KONINKLIJK NEDERLANDS
METEOROLOGISCH INSTITUUT

Verslagen V-48
(R III-231-1959)

DE HOEKTOETS

door

Dr. C. Levert

519.2

INHOUD

De Hoektoets

0. Inleiding

0.1. Het probleem.

Gegeven: N stellen getallenwaarden van n "bij elkaar" behorende stochastische variabelen $\underline{x}_1$, $\underline{x}_2$ ... $\underline{x}_n$.

Gevraagd: is er <u>associatie</u> tussen de variabelen in het n-voudige universum?

0.2. Definitie van Associatie

De definitie van <u>non-associatie</u> = onafhankelijkheid = non-correlatie is a.v.: indien de kans $\phi(a_1, a_2 \ldots a_n)\, da_1 \ldots da_n$, dat tegelijkertijd de $\underline{x}_1$ een waarde aanneemt tussen a_1 en a_1+da_1, de $\underline{x}_2$ tussen a_2 en a_2+da_2 ..... de $\underline{x}_n$ tussen a_n en a_n+da_n, geschreven kan worden als het product van n kansen $\varphi_i(a_i)\, da_i$ $(i=1,2,..n)$, waarbij iedere φ_i geen der a_j's $(j \neq i)$ bevat, en indien dit zo is voor alle waarden van de a_i's, voor welke de kansfuncties gedefinieerd zijn, dan heten de $\underline{x}_1$, $\underline{x}_2$ ... $\underline{x}_n$ <u>ongeassocieerd</u>. Het wel geassocieerd zijn kan, gezien in het licht van bovenstaande definitie, zeer veel betekenen. Doorgaans denkt men aan de lineaire regressie; dan hangt (bijv.) $\mathcal{E}\underline{x}_n$ lineair samen met de overige n-1 x_j's, d.i. $\mathcal{E}\underline{x}_n = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_{n-1} x_{n-1}$. Gegeven de waarden x_1, x_2 ... x_{n-1}, dan volgt de bijbehorende waarde van x_n nog een kansverdeling, waarin de gemiddelde waarde door $\mathcal{E}\underline{x}_n$, vollediger geschreven: $\mathcal{E}(\underline{x}_n \mid x_1, x_2, \ldots x_j, \ldots x_{n-1})$, wordt voorgesteld. Dikwijls stelt men ook nog de eis, dat de variantie σ_n in deze voorwaardelijke verdeling, vollediger geschreven: $\sigma(\underline{x}_n \mid x_1, x_2, \ldots x_j, \ldots x_{n-1})$ constant is, d.w.z. niets met de x_1, x_2 ... x_{n-1} te doen heeft. Dit is een kwestie van definitie. Men kan ook aan andere functionele verbanden tussen $\mathcal{E}\underline{x}_n$ en x_1, x_2 ... x_{n-1} denken, bijv. aan tweede-graadsregressies, aan exponentiele relaties enz. Het kan ook zijn, dat men slechts interesse heeft voor de relatie tussen σ_n aan de ene en x_1, x_2 ... x_{n-1} aan de andere zijde. Deze opmerkingen bedoelen duidelijk te maken, dat - mocht de toets tot een statistisch signifikante associatie leiden - het zonder bijv. verdere statistische onderzoekingen niet mogelijk is deze associatie te interpreteren.

0.3. Lineaire regressie

Dikwijls stuurt men onmiddellijk aan op een lineaire regressie, waarbij bijv. $\underline{x}_1$ gezien wordt als afhankelijke variabele, die "bepaald wordt" door de overige $\underline{x}_j$'s (tussen welke weer associatie aanwezig kan zijn). Men berekent dan partiële en multipele korrelatie-koëfficienten. Ook wel bouwt men een meervoudige variantie-analyse op.
Als het aantal (N) stellen en als in het bijzonder het aantal (n) variabelen groot is, worden dit omvangrijke en tijdrovende berekeningen, welke vanzelfsprekend via overzichtelijke schema's en d.m.v. speciale rekenmachines vergemakkelijkt kunnen worden. Nadat dan de korrelatie-koëfficiënten berekend zijn, raadpleegt men tabellen teneinde te beoordelen of zij (statistisch) significant van nul verschillen en wel van nul, omdat men er van uit gaat, dat bij afwezigheid van enige associatie de gemiddeld te verwachten waarden van deze koëfficiënten nul zullen zijn. Aldus komt men wellicht tot de uitspraak: "er is associatie (non-associatie)" behoudens de kans α% een foutieve konklusie te trekken. Gewoonlijk zondigt men tegen de belangrijke basisthese, dat elk der variabelen normaal verdeeld moet zijn (dus zeker contimu en zeker symmetrisch) opdat het statistisch verantwoord is bedoelde tabellen te gebruiken. Natuurlijk kan het zijn, dat men vanwege een te klein materiaal toch niet in staat is te verifiëren of aan deze voorwaarde voldaan is.
En mocht het materiaal wel voldoende groot zijn, dan ontbreekt dikwijls de tijd. Tenslotte bestaat de mogelijkheid, dat men zonder een nader onderzoek direct zeker weet, dat één of meer der variabelen niet normaal verdeeld is. Wij denken aan de dagsom neerslag, de overdagbewolking, de relatieve vochtigheid van 14 uur in januari.
In al deze gevallen verdient het aanbeveling een associatie-toets toe te passen, die
én parametervrij is (d.w.z. niets eist ten aanzien van de verdelingsfuncties)
én weinig rekenwerk vraagt (en liefst zich leent voor machinale verwerking van ponskaarten)

Pas als zulk een toets geleerd heeft, dat er zeer waarschijnlijk associatie is tussen de variabelen, heeft het zin over te gaan tot omvangrijkere berekeningen van regressiekoëfficiënten, korrelatiekoëfficienten (lineaire zowel als kromlijnige regressies) en wat dies meer zij.

Leerde de toets het tegendeel, dan ga men niet verder en heeft men zich veel werk bespaard.
Van zulke toetsen zijn er verschillende. Wij noemen o.a.

1. De χ^2-toets in twee dimensies (x en y; of x_1, x_2), in een k x ℓ-tableau. ($k \geqslant 2$; $\ell \geqslant 2$). Bij n variabelen (x_1, x_2 ... x_n), ingedeeld in resp. k_1, k_2 ... k_n klassen: $k_1 x\, k_2 x \ldots k_n$ "cellen" (hyperkubussen).
2. De toets van Schelling
3. Zekere run-toetsen
4. De hoektoets

1. De hoektoets (corner test; quadrant count test; quadrant sum test)

1.1. In dit rapport wil ik een lans breken voor de z.g. hoektoets. Hier en daar leest men van het bestaan van deze toets, doch meer ook niet. Vele leerboeken lopen er snel over heen. Waaraan dit precies ligt is mij niet duidelijk. Natuurlijk heeft de toets zijn voor- en nadelen, zoals trouwens elke toets deze bezit. Natuurlijk moet men de uitkomsten na toepassing op de juiste wijze interpreteren, maar ook geldt dit voor iedere andere toets. De "efficiency" schijnt nog niet onderzocht te zijn; ongetwijfeld een nadeel, maar van enige andere toetsen kent men ook nog niet de efficiency. De toets eist afwezigheid van gelijke waarden onder de x_1, of onder de x_2 enz., maar andere parametervrije toetsen komen ook in moeilijkheden indien het aantal gelijken "te groot" wordt. Enz.
De toets moet zeker als een ruw oriënterende toets beschouwd worden. Men ziet haar dan ook gerangschikt worden in de rubriek rough and ready tests, fasttests, short-cut tests, quick and dirty tests (zo iets als "klaar-terwijl-u-wacht"-toetsen).

Voor wiskundige afleidingen zij verwezen naar het artikel der auteurs P.S. Olmstead and J.W. Tukey in de Ann.Math.Stat. 18, 495, (1947).

1.2. De propedure.

1.2.0. Neven-onderstelling: de variabelen zijn contimm verdeeld (de kans op het optreden van twee of meer gelijken is daardoor nul).
Nulhypothese H_o: er is geen associatie.

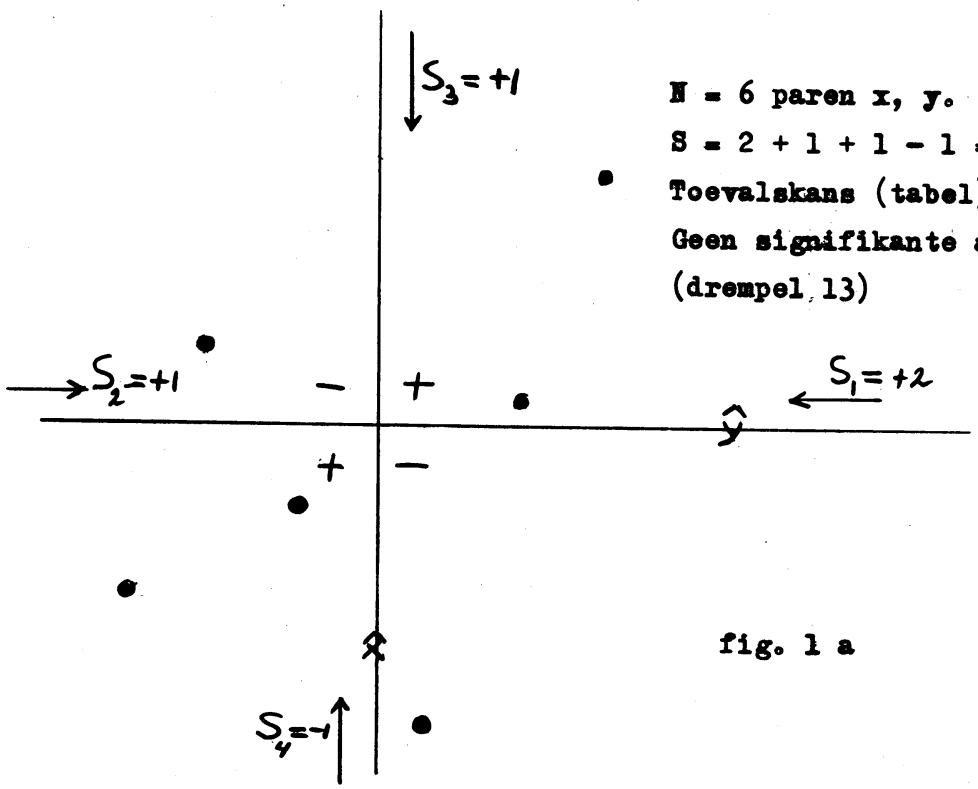
1.2.1. Voor twee variabelen x en y

Men heeft gemeten N paren x, y. Men wil H_o toetsen. Onderstel N = even, alsmede dat zowel onder de x- als onder de y-waarden geen gelijken zijn (zie overigens addendum 4.2 en 4.3.). De mediaanwaarden x en y vallen dus niet samen met een x-, resp. een y-waarde. Zet de N paren als punten uit in een Cartesisch coördinaten stelsel met lineaire schalen. Teken de twee mediaanrechten. Deze verdelen het vlak in vier kwadranten, die wij + of - kunnen noemen. De kwadranten, waarin $x > \hat{x}$ en $y > \hat{y}$ of waarin $x < \hat{x}$ en $y < \hat{y}$, heten per definitie +; de andere twee heten -. Anders gezegd (en aldus direct generaliseerbaar tot meer dan 2 variabelen): als x positief heet indien $x > \hat{x}$ (en anders negatief) en idem voor y, dan heet het punt x, y positief als het product der tekens der coördinaten positief is.

Start, van rechts komende, met de hoogste x-waarde en wandel de puntenwolk binnen. Tel alle punten, zolang ze nog hetzelfde teken hebben, tot dit teken verandert, hetgeen geschiedt als men een mediaanrechte passeren moet. Het tot deze passage getelde aantal punten heet S_1, voorzien van teken. (het teken der punten). Start nu vanuit de kleine x-waarden en handel dito: S_2. Vervolgens vanuit grote y-waarden: S_3 en vanuit kleine y-waarden: S_4. Tel op:
$S = S_1 + S_2 + S_3 + S_4$ (S kan pos., nul en neg. zijn).
In fig. 1 ziet men 2 simpele voorbeelden.
Genoemde auteurs hebben berekend aan welke waarschijnlijkheidsverdeling deze S gehoorzaamt, als een aselecte steekproef van N paren uit een biuniversum genomen wordt en H_o geldt. Zie tabel 1, waarin de kans op een $|S| \geq k$ getabelleerd is voor k = 0, 1, 2, 3 ... en N = 2, 4, 6, 8, 10, 14, ... De variantie σ^2 (S) in de waarschijnlijkheidsverdeling van S wordt eveneens genoemd.
Hoe groter de N, hoe beter kunnen wij $P[|S| \geq k]$ benaderen met de formule

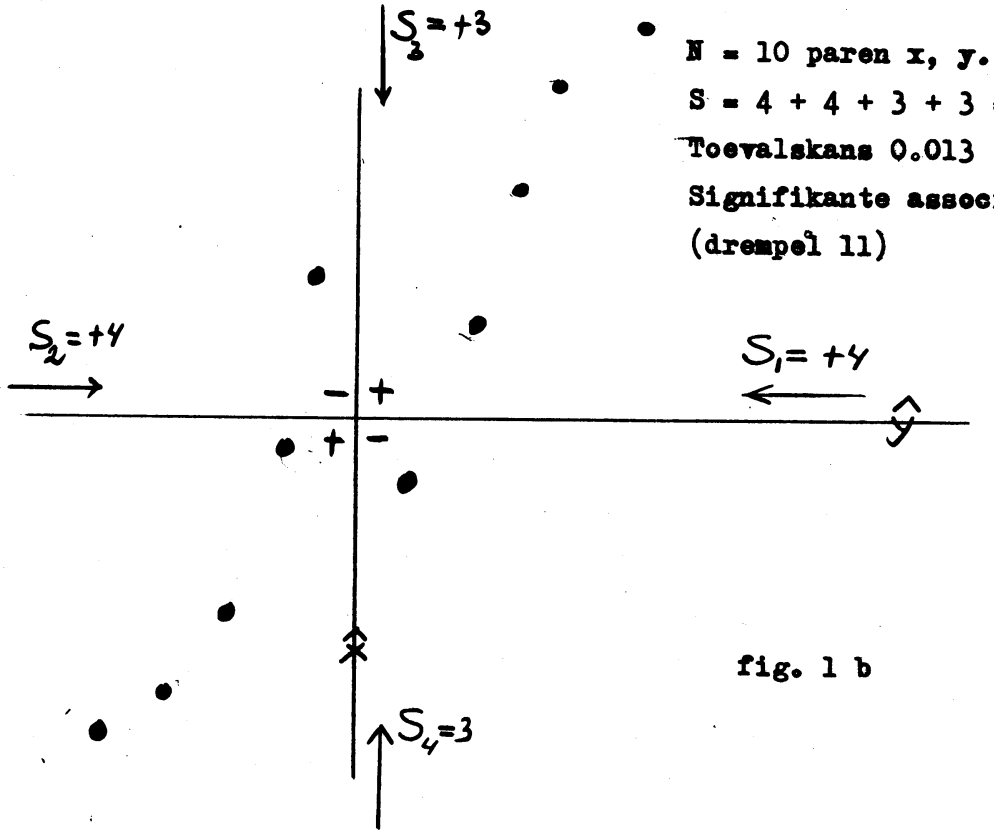
$$P = (9k^3 + 9k^2 + 168k + 208)/216.2^k$$

Hoe groter n, hoe beter is $\underline{S}$ normaal verdeeld rondom nul met variantie 24. Men kan dus ook een schatting van $P[|\underline{S}| \geq k]$ maken door in de standaardnormale verdeling van $\underline{\tau}$ te berekenen (via tabellen) met welke kans de waarde $t \equiv \{|S| - \frac{1}{2}\} : 24$ wordt overschreden.



N = 6 paren x, y.
S = 2 + 1 + 1 − 1 = 3
Toevalskans (tabel) is 0.60
Geen signifikante associatie
(drempel 13)

fig. 1 a



N = 10 paren x, y.
S = 4 + 4 + 3 + 3 = 14
Toevalskans 0.013 (tabel)
Signifikante associatie
(drempel 11)

fig. 1 b

Het valt op, dat de 5 % drempel S^* voor $N \geqslant 8$ tussen N= 10 en 11 blijft liggen, m.a.w. voor $N \geqslant 8$ een vast getal, n.l. 11, is.
Er is derhalve associatie (behoudens de kans 0.05 dat de uitspraak fout is), als $|S| \geqslant 11$. Men behoeft slechts dit ene getal 11 te onthouden.
N.B. associatie kan hier positief en negatief zijn!

1.2.2. Voor meer dan twee variabelen

Er zijn n variabelen $\underline{x}_1, \underline{x}_2 \ldots \underline{x}_n$ (en wel N stellen metingen.) In de n-dimensionale Cartesische ruimte liggen dus n assen en N punten. Er zijn n mediaanwaarden, n mediaan "hypervlakken" en 2^n hyperkwadranten. Weer kennen we aan elk der waarden van $x_1, x_2 \ldots x_n$ een + of - teken toe. Aldus draagt ook het punt P een teken en wel het product der tekens der coördinaten. Het hyperkwadrant heeft het teken der daarin gelegen punten.
Is n = 3 dan zijn er 4 + octanten: + + +; + - -; - - +; - + -
en 4 - octanten: - - -; - + +; + + -; + - +.
In 2 dimensies (2 assen) is $S = S_1 + S_2 + S_3 + S_4$
in 3 dimensies (3 assen) is $S = S_1 + S_2 + \ldots S_6$ etc.
in n dimensies (n assen) is $S = \sum_1^{2n} S_i$

De telprocedure is geheel analoog. Bijv. n = 5 variabelen. Men start vanuit de hoge x_1-waarden en wel bijv. met een punt: + - + + +
(een - punt). De kombinatie met de eerst lagere x_1 is bijv. + + + - +
(weer een - punt); vervolgens + - + + +, + + - + +, + - - + +.
Aldus 4 -punten, gevolgd door een + punt en derhalve $S_1 = -4$.
Bij overgang van het 4^{de} naar het 5^{de} punt komt men in een dotriacontant (de 5 dimensionale ruimte wordt door de 5 hypervlakken in $2^5 = 32$ dotriacontanten verdeeld) van een ander teken. Dezelfde x_1-as levert ook een tweede component S_2, etc. Ook nu blijkt de S^*-drempel bij gegeven onbetrouwbaarheid, zegge 0.05, voor voldoend grote N niet meer te veranderen, doch wel met toenemende n toe te nemen.
Zie tabel 2.

Tabel 2

	S^*-drempels								
onbetr.	n=2	3	4	5	6	7	8	9	10
0.05	11	13	15	16	18	19	20	22	23
0.01	14	16	17	19	21	22	24	25	26

Voorbeeld: de getelde S duidt bij n = 6 variabelen op associatie, behoudens de kans 0.05 op een foutieve uitspraak, indien $|S| \geqslant 18$. (en wenst men groter zekerheid, t.w. een onbetrouwbaarheid van 0.01, dan dient $|S| \geqslant 21$ te zijn).

2. Numeriek voorbeeld

Men hielp mij aan een getallenvoorbeeld met n = 5 variabelen x_1, x_2, x_3, x_4 en x_5 waarvan er één de afhankelijke is en de andere vier de verklarende variabelen zijn. Met opzet wilde ik niet weten welke dat waren. De waarden x_1 mochten met een willekeurig getal worden vermeerderd of vermenigvuldigd; de x_2-waarden eveneens, etc. De uitkomst van de toets is daar toch ongevoelig voor. Voor mij waren de getallen derhalve niets meer dan getallen: 29 stellen bij elkaar behorende getallenwaarden. Vraag: vertonen zij associatie? Zie tabel 3.

Tabel 3

	x_1	x_2	x_3	x_4	x_5
1	139	169	139	121	145
2	150	138	149	216	84
3	156	150	166	139	71
4	125	184	197	154	118
5	92	186	102	8	149
6	170	165	125	11	112
7	104	153	170	55	100
8	203	161	130	183	40
9	147	163	174	106	133
10	77	221	31	95	148
11	102	206	94	106	174
12	212	167	175	81	66
13	157	193	108	173	87
14	123	213	179	78	138
15	112	193	29	51	121
16	128	173	159	111	84
17	157	179	48	94	127
18	61	192	148	64	155
19	169	176	129	164	87
20	138	139	124	100	72
21	123	177	135	139	83
22	145	197	67	46	53
23	137	150	146	213	58
24	154	174	154	219	104
25	173	185	195	66	64
26	187	172	160	133	67
27	137	147	139	63	117
28	80	181	142	159	76
29	200	141	153	241	35

Wij merken direct op: N = 29 = oneven; bij gevolg is voor elke x_i in de naar grootte gerangschikte waarden de 15de waarde de mediaan. Het feit, dat één der waarden zelf de mediaan is (zodat men niet weet of een + of een -teken toegekend moet worden) kán moeilijkheden (twijfel omtrent de S-waarde) meebrengen, maar dit behoeft niet per se het geval te zijn.

In het gegeven getallenvoorbeeld kwamen wij inderdaad in moeilijkheden, nog verergerd doordat in enkele der 5 variabelen de mediaan zelfs meer dan enkelvoudig vertegenwoordigd is. Aldus is het wenselijk 5 stellen te schrappen. De keuze ervan is natuurlijk vrij persoonlijk, doch wanneer daarbij niet gelet wordt op de constellatie der punten in de 5 dimensionele ruimte en wanneer men - zoals hier - helemaal niet weet wat de variabelen voorstellen, kan de consequentie van deze subjectiviteit slechts klein zijn. Wij schrapten het stel met $x_1 = 156$ (louter willekeurig gekozen) en de 2 stellen met $x_4 = 106$ ($\hat{x}_4 = 106$) en de twee stellen met $x_5 = 87$ ($\hat{x}_5 = 87$)
Zo resteren er 24 stellen met nieuwe mediaanwaarden $\hat{x}'_1 = 137\frac{1}{2}$; $\hat{x}'_2 = 173\frac{1}{2}$; $\hat{x}'_3 = 144$; $\hat{x}'_4 = 106$ en $\hat{x}'_5 = 87$. Doordat voor geen der variabelen de nieuwe mediaan een der 25 waarden zelf is (de mediaan is n.l. gelegen tussen de 12 en 13^{de} waarde) kan aan elk der waarden in alle stellen een + of -teken toegekend worden.

Men rangschikt nu 5 keren, n.l. één keer naar opklimmende x_1 (tabel 4), een keer naar x_2 (tabel 5), een keer naar x_3, een keer naar x_4 en een keer naar x_5.
Het is de bedoeling alle soorten associaties te onderzoeken, d.w.z. tussen x_1 en x_2, tussen x_1 en x_3, x_1 en x_4, x_1 en x_5, x_2 en x_3 .... x_4 en x_5; de 3-voudige associaties, kort aangeduid met 123, 124, 125, 134, 135, 145; de 4-voudige associaties 1234, 1235, 1245, 1345; en tenslotte de 5-voudige associatie 12345.
Elk der associaties levert 2 S-componenten, n.l. één komende vanuit de hoge en één komende vanuit de lage x-waarde. De S-bijdragen, voorzien van teken, vindt men ook in tabel 4. Tabel 5 heeft x_2 voorop. Met resp. x_3, x_4, x_5 voorop, analoge tabellen. De aldus getelde S-componenten worden vervolgens overzichtelijk samengenomen in een eindtabel 6, waarin per kolom algebraisch gesommeerd wordt. De signifikantie-drempels zijn S^{**} = (2 v. ass.), 13 (3 v.), 15 (4 v.) en 16 (5 v.). (met 0.05 onbetrouwbaarheid). De signifikante S-waarden zijn dubbelonderstreept.
Eerst nu vernam ik de betekenis der variabelen. Tabel 7 geeft overzichtelijk de signifikant bevonden associaties. Enig commentaar: dat de hoeveelheid zon in de decaden juli II en III (negatief) gecorreleerd is met de neerslag in juni + juli is niet zo erg verwonderlijk; de decaden juli II en III zijn er twee van de zes in juni + juli.

Het schema vertelt, dat de opbrengst (x_5) geassocieerd is met de zon(x_2) in juli II + III. Voegen wij x_3 toe dan is het met de associatie gedaan; voegen wij ook x_4 toe, dan is x_1 duidelijk met de combinatie van x_2, x_3 en x_4 geassocieerd. Kennelijk heeft het weinig zin tot een 5-voudige associatie over te gaan. Een toets leert niet hoe men nu tussen x_5, x_1, x_2, x_4 en x_5, x_2, x_3, x_4 moet kiezen, al heeft ook het eerste vijftal een meer signifikante associatie dan het andere (het is echter weinig zinvol om van "meer signifikant" te spreken). Hier moet de onderzoeker op andere dan louter statistische gronden kiezen.

Het lijkt me duidelijk, dat de toets een oriënterende taak heeft. Het rekenwerk is louter telwerk, nadat tekens toegekend zijn en mocht het om grote aantallen stellen (hoe meer, hoe liever) en vele variabelen gaan, dan kan men dit werk, via ponskaarten, zelfs machinaal laten verrichten. Het is wel zaak met even aantallen stellen en met ongelijke waarden te werken.

3. Samenvatting

Het addendum geeft nog enige verdere bijzonderheden.

Voor- en nadelen van de toets tegenover elkaar stellende, komt er:

Voordelen	Nadelen
1. parametervrij (wel moet elk der variabelen continu verdeeld zijn)	1. is tenminste één der variabelen discontinu verdeeld, dan kan de hoektoets niet toegepast worden.
3/2. weinig rekenwerk: het enige rekenwerk is het toekennen van tekens en aftellen van "gelijkgetekende" combinaties.	2. heeft men juist interesse voor de grote massa der "punten" (rondom de medianen) dan moet men niet de hoektoets gebruiken.
2/3. de toets richt de aandacht op associatie tussen extreme waarden der variabelen (daarom: periferie of hoek-toets)	3. de toets levert geen associatiegraad (maar de toets is hiervoor ook niet opgezet!)
4. men kan zich na deze oriëntatie beraden op meer rekenwerk bij toepassing van andere statistische methodes.	-

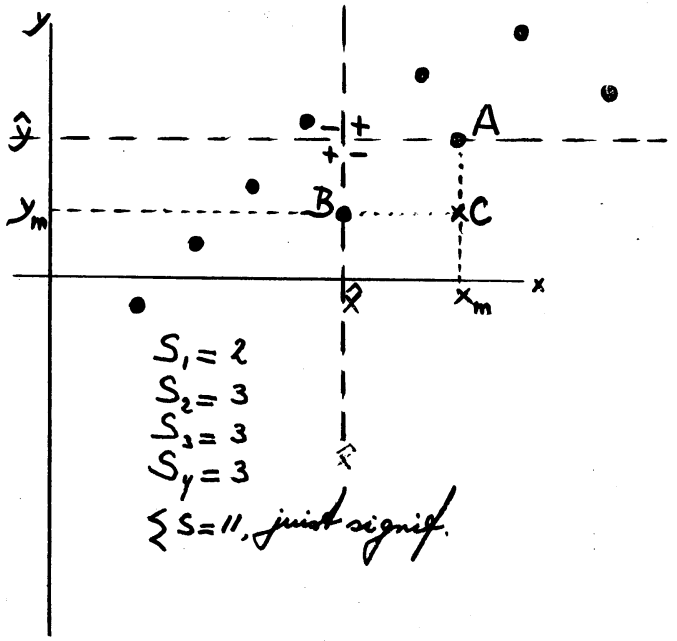
4. Addendum

4.1. Een opmerking over associatie tussen $\underline{x}_1$, $\underline{x}_2$ en $\underline{x}_3$

Olmstead en Tukey zeggen: gesteld er zullen N tripels x_{1i}, x_{2i}, x_{3i} gemeten zijn. Natuurlijk kan men in de steekproeven x_{1i}, x_{2i}; x_{1i}, x_{3i} en x_{2i}, x_{3i} de korrelatiekoëfficiënten r_1, r_2, r_3 berekenen of men kan drie keren de hoektoets toepassen. Belangrijk is: wanneer er in het geval van een driemensionale normale verdeling tussen de paren $\underline{x}_1$, $\underline{x}_2$; $\underline{x}_1$, $\underline{x}_3$; $\underline{x}_2$ en $\underline{x}_3$ drie keren geen correlatie is, dan is ook een drievoudige associatie afwezig. Omgekeerd mag men uit het feit, dat tussen deze paren drie keren associatie afwezig is, geconstateerd via bijv. de hoektoets, eerst dan tot drievoudige non-associatie concluderen als men van de normaliteit der verdelingen zeker is.

4.2. Hoe te handelen indien N = oneven in het geval van twee variabelen $\underline{x}$ en $\underline{y}$?

De mediaan $\hat{x}$ valt met de middelste x-waarde samen en idem voor $\hat{y}$. Moeilijkheden bij de telprocedure zijn er alleen dan (dus lang niet altijd) als men komt aan een punt, waarvan één der coördinaten $\hat{x}$ of $\hat{y}$ is (zeer toevallig kan ook $\hat{x} = \hat{y}$ zijn). Laat deze twee punten $A \equiv x_m, \hat{y}$. (m is één der getallen 1, 2 ... N) en $B \equiv \hat{x}, y_m$ (m als k) zijn. Zie bijgaande figuur. Schrap deze punten en vervang ze door een nieuw punt $C \equiv x_m, y_k$. Dan is N met 1 verminderd en even geworden. Pas de oude teltechniek toe en er zijn nu geen moeilijkheden meer; de mediaanrechten zijn niet van plaats veranderd.



Van rechts komende, is $S_1 = 2$ of 4? Doet punt A mee of niet in de bijdrage tot S_1?
Antw.: schrap de punten A en B en voeg C toe.
Daarna is $S_1 = 2$ enz.

Fig. 2

4.3. Hoe te handelen met gelijken onder de x-waarden (en) of onder de y-waarden in het geval van twee variabelen?

De auteurs raden aan zo min mogelijk hiermee te doen te hebben: dus allereerst geen discontinu verdeelde variabele beschouwen (aantal regendagen, aantal zonnige dagen per decade enz.) en ten tweede niet te veel af te ronden.
Overigens vonden ze half-empirisch het volgende regeltje: gesteld men wandelt vanuit de kleine x-waarde het x-y-veld binnen. Men telt de "gelijkgetekende" punten af, doch stuit dan opeens op een k-tal punten, alle met dezelfde x_k, doch met y-waarden aan weerszijden van $\hat{y}$; stel k_1 stuks met y's $> \hat{y}$ en k_2 stuks met y's $< \hat{y}$ ($k_1 + k_2 = k$). Òf de k_1 punten zijn ten gunste van de continuering der telling òf de k_2 punten zijn het niet; dat hangt van de situatie af. In elk geval: neem deze k punten samen als q= (gunstig): (1+ ongunstig), als bijdrage tot de S-component; altijd is $q < 1$.
Toelichting

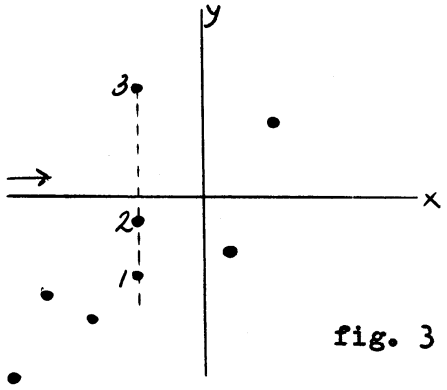


fig. 3

Hoe met de punten 1, 2 en 3?
Men komt (zie→) vanuit kleine x.

Te bedenken, dat de x-coördinaten der 3 punten 1, 2 en 3 misschien door afronding gelijk werden. Er zijn bij ongelijke x-waarden 6 mogelijkheden (deze coördinaten a_1, a_2, a_3 noemende), t.w. $a_1 > a_2 > a_3$; $a_1 > a_3 > a_2$; $a_2 < a_1 > a_3$; $a_2 > a_3 > a_1$; $a_3 > a_1 > a_2$ en $a_3 > a_2 > a_1$. Bij het gegeven schetsje levert dat 6 mogelijkheden voor S, t.w.
3 + 2= 5 3 + 1= 4 3 + 2= 5 3 + 1= 4 3 + 0= 3 3 + 0= 3
Gemiddeld (5 + 4 + 5 + 4 + 3 + 3) : 6= 4, d.i. ook 3 + (2)/(1 + 1)= 3 + 1= 4, waarin (2)/(1 + 1)= ("gunstig"): (1 + "ongunstig") = q.
Die 6 mogelijkheden werden daarbij evenwaarschijnlijk geacht.
N.B. indien er meer dan twee variabelen zijn, wordt dit alles veel moeilijker. De auteurs geven geen oplossing. Hun enige raad is ervoor te zorgen, dat er geen gelijken voorkomen, noch onder de x_1-waarden, noch onder de x_2-waarden, etc.

De toets beschouwt in hoofdzaak de situatie in de periferie der punten constellatie.

4.4. Olmsted en Tukey noemen als opmerkelijke eigenschappen van hun toets

a) speciaal gewicht wordt gelegd op de extreme waarden van de variabelen

b) gemakkelijke berekeningen

c) parametervrij

Wat a) betreft: men heeft een "stippenkaart" gemaakt van een matig groot aantal (zegge 25 à 200) paren x, y. Meestal begint men met deze kaart (puntenconstellatie) "aan te kijken". Men spreekt plastisch van "bolvormige", "sigaarvormige", "haltervormige" constellaties enz. Waardoor laat men zich daarbij leiden? Meer door de punten aan de periferie (zeer sterk gesproken: "uitbijters") dan door die der massa? Dikwijls is meer het eerste dan het tweede het geval. Soms is dit wenselijk, soms niet; dat hangt van de aard van het probleem af. Een kwantitatieve toets voor associatie, die juist de aandacht richt op de "omtrek" van de stippenwolk is er niet (volgens de auteurs). Daarom ontwikkelden zij hun cornertest.
Mocht het juist noodzakelijk zijn deze "omtrek-stippen" niet te beschouwen (misschien is er een gegronde vrees, dat juist zij zeer onnauwkeurig zijn) dan doet men goed de hoektoets niet toe te passen.
De schrijvers vertellen, dat ze niet weten hoe het met de efficiency van de toets gesteld is. Trouwens zij noemen nog meer onopgeloste mathematische vragen inzake de corner test en vragen aan de lezers die te willen beantwoorden. De mathematicus-statisticus helpe hier!

4.5 De hoektoets en een onderzoek naar de representativiteit

Gesteld men wil weten of een regenstation A "representatief" is voor een bepaald gebied rondom A_o. Wat van de neerslag? Dagsommen? Maandsommen? Stortregens? Wat is representatief?
De meningen over de juiste definitie van representatief kunnen verdeeld zijn.

a) "Klimatologische" definitie

Een gebied G rondom A_o heet "gerepresenteerd" te kunnen worden door A_o (A_o is "representatief voor" G) bijv. ten aanzien van dagsommen, wanneer, gegeven de dagsom h te A_o, die te A_i een kansverdeling rondom een gemiddelde h volgt, die voorgeschreven "nauw" is; dit dan voor alle stations uit G en elke h. [als de kansverdeling symmetrisch is, zou men kunnen eisen, dat de standaarddeviatie σ_i ervan ten hoogste 10 % van h is, voor alle i en alle h 's; men kan ook de correlatie-coëfficient tussen A_o en A_i als associatiegraad beschouwen]. De definitie impliceert verwisselbaarheid; A_i en A_o mogen voor elkaar in de plaats gesteld worden; ze kunnen elkaar vervangen.

b) andere definitie

A_o heet representatief voor G (waarin overigens nog de stations A_1, $A_2 \ldots A_p$ gelegen zijn) als er een voldoend nauwe associatie tussen A_o en $M_p = \frac{1}{p+1}(A_o + A_1 + \ldots A_p)$ is. Wat "voldoend nauw" is zal statistisch gedefinieerd moeten worden. Men moet daartoe een associatie-graad invoeren. Men zou kunnen denken aan een lineaire regressie tussen A_o en M_p, bijv. $\mathcal{E} M_p = \alpha + \beta A_o$, of aan een kwadratische regressie, bijv. $\mathcal{E} M_p = \alpha + \beta A_o + \gamma A_o^2$ enz. Het zal niet per se nodig zijn, in het geval van lineaire regressie, dat $\beta = 1$ en $\alpha = 0$ zijn, in welk geval er "verwisselbaarheid" (zie de klimatologische definitie) zou zijn. Nodig is slechts dat, gegeven de A_o, de bijbehorende M_p voldoend strak vastligt. In ieder geval: een „voldoend nauwe" associatie betekent associatie, d.w.z. men gaat, logisch, eerst dan onderzoeken of de associatie voldoende strak is, nadat „statistisch is kunnen worden bewezen", dat ze signifikant is (met 0.05 kans op een foutieve uitspraak).
Indien men meent, dat de realiteit van de associatie niet onmiddellijk voor de hand ligt, zou men haar bijv. met de hoektoets kunnen onderzoeken.
Zij de stations rondom A_o aangeduid met A_1, A_2 ...; $\overline{A_o A_i}$ neemt toe met i.

Het gebied G, "ingenomen" door de i + 1 stations, neemt dus toe met toenemende i. "Vergelijk" A_o met $M_2 = \frac{1}{2}(A_o + A_1)$; daarna A_o met $M_3 = \frac{1}{3}(A_o + A_1 + A_2)$; vervolgens A_o met $M_4 = \frac{1}{4}(A_o + A_1 + A_2 + A_3)$ etc. (de factoren $\frac{1}{2}$, $\frac{1}{3}$ ... mogen weggelaten worden) Pas steeds de associatie-toets toe (indien alles op ponskaarten plaats vindt, kan dit snel gebeuren). Ga met de "vergelijking" tussen A_o en M_j (j= 2, 3, 4...) zover, tot, op basis van het beschikbare materiaal, de associatie niet meer statistisch signifikant is. (behoudens de 0.05 kans op een foutieve konklusie). N.B.: natuurlijk kan het zijn, dat een wel signifikante associatie tussen A_o en M_k toch een in de praktijk onbruikbare relatie tussen M_k en A_o impliceert, doordat de spreiding rondom de regressierechte veel te groot is; doch deze kwestie behandelen wij hier niet; ze behoort thuis onder het chapiter "associatiegraad".

4.6. <u>De hoektoets en een onderzoek naar autocorrelatie (persistentie)</u>

Het begrip representatief is ook op de tijd te betrekken.
Beschouw bijv. dag voor dag op station S de gemiddelde overdagtemperatuur T. Is T_{i+1}(dag i + 1) afhankelijk van T_i?; idem T_{i+2} en T_i? Onderzoek derhalve de associatie tussen de paren T_1, T_2; T_2, T_3; T_3, T_4 ... Deze associatie betreft de persistentie van de orde 1. Doe het ook voor T_1, T_3 ; T_2, T_4 ; T_3, T_5 .... d.i. persistentie van orde 2 en T_1, T_4 ; T_2, T_5 enz. Men verkrijgt hiermee wel is waar niet een getal, dat de persistentie-sterkte aangeeft, maar wel een indruk van de "uitgestrektheid" (in de tijd). De persistentie in de overdagtemperatuur grijpt misschien vooruit over 6 dagen, de neerslag over 3 dagen, de zon over 2 dagen etc.

Tabel 1

Hoektoets: $P\left[|S| \geq k\right]$ onder H_o

k \ N	2	4	6	8	10	14	∞
0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1	1.000	0.750	0.933	0.904	0.911	0.912	0.912
2	1.000	0.750	0.756	0.754	0.757	0.758	0.755
3	1.000	0.417	0.600	0.600	0.601	0.604	0.600
4	1.000	0.417	0.467	0.462	0.466	0.469	0.463
5	0.000	0.333	0.311	0.351	0.352	0.355	0.347
6		0.333	0.222	0.262	0.259	0.261	0.252
7		0.333	0.156	0.182	0.187	0.188	0.178
8		0.333	0.111	0.126	0.133	0.133	0.122
9		0.000	0.100	0.084	0.093	0.092	0.081
10			0.100	0.055	0.064	0.063	0.053
11			0.100	0.038	0.044	0.043	0.034
12			0.100	0.030	0.030	0.030	0.022
13			0.000	0.029	0.019	0.020	0.013
14				0.029	0.013	0.014	0.008
15				0.029	0.010	0.010	0.005
16				0.029	0.008	0.010	0.003
17				0.000	0.008	0.004	0.002
18					0.008	0.003	0.001
19					0.008	0.002	0.001
20					0.008	0.001	0.000
21					0.000	0.001	0.000
22						0.001	0.000
23						0.001	0.000
24						0.001	0.000
25						0.001	0.000
26						0.001	0.000
27						0.001	0.000
28						0.000	0.000
$\sigma^2(S)$	16	24	$26\frac{2}{5}$	$26\frac{6}{7}$	$26\frac{16}{21}$	$26\frac{149}{429}$	24

Tabel 4

x_1	x_2	x_3	x_4	x_5	123	124	125	134	135	145	1234	1235	1245	1345	12345
64 −	192 +	148 +	64 −	155 +	−	+	−	+	−	+	+	−	+	+	+
77 −	221 +	31 −	95 −	148 +	+	+	−	−	+	+	−	+	+	−	−
80 −	181 +	142 −	159 +	76 −		−	+			+			+		
92 −	186 +	104 −	8 −	149 +						+			+		
104 −	153 −	170 +	55 −	100 +						+			−		
112 −	193 +	29 −	51 −	121 +						+					
123 −	213 +	179 +	78 −	138 +						+					
123 −	177 +	135 −	139 +	83 −						+					
125 −	184 +	197 +	159 +	118 +						−					
		enz.													
170 +	165 −	125 −	11 −	112 +											
173 +	185 +	195 +	66 −	64 −			−								
187 +	172 −	160 +	133 +	67 −			+								
200 +	141 −	153 +	241 +	35 −			+	+			−			−	+
203 +	161 −	130 −	183 +	40 −	+	−	+	−	+	−	+	−	+	+	−
212 +	167 −	175 +	81 −	66 −	−	+	+	−	−	+	+	+	−	+	−
S − bijdrage															
$x_1 < \hat{x}_1$	− 4	− 1	2	− 2	− 1	+ 2	− 2	+ 1	− 1	+ 8	+ 1	− 1	+ 4	+ 1	+ 1
$x_1 > \hat{x}_1$	− 4	1	− 1	− 5	− 1	− 1	+ 4	− 2	− 1	+ 1	+ 2	+ 1	− 1	+ 2	− 2

Tabel 5

x_2	x_1	x_3	x_4	x_5	213	214	215	234	235	245	2134	2135	2145	2345	12345
138 −	150 +	149 +	216 +	84 −	−	−	+	−	+	+	−	+	+	+	+
139 −	138 +	124 −	100 −	72 −	+	−	+	−	−	−	+	+	+	+	−
141 −	200 +	153 +	241 +	35 −		−	+	−				+	+	+	
147 −	137 −	139 −	63 −	117 +		−	+	−				−	−	−	
150 −	137 −	146 +	213 +	58 −		+	−	−							
153 −	104 −	170 +	55 −	100 +				+							
		enz.													
181 +	80 −	142 −	159 +	76 −			+						−		
184 +	125 −	197 +	154 +	118 +			−						−		
185 +	173 +	195 +	66 −	64 −			−						+		
186 +	92 −	102 −	8 −	149 +			−						+		
192 +	61 −	148 +	64 −	155 +			−						+		
193 +	112 −	29 −	51 −	121 +			−						+		
197 +	145 +	67 −	46 −	53 −		−	−			+			+		
213 +	123 −	179 +	78 −	138 +	−	+	−	−	+	−	+	−	+	−	+
221 +	77 −	31 −	195 +	148 +	+	+	−	+	−	−	−	+	+	+	−
S-bijdrage															
$x_2 < \hat{x}_2$	− 3	− 1	− 1	3	− 1	− 4	+ 4	− 5	1	1	− 1	3	3	3	1
$x_2 > \hat{x}_2$	− 2	− 1	− 7	2	1	2	− 8	1	− 1	− 2	− 1	1	7	1	− 1

Tabel 6

Samenvatting der S-componenten

	12	13	14	15	23	24	25	34	35	45	123	124	125	134	135	145	234	235	245	345	1234	1235	1245	1345	2345	12345
van + x_1	- 4	1	-1	-5							-1	1	4	-2	-1	1					2	1	-1	2		-2
- x_1	- ~~5~~4	-1	2	-2							-1	2	-2	1	-1	8					1	-1	4	1		1
+ x_2	- 2				-1	-7	2				1	2	-8				1	-1	-2		-1	1	7		1	-1
- x_2	- ~~4~~3				-1	-1	3				-4	-4	x+y				-5	1	1		-1	3	3		3	1
+ x_3		-1			3			1	1		-1			-2	-5		1	1		2	-2	-3		-1	2	-1
- x_3		2			-5			-7	-3		2			-2	-2		5	-3		3	-2	2		-2	6	-2
+ x_4			3			-1		4		-1		-1		3		-1	-1		5	-1	-1		-1	-1	4	2
- x_4			1			-1		4		-2		2		-1		1	1		-1	2	-2		4	-1	1	-4
+ x_5				1			4		-1	-3			-5		-2	5		2	-7	-2	-2	-2	3	3	-2	2
- x_5				-2			2		2	-2			3		-1	-4		1	3	-1	-1	-1	2	-1	-1	1
som	~~-12~~ -13	1	5	-8	-4	-10	11	2	-1	-8	-4	2	-~~7~~ 4	-3	-12	10	2	1	-1	3	-9	-1	21	0	16	-3

Tabel 7

	opbrengst	neerslag jun + jul	zonneschijn jul II + III	temp. neerslag winter	
	x_5	x_1	x_2	x_3	x_4
tweev.	x	x	x		
			x		
driev.	(x)	(x)		(x)	
vierv.	x	x	x		x
	x		x	x	x
vijfv.	Geen				