VAN DE REDACTIE

Het doet ons genoegen dat wij bij de lezers een artikelenreeks van de hand van de heer H.A.A. de Melverda kunnen introduceren over de Correlatie-rekening als hulpmiddel bij bedrijfseconomische analyses.

De Correlatie-rekening is een onderwerp van wiskundige aard en schrikt daarom vele niet-mathematici af. Ze kan echter een belangrijk hulpmiddel zijn bij bedrijfseconomische analysen; daarom acht de Redactie het nodig, dat zij meer dan tot nu toe het geval is door de bedrijfseconomen en de accountants wordt toegepast. Dit is mogelijk, omdat de Correlatie-rekening in waarheid voor de studerenden voor het accountantsexamen en voor de accountants geen bijzondere wiskundige aanleg en geen kennis vereist boven hetgeen zij uit anderen hoofde voor hun studie en voor hun vak moeten weten; nodig is slechts, dat men de traditionele afwerende instelling ten opzichte van wiskundige methoden overwint. De Redactie wil daarom een poging doen om de belangstelling voor de Correlatie-rekening te wekken en meent een doelmatig middel daartoe gevonden te hebben in de hierbij aangeboden artikelen.

Hoewel de omvang van de reeks — een 5 tal — publicatie in boekvorm ongetwijfeld rechtvaardigt, hebben wij aan publicatie in het Maandblad de voorkeur gegeven, omdat wij het vertrouwen koesteren, dat op deze wijze op grotere schaal kennis van deze beschouwingen zal worden genomen. Wij hebben de hoop, dat de beschouwingen van de heer de Melverda aldus gemakkelijker de barrière van de afwerende houding zullen doen overwinnen en een groter aantal lezers zullen vinden, die de moeite zullen nemen om daarvan op gezette wijze kennis te nemen.

CORRELATIEREKENING ALS HULPMIDDEL BIJ BEDRIJFSECONOMISCHE ANALYSES

door H. A. A. de Melverda

I

Inleiding

Er bestaat geen onafhankelijk verschijnsel. Men kan ervan verzekerd zijn, dat men zich vergist, als men meent, dat er een onafhankelijk verschijnsel bestaat. Als men zulk een verschijnsel meent te ontmoeten en men bestudeert het grondig en diepgaand, dan zal men niets vinden, dat aanspraak kan maken op onafhankelijkheid. In de verschijnselenwereld is elk ding (feit en omstandigheid) afhankelijk van andere dingen, en deze op hun beurt van weer andere dingen, zodat uiteindelijk elk ding afhankelijk is van alle overige dingen. Dit toont aan, dat alles in werkelijkheid Eén is. Slechts de Eénheid is onafhankelijk.

Het zintuiglijk leven neemt slechts dingen waar en is gevangen in de illusie, dat deze dingen onafhankelijk zijn. Het denken kan de afhankelijkheden van de dingen nooit volmaakt zien, want dan zou het de Eénheid moeten zien. Dat kan slechts het weten. In de denkende mens leeft echter de vurige wens, om door denken tot weten te komen, hoe onbegonnen deze taak ook is. De oogst van de inspanning, die men zich bij het kwijten van deze taak heeft gegeven, noemt men wetenschap. Alle wetenschap heeft daarom tot taak: het zoeken van afhankelijkheden tussen en eenheid in de verschijnselen.

De verschijnselen hebben, tezamen in groepen genomen, een quantitatief en daarom statistisch karakter. Daarom kan het zoeken naar afhankelijkheden tussen groepen van verschijnselen ten dele plaats vinden door slechts te letten op de samenhang, de correlatie (co-relatie), die er bestaat tussen de quantiteiten dier groepen.

Voor het begrip quantiteit is meetbaarheid verondersteld. In dit opzicht geeft het economisch leven een schat van gegevens, want hier werkt men bij uitstek met quantiteiten, terwijl door de rol van het geld de niet direct vergelijkbare of meetbare grootheden herleid worden tot meetbare en vergelijkbare quantiteiten. Het behoeft ons daarom niet te verwonderen, dat juist in de economie de afhankelijkheden van de verschijnselen voornamelijk worden gezocht als quantitatieve afhankelijkheden, waardoor economie en statistiek nauw aan elkaar verbonden worden.

In de evolutie der wetenschappen ziet men, dat men steeds eerst denkt aan een zeer speciale vorm van afhankelijkheid der verschijnselen, n.l. aan het *functioneel verband*, d.w.z. dat de afhankelijkheid nauwkeurig een bepaalde wet volgt, waarvan dus niet in het minst wordt afgeweken. Quantitatief gezien kan zulk een wet uitgedrukt worden door een algebraïsche formule, waarbij na wijziging van de waarde x van het ene verschijnsel nauwkeurig berekend kan worden, wat de waarde y *moet* zijn voor het andere verschijnsel. Doch zowel practische observatie als theoretische analyse (b.v. de natuurkunde van Planck, Einstein, Heisenberg en de Broglie) leerden inzien, dat functionele verbanden slechts bestaan als toevalligheid en dat men al blij mag zijn met de kennis, dat de waarde x van het ene verschijnsel *kan* samengaan met de waarde y van het andere verschijnsel, terwijl men heel blij mag zijn, als men daarbij nog de mate van waarschijnlijkheid kan aangeven. In de practijk blijkt er daarom slechts een *statistisch verband* tussen de verschijnselen te bestaan.

De oorzaak hiervan kan o.a. verklaard worden uit de omstandigheid, dat bepaalde verschijnselen verschillende verschijnselen kunnen oproepen; zo kan een knikker, die op een horizontaal ijzeren staafje valt, de ene keer rechts, maar de andere keer links ervan vallen, zonder dat men van te voren kan voorspellen, welke van de twee mogelijkheden zal optreden; valt de knikker nu op nog verschillende lager gelegen ijzeren staafjes, dan kan men slechts de graad van waarschijnlijkheid aangeven, dat de knikker op een bepaalde plaats op de bodem zal blijven liggen; doch de werkelijkheid kan heel anders zijn.

Een andere oorzaak ligt in de omstandigheid, dat de causaliteitsreeks niet lineair is, doch een kringloop vertoont, m.a.w. de gevolgen keren terug op hun oorzaken.

Vervolgens ligt een belangrijke oorzaak in de isolerende methode der wetenschap. Men zoekt de afhankelijkheid tussen twee of meer verschijnselen en doet net alsof de andere afhankelijkheden niet meer bestaan. Doch de werkelijkheid is anders, waardoor tussen verwachting en werkelijkheid verschillen optreden, die worden beschouwd als „storende" invloeden.

De grondoorzaak ligt in de beperktheid van ons zintuiglijk waarnemingsvermogen, en in de eigenschap van het denken (n.l. in tegenstellingen), waardoor ons de eenheid van al het gebeuren ontgaat en waardoor het zoeken naar weten slechts is als het tasten van een blinde in een onbekende wereld.

**
*

De methode, die men volgt, om de statistische verbanden weer te geven, kan bestaan in het benaderen ervan door functionele verbanden en vervolgens aan te geven, wat de mate van nauwkeurigheid is, die daarmede bereikt wordt of die daaraan kan worden gehecht. Deze methode, welke de naam draagt van *correlatierekening*, maakt deel uit van de mathematische statistiek.

De meeste leerboeken, die de correlatierekening behandelen, hebben twee nadelen: òf ze zijn niet begrijpelijk voor hen, die slechts een beperkte kennis van de wiskunde hebben, òf ze geven geen verklaring van de methode en vergenoegen zich ermede alleen de formules aan de hand te doen, waarmede gewerkt moet worden. Beide omstandigheden zijn niet bepaald bevorderlijk om deze methode van afhankelijkheidsonderzoek populair te maken.

Het doel van deze beperkte artikelenreeks is de aandacht te vestigen van de bedrijfseconoom op het grote nut, dat de correlatierekening ook voor zijn vak kan hebben, vooral voor de practische analyse. Daarbij zullen wij pogen in een zeer kort bestek een uiteenzetting te geven van deze methode, zodanig, dat beide bovengenoemde bezwaren kunnen worden ondervangen, zodat men na bestudering van de serie deze methode volledig zal hebben begrepen, en tevens in de practijk kan toepassen.

Wij zullen ons onthouden van bronvermeldingen en uiteraard geen aanspraak maken op oorspronkelijkheid. Wanneer de correlatierekening meer bekend en begrepen wordt, achten wij het doel van deze serie vervuld. Wij zullen daarbij van de lezer niet meer mathematische kennis vragen dan de elementaire lagere algebra. Hierdoor kunnen sommige bewijzen minder scherp of minder puntig worden afgeleid, doch dit geschiedt o.i. zonder veel bezwaar; het inzicht in de gedachtengang en de methode zal meestal voldoende overtuiging kunnen geven van de juistheid der berekeningen.

Enkelvoudige lineaire correlatie

Men spreekt van enkelvoudige correlatie, indien men de afhankelijkheid tussen slechts twee grootheden op het oog heeft, m.a.w. indien men wil onderzoeken wat de invloed is die de ene grootheid (Y) ondergaat als de andere grootheid (X) verandert. Het meest eenvoudige geval, dat men zich vervolgens kan indenken, is, dat de afhankelijkheid lineair is, d.w.z. de toename van Y evenredig is met de toename van X. Algebraïsch, d.w.z. als functioneel verband beschouwd, wordt het enkelvoudige lineaire verband aangegeven door de vergelijking $Y' = a + bX$. (Ter onderscheiding zullen wij de waargenomen waarden, d.w.z. die behorend bij het statistisch verband, aanduiden met Y en de benaderende waarden, d.w.z. behorende bij het functioneel verband, aanduiden met Y'; in beide verbanden nemen wij de X als zgn. onafhankelijk variabele en daarvoor dus steeds de waargenomen waarden).

Een voorbeeld van een dergelijk verband kan men vinden in de analyse van de totale kosten welke per maand in een bedrijf worden gemaakt. Het verband, dat men dan meestentijds veronderstelt te bestaan is als volgt:

Maandkosten (Y') = Vaste kosten (a) + Productiehoeveelheid (X) $\times$ Variabele kosten (b). Bij de practische analyse rijst dan de vraag: bestaat dit verband wel precies zo? Misschien zijn de variabele kosten progressief, misschien wel degressief, misschien spelen andere factoren nog een belangrijke rol.

Teneinde de methode der correlatierekening te demonstreren, zullen wij een gefingeerd practisch geval van enkelvoudige lineaire correlatie behandelen. Wij veronderstellen dan, dat wij over de maandcijfers beschikken van een bedrijf, waarin slechts één bepaald uniform product wordt vervaardigd. Wij willen vervolgens het verband analyseren, dat bestaat tussen de maandkosten en de productiehoeveelheid.

Gesteld dat wij de volgende *12* paren bij elkaar behorende getallen in duizendtallen gedurende een vol jaar hebben gevonden:

	Productiehoeveelheid	*Maandkosten*
	X	*Y*
	109	*202*
	121	*242*
	129	*261*
	138	*240*
	143	*267*
	149	*242*
	152	*254*
	160	*282*
	163	*270*
	170	*296*
	178	*280*
	188	*308*
Totaal	*1800*	*3144*
Gemiddeld per maand	*150*	*262*

Aangezien wij het verband tussen maandkosten en productiehoeveelheid onderzoeken, kan ons de tijdsorde der getallenparen onverschillig laten; bovenstaande getallenparen zijn daarom gerangschikt naar opklimming van de productiehoeveelheid. Gemakshalve zijn de getallen zo gekozen, dat het gemiddelde per maand een rond getal is. Voorts zal het wel geen toelichting behoeven, dat de kosten zo goed mogelijk per maand gelocaliseerd moeten zijn, waarbij gedacht is aan historische kosten.

Bezien wij bovenstaande getallenparen, dan treedt wel duidelijk in het licht, dat de maandkosten stijgen bij toenemende productie, doch dat een en ander, blijkens de tweede kolom, niet bepaald regelmatig plaats vindt. Teneinde het verband duidelijker zichtbaar te maken, kan een grafische voorstelling zeer goede diensten bewijzen. Elk getallenpaar correspondeert grafisch met een punt. Zodoende vinden wij een *correlatie-diagram* (scatter-diagram): zie figuur 1.

De eerste taak van de correlatierekening is nu, om een rechte lijn te bepalen, welke zodanig tussen de aangegeven punten loopt, dat deze punten zo dicht mogelijk om de lijn liggen. Een zodanige lijn wordt *regressie-lijn* genoemd.

Het is nu aanstonds duidelijk, dat het vinden van deze lijn neerkomt op het vinden van twee onbekenden, n.l. de vaste kosten per maand *(a)* en de variabele kosten per eenheid *(b)*. Grafisch bezien komt dit neer op het bepalen van het punt *P* en de helling van de regressielijn.

***

Eenvoudiger wordt het vraagstuk, als men reeds de plaats weet van

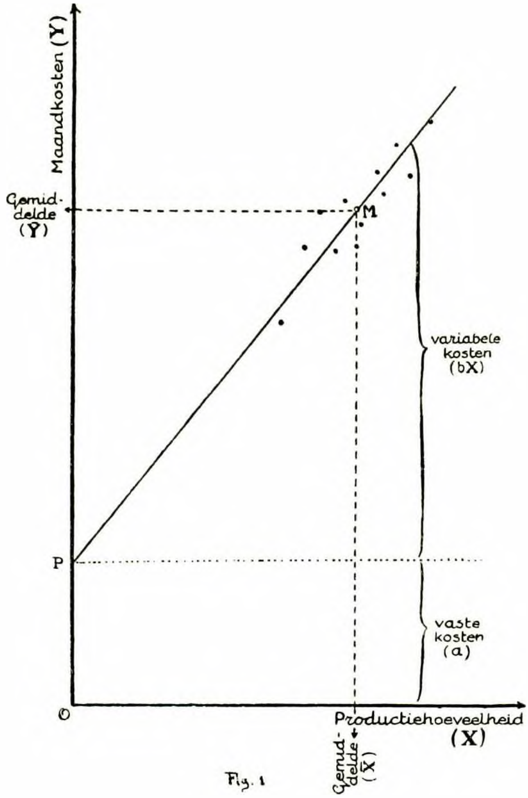

Fig. 1

één punt op de te bepalen regressielijn. Immers het vraagstuk wordt dan gereduceerd tot het bepalen van de helling van de lijn (dus de variabele kosten per eenheid). Deze vereenvoudiging blijkt inderdaad mogelijk: het trefpunt van de maandgemiddelden d.i. het punt M, moet zulk een punt zijn. Hoewel voor het exacte bewijs een weinig hogere wiskunde [1]) noodzakelijk zou zijn, zal de juistheid van deze stelling op zichzelf wel zeer plausibel zijn. Immers in de maandgemiddelden zijn alle afwijkingen ten opzichte van de regressielijn naar boven zowel als naar beneden gemiddeld en daardoor tot nul gereduceerd, hetgeen betekent, dat het punt M op de gezochte regressielijn moet liggen. (De gemiddelde waarden worden algebraïsch aangegeven met een streep boven het symbool; dus $\bar{X} = 150$ en $\bar{Y} = 262$).

Van deze omstandigheid kunnen wij grafisch goed gebruik maken door het punt M te kiezen als oorsprong van de grafiek. Daarbij hebben wij nog het bijkomstige voordeel, dat de waarden van de variabelen niet meer behoeven te worden afgepast vanaf de oorsprong, waardoor veel ruimte op de grafiek verloren gaat, zoals fig. 1 deed zien. De nieuwe grafiek komt er nu als volgt uit te zien: zie figuur 2.

---

[1]) Volgens de methode der kleinste quadraten moet $\sum (Y - a - bX)^2$ maximaal zijn, waaruit als eerste normaalvergelijking volgt

$$\sum Y - Na - b\sum X = 0, \text{ of } \bar{Y} = a + b\bar{X},$$

d.w.z. het punt M $(\bar{X}, \bar{Y})$ moet liggen op de lijn $Y' = a + bX$.

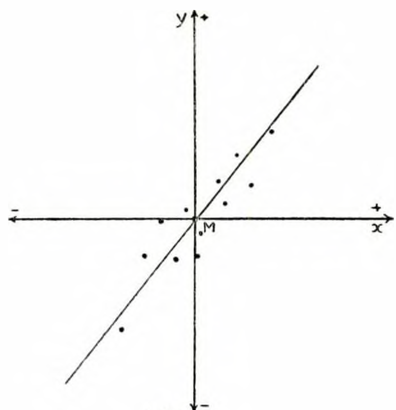

Fig. 2

Duidelijk is nu te zien, dat het nu nog slechts aankomt op de bepaling van de helling van de regressielijn. Inmiddels zal het tevens duidelijk zijn geworden, dat de getallen, waarmede wij nu werken, thans andere zijn: in plaats van de *oorspronkelijke getallen* (crude data) werken wij nu met de *afwijkingen van het gemiddelde,* welke wij vinden door de oorspronkelijke getallen te verminderen met de gemiddelde waarde. De aldus gevonden getallen zullen wij aanduiden met x en y ter onderscheiding van de oorspronkelijke getallen, die wij met hoofdletters hebben aangeduid.

	X	x	Y	y
	109	− 41	202	− 60
	121	− 29	242	− 20
	129	− 21	261	− 1
	138	− 12	240	− 22
	143	− 7	267	+ 5
	149	− 1	242	− 20
	152	+ 2	254	− 8
	160	+ 10	282	+ 20
	163	+ 13	270	+ 8
	170	+ 20	296	+ 34
	178	+ 28	280	+ 18
	188	+ 38	308	+ 46
Totaal:	1800	0	3144	0
	$\overline{X} = 150$		$\overline{Y} = 262$	

Dat de som der afwijkingen van het gemiddelde steeds gelijk aan nul moet zijn is duidelijk, want men heeft *12* × het gemiddelde afgetrokken, hetgeen juist gelijk is aan het totaal der *12* waarden.

In plaats van de oude vergelijking komt nu de vergelijking $y' = bx$ [2]). Het vraagstuk is derhalve de waarde van b hierin te bepalen. Immers, daar $\overline{Y} = a + b\overline{X}$ is, is $a = \overline{Y} - b\overline{X}$; of in ons voorbeeld: $a = 262 - b \times 150$. Hiermede zullen wij ons in het derde artikel bezig houden.

Eerst is het gewenst aan enige andere kwesties aandacht te schenken.

**
*

1. Op het eerste gezicht lijkt het nogal komisch, om zoveel drukte te maken om een lijn te trekken, die zo goed mogelijk tussen de gegeven punten doorloopt. Met een goed gevoel voor verhoudingen zal men deze lijn wel direct uit de hand kunnen schetsen en het zal niet zo heel erg zijn, of men dusdoende een beetje verschil krijgt met de theoretisch juiste lijn: het is toch maar een benadering van de werkelijkheid. Dit is ongetwijfeld juist en vermoedelijk zou de gehele correlatierekening niet uitgevonden zijn of althans niet in de practijk gebracht zijn, als alle te onderzoeken verbanden zo eenvoudig waren als wij in ons voorbeeld stelden. De practijk is dikwijls heel wat ingewikkelder.

Wanneer b.v. in ons voorbeeld twee invloeden inwerken op de maandkosten, dan is de grafische voorstelling van drie-dimensionaal karakter en het trekken van een „lijn" (thans een vlak) met de hand is al practisch ondoenlijk; werken er echter drie invloeden op de maandkosten, dan is er zelfs niet eens meer een grafische voorstelling van te maken, want het verband is dan vier-dimensionaal geworden. En toch, al kunnen wij ons geen aanschouwelijke voorstelling meer maken van dergelijke verbanden, de realiteit geeft ze ons dikwijls te zien.

Men denke slechts aan een eenvoudige vrachtrijder, wiens kosten ten dele afhankelijk zijn van het aantal ritten, ten dele van het aantal vervoerde tonnen en ten dele van het aantal afgelegde kilometers. Of men denke slechts aan een bedrijf, dat drie verschillende producten voortbrengt. Een statisticus mag niet bang zijn voor meerdimensionale verbanden, waarvan geen aanschouwelijk beeld is te maken, want dergelijke verbanden zijn toch zeer reëel. Het is duidelijk, dat voor de analyse van dergelijke verbanden een methode, gebaseerd op een bepaalde rekentechniek, noodzakelijk is.

Behalve de omstandigheid, dat het verband meervoudig (multipel) kan zijn, kan ook het verband niet-lineair zijn. Dit schept ook de noodzaak voor een bepaalde rekentechniek.

2. Vervolgens moeten wij wijzen op een verschil in methode tussen de splitsing van de kosten in vaste en variabele kosten volgens de boekhoudkundige weg en volgens de statistische weg. De boekhoudkundige weg splitst de kosten in soorten en bepaalt (?) geleid door een gevoel voor evenredigheden de kosten, die vast zullen zijn en de kosten, die variabel zullen zijn. Waar de boekhoudkundige er eigenlijk maar een slag naar slaat, gaat de statisticus methodisch te werk. Waar de boekhoudkundige eigenlijk slechts rekening houdt met proportionele afhankelijkheden, kijkt de statisticus ook naar andere (kromlijnige) afhankelijkheden.

Bovendien het is niet altijd juist alle kosten uit elkaar te plukken en vervolgens te sorteren over twee categorieën: vast en variabel. Verschillende kostensoorten zijn gedeeltelijk vast en gedeeltelijk variabel. Wij doelen hier niet alleen op het intermitterend variabel karakter van vele vaste kosten, maar ook op het intermitterend vast karakter van vele variabele kosten. Men denke slechts als voorbeeld aan de brandstofkosten

---

[2]) Hoewel dit eigenlijk niet eens meer nodig is, kan men de juistheid van deze vergelijking verifiëren als volgt. Als M een punt is van de gezochte regressielijn, dan is $\overline{Y} = a + b\overline{X}$, waaruit volgt:

$$y' = Y' - \overline{Y} = (a + bX) - (a + b\overline{X}) = b(X - \overline{X}) = bx.$$

bij een stoomketel; ook als de productie even stilstaat, moet de ketel op stoom gehouden worden; pas als de productie langer stilstaat zal men het vuur doven. Of men denke aan de reparatiekosten, die zeer dikwijls gedeeltelijk vast en gedeeltelijk variabel blijken.

Voor sommige kostensoorten neemt men a priori aan, dat ze geheel variabel zijn, zoals de directe lonen en de directe materialen. Als dit zo was, zou de grafiek, waarin de directe kosten (of eenheden) worden vergeleken met de productiehoeveelheid, volgens de correlatiediagram-methode, een lijn door de oorsprong te zien moeten geven. Is dit echter niet het geval, doch vindt men een positief vast bedrag, dan is dit soms een aanwijzing voor verspillingen. Men kan uit dit voorbeeld leren, dat het analyseren door middel van correlatie-diagrammen reeds veel kan leren; het uitrekenen van de regressielijnen is dikwijls niet eens nodig.

Het kan daarom door al deze omstandigheden voorkomen, dat er een opmerkelijk verschil bestaat tussen de boekhoudkundige en de statistische kostensplitsing. Daarbij zij nog opgemerkt, dat het verschil maakt, of de analyse betrekking heeft op de short run of op de long run. In het laatste geval verdwijnen de vaste kosten en worden alle kosten practisch variabel.

3. Voorts zij opgemerkt, dat men niet altijd de statistische analyse kan toepassen, met name in het geval, waarin er een te gering aantal getallen-paren is. Zou men b.v. slechts over de getallen van twee halve jaren beschikken, dan vindt men grafisch twee punten, waardoorheen altijd wel een rechte lijn loopt, doch waarvan de betrouwbaarheid zeer gering is. Voor statistisch onderzoek heeft men een menigte gegevens nodig.

En dat niet alleen: er moet ook een voldoende variatie in elke variabele zijn, anders komen alle punten te dicht op elkaar en wordt de conclusie minder betrouwbaar.

Ook kan het voorkomen, dat een of meer punten ver buiten het normale verloop liggen. Het kan zijn, dat dit te wijten is aan een bepaalde specifieke oorzaak, die niet heeft gewerkt bij de andere waarnemingen. Blijkt dit het geval te zijn, dan moet men dit punt van de berekeningen uitsluiten, daar één enkel dergelijk abnormaal punt de mate van correlatie zeer sterk kan verstoren. Blijkt, dat geen specifieke oorzaak kan worden gevonden, dan moet men deze punten hun rol laten spelen; anders zou men zichzelf en anderen bedriegen.

4. Heeft men twee of meer identieke getallenparen, dan moet men dit getallenpaar ook twee of meer keren opnemen. Dit kan men eenvoudig doen door de bij elkaar behorende waarden te vermenigvuldigen met de frequentie waarin zij voorkomen. Voor de berekeningen moet men dan niet vergeten dat het aantal waarnemingen gelijk is aan de som van de frequenties der identieke waarnemingen.

5. Tenslotte willen wij wijzen op een gevaar, dat elke methode in zich bergt, n.l., dat men ze gaat toepassen als een routine. De correlatierekening is niet een werktuig om afhankelijkheden gedachteloos op te sporen, want dan kan men zeer goed irrationele correlaties vinden. Als verschijnsel *A* de verschijnselen *B* en *C* oproept, dan kan men tussen de laatste verschijnselen een correlatie vinden, die irrationeel is, omdat men de gemeenschappelijke oorzaak *A* buiten spel heeft gelaten.

Ook moet men niet steeds aan het rekenen slaan, want de berekeningen zijn doorgaans nogal tijdrovend. Men moet echter wel een sterke feeling voor afhankelijkheden trachten te ontwikkelen. Deze feeling is zeer belangrijk, want ze doet de dingen zien zoals ze zijn: afhankelijk.