hier ligt de zaak dan ongeveer zo dat de intelligentie *vóór het tijdstip* waarop een bepaald soort onderwijs wordt aangevangen bepaalt of dit onderwijs zal worden gevolgd; de *stijging* in intelligentie *na afloop* van de voor dit onderwijs bestemde periode zal

(o. m.) veroorzaakt zijn door het onderwijs in kwestie.

§ 9. *Directe en indirecte samenhang*

De samenhang tussen twee grootheden kan voorts zijn *direct* of *indirect*, al naarmate er in het mechanisme dat de grootheden logisch verbindt geen dan wel een of meer tussenschakels bestaan. De hierboven gereleveerde samenhang tussen productie en prijs was een directe. Verschillende indirecte relaties kunnen daarnaast optreden. Zo zal, zoals reeds gezegd, bij een grote oogst van tarwe de prijs van tarwe laag zijn; daardoor zal dan tevens de prijs van rogge laag worden, omdat de vraag naar rogge door de overvloedige tarweoogst zal verminderen. Er is dus verband tussen tarweoogst (A) en roggeprijs (B) met als tussenschakel de tarweprijs (C). Schematisch voorgesteld aldus:

$$A \rightarrow C \rightarrow B$$

Er zal ook verband zijn tussen de tarweprijs en de productie van rogge in het volgend jaar (D) volgens het schema:

$$A_1 \rightarrow C_1 \rightarrow B_1 \rightarrow D_2$$

We hebben dan een relatie met 2 tussenketens.

De ketens behoeven niet *gelijk gericht* te zijn. Het kan ook zijn dat twee grootheden A en B beide gevolg van een derde (C) zijn. Zo zal een vermeerdering van kapitaalinvesteringen (C)

zowel een stijging van de prijzen van ijzer (A) als van de algemene werkgelegenheid (B) ten gevolge hebben. Te zeggen dat de werkgelegenheid is gestegen, omdat de ijzerprijzen zijn gestegen of ook omgekeerd, dat de ijzerprijzen zijn gestegen omdat de werkgelegenheid is gestegen, zou voorbarig zijn. De logische structuur, zoals we die hier beschreven hebben zou er schematisch aldus uit zien:

$$C \begin{matrix} \nearrow A \\ \searrow B \end{matrix}$$

Daarbij zouden ook nog wel andere relaties kunnen komen, doch daarover worde thans niet gesproken.

§ 10. *Enkelvoudige en meervoudige samenhang*

De oorzaken die op een gegeven verschijnsel hun directe invloeden uitoefenen kunnen *verschillende* zijn. Zo kan de stand van het loon (C) veranderd worden doordat er veranderingen in de prijzen (B) optreden, maar ook omdat er veranderingen in de werkgelegenheid (A) optreden [1]). Schematisch:

$$\begin{matrix} A \searrow \\ B \nearrow \end{matrix} C$$

[1]) Van complicaties als behandeld in De Nederlandsche Conjunctuur van Juni en September 1933, resp. blz. 12 en blz. 13, worde hier afgezien.

In zekere zin hebben we hier dus het omgekeerde van het logische verband als in de tweede helft van de vorige paragraaf werd besproken. Onze belangstelling is thans gericht op de grootheid *C*, in plaats van op *A*, en het belang van deze gevallen zal blijken bij de quantitatieve behandeling van de samenhang tussen grootheden.

De *meervoudige samenhang*, waarover we nu spreken in tegenstelling tot enkelvoudige in de tot hier behandelde gevallen, kan natuurlijk ook een groter aantal oorzaken dan twee vertonen.

Vooral in de sociale en anthropologische wetenschappen, waarbij gewoonlijk een groot aantal verschijnselen een rol spelen in één vraagstuk, is dit type van samenhang van groot belang.

Alles wat hier gezegd is over de vorm van de samenhang die er kan bestaan tussen twee of meer grootheden betreft dus, zoals reeds even werd gezegd, de logische structuur van het object dat men gaat onderzoeken; het zal duidelijk zijn, dat deze eerst bij wijze van hypothese moet zijn vastgesteld, voordat de statistiek haar controlerende functie kan gaan waarnemen. In een concreet voorbeeld kunnen dan verschillende combinaties voorkomen van de beschouwde eenvoudigste typen. Als voorbeeld daarvan zij hier een schema (zie fig. 1)

vermeld dat door de auteur werd opgesteld als uitgangspunt voor een aantal onderzoekingen betreffende het conjunctuurvraagstuk. In de linker (1e) kolom van bijgaand diagram zijn de drie grootheden opgesomd die in aanmerking werden genomen. Voor elke grootheid is in de daarmee corresponderende rij (= horizontale kolom) een serie punten getekend die door hun plaats aangeven het kwartaal waarop zij betrekking hebben. De peilen die het diagram

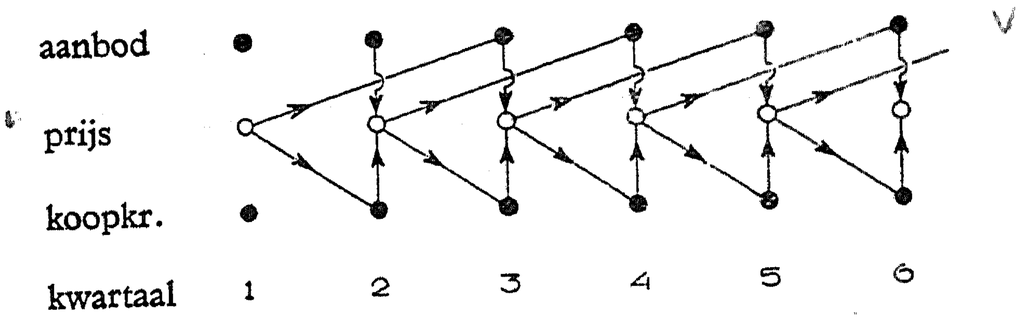

Fig 1. Logische structuur van een object. Voorbeeld: een vereenvoudigd model van het conjunctuurmechanisme in de economie.

doorkruisen, geven evenals hierboven aan, welke causale relaties werden aangenomen. Voor de kenner van de vraagstukken waarom het hier gaat zal duidelijk zijn, dat een zeer vereenvoudigd schema is gebruikt. Het is juist eenvoudig genoeg om mathematisch te kunnen worden gehanteerd en gecompliceerd genoeg om er enkele fundamentele problemen van conjunctuurpolitiek mee te kunnen belichten.

Een tweede, aan de psychologie ontleend, voorbeeld is gelegen in figuur 2, op pag. 44

afgedrukt, en ontleend aan Bigot en Kohnstamm, Hoofdstukken uit de Psychologie, blz. 128. Het gaat hier, zoals uit het onderschrift moge

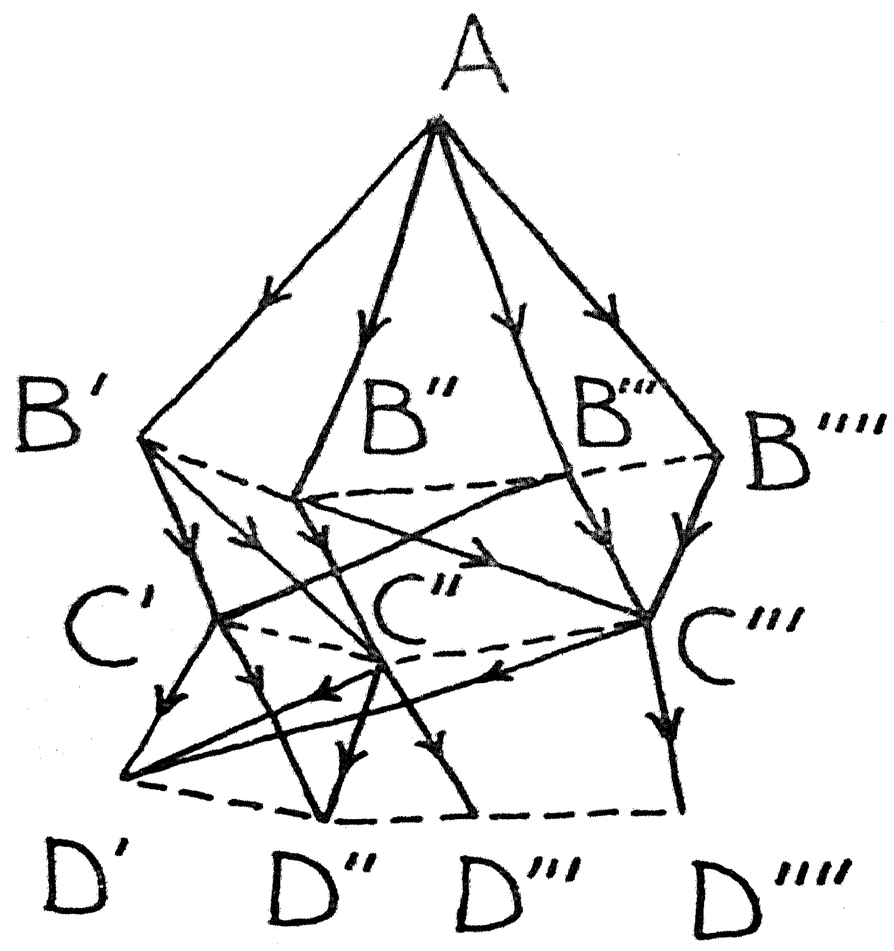

Fig. 2. Logische structuur van een object. Voorbeeld: de intelligentie. Eén bepaald intelligentieniveau *B*, *C* of *D* bevat nog verschillende toestanden. Voor het niveau is het geenszins onverschillig of de aangegeven verbindingen der niveau's alle of slechts gedeeltelijk functioneren. (Ontleend aan Bigot en Kohnstamm, Hoofdstukken uit de Psychologie).

blijken, om een lichtelijk verschillende kwestie, die echter toch duidelijk verwantschap vertoont met het voorgaande.

II. Quantitatieve beschrijving van de samenhang

§ 11. *Enkelvoudige functionele samenhang*

Wanneer het qualitatieve „voorwerk" is gedaan, kan de eigenlijke contrôle aanvangen. Deze gaat steeds uit van de grondgedachte, dat de wijzigingen in een „*te verklaren*" *reeks* cijfers, betrekking hebbend op het object dat men onderzoekt, moeten zijn te reconstrueren uit de wijzigingen die men in de „*verklarende reeksen*" aantreft. De verklarende reeksen zijn dan de cijfers betrekking hebbend op de verschijnselen, waarvan men eerst door redenering heeft vastgesteld, dat ze oorzaken moeten zijn van de wijzigingen in het object. Is men b.v. door economische redenering tot het inzicht gekomen dat de schommelingen in het verbruik van rundvlees worden bepaald door de veranderingen in:

1. het inkomen;
2. de prijs van rundvlees;
3. de prijs van varkensvlees;
4. de algemene kosten van levensonderhoud;

dan bestaat de contrôle hierin dat men de schommelingen in het verbruik van rundvlees tracht „op te bouwen" (in een mathematische betekenis) uit de overeenkomstige schommelingen die in de vier genoemde andere groot-

heden voorkomen. Die contrôle is dus essentiëel quantitatief [1]).

In de realiteit lukt het praktisch nooit om elke beweging tot in kleine details op de aangegeven wijze te „verklaren". Men neemt dan gewoonlijk aan — en terecht — dat naast de reeds beschouwde oorzaken nog andere werken, die de nog niet verklaarde afwijkingen tot gevolg hebben. Deze andere oorzaken zullen meestal minder belangrijk zijn; waren ze belangrijk, dan zou men ze in de beschouwing moeten opnemen. Ze zijn gewoonlijk ook talrijk en min of meer van elkaar onafhankelijk; en deze omstandigheden rechtvaardigen de werkhypothese dat de niet beschouwde oorzaken zgn. „toevallige afwijkingen" tot gevolg hebben. Over deze „toevallige afwijkingen" zal nog zeer uitvoerig sprake zijn in het derde gedeelte van dit werkje. Wij stellen ons in dit hoofdstuk op het standpunt, dat ze afwezig zijn. Er zijn dan alleen de „belangrijke" oorzaken, die de zgn. „systematische afwijkingen" tot gevolg hebben. Om deze reden luidt ook de titel van dit gedeelte „systematische samenhang". In

[1]) Dit woord wekt vaak misverstand. Men denkt bij quantitatief vaak aan quantitatief volgens zeer eenvoudige, „starre", rekenregels. Dat is in 't geheel niet nodig. De wijze waarop de ene grootheid de andere beïnvloedt kan desnoods zeer grillig en ingewikkeld zijn. Ze moet echter bepaald zijn; zonder die veronderstelling zou zich alles laten bewijzen.

de wiskunde is men gewend om dan te spreken van *functionele samenhang*. Gewoonlijk zal deze ook zijn een *eenduidige* functionele samenhang; d.w.z. dat bij een gegeven waardencombinatie van de verklarende reeksen slechts één waarde van de te verklaren reeks behoort. Er kan in bepaalde gevallen ook wel optreden een *meerduidig* functioneel verband, d.w.z. dat bij een gegeven waardencombinatie der verklarende reeksen meerdere waarden van de te verklaren reeks behoren.

De eenvoudigste gevallen, die wij hierbij zullen hebben te beschouwen, zijn die waarbij slechts *enkelvoudig verband* aanwezig is (§ 10). Deze zullen nu eerst worden behandeld. Dit kan gemakkelijk gebeuren aan de hand van vlakke grafische voorstellingen, waarbij langs de horizontale as de cijfers betreffende de oorzaakgrootheid, langs de vertikale de cijfers betreffende het gevolg worden afgezet. Laten we beginnen met daarbij de grootheden in hun absolute of natuurlijke maatstaven uit te drukken. Elk stel van 2 bijeenbehorende getallen wordt dan door een punt in de grafiek voorgesteld, waarvan die getallen de coördinaten zijn. Het feit van functionele samenhang tussen twee verschijnselen komt nu grafisch hierin tot uitdrukking, dat de verschillende punten op één lijn liggen.

Een eerste onderscheid dat nu te maken is, is dat tussen *rechtlijnig* en *kromlijnig* verband,

al naarmate bedoelde lijn recht of krom is. Veelal laat de nauwkeurigheid van practische waarnemingen niet toe om empirisch vast te stellen of de lijn waaromheen zich de punten groeperen een rechte dan wel een kromme is. Voorts laat zich een kromme lijn over een korte afstand gewoonlijk door een rechte benaderen. Deze beide omstandigheden zijn oorzaak, dat het voor de practijk van de hier bedoelde onderzoekingen zeer belangrijk is, het speciale geval van rechtlijnig verband — dat een eenvoudiger behandeling met formules toelaat — uitvoerig te bespreken.

§ 12. *Het rechtlijnig verband; de regressie-coëfficiënt*

Het *rechtlijnig verband* kan nog nader onderscheiden worden — wat voor sommige niet-mathematische lezers wel gewenst zal zijn — in drie soorten:

a. *gelijkheid* van de twee grootheden (d.w.z. de bijeenbehorende waarden der twee grootheden worden telken male door hetzelfde getal uitgedrukt): de grafische voorstelling is dan een lijn door de oorsprong (wanneer gelijke schalen worden gebezigd, onder 45° met de assen);

b. evenredigheid der twee grootheden (de ene grootheid is steeds een zeker aantal malen de andere grootheid): de grafische voorstelling

is eveneens een lijn door de oorsprong (bij gebruik van gelijke schalen nu echter onder een andere hoek dan 45° met de assen);

c. algemene lineaire samenhang: de lijn gaat nu niet door de oorsprong. De getallen, die nu de ene grootheid uitdrukken worden gevonden door een zeker aantal malen de tweede te nemen en daarbij nog een vast bedrag op te tellen (af te trekken). In formules uitgedrukt zijn de gevallen gekenmerkt door:

$$1)\quad X = Y$$
$$2)\quad X = bY$$
$$3)\quad X = bY + a$$

Anders uitgedrukt, kan men ook zeggen dat voor alle drie de gevallen de formule $X = bY + a$ gelden, doch dat voor:

$$1)\quad b = 1 \quad a = 0$$
$$2)\quad b \neq 1 \quad a = 0$$
$$3)\quad a \neq 0$$

geldt.

Opgemerkt kan nog worden, dat in geval 3) de opeenvolgende *verschillen* van de grootheden A en B weer wèl evenredig zijn en dus onder geval 2) of zelfs 1) behoren.

In sommige practische gevallen is het er slechts om te doen, te zien of er inderdaad lineair verband bestaat tussen de twee grootheden; en als dat dan aanwezig blijkt, vindt men empirisch, hoe groot b en a zijn, dus ook,

in welk der gevallen 1), 2) of 3) men verkeert. Dat is b.v. zo, wanneer men in de economie met een aanbodlijn te doen heeft. De redenering, ontleend aan de economie, levert ons daar slechts de verwachting dat er verband is, doch niet dat daarbij bovendien *b* en *a* een bepaalde waarde zouden moeten hebben (hoogstens, vooral ten aanzien van *b*, dat ze een bepaald teken zouden moeten hebben). In andere gevallen echter is het er juist om te doen, te vinden dat er gelijkheid bestaat tussen de twee reeksen. Zo b.v. bij het toekennen van cijfers aan leerlingen. Het is daar niet voldoende — om een goede beoordeling te hebben van een groep leerlingen — dat b.v. het hoofd van de lagere school en het toelatingsexamen de verschillende leerlingen in dezelfde volgorde plaatsen, doch het is gewenst dat ze ook minstens de scheiding tussen voldoende en onvoldoende op dezelfde plaats in de reeks leerlingen aanbrengen, en liefst gelijke cijfers geven. Het probleem is hier dus lichtelijk verschillend.

Bij de eerstgenoemde problemen heeft men er veel nut van, wanneer men in plaats van de absolute of natuurlijke maatstaf der grootheden gebruikt een maatstaf die zijn beginpunt heeft bij het gemiddelde der beschouwde grootheid. Men zegt dan, zoals reeds werd opgemerkt, dat de cijfers „afwijkingen van het gemiddelde" voorstellen. We zullen in het vervolg deze

afwijkingen steeds door *kleine letters*, in tegenstelling tot de voor de absolute cijfers gebruikte hoofdletters, weergeven. Voor deze afwijkingen van het gemiddelde bestaan er weer, zoals men gemakkelijk inziet, slechts de gevallen 1) en 2).

Voor bepaalde doeleinden, in hoofdstuk III nader te bespreken, gaat men nog verder en kiest men ook nog een andere *eenheid*. En wel kiest men dan de zgn. *standaarddeviatie* (zoals reeds opgemerkt, ook wel *standaardafwijking* of *middelbare afwijking* genoemd) tot eenheid. Zijn de afwijkingen van het gemiddelde resp. x_1, $x_2 \ldots . x_n$, waarbij dus n is het aantal waarnemingen, dan wordt de standaarddeviatie, die bijna steeds door σ_x wordt weergegeven, gedefiniëerd door:

$$\sigma_x^2 = \frac{1}{n}\left(x_1^2 + x_2^2 + \ldots . x_n^2\right)$$

of

$$n\sigma_x^2 = \sum_1^n{}^i x_i^2.$$

We zullen de grootheden, in deze eenheden, en vanaf haar gemiddelde gemeten, aanduiden door Griekse letters; zodat dus b.v. geldt:

$$\xi_1 = \frac{x_1}{\sigma_x}, \xi_2 = \frac{x_2}{\sigma_x}, \ldots . \xi_n = \frac{x_n}{\sigma_x}$$

of algemeen

$$\xi_i = \frac{x_i}{\sigma_x} \quad (i = 1 \ldots . n)$$

en we zeggen dat de ξ's „op standaarddeviatie staan". Anders uitgedrukt: de standaarddeviatie

van de ξ's zelf is 1. Tussen twee grootheden op standaarddeviatie, die lineair verband vertonen, kan niet anders dan *gelijkheid* bestaan, zoals men gemakkelijk narekent. Hier kan dus alleen geval 1) bestaan.

De coëfficiënt b, hierboven ingevoerd in de vergelijking die het verband tussen X en Y uitdrukte, noemt men de *regressiecoëfficiënt* [1]). Deze geeft dus aan met hoeveel van de gekozen eenheden X stijgt, wanneer Y met één eenheid stijgt. Kiest men als eenheden de resp. standaarddeviaties van X en Y, dan is de regressiecoëfficiënt noodzakelijkerwijze 1.

§ 13. *Kromlijnig verband; meervoudig verband*

Kromlijnig verband

Zoals reeds is opgemerkt, zijn er ook wel

[1]) De naam regressiecoëfficiënt is op merkwaardige wijze in de statistiek in gebruik gekomen. Met regressie werd door Galton aangeduid het verschijnsel dat de kinderen bepaalde kenmerken van hun ouders in 't algemeen ook vertonen doch — alweer in 't algemeen genomen — in mindere mate, zodat er dus een neiging tot een „terugkeer", een „regressie" naar het gemiddelde aanwezig is. De regressiecoëfficiënt gaf in zijn geval de verhouding weer waarmede die regressie zich voordoet. Voor het begrip van deze coëfficiënt doet het er echter natuurlijk niets toe of er een dergelijke „terugkeer" is; ook als er een steeds verdere verwijdering van het gemiddelde zou plaats hebben, zou een dergelijke coëfficiënt betekenis hebben. In de naam regressiecoëfficiënt zit dus reeds een bepaalde theorie van de erfelijkheid besloten.

gevallen waarin de nauwkeurigheid van meting groot genoeg is, om te kunnen vaststellen dat het verband tussen twee grootheden niet lineair, maar kromlijnig is. In dat geval is de grafische voorstelling dus een kromme lijn; deze kan zowel een mathematische als een empirische kromme zijn. Er is hier geen aanleiding om van „de" regressiecoëfficiënt te spreken. Veelal rekent men bij het nader beschouwen van deze soort samenhang ook met afwijkingen van het gemiddelde. De kromme lijn, die het verband voorstelt tussen de aldus gemeten grootheden, zal echter in 't algemeen niet (zoals bij de rechte lijn) door de oorsprong van het assenkruis gaan.

Meervoudig verband

Bij meervoudig verband zou men in principe van een meerdimensionale (in casu maximaal driedimensionale) grafische voorstelling gebruik kunnen maken, doch dat ontmoet veelal bezwaren; alleen voor demonstratiedoeleinden hebben dergelijke voorstellingen een grote waarde. Er zijn echter ook methoden, om een samenhang waarin meer dan twee grootheden een rol spelen in tweedimensionale (vlakke) grafieken af te beelden. Enkele voorbeelden zullen daarvan hier behandeld worden.

Een eerste voorbeeld is de wijze waarop men reeds in de (theoretische) economie gewend was om de vraagfunctie grafisch te bestuderen,

wanneer niet alleen de prijs, maar ook de een of andere verdere oorzaak van vraagveranderingen in de beschouwingen werd opgenomen. Om de gedachten te bepalen, nemen we aan dat die verdere oorzaak alleen het inkomen van de betr. groep mensen is. De gevraagde hoeveelheid goederen X hangt dan dus af van de prijs Y en het inkomen Z; in algemeen wiskundige symbolen: $X = f(Y, Z)$

Zouden we aannemen dat het verband ook thans lineair is, zowel in Y als in Z, dan zou deze algemene formule de vorm aannemen:

$$X = b_{12}Y + b_{13}Z + a$$

Voor het gemak van het spreken zullen we verder deze formule gebruiken. De lezer zal bemerken dat onze opmerkingen echter evenzeer gelden voor de algemene formule.

Wil men nu de aangegeven samenhang systematisch gaan bestuderen, dan is het geschikt om aan Z eerst een vaste waarde Z_0 te geven en na te gaan welk verband er nu nog bestaat tussen X en Y. Dit wordt gegeven door

$$X = b_{12}Y + (b_{13}Z_0 + a)$$

waarbij we de laatste twee termen tussen haakjes hebben gezet om aan te duiden dat deze te zamen nu een constante term vormen. Het verband tussen X en Y, bij gegeven Z_0, wordt dan nu door een lijn voorgesteld in fig. 3 aangegeven met $X = f(Y, Z_0)$. In het bijzondere geval dat we hier beschouwen is dat een rechte lijn, die

van de Y-as afsnijdt een segment $b_{13}Z_0 + a$.

Geven we nu Z een andere waarde Z_1, dan krijgen we voor de gevallen die daaronder behoren een verband

$$X = b_{12}Y + (b_{13}Z_1 + a)$$

of algemeen $X = f(Y, Z_1)$ in de figuur aangegeven door de aldus aangeduide lijn. Deze is

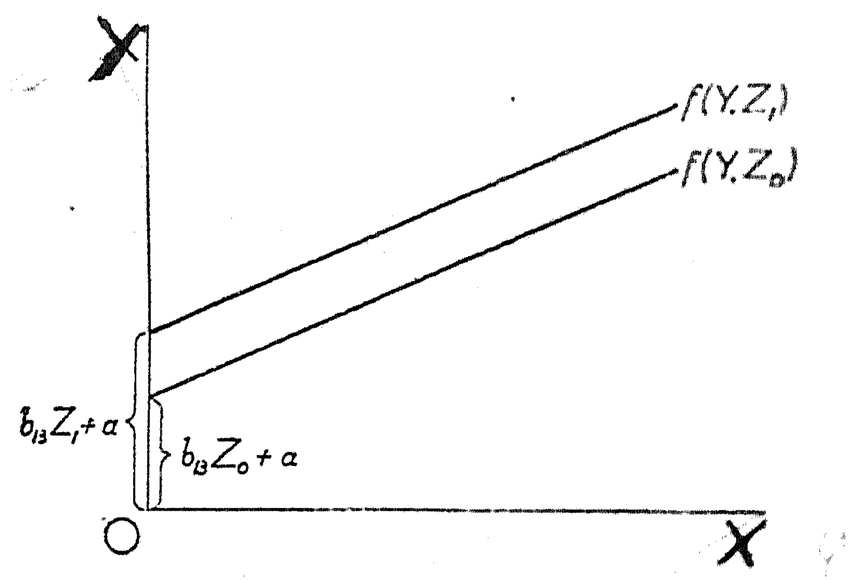

Fig. 3. Grafische voorstelling van een verband tussen drie grootheden, $X = f(Y, Z)$, voor twee waarden van Z, nl. Z_0 en Z_1.

in ons speciale geval een rechte, die een bedrag $b_{13}Z_1 + a$ van de Y-as afsnijdt. We kunnen zo een systeem van lijnen verkrijgen dat als geheel een denkbeeld geeft van het meervoudig verband [1]).

[1]) Dit systeem kan beschouwd worden als de projectie op het XOY-vlak van een systeem van „lijnen van constante Z" in de driedimensionale voorstelling getrokken op het oppervlak $X = f(Y, Z)$.

De verschuiving van de „vraaglijn bij constante Z" [1]) bij verandering van de waarde van Z is een in de literatuur vaak besproken verschijnsel, onder de Engelse naam „shifting" bekend. In ons eenvoudige geval is deze shifting evenredig met de verandering in Z.

2) Een tweede methode, meer bekend uit de statistische onderzoekingen van Amerikaanse landbouwstatistici (speciaal Ezekiel) maakt gebruik van twee grafische voorstellingen (zie fig. 4). In de bovenste worden met elkaar vergeleken de grootheden X en Y (waarbij, om aan te sluiten bij de genoemde schrijvers, nu Y langs de horizontale as is afgezet). De punten P_1, P_2, enz., die elk als coördinaten hebben bijeenbehorende waarden van X en Y (P_1 heeft de coördinaten X_1 en Y_1, P_2 : X_2 en Y_2, enz.), liggen nu niet op één lijn, want al naarmate de Z_1, Z_2, enz. verschillend zijn, zijn er afwijkingen van de lijn

$$X' = b_{12}Y + a$$

van verschillende omvang aanwezig. Anders gezegd, er is een onderscheid tussen X en deze X', dat afhankelijk is van Z, want

$$X - X' = b_{13}Z$$

[1]) Zoals men weet, wordt in de economie de vraaglijn ook onder deze voorwaarde gedefiniëerd en is de toevoeging dus eigenlijk overbodig: men behoeft slechts de waarde van Z te vermelden.

Om die reden worden nu de waarden $X—X'$, die ook wel de *residuen* genoemd worden, in de onderste grafiek als ordinaat vergeleken met de

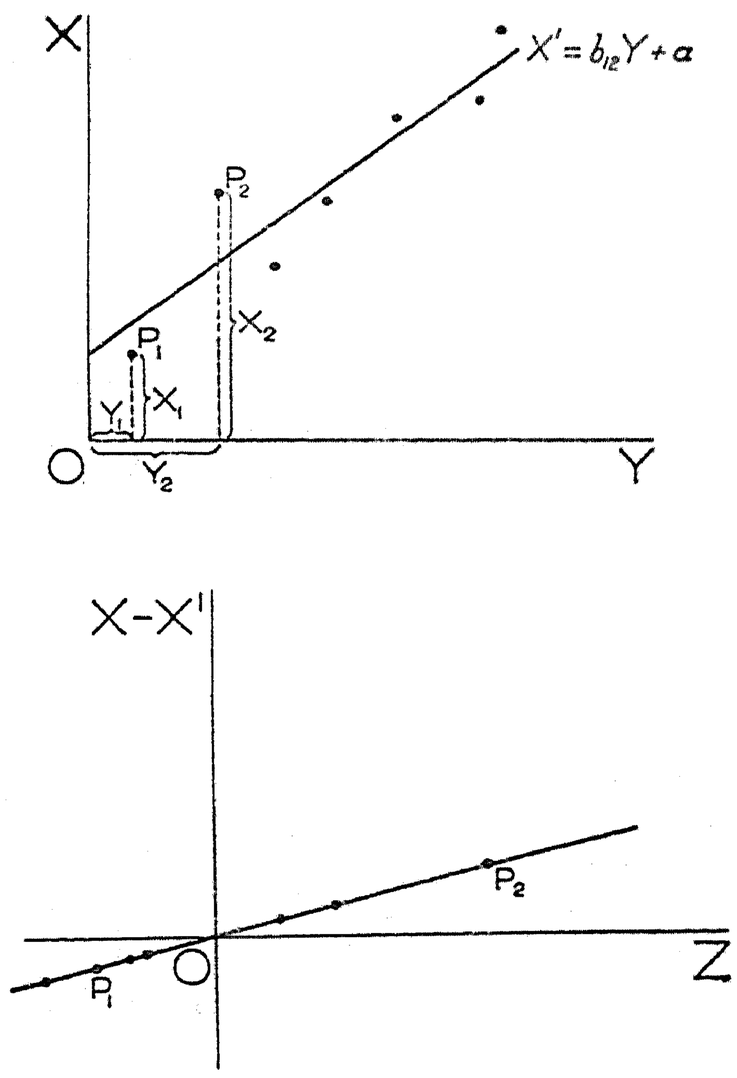

Fig. 4. Methode van Ezekiel voor het grafisch voorstellen van een functie van twee veranderlijken $X(Y,Z)$.

abscissen Z. Tussen deze residuen en de Z-coördinaten bestaat nu een functioneel verband.

Veelal wordt deze methode niet voor de

absolute waarden X, Y en Z toegepast, doch voor de afwijkingen van het gemiddelde x, y en z.

Het spreekt van zelf dat men de hier aangeroerde vraagstukken ook algebraïsch kan behandelen. Belangwekkende voorbeelden daarvan zijn gegeven voor pedagogische vraagstukken, b.v. bij behandeling volgens de methode van Spearman [1]). Men neemt daarbij aan dat de rapportcijfers x_i voor verschillende vakken ($i = 1$, 2 en 3 b.v.) zich elk laten weergeven door een lineaire functie van een algemene grootheid g en een grootheid s_i die voor elk vak verschillend is; aldus:

$$x_1 = c_1\, g + c_1'\, s_1 \qquad (13.1)$$
$$x_2 = c_2\, g + c_2'\, s_2 \qquad (13.2)$$
$$x_3 = c_3\, g + c_3'\, s_3 \qquad (13.3)$$

Daarbij stelt dan g voor een soort van algemene intelligentie, s_1 een speciale aanleg voor vak No. 1, s_2 voor vak 2, enz. De g is voor elke leerling verschillend, maar voor alle vakken dezelfde; de coëfficiënten zijn daarentegen voor alle leerlingen dezelfde. Men stelt nu de verdere voorwaarden dat de grootheden s_1, s_2, s_3

[1]) Voor een zeer duidelijke uiteenzetting zie men b.v. Dr. B. M. van Dalfsen, De correlatie tusschen Meetkunde en Fransch, Paedagogische Studiën X (1929) blz. 231 en De samenhang der Rapportcijfers voor de verschillende leervakken eener H. B. S., Paed. Studiën XI (1930), blz. 230, speciaal blz. 247—250.

en g onderling onafhankelijk zijn, hetgeen wordt vertaald door de eis dat de correlatiecoëfficiënten tussen deze grootheden onderling nul zijn (vgl. voor de betekenis van deze term § 21 en volgende), hetgeen aanleiding is tot de vgln:

$$\Sigma s_1 s_2 = 0 \quad \Sigma s_1 s_3 = 0 \quad \Sigma s_2 s_3 = 0$$
$$\Sigma s_1 g = 0 \quad \Sigma s_2 g = 0 \quad \Sigma s_3 g = 0$$

waarbij de sommatie plaats heeft over alle leerlingen. Bovendien wordt de eis gesteld dat de standaarddeviaties van g, s_1, s_2 en s_3 gelijk aan de eenheid zijn. Door deze eisen laten zich de coëfficiënten c_1 t/m c_3' berekenen.

Ofschoon bij de desbetreffende formuleringen een ruim gebruik gemaakt wordt van de begrippen der correlatierekening, is de probleemstelling toch essentiëel anders dan de problemen die wij te dien opzichte later zullen behandelen. Men veronderstelt nl. dat de vgln. (13.1), (13.2) en (13.3) exact gelden voor elken leerling en elk vak, terwijl bij de vraagstukken die later besproken worden, aangenomen zal worden dat er een verschil bestaat tussen de rechter- en de linkerzijde dezer vergelijkingen. De „restloze" verklaring van de grootheden x_1, x_2 en x_3 in het bovenstaande geval is alleen mogelijk doordat men grootheden s_1, s_2 en s_3 introduceert in even groten getale als de grootheden x_1, x_2 en x_3; in de later te bespreken problemen wordt aangenomen dat de grootheden s reeds a priori gegeven zijn.

De hier gevolgde behandelingswijze komt sterk overeen met de in § 33 te bespreken probleemstelling van Frisch.

Opgemerkt moge nog worden, dat de in de vergelijkingen (13.1), (13.2) en (13.3) gegeven opzet slechts mogelijk is bij juist *drie* vakken. Als men meer vakken beschouwt, moeten andere wegen ingeslagen worden; één daarvan is, dat naast de algemene factor g nog „groepsfactoren" worden ingevoerd. In het reeds genoemde artikel van den heer Van Dalfsen worden nog ingevoerd een factor x, aangeduid als wiskunde-aanleg, die tot uiting komt in de cijfers voor algebra, meetkunde en tekenen; een factor y, genoemd talen-aanleg, tot uiting komende bij Nederlands, Frans en Duits, en een factor z, die de schrijver aantrof bij de vakken algebra en vreemde talen en die hij accuratesse noemde.

§ 14. *Multicollineariteit*

Bij gevallen waarin men met meer dan drie veranderlijken werkt is uiteraard een grafische voorstelling op de gewone wijze onmogelijk. En bij een groeiend aantal veranderlijken worden ook de andere grafische hulpmiddelen steeds ingewikkelder te hanteren. Veelal neemt men dan zijn toevlucht tot de zuiver algebraïsche methode van berekenen, welke in § 19, bij de correlatierekening, zal worden behandeld. Deze stelt ons, zeer in 't algemeen gesproken, in staat

om de regressievergelijkingen te vinden ook bij een willekeurig aantal veranderlijken.

Frisch (Statistical Confluence Analysis by Means of Complete Regression Systems) [1]) wijst op de gevaren die deze methode meebrengt, wanneer zgn. *multicollineariteit* optreedt, d. w. z. dat er tussen de veranderlijken die men als „verklarende reeksen" heeft aangenomen, betrekkingen (en wel speciaal lineaire) bestaan. Enige der coëfficiënten kunnen daardoor volkomen onbepaald worden. In de realiteit, waar altijd afwijkingen van de volkomen functionele samenhang voorkomen, verbergen deze afwijkingen dan vaak de onbepaaldheid doordat men b.v. niet een uitkomst $\frac{0}{0}$ verkrijgt, doch een uitkomst $\frac{\text{kleine afwijking}}{\text{andere kleine afwijking}}$, waardoor een schijnbare bepaaldheid ontstaat die echter volkomen misleidend is.

In de terminologie van de hoger-dimensionale meetkunde kan men dit aldus uitdrukken dat de waarnemingen in plaats van te liggen in één $(n-1)$-dimensionale ruimte (bij $n-1$ verklarende reeksen en een n-dimensionale ruimte om het verband voor te stellen), doch in een ruimte van lagere dimensie. Het eenvoudigst denkbare geval is dat van $n = 3$. In 't algemeen

[1]) Universitetets Økonomiske Institutt, Oslo 1934, Publ. Nr. 5.

vgl collegeleer matrix

gesproken zal een (lineair) verband tussen 2 verklarende en een te verklaren reeks worden voorgesteld door de punten van een plat vlak; wanneer echter tussen de twee verklarende reeksen nog een verband bestaat, degenereert dat platte vlak tot een rechte lijn. Frisch geeft in zijn genoemd werk methoden aan om hier de juiste conclusie te trekken. Deze methoden kunnen hier echter niet nader worden besproken.

TWEEDE GEDEELTE

SYSTEMATISCHE SAMENHANG TUSSEN TWEE EN MEER VERSCHIJNSELEN

I. QUALITATIEVE BESCHRIJVING VAN DE SAMENHANG

§ 7. *Doel van het meten*

Het meten van de omvang van bepaalde verschijnselen is in veel gevallen doel op zichzelf. Men houdt statistieken bij van prijzen, van lonen, van verkochte hoeveelheden, van het volksinkomen, van de betalingsbalans, omdat men zich voor de getallen betreffende deze verschijnselen op zich zelf interesseert.

In een toenemend aantal gevallen is echter het doel van de meting een verder liggend: men wenst de samenhang tussen meer dan een verschijnsel statistisch te onderzoeken. Dat wil zeggen: *men wenst statistisch te controleren of een op andere wijze aannemelijk gemaakte samenhang inderdaad aanwezig is.* Deze andere wijze van aannemelijk maken is gewoonlijk ontleend aan een der wetenschappen die zich met de bedoelde verschijnselen inlaat. Men zal b.v. op grond van de wetenschap der economie een

verband vermoeden tussen de prijs van een artikel en de productie daarvan. Of men zal op grond van de biologie een verband aanwezig achten tussen de temperatuur van het Zuiderzeewater en de ansjovisvangst.

Uitdrukkelijk wordt hier vastgesteld, dat de rol van de statistiek in deze slechts een controlerende kan zijn en het statistisch onderzoek moet worden voorafgegaan door het onderzoek binnen de betreffende tak van wetenschap. Men heeft wel gemeend, dat men langs statistische weg een verband kan bewijzen. Wat men kan bewijzen is echter nooit meer dan een overeenstemming in de bewegingen van twee of meer verschijnselen. In welke richting het causaal verband echter in zo'n geval loopt is statistisch bijna nooit met zekerheid uit te maken. Men kan alleen constateren of een bepaalde richting al dan niet met de feiten in overeenstemming is. Zo kan natuurlijk niet de oorzaak op een later tijdstip plaats hebben dan het gevolg, uit welke omstandigheid soms conclusies van betekenis zijn te trekken.

Een en ander zal nog duidelijker worden door een nadere analyse van de *logische structuurvormen* waarop men zoal kan stuiten. We zouden de hier volgende ontleding ook kunnen noemen de qualitatieve beschrijving van de samenhang. Deze moet, zoals reeds gezegd, steeds gebaseerd

worden op de wetenschap welke zich met de verschijnselen in kwestie bezighoudt.

§ 8. *Richting van het causale verband*

We hebben reeds even de kwestie aangeroerd, *in welke richting* het causale verband gedacht moet worden. Als twee grootheden samenhangen, kan het zijn dat A oorzaak en B gevolg is (symbolisch aan te geven door $A \rightarrow B$); het kan ook dat B oorzaak en A gevolg is ($B \rightarrow A$). Veelal liggen de zaken echter niet zo eenvoudig. Men pleegt dan nog al eens het woord wisselwerking te gebruiken. Daarmee wordt echter vaak het vraagstuk weggepraat. In vele gevallen kan men daar tot een scherper stellen van de logische structuur komen door een nadere onderscheiding in te voeren, speciaal door onderscheid te maken naar het tijdstip waarop men elk der verschijnselen A en B meet. We zullen dan met A_1 aangeven de omvang van A op tijdstip 1, met A_2 op tijdstip 2, en evenzo de symbolen B_1, B_2, enz. gebruiken. Als men met de symbolen A, B, enz. verschijnselen aanduidt, geven A_1, A_2, B_1, B_2, enz. *gebeurtenissen* aan (in de tijd gelocaliseerde verschijnselen). Het kan nu b.v. zo staan met de samenhang tussen A en B, dat A_1 bepalend is voor B_1 (A_1 oorzaak en B_1 gevolg), doch B_1 bepalend voor A_2 (B_1 oorzaak en A_2 gevolg), A_2 voor B_2, B_2 voor A_3 enz. Een en ander

is te verenigen in het volgende schema:

$$\begin{array}{ccccccccc} A_1 & & A_2 & & A_3 & & A_4 & & \\ \downarrow & \nearrow & \downarrow & \nearrow & \downarrow & \nearrow & \downarrow & \nearrow & \text{enz.} \\ B_1 & & B_2 & & B_3 & & B_4 & & \end{array}$$

Iets dergelijks doet zich voor bij de productie (A) en de prijzen (B) van sommige landbouw- en veeteeltproducten. De productie in het eerste jaar bepaalt door het „vraagverband" de prijshoogte in dat jaar (als er veel geproduceerd is, kan dit door de consumenten alleen opgenomen worden bij een lage prijs); de prijshoogte in dat jaar bepaalt (soms naast andere oorzaken) de omvang van de productie in het volgend jaar; enz. Noch prijs, noch productie is oorzaak zonder nadere aanduiding.

Een ander voorbeeld kan ontleend worden aan de psychologie of pedagogie als men de samenhang bespreekt tussen intelligentie en genoten onderwijs. De intelligentie van een bepaalde persoon kan zowel gevolg zijn van genoten onderwijs als ook oorzaak daarvan, zo wordt betoogd. Ook hier ligt de zaak dan ongeveer zo dat de intelligentie *vóór het tijdstip* waarop een bepaald soort onderwijs wordt aangevangen bepaalt of dit onderwijs zal worden gevolgd; de *stijging* in intelligentie *na afloop* van de voor dit onderwijs bestemde periode zal

(o. m.) veroorzaakt zijn door het onderwijs in kwestie.

§ 9. *Directe en indirecte samenhang*

De samenhang tussen twee grootheden kan voorts zijn *direct* of *indirect*, al naarmate er in het mechanisme dat de grootheden logisch verbindt geen dan wel een of meer tussenschakels bestaan. De hierboven gereleveerde samenhang tussen productie en prijs was een directe. Verschillende indirecte relaties kunnen daarnaast optreden. Zo zal, zoals reeds gezegd, bij een grote oogst van tarwe de prijs van tarwe laag zijn; daardoor zal dan tevens de prijs van rogge laag worden, omdat de vraag naar rogge door de overvloedige tarweoogst zal verminderen. Er is dus verband tussen tarweoogst (A) en roggeprijs (B) met als tussenschakel de tarweprijs (C). Schematisch voorgesteld aldus:

$$A \rightarrow C \rightarrow B$$

Er zal ook verband zijn tussen de tarweprijs en de productie van rogge in het volgend jaar (D) volgens het schema:

$$A_1 \rightarrow C_1 \rightarrow B_1 \rightarrow D_2$$

We hebben dan een relatie met 2 tussenketens.

De ketens behoeven niet *gelijk gericht* te zijn. Het kan ook zijn dat twee grootheden A en B beide gevolg van een derde (C) zijn. Zo zal een vermeerdering van kapitaalinvesteringen (C)

zowel een stijging van de prijzen van ijzer (A) als van de algemene werkgelegenheid (B) ten gevolge hebben. Te zeggen dat de werkgelegenheid is gestegen, omdat de ijzerprijzen zijn gestegen of ook omgekeerd, dat de ijzerprijzen zijn gestegen omdat de werkgelegenheid is gestegen, zou voorbarig zijn. De logische structuur, zoals we die hier beschreven hebben zou er schematisch aldus uit zien:

$$C \begin{matrix} \nearrow A \\ \searrow B \end{matrix}$$

Daarbij zouden ook nog wel andere relaties kunnen komen, doch daarover worde thans niet gesproken.

§ 10. *Enkelvoudige en meervoudige samenhang*

De oorzaken die op een gegeven verschijnsel hun directe invloeden uitoefenen kunnen *verschillende* zijn. Zo kan de stand van het loon (C) veranderd worden doordat er veranderingen in de prijzen (B) optreden, maar ook omdat er veranderingen in de werkgelegenheid (A) optreden [1]). Schematisch:

$$\begin{matrix} A \searrow \\ B \nearrow \end{matrix} C$$

[1]) Van complicaties als behandeld in De Nederlandsche Conjunctuur van Juni en September 1933, resp. blz. 12 en blz. 13, worde hier afgezien.

In zekere zin hebben we hier dus het omgekeerde van het logische verband als in de tweede helft van de vorige paragraaf werd besproken. Onze belangstelling is thans gericht op de grootheid *C*, in plaats van op *A*, en het belang van deze gevallen zal blijken bij de quantitatieve behandeling van de samenhang tussen grootheden.

De *meervoudige samenhang*, waarover we nu spreken in tegenstelling tot enkelvoudige in de tot hier behandelde gevallen, kan natuurlijk ook een groter aantal oorzaken dan twee vertonen.

Vooral in de sociale en anthropologische wetenschappen, waarbij gewoonlijk een groot aantal verschijnselen een rol spelen in één vraagstuk, is dit type van samenhang van groot belang.

Alles wat hier gezegd is over de vorm van de samenhang die er kan bestaan tussen twee of meer grootheden betreft dus, zoals reeds even werd gezegd, de logische structuur van het object dat men gaat onderzoeken; het zal duidelijk zijn, dat deze eerst bij wijze van hypothese moet zijn vastgesteld, voordat de statistiek haar controlerende functie kan gaan waarnemen. In een concreet voorbeeld kunnen dan verschillende combinaties voorkomen van de beschouwde eenvoudigste typen. Als voorbeeld daarvan zij hier een schema (zie fig. 1)

vermeld dat door de auteur werd opgesteld als uitgangspunt voor een aantal onderzoekingen betreffende het conjunctuurvraagstuk. In de linker (1e) kolom van bijgaand diagram zijn de drie grootheden opgesomd die in aanmerking werden genomen. Voor elke grootheid is in de daarmee corresponderende rij (= horizontale kolom) een serie punten getekend die door hun plaats aangeven het kwartaal waarop zij betrekking hebben. De peilen die het diagram

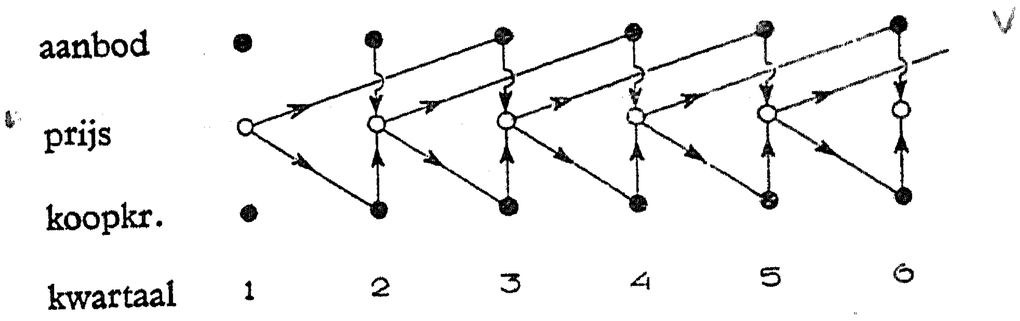

Fig 1. Logische structuur van een object. Voorbeeld: een vereenvoudigd model van het conjunctuurmechanisme in de economie.

doorkruisen, geven evenals hierboven aan, welke causale relaties werden aangenomen. Voor de kenner van de vraagstukken waarom het hier gaat zal duidelijk zijn, dat een zeer vereenvoudigd schema is gebruikt. Het is juist eenvoudig genoeg om mathematisch te kunnen worden gehanteerd en gecompliceerd genoeg om er enkele fundamentele problemen van conjunctuurpolitiek mee te kunnen belichten.

Een tweede, aan de psychologie ontleend, voorbeeld is gelegen in figuur 2, op pag. 44

afgedrukt, en ontleend aan Bigot en Kohnstamm, Hoofdstukken uit de Psychologie, blz. 128. Het gaat hier, zoals uit het onderschrift moge

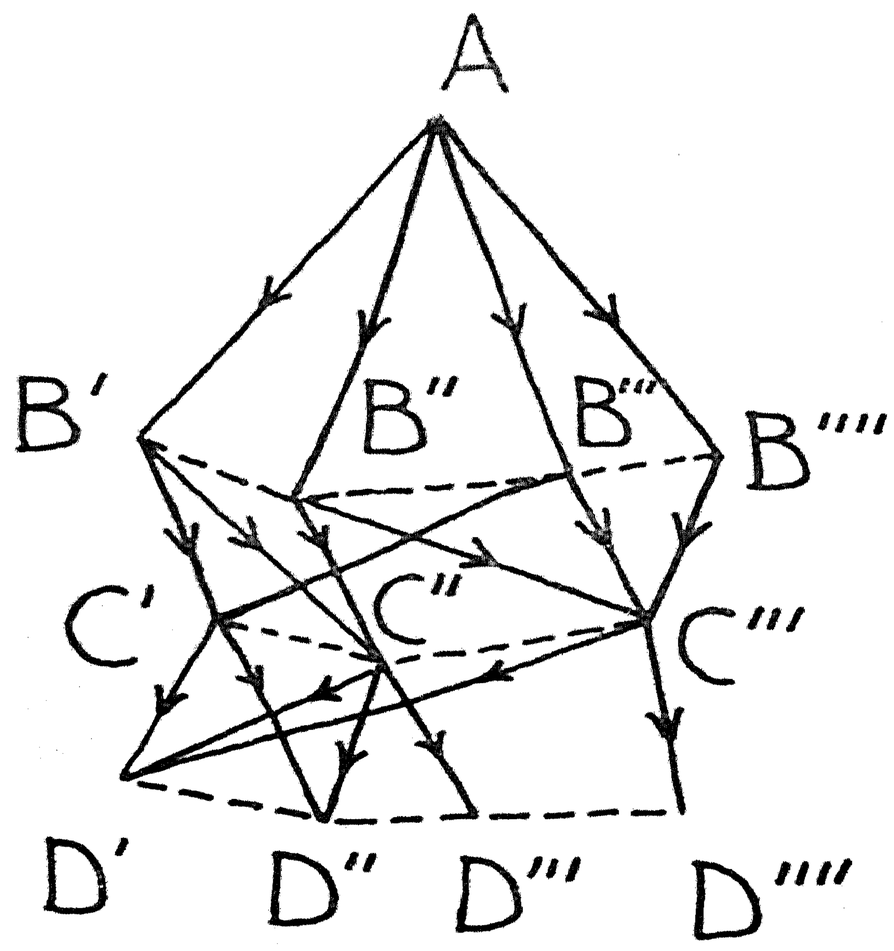

Fig. 2. Logische structuur van een object. Voorbeeld: de intelligentie. Eén bepaald intelligentieniveau *B*, *C* of *D* bevat nog verschillende toestanden. Voor het niveau is het geenszins onverschillig of de aangegeven verbindingen der niveau's alle of slechts gedeeltelijk functioneren. (Ontleend aan Bigot en Kohnstamm, Hoofdstukken uit de Psychologie).

blijken, om een lichtelijk verschillende kwestie, die echter toch duidelijk verwantschap vertoont met het voorgaande.

II. Quantitatieve beschrijving van de samenhang

§ 11. *Enkelvoudige functionele samenhang*

Wanneer het qualitatieve „voorwerk" is gedaan, kan de eigenlijke contrôle aanvangen. Deze gaat steeds uit van de grondgedachte, dat de wijzigingen in een „*te verklaren*" *reeks* cijfers, betrekking hebbend op het object dat men onderzoekt, moeten zijn te reconstrueren uit de wijzigingen die men in de „*verklarende reeksen*" aantreft. De verklarende reeksen zijn dan de cijfers betrekking hebbend op de verschijnselen, waarvan men eerst door redenering heeft vastgesteld, dat ze oorzaken moeten zijn van de wijzigingen in het object. Is men b.v. door economische redenering tot het inzicht gekomen dat de schommelingen in het verbruik van rundvlees worden bepaald door de veranderingen in:

1. het inkomen;
2. de prijs van rundvlees;
3. de prijs van varkensvlees;
4. de algemene kosten van levensonderhoud;

dan bestaat de contrôle hierin dat men de schommelingen in het verbruik van rundvlees tracht „op te bouwen" (in een mathematische betekenis) uit de overeenkomstige schommelingen die in de vier genoemde andere groot-

heden voorkomen. Die contrôle is dus essentiëel quantitatief [1]).

In de realiteit lukt het praktisch nooit om elke beweging tot in kleine details op de aangegeven wijze te „verklaren". Men neemt dan gewoonlijk aan — en terecht — dat naast de reeds beschouwde oorzaken nog andere werken, die de nog niet verklaarde afwijkingen tot gevolg hebben. Deze andere oorzaken zullen meestal minder belangrijk zijn; waren ze belangrijk, dan zou men ze in de beschouwing moeten opnemen. Ze zijn gewoonlijk ook talrijk en min of meer van elkaar onafhankelijk; en deze omstandigheden rechtvaardigen de werkhypothese dat de niet beschouwde oorzaken zgn. „toevallige afwijkingen" tot gevolg hebben. Over deze „toevallige afwijkingen" zal nog zeer uitvoerig sprake zijn in het derde gedeelte van dit werkje. Wij stellen ons in dit hoofdstuk op het standpunt, dat ze afwezig zijn. Er zijn dan alleen de „belangrijke" oorzaken, die de zgn. „systematische afwijkingen" tot gevolg hebben. Om deze reden luidt ook de titel van dit gedeelte „systematische samenhang". In

[1]) Dit woord wekt vaak misverstand. Men denkt bij quantitatief vaak aan quantitatief volgens zeer eenvoudige, „starre", rekenregels. Dat is in 't geheel niet nodig. De wijze waarop de ene grootheid de andere beïnvloedt kan desnoods zeer grillig en ingewikkeld zijn. Ze moet echter bepaald zijn; zonder die veronderstelling zou zich alles laten bewijzen.

de wiskunde is men gewend om dan te spreken van *functionele samenhang*. Gewoonlijk zal deze ook zijn een *eenduidige* functionele samenhang; d.w.z. dat bij een gegeven waardencombinatie van de verklarende reeksen slechts één waarde van de te verklaren reeks behoort. Er kan in bepaalde gevallen ook wel optreden een *meerduidig* functioneel verband, d.w.z. dat bij een gegeven waardencombinatie der verklarende reeksen meerdere waarden van de te verklaren reeks behoren.

De eenvoudigste gevallen, die wij hierbij zullen hebben te beschouwen, zijn die waarbij slechts *enkelvoudig verband* aanwezig is (§ 10). Deze zullen nu eerst worden behandeld. Dit kan gemakkelijk gebeuren aan de hand van vlakke grafische voorstellingen, waarbij langs de horizontale as de cijfers betreffende de oorzaakgrootheid, langs de vertikale de cijfers betreffende het gevolg worden afgezet. Laten we beginnen met daarbij de grootheden in hun absolute of natuurlijke maatstaven uit te drukken. Elk stel van 2 bijeenbehorende getallen wordt dan door een punt in de grafiek voorgesteld, waarvan die getallen de coördinaten zijn. Het feit van functionele samenhang tussen twee verschijnselen komt nu grafisch hierin tot uitdrukking, dat de verschillende punten op één lijn liggen.

Een eerste onderscheid dat nu te maken is, is dat tussen *rechtlijnig* en *kromlijnig* verband,

al naarmate bedoelde lijn recht of krom is. Veelal laat de nauwkeurigheid van practische waarnemingen niet toe om empirisch vast te stellen of de lijn waaromheen zich de punten groeperen een rechte dan wel een kromme is. Voorts laat zich een kromme lijn over een korte afstand gewoonlijk door een rechte benaderen. Deze beide omstandigheden zijn oorzaak, dat het voor de practijk van de hier bedoelde onderzoekingen zeer belangrijk is, het speciale geval van rechtlijnig verband — dat een eenvoudiger behandeling met formules toelaat — uitvoerig te bespreken.

§ 12. *Het rechtlijnig verband; de regressie-coëfficiënt*

Het *rechtlijnig verband* kan nog nader onderscheiden worden — wat voor sommige niet-mathematische lezers wel gewenst zal zijn — in drie soorten:

a. *gelijkheid* van de twee grootheden (d.w.z. de bijeenbehorende waarden der twee grootheden worden telken male door hetzelfde getal uitgedrukt): de grafische voorstelling is dan een lijn door de oorsprong (wanneer gelijke schalen worden gebezigd, onder 45° met de assen);

b. evenredigheid der twee grootheden (de ene grootheid is steeds een zeker aantal malen de andere grootheid): de grafische voorstelling

is eveneens een lijn door de oorsprong (bij gebruik van gelijke schalen nu echter onder een andere hoek dan 45° met de assen);

c. algemene lineaire samenhang: de lijn gaat nu niet door de oorsprong. De getallen, die nu de ene grootheid uitdrukken worden gevonden door een zeker aantal malen de tweede te nemen en daarbij nog een vast bedrag op te tellen (af te trekken). In formules uitgedrukt zijn de gevallen gekenmerkt door:

$$1)\quad X = Y$$
$$2)\quad X = bY$$
$$3)\quad X = bY + a$$

Anders uitgedrukt, kan men ook zeggen dat voor alle drie de gevallen de formule $X = bY + a$ gelden, doch dat voor:

$$1)\quad b = 1 \quad a = 0$$
$$2)\quad b \neq 1 \quad a = 0$$
$$3)\quad a \neq 0$$

geldt.

Opgemerkt kan nog worden, dat in geval 3) de opeenvolgende *verschillen* van de grootheden A en B weer wèl evenredig zijn en dus onder geval 2) of zelfs 1) behoren.

In sommige practische gevallen is het er slechts om te doen, te zien of er inderdaad lineair verband bestaat tussen de twee grootheden; en als dat dan aanwezig blijkt, vindt men empirisch, hoe groot b en a zijn, dus ook,

in welk der gevallen 1), 2) of 3) men verkeert. Dat is b.v. zo, wanneer men in de economie met een aanbodlijn te doen heeft. De redenering, ontleend aan de economie, levert ons daar slechts de verwachting dat er verband is, doch niet dat daarbij bovendien *b* en *a* een bepaalde waarde zouden moeten hebben (hoogstens, vooral ten aanzien van *b*, dat ze een bepaald teken zouden moeten hebben). In andere gevallen echter is het er juist om te doen, te vinden dat er gelijkheid bestaat tussen de twee reeksen. Zo b.v. bij het toekennen van cijfers aan leerlingen. Het is daar niet voldoende — om een goede beoordeling te hebben van een groep leerlingen — dat b.v. het hoofd van de lagere school en het toelatingsexamen de verschillende leerlingen in dezelfde volgorde plaatsen, doch het is gewenst dat ze ook minstens de scheiding tussen voldoende en onvoldoende op dezelfde plaats in de reeks leerlingen aanbrengen, en liefst gelijke cijfers geven. Het probleem is hier dus lichtelijk verschillend.

Bij de eerstgenoemde problemen heeft men er veel nut van, wanneer men in plaats van de absolute of natuurlijke maatstaf der grootheden gebruikt een maatstaf die zijn beginpunt heeft bij het gemiddelde der beschouwde grootheid. Men zegt dan, zoals reeds werd opgemerkt, dat de cijfers „afwijkingen van het gemiddelde" voorstellen. We zullen in het vervolg deze

afwijkingen steeds door *kleine letters*, in tegenstelling tot de voor de absolute cijfers gebruikte hoofdletters, weergeven. Voor deze afwijkingen van het gemiddelde bestaan er weer, zoals men gemakkelijk inziet, slechts de gevallen 1) en 2).

Voor bepaalde doeleinden, in hoofdstuk III nader te bespreken, gaat men nog verder en kiest men ook nog een andere *eenheid*. En wel kiest men dan de zgn. *standaarddeviatie* (zoals reeds opgemerkt, ook wel *standaardafwijking* of *middelbare afwijking* genoemd) tot eenheid. Zijn de afwijkingen van het gemiddelde resp. x_1, $x_2 \ldots . x_n$, waarbij dus n is het aantal waarnemingen, dan wordt de standaarddeviatie, die bijna steeds door σ_x wordt weergegeven, gedefiniëerd door:

$$\sigma_x^2 = \frac{1}{n}\left(x_1^2 + x_2^2 + \ldots . x_n^2\right)$$

of

$$n\sigma_x^2 = \sum_1^n{}^i x_i^2.$$

We zullen de grootheden, in deze eenheden, en vanaf haar gemiddelde gemeten, aanduiden door Griekse letters; zodat dus b.v. geldt:

$$\xi_1 = \frac{x_1}{\sigma_x}, \xi_2 = \frac{x_2}{\sigma_x}, \ldots . \xi_n = \frac{x_n}{\sigma_x}$$

of algemeen

$$\xi_i = \frac{x_i}{\sigma_x} \quad (i = 1 \ldots . n)$$

en we zeggen dat de ξ's „op standaarddeviatie staan". Anders uitgedrukt: de standaarddeviatie

van de ξ's zelf is 1. Tussen twee grootheden op standaarddeviatie, die lineair verband vertonen, kan niet anders dan *gelijkheid* bestaan, zoals men gemakkelijk narekent. Hier kan dus alleen geval 1) bestaan.

De coëfficiënt b, hierboven ingevoerd in de vergelijking die het verband tussen X en Y uitdrukte, noemt men de *regressiecoëfficiënt* [1]). Deze geeft dus aan met hoeveel van de gekozen eenheden X stijgt, wanneer Y met één eenheid stijgt. Kiest men als eenheden de resp. standaarddeviaties van X en Y, dan is de regressiecoëfficiënt noodzakelijkerwijze 1.

§ 13. *Kromlijnig verband; meervoudig verband*

Kromlijnig verband

Zoals reeds is opgemerkt, zijn er ook wel

[1]) De naam regressiecoëfficiënt is op merkwaardige wijze in de statistiek in gebruik gekomen. Met regressie werd door Galton aangeduid het verschijnsel dat de kinderen bepaalde kenmerken van hun ouders in 't algemeen ook vertonen doch — alweer in 't algemeen genomen — in mindere mate, zodat er dus een neiging tot een „terugkeer", een „regressie" naar het gemiddelde aanwezig is. De regressiecoëfficiënt gaf in zijn geval de verhouding weer waarmede die regressie zich voordoet. Voor het begrip van deze coëfficiënt doet het er echter natuurlijk niets toe of er een dergelijke „terugkeer" is; ook als er een steeds verdere verwijdering van het gemiddelde zou plaats hebben, zou een dergelijke coëfficiënt betekenis hebben. In de naam regressiecoëfficiënt zit dus reeds een bepaalde theorie van de erfelijkheid besloten.

gevallen waarin de nauwkeurigheid van meting groot genoeg is, om te kunnen vaststellen dat het verband tussen twee grootheden niet lineair, maar kromlijnig is. In dat geval is de grafische voorstelling dus een kromme lijn; deze kan zowel een mathematische als een empirische kromme zijn. Er is hier geen aanleiding om van „de" regressiecoëfficiënt te spreken. Veelal rekent men bij het nader beschouwen van deze soort samenhang ook met afwijkingen van het gemiddelde. De kromme lijn, die het verband voorstelt tussen de aldus gemeten grootheden, zal echter in 't algemeen niet (zoals bij de rechte lijn) door de oorsprong van het assenkruis gaan.

Meervoudig verband

Bij meervoudig verband zou men in principe van een meerdimensionale (in casu maximaal driedimensionale) grafische voorstelling gebruik kunnen maken, doch dat ontmoet veelal bezwaren; alleen voor demonstratiedoeleinden hebben dergelijke voorstellingen een grote waarde. Er zijn echter ook methoden, om een samenhang waarin meer dan twee grootheden een rol spelen in tweedimensionale (vlakke) grafieken af te beelden. Enkele voorbeelden zullen daarvan hier behandeld worden.

Een eerste voorbeeld is de wijze waarop men reeds in de (theoretische) economie gewend was om de vraagfunctie grafisch te bestuderen,

wanneer niet alleen de prijs, maar ook de een of andere verdere oorzaak van vraagveranderingen in de beschouwingen werd opgenomen. Om de gedachten te bepalen, nemen we aan dat die verdere oorzaak alleen het inkomen van de betr. groep mensen is. De gevraagde hoeveelheid goederen X hangt dan dus af van de prijs Y en het inkomen Z; in algemeen wiskundige symbolen: $X = f(Y, Z)$

Zouden we aannemen dat het verband ook thans lineair is, zowel in Y als in Z, dan zou deze algemene formule de vorm aannemen:

$$X = b_{12}Y + b_{13}Z + a$$

Voor het gemak van het spreken zullen we verder deze formule gebruiken. De lezer zal bemerken dat onze opmerkingen echter evenzeer gelden voor de algemene formule.

Wil men nu de aangegeven samenhang systematisch gaan bestuderen, dan is het geschikt om aan Z eerst een vaste waarde Z_0 te geven en na te gaan welk verband er nu nog bestaat tussen X en Y. Dit wordt gegeven door

$$X = b_{12}Y + (b_{13}Z_0 + a)$$

waarbij we de laatste twee termen tussen haakjes hebben gezet om aan te duiden dat deze te zamen nu een constante term vormen. Het verband tussen X en Y, bij gegeven Z_0, wordt dan nu door een lijn voorgesteld in fig. 3 aangegeven met $X = f(Y, Z_0)$. In het bijzondere geval dat we hier beschouwen is dat een rechte lijn, die

van de Y-as afsnijdt een segment $b_{13}Z_0 + a$.

Geven we nu Z een andere waarde Z_1, dan krijgen we voor de gevallen die daaronder behoren een verband

$$X = b_{12}Y + (b_{13}Z_1 + a)$$

of algemeen $X = f(Y, Z_1)$ in de figuur aangegeven door de aldus aangeduide lijn. Deze is

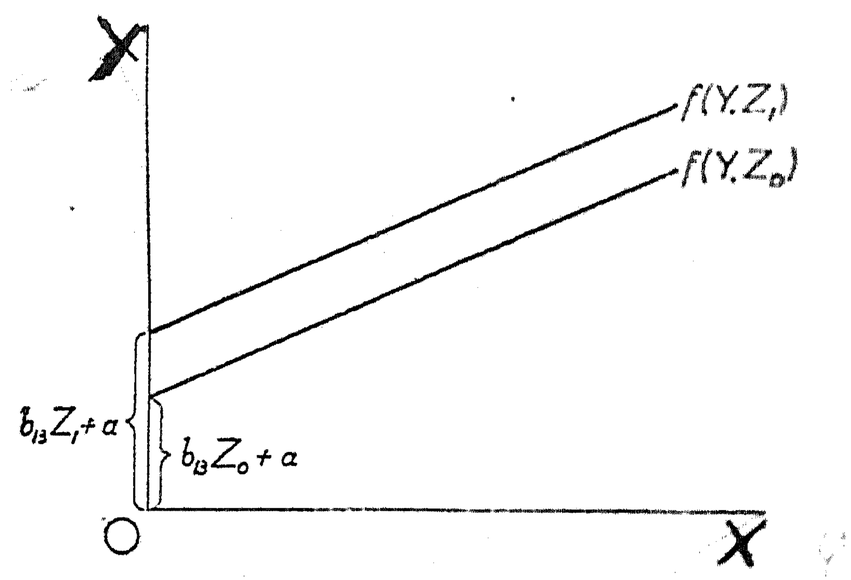

Fig. 3. Grafische voorstelling van een verband tussen drie grootheden, $X = f(Y, Z)$, voor twee waarden van Z, nl. Z_0 en Z_1.

in ons speciale geval een rechte, die een bedrag $b_{13}Z_1 + a$ van de Y-as afsnijdt. We kunnen zo een systeem van lijnen verkrijgen dat als geheel een denkbeeld geeft van het meervoudig verband [1]).

[1]) Dit systeem kan beschouwd worden als de projectie op het XOY-vlak van een systeem van „lijnen van constante Z" in de driedimensionale voorstelling getrokken op het oppervlak $X = f(Y, Z)$.

De verschuiving van de „vraaglijn bij constante Z" [1]) bij verandering van de waarde van Z is een in de literatuur vaak besproken verschijnsel, onder de Engelse naam „shifting" bekend. In ons eenvoudige geval is deze shifting evenredig met de verandering in Z.

2) Een tweede methode, meer bekend uit de statistische onderzoekingen van Amerikaanse landbouwstatistici (speciaal Ezekiel) maakt gebruik van twee grafische voorstellingen (zie fig. 4). In de bovenste worden met elkaar vergeleken de grootheden X en Y (waarbij, om aan te sluiten bij de genoemde schrijvers, nu Y langs de horizontale as is afgezet). De punten P_1, P_2, enz., die elk als coördinaten hebben bijeenbehorende waarden van X en Y (P_1 heeft de coördinaten X_1 en Y_1, P_2 : X_2 en Y_2, enz.), liggen nu niet op één lijn, want al naarmate de Z_1, Z_2, enz. verschillend zijn, zijn er afwijkingen van de lijn

$$X' = b_{12}Y + a$$

van verschillende omvang aanwezig. Anders gezegd, er is een onderscheid tussen X en deze X', dat afhankelijk is van Z, want

$$X - X' = b_{13}Z$$

[1]) Zoals men weet, wordt in de economie de vraaglijn ook onder deze voorwaarde gedefiniëerd en is de toevoeging dus eigenlijk overbodig: men behoeft slechts de waarde van Z te vermelden.

Om die reden worden nu de waarden $X—X'$, die ook wel de *residuen* genoemd worden, in de onderste grafiek als ordinaat vergeleken met de

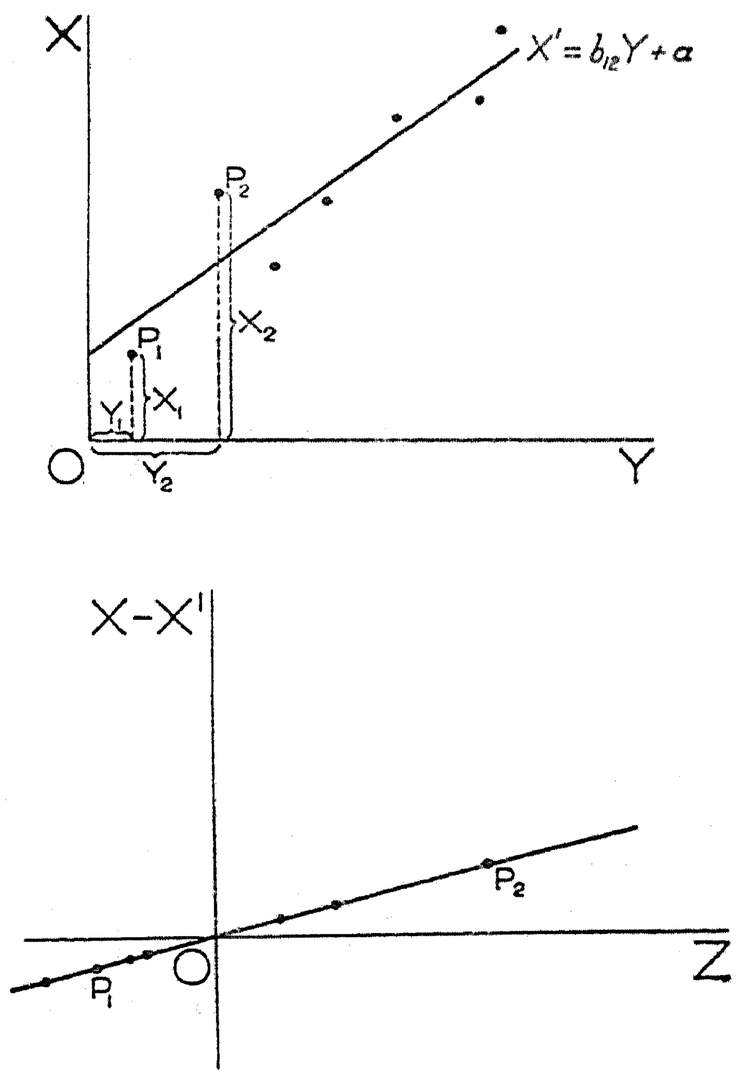

Fig. 4. Methode van Ezekiel voor het grafisch voorstellen van een functie van twee veranderlijken $X(Y,Z)$.

abscissen Z. Tussen deze residuen en de Z-coördinaten bestaat nu een functioneel verband.

Veelal wordt deze methode niet voor de

absolute waarden X, Y en Z toegepast, doch voor de afwijkingen van het gemiddelde x, y en z.

Het spreekt van zelf dat men de hier aangeroerde vraagstukken ook algebraïsch kan behandelen. Belangwekkende voorbeelden daarvan zijn gegeven voor pedagogische vraagstukken, b.v. bij behandeling volgens de methode van Spearman [1]). Men neemt daarbij aan dat de rapportcijfers x_i voor verschillende vakken ($i = 1$, 2 en 3 b.v.) zich elk laten weergeven door een lineaire functie van een algemene grootheid g en een grootheid s_i die voor elk vak verschillend is; aldus:

$$x_1 = c_1\, g + c_1'\, s_1 \qquad (13.1)$$
$$x_2 = c_2\, g + c_2'\, s_2 \qquad (13.2)$$
$$x_3 = c_3\, g + c_3'\, s_3 \qquad (13.3)$$

Daarbij stelt dan g voor een soort van algemene intelligentie, s_1 een speciale aanleg voor vak No. 1, s_2 voor vak 2, enz. De g is voor elke leerling verschillend, maar voor alle vakken dezelfde; de coëfficiënten zijn daarentegen voor alle leerlingen dezelfde. Men stelt nu de verdere voorwaarden dat de grootheden s_1, s_2, s_3

[1]) Voor een zeer duidelijke uiteenzetting zie men b.v. Dr. B. M. van Dalfsen, De correlatie tusschen Meetkunde en Fransch, Paedagogische Studiën X (1929) blz. 231 en De samenhang der Rapportcijfers voor de verschillende leervakken eener H. B. S., Paed. Studiën XI (1930), blz. 230, speciaal blz. 247—250.

en g onderling onafhankelijk zijn, hetgeen wordt vertaald door de eis dat de correlatiecoëfficiënten tussen deze grootheden onderling nul zijn (vgl. voor de betekenis van deze term § 21 en volgende), hetgeen aanleiding is tot de vgln:

$$\Sigma s_1 s_2 = 0 \quad \Sigma s_1 s_3 = 0 \quad \Sigma s_2 s_3 = 0$$
$$\Sigma s_1 g = 0 \quad \Sigma s_2 g = 0 \quad \Sigma s_3 g = 0$$

waarbij de sommatie plaats heeft over alle leerlingen. Bovendien wordt de eis gesteld dat de standaarddeviaties van g, s_1, s_2 en s_3 gelijk aan de eenheid zijn. Door deze eisen laten zich de coëfficiënten c_1 t/m c_3' berekenen.

Ofschoon bij de desbetreffende formuleringen een ruim gebruik gemaakt wordt van de begrippen der correlatierekening, is de probleemstelling toch essentiëel anders dan de problemen die wij te dien opzichte later zullen behandelen. Men veronderstelt nl. dat de vgln. (13.1), (13.2) en (13.3) exact gelden voor elken leerling en elk vak, terwijl bij de vraagstukken die later besproken worden, aangenomen zal worden dat er een verschil bestaat tussen de rechter- en de linkerzijde dezer vergelijkingen. De „restloze" verklaring van de grootheden x_1, x_2 en x_3 in het bovenstaande geval is alleen mogelijk doordat men grootheden s_1, s_2 en s_3 introduceert in even groten getale als de grootheden x_1, x_2 en x_3; in de later te bespreken problemen wordt aangenomen dat de grootheden s reeds a priori gegeven zijn.

De hier gevolgde behandelingswijze komt sterk overeen met de in § 33 te bespreken probleemstelling van Frisch.

Opgemerkt moge nog worden, dat de in de vergelijkingen (13.1), (13.2) en (13.3) gegeven opzet slechts mogelijk is bij juist *drie* vakken. Als men meer vakken beschouwt, moeten andere wegen ingeslagen worden; één daarvan is, dat naast de algemene factor g nog „groepsfactoren" worden ingevoerd. In het reeds genoemde artikel van den heer Van Dalfsen worden nog ingevoerd een factor x, aangeduid als wiskunde-aanleg, die tot uiting komt in de cijfers voor algebra, meetkunde en tekenen; een factor y, genoemd talen-aanleg, tot uiting komende bij Nederlands, Frans en Duits, en een factor z, die de schrijver aantrof bij de vakken algebra en vreemde talen en die hij accuratesse noemde.

§ 14. *Multicollineariteit*

Bij gevallen waarin men met meer dan drie veranderlijken werkt is uiteraard een grafische voorstelling op de gewone wijze onmogelijk. En bij een groeiend aantal veranderlijken worden ook de andere grafische hulpmiddelen steeds ingewikkelder te hanteren. Veelal neemt men dan zijn toevlucht tot de zuiver algebraïsche methode van berekenen, welke in § 19, bij de correlatierekening, zal worden behandeld. Deze stelt ons, zeer in 't algemeen gesproken, in staat

om de regressievergelijkingen te vinden ook bij een willekeurig aantal veranderlijken.

Frisch (Statistical Confluence Analysis by Means of Complete Regression Systems) [1]) wijst op de gevaren die deze methode meebrengt, wanneer zgn. *multicollineariteit* optreedt, d. w. z. dat er tussen de veranderlijken die men als „verklarende reeksen" heeft aangenomen, betrekkingen (en wel speciaal lineaire) bestaan. Enige der coëfficiënten kunnen daardoor volkomen onbepaald worden. In de realiteit, waar altijd afwijkingen van de volkomen functionele samenhang voorkomen, verbergen deze afwijkingen dan vaak de onbepaaldheid doordat men b.v. niet een uitkomst $\frac{0}{0}$ verkrijgt, doch een uitkomst $\frac{\text{kleine afwijking}}{\text{andere kleine afwijking}}$, waardoor een schijnbare bepaaldheid ontstaat die echter volkomen misleidend is.

In de terminologie van de hoger-dimensionale meetkunde kan men dit aldus uitdrukken dat de waarnemingen in plaats van te liggen in één $(n-1)$-dimensionale ruimte (bij $n-1$ verklarende reeksen en een n-dimensionale ruimte om het verband voor te stellen), doch in een ruimte van lagere dimensie. Het eenvoudigst denkbare geval is dat van $n = 3$. In 't algemeen

[1]) Universitetets Økonomiske Institutt, Oslo 1934, Publ. Nr. 5.

vgl collegeleer matrix

gesproken zal een (lineair) verband tussen 2 verklarende en een te verklaren reeks worden voorgesteld door de punten van een plat vlak; wanneer echter tussen de twee verklarende reeksen nog een verband bestaat, degenereert dat platte vlak tot een rechte lijn. Frisch geeft in zijn genoemd werk methoden aan om hier de juiste conclusie te trekken. Deze methoden kunnen hier echter niet nader worden besproken.