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

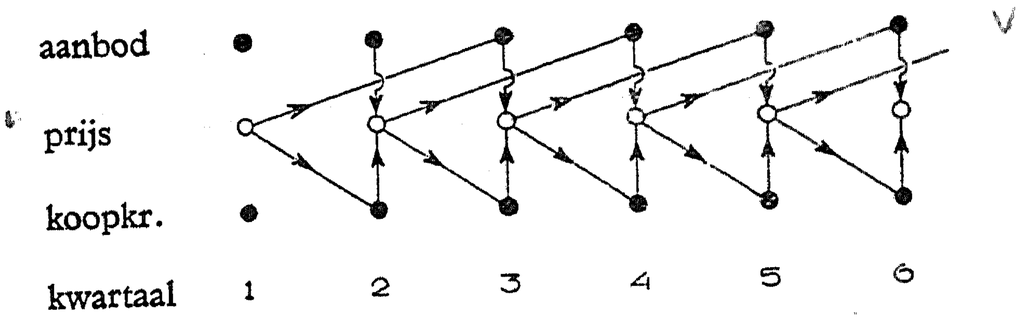

Fig 1. Logische structuur van een object. Voorbeeld: een vereenvoudigd model van het conjunctuurmechanisme in de economie.

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

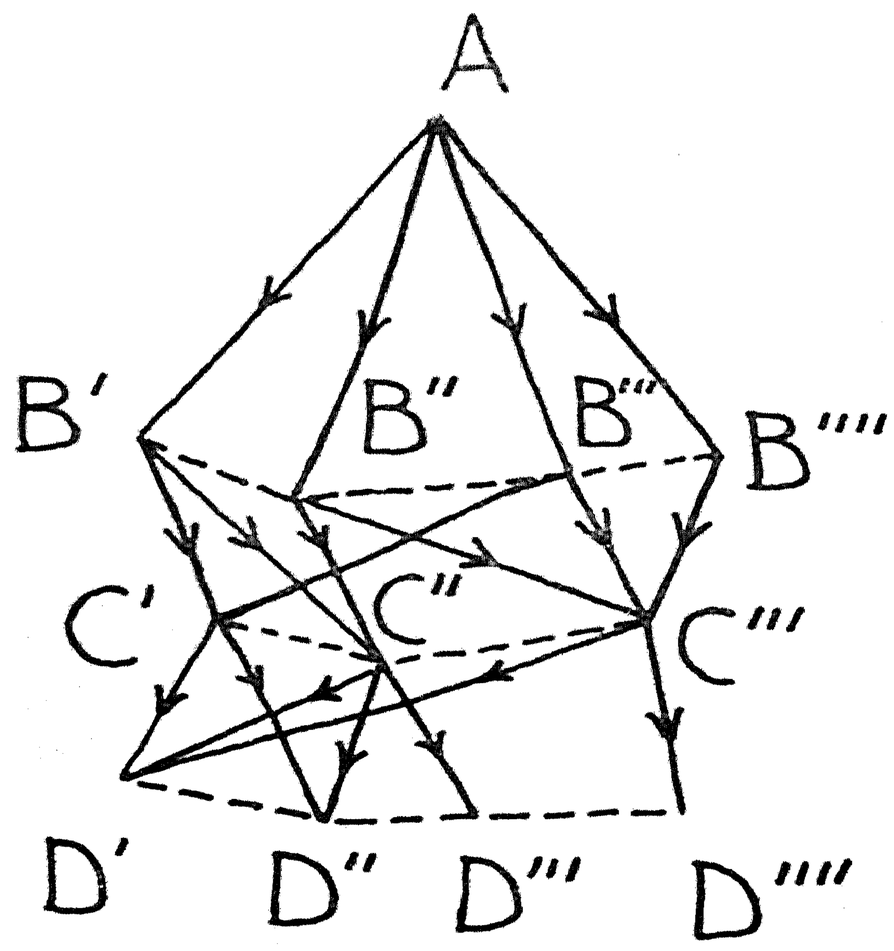

Fig. 2. Logische structuur van een object. Voorbeeld: de intelligentie. Eén bepaald intelligentieniveau *B*, *C* of *D* bevat nog verschillende toestanden. Voor het niveau is het geenszins onverschillig of de aangegeven verbindingen der niveau's alle of slechts gedeeltelijk functioneren. (Ontleend aan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

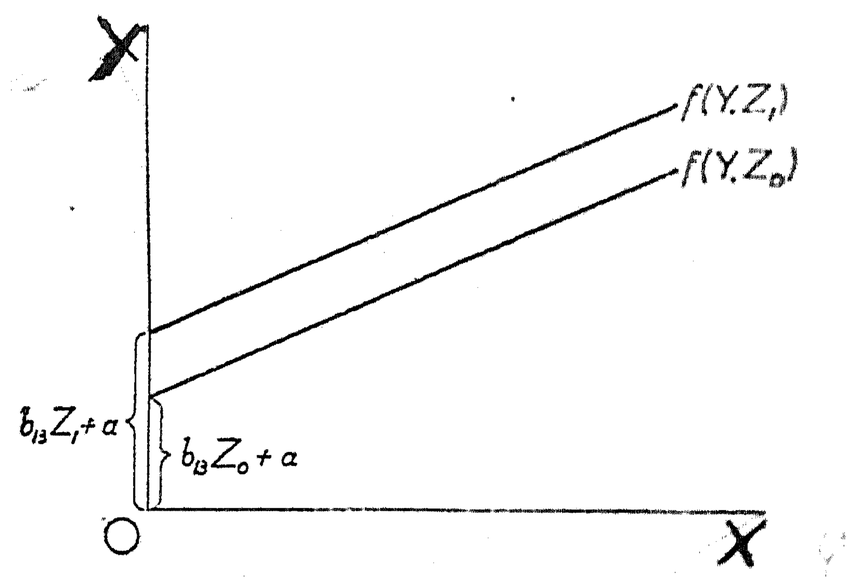

Fig. 3. Grafische voorstelling van een verband tussen drie grootheden, $X = f(Y, Z)$, voor twee waarden van Z, nl. Z_0 en Z_1.

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

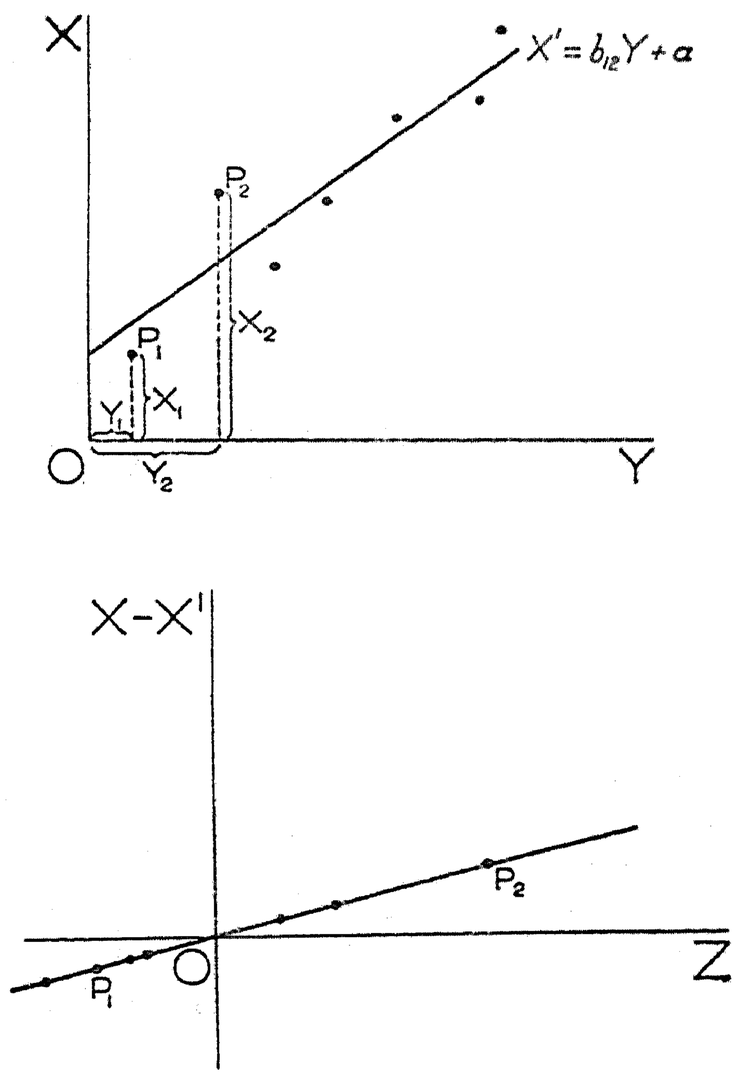

Fig. 4. Methode van Ezekiel voor het grafisch voorstellen van een functie van twee veranderlijken $X(Y,Z)$.

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.