MATHEMATISCH CENTRUM
2de Boerhaavestraat 49
A M S T E R D A M O.
Statistische Afdeling
Rapport S59.

Maart 1951

Cursus "Parametervrije Methoden".

I. Kendall's rangcorrelatie-coefficient τ.

door

Dr J. Hemelrijk Jr.

§1. Inleiding

Bij de definitie van de rangcorrelatiecoëfficiënt τ van Kendall wordt, evenals bij vele andere parametervrije methoden, alleen de volgorde van grootte van één of meer series waarnemingen in de beschouwingen betrokken, terwijl de grootte der waarnemingen verder buiten beschouwing wordt gelaten. Dit heeft het nadeel, dat in vele gevallen slechts van een gedeelte der gegevens gebruik wordt gemaakt. Daartegenover staat het voordeel van een zeer algemene geldigheid der methoden: ten eerste kunnen ook gegevens, waarbij redelijkerwijs alleen een volgorde bepaald kan worden, met deze methoden verwerkt worden en ten tweede zal het in het algemeen niet nodig zijn veel onderstellingen te maken omtrent de vorm van de in de beschouwingen optredende waarschijnlijkheidsverdelingen.

Voorbeeld 1: Om de kleurgevoeligheid van een modeontwerpster te toetsen, worden haar een twaalftal schijfjes gegeven, die alle blauw gekleurd zijn, maar van licht tot donkerblauw variëren. De objectieve volgorde, van licht tot donker, wordt colorimetrisch bepaald en is dus bekend. De mode-ontwerpster moet de schijven naar opklimmende volgorde rangschikken en men wenst een maat aan te geven voor de overeenstemming tussen de door haar gegeven volgorde en de juiste.

Het is duidelijk, dat men in dergelijke gevallen rangordemethoden nodig heeft. Omgekeerd zullen deze methoden het ontwerpen van experimenten als vermelde vergemakkelijken en bevorderen.

Voorbeeld 2: Aan een aantal firma's wordt, vertrouwelijk, de vraag voorgelegd, welk dividend zij over het lopende jaar denken te kunnen uitkeren. Men vermoedt, dat er speciaal bij de firma's, die een hoog dividend verwachten, een zekere weerstand

bestaat tegen het geven van deze inlichtingen en, daar men na een bepaalde tijd de binnengekomen antwoorden wenst te verwerken, is het van belang te weten, of dit vermoeden juist is of niet. Dit kan men nu uitmaken, door na te gaan of er een correlatie is tussen het tijdstip van ontvangst van de binnengekomen antwoorden en de hoogte van het opgegeven dividend. Daarbij moet men echter beschikken over een toetsingsmethode, om de hypothese te toetsen, dat het vermoeden onjuist is, d.w.z. dat er geen verband is tussen het tijdstip van ontvangst en de hoogte van het dividend.

Het is duidelijk, dat het in een dergelijk geval niet verantwoord is te werken met de gewone correlatiecoëfficiënt en de daarbij behorende toetsingsmethode, daar men dan zou moeten onderstellen, dat de dividend-hoogte en het tijdstip van ontvangst van het antwoord een twee-dimensionale normale waarschijnlijkheidsverdeling zouden bezitten.

Beide voorbeelden, evenals de meeste der volgende en de besproken methoden zijn ontleend aan:

M.G. Kendall, Rank Correlation Methods, London 1948.

Dit boek bevat tevens een vrijwel volledige lijst van de litteratuur over het besproken onderwerp.

Wij bespreken eerst het geval, waarbij alle rangnummers verschillend zijn.

§2. Definitie van τ en S.

Wij zullen de berekening van τ aan een voorbeeld demonstreren en kiezen daarvoor voorbeeld 1. Wij geven de schijfjes twee rangnummers: ten eerste het juiste rangnummer (A) in volgorde van opklimmende kleurenintensiteit en ten tweede het door de proefpersoon aan de schijfjes toegekende rangnummer (B). Stel de onderstaande rangnummers zijn verkregen:

A: 1 2 3 4 5 6 7 8 9 10 11 12

B: 1 4 7 2 3 5 8 12 10 6 11 9.

Wij beschouwen nu alle mogelijke paren schijfjes (dus 66 paren in totaal) en tellen het aantal paren, waarbij de proefpersoon de juiste volgorde heeft aangewezen. Dit is bv. het geval bij de paren met A-rangnummers: (1,2), (1,3),...,(1,12) en ook bv. met het paar (2,3), maar niet met het paar met A-rangnummers (2,4). Dit aantal paren, waarvan de rangschikking juist beoordeeld is, noemen wij R '). In ons geval is R= 52. Dit kan men

---

') Kendall gebruikt de letter P, die wij liever uitsluitend als waarschijnlijkheidssymbool gebruiken.

het gemakkelijkst berekenen door, bij ieder der B-rangnummers na te gaan, hoeveel grotere B-rangnummers er achter staan, en deze aantallen op te tellen. Het is duidelijk, dat R minstens de waarde 0 en hoogstens de waarde 66 bezit.

Indien wij de eerste rij rangnummers niet in de natuurlijke volgorde hebben gezet, dus een willekeurige, maar voor de rijen A en B gelijke, permutatie op bovenstaand voorbeeld toepassen, blijft de R onveranderd; alleen de berekening wordt ingewikkelder.

Hebben wij bv. de gegevens als volgt opgeschreven:

A: 1 3 7 2 5 12 9 8 11 4 6 10
B: 1 7 8 4 3 9 10 12 11 2 5 6 ,

zodat de boven elkaar staande paren rangnummers dezelfde zijn als eerst, dan kan men niet de eenvoudige berekening van R toepassen, die boven aangegeven is, maar moet men direct op de definitie teruggrijpen: beschouwen wij b.v. de twee op de 4^{e} en 5^{e} plaats staande paren:

A: 5 12
B: 3 9 ,

dan zien we, dat voor deze twee paren de volgorde in A overeenstemt met die in B, zodat dit paar één van de 52 paren is, dat een bijdrage 1 tot R levert. Gaat men dit voor alle 66 paren na, dan vindt men weer R= 52. Het voordeel van de rangschikking, waarbij de rij A (of B) in de natuurlijke volgorde staat, is duidelijk. De berekening van R kan nog op verschillende andere wijzen worden uitgevoerd, die wij hier niet zullen bespreken.

Algemeen kunnen we R dus als volgt definiëren:

R is het aantal der paren rangnummers in rij B, waarvan de volgorde naar grootte overeenstemt met de volgorde naar grootte van de overeenkomstige rangnummer-paren van rij A.

R bezit de volgende eigenschappen:

I. Indien de rijen A en B de lengte n bezitten, ligt R tussen 0 en $\frac{1}{2}n(n-1)$.

II. Indien de rijen A en B geheel overeenstemmen, is $R=\frac{1}{2}n(n-1)$; ook het omgekeerde geldt.

III. Indien de rijen A en B precies tegengesteld zijn, is R=0; ook het omgekeerde geldt.

Kendall wenst echter een correlatie-coëfficiënt te gebruiken, die tussen -1 en +1 loopt en definieert daarom:

(1) $$\tau = \frac{4R}{n(n-1)} - 1.$$

Wij hebben dus voor τ:

Ia: τ ligt tussen -1 en +1.

IIa: Indien de rijen A en B geheel overeenstemmen is $\tau = +1$; ook het omgekeerde geldt.

IIIa: Indien de rijen A en B precies tegengesteld zijn, is $\tau = -1$; ook het omgekeerde geldt..

Verder gebruikt Kendall vaak de grootheid S, die gedefinieerd is als:

(2) $$S = 2R - \frac{1}{2}n(n-1),$$

zodat S tussen $-\frac{1}{2}n(n-1)$ en $+\frac{1}{2}n(n-1)$ ligt, terwijl het verband tussen S en τ gegeven wordt door

(3) $$\tau = \frac{2S}{n(n-1)} .$$

De definitie van S kan ook als volgt gegeven worden:

S is het aantal der paren rangnummers in rij B, waarvan de volgorde naar grootte overeenstemt met de volgorde naar grootte van de overeenkomstige rangnummerparen in A, verminderd met het aantal der paren, waarvoor dit niet het geval is.

In voorbeeld 1 hebben wij dus: S= 38 en τ = 0,58.

§3. De verdeling van S onder een nulhypothese.

Het belangrijkste voordeel van Kendall's methode ligt daarin, dat de waarschijnlijkheidsverdeling van S ') bekend is, onder de volgende nulhypothese:

H_o: De rij rangnummers B vertoont geen enkel verband met de rij rangnummers A; dat wil zeggen, dat bij gegeven rij A iedere permutatie van de rij B even waarschijnlijk is.

Voor onze beide voorbeelden betekent dit, dat, bij het eerste voorbeeld, de proefpersoon de verschillende kleuren in het geheel niet kan onderscheiden en de schijfjes in een willekeurige volgorde rangschikt, en bij het tweede voorbeeld, dat de volgorde van ontvangst van het antwoord geen samenhang vertoont met het dividendpercentage.

') S wordt in deze paragraaf als een stochastische grootheid beschouwd; dit stochastische karakter geven wij aan door onderstreping. Niet onderstreepte letters geven niet-stochastische grootheden aan, bv. door stochastische grootheden aangenomen waarden.

Voor kleine waarden van n kan men de verdeling exact berekenen met behulp van een recursieformule.

Stellen wij de kans, onder de nulhypothese H_o, dat $\underline{S}$ de waarde S aanneemt, wanneer de rijen A en B de lengte n hebben, voor door:

$$(4) \qquad P\left[\underline{S}=S \mid H_o;\ n\right],$$

en zetten wij de rij A steeds in de natuurlijke volgorde, dan hebben wij voor n=2 :

$$(5) \quad P\left[\underline{S}=-1 \mid H_o;2\right] = P\left[\underline{S}=+1 \mid H_o;2\right] = \frac{1}{2},$$

zoals direct duidelijk is.

Gaan wij nu op n=3 over, dus nemen wij er het rangnummer 3 bij in de rij B, dan is er, volgens H_o, gelijke waarschijnlijkheid, dat dit op de eerste, de tweede of de derde plaats komt te staan; dit betekent dus, dat het gelijke kans bezit om in één van de drie open plaatsen (links van beide rangnummers 1 en 2, tussen beide in, of rechts van beide) terecht te komen. Komt het geheel links terecht, dan neemt S met twee af, geheel rechts, dan neemt S met twee toe, komt het ertussen, dan blijft S gelijk. Wij kunnen dus de verdeling voor n=3 uit die voor n=2 afleiden volgens de formule:

$$(6) \quad P\left[\underline{S}=S \mid H_o;3\right] = \frac{1}{3}\Big\{ P\left[\underline{S}=S-2 \mid H_o;2\right] + P\left[\underline{S}=S \mid H_o;2\right] + P\left[\underline{S}=S+2 \mid H_o;\ 2\right]\Big\}.$$

Nemen wij bv. S=-3, dan is volgens (5) van de drie termen van het tweede lid alleen de laatste $\neq 0$, en wij vinden

$$(7) \qquad P\left[\underline{S}=-3 \mid H_o;\ 3\right] = \frac{1}{3}\,\frac{1}{2} = \frac{1}{6}.$$

Voor S=-1 zijn de laatste twee termen gelijk aan $\frac{1}{2}$, dus vinden wij

$$(8) \qquad P\left[\underline{S}=-1 \mid H_o;\ 3\right] = \frac{1}{3}.$$

De algemen recursieformule luidt:

$$(9) \quad P\left[\underline{S}=S \mid H_o;\ n+1\right] = \frac{1}{n+1} \sum_{\nu=0}^{n} P\left[\underline{S}=S-n+2\nu \mid H_o;\ n\right],$$

en kan op dezelfde wijze worden afgeleid.

In figuur 1 is de verdeling van $\underline{S}$ voor enkele waarden van n getekend.

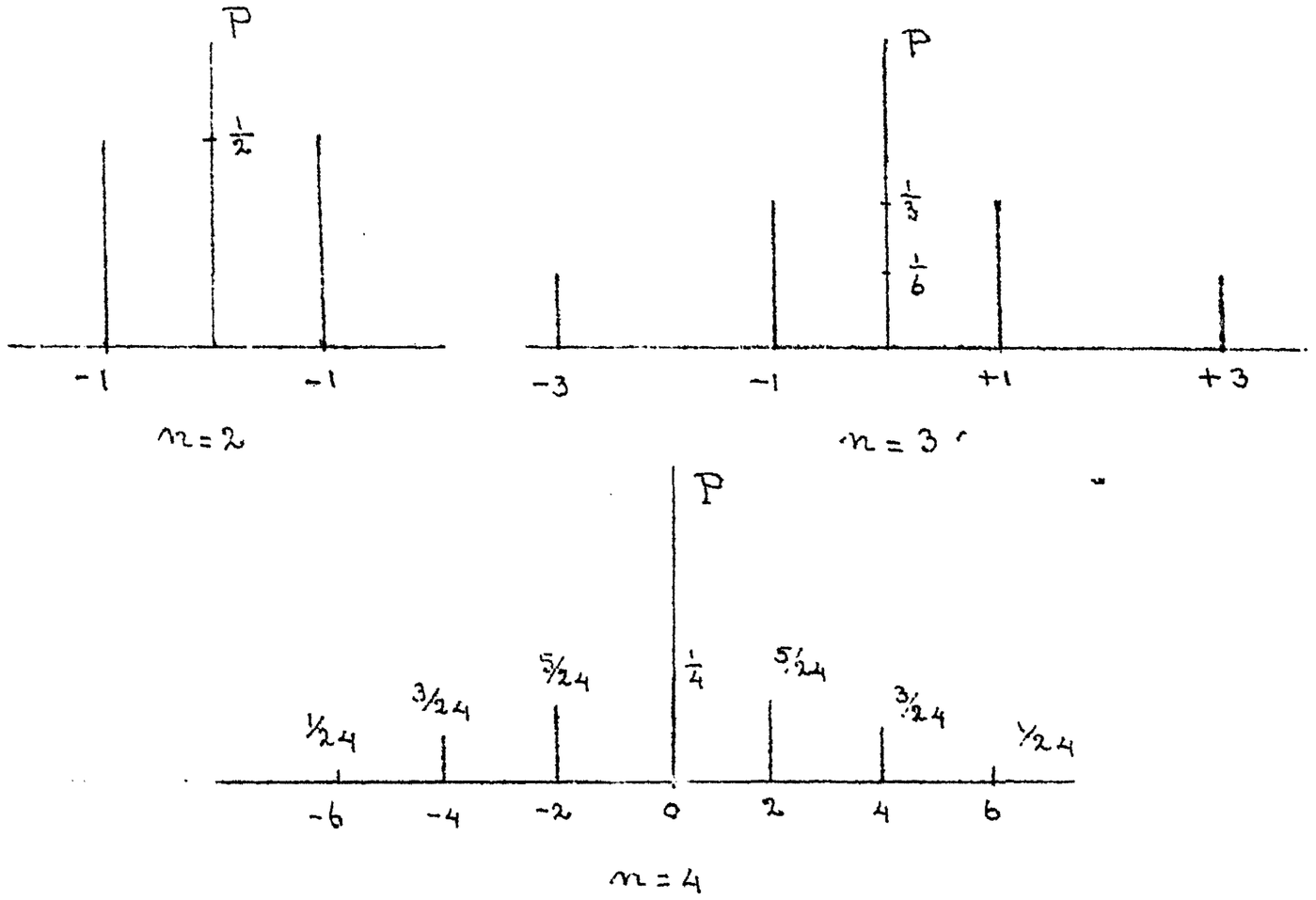
Bijlage 1.

Tabel van enkele quantilen van de waarschijnlijkheidsverdeling van S onder de nulhypothese H_o.

Deze tabel bevat de kleinste waarden van S, waarvoor geldt:

$$P\left[\underline{S} \geqq S \mid H_o \; ; n\right] \leqq \alpha .$$

n	α =0,005	α =0,01	α =0,025	α =0,05	α =0,10	n
4	-	-	-	6	6	4
5	-	10	10	8	8	5
6	15	13	13	11	9	6
7	17	17	15	13	11	7
8	22	20	18	16	12	8
9	26	24	20	18	14	9
10	29	27	23	21	17	10
11	33	31	27	23	19	11
12	38	34	30	26	20	12
13	42	40	34	28	24	13
14	47	43	37	31	25	14
15	53	47	41	35	29	15
16	58	52	46	38	30	16
17	62	58	50	42	34	17
18	67	63	53	45	35	18
19	73	67	57	49	39	19
20	80	72	62	52	42	20
21	84	78	66	56	44	21
22	91	83	71	61	47	22
23	97	89	75	65	51	23
24	104	94	80	68	54	24
25	110	100	86	72	58	25
26	115	105	91	77	61	26
27	123	111	95	81	63	27
28	130	118	100	86	68	28
29	136	124	106	90	70	29
30	143	131	111	95	73	30
31	151	137	117	99	77	31
32	158	144	122	104	82	32
33	164	150	128	108	84	33
34	171	157	133	113	89	34
35	179	163	139	117	93	35
36	188	170	144	122	96	36
37	194	178	150	128	100	37
38	203	185	157	133	105	38
39	211	191	163	137	109	39
40	218	200	170	142	112	40



figuur 1
Verdeling van S.

De exacte verdeling van S onder hypothese H_o is door Kendall berekend tot n= 10 toe en door de Rekenafdeling van het Mathematisch Centrum verder tot n= 40. Op bijlage 1 vindt men een gedeelte van de resultaten van deze berekeningen van de rekenafdeling in een tabel samengevat.

De verdeling van S bezit de volgende eigenschappen:

(a). De verdeling is symmetrisch ten opzichte van 0. Indien $\frac{1}{2}n(n-1)$ even resp. oneven is, kan S slechts even resp. oneven waarden aannemen.

(b). De verdeling nadert voor $n \to \infty$ tot een normale verdeling met gemiddelde 0 en spreidingskwadraat

(10). $$\sigma_S^2 = \frac{1}{18}n(n-1)(2n+5).$$

Wij bewijzen eerst de symmetrie van de verdeling. Daartoe denken wij ons rij A weer in de natuurlijke volgorde geplaatst. Bij iedere rangschikking van rij B beschouwen wij nu de tegenovergestelde rangschikking, die ontstaat, als we rij B achterstevoren leggen. Door dit achterstevoren leggen verandert S van teken, maar niet van grootte, daar het teken van de door een willekeurig paar geleverde bijdrage tot S van teken verandert. Hieruit volgt de symmetrie van de verdeling van S ten opzichte van 0.

De tweede onder (a) genoemde eigenschap volgt hieruit, dat iedere rangschikking van B uit iedere andere verkregen kan worden door achtereenvolgens verwisselen van paren elemen-

ten van B. Bij een dergelijke verwisseling verandert echter het teken van de door dat paar geleverde bijdrage, zodat S met 2 toe of afneemt.

Het bewijs van de onder (b) genoemde asymptotische normaliteit zullen wij niet geven, daar dit nogal ingewikkeld is. Wij geven wel een schets van de <u>afleiding van vergelijking (10) voor de spreiding</u>.

Hiertoe voeren wij de volgende notatie in:

1) a_{ij}= +1 resp. -1, indien het rangnummer, dat op de i^{e} plaats in rij A staat kleiner resp. groter is dan het op de j^{e} plaats staande rangnummer. (Staat rij A in de natuurlijke volgorde, dan is dus a_{ij}= 1 als $i < j$ is en a_{ij}= -1 als $i > j$ is). Voor i= j definiëren wij a_{ij}= 0.

2) Een analoge definitie geldt voor b_{ij} in verband met rij B.

3) Het symbool $\sum$, zonder aanduiding van sommatie-indices, geeft aan, dat over alle waarden der indices gesommeerd wordt.

4) Het symbool $\sum'$, zonder aanduiding van sommatie-indices, geeft aan, dat over alle indices gesommeerd wordt, maar dat daarbij alleen die termen genomen worden, waarin verschillende indices verschillende waarden bezitten.

5) Wij definiëren tenslotte:

$$c_{ij} = a_{ij} b_{ij} \tag{11}$$

en hebben dan

$$c = \sum c_{ij} = 2S, \tag{12}$$

zoals direct uit de definities van c_{ij} en S blijkt.

Houden wij rij A weer in de natuurlijke volgorde, dan geldt:

$$\sum_{j=1}^{n} a_{ij} = (n-i)-(i-1) = n+1-2i \tag{13}$$

en

$$\sum a_{ij}^2 = n(n-1), \tag{14}$$

daar

$$a_{ij}^2 = \begin{cases} 1 & \text{als } i \neq j \\ 0 & \text{als } i = j. \end{cases}$$

Verder is (wegens $\sum i^2 = \frac{1}{6} n(n+1)(2n+1)$):

$$\sum a_{ij} a_{ik} = \sum a_{ij} \sum_{k=1}^{n} a_{ik} = \sum a_{ij}(n+1-2i) = \sum (n+1-2i)^2 = \tag{15}$$

$$= \sum \{(n+1)^2 - 4i(n+1) + 4i^2\} = n(n+1)^2 - 4(n+1)\cdot\tfrac{1}{2}n(n+1) + \tfrac{2}{3}n(n+1)$$

$$(2n+1) = \tfrac{1}{3} n(n^2-1).$$

Geven we verder, met het symbool $\mathcal{E}$ het gemiddelde over alle n! permutaties van de B-rij aan, dan is, zoals boven bewezen is:

(16) $$\mathcal{E}\, \underline{c} = 0.$$

Hierbij valt op te merken, dat wij nu $\underline{c} = 2\underline{S}$ als een stochastische grootheid beschouwen, waarbij wij aan ieder der n! mogelijke permutaties van de B-rij de waarschijnlijkheid $\frac{1}{n!}$ toekennen (dit is precies de hypothese H_o). Ook de grootheden $\underline{b}_{ij}$ zijn nu stochastisch (zij nemen ieder de waarde -1 of +1 aan), maar daar de A-rij in de natuurlijke volgorde vastgehouden wordt, zijn de a_{ij} constant (vgl. de definitie).

Ter berekening van $\mathcal{E}\underline{c}^2$ splitsen wij eerst c^2 in een aantal termen:

$$(\sum c_{ij})^2 = \sum' c_{ij}\, c_{kl} + \sum' c_{ij}\, c_{ik} + \sum' c_{ji}\, c_{ki} + \sum' c_{ij}\, c_{ki} + \sum' c_{ji}\, c_{ik} + \sum c_{ij}\, c_{ij} + \sum c_{ij}\, c_{ji}.$$

Nu is echter

(17) $$c_{ij} = c_{ji},$$

zoals direct uit de definitie volgt, dus is:

$$\sum' c_{ij}\, c_{ik} = \sum' c_{ji}\, c_{ki} = \sum' c_{ij}\, c_{ki} = \sum' c_{ji}\, c_{ik}$$

en

$$\sum c_{ij}^2 = \sum c_{ij}\, c_{ji},$$

dus is:

(18) $$c^2 = \sum' c_{ij}\, c_{kl} + 4 \sum' c_{ij}\, c_{ik} + 2 \sum c_{ij}^2.$$

Derhalve is:

$$\mathcal{E}\, c^2 = \mathcal{E} \sum' a_{ij}\, a_{kl}\, \underline{b}_{ij}\, \underline{b}_{kl} + 4\, \mathcal{E} \sum' a_{ij}\, a_{ik}\, \underline{b}_{ij}\, \underline{b}_{ik} + 2\, \mathcal{E} \sum a_{ij}^2\, \underline{b}_{ij}^2 =$$

$$= \sum' a_{ij}\, a_{kl}\, \mathcal{E}\, \underline{b}_{ij}\, \underline{b}_{kl} + 4 \sum' a_{ij}\, a_{ik}\, \mathcal{E}\, \underline{b}_{ij}\, \underline{b}_{ik} + 2 \sum a_{ij}^2\, \mathcal{E}\, \underline{b}_{ij}^2.$$

Hierin moeten de grootheden $\mathcal{E}\, \underline{b}_{ij}\, \underline{b}_{kl}$, $\mathcal{E}\, \underline{b}_{ij}\, \underline{b}_{ik}$ en $\mathcal{E}\, \underline{b}_{ij}^2$ nog berekend worden. Daar hierbij het symbool $\mathcal{E}$ de betekenis heeft van $\frac{1}{n!}$ maal de som van de waarden, die $\underline{b}_{ij}^2$ bij alle permutaties aanneemt, is

(19) $$\mathcal{E}\, \underline{b}_{ij}^2 = 1, \text{ voor iedere i en j met } i \neq j, \text{ daar}$$

bij ieder der permutaties $b_{ij}^2 = (\pm 1)^2 = 1$ is.

Verder is:

(20) $$\mathcal{E}\, \underline{b}_{ij} \underline{b}_{kl} = 0, \quad \text{voor iedere i,j,k en l, als}$$

i,j,k en l verschillend zijn. Immers bij iedere permutatie is een andere aan te wijzen, waarbij het $i^{\underline{e}}$ en $j^{\underline{e}}$ rangnummer ver-

wisseld zijn, zodat de bijdragen tot de som elkaar, wegens het tegengestelde teken, opheffen.

De berekening van $\mathcal{E}\underline{b}_{ij}\underline{b}_{ik}$ is iets ingewikkelder. In plaats van het permuteren der rangnummers van rij B, en het in het oog houden van vaste plaatsen (nl. de $i^{\underline{e}}$, $j^{\underline{e}}$ en $k^{\underline{e}}$ plaats), kunnen wij ook op alle mogelijke wijzen 3 verschillende plaatsen aanwijzen en vervolgens de bijdragen van de op die plaatsen staande rangnummers als bijdragen tot $\mathcal{E}\underline{b}_{ij}\underline{b}_{ik}$ optellen en vervolgens delen door het aantal wijzen, waarop we deze 3 plaatsen aan kunnen wijzen. Het maakt echter geen verschil, of wij dit in de B-rij (in een vaste volgorde), dan wel in de A-rij doen, daar beide de rangnummers 1,...,n bevatten. De eerste plaats kunnen we op n wijzen kiezen, de tweede vervolgens nog op (n-1) wijzen en de derde op (n-2). Bij een bepaalde keuze krijgen wij dus als bijdrage tot $\mathcal{E}\underline{b}_{ij}\underline{b}_{ik}$, een term van de vorm:

$$\frac{1}{n(n-1)(n-2)}\, a_{\alpha\beta}\, a_{\alpha\gamma}\ ,$$

zodat wij vinden:

$$(21)\qquad \mathcal{E}\,\underline{b}_{ij}\,\underline{b}_{ik} = \frac{(n-3)!}{n!}\sum{}' a_{\alpha\beta}\, a_{\alpha\gamma}$$

voor iedere i,j en k, met $i \neq j \neq k \neq i$.

Derhalve wordt:

$$\mathcal{E}\,\underline{c}^2 = \frac{4(n-3)!}{n!}\left(\sum{}' a_{ij}\, a_{ik}\right)^2 + 2\sum a_{ij}^2 .$$

Nu is (vgl. (14) en (15)):

$$\sum{}' a_{ij}\, a_{ik} = \sum a_{ij}\, a_{ik} - \sum a_{ij}^2 = \tfrac{1}{3}n(n^2-1) - n(n-1),$$

dus

$$(22)\quad \mathcal{E}\,\underline{c}^2 = \frac{4(n-3)!}{n!}\left\{\tfrac{1}{3}n(n^2-1) - n(n-1)\right\}^2 + 2n(n-1) =$$

$$= \frac{4n}{9(n-1)(n-2)}\left(n^2-3n+2\right)^2 + 2n(n-1) =$$

$$= \frac{2n(n-1)(2n+5)}{9}\ .$$

Daar $\underline{c} = 2\underline{S}$ is en $\mathcal{E}\,\underline{S} = 0$ is, is dus:

$$(23)\quad \sigma_{\underline{S}}^2 = \mathcal{E}\,\underline{S}^2 = \frac{n(n-1)(2n+5)}{18}\ .$$

Wegens (3) is dus

$$(24)\qquad \sigma_{\underline{I}}^2 = \frac{4}{n^2(n-1)^2}\,\sigma_{\underline{S}}^2 = \frac{2(2n+5)}{9n(n-1)}\ .$$

§4. Toetsing van de nulhypothese.

Op grond van de resultaten van de vorige paragraaf kunnen wij nu de hypothese H_o toetsen aan een experimenteel gevonden resultaat. Wij zullen dit aan voorbeeld 2 toelichten (vgl. §1).

Stel, dat in volgorde van het tijdstip van ontvangst van het antwoord de volgende percentages worden opgegeven:

$7\frac{1}{2}$ $6\frac{1}{2}$ 6 8 $12\frac{1}{2}$ 4 $4\frac{1}{2}$ 7 $8\frac{1}{2}$ $5\frac{1}{2}$ 9 10 5 $10\frac{1}{2}$ $9\frac{1}{2}$,

zodat de twee rijen rangnummers zijn (A voor tijdstip van ontvangst en B voor het dividendpercentage):

A: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
B: 8 6 5 9 15 1 2 7 10 4 11 13 3 14 12 ,

dan is dus

$R = 7+8+8+6+0+9+8+5+4+5+3+1+2+0+0 = 65$

dus $S = 2 \times 65 - \frac{1}{2}.14.15 = 25$.

Wij zien in bijlage 1, dat de overschrijdingskans, behorende bij deze waarde, $> 0{,}10$ is, zodat er geen reden is, aan te nemen, dat de hoogte van het dividendpercentage invloed uitoefent op het tijdstip van de inzending van het antwoord, te meer niet, daar deze overschrijdingskans éénzijdig is, zodat de tweezijdige $> 0{,}20$ is.

Om ook de benaderingsmethode, berustende op de approximatie van de verdeling van S met een normale verdeling met gemiddelde 0 en spreiding (23), te illustreren, berekenen wij de overschrijdingskans ook op die wijze. Wij vervangen daarbij dus de discontinue verdeling van S door een normale verdeling met gemiddelde 0 en, in het onderhavige geval:

$$\sigma^2 = \frac{15.14.35}{18} = 408,$$

dus $\sigma = 20{,}2$. Derhalve is

$$(25) \qquad \frac{S}{\sigma_S} = \frac{25}{20{,}0} \approx 1{,}2$$

en de, in een tabel van de normale verdeling opgezochte, éénzijdige overschrijdingskans bedraagt ongeveer 0,12.

Continuïteitscorrectie. Bij de benadering van een discrete waarschijnlijkheidsverdeling door een continue past men vaak een zogenaamde "continuïteitscorrectie" toe, die daaruit bestaat, dat men niet de gevonden waarde van de toetsingsgrootheid gebruikt, maar het gemiddelde van deze waarde en de meest nabijliggende waarde aan de zijde van het gemiddelde. In dit

geval houdt deze correctie dus in, dat men de absolute waarde van de gevonden S met 1 vermindert, daar S alleen even of alleen oneven waarden aanneemt. Wij zullen hier niet op de achtergrond van deze correctie ingaan. In bovenstaand voorbeeld betekent het, dat wij in (25) S vervangen door S-1 (dus 25 door 24), hetgeen voor het resultaat van de toets geen verschil maakt.

§5 Gelijke rangnummers.

Ook het geval, dat er onder de rangnummers gelijke voorkomen, is door Kendall en andere onderzoekers beschouwd. Gelijke rangnummers (Engels:"ties") kunnen voorkomen: a. doordat bij een subjectieve beoordeling ener volgorde tussen twee of meer der beschouwde objecten geen onderscheid wordt gemaakt, en b. doordat onder een aantal quantitatieve waarnemingen gelijke optreden. Doordat de aan dit verschijnsel verbonden moeilijkheden grotendeels zijn opgelost, kan de methode der rangcorrelatie ook op discreet verdeelde stochastische grootheden en op gegroepeerde waarnemingen worden toegepast. In het volgende worden de resultaten der genoemde onderzoekingen zonder bewijs gegeven.

Indien aan twee of meer der objecten van één rij gelijke rangnummers worden toegekend, geeft men al deze objecten als rangnummer het gemiddelde van de rangnummers, die zij tezamen in beslag genomen zouden hebben.

Voorbeeld 3: Volgens de kwartaalverslagen van het Bureau van Statistiek der Gemeente Amsterdam zijn er in de verschillende maanden van de jaren 1948 en 1949 de volgende aantallen "Baldadige alarmeringen. waarop de brandweer is uitgerukt" geweest:

	Jan	Feb	Mrt	Apr	Mei	Juni	Juli	Aug	Sept	Oct	Nov	Dec
1948:	1	5	1	3	6	4	5	8	12	2	3	2
1949:	1	4	1	2	1	3	4	4	5	1	3	4

Bij rangschikking naar opklimmende grootte van dit aantal, zijn in 1948 de maanden Januari en Maart gelijk om krijgen, daar zij het kleinste aantal vertonen, beide het rangnummer $\frac{1}{2}(1+2)=1\frac{1}{2}$. October en December volgen, met rangnummer $\frac{1}{2}(3+4)=3\frac{1}{2}$, enz. In 1949 krijgen de maanden Januari, Maart, Mei en October alle vier het rangnummer $\frac{1}{4}(1+2+3+4)=$ $=\frac{1}{2}(1+4)=2\frac{1}{2}$, April krijgt rangnummer 5, enz.

Wij krijgen dan, als wij de maanden voor 1948 in opklimmende volgorde zetten:

	Jan	Mrt	Oct	Dec	Apr	Nov	Juni	Feb	Juli	Mei	Aug	Sep
1948:	1½	1½	3½	3½	5½	5½	7	8½	8½	10	11	12
1949:	2½	2½	2½	9½	5	6½	6½	9½	9½	2½	9½	12.

De waarde van S wordt nu op de volgende wijze berekend: indien van een tweetal onder elkaar staande paren één der paren rangnummers (of beide) uit twee gelijke rangnummers bestaat, is de bijdrage tot S gelijk aan 0. Dit is gebaseerd op de gedachte, dat men beide mogelijke rangschikkingen van de twee gelijken beschouwt en het gemiddelde van de dan verkregen bijdragen tot S als bijdrage neemt. Deze mogelijke rangschikkingen hebben echter steeds 0 als gemiddelde, daar steeds de ene een waarde +1 geeft en de tegengestelde de waarde -1.

Indien wij dus in ons voorbeeld S willen berekenen, geeft het paar Januari-Maart een bijdrage 0, Januari-October eveneens, Januari-December tot en met Januari-Juli geven +1, Januari-Mei weer 0; en Januari-Augustus en Januari-September +1. Vervolgens vergelijken wij Maart met alle erachterstaande maanden: October en Mei geven weer 0 en de overige +1. Zo gaan wij door; de eerste bijdrage -1, die wij ontmoeten is die van het paar December-April. Wij krijgen, bij deze successieve vergelijking van alle maanden met de rechts daarvan staande maanden:

Januari : +8
Maart : +8
October : +7
December : -4+1 = -3
April : +5-1 = +4
November : +4-1 = +3
Juni : +4-1 = +3
Februari : -1+1 = 0
Juli : -1+1 = 0
Mei : +2
Augustus : 1
September : 0,

zodat S = 33.

De waarschijnlijkheidsverdeling van S onder de nulhypothese H_o. inhoudende dat de beide rangschikkingen onafhankelijk van elkaar zijn, is nu weer van veel belang.

Het optreden der gelijke rangnummers verandert deze waarschijnlijkheidsverdeling echter. Zij is voor een aantal gevallen exact berekend door Sillitto [1]. Daar deze verdeling echter van de stelsels gelijken in ieder der rijen rangnummers afhangt, is een enigszins volledige tabellering nauwelijks uitvoerbaar. Sillitto geeft tabellen tot n = 10 toe, waarbij tweetallen en drietallen gelijken in beide rijen worden toegelaten. Daar wij in ons voorbeeld n=12 hebben, kunnen wij deze tabellen niet gebruiken en moeten wij volstaan met een benadering van de verdeling van S. De verdeling van S is, evenals wanneer er geen gelijken zijn, symmetrisch t.o.v. 0; dit kan op dezelfde wijze bewezen worden als voor het geval, dat er geen gelijke zijn. Het spreidingskwadraat van S wordt nu echter:

$$(26)\quad \sigma_S^2 = \frac{1}{18}\left\{n(n-1)(2n+5) - \sum_t t(t-1)(2t+5) - \sum_u u(u-1)(2u+5)\right\} +$$
$$+ \frac{1}{9n(n-1)(n-2)} \sum_t t(t-1)(t-2) \cdot \sum_u u(u-1)(u-2) +$$
$$+ \frac{1}{2n(n-1)} \sum_t t(t-1) \cdot \sum_u u(u-1),$$

waarin n de lengte der twee rijen aangeeft en t (resp. u) het aantal van ieder der stelsels gelijke rangnummers in de eerste (resp. tweede) rij voorstelt.

In ons voorbeeld neemt t dus de waarden 2, 2, 2, 1, 1 en 1 aan en u de waarden 4, 4, 1, 2, 1. De termen en factoren van (26) met t=1 of u=1 bezitten steeds de waarde 0, zodat formule (26), indien er geen gelijke rangnummers zijn, in (23) overgaat. De zojuist genoemde waarden van t en u geven tezamen met n=12:

$$\sum_t t(t-1)(2t+5) = 4(2.1.9) = 72.$$

$$\sum_u u(u-1)(2u+5) = 2.(4.3.13) + 2.1.9 = 330.$$

$$\sum_t t(t-1)(t-2) = 0.$$

$$\sum_t t(t-1) = 4.2 = 8 \quad \text{en} \quad \sum_u u(u-1) = 2.(4.3) + 2 = 26$$

dus

$$\sigma_S^2 = \frac{1}{18}\left\{12.11.29 - 72 - 330\right\} + \frac{8.26}{24.11} = 191$$

$$\sigma_S = 13,8.$$

---

1) G.P.Sillitto, The distribution of Kendall's coefficient of rank correlation in rankings containing ties, Biometrika 4 (1947) p. 36.

Daar men verder kan bewijzen, dat de verdeling van S, indien er geen al te lange series gelijken voorkomen, voor n> 10 vrij goed door de aangepaste normale verdeling benaderd wordt, kunnen wij de bij de waarde S=33 behorende overschrijdingskans (met continuïteitscorrectie), berekenen door de bij de grootheid

$$\frac{S-1}{\sigma_S} = \frac{32}{13,8} = 2,32$$

behorende overschrijdingskans in een tabel der normale verdeling (met gemiddelde 0 en spreiding 1) op te zoeken. Wij vinden dan voor de tweezijdige overschrijdingskans de waarde 0,02, hetgeen wijst op een significante overeenstemming tussen de beide rijen.

Opmerkingen: 1. Een waarschuwing.

Het is niet overbodig aan dit resultaat een waarschuwing toe te voegen. Men zou nl. geneigd zijn te concluderen, dat de baldadigheid, of het goede vertrouwen van de brandweer (of beide) een periodiciteit vertonen. Immers, de toegepaste toets heeft bij dit voorbeeld het karakter van een periodiciteitstoets, met een gegeven periode van 1 jaar. Het voorbeeld werd echter gekozen, toen bij de beschouwing van de twee getallen reeksen de overeenkomst opviel. Deze overeenkomst werd vervolgens op bovenstaande wijze getoest. Dit kiezen van de te toetsen hypothese en de toetsingsmethode naar aanleiding van een gevonden resultaat is echter zeer riskant, daar men dan feitelijk geen idee meer heeft van de onbetrouwbaarheid van de toets, die echter zeker niet meer gegeven wordt door de gebruikte onbetrouwbaarheidsdrempel. Dit is niet bezwaarlijk, indien men de methode slechts gebruikt om aanwijzingen op te sporen voor verder onderzoek, dat deze aanwijzingen bevestigen of weerleggen kan. Men dient echter zeer voorzichtig te zijn, zolang dit verder onderzoek niet plaats gevonden heeft. Wat de "baldadige alarmeringen" betreft, zijn mij nog niet alle gegevens van 1950 bekend. De maanden Januari en Maart bezaten echter van de eerste 6 maanden het hoogste aantal (nl. 5) dergelijke alarmeringen, in plaats van het laagste. Het is dus zeer wel mogelijk, dat de tussen 1948 en 1949 geconstateerde overeenstemming een toevallige zal blijken te zijn geweest.

2. In een geval als het gebruikte voorbeeld wordt S beschouwd als een coëfficiënt van overeenstemming tussen twee rijen rangnummers, waarbij het optreden van gelijke rangnummers op natuurlijke wijze optreedt. Men zal dus t weer zodanig

wensen te definiëren, dat $\tau = 1$ wordt, indien de rijen volledig overeenstemmen, ook wat betreft de gelijken. Kendall geeft als algemene vorm van de correlatiecoëfficiënt de formule:

$$\tau = \frac{\sum a_{ij} b_{ij}}{\sqrt{\sum a_{ij}^2 \cdot \sum b_{ij}^2}} \qquad (27)$$

waarin a_{ij} en b_{ij} de in §3 (blz.7) gedefiniëerde grootheden zijn. Inderdaad wordt dan $\tau = 1$, indien de rijen volledig overeenstemmen en -1, indien zij geheel tegengesteld zijn, ook indien gelijken in beide rijen optreden, mits overeenkomstige paren in beide rijen gelijk zijn (dus als $a_{ij}=0$ en $b_{ij}=0$ steeds gelijk optreden). Zijn er geen gelijken, dan gaat (27) over in formule (3) van blz. 4, de oude definitie van τ. Worden de aantallen gelijken in de beide rijen weer aangegeven met t (voor de eerste) en met u (voor de tweede rij), dan kan men (27) ook schrijven als:

$$\tau = \frac{2S}{\sqrt{\{n(n-1)-2T\}\{n(n-1)-2U\}}} \qquad (28)$$

waarin

$$2T = \sum_t t(t-1) \qquad (29)$$

en

$$2U = \sum_u u(u-1) \qquad (30)$$

is. In ons voorbeeld was S=33, n=12, 2T=8 en 2U=26, dus wordt

$$\tau = \frac{66}{(11.12-8)(11.12-26)} = 0,58.$$

Beschouwen wij echter voorbeeld 1 (zie pag. 1), waarbij een objectieve volgorde bestaat zonder gelijken, dan is het optreden van gelijken te wijten aan een gebrek aan onderscheidingsvermogen van de proefpersoon. Kendall raadt aan, om in dat geval voor τ niet formule (28) te gebruiken, maar formule (3), omdat dit gebrek aan onderscheidingsvermogen dan tot uitdrukking komt in een vermindering van de waarde van τ. Tevens kan men dan de gevonden waarde van τ beschouwen als de gemiddelde waarde over alle mogelijke rangschikkingen, die men aan de gelijken zou kunnen toekennen (zonder de volgorde ten opzichte van andere rangnummers te veranderen). Wij moeten er echter op wijzen, dat de boven beschreven toets overeenkomt met het gebruik van de in formule (28) gedefiniëerde τ, daar deze toets in feite een voorwaardelijke toets is, met als voorwaarde het optreden van zoveel gelijken in beide rijen, als bij het experiment gevonden zijn. Een toets, die, ook wanneer gelijke rangnummers optreden,

overeenkomt met het gebruik van formule (3) is niet bekend en zal ook moeilijk te vinden zijn, daar dan de kans op het optreden van gelijken bekend zou moeten zijn. Dit is voor de boven beschreven toets niet nodig.

§6. Enkele verdere resultaten.

1. Indien uit een eindige collectie van N paren (x_i, y_i), waarvoor de rangcorrelatiecoëfficiënt tussen de rij der waarden x_i en die der y_i de waarde τ bezit, een steekproef wordt genomen zonder teruglegging en van deze steekproef van getallenparen de rangcorrelatiecoëfficiënt gelijk aan $\underline{t}$ is ($\underline{t}$ neemt voor verschillende steekproeven verschillende waarden aan en heeft dus een waarschijnlijkheidsverdeling), dan geldt:

$$\sigma_{\underline{t}}^2 \leqq \frac{2}{n}(1-\tau^2), \qquad (31)$$ [1]

waarin n de uitgebreidheid van de steekproef is. Met behulp hiervan kan men het verschil tussen twee gevonden rangcorrelatiecoëfficiënten op significantie toetsen, terwijl men met behulp van één steekproefwaarde t, een betrouwbaarheidsinterval voor τ kan construeren. Daar (31) een ongelijkheid is en de (in het algemeen onbekende) collectie-grootheid τ in het rechterlid voorkomt, is de toets noodgedwongen van een ietwat primitief en weinig scherp karakter, terwijl het betrouwbaarheidsinterval wijder is dan het geval zou zijn, indien meer over de verdeling van $\underline{t}$ bekend was. Men kan weliswaar de normale verdeling als een benadering van de verdeling van $\underline{t}$ (gebruiken), daar deze laatste voor n (en N) $\to \infty$ tot een normale verdeling nadert, maar voor de spreiding van deze benaderende normale verdeling kent men slechts de door (31) gegeven bovengrens.

2. Indien men een steekproef $(x_1,y_1),\ldots,(x_n,y_n)$ uit een tweedimensionale normale verdeling neemt, en van deze steekproef de rangcorrelatiecoëfficiënt $\underline{t}$ berekt, kan men de grootheid

$$r_{\underline{t}}' = \sin \frac{\pi \underline{t}}{2} \qquad (32)$$

als schatting van de (gewone) correlatiecoëfficiënt ρ gebruiken.

---

[1] Deze grootheid $\underline{t}$ wèl te onderscheiden van de in §5 met dezelfde letter aangegeven aantallen gelijken in de eerste rij. Nu gebruiken we de t als de met τ overeenkomende latijnse letter.

Van de eigenschappen van deze schatting is slechts weinig bekend. Ter vergelijking met de gebruikelijke schatting $\underline{r}$ van ρ (met $\underline{r}$ wordt de gewone correlatiecoëfficiënt van de steekproef bedoeld) vermelden wij slechts, dat voor grote steekproeven bij benadering geldt:

$$(33)\quad \frac{\sigma_{\underline{r}}^2}{\sigma_{\underline{r}^2}} = \begin{cases} 0{,}43 & \text{als } \rho = 0 \text{ is,} \\ 0{,}28 & \text{als } \rho = 0{,}9 \text{ is.} \end{cases}$$

Dit geeft enige indruk van het verlies aan precisie, dat optreedt, indien men de normaliteit van een verdeling, als deze aanwezig is, verwaarloost. Dit verlies is veel geringer voor $\rho = 0$ dan voor grotere ρ .

Tenslotte vermelden wij nog dat de later in deze cursus te behandelen toets van Wilcoxon voor het probleem van twee steekproeven berust op een grootheid, die op eenvoudige wijze uit de rangcorrelatiecoëfficiënt kan worden afgeleid, zodat deze toets in zekere zin als een toepassing van de hier behandelde theorie kan worden gezien. Voor dit speciale geval is de theorie echter door andere onderzoekers onafhankelijk van de theorie der rangcorrelatie ontwikkeld.

MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 59

Cursus Parametervrije Methoden

II. Methode van de m rangschikkingen

door
Ph.van Elteren

April 1951

1. Inleiding.

1.1. Definitie van S en W.

Voor een eerste orientatie in de methode gaan wij uit van een voorbeeld. Wij stellen ons voor een klasse met 5 leerlingen, genaamd a, b, c, d en e. Aan drie leraren A, B en C wordt gevraagd hun oordeel over de capaciteiten van de leerlingen uit te drukken in rangnummers. Aan de hand van het verkregen resultaat willen wij vaststellen, of de leraren het in grote lijnen eens zijn in hun oordeel, dan wel uiteenlopende opinies hebben.

Het volgende schema vertoont volkomen eensgezindheid tussen de leraren

(II.1)

	a	b	c	d	e
A	5	3	1	2	4
B	5	3	1	2	4
C	5	3	1	2	4
	15	9	3	6	12

.

Wij zien, dat de kolomtotalen sterk uiteenlopen. Het is ook mogelijk, dat de leraren minder eensgezind zijn, zoals bijvoorbeeld in het volgende geval:

(II.2)

	a	b	c	d	e
A	5	3	1	2	4
B	3	1	4	5	2
C	5	4	2	1	3
	13	8	7	8	9

.

Het blijkt, dat de kolomtotalen veel dichter bij elkaar liggen, dan in schema (II.1).

---

1) De inhoud van deze voordracht is grotendeels ontleend aan M.G. Kendall (1948) "Rank Correlation methods", hoofdstukken 6 en 7.

Wij komen er zodoende toe een maat te zoeken voor de eensgezindheid van de leraren, die afhangt van de "variatie" in de kolomtotalen. Zo een maat verkrijgen wij, door de kolomtotalen te verminderen met hun gemiddelde - in dit geval: 9 - de aldus verkregen gereduceerde kolomtotalen te quadrateren en daarna op te tellen. Deze maat duiden we verder aan met de letter S.

Wij vinden nu bij schema (II.1):

$$S = (15-9)^2 + (9-9)^2 + (3-9)^2 + (6-9)^2 + (12-9)^2 =$$
$$= 6^2 + 0 + (-6)^2 + (-3)^2 + 3^2 = 90,$$

en bij schema (II.2):

$$S = 4^2 + 1^2 + 2^2 + 1^2 + 0 = 22.$$

Indien men de maximale waarde $S_{max.}$ van S kent, die bij het gegeven schema kan voorkomen, dan geeft de werkelijk gevonden S een indruk van de overeenstemming. Indien de gevonden S dicht bij $S_{max.}$ ligt, is de overeenstemming goed, ligt zij er ver van af, dan is de overeenstemming slecht. Het ligt daarom voor de hand een maat in te voeren, die gelijk is aan de verhouding van S en $S_{max.}$. Deze maat wordt de overeenstemmingscoëfficiënt genoemd, en aangeduid met de letter W. De grootheid W kan dan varieren tussen 0 en 1. De overeenstemming zal beter zijn, naarmate W dichter bij 1 ligt.

De S_{max} behorende bij schema(II 2) is gelijk aan de S, die we gevonden hebben bij schema (II.1) terwijl de S van schema(II 2) gelijk was aan 22; de overeenstemmingscoëfficiënt voor schema(II 2) is dus: $W = \frac{22}{90} = 0,24$.

Wij zullen ons herinneren, dat bij de "gewone" rangcorrelatie, waarbij slechts twee rangschikkingen voorkwamen, een rangcorrelatiecoëfficiënt τ werd gedefiniëerd, die varieerde tussen -1 en +1. Deze coëfficiënt was +1 als de rangschikkingen volkomen identiek waren en was -1 als zij tegengesteld waren. Drie of meer volkomen tegengestelde rangschikkingen zijn niet mogelijk; dit verschil met de gewone rangcorrelatie komt tot uiting in het feit, dat de hier gedefinieerde overeenstemmingscoëfficiënt varieert tussen 0 en 1, in plaats van tussen -1 en +1.

1.2. Principe van de toetsingsmethode.

Hoewel de grootheid W een plausibele maat is voor de overeenstemming, hebben wij met de definitie van deze grootheid, ons probleem nog niet opgelost. Wij zullen aan iedere gevonden waarde van W of S een objectieve interpretatie moeten geven. Hiertoe gaan wij met de statistische grootheden W of S de hypothese H_o toetsen, dat in iedere

rangschikking alle permutaties van de rangnummers even waarschijnlijk zijn, terwijl de rangschikkingen onderling onafhankelijk zijn. In het voorbeeld van 1.1 zou H_o inhouden, dat de leraren de rangnummers door loting hadden toegekend, of dat alle leerlingen in feite even begaafd zijn, zodat het door de leraren gemaakte onderscheid op het toeval berust.

Uitgaande van de hypothese H_o bezitten W en S een waarschijnlijkheidsverdeling, zodat wij deze letters nu onderstrepen en wij nu de kans berekenen, dat de gevonden waarde van W of S of een grotere waarde optreedt. Is deze kans groot, dan is gebleken, dat de uitkomst van het onderzoek aan het toeval kan worden toegeschreven, zodat H_o niet verworpen behoeft te worden. Is de kans echter klein, dan zal men niet geneigd zijn, om het resultaat aan toevalsfactoren toe te schrijven; wij spreken dan van een significante overeenstemming (H_o wordt dan wèl verworpen). Als grens tussen significant en niet significant kiest men zeer vaak een kans 0,05 of 0,01. Deze grens noemen wij de onbetrouwbaarheidsdrempel.

Bij de berekeneing van de waarschijnlijkheidsverdeling van W of S, dienen wij te bedenken, dat de waarde van deze grootheden niet verandert, als wij twee kolommen verwisselen. Dit betekent, dat wij de rangschikking in bijvoorbeeld de eerste rij onveranderd kunnen laten en slechts hebben te letten op de permutaties, die mogelijk zijn in de overige rijen. Desondanks zijn er in ons eenvoudige schema (II.1) nog: $(5!)^2 = 14400$ mogelijkheden. Men is er in geslaagd, om voor dit kleine schema, de verdeling van S exact te berekenen, volgens een verderop te behandelen methode. Bij schema's van iets grotere omvang, neemt men zijn toevlucht tot nog te bespreken benaderingsmethoden. Hieronder volgt een uittreksel uit de tabel, welke in Kendall (1948)[1] voor dit schema (3 rijen, 5 kolommen) wordt vermeld:

---

1) Boeken, artikelen en tabellen vermeld in de litteratuurlijst, worden in de tekst aangeduid met de naam van de schrijver en het jaar van verschijnen.

Tabel II.1

Verdeling van S onder hypothese H_o, voor m=3, n=5.

S	$P[\underline{S} = S]$	$P[\underline{S} \geqq S]$
0	0,000	1,
2	0,012	1,000
4	0,016	0,988
6	0,031	0,972
8	0,027	0,941
10	0,069	0,914
12	0,014	0,845
⋮	⋮	⋮
22	0,054	0,649
⋮	⋮	⋮
62	0,011	0,056
64	0,007	0,038
⋮	⋮	⋮
74	0,007	0,015
76	0,003	0,008
⋮	⋮	⋮
86	0,00083	0,00090
90	0,00007	0,00007

De eerste kolom bevat de waarden die $\underline{S}$ kan aannemen. De tweede kolom bevat de kansen, dat $\underline{S}$ ieder van die waarden aanneemt, (indien H_o geldt), d.w.z. het aantal mogelijkheden, waarbij $\underline{S}$ gelijk is aan een gegeven waarde S, gedeeld door het totaal aantal mogelijkheden (14400). De derde kolom wordt uit de tweede verkregen, door de getallen van de tweede kolom te sommeren vanaf de onderste regel tot en met de regel die men beschouwt. Deze bevat dus voor iedere mogelijke S de kans (onder hypothese H_o), dat $\underline{S}$ de waarde S bereikt of overschrijdt, deze kans noemen wij overschrijdingskans van $\underline{S}$.

Aan de hand van de tabel constateren wij:

1) $\underline{S}$ kan in dit geval alléén even waarden aannemen.
2) De verdeling van $\underline{S}$ heeft een zeer onregelmatig karakter.
3) De bij schema (II.2) gevonden waarde van $\underline{S}$ is geenszins significant. (Overschrijdingskans 0,65).
4) Om de hypothese H_o met een onbetrouwbaarheidsdrempel 0,05 te kunnen verwerpen, had men een S minstens gelijk aan 64 moeten vinden; pas als $S \geqq 76$ kan men verwerpen met een onbetrouwbaarheidsdrempel 0,01.

2. De exacte verdeling van S.

2.1. Wiskundige formules voor S en W.

Alvorens iets naders te kunnen vertellen over de verdeling van S, dienen wij de definities van S en W in een algemeen geldende mathematische vorm te gieten.

Wij stellen ons voor een schema met m rijen en n kolommen. De rangnummers vervangen wij door de letter a met twee indices, een rij- en een kolom-index; a_{ij} is dus het rangnummer dat staat in de i-de rij en de j-de kolom, hierbij kan i gelijk zijn aan 1,2,...,m en j aan 1,2,...,n. Het schema heeft dan de volgende gedaante:

(II.3)
$$\begin{array}{cccc} a_{11} & a_{12} & \cdots\cdots & a_{1n} \\ a_{21} & a_{22} & \cdots\cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots\cdots & a_{mn} \end{array}$$

In het geval van totale overeenstemming wordt het schema:

$$\begin{array}{cccccc} 1 & 2 & 3 & \cdots\cdots & n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 2 & 3 & \cdots\cdots & n \\ \hline 1m & 2m & 3m & \cdots\cdots & nm \end{array}$$

Het gemiddelde van de kolomtotalen is dan gelijk aan m x het gemiddelde $\bar{j}$ van de getallen 1,2,...,n.

Er geldt:

(II.5) $$\bar{j} = \frac{1}{n}(1 + 2 + 3 + \ldots + n) = \frac{1}{2}(n+1),$$

en

(II.6) $$S_{max} = (m-mj)^2 + (2m - mj)^2 + \ldots + (nm - mj)^2 =$$
$$= m^2\left\{(1-j)^2 + (2-j)^2 + \ldots + (n-j)^2\right\} =$$
$$= m^2 \sum_{j=1}^{n} (j - \bar{j})^2 .$$

Wij maken nu gebruik van de algemene stelling: Als

(II.7) $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is, dan geldt:

(II.8) $$\sum_{i=1}^{n}(x_i - x)^2 = \sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2 .$$

Dan volgt uit (II.6):

$$S_{max} = m^2 \sum_{j=1}^{n} j^2 - n\bar{j}^2 = \tag{II.9}$$

$$= m^2\left\{ \frac{1}{6} n(n+1)(2n+1) - \frac{1}{4}(n+1)^2\right\} =$$

$$= \frac{1}{12} m^2 n(n^2-1) .$$

Indien er geen totale overeenstemming is, zal men voor de berekening van S eveneens gebruik kunnen maken van formule (II.8). In het algemeen is het dan echter eenvoudiger om eerst de gereduceerde kolomtotalen te bepalen.

Men kan S eenvoudig in formule-vorm brengen, indien men alle rangnummers vermindert met hun gemiddelde, $\frac{1}{2}(n+1)$. Deze gereduceerde rangnummers stellen wij voor door x_{ij}, zodat dus geldt:

$$x_{ij} = a_{ij} - \tfrac{1}{2}(n+1) . \tag{II.10}$$

Deze gereduceerde rangnummers vormen het schema:

$$\begin{array}{cccc} x_{11} & x_{12} & \cdots\cdots & x_{1n} \\ x_{21} & x_{22} & \cdots\cdots & x_{2n} \\ & & & \\ \underline{x_{m1}} & \underline{x_{m2}} & \cdots\cdots & \underline{x_{mn}} \\ s_1 & s_2 & & s_n \end{array} \tag{II.11}$$

De kolomtotalen $s_1, s_2, \ldots, s_n$ van dit schema zijn de gereduceerde kolomtotalen van schema (II.3), zodat:

$$S = s_1^2 + s_2^2 + \cdots\cdots + s_n^2 .$$

Dus als:

$$s_j = \sum_{i=1}^{m} x_{ij} \quad \text{voor} \quad j = 1,2,\ldots,n, \tag{II.12}$$

dan

$$S = \sum_{j=1}^{n} s_j^2 . \tag{II.13}$$

In verband met de definitie van W en vergelijking (II.9) geldt:

$$W = \frac{S}{S_{max}} = \frac{12\,S}{m^2 n(n^2-1)} \tag{II.14}$$

Voorbeeld:

Bij het rangnummerschema:

vindt men met behulp van formule (II.8):

$$S = 6^2 + 4^2 + 8^2 + 15^2 + 20^2 + 17^2 - 7 \times 12^2 = 218$$

en met behulp van de gereduceerde rangnummers:

-3	-2	-1	0	+1	+2	+3
-2	-3	-1	+1	0	+3	+2
-1	-3	-2	+1	+2	+3	0
-6	-8	-4	+2	+3	+8	+5

Dus is:

$$S = 6^2 + 8^2 + 4^2 + 2^2 + 3^2 + 8^2 + 5^2 = 218$$

Verder geldt:

$$S_{max} = \frac{1}{12} \times 3^2 \times 7 \times (7^2 - 1) = 252$$

dus is:

$$W = \frac{S}{S_{max}} = \frac{218}{252} = 0{,}865.$$

2.2. Algemene opmerkingen over de verdeling van S.

Het gemiddelde van de rangnummers is gelijk aan $\frac{1}{2}(n+1)$, en is dus geheel als n oneven is, en een geheel getal $+\frac{1}{2}$ als n even is. In verband met betrekking (II.10) geldt ditzelfde ook voor alle x_{ij}. Daaruit volgt dat de s_j hetzij alle geheel zijn, hetzij alle een geheel getal $+\frac{1}{2}$ zijn.

Wij veronderstellen nu, dat in de i-de rij $x_{ij} = a$ en $x_{ik} = b$ is en gaan nu beide gereduceerde rangnummers verwisselen. Dan veranderen de gereduceerde kolomtotalen s_j en s_k. En wel gaat s_j over in:

(II.14) $$s_j' = s_j + b - a$$

en s_k gaat over in:

(II.15) $$s_k' = s_k + a - b,$$

terwijl alle andere gereduceerde kolomtotalen bij deze verwisseling ongewijzigd blijven.

De wijziging die in S optreedt is gelijk aan:

(II.16) $$S' - S = (s_j')^2 + (s_k')^2 - s_j^2 - s_k^2 = (s_j+b-a)^2+(s_k+a-b)^2-$$

$$- s_j^2 - s_k^2 = 2(s_k-s_j)(a-b)+ 2(a-b)^2 .$$

Aangezien uit het voorafgaande blijkt, dat zowel a-b als $s_k - s_j$ gehele getallen zijn, is S'-S een <u>even</u> getal.

Aangezien alle mogelijkheden kunnen worden verkregen door een reeks van rangnummers - verwisselingen, volgt hieruit, dat de opeenvolgende mogelijke waarden van S, steeds een <u>even</u> getal verschillen. Wij constateerden dit reeds bij ons voorbeeld in §1.

Gebruikmakend van het bovenstaande en van formule (II.9), kunnen wij gemakkelijk aantonen, dat geldt:

Alle S-waarden zijn <u>even</u> als:

1e. m <u>even</u> is, of

2e. m en n beide <u>oneven</u> zijn, of

3e. m <u>oneven</u> en n een <u>3-voud</u> is. (bv. m = 3, n = 8).

Alle S-waarden zijn <u>oneven</u> als:

m <u>oneven</u> en n een 4-voud doch geen 8-voud is. (bv. m = 3, n = 4).

Alle S-waarden zijn een <u>geheel getal $+\frac{1}{2}$</u> als:

m <u>oneven</u> en n <u>even</u> doch geen 4-voud is. (bv. m = 3, n = 6).

<u>2.3</u>. <u>Methode ter berekening van de exacte verdeling van S</u>.

Wij gaan uit van een zeer eenvoudig schema met 2 rijen en 3 kolommen. Bij gebruik van gereduceerde rangnummers wordt het geval van maximale overeenstemming:

$$\begin{array}{ccc} -1 & 0 & 1 \\ -1 & 0 & 1 \\ \hline -2 & 0 & 2 \end{array} \qquad S_{max} = (-2)^2 + 2^2 = 8 \ .$$

Wij gaan nu de elementen van de 2-de rij permuteren en verkrijgen dan de volgende mogelijkheden

(II.13)

-2	0	2	S = 8
-2	1	1	S = 6
-1	-1	2	S = 6
-1	1	0	S = 2
0	-1	1	S = 2
0	0	0	S = 0 .

Onder de hypothese H_o, wordt dus de verdeling van S gegeven door de volgende tabel:

	S	$P[\underline{S}=S]$		$P[\underline{S} \geqq S]$
	0....	$\frac{1}{6}$	....	1
(II.19)	2....	$\frac{2}{6}$	....	$\frac{5}{6}$
	6....	$\frac{2}{6}$	....	$\frac{3}{6}$
	....	$\frac{2}{6}$	....	$\frac{1}{6}$

Wij voegen nu een derde rij toe. Indien de twee oorspronkelijke rijen samen de (gereduceerde) kolomtotalen -2, 0, 2 geven, ontstaan door toevoeging van de 3de rij de mogelijkheden:

	-3	0	3	S=18
	-3	1	2	S=14
(II.20)	-2	-1	3	
	-2	1	1	S=6
	-1	-1	2	
	-1	0	1	S=2

Gaan we evenzo te werk uitgaande van de andere mogelijkheden van (II.18), dan verkrijgen we tenslotte de volgende tabel voor de verdeling van $\underline{S}$, onder de hypothese H_o:

Tabel (II.2)

8	$P[\underline{S}=S]$	$P[\underline{S} \geqq S]$
0	$\frac{2}{36}$	1
2	$\frac{15}{36}$	$\frac{34}{36}$
6	$\frac{6}{36}$	$\frac{19}{36}$
8	$\frac{6}{36}$	$\frac{13}{36}$
14	$\frac{6}{36}$	$\frac{7}{36}$
18	$\frac{1}{36}$	$\frac{1}{36}$

Door toevoeging van nieuwe rijen kan men aldus in principe de verdeling van $\underline{S}$ bepalen als n=3 en m=4,5, 6 enz.(Als n=2 komt deze methode overeen met de zogenaamde tekentoets, die later in deze cursus behandeld zal worden tezamen met de symmetrietoetsen).

Het aantal mogelijkheden neemt echter snel toe, zodat de methode in de practijk reeds voor vrij kleine m onuitvoerbaar wordt. Dit geldt nog sterker, indien men niet uitgaat van n=3 maar van n=4 of meer. In M.G.Kendall (1948) komen tabellen voor van de exacte verdeling van $\underline{S}$ voor de gevallen

n=3 m=2,3,...,10
n=4 m=2,3,4,5,6
n=5 m=3

Men kan gemakkelijk het volgende aantonen:

1^o Bij een gegeven <u>even</u> aantal kolommen n, zullen alle S-waarden die optreden bij een schema met een bepaald aantal rijen m_o, ook optreden bij alle schema's met een even

aantal rijen meer.

2° Bij een gegeven <u>oneven</u> aantal kolommen n, zullen alle S-waarden, die optreden bij een schema met een bepaald aantal rijen m_o, ook optreden bij alle schema's met meer rijen.

<u>Bewijs</u>: 1° Indien bij $m=m_o$ de gereduceerde kolomtotalen zijn:

$$S_1 \quad S_2 \quad S_3 \ldots\ldots\ldots S_n$$

vindt men dezelfde kolomtotalen na toevoeging van de volgende twee rijen gereduceerde rangnummers:

$$-\frac{n-1}{2};\quad -\frac{n-3}{2};\quad -\frac{n-5}{2}\ldots\ldots\ldots;\quad +\frac{n-1}{2}$$
$$+\frac{n-1}{2};\quad +\frac{n-3}{2};\quad +\frac{n-5}{2}\ldots\ldots\ldots;\quad -\frac{n-1}{2}\ .$$

2° Men kan in een schema met m rijen de kolommen zodanig permuteren, dat de volgorde der gereduceerde rangnummers in de laatste rij wordt:

(a) $-\frac{n-1}{2}; -\frac{n-5}{2}; -\frac{n-9}{2}; \ldots\ldots\ldots; +\frac{n-1}{2}; -\frac{n-3}{2}; -\frac{n-7}{2}; \ldots\ldots\ldots; +\frac{n-3}{2}$.

Indien men deze laatste <u>rij</u> vervangt door:

(b) $-\frac{n-1}{2};\quad -\frac{n-3}{2};\quad -\frac{n-5}{2};\ldots\ldots\ldots;\ 0\ ;\ +1;\ +2;\ \ldots\ldots;+\frac{n-1}{2}$.

en toevoegt de rij:

(c) $0;\quad +1;\quad +2;\ldots\ldots\ldots;\quad +\frac{n-1}{2};\quad -\frac{n-1}{2};\quad -\frac{n-3}{2};\ldots\ldots;\quad -1,$

verkrijgt men een schema met m+1 rijen, waarvan de kolomtotalen hetzelfde zijn als van het oorspronkelijk schema met m rijen. Immers door optelling van de overeenkomstige elementen uit de rijen (b) en (c) verkrijgt men de rij (a) in de juiste volgorde.

§ 3. <u>Rangschikkingen met gelijken</u>.

<u>3.1. Inleiding</u>.

In § 1 hebben wij stilzwijgend ondersteld, dat iedere leraar aan elk van de leerlingen verschillende rangnummers toekende. Het is echter zeer wel mogelijk, dat een leraar niet tot een besluit kan komen, wie van twee of meerdere leerlingen de beste is en hen gelijke rangnummers wil geven. Wil men in een dergelijk geval de methode der m-rangschikkingen toepassen, dan dient men de gelijke rangnummers zo te kiezen, dat het gemiddelde van alle rangnummers onveranderd blijft. Indien b.v. de drie beste leerlingen van een klasse hetzelfde rangnummer moeten hebben, zullen wij hiervoor het rangnummer 2 kiezen.

Wij kiezen als voorbeeld het volgende schema:

(II,21)

	a	b	c	d	e
A	$1\frac{1}{2}$	$1\frac{1}{2}$	3	4	5
B	1	2	4	5	3
C	3	3	3	1	5
	$5\frac{1}{2}$	$6\frac{1}{2}$	10	10	13

$S = (3\frac{1}{2})^2 + (2\frac{1}{2})^2 + 1^2 + 1^2 + 4^2 = 36$.

Ook voor rangschikkingen met gelijke rangnummers definiëren wij een coëfficiënt van overeenstemming W. De definitie van W is in dit geval echter enigszins anders, hier geldt nl.:

(II.22) $\quad W = \frac{S}{mS'}$,

waarin

S= som van de quadraten der gereduceerde kolomtotalen.
m= aantal rijen, en
S'= som van de quadraten van de gereduceerde rangnummers in alle <u>rijen</u> is.

S' is een grootheid, die niet verandert als de rangnummers gepermuteerd worden.

Voor schema (II,21) geldt:

$S' = 2x(1\frac{1}{2})^2 + 1^2 + 2^2 + 2(2^2+1^2) + 2x2^2 = 27\frac{1}{2}$,

$mS' = 3x27\frac{1}{2} = 82\frac{1}{2}$,

$W = \frac{S}{mS'} = \frac{36\frac{1}{2}}{82\frac{1}{2}} = 0,44.$

De definitie van W is zo gekozen, dat W=1 is indien er volledige overeenstemming is tussen de verschillende rijen, d.w.z. ook in de gelijke rangnummers. De definitie geldt ook, als er geen gelijken voorkomen.

3.2 <u>Correcties in de formules vanwege het optreden van gelijke rangnummers</u>.

We gaan eerst na wat er gebeurt met de som van de quadraten der gereduceerde rangnummers van een rij, als t rangnummers gelijk worden.

Uit relatie (II,8) volgt, dat de verandering die optreddt in de som van de quadraten der gereduceerde rangnummers, gelijk is aan de verandering die optreedt in de som der quadraten der rangnummers zelf, daar het gemiddelde der rangnummers niet verandert door het optreden van gelijken. Deze verandering is, indien wij veronderstellen, dat de rangnummers

$$y+1,\ y+2,\ \ldots\ldots\ldots y+t$$

gelijk worden aan

$$\underline{y + \frac{t+1}{2}} .$$

(II,23). $(y+1)^2+ (y+2)^2+\ldots+(y+t)^2- t(y+ \frac{t+1}{2})^2=$

$= ty^2+ 2(1+2+\ldots+t)y+ 1^2+2^2+\ldots+t^2-ty^2-t(t+1)y + \frac{(t+1)^2}{4}$

$= \sum_{K=1}^{t} K^2- \frac{1}{4}(t+1)^2= \frac{1}{12}t(t^2-1)$. (zie (II,9)).

Indien in de i-de rij, groepen van t_1, t_2,...,t_p gelijke rangnummers voorkomen zal de totale correctie in de som der quadraten van de gereduceerde rangnummers worden:

(II,24) $T_i= \frac{1}{12}\sum_{\nu=1}^{p} t_\nu (t_\nu^2-1)$.

Wij merken hierbij op, dat T_i alleen afhankelijk is van de grootte van de groepen gelijke rangnummers, niet van de vraag wèlke t-tallen rangnummers gelijk zijn.

<u>Bijvoorbeeld</u>:

Voor de rij

$$1;\ 2\tfrac{1}{2};\ 2\tfrac{1}{2};\ 3;\ 5;\ 5;\ 5$$

geldt:

$$T= \tfrac{1}{12}\cdot 2(2^2-1)+ \tfrac{1}{12}\cdot 3(3^2-1)= 2{,}5$$

Dit is eveneens het geval bij de rij:

$$2;\ 2;\ 2;\ 4\tfrac{1}{2};\ 4\tfrac{1}{2};\ 6;\ 7$$

De correctie in de som der quadraten S' van de gereduceerde rangnummers uit <u>alle</u> rijen wordt nu:

$$\sum_{i=1}^{m} T_i$$

Indien er geen gelijke rangnummers voorkomen geldt:

(II,25)
$$S'= m\sum_{j=1}^{n} x^2_{ij}= m\sum_{j=1}^{n}\left\{a_{ij}- \tfrac{1}{2}(n+1)\right\}^2=$$
$$= m\sum_{j=1}^{n}(j-\bar{j})^2= \frac{1}{12}mn(n^2-1).$$

Indien er wel gelijke rangnummers voorkomen geldt dus:

(II,26) $S'= \frac{1}{12}mn(n^2-1)-\sum_{i=1}^{m} T_i$,

zodat tenslotte geldt:

(II;27) $$W=\frac{S}{mS'}= \frac{S}{\frac{1}{12}m^2n(n^2-1)-m\sum_{i=1}^{m} T_i}$$

<u>Voorbeeld</u>:

Bij het rangnummerschema:

(II,28)

1	2	4	4	4	6	7
$1\frac{1}{2}$	3	$1\frac{1}{2}$	$4\frac{1}{2}$	7	$4\frac{1}{2}$	6
2	1	3	5	6	4	7
$4\frac{1}{2}$	6	$8\frac{1}{2}$	$13\frac{1}{2}$	17	$14\frac{1}{2}$	20

vindt men met behulp van formule(II.8):

$$S=(4\tfrac{1}{2})^2+6^2+(8\tfrac{1}{2})^2+(13\tfrac{1}{2})^2+17+(14\tfrac{1}{2})^2+20^2-7\times12^2=202$$

$$T_1=\tfrac{1}{12}\times3(3^2-1)=2$$

$$T_2=\tfrac{1}{12}\times2(2^2-1)+\tfrac{1}{12}\times2(2^2-1)=1$$

$$T_3=0$$

$$S'=\tfrac{1}{12}\times3\times7(7^2-1)-(2+1)=81$$

$$W=\frac{202}{3\times81}=\frac{202}{243}=0{,}831$$

3.3 Exacte verdeling van S indien er gelijke rangnummers zijn.

Indien een oneven aantal rangnummers gelijk wordt, blijven alle rangnummers gehele getallen. Zoals aangetoond kan worden, is op dat geval hetgeen wij in 2.2 beweerd hebben, over de waarden, die S kan aannemen, eveneens van toepassing.

Indien echter een even aantal rangnummers gelijk wordt, treden er rangnummers op, die gelijk zijn aan een geheel getal $+\frac{1}{2}$, terwijl andere rangnummers geheel zijn, zodat het betoog van 2.2 niet meer geldt.

Wij geven als voorbeeld de verdeling van S voor de rangnummers:

(II,28)

1^orij:	1	$2\frac{1}{2}$	$2\frac{1}{2}$
2^orij:	1	2	3
3^orij:	1	2	3

Wij vinden, volgens de methode beschreven in 2.3, uitgaande van de hypothese, dat alle permutaties van de door het schema (II,29) gegeven rangnummers even waarschijnlijk zijn, de volgende tabel:

Tabel (II,3)

S	$P[S = S]$	$P[S \geqq S]$
$\frac{1}{2}$	$\frac{1}{18}$	1
$1\frac{1}{2}$	$\frac{3}{18}$	$\frac{17}{18}$
$2\frac{1}{2}$	$\frac{1}{18}$	$\frac{14}{18}$
$3\frac{1}{2}$	$\frac{4}{18}$	$\frac{13}{18}$
$4\frac{1}{2}$	$\frac{2}{18}$	$\frac{9}{18}$
$6\frac{1}{2}$	$\frac{1}{18}$	$\frac{7}{18}$
$8\frac{1}{2}$	$\frac{1}{18}$	$\frac{6}{18}$
$9\frac{1}{2}$	$\frac{1}{18}$	$\frac{5}{18}$
$10\frac{1}{2}$	$\frac{2}{18}$	$\frac{4}{18}$
$13\frac{1}{2}$	$\frac{1}{18}$	$\frac{2}{18}$
$15\frac{1}{2}$	$\frac{1}{18}$	$\frac{1}{18}$

Hoewel slechts een kleine verandering in het rangnummerschema is aangebracht, wijkt de verkregen verdeling toch sterk af van de verdeling voor het overeenkomstige schema zonder gelijke rangnummers. Dit blijkt duidelijk als we een grafiek vervaardigen, van $P[\underline{S}=S]$ uit tabel (II,2) en uit tabel (II,3). (Fig. II, 1,2 en 3).

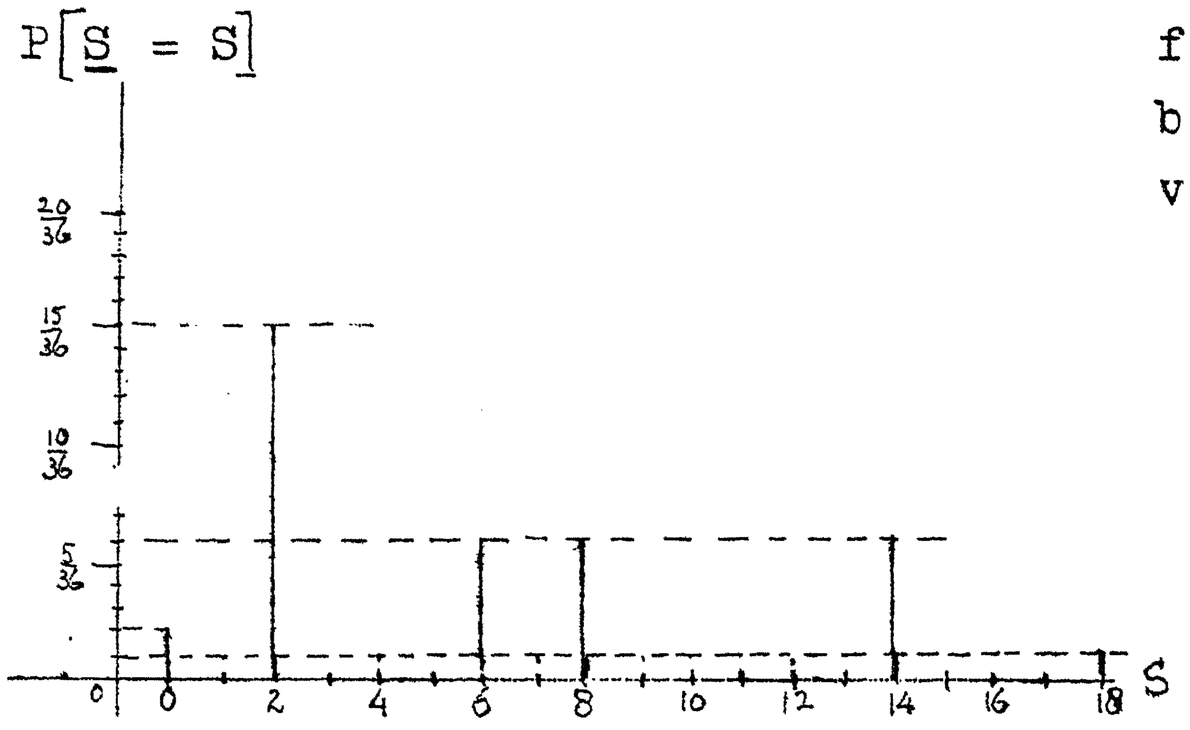


fig.II.1 Verdeling van $\underline{S}$ bij 3 rangschikkingen van 3 ongelijke rangnummers

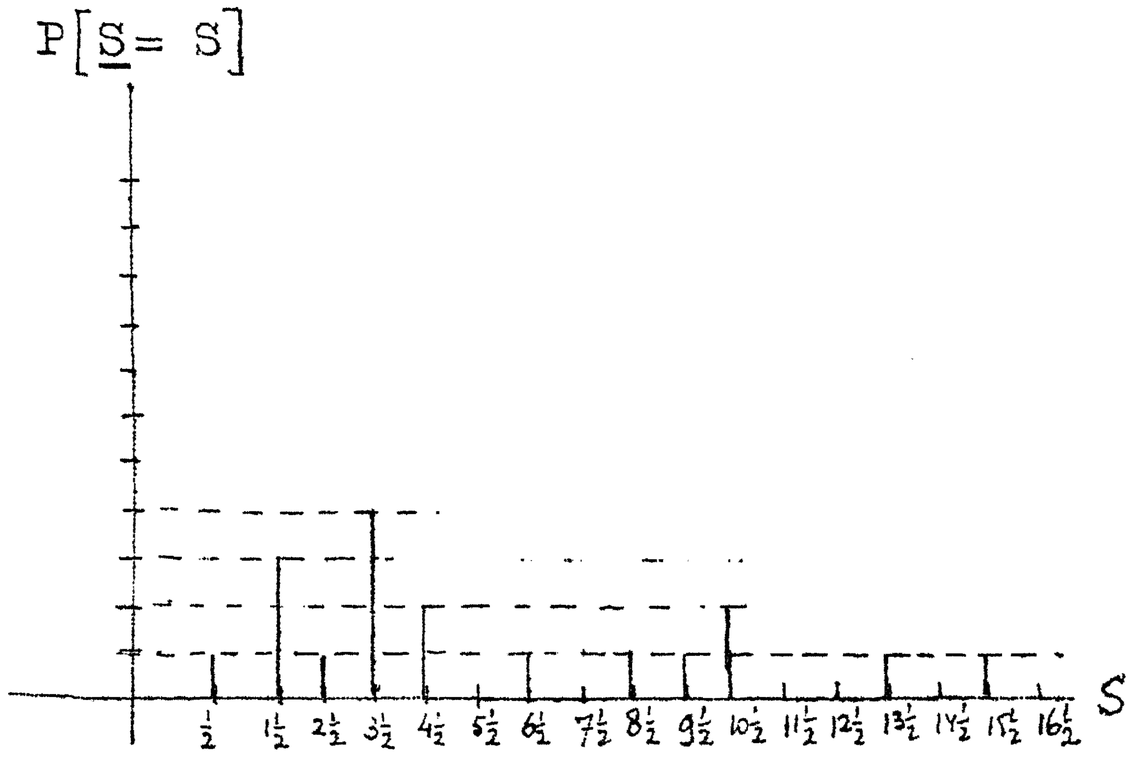


Fig.II.2 Verdeling van $\underline{S}$ bij 3 rangschikkingen: waarbij in één rangschikking de twee hoogste rangnummers gelijk zijn.

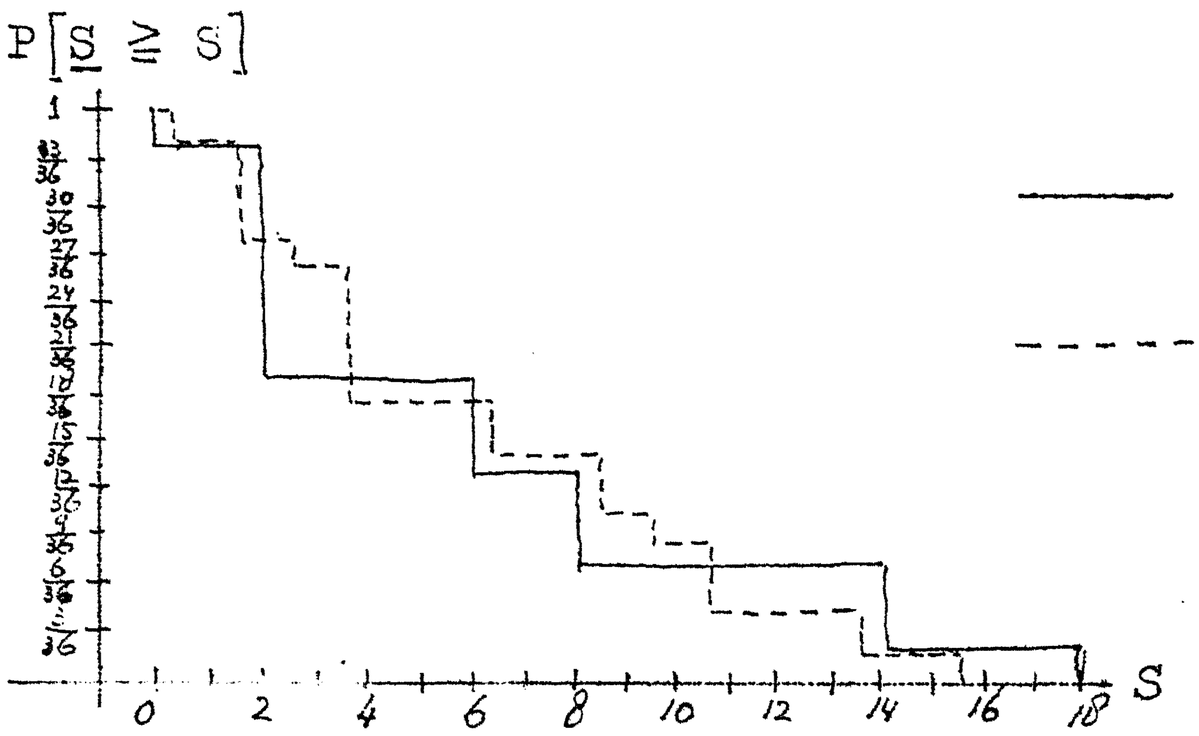


Fig II.3 Overschrijdingskansen $P\ \underline{S}\ \ S$, voor
——— 3 rangschikkingen van 3 ongelijke rangnummers
- - - - - 3 rangschikkingen in één waarvan de twee hoogste rangnummers gelijk zijn

Bij beschouwing van fig.II.3 constateren wij dat de overschrijdingskansen voor $S > 10\frac{1}{2}$ het grootste zijn in het geval van ongelijke rangnummers. Dit was te verwachten; want indien er gelijke rangnummers voorkomen, is S gemiddeld lager, dan indien er geen gelijke rangnummers voorkomen.

Er bestaan, voorzover bekend, geen tabellen voor de exacte verdeling van S voor rangschikkingen met gelijken. In plaats daarvan gebruikt men de overeenkomstige tabellen voor rangschikkingen zonder gelijken. Men vindt zodoende bij grote waarden van S te grote overschrijdingskansen en zal dus minder snel tot verwerping van H_o overgaan, dan indien men van de juiste tabellen gebruik gemaakt had; men maakt dus een fout in de "veilige richting". Dit is niet meer juist voor kleinere S (zie b.v. fig.II. 3), in de regel zijn daarbij echter de overschrijdingskansen zo groot, dat een nauwkeurige bepaling daarvan, althans voor het toetsen van de hypothese H_o, weinig zin heeft.

4. Benaderingen voor de verdeling van S of van W.

4.1. Algemeen principe van de benaderingsmethoden.

Als benaderde verdelingen worden continue verdelingen gebruikt, die aangepast zijn met het gemiddelde en eventueel het tweede moment, terwijl tenminste het derde en vierde gereduceerde moment van de aangepaste verdeling voor grote m en n asymptotisch gelijk zijn aan de overeenkomstige momenten van de verdeling van S of van W.

4.2. Benadering met de Bêta-verdeling.

De eerste verdeling, die men als benadering kan gebruiken is de zogenaamde Bêta-verdeling, die wordt aangepast met het gemiddelde en het tweede moment. Deze aanpassing wordt bereikt door de parameters p en q van de Bêta-verdeling als volgt te kiezen:

(II,28) $$p = \tfrac{1}{2}(n-1) - \frac{1}{m}$$

en

(II.29) $$q = (m-1)p = \tfrac{1}{2}(m-1)(n-1) - \frac{m-1}{m}\ .$$

Door deze keuze worden het gemiddelde en de spreiding van de Bêta-verdeling gelijk aan de overeenkomstige grootheden van de verdeling van W. Bij nader onderzoek blijken het derde en het vierde gereduceerde moment van de verdeling van W en de aangepaste Bêta-verdeling bij benadering eveneens gelijk te zijn, tenzij m(n-1) zeer klein is. De afwijking in het derde moment is kleiner dan 10 procent van de juiste waarde als $m(n-1) > 18$ is en kleiner dan één procent als $m(n-1) > 198$ is.

Dit alles geldt, indien er geen gelijken zijn, onder de hypothese H_o, dat alle permutaties van de rangnummers in iedere rij even waarschijnlijk zijn. Zijn er wel gelijken, dan blijft de B-verdeling een goede benadering, doch men zal, indien het aantal gelijken groot is, de parameters p en q in afwijking van de formules (II.28) en (II.29) als volgt moeten kiezen:

$$(II.30) \qquad p = \frac{m-1}{m^3\tilde{\mu}_2(\underline{W})} - \frac{1}{m}$$

$$(II.31) \qquad q = (m-1)p = \frac{(m-1)^2}{m^3\tilde{\mu}_2(\underline{W})} - \frac{m-1}{m} \ ,$$

waarin

$$(II.32) \qquad \tilde{\mu}_2(\underline{W}) = \frac{4}{m^2(n-1)} \times \frac{\sum_{i=1}^{m} \sum_{k=1}^{i-1} \tilde{\mu}_{2,i}\tilde{\mu}_{2,k}}{\left\{\sum_{i=1}^{m} \tilde{\mu}_{2,i}\right\}^2} \quad \text{is.}$$

Onder $\tilde{\mu}_{2,i}$ verstaat men in deze formule $\frac{1}{n}$ x de som van de quadraten der gereduceerde rangnummers (dit zijn de rangnummers verminderd met hun gemiddelde) in de i-de rij. Uit onze formules in §3.2 volgt direct:

$$(II.33) \qquad \tilde{\mu}_{2,i} = \frac{1}{n}\left\{\frac{1}{12}\cdot n(n^2-1) - T_i\right\} = \frac{1}{12}(n^2-1) - \frac{1}{n}T_i$$

Het blijkt dus, dat indien er gelijken voorkomen, deze verdeling slechts afhangt van de grootte der T_i . Men kan dus zeggen, dat, zo er veel gelijken voorkomen, de B-verdeling met parameters p en q gegeven door (II.30) en (II.31) bij benadering de verdeling van $\underline{W}$ voorstelt, onder de hypothese H_o, dat in iedere rij alle permutaties van de rangnummers even waarschijnlijk zijn en onder de voorwaarde, dat alle veranderingen, wat betreft het aantal en de groepering der gelijken zijn toegestaan, waarbij de grootheid T_i van die rij niet verandert.

Men vindt de B-verdeling getabelleerd in K. Pearson (1934) en wel voor de volgende waarden van p en q:

0 ; 0,5 ; 1 ; 1,5 ; 2 ... 10,5 ; 11 ; 12 ; 13 ; .... ; 50 .

4.3. Benadering met de z-verdeling van Fisher.

In de practijk zal men zeer vaak de tabellen van K. Pearson voor de B-verdeling niet kunnen gebruiken, omdat q = (m-1)p slechts bij een zeer kleine m kleiner zal zijn dan 50. Men maakt daarom meer gebruik van de tafels van de verdeling van e^{2z} of van z. De z is een grootheid, die samenhangt met de normale verdeling, en waarvan de

verdeling, die een eenvoudige transformatie is van de Bêta-verdeling, berekend is door Fisher. Men vindt haar in ieder leerboek betreffende Wiskundige Statistiek behandeld.

Het blijkt nu, dat

$$(II.39) \qquad V = (m-1)\frac{W}{1-W}$$

bij benadering verdeeld is als e^{2z}, indien men voor de parameters (de z.g. aantallen vrijheidsgraden) dezer verdeling kiest:

$$(II.35) \quad \begin{cases} \nu_1 = (n-1) - \frac{2}{m} \\ \nu_2 = (m-1)\nu_1 = (m-1)(n-1) - \frac{2(m-1)}{m} . \end{cases}$$

En als er veel gelijken voorkomen:

$$(II.36) \qquad \nu_1 = \frac{2(m-1)}{m^3 \tilde{\mu}_2(\underline{W})} - \frac{2}{m} \text{ en } \nu_2 = (m-1)\,\nu_1$$

waarin $\tilde{\mu}_2(\underline{W})$ gegeven wordt door (II.32).

Een tabel van de waarden van e^{2z} met eenzijdige overschrijdingskansen :

$$(II.37) \quad \begin{cases} 0{,}2 \;;\; 0{,}1 \;;\; 0{,}05 \;;\; 0{,}01 \;;\; 0{,}001 \\ \text{voor :} \\ \nu_1 = 1 \;;\; 2 \;;\ldots;\; 6 \;;\; 8 \;;\; 12 \;;\; 24 \;;\infty \\ \text{en} \\ \nu_2 = 1 \;;\; 2 \;;\; \ldots;\; 30 \;;\; 40 \;;\; 60 \;;\; 120 \;;\infty \end{cases}$$

komt voor in Fisher en Yates (1949) Tabel V.

Verder vindt men de waarden van e^{2z}. met overschrijdingskansen 0,05 en 0,01 voor:

$$(II.38) \quad \begin{cases} \nu_1 = 1 \;;\; 2 \;;\; \ldots;\; 12 \;;\; 14 \;;\; 16 \;;\; 20 \;;\; 24 \;;\; 30;\; 40;\; 50; \\ \qquad 75;\; 100 \;;\; 200 \;;\; 500 \;;\infty \\ \text{en} \\ \nu_2 = 1 \;;\; 2 \;;\ldots;\; 30 \;;\; 32 \;;\; \ldots;\; 50 \;;\; 55 \;;\; \ldots;\; 70;\; 80 \;; \\ \qquad 100 \;;\; 125 \;;\; 150 \;;\; 200 \;;\; 400 \;;\; 1000 \end{cases}$$

in Hoel (1947), Tabel V.

Uit het feit dat V bij benadering verdeeld is als e^{2z}, volgt dat:

$$(II.39) \qquad \frac{1}{2} \ln V = \frac{1}{2} \ln \frac{(m-1)W}{1-W}$$

bij benadering verdeeld is als z, eveneens met parameters ν_1 en ν_2 gegeven door (II.35) of (II.36).

In tabel V van Fisher en Yates (1949), vindt men de waarden van z, die corresponderen met de waarden van e^{2z}, met overschrijdingskansen en parameterwaarden gegeven door (II.37). Een kleinere tabel, welke slechts de punten met overschrijdingskansen 0,05 en 0,01 bevat, voor alle parameterwaarden vermeld in (II.35), behoudens $\nu_2 = 40$ en $\nu_2 = 120$, treft men aan in Fisher (1948) Tabel VI, welke tabel gereproduceerd is in Kendall (1948) Appendix Tabel 7 en in Kendall (1947) Appendix Tabellen 4 en 5.

Voorbeeld:

Indien voor een rangnummersschema zonder gelijken geldt:

$$m = 20 \qquad n = 11 \qquad W = 0,1$$

vindt men:

$$\nu_1 = (11-1) - \frac{2}{20} = 9,9$$

$$\nu_2 = (m-1)\nu_1 = 19 \times 9,9 = 188,1$$

$$V = (m-1)\frac{W}{1-W} = 19 \times \frac{0,1}{0,9} = 2,11$$

In de tabel van Hoel vinden we:

ν_1	ν_2	0,05-punt [1]	0,01-punt [1]
10	150	1,89	2,44
10	200	1,87	2,41

Wij constateren, dat $0,05 > P[V > 2,11] > 0,01$ is.

De verkregen uitkomst is dus significant, als we als onbetrouwbaarheidsdrempel 0,05, doch niet significant als we als onbetrouwbaarheidsdrempel 0,01 aanhouden.

Ter contrôle gaan wij ook nog als volgt te werk:

$z = \frac{1}{2} \ln 2,11 = 0,373$.

Wij vinden in tabel 7 van Kendall (1948)

ν_1	ν_2	0,05-punt [1]	0,01-punt [1]
8	60	0,3702	0,5189
12	60	0,3255	0,4574
8	∞	0,3309	0,4604
12	∞	0,2804	0,3908

---

[1]) Onder het 0,05 (0,01)-punt verstaan wij de waarde van de variabele, waarbij een overschrijdingskans 0,05 (0,01) behoort.

Dus is:

$$0{,}05 > P[\underline{z} \geqq 0{,}373] > 0{,}01$$

waaruit wij dezelfde conclusie trekken als boven.

4.4. Benadering met de χ^2-verdeling.

Deze benadering berust op het feit, dat:

$$s_j = \sum_{i=1}^{m} x_{ij} \qquad \text{((II.12) zie blz. 23)}$$

een verdeling bezit, die voor iedere j convergeert naar een normale verdeling als m toeneemt en wel onder de hypothese H_o, dat alle permutaties van de rangnummers in iedere rij even waarschijnlijk zijn.

Het gemiddelde van deze normale verdeling is gelijk aan 0, evenals het gemiddelde van $\underline{s}_j$.

Er geldt namelijk:

$$\text{(II.40)} \qquad \mathcal{E}\,\underline{s}_j = \mathcal{E} \sum_{i=1}^{m} x_{ij} = \sum_{i=1}^{m} \frac{1}{n} \sum_{j=1}^{n} x_{ij} = 0$$

De spreiding van de normale verdeling is gelijk aan de spreiding van $\underline{s}_j$, dus gelijk aan $\sqrt{\mathcal{E}\,\underline{s}_j^2}$ waarin:

$$\mathcal{E}\,\underline{s}_j^2 = \mathcal{E}\Big(\sum_{i=1}^{m} x_{ij}\Big)^2 = \mathcal{E} \sum_{i=1}^{m} x_{ij}^2 + 2\,\mathcal{E} \sum_{i=1}^{m} \sum_{k=1}^{i-1} x_{ij}\, x_{kj} =$$

$$= \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}^2 + \frac{2}{n^2} \sum_{i=1}^{m} \sum_{k=1}^{i-1} \sum_{j=1}^{n} \sum_{l=1}^{n} x_{ij}\, x_{kl}$$

Omdat $\sum_{j=1}^{n} x_{ij} = 0$ is, is de tweede term van het rechterlid gelijk aan nul terwijl:

$$\sum_{j=1}^{n} x_{ij}^2 = \frac{1}{12}\, n(n^2-1) - T_i \text{ is (zie II.24)},$$

zodat geldt:

$$\text{(II.41)} \qquad \mathcal{E}\,\underline{s}_j^2 = \frac{1}{12}\, m(n^2-1) - \frac{1}{n} \sum_{i=1}^{m} T_i$$

Men kan nu bewijzen, dat de hogere gestandardiseerde momenten van de verdeling van $\underline{s}_j$ voor $m \to \infty$ convergeren naar de overeenkomstige momenten van de normale verdeling met gemiddelde nul en spreiding één, zodat $\underline{s}_j$ asymptotisch normaal verdeeld is, als $m \to \infty$.

De verdeling van de som van de quadraten van n onafhankelijk normaal verdeelde variabelen (met gemiddelde 0 en spreiding 1), staat in de wiskundige statistiek bekend als de χ^2-verdeling met

n vrijheidsgraden. Hieruit volgt dat de verdeling van $\underline{S} = \sum_{j=1}^{n} \underline{s}_j^2$ bij benadering het karakter zal hebben van een χ^2-verdeling en wel, omdat

$$\sum_{j=1}^{n} s_j = 0 \qquad \text{(II.42)}$$

en dus slechts n-1 grootheden s_j onafhankelijk zijn, een χ^2-verdeling met n-1 vrijheidsgraden.
Nu geldt, voor een χ^2-verdeling met n-1 vrijheidsgraden

$$\mathcal{E}\,\chi^2 = n-1, \qquad \text{(II.43)}$$

terwijl

$$\mathcal{E}\,\underline{S} = \sum_{j=1}^{n} \mathcal{E}\,\underline{s}_j^2 = n\,\mathcal{E}\,\underline{s}_j^2 \text{ is,} \qquad \text{(II.44)}$$

zodat

$$\chi_r^2 = \frac{n-1}{n\,\mathcal{E}\,\underline{s}_j^2} \cdot S \qquad \text{(II.45)}$$

bij benadering verdeeld is als χ^2 met <u>n-1</u>-vrijheidsgraden; $\mathcal{E}\,\underline{s}_j^2$ wordt hierin gegeven door (II.41).

Indien er geen gelijken voorkomen gaat (II.45) over in:

$$\chi_r^2 = \frac{12}{mn(n+1)}\, S = m(n-1)W \,. \qquad \text{(II.46)}$$

Zijn er wel gelijken, dan geldt:

$$\chi_r^2 = \frac{1}{\frac{1}{12}mn(n+1) - \frac{1}{n-1}\sum_{i=1}^{m} T_i} \cdot S \qquad \text{(II.47)}$$

Uit deze laatste formule blijkt, dat volgens deze benadering de verdeling van $\underline{S}$ onafhankelijk is van het aantal gelijken en hun verdeling over het schema, zolang $\sum_{i=1}^{m} T_i$ niet verandert. Wij zien dus, dat χ_r^2 gegeven door (II.47) voor grote m bij benadering verdeeld is als χ^2 met n-1 vrijheidsgraden, onder de hypothese H_o, dat alle permutaties der rangnummers even waarschijnlijk zijn, terwijl al die veranderingen wat betreft het aantal en de groepering van de gelijken in het schema zijn toegestaan, die $\sum_{i=1}^{m} T_i$ ongewijzigd laten.

<u>Voorbeeld</u>:

28 rangschikkingen van 13 (geen [illegible]

$S = 11440$

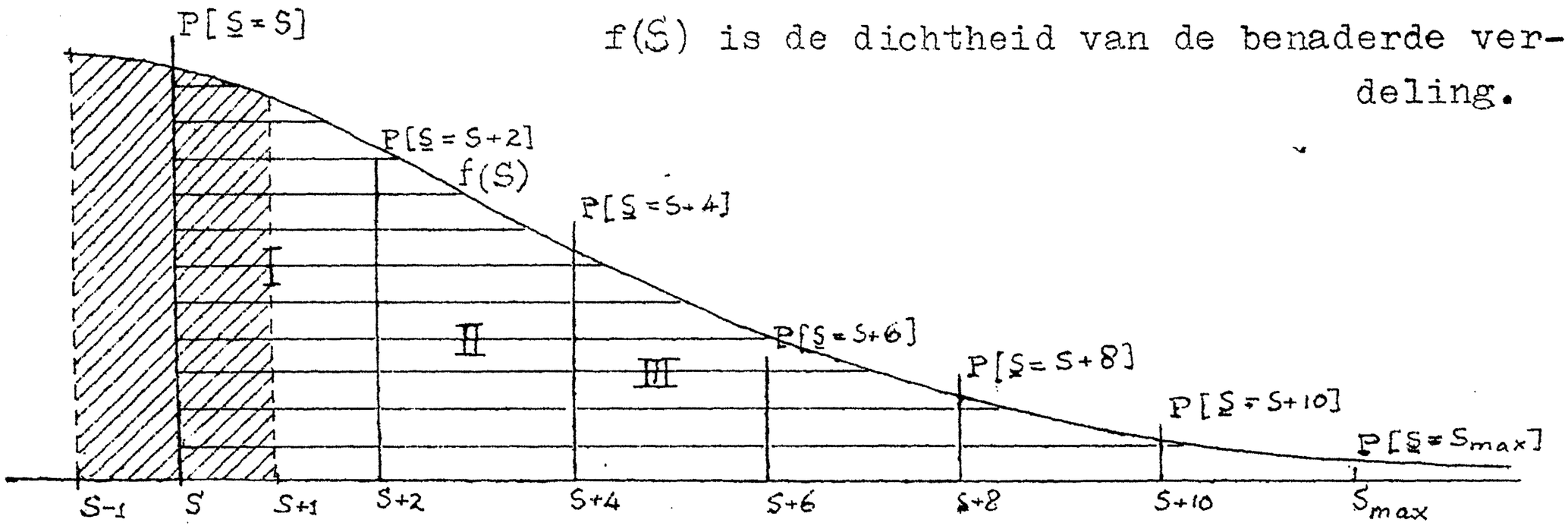
$\chi_r^2 = \frac{12 \times 11440}{28 \times 13 \times 14} = 27$ n-1 = 12 vrijheidsgraden .

Het 0,01-punt is in dit geval 26,217, zodat men aan de hand van het resultaat de hypothese H_o met de onbetrouwbaarheid 0,01 kan verwerpen.

4.5. Continuïteitscorrecties.

Bij de overgang van de exacte verdeling van S op een continue benadering, doet zich nog een moeilijkheid voor. Indien wij in de formules voor de continue variabele de waarde van S invullen, die wij bij het experiment gevonden hebben, dan nemen wij aan dat het oppervlak onder de (benaderende dichtheids-)kromme rechts van deze S-waarde, ongeveer gelijk is aan de (exacte) overschrijdingskans van deze S-waarde, d.w.z. gelijk aan de som der discrete kansen $P[\quad = \quad]$, $P[\quad = \quad+2]$... $P[\quad = s_{max}]$.

Fig. (II.4)
Continuïteitscorrectie
f(S) is de dichtheid van de benaderde verdeling.



We beschouwen hier speciaal het geval, dat er geen gelijke rangnummers voorkomen, zodat de verschillen tussen de S-waarden even zijn.

Wij hebben in fig. 4 de rechterstraart geschetst van de exacte en de benaderende verdeling; hierin wordt $P[\quad \geq \quad]$ volgens de exacte verdeling voorgesteld door de som van de verticale dikke lijnstukken en dezelfde kans volgens de benaderende verdeling door de oppervlakte van het horizontaal gearceerde gedeelte. De consequentie van deze methode is, dat men $P[\quad = \quad]$ gelijk stelt de oppervlakte van strook I, $P[\quad = s+2]$ aan de oppervlakte van strook II, enz. Op grond van symmetrieoverwegingen, geeft men er veelal de voorkeur aan, om aan te nemen, dat $P[\quad = \quad]$ gelijk is aan de oppervlakte van de schuin gearceerde strook tussen s-1 en s+1. Hieruit volgt dat men de exacte overschrijdingskans van S, benadert met "continue" overschrijdingskans van s-1.

De strook om s=0 ligt tussen -1 en +1, de strook om $s = s_{max}$ tussen $s_{max}-1$ en $s_{max}+1$, men smeert de discreet verdeelde waarschijnlijkheid [illegible] dus a.h.w. uit over het interval $(-1, s_{max}+1)$, dat de lengte $s_{max}+2$ heeft. Op grond hiervan verwacht men, dat de

benadering beter zal worden indien men ook in de noemer van W een correctie ter grootte van 2 aanbrengt, zodat men

$$(II.48) \qquad W' = \frac{s-1}{max^{+2}} = \frac{s-1}{\frac{1}{12}m^2n(n^2-1)+2}$$

gebruikt in plaats van W.

In de practijk blijkt inderdaad, dat men door toepassing van deze correcties voor kleine m en n in het bijzonder met de Bêta-verdeling en de z-verdeling, betere benaderingen van de exacte overschrijdingskansen verkrijgt. Voor grote waarden van m en n, is de invloed van de continuïteitscorrectie te verwaarlozen.

Het heeft geen zin om de continuïteitscorrecties toe te passen, indien er rijen met gelijke rangnummers voorkomen, omdat dan ook de exacte tabel maar bij benadering juist is, terwijl de werkelijke verdeling onbekend is, zodat wij niet kunnen nagaan of de continuïteitscorrectie in dit geval een gunstige invloed heeft of niet.

4.6. Onderlinge vergelijking van de benaderende en de exacte verdelingen.

Wij zullen deze vergelijking uitvoeren aan de hand van tabel (II.4). Deze tabel bevat:

In kolom 1. Het aantal rangschikkingen m;
" " 2. Het aantal rangnummers n, in iedere rangschikking;
" " 3. S-waarden, welke volgens de exacte verdeling onder de hypothese H_o een overschrijdingskans in de buurt van 0,05 bezitten;
" " 4. De overschrijdingskansen, behorende bij de S-waarden in kolom 3;
" " 5. De S-waarde met overschrijdingskans 0,05, indien men uitgaat van de veronderstelling dat $\frac{1}{2}$ ln $\frac{(m-1)W'}{1-W'}$ verdeeld is als de z van Fisher met

$$\begin{cases} \nu_1 = n-1 - \frac{2}{m} \\ \nu_2 = (m-1)\nu_1 \end{cases}$$

en:

$$W' = \frac{s-1}{\frac{1}{12}m^2(n^2-1)+2}$$

zodat we hier de continuïteitscorrectie toepassen (zie II.48);

In kolom 6. De S-waarde met overschrijdingskans 0,05 indien men uitgaat van de veronderstelling, dat $m(n-1)W$ verdeeld is als χ^2 met n-1 vrijheidsgraden, waarbij W berekend is volgens

$$W = \frac{12 \cdot}{m^2 n(n^2-1)}$$

zonder continuïteitscorrectie;

" " 7. S-waarden met overschrijdingskans $\sim$ 0,01 (analoog aan kolom 3);

" " 8. Overschrijdingskansen behorende bij de S-waarden in kolom 7;

" " 9. S-waarde met overschrijdingskans 0,01, onder dezelfde veronderstelling als in kolom 5;

" " 10. S-waarde met overschrijdingskans 0,01 onder dezelfde veronderstelling als in kolom 6.

De gegevens voor de kolommen 3, 4, 5, 7, 8 en 9 zijn rechtstreeks ontleend aan Kendall (1948), tabellen 5 en 6. De kolommen 6 en 10, zijn berekend met behulp van tabel 8, door toepassing van:

$$S = \frac{mn(n+1)\chi^2}{12}$$

Zie formule (II.46).

In tabel (II.4) constateren wij het volgende:

1). Over het algemeen sluiten de 0,05 en 0,01 punten (S-waarden met overschrijdingskansen 0,05 en 0,01) bepaald met behulp van de z-benadering zeer goed aan bij de overeenkomstige punten van de exacte verdeling. (Mits men voor kleine m en n continuïteitscorrecties toepast).

2). De 0,05- en in het bijzonder de 0,01-punten, bepaald met behulp van de χ^2- benadering liggen over het algemeen te hoog. Het is daarom beter hier geen continuïteits-correctie toe te passen, aangezien de punten dan nog hoger komen te lig gen. Indien echter m relatief groot is t.o.v. n, is de χ^2- benadering beter dan z-benadering (zie b.v. de 0,01-punten bij m = 9 en 10 en n = 3).

Wij concluderen hieruit:

In de gevallen, waarvoor geen exacte tabellen beschikbaar zijn, vinden wij bij toepassing van de χ^2-benadering, over het algemeen een te grote waarde voor de overschrijdingskans, zodat wij niet ten gevolge van het verschil tussen benadering en werkelijke verdeling ten onrechte tot de verwerping van H_o, indien deze juist is, zullen overgaan (wat de fouten van de eerste soort betreft geeft deze benadering dus resultaten "aan de veilige kant"). Deze benadering heeft het voordeel, dat zij tot minder rekenwerk aanleiding geeft dan de

Tabel (II.4)

Kolom: 1	2	3	4	5	6	7	8	9	10
				0,05 punt				0,01 punt	
m	n	S	$P[\ldots \geq \ldots]$	Benaderd met z	Benaderd met χ^2	S	$P[\ldots \geq \ldots]$	Benaderd met z	Benaderd met χ^2
3	5	62	0,056	64,4	71,1	74	0,015	75,6	99,6
		64	0,045			76	0,0078		
4	4	50	0,052	49,5	52,1	62	0,012	61,4	75,6
		52	0,036			64	0,0069		
5	4	61	0,055	62,6	65,0	81	0,012	80,5	94,5
		65	0,044			83	0,0087		
6	4	74	0,056	75,7	78,2	100	0,010	99,5	113,5
		76	0,043			102	0,0096		
6	5			136,1	142,3			176,1	199,2
6	6			221,4	232,5			282,4	316,8
6	7			335,2	352,6			422,6	470,7
8	3	42	0,079	48,1	47,9	62	0,018	66,8	73,7
		50	0,047			72	0,0099		
8	4			101,7	104,2			137,4	151,3
8	5			183,7	189,8			242,7	265,5
8	6			299,0	312,0			388,3	422,5
8	7			453,1	470,0			579,9	627,7
9	3	54	0,057	54,0	53,9	78	0,010	75,9	82,9
		56	0,048			86	0,0060		
10	3	56	0,066	60,0	60,0	86	0,012	85,1	92,1
		62	0,046			96	0,0075		
10	4			127,8	130,3			175,3	189,1
10	5			231,2	237,2			309,1	331,9
10	6			376,7	387,5			494,0	528,0
10	7			571,0	587,7			737,0	784,6
15	3			89,8	89,9			131,0	138,2
15	4			192,9	195,4			269,8	283,6
15	5			349,8	355,7			475,2	497,9
15	6			570,5	581,2			758,2	791,8
15	7			864,9	881,4			1129,5	1176,8
20	3			119,7	119,8			177,0	184,2
20	4			258,0	260,4			364,2	378,2
20	5			468,5	474,4			641,2	664,0
20	6			764,4	774,9			1022,2	1056,3
20	7			1158,7	1174,9			1521,9	1569,1

z-benadering; zij is tevens de beste benadering als het aantal rijen veel groter is dan het aantal kolommen. In andere gevallen is de z-benadering echter scherper. Wij zullen haar dus in het bijzonder toepassen indien de χ^2-benadering een juist niet-significante uitkomst geeft. Als men werkt met onbetrouwbaarheidsdrempel 0,01, en $n > m$ is, is de χ^2-benadering dermate onscherp, dat het geschikter is direct de z-benadering te kiezen.

5. Schatting van de ware volgorde.

De methode van de m rijen levert ons niet alleen een toets voor overeenstemming tussen de m rangschikkingen van een schema, doch ook een schatting van een objectieve volgorde der objecten. Deze schatting verkrijgt men, indien men de objecten rangschikt volgens de opklimmende grootte van de kolom-totalen; zij neemt, zoals we zullen zien, tussen alle mogelijke schattingen van een objectieve volgorde, een bijzondere plaats in.

Wij beschouwen hiertoe het volgende schema van niet-gereduceerde rangnummers:

(II.49)
$$\begin{array}{cccc} a_{11} & a_{12} & \cdots\cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots\cdots & a_{mn} \\ \hline S_1 & S_2 & & S_m \end{array}$$

De letters $S_1,\ldots,S_m$ stellen hier dus de niet-gereduceerde kolom-totalen voor. Tevens beschouwen wij een willekeurige rangschikking als schatting van de volgorde der objecten; in deze rangschikking draagt object j het rangnummer $X_j (j = 1,2,\ldots,n)$. Indien wij nu een tweede schema opbouwen, bestaande uit m rijen: $X_1, X_2, \ldots X_n$ dan zullen de kolomtotalen van dit tweede schema worden:

$$mX_1,\ldots \qquad mX_n .$$

Nu is de som van de quadraten van de verschillen tussen de kolomtotalen van dit schema en van schema (II.49) minimaal als we voor

$$X_1,\ldots,X_m$$

de rangschikking kiezen volgens de opklimmende grootte der S_j.

Bewijs:

Genoemde som wordt:

$$\sum = \sum_{j=1}^{n} (S_j - m X_j)^2 =$$

$$\sum_{j=1}^{n} S_j^2 - 2m \sum_{j=1}^{n} S_j X_j + m^2 \sum_{j=1}^{n} X_j^2 .$$

De eerste en laatste term van deze uitdrukking zijn constant, want $\sum_{j=1}^{n} S_j^2$ is gegeven door schema (II.49) en

$$\sum_{j=1}^{n} X_j^2 = 1^2 + 2^2 + \ldots + n^2 = \frac{1}{6} n(n+1)(2n+1) .$$

zodat $\sum$ minimaal is, als $\sum_{j=1}^{n} S_j X_j$ maximaal is. Dit is het geval, indien voor ieder paar gehele getallen: $1 \leqq j \leqq n$, $1 \leqq K \leqq n$, geldt: $X_j < X_k$ als $S_j < S_k$, dus indien de rangschikking $X_1, \ldots, X_n$ de rangschikking volgens opklimmende grootte van de kolomtotalen voorstelt.

Men dient het volgende wel te bedenken: indien de overeenstemmingstoets een significant resultaat heeft opgeleverd, dan betekent dit nog niet dat, om bij ons oorspronkelijke voorbeeld te blijven, leerling a significant "beter" is dan leerling b, indien leerling a in de boven beschreven volgorde-schatting een lager rangnummer heeft dan b. De toets leert ons, indien we een significant resultaat gevonden hebben, slechts dat de leerlingen niet alle even begaafd zijn en geeft ons een schatting van de volgorde van begaafdheid. Wil men echter individuele verschillen nader onderzoeken, dan kan men gebruik maken van andere toetsingsmethoden, zoals bv. symmetrietoetsen, waarvan er enige later in deze cursus zullen worden besproken.

6. Voorbeeld uit de praktijk.

Van drie verschillende konijnen zijn op vaste tijden (2 u, 8 u, 14 u en 20 u) op 3 opeenvolgende dagen bloedmonsters genomen. Deze monsters zijn in telkamers gebracht waarbij met ieder konijn een bepaalde telkamer correspondeerde. De telkamers duiden wij aan met I, II en III. In ieder van de telkamers werd het aantal erythrocyten (rode bloedlichaampjes) per volume-eenheid op twee verschillende plaatsen gemeten. Deze plaatsen duiden wij aan met indices 1 en 2 bij de telkamernummers.

Tabel (II.5)

	2 u	8 u	14 u	20 u
I_1	321	297,6	272,6	298
II_1	295	278	276,3	289,6
III_1	265	234,6	247,3	257,6
I_2	316,5	301,3	269	299,6
II_2	294,5	272	274	290,6
III_2	258	236,3	237,3	258

Wij willen allereerst nagaan of er een zekere dagperiodiciteit in de waarnemingen zit. Wij gaan daartoe in iedere rij volgens opklimmende grootte nummeren.

Het resultaat is:

Tabel (II.6)

	2u	8u	14u	20u
I_1	1	3	4	2
II_1	1	3	4	2
III_1	1	4	3	2
I_2	1	2	4	3
II_2	1	4	3	2
III_2	$1\frac{1}{2}$	4	3	$1\frac{1}{2}$
	$6\frac{1}{2}$	20	21	$12\frac{1}{2}$

Men vindt: $S = 8\frac{1}{2}^2 + 5^2 + 6^2 + 2\frac{1}{2}^2 = 139,5$.

Met behulp van tabel 5c uit Kendall (1948) vindt men voor de exacte overschrijdingskans: 0,0003. Er is dus een zeer duidelijke dagperiodiciteit.

Het aantal erythrocyten is het geringste in de nacht, stijgt tussen 2 u en 8 u, en begint na de middag af te nemen.

Wij kunnen echter aan de cijfers van tabel (II.5) ook een onderzoek instellen, naar de in de verschillende telkamers getelde aantallen erythrocyten. Wij nummeren dan de getallen in iedere kolom volgens opklimmende grootte.

Het resultaat is:

Tabel (II.7)

	I_1	II_1	III_1	I_2	II_2	III_2
2 u	6	4	2	5	3	1
8 u	5	4	1	6	3	2
14 u	4	6	2	3	5	1
20 u	5	3	1	6	4	2
	20	17	6	20	15	6

Men vindt nu: $S = 6^2 + 3^2 + 8^2 + 6^2 + 1 + 8^2 = 210$.

In tabel 6 van Kendall (1948) vindt men dat voor dit geval het 0,01-punt ligt bij: 175,3, zodat tussen de aantallen erythrocyten per telkamerdeel, geteld op verschillende uren een overeenstemming bestaat, die significant is bij een onbetrouwbaarheidsdrempel 0,01. Bij nadere beschouwing van de tabel blijkt, dat de getelde hoeveelheid erythrocyten het grootste was in telkamer I en het kleinste in telkamer III, doch dat het onderscheid tussen de beide delen van één telkamer zeer gering is; het ligt dus voor de hand om de geconstateerde verschillen toe te schrijven aan het feit, dat in iedere telkamer het bloed van een ander konijn werd onderzocht.

Litteratuur:

M.G. Kendall (1947), The advanced theory of Statistics (3rd edition). London, Charles Griffin & Co.

M.G. Kendall (1948), Rank Correlationmethods. London, Charles Griffin & Co.

R.A. Fisher (1948), Statistical Methods (10th edition). London-Edinburgh, Oliver & Boyd.

R.A. Fisher and F. Yates (1949), Statistical Tables (3rd edition). London-Edinburgh, Oliver & Boyd.

P.G. Hoel (1947), Introduction to Mathematical Statistics. New York, J. Wiley & Sons.

K. Pearson (1934), Tables of the Incomplete B-function. Cambridge University Press.

Cursus Parametervrije methoden.

Errata in Hoofdstuk II.

Pag.	Regel	Correctie.
20	15 v.o.	"berekeneing" moet zijn "berekening";
22	1 v.o.	$\sum_{i=1}^{n}(x_i-x)^2$ moet zijn $\sum_{i=1}^{n}(x_i-\bar{x})^2$;
24	8 v.o.	(II.14) moet zijn (II.15a);
24	6 v.o.	(II.15) moet zijn (II.15b);
25	4 v.b.	"rangnummers-verwisselingen" moet zijn "rangnummerverwisselingen";
25	12 v.o.	Voor het schema toevoegen: "(II.17)";
26	5 v.b.	"8... $\frac{2}{6}$ ... $\frac{1}{6}$" moet zijn: "8 ... $\frac{1}{6}$ ... $\frac{1}{6}$" ;
27	5 v.o.	"m-rangschikkingen" moet zijn: "m rangschikkingen";
28	7 v.b.	"36" moet zijn "$36\frac{1}{2}$";
32	14 v.b.	"verdeling" moet zijn: "verdelingen ge-";
37	3 v.o.	achter "geen" toevoegen: "gelijke rangnummers)";
38	10 v.b.	In de open plekken tussen de teksthaken [ ] achter het
	15 v.o.	symbool P invullen: "S";
	11 v.o.	
	10 v.o.	
	8 v.o.	
	4 v.o.	Tweemaal achter "om" invullen "S";
	4 t/m 1 v.o.	Vóór index "max" invullen "S";
	11 v.o.	Achter "stelt" invullen "aan";
	5 v.o.	"s-1" moet zijn "S-1";
39	3 v.b.	Er moet staan: (II.48) $W' = \frac{S-1}{S_{max}+2} = \frac{S-1}{\frac{1}{12}m^2n(n^2-1)+2}$;
39	3 v.o.	Er moet staan: $W' = \frac{S-1}{\frac{1}{12}m^2n(n^2-1)+2}$;
40	4 v.b.	Teller van de breuk moet zijn: 12 S
41	——	Boven kolom 4 en kolom 8 moet staan: $P[\underline{S} \geqq S]$;
44	11 v.o.	Achter "aan de" invullen "hand".

Cursus Parametervrije methoden.

III. De toets van Wilcoxon voor het probleem van twee steekproeven.

Door H.R. van der Vaart.

Mei 1951.

§ 1. Inleiding.

1.1 Het probleem, zoals de practijk dit stelt.

Vaak wordt een statistisch antwoord gezocht op vragen als deze: m duiven zijn gedurende enige tijd gevoederd met een diëet A, en n (andere) duiven met een diëet B. Van elke duif is daarna bepaald, welk percentage van zijn lichaam uit vet bestond. Is nu met behulp van die m+n getallen (percentages) uit te maken, of die diëten een verschillende invloed op het vetgehalte van die duiven hebben gehad?

Problemen van deze soort duidt men aan als: het probleem van twee steekproeven. De onderzoekers, die een dergelijke vraag stellen, wensen — enigszins vaag uitgedrukt — te weten of "in doorsnee" diëet A een hoger (of lager) vetgehalte oplevert als eindresultaat dan diëet B, dan wel of beide diëten eenzelfde vetgehalte geven. Een andere vraag die in dit verband mogelijk zou zijn, n.l. of diëet A misschien een grotere (of kleinere) spreiding in de vetgehaltes geeft dan B, interesseert degenen, die dergelijke problemen onderzoeken, in het algemeen minder. De toets van Wilcoxon dient nu in hoofdzaak juist om het bedoelde verschil in niveau te ontdekken. Zij doet dit door gebruik te maken van een grootheid U, die we in § 1.2 zullen definiëren. Vooraf merken wij nog op dat bij de volgende beschouwingen vooronderstelt wordt, dat de (m+n) waarnemingen (stochastisch) onafhankelijk zijn, d.w.z. dat het resultaat van één van de waarnemingen niet beïnvloed wordt door het resultaat van één of meer van de overigen.

1.2 Definitie van de grootheid U.

We noemen de m,(resp. n) getallen, die de vetpercentages bij gebruik van het diëet A(resp. B) aangeven, x_1 , x_2 , ... , x_m resp.

y_1 , y_2 , ..., y_n. Voor de lig_ing van deze (m+n) getallen t.o.v. elkaar zijn er dan vele mogelijkheden. De uitersten zijn: "elk van de y-waarden is kleiner dan elk van de x-waarden" tegenover "elk van de y-waarden is groter dan elk van de x-waarden".

Wilcoxon's toets nu (zie (10))[1], in de vorm, die Mann en Whitney (7) er aan geven, (vgl. ook (1)) is gebaseerd op het getal U.[2] We geven aan de x-waarden de index i, (i = 1,...,m) en aan de y-waarden de index j (j = 1,...,n). U is dan het aantal paren (i,j), waarvoor geldt $y_j \leq x_i$. De beide zo juist genoemde uitersten geven blijkbaar U = mn, resp. U = 0. Blijkens de definitie van U, die in deze vorm alleen geldt als voor geen paar (i,j) $y_j = x_i$ is, is U een geheel getal met $0 \leq U \leq mn$. Het geval, dat voor sommige paren (i,j) wel $y_j = x_i$ is, wordt in § 5.4 behandeld.

Men voelt wel aan, dat men de hypothese, dat de diëten geen verschillen veroorzaken, niet zal kunnen verwerpen, als U "dicht bij" $\frac{1}{2}$mn ligt, en niet zal kunnen handhaven, als U "dicht bij" mn of 0 ligt. We dienen deze opmerking èn sommige beschouwingen van § 1.1 thans te preciseren.

1.3 Getoetste hypothese; alternatieve hypothesen.

Als we de hypothese, dat de diëten geen verschillen veroorzaken, "niet kunnen handhaven", moet iets anders er voor in de plaats komen. Dit andere pleegt men te noemen: de toelaatbare alternatieve hypothesen, die we tezamen aanduiden met de letter H. Men kan nu de toelaatbare alternatieve hypothesen bij Wilcoxon's toets ruwweg beschrijven als: "de resultaten van diëet B zijn verschoven ten opzichte van de resultaten van diëet A". Voorlopig laten we het hierbij, wat de alternatieven betreft. Deze kwestie wordt in § 3 nog nader toegelicht. Slechts merken we nog één ding op. Alvorens men Wilcoxon's toets gaat toepassen, dient men zich eerst te vergewissen, of de alternatieve hypothesen, die het probleem in kwestie om vakwetenschappelijke, dus niet-statistische redenen, toelaat, samenvallen met de in § 3 besproken, om statistisch-theoretische redenen, toelaatbare alternatieve hypothesen. Deze opmerking geldt trouwens voor elke statistische toets, waarbij iets van de alternatieven bekend is.

Nu de getoetste hypothese zelf. Deze zullen we met H_0 aangeven. Zij houdt in, dat er geen verschuiving optreedt of wel, precies gezegd: H_0 houdt in, dat x_1 ,..., x_m en y_1 ,..., y_n twee (random) steekproeven zijn uit éénzelfde verdeling.

---

1) Een enkel getal tussen haakjes verwijst naar de literatuurlijst op blz. 67

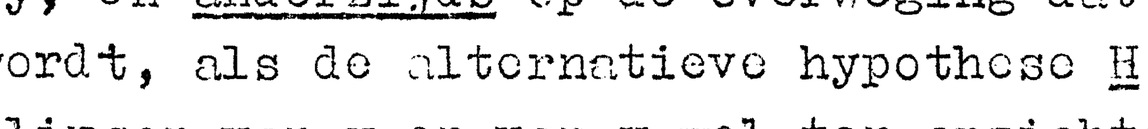
2) De definitie hiervan volgt.

Dit wil het volgende zeggen: Als voor elke i geldt, dat $P\left[\underline{x}_i \leq z\right] = F(z)$ en voor elke j, dat $P\left[\underline{y}_j \leq z\right] = G(z)$, dan houdt H_o in dat $\underline{F(z) \equiv G(z)}$.

Hierbij moet aangetekend worden, dat F en G continu verondersteld moeten kunnen worden, opdat men Wilcoxon's toets kan toepassen. De hieronder gegeven redeneringen maken namelijk gebruik van: $P\left[\underline{x}_i = \underline{y}_j\right] = 0$ voor elk paar (i,j), d.w.z. dat geen x gelijk zal zijn aan een y. Bovendien moeten de integralen in het begin van 2.1 door andere uitdrukkingen worden vervangen als F en G niet continu zijn. Met behulp van een kunstgreep kan deze continuiteitsvoorwaarde nog door een minder stringente voorwaarde worden vervangen. Vergelijk hierover § 5.4.

1.4 Kritiek gebied.

Wanneer verwerpen we de hypothese H_o nu? Dit volgt uit de opmerking aan het eind van § 1.2: Als de voor U gevonden waarde dicht bij mn of bij 0 ligt.



Dit wil zeggen: we kiezen een getal s met $0 \leq s < \frac{mn}{2}$ en spreken het volgende af: H_o wordt verworpen als $0 \leq U \leq s$ of $mn-s \leq U \leq mn$. Deze twee intervallen voor U vormen dus samen het zgn. kritieke gebied w. Dat het kritieke gebied w van deze toets opgebouwd wordt uit twee dergelijke intervallen, uit twee dergelijke staarten, berust enerzijds op de overweging, dat het wel zeer onwaarschijnlijk is dat zowat alle y-waarden groter resp. kleiner zouden zijn dan zowat alle x-waarden, als de getoetste hypothese H_o juist is, dus als de verdeling van x niet verschoven is t.o.v. die van y, en anderzijds op de overweging dat de kans daarop groter wordt, als de alternatieve hypothese H juist is, dus als de verdelingen van x en van y wel ten opzichte van elkaar verschoven zijn.

Hieruit volgt, dat in verband met de eigenschappen van de toets de volgende twee waarschijnlijkheden van groot belang zijn: $P\left[\underline{U} \in w \mid H_o\right]$ en $P\left[\underline{U} \in w \mid H\right]$. Over de waarschijnlijkheid $P\left[\underline{U} \in w \mid H_o\right]$ zullen we het hebben in § 2, van de waarschijnlijkheid $P\left[\underline{U} \in w \mid H\right]$ behandelen we enkele eigenschappen in § 3. Vooraf merken we nog het volgende op. We willen natuurlijk niet graag een verkeerde beslissing nemen. Daar ($U \in w$) betekent: verwerping van H_o en (U niet $\in w$): niet verwerping van H_o, zou het dus ideaal zijn, als de intervallen, waaruit w is opgebouwd, zó gekozen konden

worden, dat $P[\underline{U} \in w \mid H_o] = 0$ (1.4,1)

en dat $P[\underline{U} \in w \mid H] = 1.$ (1.4,2)

Het is echter duidelijk, dat (1.4,1) slechts is te bereiken door w leeg te maken (geen enkele U-waarde in w), en (1.4,2) door voor w te nemen $[o \leqq U \leqq mn]$, zodat aan deze eisen nooit tegelijk voldaan kan worden. Al wat we kunnen doen, is dus s zo te kiezen, dat $P[\underline{U} \in w \mid H_o] = \alpha$ zo <u>dicht mogelijk bij 0</u> ligt — dus b.v. $\alpha = 0{,}01$ of $\alpha = 0{,}05$ of zo klein als de desbetreffende onderzoeker nodig acht — en $P[\underline{U} \in w \mid H]$ zo <u>dicht mogelijk bij 1</u>.

Wat de <u>tweede</u> eis betreft, hieraan voldoet een "goede" statistische toets beter dan een "slechte" en voorts voldoen grotere steekproeven hieraan meestal beter dan kleine (vgl. ook § 3). Wilcoxon's toets is gebleken een meer dan redelijk "goede" toets te zijn.

Wat de <u>eerste</u> eis betreft, bij gegeven toetsingstype (in dit geval: Wilcoxon) en gegeven steekproefgrootten (m waarden van x en n waarden van y) hangt $P[\underline{U} \in w \mid H_o]$ geheel van s af (welk getal de grenzen van w bepaalt, zie begin van 1.4). Voor de berekening van $P[\underline{U} \in w \mid H_o]$ heeft men de waarschijnlijkheidsverdeling van $\underline{U}$ onder de hypothese H_o nodig. Deze gaan we thans behandelen.

<u>§ 2. De waarschijnlijkheidsverdeling van U onder de getoetste hypothese H_o.</u>

<u>2.1 De exacte verdeling.</u>

We willen nu berekenen $P_{m,n}[\underline{U} = U \mid H_o]$. Hierin geeft H_o dus aan: $F \equiv G$, U is een geheel getal met $0 \leqq U \leqq mn$ en de indices m,n geven te kennen, dat er m x-waarden zijn (met cumulatieve verdelingsfunctie F(x)) en n y-waarden (met verdelingsfunctie G(y)). We beginnen eens met het geval $m = n = 2$. Er zijn dan de volgende mogelijkheden van volgorden:

$x_1 < x_2 < y_1 < y_2$

$x_2 < x_1 < y_1 < y_2$

$x_1 < x_2 < y_2 < y_1$

$x_2 < x_1 < y_2 < y_1$

$x_1 < y_1 < x_2 < y_2$

Er zijn evenveel zulke rijtjes als er permutaties zijn van 2+2=4 elementen, in het algemeen: van (m + n) elementen. Dus er

zijn 4!, in het algemeen (m + n)! van zulke rijtjes. Wegens de continuïteit van F geldt:

$$P[x_1 < x_2 < y_1 < y_2] = P[x_1 \leqq x_2 \leqq y_1 \leqq y_2] =$$

$$= \int_{-\infty}^{+\infty} dF(y_2) \int_{-\infty}^{y_2} dF(y_1) \int_{-\infty}^{y_1} dF(x_2) \int_{-\infty}^{x_2} dF(x_1) = (\text{subst. } x_1 = x_2,\ x_2 = x_1) =$$

$$= \int_{-\infty}^{+\infty} dF(y_2) \int_{-\infty}^{y_2} dF(y_1) \int_{-\infty}^{y_1} dF(x_1) \int_{-\infty}^{x_1} dF(x_2) = P[x_2 < x_1 < y_1 < y_2] = \text{enz.}$$

Dus de waarschijnlijkheid van elk van de (m + n)! permutaties van $x_1,\ldots,x_m,y_1,\ldots,y_n$ is even groot. Samen zijn deze waarschijnlijkheden gelijk aan 1. Dus

$$P[x_1 < x_2 < y_1 < y_2] = \frac{1}{(m+n)!} = P[\text{enz.}] \ .$$

Nu hebben de eerste vier van de hierboven gegeven rijtjes één ding gemeen, waarin ze bovendien van de volgende verschillen. Als we de indices weglaten, wordt elk van de eerste vier n.l. xxyy. Het vijfde daarentegen wordt xyxy. We zeggen: de eerste vier rijtjes vormen éénzelfde x-y-rangschikking. Hoeveel rijtjes met waarschijnlijkheid = $\frac{1}{(m+n)!}$ vormen tezamen één x-y-rangschikking?

Over de m x-plaatsen kunnen $x_1,\ldots,x_m$ gepermuteerd worden (m! permuties). Over de n y-plaatsen geldt hetzelfde voor $y_1,\ldots,y_n$ (n! permuties). Elk van de m! x-permuties kan worden gecombineerd met elk van de n! y-permuties. Dus m!n! rijtjes met waarschijnlijkheid = $\frac{1}{(m+n)!}$ vormen tezamen één x-y-rangschikking.

Dit geldt voor elk van de denkbare x-y-rangschikkingen. Dus voor elke denkbare x-y-rangschikking geldt:

$$P_{m,n}[\text{ één bepaalde x-y-rangschikking}] = \frac{m!n!}{(m+n)!}$$

We behoeven dus slechts te tellen, hoeveel x-y-rangschikkingen gemeen hebben dat voor hen $\underline{U} = U$ is, om $P[\underline{U} = U]$ te kennen. Als dit aantal n.l. gelijk is aan $p'_{m,n}(U)$ is, dan is

$$P[\underline{U} = U] = \frac{m!n!}{(m+n)!}\, p'_{m,n}(U). \qquad (2.1,1)$$

Voor dit getal $p'_{m,n}(U)$ is gemakkelijk een recurrente betrekking te vinden, gelijk Mann en Whitney opmerkten. Alle $p'_{m,n}(U)$ x-y-rangschikkingen vallen n.l. uiteen in:

a x-y-rangschikkingen, waarvan de laatste een y is,
b x-y-rangschikkingen, waarvan de laatste een x is.

Het aantal van groep a telt men door die laatste y weg te laten: $p'_{m,n-1}(U)$. Het aantal in groep b telt men door die laatste x weg te laten.: $p'_{m-1,n}(U-n)$

Zo vindt men:

$$p'_{m,n}(U) = p'_{m-1,n}(U-n) + p'_{m,n-1}(U) \qquad (2.1,2)$$

Door in het voorgaande "laatste" te vervangen door "eerste", vindt men

$$p'_{m,n}(U) = p'_{m,n-1}(U-m) + p'_{m-1,n}(U) \qquad (2.1,2a)$$

Uit elke x-y-rangschikking, waarvoor $\underline{U} = U$ is, kan men door van achter naar voren op te schrijven één en slechts één x-y-rangschikking verkrijgen, waarvoor $\underline{U} = mn - U$ is.

Voorbeeld: x x y x y y x y x x

Hierbij blijven m en n ongewijzigd. Hieruit volgt:

$$p'_{m,n}(U) = p'_{m,n}(mn - U)$$

dus, daar

$$p'_{m,n}(U) = \frac{(m+n)!}{m!\,n!}\, P\left[\underline{U} = U\right],$$

is

$$P_{m,n}\left[\underline{U} = U\right] = P_{m,n}\left[\underline{U} = mn - U\right] \qquad (2.1,3),$$

d.w.z. de verdeling van U is symmetrisch t.o.v. $\frac{mn}{2}$, zodat voor het gemiddelde geldt:

$$\mu = E(\underline{U}) = \frac{mn}{2} \qquad (2.1,4)$$

Ook kan men uit elke x-y-rangschikking, waarvoor $\underline{U} = U$, door elke x door y en elke y door x te vervangen één en slechts één x-y-rangschikking verkrijgen, waarvoor $\underline{U} = mn - U$.

Voorbeeld x x y x y y y x y x .

Hierbij komen er n x-waarden in plaats van m en m y-waarden in plaats van n.

Hieruit volgt

$$p'_{m,n}(U) = p'_{n,m}(mn - U),$$

dus

$$P_{m,n}\left[\underline{U} = U\right] = P_{n,m}\left[\underline{U} = mn - U\right] \qquad (2.1,5)$$

Uit (2.1,3) volgt

$$P_{n,m}\left[\underline{U} = mn - U\right] = P_{n,m}\left[\underline{U} = U\right], \text{ zodat}$$

$$P_{m,n}\left[\underline{U} = U\right] = P_{n,m}\left[\underline{U} = U\right]. \qquad (2.1,6)$$

Hieruit volgt, dat onder de getoetste hypothese H_o de verdeling van $\underline{U}$ voor m x-waarden en n y-waarden dezelfde is als voor n x-waarden en m y-waarden. Dit is van veel belang voor de tabellering, daar deze hierdoor aanzienlijk bekort kan worden; immers men behoeft nu slechts voor $m \leqq n$ te tabelleren.

Door de redenering, die tot (2.1,2) voerde, toe te passen voor $U < n$, ziet men, dat tot de beginvoorwaarden van (2.1,2) behoort:

$$p'_{m,n}(k) = 0, \text{ als } k < 0 \qquad (2.1,7)$$

Wegens (2.1,3) volgt hieruit:

$$p'_{m,n}(k) = 0, \text{ als } k > mn \qquad (2.1,7a)$$

Door de redeneringen, die tot (2.1,2) en (2,1,2a) voeren, toe te passen voor $m = 1$ en / of $n = 1$, ziet men dat de beginvoorwaarden gecompleteerd worden door

$$p'_{0,i}(k) = p'_{i,0}(k) = \begin{cases} 0, \text{ als } k \neq 0 \\ 1, \text{ als } k = 0 \end{cases} \qquad (2.1,8)$$

Zonder meer is duidelijk, dat

$$p'_{m,n}(0) = p'_{m,n}(1) = p'_{m,n}(mn) = p'_{m,n}(mn-1) = 1. \qquad (2.1,9)$$

Met behulp van de voorgaande formules kan men in principe voor elke gehele $m > 0$, $n > 0$ en $U \geqq 0$ de grootheden $p'_{m,n}(U) = p'_{m,n}(mn-U)$, dus ook $P_{m,n}[\underline{U} = U] = P_{m,n}[\underline{U} = mn-U]$, en $P_{m,n}[\underline{U} \leqq U] = P_{m,n}[\underline{U} \geqq mn-U]$ berekenen.

We zullen enkele voorbeelden geven, opdat de lezer zichzelf kan controleren:

m	n	U	$p'_{m,n}(U)$	$P_{m,n}[\underline{U}=U]$	$P_{m,n}[\underline{U} \leqq U]$
1	1	0	1 (uit (2.1,9))	$\frac{1}{2}$	$\frac{1}{2}$
		1	1 (uit (2.1,9))	$\frac{1}{2}$	1
1	2	0	1	$\frac{1}{3}$	$\frac{1}{3}$
		1	1	$\frac{1}{3}$	$\frac{2}{3}$
1)		2	1 (uit (2.1,5))	$\frac{1}{3}$	1
2	2	0	1	$\frac{1}{6}$	$\frac{1}{6}$
		1	1	$\frac{1}{6}$	$\frac{1}{3}$
		2	2 (uit (2.1,2))	$\frac{2}{6}$	$\frac{2}{3}$
		3	1	$\frac{1}{6}$	$\frac{5}{6}$
		4	1	$\frac{1}{6}$	1

1) Wegens (2.1,6) behoeft het geval m=2, n=1; het geval m=3, n=1 enz. niet getabelleerd te worden.

m	n	U	$p'_{m,n}(U)$	$P_{m,n}[\underline{U}=U]$	$P_{m,n}[\underline{U} \leqq U]$
1	3	0	1	$\frac{1}{4}$	$\frac{1}{4}$
		1	1	$\frac{1}{4}$	$\frac{1}{2}$
		2	1	$\frac{1}{4}$	$\frac{3}{4}$
		3	1	$\frac{1}{4}$	1
2	3	0	1	$\frac{1}{10}$	0,1
		1	1	$\frac{1}{10}$	0,2
		2	2	$\frac{2}{10}$	0,4
		3	2	$\frac{2}{10}$	0,6
3	3	0	1	$\frac{1}{20}$	0,05
		1	1	$\frac{1}{20}$	0,10
		2	2	$\frac{2}{20}$	0,20
		3	3	$\frac{3}{20}$	0,35
		(5	3	$\frac{3}{20}$	0,65)

1)

enz.

In overeenstemming met § 1.4 kunnen we nu s zo kiezen dat $P[\underline{U} \in w | H_o] = P[\underline{U} \leqq s \text{ of } \underline{U} \geqq mn-s | H_o] \leqq \alpha \leqq 0,01$ of $0,05$.

Als we dan voor U een waarde vinden, die tussen de grenzen s en mn-s ligt, verwerpen we de hypothese H_o niet; anders wel.

Nadat we hebben uiteengezet, hoe de verdeling van $\underline{U}$ onder de hypothese H_o numeriek berekend kan worden, berekenen we nog

$$\mu = E(\underline{U}) \text{ en } \sigma^2 = E[(\underline{U}-\mu)^2] = E(\underline{U}^2) - \mu^2.$$

We zagen reeds dat $\mu = \frac{mn}{2}$ (2.1,4), maar zullen dit nog op een andere manier bewijzen en tevens σ^2 berekenen (Deze berekeningen

---

1) Wegens de symmetrie van de verdeling hoeft $U > \frac{mn}{2}$ niet getabelleerd te worden.

berekeningen staan ongeveer zo in van Dantzig (2), punt 3).

Stel $\varphi(z) = \begin{cases} 1 \text{ voor } z \geq 0 \\ 0 \text{ voor } z < 0 \end{cases}$ en $\underline{z}_{ij} = i(\underline{x}_i - \underline{y}_j)$, dan geldt voor elk paar (i,j) met $y_j < x_i$, dat $z_{ij} = 1$ en voor elk paar (i,j) met $y_j > x_i$, dat $z_{ij} = 0$.

Bijgevolg is $\underline{U} = \sum_1^m \sum_1^n \underline{z}_{ij}$.

Daar wegens de continuïteit

$$P[\underline{y}_j = \underline{x}_i \mid H_o] = P[\underline{z}_{ij} = 0 \mid H_6] = 0$$

en daar

$$P[\underline{y}_j < \underline{x}_i \mid H_o] = P[\underline{z}_{ij} = 1 \mid H_o] = \frac{1}{2} = P[\underline{z}_{ij} = 0 \mid H_o] = P[\underline{y}_j > \underline{x}_i \mid H_o]$$

is $\mathcal{E}(\underline{z}_{ij}) = \frac{1}{2}$ en

$$\mathcal{E}(\underline{U}) = \mathcal{E}\left(\sum_1^m \sum_1^n \underline{z}_{ij}\right) = \sum_1^m \sum_1^n \mathcal{E}(\underline{z}_{ij}) = \frac{mn}{2} \qquad (2.1,10)$$

Voorts is

$$\sigma^2 = \mathcal{E}\left[\left(\sum_1^m \sum_1^n \underline{z}_{ij}\right)^2\right] - \mu^2 = -\mu^2 + \sum_1^m \sum_1^n \sum_1^m \sum_1^n \mathcal{E}(\underline{z}_{ij}\underline{z}_{kl}) \qquad (2.1,11)$$

Nu geldt voor onafhankelijke stochastische grootheden $\underline{v}$ en $\underline{w}$, dat $\mathcal{E}(\underline{v}\underline{w}) = \mathcal{E}(\underline{v})\mathcal{E}(\underline{w})$. Daar $\underline{z}_{ij}$ en $\underline{z}_{kl}$ alleen dan onafhankelijk zijn, als $i \neq k$ en $j \neq l$, moet de 4-voudige som in (2.1,11) gesplitst worden in de sommen

$$\sum_1^m \sum_1^n \sum_1^m \sum_{\substack{1 \\ i \neq k}}^n{}^{j \neq l} = \sum_{i,j,k,l}$$

$$\sum_1^m \sum_1^n \sum_1^m \sum_{\substack{1 \\ i = k}}^n{}^{j \neq l} = \sum_{i,j,i,l}$$

$$\sum_1^m \sum_1^n \sum_1^m \sum_{\substack{1 \\ i \neq k}}^n{}^{j = l} = \sum_{i,j,k,j}$$

$$\sum_1^m \sum_1^n \sum_1^m \sum_{\substack{1 \\ i = k}}^n{}^{j = l} = \sum_{i,j,i,j}$$

$m(m-1)n(n-1)$	termen
$mn(n-1)$	termen
$m(m-1)n$	termen
mn	termen
$m^2n(n-1) + m^2n = m^2n^2$	termen.

Nu geldt:

$$\mathcal{E}(\underline{z}_{ij}\underline{z}_{kl}) = \mathcal{E}(\underline{z}_{ij})\mathcal{E}(\underline{z}_{kl}) = \frac{1}{4}.$$

$$\mathcal{E}(\underline{z}_{ij}\underline{z}_{il}) = P[\underline{z}_{ij} = 1 \text{ èn } \underline{z}_{il} = 1] = P[\underline{y}_j < \underline{x}_i \text{ èn } \underline{y}_l < \underline{x}_i \mid H_o] =$$

$$\int_{-\infty}^{+\infty} dF(x_i) \int_{-\infty}^{x_i} dF(y_j) \int_{-\infty}^{x_i} dF(y_l) = \int_{-\infty}^{+\infty} F^2(x_i).dF(x_i) = \frac{1}{3}$$

$$\mathcal{E}(\underline{z}_{ij}\underline{z}_{kj}) = P[\underline{y}_j < \underline{x}_i \text{ èn } y_j < x_k \mid H_o] = \int_{-\infty}^{+\infty} dF(y_j) \int_{y_j}^{+\infty} dF(x_i) \int_{y_j}^{+\infty} dF(x_k) =$$

$$= \int_{-\infty}^{+\infty} [1 - F(y_j)]^2 dF(y_j) = \frac{1}{3}.$$

$$\mathcal{E}(\underline{z}_{ij}\underline{z}_{ij}) = P[z_{ij} = 1] = P[\underline{y}_j < \underline{x}_i \mid H_o] = \frac{1}{2}.$$

Zodat $\mathcal{E}(\underline{U}^2) = \frac{1}{4}mn(m-1)(n-1) + \frac{1}{3}mn(m+n-2) + \frac{1}{2}mn =$

$$= \frac{1}{12}mn[3mn - 3m - 3n + 3 + 4m + 4n - 8 + 6] = \frac{1}{12}mn[3mn + m + n + 1]$$

en $\sigma^2 = \mathcal{E}(\underline{U}^2) - \mu^2 = \frac{1}{12}\,mn(3mn+m+n+1) - \frac{m^2n^2}{4} =$

$$\sigma^2 = \frac{1}{12}\,mn(m+n+1) \qquad (2.1,12)$$

2.2 Benadering voor grote steekproeven.

Men kan bewijzen, dat voor grotere m en n (in de practijk >10) en wanneer $\frac{m}{n}$ niet te veel van 1 verschilt, de volgende benaderingen zeer bruikbaar zijn:

$$\underline{a}.\quad P_{m,n}\left[\underline{U} \leqq s\right] \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{v} e^{-\frac{1}{2}t^2}\,dt \qquad (2.2,1a)$$

$$\text{waar}\quad v = \frac{s+\frac{1}{2}-\mu}{\sigma} = \frac{s+\frac{1}{2}-\frac{mn}{2}}{\sqrt{\frac{1}{12}mn(m+n+1)}}\,.$$

$$\underline{b}.\quad P_{m,n}\left[\underline{U} \geqq s'\right] \approx \frac{1}{\sqrt{2\pi}} \int_{v'}^{\infty} e^{-\frac{1}{2}t^2}\,dt \qquad (2.2,1b)$$

$$\text{waar } v' = \frac{s'-\frac{1}{2}-\mu}{\sigma} = \frac{s'-\frac{1}{2}-\frac{mn}{2}}{\sqrt{\frac{1}{12}mn(m+n+1)}}$$

(De $+\frac{1}{2}$, resp. $-\frac{1}{2}$, in de teller is de z.g. continuïteitscorrectie)

Kemperman ((6),p.9) heeft gevonden, dat voor $m \gg n$, zelfs voor $n < 10$ een zeer bruikbare benadering gegeven wordt door:

$$P_{m,n}\left[\underline{U} \leqq s\right] \approx \frac{1}{n!} \sum_{0 \leqq k < V} (-1)^k \binom{n}{k} (V-k)^n \qquad (2.2,2)$$

$$\text{waarin}\quad V = \frac{n}{2} + \frac{s+\frac{1}{2}-\frac{mn}{2}}{\sqrt{m(m+n+1)}}$$

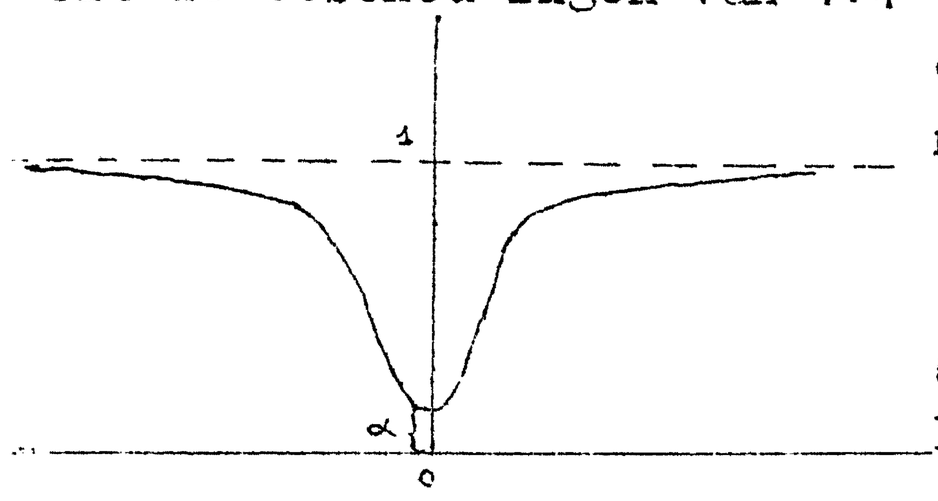
Daar het om redenen, die in § 3 kort besproken zullen worden, echter wenselijk is om, indien enigszins mogelijk, $\frac{m}{n}$ niet veel van 1 te laten verschillen, zal men deze ingewikkelde benadering (2.2,2) niet vaak nodig hebben.

§ 3. Enkele opmerkingen over het onderscheidingsvermogen, d.i. over de waarschijnlijkheidsverdeling van U onder de alternatieve hypothesen H.

Deze opmerkingen sluiten aan bij het in § 1.4 besprokene. Zij zullen gemaakt worden aangaande $P[\underline{U} \in w|H]$ voor kleine steekproeven (cf. van der Vaart (9))) en voor grote (cf. van Dantzig (2)).

Wat de kleine steekproeven betreft, hierbij zullen de besproken alternatieve hypothesen beperkt worden tot starre verschuivingen van de verdelingsfunctie, d.w.z. tot $G(z) = F(z+\mu)$ met $\mu \neq 0$, waarbij, evenals in § 3, F de verdelingsfunctie van x en G die van y is. Voor $P[\underline{U} \in w|H]$ kan men dan schrijven $P[\underline{U} \in w|\mu]$.

Uit de beschouwingen van 1.4 blijkt dan, dat bijgaande figuur een indruk geeft van het zó ideaal mogelijke verloop van $P[\underline{U} \in w|\mu]$:



a. voor $\mu=0$ is $P \leqq \alpha$, α klein.

b. voor $\mu \neq 0$ sluit P zo dicht mogelijk aan bij het ideale gedrag $P = 1$, d.w.z.:

b.1. voor $\mu \neq 0$ is $P[\underline{U} \in w|\mu] > P[\underline{U} \in w|\ 0]$.

b.2. de grafiek van P vertoont een scherpe "punt" bij $\mu=0$ d.w.z.

$\left(\frac{d^2P}{d\mu^2}\right)_{\mu=0}$ is "zo groot mogelijk".

Punt a is in § 2 uitvoerig behandeld.

Aangaande b.1 is op te merken, dat in het geval $m \neq n$ bewezen is, dat voor sommige verdelingsfuncties F(x), waarvan de eerste afgeleide — de waarschijnlijkheidsdichtheid F'(x) — geen punt van symmetrie heeft , deze eigenschap niet geldt, daar vóór deze functies bovenstaande grafiek voor $\mu = 0$ een niet-horizontale raaklijn heeft. Hieruit volgt, dat men, indien enigszins mogelijk, m = n moet nemen, d.w.z. de uitgebreidheid van de beide steekproeven (die der x en die der y) even groot moet maken, omdat anders, zelfs voor deze zeer beperkte klasse van alternatieve hypothesen, aan b.1 niet is voldaan.

Onder deze voorwaarde is er namelijk slechts één bepaalde, nogal abnormale, en wel U-vormige, verdeling gevonden kunnen worden, waarvoor de eigenschap b.1 niet gold, zodat men, als m = n is, tamelijk wel erop kan vertrouwen, dat aan b.1 is voldaan.

Aangaande b.2: hierin verschillen "goede" en "slechte" statistische toetsen van elkaar (vgl. § 1.4). Wanneer nu de alternatieve hypothesen bestaan uit $G(z) = F(z+\mu)$ met $\mu \neq 0$, terwijl bovendien nog normaliteit wordt ondersteld, d.w.z.:

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}t^2}\, dt,$$

kan men, zoals bekend, de hypothese $\mu = 0$ toetsen tegen de alternatieven $\mu \neq 0$ met de toets van Student, die is immers een toets is voor het verschil van de gemiddelden van twee normale verdelingen met dezelfde spreiding.

Het kritieke gebied w_{st} van deze toets is van de vorm $|\underline{t}| \geqq t_{\frac{1}{2}\alpha}$, waarbij:

1: $\alpha = P[\underline{t} \in w_{st}|H_0]$;

2: $$t=\sqrt{\frac{mn(m+n-2)}{m+n}}\cdot\frac{\bar{y}-\bar{x}}{\sqrt{(x_i-\bar{x})^2+(y_j-\bar{y})^2}}$$

met $\bar{x}=\frac{1}{m}\sum_1^m x_i$ en $\bar{y}=\frac{1}{n}\sum_1^n y_j$; en

3: $$\tfrac{1}{2}\alpha=\frac{1}{\sqrt{(m+n-2)\pi}}\cdot\frac{\Gamma\left(\frac{m+n-1}{2}\right)}{\Gamma\left(\frac{m+n-2}{2}\right)}\int_{t_{\frac{1}{2}\alpha}}^{\infty}\left(1+\frac{x^2}{m+n-2}\right)^{-\frac{m+n-1}{2}}dx$$

Nu is deze toets van Student onder al die toetsen van de hypothese H_o (inhoudende dat $\mu=0$ is), die aan eis b.1 voldoen, de toets, waarbij aan eis b.2 het best voldaan is; tenminste als de onderstelling van normaliteit en gelijkheid der spreiding vervuld is. Dat wij in § 1.4 opmerkten, dat Wilcoxon's toets meer dan redelijk goed was, berust op de volgende tabel, waaruit weliswaar blijkt, dat Wilcoxon iets minder"scherp"is dan Student, maar dat het verschil zeer gering is. Tegenover dit onbelangrijk kleine verschil staat dan de grote winst, dat Wilcoxon ook voor niet- normale verdelingen volkomen exact is - dezelfde winst, die voor alle parametervrije methoden geldt

Tabel van $\alpha''=\left(\frac{d^2P}{d\mu^2}\right)_{\mu=0}$

m	n	α [1]	$\alpha''_{Wi}(0)$	$\alpha''_{St}(0)\ \&\ \alpha''_{wi}(0)$
2	2	1/3	0,3675	0,00282
2	3	1/5	0,39132	0,01138
2	4	2/15	0,36755	0,01863
2	4	4/15	0,51398	0,02672
3	3	1/10	0,34698	0,01261
3	3	1/5	0,52282	0,02024

Voor m en $n\to\infty$ blijkt bovendien te gelden $\frac{\alpha''_{wi}(0)}{\alpha''_{st}(0)}\approx\frac{3}{\pi}$

Wat overigens de grote steekproeven betreft, de eigenschap $P[\underline{U}\in w|H]=1$, waaraan voor kleinere steekproeven niet voldaan is, blijkt asymptotisch wel te gelden.

---

[1] Voor de grootte van α bezit men bij dergelijke kleine aantallen slechts een geringe keuze; vandaar de min of meer bizarre waarden in deze kolom.

Men duidt deze eigenschap aan met de uitdrukking: de toets is asymptotisch onderscheidend voor H. Prof. van Dantzig (2) heeft nl. behalve enkele meer gecompliceerde resultaten, het volgende bewezen:

Als $m \to \infty$, $n \to \infty$ (met m/n constant), dan is:

$$\lim. P[\underline{U} \in w \mid H] = 1,$$

waarbij H zelfs veel ruimer is dan starre verschuivingen. Deze eigenschap geldt nl. als H wordt gedefinieerd door:

$$P[\underline{y} < \underline{x}] = P[\underline{x}-\underline{y} > 0] \neq \tfrac{1}{2}, \text{ d.w.z.}$$

de alternatieven zijn te beschrijven als: de mediaan van $(\underline{x}-y) \neq 0$ [1].

In aansluiting op het gezegde aan het slot van de eerste alinea van 1.3 volgt hieruit, dat men zich vóór toepassing van Wilcoxon's toets wel moet overtuigen of het onderzoek er iets aan heeft, wanneer men zijn conclusie zou gieten in de vorm: de kans dat $\underline{y} < \underline{x}$ is kleiner dan $\frac{1}{2}$, dan wel $> \frac{1}{2}$. Het wil ons echter voorkomen, dat dát resultaat voortreffelijk aansluit bij het vage idee, dat de in 1.1 bedoelde onderzoekers koesteren. Voor onze duiven betekent het immers, dat, de kans, dat een met diëet B gevoede duif een groter vetgehalte krijgt dan een met diëet A gevoede, kleiner is dan $\frac{1}{2}$.

§ 4. Voorbeeld.

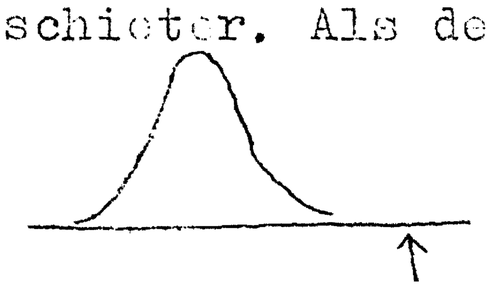
Wij zullen nu een voorbeeld behandelen, waarbij drie groepen duiven werden onderzocht op hun vetgehalte, nadat ze gedurende enige tijd elk met een ander diëet waren gevoederd.
De getallen waren als volgt:

Groep I: 1,42/2,22/1,68/1,68/2,60/2,54/7,80/2,28/;
Groep II: 3,36/3,77/4,47/3,89/1,63/4,75/;
Groep III: 3,02/3,42/6,43/2,93/2,23/7,85/.

Er doet zich hier nog een probleem voor aangaande het getal 7,80 in groep I. Dit getal wijkt wel erg sterk af van de andere getallen van groep I. Men zou geneigd kunnen zijn, deze waarneming als "uitschieter" te betitelen en niet te gebruiken. Maar hieraan zijn, zoals steeds, bezwaren verbonden.

---

[1] De getoetste hypothese H_o moet omschreven worden als $F \equiv G$. De formulering $P[\underline{x} - \underline{y} > 0] = \frac{1}{2}$ (mediaan van $(\underline{x} - \underline{y}) = 0$) zou te ruim zijn, daar de verdeling van $\underline{U}$ is afgeleid onder de hypothese $F \equiv G$.

Ten eerste: hoewel groep III als geheel iets hoger ligt (op het eerste gezicht) dan groep I, maakt toch het voorkomen van 6 6,43 én 7,85 in groep III wat huiverig om in groep I 7,80 als uitschieter weg te gooien. Ten tweede: om op statistische gronden te beslissen of een bepaalde waarneming een uitschieter is of niet, moet men vanzelfsprekend vrij stringente onderstellingen aangaande de verdelingsfunctie maken. Als de waarschijnlijkheidsdichtheid aldus loopt  , is een getal bij het pijltje geen uitschieter. Als de waarschijnlijkheidsdichtheid aldus loopt,  echter wel.

Gebruikt men nu de toets van Student, dan veronderstelt men toch al normaliteit en kan men deze onderstelling ook wel gebruiken om een uitschietertoets als die van Grubbs (3) toe te passen. Gebruikt men echter Wilcoxon, dan zou een dergelijke handelwijze inconsequent zijn, omdat Wilcoxon geen normaliteit veronderstelt. Nu zal echter blijken, dat Wilcoxon's toets veel minder in de war wordt gestuurd door dat ene getal 7.80 dan Student. Dit is ook theoretisch onmiddellijk in te zien. Men vindt immers b.v. bij vergelijkingen van groep I en II dezelfde waarde s voor $\underline{U}$, onverschillig of er nu 7,80 of 4,80 staat. Voor Student maakt het echter een groot verschil, of er een bijdrage $(7,80-2,8)^2$ of $(4,30-2,1)^2$ aan de kwadratensom wordt geleverd. (2,8 resp. 2,1 is ongeveer het gemiddelde van groep I). Wij kunnen dus concluderen, dat men zich bij Wilcoxon veel minder het hoofd hoeft te breken over het al of niet optreden van uitschieters.

Hieronder worden nu Wilcoxon en Student toegepast voor vergelijking van de volgende paren van waarnemingsgroepen:
I-II, I'(=I zonder 7,80)-II, I-III, I'-III, II-III, I-(II+III), I'-(II+III).

Onder k_{wi} en k_{st} zijn getabelleerd de (tweezijdige) overschrijdingskansen, berekend met de toets van Wilcoxon, resp. Student. D.i. dus:

$$k_{wi}=P\left[\left|\underline{U}-\mu\right| \geqq \left|U-\mu\right| \;\middle|\; H_o\right] \text{ resp. } k_{st}= P\left[\underline{t} \geqq |t| \;\middle|\; H_o\right],$$

waarin U resp. t de bij de proef gevonden waarden van $\underline{U}$ resp. Student's $\underline{t}$ zijn.

	m	n	U	k_{wi}	k_{st}	t
I-II	8	6	12	0,1419	0,374	0,9243
I'-II	7	6	6	0,0350	0,005	3,4704
I-III	8	6	8	0,0426	0,212	1,3183
I'-III					0,025	2,5843
II-III	6	6	19[1)]	0,9372	0,532	0,6488
I-(II+III)	8	12	20	0,0338	0,178	1,4039
I'-(II+III)	7	12	9	0,0060	0,012	2,8341

1) Voor het bepalen van de overschrijdingskans volgt uit 19 de waarde 36-19=17 als s-waarde $< \frac{mn}{2}$.

§ 5. Slotopmerkingen.

5.1 Vergelijking van Wilcoxon's toets met die van Student.

Wanneer we nu terugzien op het behandelde, kunnen we het volgende opmerken, wanneer we Wilcoxon's toets voor twee steekproeven vergelijken met die van Student:

1. Wilcoxon vereist veel minder rekenwerk dan Student. Student te berekenen zonder (zelfs met) rekenmachine is bijna steeds een straf. De bepaling van de waarde, die U voor het onderzochte steekproevenpaar aanneemt vraagt niets anders dan na te gaan of elk van n getallen groter of kleiner is dan elk van m getallen.

2. Om Student te kunnen toepassen moet men normaliteit aannemen. Aangezien men hierover slechts zelden gegevens bezit, weet men meestal eigenlijk niet, in hoeverre de berekende overschrijdingskansen juist zijn.

3. Toch is langs theoretische weg aangetoond, dat in de gevallen waarin de toepassing van Student verantwoord is, Wilcoxon een bijna even scherpe toets is.

4. Practisch bleek Wilcoxon bovendien nog te verkiezen boven Student wegens geringere gevoeligheid ten opzichte van uitschieters.

5. De alternatieve hypothesen, ten opzichte waarvan Wilcoxon bruikbaar is, zijn ruim en komen overeen met de verlangens der practijk.

5.2 Het optreden van gelijke waarnemingen; inleiding.

Tot nu toe hebben wij steeds ondersteld, dat er onder de waarnemingen geen gelijke optreden. In de practijk komt het echter toch wel herhaaldelijk voor, dat men gelijke waarnemingen doet. Dit kan twee oorzaken hebben, die we aan 2 biologische voorbeelden zullen toelichten.

I. Men wil weten, of er verschil is in de lengte van een bepaald bot tussen vrouwelijke en mannelijke orang oetan's.
II. Men wil weten, of er verschil is in het aantal vinstralen van de rugvin van 2 nauwverwante vissoorten.

In geval II kunnen de uitkomsten der waarnemingen uit niets anders bestaan dan een rij gehele getallen. Het is onmogelijk dat men 7½ vinstraal telt: er zijn er 7 of 8. Deze grootheid (nl. het aantal vinstralen) is een discrete veranderlijke. Zijn cumulatieve verdelingsfunctie is persé discontinu (een trapfunctie) en men zou zich te zeer van de werkelijkheid verwijderen door hiervoor een continue verdelingsfunctie als statistisch model te willen gebruiken.
Het is duidelijk dat de kans op gelijke waarnemingen nu niet gelijk 0 is (stel: vissoort A heeft 8, 9, 10, 11 of 12 vinstralen, en er wordt een steekproef genomen van 20 exemplaren) en d dat de voorgaande theorie dus niet van toepassing is. Door Hemelrijk is echter gevonden dat er een zeer nauw verband bestaat tussen Kendall's rangcorrelatie en de toets van Wilcoxon. Dit verband bestaat zowel bij discontinue als bij continue verdelingsfct. Waar nu Kendall zijn rangcorrelatie reeds heeft uitgebreid tot gevallen als onder II genoemd, kan daarvan gebruik gemaakt worden om ook aan Wilcoxon's toets een zodanige uitbreiding te geven. Dit zal behandeld worden in § 5.4.

In geval I ligt de zaak echter anders. De lengte van zo 'n bot is niet een discrete veranderlijke. Deze lengte is een grootheid, die in een zeker interval alle waarden kan aannemen en zich niet beperkt tot een discrete reeks. Zijn cumulatieve verdelingsfunctie is persé continu en men zou zich van de werkelijkheid te zeer verwijderen door een discontinue verdelingsfunctie als statistisch model te gebruiken. Toch kunnen hierbij gelijke waarnemingen optreden en wel doordat de waarnemingen in klassen worden ingedeeld. We zullen hierover enkele opmerkingen maken in § 5.3.

5.3 Geval I van gelijke waarnemingen.

Stel, dat de uiterste grenzen van de lengte van het bot in kwestie zijn 40 cm. en 55 cm., voorts, dat onze steekproef bestaat uit 20 waarnemingen en dat de waarnemingen gedaan werden in centimeters nauwkeurig (geen cijfers achter de komma bepaald). Het is duidelijk, dat er dan "gelijke" waarnemingen zullen optreden. Het is evenzeer duidelijk, dat als er één cijfer achter de komma was bepaald, er waarschijnlijk geen "gelijke" waarnemingen zouden zijn opgetreden.

Nu betekent het waarnemen in centimeters nauwkeurig niets anders dan dat de waarnemingen worden ingedeeld in klassen, in intervallen, waarvan er één gegeven wordt door: 44,50-45,49. En de "gelijkheid van twee waarnemingen betekant nu, dat ze beide in dezelfde klassen zijn gevallen, Dat wil zeggen: ze zijn wel schijnbaar gelijk en zoals de waarnemingen er nu liggen, hebben we geen middel om uit te maken, welke de grootste is, maar in wezen zijn ze ongelijk: ze liggen wel in dezelfde klasse, maar niet op dezelfde plaats in die klasse. Er zijn twee gevallen te onderscheiden:

A. We zijn zo gelukkig, dat geen x-waarde "gelijk " is aan een y-waarde. Er zijn slechts onderling"gelijke" x-waarden en o onderling "gelijke" y-waarden ("gelijk"=in dezelfde klasse).

B. Helaas zijn er ook paren (i,j), waarvoor x_i en y_j in dezelfde klasse zijn gevallen.

Ad A: In dit geval zijn er geen moeilijkheden. We berekenen U voor onze steekproeven net als hiervoor en ook de verdeling van U, evenals de μ en de σ blijven ongewijzigd. Dit geval staat in tegenstelling tot de door Hemelrijk in § 5.4 besproken theorie voor het geval II van § 5.2 (discontinue verdelingsfct), waarbij men rekening moet houden met alle gelijke waarnemingen, onafhankelijk van de steekproef, waaruit zij getrokken zijn.

Ad B: Het geval dat (schijnbaar) voor één of meer paren (i,j) geldt, dat $x_i=y_j$. We kunnen thans tengevolge van de waarneming in te weinig decimalen niet uitmaken of $x_i > y_j$, dan wel $x_i < y_j$. Slechts weten we, dat de kans, dat het gelijkteken geldt, gelijk aan 0 is. Maar aan deze wetenschap hebben we niets voor de bepaling van de bijdrage, die dergelijk paren (i,j) aan U leveren. Dit houdt een aansporing in tot zo nauwkeurig mogelijk waarnemen en een afwijzing van de gedachte, dat statistiek toch maar een vak is dat met onzekerheden werkt en waarbij men zich "dus" onnauwkeurig werken kan permitteren. Intussen kan men, als er niet te veel paren van gelijke x en y zijn, als volgt redeneren: bereken tweemaal de waarde U uit de steekproeven, één keer door van alle paren van gelijke x en y aan te nemen dat $y < x$ en één keer idem met $y > x$. De (onbekende en niet nader te bepalen) waarde, die U voor de beschouwde steekproeven eigenlijk heeft, ligt dan in het gesloten interval met deze beide berekende U-waarden als eindpunten. Als dit gehele interval in het kritieke gebied ligt of, anders gezegd, als deze beide U-waarden significant zijn, is ons experiment zeker significant. Als het gehele interval buiten het kritieke gebied ligt, is ons experiment zeker niet significant. Als het gedeeltelijk erin en gedeeltelijk erbuiten ligt, zit men op de grens.

Vanzelfsprekend wordt hierbij de bij het kritieke gebied behorende kans berekend met behulp van de in vorige §§ afgeleide waarschijnlijkheidsverdeling van U onder de getoetste hypothese H_o, zonder enige wijziging. D.w.z. de overschrijdingskansen van beide berekende U-waarden bepaalt men voor grote m en n met behulp van $\mu=\frac{1}{2}mn$ en $\sigma^2=\frac{1}{12}mn(m+n+1)$ en voor kleine m en n met de voor continue verdelingsfct. van x en y afgeleide tabel voor U. (Van der Vaart (8)).
Als een soort benaderende bepaling van de plaats van bovenbedoeld U-interval kan men, indien er weinig paren van gelijke x en y zijn, desnoods volstaan met de regel :

Elk paar (i,j) met $y_j < x_i$ geeft een bijdrage +1 tot U
" " (i,j) " $y_j = x_i$ " " " $+\frac{1}{2}$ " U
" " (i,j) " $y_j > x_i$ " " " 0 " U,

waarbij de significantie van U bepaald kan worden als tot nu toe. Men krijgt dan een soort gemiddelde significantie, maar het is toch veiliger om beide uiterste U-waarden te gebruiken.

5.4. Geval II van gelijke waarnemingen en het verband met de methode der rangcorrelatie,

door

Dr J.Hemelrijk Jr.

Indien onze waarnemingen van niet-continu verdeelde stochastische grootheden afkomstig zijn, kunnen wij, hoewel de theorie van de toets van Wilcoxon voor dit geval nog slechts gedeeltelijk ontwikkeld is, de toets toch toepassen, althans indien er niet te veel gelijke waarnemingen optreden. Daarbij maken wij dan gebruik van het verband met de rangcorrelatietheorie van Kendall.

Indien er gelijken optreden, definiëren wij U als volgt:

Elk paar (i,j) met $y_j < x_i$ geeft een bijdrage +1 tot U;
" " (i,j) " $y_j = x_i$ " " " $+\frac{1}{2}$ " U;
" " (i,j) " $y_j > x_i$ " " " 0 " U.

De op deze wijze gedefiniëerde U blijkt op de volgende wijze herleid te kunnen worden tot de grootheid S, die Kendall in de theorie der rangcorrelatie gebruikt (vgl. Hoofdstuk I van deze cursus; vooral ook § 5): als eerste rij(A) van rangnummers nemen wij de rangnummers der waarnemingen $x_1,\ldots,x_m$ en $y_1,\ldots,y_n$, bij rangschikking naar opklimmende grootte, waarbij wij de gelijken behandelen op de in hoofdstuk I, § 5 beschreven wijze. Om de tweede rij (B) van rangnummers te verkrijgen onderscheiden wij de waarnemingen slechts naar de steekproef, waaruit zij afkomstig zijn. Geven wij dit aan met x resp. y en

zetten wij de rij A in de natuurlijke volgorde, dan verkrijgen wij (met eventueel gelijke rangnummers in rij A) een schema van de volgende aard:

A: 1 2 3 - - - - - (m+n-1) (m+n)
B: x x y - - - - - y x.

Wij geven nu in rij B aan alle x- en gelijke rangnummers en eveneens aan alle y's, waarbij wij de x- en de kleinste rangnummers geven. Deze krijgen dus alle als rangnummer:

$$\frac{1}{m}(1+2+\ldots+m)=\frac{m+1}{2},$$

terwijl iedere y als rangnummer krijgt:

$$\frac{1}{n}\left\{(m+1)+(m+2)+\ldots+(m+n)\right\}=m+\frac{n+1}{2}$$

Gaan wij nu na, welke bijdrage een paar (h,k) van rangnummers uit rij B aan de grootheid S van Kendall geeft, dan zien wij in de eerste plaats, dat deze gelijk aan nul is, indien de elementen, die op de h^{e} en k^{e} plaats in rij B staan, gelijk zijn, dus indien beide uit dezelfde steekproef afkomstig zijn. Wij behoeven dus slechts die paren (h,k) te beschouwen, waarbij de ene een x en de andere een y is en deze krijgen we juist één maal, indien wij de paren (i,j) beschouwen, behorende bij x_i en y_j. Een dergelijk paar geeft tot S de bijdrage 0, indien $x_i=y_j$ is, daar dan de bijbehorende rangnummers in rij A gelijk zijn. Wij vinden zo, dat S op de volgende wijze ontstaat:

Elk paar (i,j) met $y_j < x_i$ geeft een bijdrage -1 tot S.
" " (i,j) " $y_j=x_i$ " " " 0 " S.
" " (i,j) " $y_j > x_i$ " " " +1 " S.

Vergelijken wij dit met de wijze waarop U ontstaat, dan zien wij, dat

$$2U+S=mn \qquad (5.3,1)$$

is, nl. gelijk aan het aantal paren (i,j); immers voor ieder dergelijk paar is de bijdrage tot 2U+S gelijk aan +1.

Aangezien echter de "nulhypothese" bij de theorie der rangcorrelatie precies overeenkomt met de "nulhypothese" bij Wilcoxon, kunnen wij alle door Kendall voor de verdeling van $\underline{S}$ bewezen eigenschappen zonder meer omrekenen in de overeenkomstige eigenschappen voor $\underline{U}$.

Kendall bewees, dat $\mathcal{E}\,\underline{S}=0$ is (zie hoofdst. I, p.6), hetgeen volgens (5.3,1) inhoudt, dat

$$\mathcal{E}\,\underline{U}=\tfrac{1}{2}\left\{mn-\mathcal{E}\,\underline{S}\right\}=\tfrac{1}{2}mn \qquad (5.3,2)$$

is. Deze formule ondergaat dus (vgl.(2.1,4)) door het optreden van gelijken geen wijziging. Verder volgt uit (5.3,1):

$$\sigma^2_{\underline{U}} = \frac{1}{4}\,\sigma^2_{\underline{S}}\,, \qquad (5.3,3).$$

zodat wij $\sigma^2_{\underline{u}}$ kunnen berekenen uit formule (26) op pag.13 van deze cursus, waarin nu u de twee waarden m en n aanneemt, daar rij B bestaat uit 2 groepen van m resp. n gelijke rangnummers. De waarden, die t aanneemt, volgen uit rij A; wij merken op, dat men daarbij rekening moet houden met alle gelijke waarnemingen, onafhankelijk van de steekproef, waaruit zij getrokken zijn. In (26) wordt nu dus:

$$\begin{aligned} \textstyle\sum u(u-1)(2u-5) &= m(m-1)(2m-5)+n(n-1)(2n-5),\\ \textstyle\sum u(u-1)(u-2) &= m(m-1)(m-2)+n(n-1)(n-2),\\ \text{en}\quad \textstyle\sum u(u-1) &= m(m-1)+n(n-1), \end{aligned}$$

terwijl in (26) n vervangen moet worden door m+n.

Voeren wij deze substitutie uit, dan verkrijgen wij na enige herleiding, rekening houdende met de factor 1/4 in (5.3,3):

$$\left.\begin{aligned} \sigma^2_{\underline{U}} = {} & \frac{1}{12}\,mn(m+n+1) - \frac{1}{72}\textstyle\sum t(t-1)(2t+5) + \\ & + \frac{m(m-1)(m-2)+n(n-1)(n-2)}{36(m+n)(m+n-1)(m+n-2)}\textstyle\sum t(t-1)(t-2) + \\ & + \frac{m(m-1)+n(n-1)}{8(m+n)(m+n-1)}\textstyle\sum t(t-1) \end{aligned}\right\} \qquad (5.3,4)$$

Indien er geen gelijke waarnemingen zijn blijft hiervan alleen de eerste term over, zodat de formule overgaat in (2.1,12)[1]. Daar voor $t \geqq 2$ en $m+n \geqq 2$ geldt:

$$\frac{1}{18}(2t+5) > \frac{m(m-1)(m-2)+n(n-1)(n-2)}{9(m+n)(m+n-1)(m+n-2)}\,(t-2) + \frac{m(m-1)+n(n-1)}{2(m+n)(m+n-1)}\,,$$

zoals men met volledige inductie gemakkelijk kan bewijzen, volgt uit (5.3,4), dat $\sigma^2_{\underline{u}}$ door het optreden van gelijke waarnemingen <u>kleiner</u> wordt.

In verband met deze uitkomsten kan men bij het optreden van gelijke waarnemingen de in § 2.2 beschreven benaderingsmethoden <u>a</u> en <u>b</u> volgen, daarin vervangende door de uit (5.3,4) volgende waarde. Indien er weinig gelijken zijn overweegt de eerste term in het rechterlid van (5.3,4), zodat men dan de benaderingsmethoden zonder wijziging kan gebruiken. Men vindt dan een grotere overschrijdingskans dan bij het gebruik van (5.3,4), zodat de kans op ten onrechte verwerpen van H_o, indien deze juist is, kleiner wordt, maar het onderscheidings-

---

[1] Dit betekent, dat in dit geval ook voor discrete verdelingen de toets in zijn oorspronkelijke vorm kan worden toegepast, indien men hem beschouwt als een voorwaardelijke toets, met als voorwaarde het niet optreden van gelijke waarnemingen (vgl. de opmerking aan het einde van § 1.3).

vermogen eveneens [1].

Indien m en n $\leq$ 10 zijn kan men, daar de spreiding van $\underline{U}$ door het optreden van gelijken verkleind wordt, en aannemende, dat door het optreden van gelijken de vorm van de verdeling van $\underline{U}$ niet zo sterk wordt beïnvlood, dat ondanks het kleiner worden van σ toch de overschrijdingskans groter zou worden, ook gebruik maken van de in Van der Vaart (8) opgenomen tabellen. Daarbij geldt dan hetzelfde voorbehoud als in de vorige alinea.

5.5 Eénzijdige en tweezijdige toetsing.

Zoals in 1.4 uiteengezet werd, bestaat het kritieke gebied uit 2 staartintervallen voor U:

Het is soms verantwoord om niet dergelijke tweezijdige overschrijdingskansen te berekenen, maar éénzijdige, of met andere woorden, een kritiek gebied te gebruiken, dat hetzij de vorm $0 \leq \underline{U} \leq s$, hetzij de vorm $mn-s \leq \underline{U} \leq mn$ heeft, in plaats van te bestaan uit de vereniging van deze beide intervallen. Dit hangt af van wat men a priori over de alternatieve hypothesen kan zeggen. Vergelijk hierover bv. Van der Vaart (8), p.11, § 4, C, 3, opm.3, of Hemelrijk en Van der Vaart (4).
Hier zullen we niet nader op deze kwestie ingaan. Indien men er niet zeker van is, dat het gebruik van éénzijdige toetsing verantwoord is, wordt het gebruik van tweezijdige aanbevolen. Indien men tabellen gebruikt, waarin de overschrijdingskansen éénzijdig zijn getabelleerd, dient men deze met 2 te vermenigvuldigen, om de tweezijdige te verkrijgen.

5.6 Tabellen.

De verdeling van $\underline{U}$ is exact getabelleerd voor $m \leq 8$, $n \leq 8$ door Mann en Whitney [7] , en voor $m \leq 10$, $n \leq 10$ door de Rekenafdeling van het Mathematisch Centrum (H.R.van der Vaart [8]).
Een hulptabel voor de grootheden

$$\tfrac{1}{2}mn \qquad \text{en} \sqrt{\tfrac{1}{12}\, mn(m+n+1)},$$

voor waarden van m en n beneden de 100, eveneens berekend door de Rekenafdeling van het Mathematisch Centrum, is als bijlage II aan de syllabus van deze cursus toegevoegd. In bovenstaande formules zijn m en n verwisselbaar. Daarom is in de bijlage niet aangegeven, dat één van beide argumenten de m en het andere n zou zijn, men kan dit naar willekeur bepalen.

LITERATUUR:

(1) Dantzig, D. van, Kadercursus Mathematische Statistiek 1947-1950, Hoofdstuk VI, § 3 (Mathematisch Centrum, Amsterdam).

(2) Dantzig, D. van, On the consistency and the power of Wilcoxon's two sample test : Proc.Kon.Ned.Ak.van Wetensch., Ser.A, 54 (1951), pp. 3 - 10; Indagationes Mathematicae 13 (1951), p. 1-8.

(3) Grubbs, F.E., Sample criteria for testing outlying observations, Annals of Mathem.Statistics 21 (1950), pp. 27- 58.

(4) Hemelrijk, J., en van der Vaart, H.R., Het gebruik van een- en tweezijdige overschrijdingskansen voor het toetsen van hypothesen; Statistica 4 (1950), 54-66.

(5) Hemelrijk, J., Kendall's rangcorrelatie-coëfficiënt τ , Hoofdstuk I van de voordrachten in de cursus "Parametervrije Methoden", Rapport S 59 van de Statistische Afdeling van het Mathematisch Centrum.

(6) Kemperman, J.H., De verdelingsfunctie van het aantal inversies in de test van Mann en Whitney, Rapport TW no. 7 van de Afdeling Toegepaste Wiskunde van het Mathematisch Centrum, Amsterdam.

(7) Mann, H.B. and Whitney, D.R., On a test of whether one of two random variables is stochastically larger than the other,Ann.of Math.Stat., 18 (1947), 50 - 60.

(8) Vaart, H.R. van der, Gebruiksaanwijzing voor de toets van Wilcoxon, Rapport S 32 (M 4) van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam.

(9) Vaart, H.R. van der, Some remarks on the power function of Wilcoxon's test for the problem of two samples, I, II, Proc.Kon.Ned.Ak. van Wetensch. 53 (1950), 494 - 506, 507 - 520, en Indagationes Mathematicae 12 (1950), 146 - 158, 159 - 172.

(10) Wilcoxon, F., Individual comparisons by ranking methods, Biometrics Bull. 1 (1945), 80 - 83.

De publicaties over de duiven zijn ontleend aan:
M.Gruber. Fat synthesis in thiamine Deficiency,Biochimica et Biophysica Acta, (in druk)

<u>BIJLAGE II</u> bij S 59

Tabel van $\mu = \frac{1}{2}\ m\ n$ en $\sigma = \sqrt{\frac{1}{12}\ m\ n(m+n+1)}$.

	11		13		15		17	
	μ	σ	μ	σ	μ	σ	μ	σ
1	5,5	3,4	6,5	4,0	7,5	4,6	8,5	5,2
2	11,0	5,1	13,0	5,9	15,0	6,7	17,0	7,5
3	16,5	6,4	19,5	7,4	22,5	8,4	25,5	9,4
4	22,0	7,7	26,0	8,8	30,0	10,0	34,0	11,2
5	27,5	8,8	32,5	10,1	37,5	11,5	42,5	12,8
6	33,0	10,0	39,0	11,4	45,0	12,8	51,0	14,3
7	38,5	11,0	45,5	12,6	52,5	14,2	59,5	15,7
8	44,0	12,1	52,0	13,8	60,0	15,5	68,0	17,2
9	49,5	13,2	58,5	15,0	67,5	16,8	76,5	18,6
10	55,0	14,2	65,0	16,1	75,0	18,0	85,0	19,9
11	60,5	15,2	71,5	17,3	82,5	19,3	93,5	21,3
13			84,5	19,5	97,5	21,7	110,5	23,9
15					112,5	24,1	127,5	26,5
17							144,5	29,0

	20		25		30		35	
	μ	σ	μ	σ	μ	σ	μ	σ
1	10,0	6,1	12,5	7,5	15,0	8,9	17,5	10,4
2	20,0	8,8	25,0	10,8	30,0	12,8	35,0	14,9
3	30,0	11,0	37,5	13,5	45,0	16,0	52,5	18,5
4	40,0	12,9	50,0	15,8	60,0	18,7	70,0	21,6
5	50,0	14,7	62,5	18,0	75,0	21,2	87,5	24,4
6	60,0	16,4	75,0	20,0	90,0	23,6	105,0	27,1
7	70,0	18,1	87,5	21,9	105,0	25,8	122,5	29,6
8	80,0	19,7	100,0	23,8	120,0	27,9	140,0	32,0
9	90,0	21,2	112,5	25,6	135,0	30,0	157,5	34,4
10	100,0	22,7	125,0	27,4	150,0	32,0	175,0	36,6
11	110,0	24,2	137,5	29,1	165,0	34,0	192,5	38,8
13	130,0	27,2	162,5	32,5	195,0	37,8	227,5	43,1
15	150,0	30,0	187,5	35,8	225,0	41,5	262,5	47,2
17	170,0	32,8	212,5	39,0	255,0	45,2	297,5	51,3
19	190,0	35,6	237,5	42,2	285,0	48,7	332,5	55,2
20	200,0	37,0	250,0	43,8	300,0	50,5	350,0	57,2
25			312,5	51,5	375,0	59,2	437,5	66,7
30					450,0	67,6	525,0	76,0
35							612,5	84,1

	40		45		50		55	
	μ	σ	μ	σ	μ	σ	μ	σ
1	20,0	11,8	22,5	13,3	25,0	14,7	27,5	16,2
2	40,0	16,9	45,0	19,0	50,0	21,0	55,0	23,1
3	60,0	21,0	67,5	23,5	75,0	26,0	82,5	28,5
4	80,0	24,5	90,0	27,4	100,0	30,3	110,0	33,2
5	100,0	27,7	112,5	30,9	125,0	34,2	137,5	37,4
6	120,0	30,7	135,0	34,2	150,0	37,7	165,0	41,3
7	140,0	33,5	157,5	37,3	175,0	41,1	192,5	45,0
8	160,0	36,2	180,0	40,2	200,0	44,4	220,0	48,4
9	180,0	38,7	202,5	43,1	225,0	47,4	247,5	51,8
10	200,0	41,2	225,0	45,8	250,0	50,4	275,0	55,0
11	220,0	43,7	247,5	48,5	275,0	53,3	302,5	58,1
13	260,0	48,4	292,5	53,6	325,0	58,9	357,5	64,1
15	300,0	52,9	337,5	58,6	375,0	64,2	412,5	69,9
17	340,0	57,3	382,5	63,4	425,0	69,4	467,5	75,4
19	380,0	61,6	427,5	68,0	475,0	74,4	522,5	80,8
20	400,0	63,8	450,0	70,4	500,0	76,9	550,0	83,5
25	500,0	74,2	562,5	81,6	625,0	89,0	687,5	96,3
30	600,0	84,3	675,0	92,5	750,0	100,6	825,0	108,7
35	700,0	94,2	787,5	103,1	875,0	112,0	962,5	120,8
40	800,0	103,9	900,0	113,6	1000,0	123,2	1100,0	132,7
45			1012,5	123,9	1125,0	134,2	1237,5	144,3
50					1250,0	145,1	1375,0	155,9
55							1512,5	167,3

	60		65		70		75	
	μ	σ	μ	σ	μ	σ	μ	σ
1	30,0	17,6	32,5	19,1	35,0	20,5	37,5	21,9
2	60,0	25,1	65,0	27,1	70,0	29,2	75,0	31,2
3	90,0	31,0	97,5	33,5	105,0	36,0	112,5	38,5
4	120,0	36,1	130,0	38,9	140,0	41,8	150,0	44,7
5	150,0	40,6	162,5	43,9	175,0	47,1	187,5	50,3
6	180,0	44,8	195,0	48,4	210,0	51,9	225,0	55,4
7	210,0	48,8	227,5	52,6	245,0	56,4	262,5	60,3
8	240,0	52,5	260,0	56,6	280,0	60,7	300,0	64,8
9	270,0	56,1	292,5	60,5	315,0	64,8	337,5	69,2
10	300,0	59,6	325,0	64,2	350,0	68,7	375,0	73,3
11	330,0	62,9	357,5	67,7	385,0	72,5	412,5	77,3
13	390,0	69,4	422,5	74,6	455,0	79,8	487,5	85,0
15	450,0	75,5	487,5	81,1	525,0	86,8	562,5	92,4
17	510,0	81,4	552,5	87,4	595,0	93,4	637,5	99,4
19	570,0	87,2	617,5	93,5	665,0	99,9	712,5	106,2
20	600,0	90,0	650,0	96,5	700,0	103,0	750,0	109,5
25	750,0	103,7	812,5	111,0	875,0	118,3	937,5	125,6
30	900,0	116,8	975,0	124,9	1050,0	133,0	1125,0	141,0
35	1050,0	129,6	1137,5	138,4	1225,0	147,1	1312,5	155,8
40	1200,0	142,1	1300,0	151,6	1400,0	160,9	1500,0	170,3
45	1350,0	154,4	1462,5	164,5	1575,0	174,5	1687,5	184,5
50	1500,0	166,6	1625,0	177,2	1750,0	187,9	1875,0	198,4
55	1650,0	178,6	1787,5	189,9	1925,0	201,1	2062,5	212,2
60	1800,0	190,5	1950,0	202,4	2100,0	214,1	2250,0	225,8
65			2112,5	214,8	2275,0	227,1	2437,5	239,3
70			2450,0	240,0	2625,0	252,7		
75					2812,5	266,0		

	80		85		90		100	
	μ	σ	μ	σ	μ	σ	μ	σ
1	40,0	23,4	42,5	24,8	45,0	26,3	50,0	29,2
2	80,0	33,3	85,0	35,3	90,0	37,4	100,0	41,4
3	120,0	41,0	127,5	43,5	135,0	46,0	150,0	51,0
4	160,0	47,6	170,0	50,5	180,0	53,4	200,0	59,2
5	200,0	53,5	212,5	56,8	225,0	60,0	250,0	66,5
6	240,0	59,0	255,0	62,5	270,0	66,1	300,0	73,1
7	280,0	64,1	297,5	67,9	315,0	71,7	350,0	79,4
8	320,0	68,9	340,0	73,0	360,0	77,1	400,0	85,2
9	360,0	73,5	382,5	77,8	405,0	82,2	450,0	90,8
10	400,0	77,9	425,0	82,5	450,0	87,0	500,0	96,2
11	440,0	82,1	467,5	86,9	495,0	91,7	550,0	101,3
13	520,0	90,3	552,5	95,5	585,0	100,7	650,0	111,1
15	600,0	98,0	637,5	103,6	675,0	109,2	750,0	120,4
17	680,0	105,4	722,5	111,4	765,0	117,4	850,0	129,3
19	760,0	112,6	807,5	118,9	855,0	125,2	950,0	137,8
20	800,0	116,0	850,0	122,5	900,0	129,0	1000,0	142,0
25	1000,0	132,9	1062,5	140,2	1125,0	147,5	1250,0	162,0
30	1200,0	149,0	1275,0	157,0	1350,0	165,0	1500,0	181,0
35	1400,0	164,5	1487,5	173,2	1575,0	181,9	1750,0	199,2
40	1600,0	179,6	1700,0	188,9	1800,0	198,2	2000,0	216,8
45	1800,0	194,4	1912,5	204,3	2025,0	214,2	2250,0	234,0
50	2000,0	209,0	2125,0	219,5	2250,0	230,0	2500,0	250,8
55	2200,0	223,3	2337,5	234,4	2475,0	245,4	2750,0	267,4
60	2400,0	237,5	2550,0	249,1	2700,0	260,7	3000,0	283,7
65	2600,0	251,5	2762,5	263,7	2925,0	275,8	3250,0	299,9
70	2800,0	265,5	2975,0	278,1	3150,0	290,7	3500,0	315,8
75	3000,0	279,3	3187,5	292,5	3375,0	305,6	3750,0	331,7
80	3200,0	293,0	3400,0	306,7	3600,0	320,3	4000,0	347,4
85			3612,5	320,9	3825,0	335,0	4250,0	363,0
90					4050,0	349,6	4500,0	378,4
100							5000,0	409,3

MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 59

Cursus Parametervrije Methoden

IV. Parametervrije methoden in de regressieanalyse

door

H. Theil

September 1951

I. INLEIDING.

1. De opzet.

Gegeven (of verondersteld) zij, dat er tussen twee variabelen, ξ en η, een rechtlijnig verband bestaat:

(1.1.1) $$\eta = \alpha + \beta \xi$$

ξ kan bijv. de temperatuur voorstellen en η de lengte van een metalen staaf bij de temperatuur ξ. Of ook: ξ is het inkomen en η het bedrag, dat door een gezin van bepaalde samenstelling aan voedsel wordt uitgegeven.

Laten we verder veronderstellen, dat n waarden van en de bijbehorende waarden van gemeten zijn. Noemen we deze waarden $\xi_1, \ldots, \xi_n, \eta_1, \ldots, \eta_n$ dan hebben we dus

(1.1.2) $$\eta_i = \alpha + \beta \xi_i \qquad (i = 1, \ldots, n)$$

Wanneer we de paren $(\xi_1, \eta_1), \ldots, (\xi_n, \eta_n)$ in een rechthoekig coördinatenstelsel afzetten, moeten we dus n punten vinden, die alle op een rechte lijn liggen. Deze rechte bepaalt de constanten α en β volkomen (voor opgesteld, dat $n \geq 2$ is)

Wij keren nu terug tot de voorbeelden, in het begin genoemd. Stel, dat men in werkelijkheid de staaf op n verschillende temperaturen brengt; stel ook, dat men inderdaad het bedrag, aan voedsel uitgegeven, van n gezinnen (met verschillend inkomen) uit huishoudrekeningen berekent. Dan zal men in het algemeen in beide gevallen vinden, dat de waargenomen punten slechts "ongeveer", niet exact, op een rechte liggen. Trekt men toch een rechte lijn door een dergelijke "puntenwolk", dan vindt men dus "afwijkingen" tussen deze punten en de punten van de rechte.
Deze afwijkingen kunnen voortkomen uit twee geheel verschillende oorzaken:

a. Het kan zijn, dat η niet slechts afhankelijk is van ξ , maar van nog meer variabelen. Dit is zonder meer duidelijk bij het voorbeeld van de uitgaven aan voedsel. Immers, afgezien van het inkomen, zullen ook de preferenties van de gezinnen hun voedseluitgaven beïnvloeden. Men kan dit in beginsel opvangen door η inderdaad als een functie van verscheidene variabelen te schrijven; maar vaak zijn deze variabelen niet beschikbaar, soms ook zijn zij principieël onmeetbaar. Bovendien zijn in vele gevallen deze variabelen zeer talrijk en in hun afzonderlijke invloed op η zo gering, dat het ondoenlijk zou zijn hen alle op te nemen. Houden we het dan dus op de enkele variabele ξ , dan moeten we voor (1.1.2) schrijven:

(1.1.3) $$\eta_i = \alpha + \beta \xi_i + w_i, \qquad (i = 1, \ldots\ldots\ldots, n)$$

waarbij de grootheden w_i de gezamenlijke invloed der "overige" variabelen op η_i voorstellen. Men noemt ze <u>Storingstermen,</u> (Engels: Disturbances).

b. Het kan zijn, dat de 2n waarden $\xi_1, \ldots\ldots, \xi_n, \eta_1, \ldots\ldots \eta_n$ niet volkomen nauwkeurig gemeten zijn. In plaats van deze waarden meet mer dan:

(1.1.4) $$x_i = \xi_i + u_i \; ; \qquad (i = 1, \ldots\ldots\ldots, n)$$

(1.1.5) $$y_i = \eta_i + v_i \; ; \qquad (i = 1, \ldots\ldots\ldots, n)$$

waarbij u_i en v_i <u>meetfouten</u> genoemd worden.

Gedurende de laatste jaren treedt, in het bijzonder in de economie, de gedachte naar voren, dat de storingstermen essentiëler zijn dan de meetfouten in deze zin, dat, ook al zouden de metingen wel volkomen nauwkeurig zijn, de waargenomen punten niet exact aan de gepostuleerde rechte (1.1.1) voldoen, dank zij de storingstermen. Wij zullen echter beide types van afwijkingen beschouwen.

1.2. <u>Parameters en stochastische grootheden.</u>

Het doel van de regressieanalyse is in de eerste plaats om tot uitspraken te komen omtrent de ligging van de rechte (1.1.1) Aangezien deze ligging volkomen bepaald wordt door α en β moet dus getracht worden deze grootheden te bepalen; men noemt α en β onbekende parameters.

Wil men hierin slagen, dan is het noodzakelijk, dat zekere veronderstellingen worden gemaakt omtrent u_i, v_i en w_i. Men pleegt aan te nemen, dat zij trekkingen zijn uit bepaalde verdelingen. Dan volgt uit (1.1.3), (1.1.4) en (1.1.5), dat niet slechts zij, maar ook de grootheden η_i, x_i en y_i stochastische grootheden zijn. Dit geven wij aan door <u>onderstreping</u> der symbolen. Waarneming van stochastische grootheden worden door dezelfde letters, niet onderstreept, aangegeven.

Van de n waarden ξ_i zullen wij niet aannemen, dat zij een verdeling hebben. Zij zijn dus parameters, en wel onbekende parameters, tenzij alle u_i gelijk nul zijn.

Tenslotte zij opgemerkt, dat er nog een derde type van parameters optreedt, nl. die welke de verdelingsfuncties der u_i, v_i en w_i beschrijven.

1.3. De methode der kleinste kwadraten.

Van de ingevoerde grootheden ξ_i, η_i, α, β, u_i, v_i, w_i, $x_i, y_i (i=1,\ldots,n)$ zijn slechts x_i en y_i observeerbaar. Blijkens (1.1.3) (1.1.4) en (1.1.5) voldoen zij aan

(1.3.1) $$y_i = \alpha + \beta x_i - \beta u_i + v_i + w_i.$$

Wij veronderstellen nu:

a) $u_i \equiv 0 \quad (i=1,\ldots,n)$

b) $(v_i + w_i)$ zijn normaal verdeeld met verwachting 0 en variantie σ^2 onafhankelijk van i $(i=1,\ldots,n)$

c) $(v_i + w_i)$ en $(v_j + w_j)$ zijn ongecorreleerd:

$$\mathcal{E}(v_i + w_i)(v_j + w_j) = 0. \qquad (i,j=1,\ldots,n;\ i \neq j)$$

Is aan deze veronderstellingen voldaan, dan kan men α en β schatten volgens de methode van de kleinste kwadraten, d.w.z. door

$$\sum_{i=1}^{n} (y_i - a - b x_i)^2,$$

waarin (x_i, y_i) de waargenomen punten zijn, minimaal te maken en a en b op te lossen; men noemt a en b de kleinste-kwadraten-schattingen van α resp β. Als uitkomst vindt men:

(1.3.2) $$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}\, x};$$

(1.3.3) $$a = \bar{y} - b\bar{x},$$

waarbij

$$n\bar{x} = \sum_i x_i;$$

$$n\bar{y} = \sum_i y_i.$$

Daar a en b funties zijn van de waarnemeingen, zijn zij evenals deze waarnemingen stochastische grootheden. Men kan nu aantonen, dat onder bovengenoemde veronderstellingen geldt:

1) a en b zijn zuivere schattingen van α resp β. $(\mathcal{E}a = \alpha;\ \mathcal{E}b = \beta)$

2) Van alle schattingen van α resp β, die zuiver zijn en die lineair in de y_i zijn, zijn a en b het doeltreffendst (d.w.z. hun spreiding is het kleinst).

3) a en b zijn normaal verdeeld met verwachting α resp β en variantie $\frac{1}{n} \cdot \frac{\sigma^2}{\mathrm{var}\, x} \cdot \frac{1}{n} \sum_i x_i^2$ resp. $\frac{1}{n} \frac{\sigma^2}{\mathrm{var}\, x}$ [1]

Ook voor σ^2, indien deze grootheid onbekend is, kan men een

1) In dit geval zijn de grootheden x_i bekende parameters, dus niet stochastisch.

schatting geven. Met behulp van Student-verdelingen kan men dan betrouwbaarheidsintervallen voor α en β vinden. Voor een uitgebreider overzicht zie men b.v. J. Hemelrijk (1949b).

1.4. Parametervrije methoden.

Aan de methode van de kleinste kwadraten kleven de volgende bezwaren:

A. Men weet niet steeds zeker, of de stochastische grootheden $(\eta_i + w_i)$ alle normaal verdeeld zijn. Weliswaar behouden de varianties van de kleinste-kwadraten-schattingen a en b dezelfde waarden, ook al zijn $(\eta_i + w_i)$ niet normaal verdeeld; maar de eigenschappen, dat hun verwachting nul is, hun variantie σ^2 en dat zij ongecorrelleerd zijn, zijn onvoldoende om met behulp van deze schattingen tot betrouwbaarheidsintervallen te geraken.

B. Vaak is het zo, dat de variantie van $(\eta_i + w_i)$ niet onafhankelijk is van x_i. Zo zal men in het algemeen bij de bestudering van huishoudrekeningen vinden (bijv. wanneer men zekere uitgaven met het inkomen correleert), dat de spreiding der storingstermen toeneemt naarmate de afhankelijke variabele η_i zelf "gemiddeld" groter wordt.

C. In vele gevallen is ook de onafhankelijke variabele, ξ, niet geheel nauwkeurig gemeten. De schattingen a en b zijn dan niet langer zuiver. Weliswaar kan men hierin voorzien, als de verhouding van de spreiding van η_i tot die van $(\eta_i + w_i)$ bekend is, maar in het algemeen kent men deze verhouding niet.

D. Ook is de veronderstelling van de ongecorreleerdheid van $(\eta_i + w_i)$ en $(\eta_j + w_j)$ vaak dubieus, i.h.b. bij de analyse van tijdreeksen. Bij deze reeksen plegen de indices i en j verschillende tijdstippen of tijdsintervallen aan te geven. Men zal aanvoelen, dat de successieve storingstermen w_i en w_{i+1} vaak positief gecorreleerd zijn, omdat zij immers de invloed der "overige" variabelen weergeven, die voor successieve tijdstippen of -intervallen veelal niet onafhankelijk zijn.

Aan sommige dezer bezwaren komen de parametervrije methoden in de regressieanalyse tegemoet; "parametervrij " heeft hier betrekking op de parameter van het derde type, nl. de parameter, die de verdeling van $\eta_i, \xi_i, w_i \; (i=1,\ldots,n)$ beschrijven. Hierbij zij vooropgesteld, dat geen dezer methoden aan het bezwaar D tegemoet komt, terwijl slechts een enkele voorziet in het bezwaar B.

De parametervrije methoden, die de laatste jaren zijn gepubliceerd, vindt men op de volgende plaatsen: A. Wald (1940), K.R. Nair and M.P.Shrivastava (1942), K.R. Nair and K.S. Banerjee (1942), G.W. Housner and J.F. Brennan (1948), M.S. Bartlett (1949), J. Hemelrijk (1949a), A.M.Mood(1950), H. Theil (1950), E.L. Scott (1950) en E.F. Drion (1951).

Wij zullen vier methoden behandelen, t.w. twee voor puntschattingen en twee voor betrouwbaarheidsintervallen.

2. DE METHODE VAN A. WALD.

2.1. Veronderstellingen.

A. Wald (1940) voert de volgende veronderstellingen in:

a) De stochastische grootheden $u_1, \ldots, u_n$ hebben alle dezelfde verdeling en zijn ongecorreleerd. Hun verwachting is nul, hun variantie is eindig.

B) De stochastische grootheden $(v_1+w_1), \ldots (v_n+w_n)$ hebben alle dezelfde verdeling en zijn ongecorreleerd. Hun verwachting is nul, hun variantie is eindig.

c) De stochastische grootheden u_i en (v_j+w_j) zijn ongecorreleerd $(i, j = 1, \ldots, n)$

Aan deze voorwaarden zal in 2.2 nog een vierde worden toegevoegd.

Deze voorwaarden komen dus tegemoet aan de bezwaren A en C, niet echter aan B en D.

2.2. Schatting van β ; haar bruikbaarheid.

Wald nummert de waargenomen punten (x_i, y_i) naar opklimmende waarden van x. Dus geldt:

$$x_1 \leq x_2 \leq \ldots \leq x_n$$

Hij neemt aan, dat n even is (is n oneven, dan kan men het punt met de middelste x-waarde buiten beschouwing laten)

Als schatting van β introduceert bij

$$b_W = \frac{\sum_{i=1}^{m} y_i - \sum_{i=m+1}^{n} y_i}{\sum_{i=1}^{m} x_i - \sum_{i=m+1}^{n} x_i} \qquad (2.2.1)$$

waarbij $m=\frac{1}{2}n$. Deze schatting kan men meetkundig interpreteren als de richtingscoëfficiënt van de rechte, die het zwaartepunt van de linkerhelft der waargenomen punten (d.w.z. van de m punten met de laagste x-waarden) verbindt met het zwaartepunt van de rechterhelft der waargenomen punten.

Van deze schatting bewijst Wald de bruikbaarheid. Daartoe voert hij in :

$$c_1 = \frac{(x_1 + \ldots + x_m) - (x_{m+1} + \ldots + x_n)}{n} \qquad (2.2.2)$$

$$c_2 = \frac{(y_1 + \ldots + y_m) - (y_{m+1} + \ldots + y_n)}{n} \qquad (2.2.3)$$

$$\gamma_1 = \frac{(\xi_1 + \ldots + \xi_m) - (\xi_{m+1} + \ldots + \xi_n)}{n} \qquad (2.2.4)$$

$$\gamma_2 = \frac{\{(\eta_1 - w_1) + \ldots + (\eta_m - w_m)\} - \{(\eta_{m+1} + w_{m+1}) + \ldots + (\eta_n + w_n)\}}{n} \qquad (2.2.5)$$

Uit (1.1.3), (2.2.4) en (2.2.5) volgt:

(2.2.6) $$\frac{\gamma_2}{\gamma_1} = \beta$$

Substituëert men in deze formules voor de waarnemingen weer de stochastische grootheden, waarvan zij afkomstig zijn, dan worden de meeste van deze grootheden weer stochastisch.
Nu is de variantie van ($\underline{c}_1 - \gamma_1$) gelijk aan $\frac{1}{n}$ var $\underline{u}$ en de variantie van ($\underline{c}_2 - \gamma_2$) aan $\frac{1}{n}$ var ($\underline{v} + \underline{w}$). Voor $n \to \infty$ naderen $\underline{c}_1$ en $\underline{c}_2$ in waarschijnlijkheid tot γ_1 resp γ_2.

Aangezien $\underline{b}_w$ gelijk is aan $\underline{c}_2 / \underline{c}_1$, nadert $\underline{b}_w$ dus tot $\gamma_2/\gamma_1 = \beta$ vooropgesteld dat $\underset{p}{\lim} \underline{c}_1 \neq 0$. Men zal inzien, dat aan dit laatste steeds is voldaan.[2)]

In het bovenstaande is impliciet gebruik gemaakt van de veronderstelling, dat de rangschikking van de geobserveerde punten (x_i , y_i) naar opklimmende waarde van x onafhankelijk van de meetfouten $\underline{u}_i$ zijn. In feite is slechts nodig, dat de verdeling der punten in de twee groepen onafhankelijk is van deze meetfouten. Opdat de juistheid van bovenstaande beschouwingen gehandhaafd blijft moet dus aan de volgende voorwaarde voldaan zijn:

d) <u>De waarschijnlijkheid, dat de indeling in 2 groepen naar opklimmende x-waarden afwijkt van die naar opklimmende ξ-waarden, is nul.</u>

Is hieraan voldaan, dan kunnen we staande houden, dat deze indeling slechts berust op de ξ's en dus onafhankelijk is van de u's. Voldoende hiervoor is, dat de range $R_{\underline{u}}$ der $\underline{u}$'s eindig is en kleiner dan ($\xi_{m+1} - \xi_m$), waarbij de index-nummering plaats heeft overeenkomstig opklimmende ξ waarden:

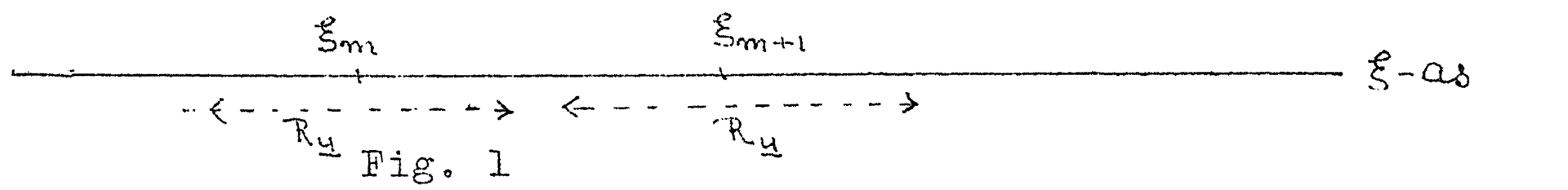


Fig. 1

Hiertoe is weer voldoende, dat ($x_{m+1} - x_m$) minstens het dubbele van deze range is. Immers, indien hieraan is voldaan, dan is het

---

2) Zie A. Wald (1950), pp. 287 en 297. De hier gegeven behandeling wijkt iets af van die van Wald. Wald geeft ook bruikbare schattingen van α , var $\underline{u}$ en var ($\underline{v}$ + $\underline{w}$). Verder geeft hij betrouwbaarheidsintervallen onder normaliteitsveronderstellingen; zie hiertoe i.h.b. loc.cit., pp. 294 e.v.

uitgesloten, dat een andere trekking van meetfouten twee x-waarden nl. x'_m en x'_{m+1}, zouden opleveren, waarvan de volgorde verwisseld is. Figuur 2 geeft een voorbeeld, waarin deze verwisseling wel optreedt.

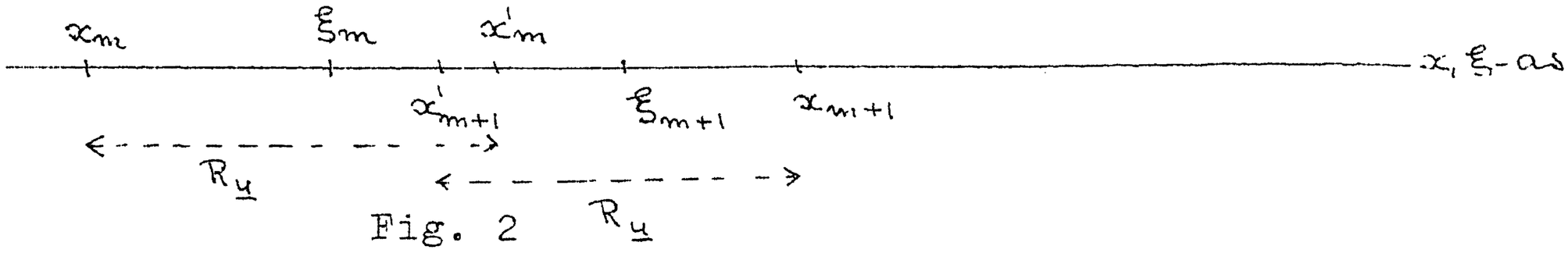


Fig. 2

De schatting $\underline{b}_w$ kan niet bogen op zuiverheid, tenzij $\underline{u}_i \equiv o$ voor alle i ; echter wel op zgn. "mediaan-zuiverheid", d.w.z. de mediaan van haar verdeling is gelijk aan:

(2.2.7) $$\text{Med } \underline{b}_w = \beta \ .$$

Men zie hiertoe J. Hemelrijk (1949 b), pp 17 e.v.

3. DE METHODE VAN HOUSNER EN BRENNAN.

3.1. Veronderstellingen.

Housner en Brennan (1948) maken gebruik van de volgende voorwaarden:

a) voorwaarde a van Wald;

b) voorwaarde b van Wald;

c) voorwaarde c van Wald;

d) Is $\xi_i < \xi_j$, dan is $P\left[\underline{x}_i < \underline{x}_j\right] = 1$.

Voorwaarde d gaat iets verder dan Wald's voorwaarde d. Zij kan op dezelfde wijze worden geconcretiseerd als voor laatstgenoemde voorwaarde aan het einde van 2.2 is gedaan; de index m moet dan alle waarden 1,.....,n-1 doorlopen. Het blijkt, dat aan de bezwaren B en D niet tegemoet wordt gekomen, wel aan A en C, maar aan C in mindere mate dan bij Wald.

3.2. Schatting van β .

De auteurs voeren de volgende schatting in:

$$(3.2.1)\qquad b_{HB} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1}(y_i - y_j)}{\sum_{i=1}^{n}\sum_{j=1}^{i-1}(x_i - x_j)}$$

de indexx nummering heeft hier plaats naar opklimmende x-waarden .

Na enig rekenen vindt men, dat b_{HB} voldoet aan

$$(3.2.2)\qquad b_{HB} = \frac{\sum_{i=1}^{n} i(y_i - \bar{y})}{\sum_{i=1}^{n} i(x_i - \bar{x})} = \frac{\sum_i (i-\bar{i})(y_i - \bar{y})}{\sum_i (i-\bar{i})(x_i - \bar{x})} = \frac{cov(i,y)}{cov(i,x)},$$

waarbij

$$(3.2.3)\qquad \bar{i} = \frac{1}{n}\sum i = \frac{1}{2}(n+1)$$

Wij zullen nu aantonen, dat $\underline{b}_{HB}$ een bruikbare schatting van β is.

Blijkens (1.1.3) en (3.2.2) geldt:

$$(3.2.4)\qquad \underline{b}_{HB} = \beta + \frac{\sum_{i=1}^{n} i(\underline{z}_i - \underline{z})}{\sum\sum_{i>j}(\underline{x}_i - \underline{x}_j)}$$

waarbij

$$(3.2.5)\qquad \underline{z}_i = -\beta\,\underline{u}_i + \underline{v}_i + \underline{w}_i$$

De noemer van de breuk in (3.2.4) is gelijk aan de som van $\frac{1}{2}n(n-1)$ termen ($\underline{x}_i - \underline{x}_j$), die alle positief zijn, dank zij de rangschikking. Minimaal is elk zo'n term gelijk aan

($\xi_i - \xi_j - R_u$).

(vgl. het slot van 2.2). Aangezien R_u en de ξ's vaste getallen zijn, niet slechts voor de getrokken steekproef, waar voor alle eventuele steekproeven, is de noemer dus van de orde van n^2 (symbolisch: $\mathcal{O}(n^2)$)
Voor de teller vindt men na enig rekenen:

$$\sum_i i(\underline{x}_i - \bar{\underline{x}}) = \tfrac{1}{2}\{-(n-1)\underline{x}_1 - (n-3)\underline{x}_2 - \dots + (n-3)\underline{x}_{n-1} + (n-1)\underline{x}_n\},$$

zodat we hebben

$$\operatorname{var}\{\sum_i i(\underline{x}_i - \bar{\underline{x}})\} = \tfrac{1}{2}\{1^2 + 3^2 + \dots + (n-1)^2\}\operatorname{var}\underline{x},$$

waarbij var $\underline{z}$ = β^2. var $\underline{u}$ + var $(\underline{v} + \underline{w})$, indien n even is ; is n oneven, dan geldt:

$$\operatorname{var}\{\sum_i i(\underline{x}_i - \bar{\underline{x}})\} = \tfrac{1}{2}\{2^2 + 4^2 + \dots + (n-1)^2\}\operatorname{var}\underline{x}.$$

In beide gevallen geldt

$$\operatorname{var}\{\sum_i i(\underline{x}_i - \bar{\underline{x}})\} < \tfrac{1}{2}\{1^2 + 2^2 + 3^2 + \dots + n^2\}\operatorname{var}\underline{x} =$$
$$= \tfrac{1}{12}n(n+1)(2n+1)\operatorname{var}\underline{x} = \mathcal{O}(n^3).$$

De breuk van (3.2.4) heeft dus een teller, waarvan de spreiding van de orde $n^{3/2}$ is, en een noemer, die van de orde n^2 is. Deze breuk convergeert dus in waarschijnlijkheid naar nul, zodat ook $\underline{b}_{HB} - \beta$ naar nul convergeert, aldus is $\underline{b}_{HB}$ een bruikbare schatting van β.
Evenals $\underline{b}_W$ is $\underline{b}_{HB}$ niet zuiver (tenzij alle $\underline{u}_i \equiv 0$), wel bruikbaar en mediaan zuiver (zie J. Hemelrijk (1949 b), pp. 17 e.v.)

4. METHODEN DER HELLINGEN: ONVOLLEDIGE METHODE.

4.1. Veronderstellingen.

Deze methode maakt gebruik van de volgende veronderstellingen: [3]

a) De n tripels $(\underline{u}_i, \underline{v}_i, \underline{w}_i)$ zijn onafhankelijk verdeeld.

b) Een van de twee volgende voorwaarden is vervuld:

b1) de stochastische grootheden $\underline{x}_i = -\beta\underline{u}_i + \underline{v}_i + \underline{w}_i$ hebben dezelfde continue verdelingsfunctie;

b2) de stochastische grootheden $\underline{x}_i$ hebben continue verdelingsfuncties, die alle symmetrisch zijn en eenzelfde mediaan ξ bezitten.

c) Als $\xi_i < \xi_j$, dan is $P[\underline{x}_i < \underline{x}_j] = 1$

Blijkens voorwaarde a wordt niet tegemoet gekomen aan het bezwaar D. Wel echter wordt het bezwaar B ondervangen, vooropgesteld althans, dat de verdeling van $\underline{x}_i$ symmetrisch is (en de verdelingsfunctie continu); dit volgt uit voorwaarde b2, want deze eist niet, dat alle $\underline{x}_i$ dezelfde verdeling hebben. Aan het bezwaar A wordt geheel, aan het bezwaar C in dezelfde mate als bij Hodges en Brennan tegemoet gekomen.

---

3) Zie H. Theil (1950).

4.2 Betrouwbaarheidsinterval voor β.

Deze methode leidt niet slechts tot een puntschatting van β , maar ook tot een betrouwbaarheidsinterval. In deze paragraaf bepalen zij ons tot het interval.

We rangschikken de punten weer naar opklimmende x-waarden en laten eventueel, nl. bij oneven n, het punt met de middelste x-waarde buiten beschouwing. Dan beschouwen we de richtingscoëfficiënt van de rechte, die de punten met rangnummers i en m+i ($m = \frac{1}{2}n$) verbindt:

$$(4.2.1)\quad \underline{D}(i, m+i) = \frac{\underline{y}_i - \underline{y}_{m+i}}{\underline{x}_i - \underline{x}_{m+i}} = \beta + \frac{\underline{z}_{m+i} - \underline{z}_i}{\underline{x}_{m+i} - \underline{x}_i}$$

(Vgl. (3.2.5) voor de betekenis van $\underline{z}_i$.)

Wij zullen nu aantonen, dat de mediaan van $\underline{D}(i, m+i)$ gelijk is aan β . Daartoe is noodzakelijk en voldoende:

$$(4.2.2)\quad P\left[\frac{\underline{z}_{m+i} - \underline{z}_i}{\underline{x}_{m+i} - \underline{x}_i} > 0\right] = P\left[\frac{z_{m+i} - z_i}{x_{m+i} - x_i} < 0\right] = \tfrac{1}{2}$$

Uit de wijze van rangschikking en uit voorwaarde c volgt, dat $(\underline{x}_{m+i} - \underline{x}_i)$ steeds positief is. Opdat β de mediaan van $\underline{D}(i, m+i)$ is, is dus noodzakelijk en voldoende:

$$(4.2.3)\quad P[\underline{z}_{m+i} - \underline{z}_i > 0] = P[\underline{z}_{m+i} - \underline{z}_i < 0] = \tfrac{1}{2}.$$

Nu zijn blijkens voorwaarde a de stochastische grootheden $\underline{z}_i$ en $\underline{z}_{m+i}$ onafhankelijk verdeeld. Dan zijn elk der voorwaarden b1 en b2 voldoende voor de juistheid van (4.2.3), hetgeen men als volgt kan inzien. In het geval b1 is de simultane verdelingsdichtheid van $\underline{z}_i$ en $\underline{z}_{m+i}$ symmetrisch rondom de rechte $\underline{z}_i = \underline{z}_{m+i}$. Geeft een trekking van een waardenpaar $(\underline{z}_i, \underline{z}_{m+i})$ tot uitkomst, dat het punt $(\underline{z}_i, \underline{z}_{m+i})$ boven deze rechte valt (waarbij $\underline{z}_i$ langs de horizontale as wordt gemeten) dan is $\underline{z}_{m+i} > \underline{z}_i$; valt het punt beneden deze rechte, dan is $\underline{z}_{m+i} < \underline{z}_i$; de kans op elk van deze beide gebeurtenissen is blijkbaar $\frac{1}{2}$. Men zie Fig. 3 (P en P' liggen gespiegeld t.o.v. de rechte):

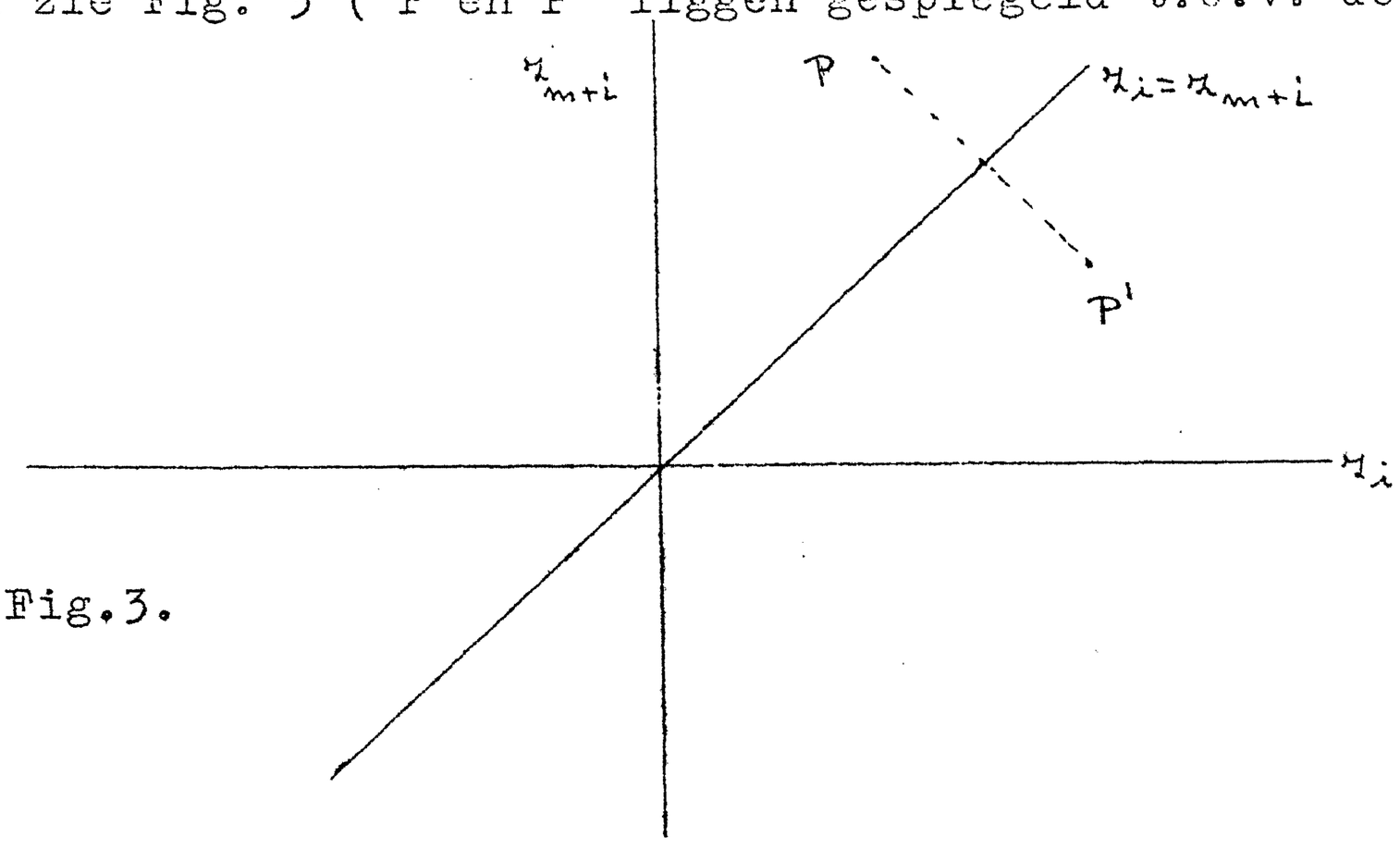


Fig.3.

In het geval b2 is de simultane verdelingsdichtheid van $\underline{z}_i$ en $\underline{z}_{m+i}$ symmetrisch rondom de rechten $z_i = \zeta$ en $z_{m+i} = \zeta$. Dus is de simultane dichtheid van $(z_i - \zeta)$ en $(z_{m+i} - \zeta)$ symmetrisch rond de oorsprong Dan is opnieuw de kans om een punt boven de rechte $z_i = z_{m+i}$ aan te treffen gelijk aan de kans om een punt er beneden te treffen. Men zie Fig. 4 (Q en Q liggen gespiegeld t.o.v. de oorsprong):

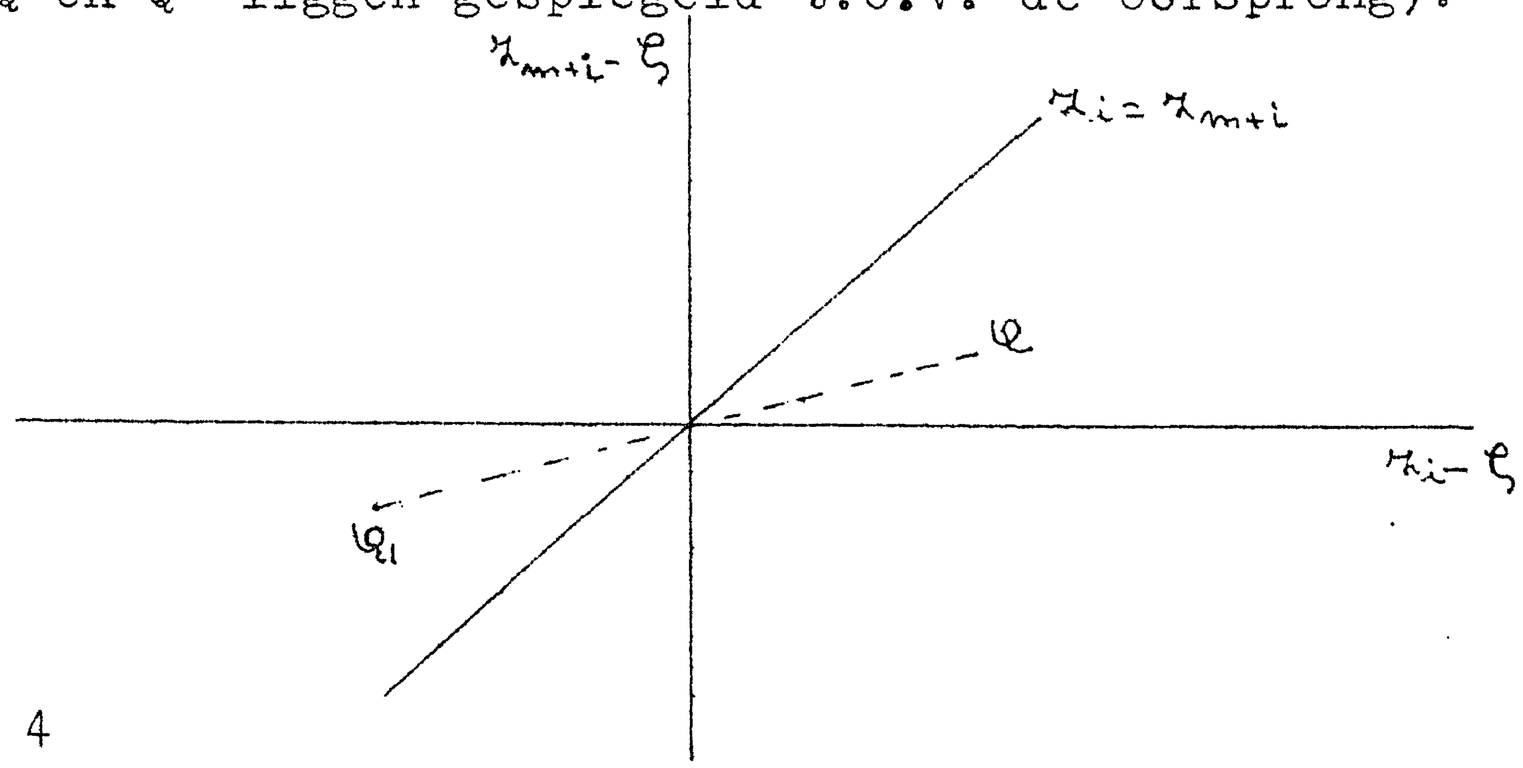


Fig. 4

Wij hebben dus bewezen, dat voor $i = 1, \ldots, m$ geldt:

(4.2.4) $P[\underline{D}(i, m+i) < \beta] = P[\underline{D}(i, m+i) > \beta] = \frac{1}{2}$.

Aldus hebben we m grootheden, die elk tot mediaan hebben. De kans dat deze alle kleiner zijn dan β , is gelijk aan $\left(\frac{1}{2}\right)^m$; de kans, dat juist $r < m$ van deze grootheden $< \beta$ zijn en de overige $(m-r)$ dus $> \beta$ is $2^{-m}\binom{m}{r}$

Met behulp van de binomiale verdeling (met parameter $\frac{1}{2}$) kan dan een betrouwbaarheidsinterval worden berekend. Daartoe nummeren we de m grootheden $\underline{D}$ (i, m+i) naar opklimmende grootte:

(4.2.4) $\underline{D}_1 < \underline{D}_2 < \underline{D}_3 < \cdots < \underline{D}_m$

Het betrouwbaarheidsinterval wordt dan gevormd door de eerste (r-1) en de laatste (r-1) van deze D's te laten vallen. Wij krijgen dan het interval $(\underline{D}_r, \underline{D}_{m+r-1})$, waarvoor geldt

$$P[\underline{D}_r \leqq \beta \leqq \underline{D}_{m-r+1} | \beta] = 1 - 2^{1-m} \sum_{s=0}^{r-1} \binom{m}{s}$$

Het rechterlid is voor m=1, - - - - - - - ,200 getabelleerd door A. van Wijngaarden (1950). Een klein gedeelte van deze tabel vindt men in bijlage III aan het eind van dit hoofdstuk.

4.3. Voorbeelden.

Teneinde te laten zien, dat deze methode (ondanks vrij sterk gespreide puntenwolken) toch tot niet zeer wijde betrouwbaarheidsintervallen kan leiden, geven wij enkele voorbeelden aan de hand van huishoudrekeningen. Tussen het inkomen (ξ) en de totale uitgaven (η) van diverse gezinnen is een lineair verband gepostuleerd.

In de hieronderstaande tabel vindt men betrouwbaarheidsgrenzen voor β bij verschillende overschrijdingskansen, en wel voor hoofdarbeiders, handarbeiders en landarbeiders (van het landelijk budget onderzoek 1935-1936) [4]

		Onbetrouwbaarheid	Betrouwbaarheidsinterval
Hoofdarbeiders	r=39	0,010	0,84 - 1,01
m = 103	40	0,018	0,85 - 1,00
	41	0,030	0,86 - 1,00
	42	0,048	0,87 - 1,00
	43	0,076	0,87 - 1,00
Handarbeiders	r=55	0,011	0,87 - 0,965
m = 139	56	0,017	0,87 - 0,96
	57	0,027	0,87 - 0,96
	58	0,041	0,88 - 0,96
	59	0,062	0,885 - 0,95
	60	0,089	0,89 - 0,94
Landarbeiders	r=5	0,012	0,64 - 1,045
m= 20	6	0,041	0,65 - 1,04

4) Men dient deze toepassing slechts als een numeriek voorbeeld te beschouwen; er is nl. alle aanleiding om behalve het inkomen ook de gezinsgrootte als variabele op te nemen, het geen hier niet is gedaan.
Schrijver dezer is dank verschuldigd aan de Heer H.S. Houthakker, die met hem de berekeningen heeft uitgevoerd.

5. METHODEN DER HELLINGEN: VOLLEDIGE METHODE.

5.1. Veronderstellingen.

De vorige methode is "onvolledig" genoemd, omdat zij niet volledig gebruik maakt van de $\binom{n}{2}$ hellingen.

(5.1.1) $$\underline{D}(i,j) = \frac{y_i - y_j}{x_i - x_j} \qquad (i = 1, \ldots, j-1;\ j = 1, \ldots, n)$$

Dit wordt wel gedaan door de "volledige" methode. Zij is gebaseerd op de volgende veronderstellingen:

a) Voorwaarde a van 4.1.

b) Voorwaarde b1 van 4.1.

c) Voorwaarde c van 4.1.

Deze methode komt dus, in tegenstelling tot de onvolledige methode, niet tegemoet aan bezwaar B. Wat deze bezwaren betreft, is zij ongeveer gelijkwaardig met de methode van Housner en Brennan.

5.2. Betrouwbaarheidsinterval.

Ook deze methode leidt tot een interval-schatting. Nummering heeft weer plaats naar opklimmende x-waarden.

Wij schrijven (5.1.1) in de vorm (vgl.(3.2.5))

(5.2.1) $$\underline{D}(i,j) = \beta + \frac{z_i - z_j}{x_i - x_j} \qquad (i < j)$$

Nu is $\underline{D}(i,j) > \beta$ dan en slechts dan, wanneer $z_i < z_j$; immers, uit de wijze van rangschikking van de punten (x_i, y_i) volgt, dat $i < j$ impliceert: $x_i < x_j$. Laten we nu als nulhypothese invoeren, dat β gelijk is aan een zeker getal, bijv. β' . Dan kan men het aantal $\underline{D}(i,j)$'s, dat kleiner is dan β' (hetgeen onder de nulhypothese identiek is met het aantal gevallen, waarvoor $z_i < z_j$) eenvoudig berekenen. Dit aantal hangt zowel af van β' als van de steekproef. Dit geval is dus een stochastische grootheid; wij zullen het weergeven met $q(\beta')$ Wij zullen nu de hypothese, dat $\beta = \beta'$ is, verwerpen wanneer er een significante rangcorrelatie bestaat tussen de rangschikking van de geobserveerde punten (x_i, y_i) naar opklimmende x-waarden, en de nummering van deze punten naar opklimmende z-waarden. Dat dit redelijk is, zal men inzien aan de hand van figuur 5.

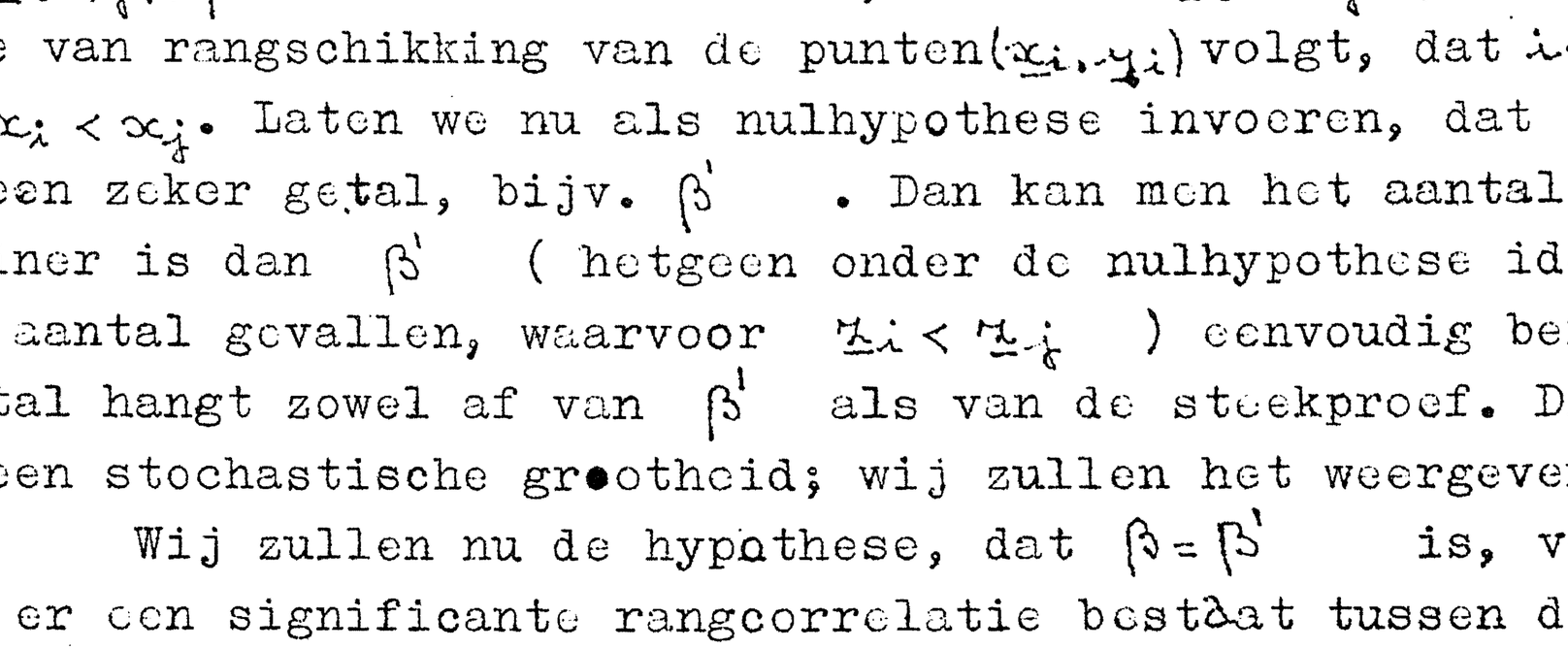


Fig. 5

In het hier getekende geval (dat suggereert, dat β' te klein is genomen) bestaat er inderdaad een (volkomen) rangcorrelatie tussen beide rangschikkingen; wij hebben immers

$$x_1 < x_2 < x_3 < x_4$$

$$y_1 < y_2 < y_3 < y_4$$

Voorlopig laten wij in het midden, in hoeverre hier ($n = 4$) reeds van een significante rangcorrelatie gesproken kan worden.

Het verband tussen deze alinea en de voorafgaande is gelegen in Kendall's rangcorrelatiecoëfficiënt. Deze auteur berekent immers, om de rangcorrelatie tussen de reeksen $x_1, \ldots, x_n$ en $y_1, \ldots, y_n$ na te gaan, het aantal gevallen, waarin $x_i < x_j$ samengaat met $y_i < y_j$, d.w.z. hij berekent: $q(\beta')$. Daarbij is onder de hypothese, dat de verdeling van y_i onafhankelijk is van i (of ook: onder de hypothese, dat $\beta = \beta'$) de verdeling van $q(\beta')$, dus van het aantal keren, dat $D(i,j) > \beta'$, bekend. Kendall's rangcorrelatiecoëfficiënt is nl. niets anders dan

$$\tau = 1 - \frac{2q(\beta')}{\binom{n}{2}}$$

Wij nemen dan een geheel getal r zodanig, dat

$$P[q(\beta') \leqq r-1 \mid \beta = \beta'] = \tfrac{1}{2} p(n,r)$$

waarbij p een kleine fractie (de onbetrouwbaarheid) is. Vervolgens nummeren we de hellingen $D(i,j)$ naar opklimmende grootte:

$$D^{(1)} < D^{(2)} < \cdots\cdots < D^{(N)},$$

waarbij $N = \frac{1}{2}n(n-1)$. Dan is $(D^{(r)}, D^{(N-r+1)})$ een betrouwbaarheidsinterval met onbetrouwbaarheid p:

$$P[D^{(r)} \leqq \beta \leqq D^{(N-r+1)} \mid \beta] = 1 - p(n,r)$$

Door A. van Wijgaarden (te verschijnen) is p(n,r) getabelleerd voor $n = 1, \ldots, 40$.

6. ENKELE OPMERKINGEN.

6.1. Schattingen van α.

Vrijwel algemeen geldt, dat, als een methode b als schatting van β geeft, voor α die schatting a wordt gekozen, die voldoet aan

$$\bar{y} = a + b\bar{x}$$

Zowel Wald als Housner en Brennan tonen aan, dat hun a een bruikbare schatting is.

Bij de methoden der hellingen kan voor α en β gezamenlijk een rechthoekig betrouwbaarheidsgebied $a^0 \leqq \alpha \leqq a', b^0 \leqq \beta \leqq b'$ gevonden worden.

6.2 Meetfouten in ξ.

Elk der vier behandelde methoden laat meet fouten u in de ξ's toe, echter niet in onbeperkte mate. De methode van Wald is het meest tolerant; zij eist slechts, dat de indeling in de twee groepen steeds onafhankelijk van de meetfouten u plaats heeft. De overige drie methoden eisen een strikte identiteit tussen de ξ- volgorde en de x-volgorde, niet slechts voor de x-en van de onderzochte steekproef, maar van alle eventuele steekproeven. Is men niet zeker van de volgorde van enkele der punten, dan kan men het beste de nummering van de betreffende x-en "random" kiezen, onafhankelijk van hun u's. De methode verliest slechts zijn bruikbaarheid, indien men de volgorde in het geheel niet kent.

Het is denkbaar, dat de "vertekening" van het resultaat optredende wanneer de u's te groot zijn en wanneer men niet door "randen" te ordenen er voldoende mee rekening heeft gehouden, van dezelfde orde van grootte is als bij de kleinste-kwadraten-schatting (1.3.2). Dit is nog niet onderzocht.

6.3. De eenvoud der berekeningen.

Tenzij de x-en alle eenvoudige waarden (d.w.z. kleine gehele getallen) aannemen, is de kleinste-kwadraten-schatting minder eenvoudig te berekenen dan de schatting van Wald en die van Housner en Brennan. Bij Wald behoeft men slechts de punten te ordenen, gemiddelden te bepalen, a te trekken en te delen; bij Housner en Brennan kan men, na de ordening van de punten, volstaan de waarden x_i en y_i met de eenvoudige getallen $(i-\bar{x})$ te vermenigvuldigen. Dit wordt echter van minder belang indien men over een goede rekenmachine beschikt.

Bij de methoden der hellingen kan men veelal het eenvoudigst de richtingscoëfficiënt van de afzonderlijke hellingen uit een spreidingsdiagram aflezen. Wanneer men evenwijdig aan elk dezer hellingen voerstralen uit de oorsprong trekt, kan men eenvoudig door het aftellen van deze voerstralen (naar opklimmende hoeken met de x-as) de grenzen van het betrouwbaarheidsinterval vinden. Voor grotere n wordt het werk nodig voor de volledige methode, prohibitief. Bij n = 40 bijv. moeten $\binom{40}{2}$ =780 hellingen worden beschouwd.

Voor de onvolledige methode kunnen ook kleine kaarten handig zijn. Daartoe schrijve men op het eerste kaartje x_1 en y_1, ..., op het m-de kaartje x_m en y_m, daarna weer op het eerste kaartje x_{m+1} en y_{m+1}, op het m-de kaartje x_n en y_n ; vervolgens berekent men op het eerste kaartje D(1,m+1), ..., op het m-de D(m,n). Tenslotte rangschikt men de kaarten naar opklimmende waarden van D en kiest men de r-de en (m-r+1)-ste.

6.4. Doeltreffendheid der puntschattingen.

Het algemene probleem van de doeltreffendheid van de hierboven beschouwde puntschattingen is zeer gecompliceerd, omdat zij op een weinig handelbare wijze afhangt van de $\xi_i (i=1,\dots,n)$ Wij zullen daarom ons beperken tot het volgende geval:

a) Alle ξ_i liggen op gelijke afstanden (zij zijn aequidistant) d.w.z. $\xi_i = c + di$. Wanneer we dan n oneven veronderstellen, kunnen we zonder verlies van algemeenheid de ξ's de waarden

$$-m, -m+1, \dots, -1, 0, 1, \dots, m-1, m$$

laten aannemen.

b) Alle meetfouten $\underline{u}_i$ zijn $\equiv 0$

c) Alle stochastische grootheden $(\underline{v}_i + \underline{w}_i)$ zijn ongecorreleerd en hebben σ^2 tot variantie.

In dit geval is de kleinste-kwadraten-schatting $\underline{b}$ het doeltreffendst, zoals reeds in § 13 is opgemerkt.

Haar variantie is

$$(6.4.1) \qquad \operatorname{var} \underline{b} = \frac{1}{n} \frac{\sigma^2}{\operatorname{var} x} = \frac{3\sigma^2}{n(m+1)(2m+1)}$$

zoals men gemakkelijk kan narekenen.

De doeltreffendheid van Wald's schatting is door Bartlett (1949) nader onderzocht. Daartoe generaliseerde hij deze schatting door niet slechts de helling van de rechte van het zwaartepunt van de linkerhelft der punten naar dat van de rechterhelft te beschouwen: hij deelt de punten in drie groepen in (naar opklimmende x-waarden), zodanig dat de eerste en de derde groep elk k punten $(1 \leqq k \leqq m)$ bevat, en neemt als schatting

$$(6.4.2) \qquad \underline{b}_k = \frac{\sum_{1}^{k} \underline{y}_i - \sum_{n-k+1}^{n} \underline{y}_i}{\sum_{1}^{k} \underline{x}_i - \sum_{n-k+1}^{n} \underline{x}_i}$$

dus de helling van de rechte, die de zwaartepunten van de eerste en de derde groep verbindt. Men kan eenvoudig narekenen, dat de variantie van $\underline{b}_k$ onder de veronderstellingen van deze paragraaf gelijk is aan

$$(6.4.3) \qquad \operatorname{var} \underline{b}_k = \frac{2\sigma^2}{k(2m-k+1)^2},$$

zodat de doeltreffendheid gegeven wordt door

$$(6.4.4) \qquad E_k = \frac{\operatorname{var} \underline{b}}{\operatorname{var} \underline{b}_k} = \frac{3k(2m+1-k)^2}{2m(m+1)(2m+1)}$$

Bij Wald is k = m; dus is de doeltreffendheid van Wald's schatting:

$$(6.4.5) \qquad E_w = \frac{3}{4} \cdot \frac{n+1}{n} \gtrsim \frac{3}{4}.$$

Bartlett bepaalt nu k zodanig, dat F_k maximaal is; differentiëren en nul stellen $k = \frac{1}{3}n$. Daarom stelt hij voor de punten in 3 gelijke groepen te verdelen. De middelste groep heeft dan geen enkele invloed op de grootte van deze schatting $\underline{b}_B$, behoudens haar bijdrage tot de nummering. Haar doeltreffendheid bedraagt

$$(6.4.6) \qquad F_B = \frac{8}{9}\,\frac{n^2}{n^2-1} \gtrsim \frac{8}{9}$$

Eenvoudig blijkt, dat onder de hier gebruikte veronderstellingen de schatting van Housner en Brennan identiek is met de kleinste- kwadraten-schatting. Immers, we hebben hier (vgl. (3.2.2)):

$$(6.4.7) \qquad \underline{b}_{HB} = \frac{cov(i,y)}{cov(i,x)} = \frac{cov(x,y)}{var\,x},$$

zodat de doeltreffendheid van deze schatting in dit geval identiek gelijk aan 1 is.

Ook met behulp van de methoden der hellingen kan men tot puntschattingen komen. Voor de schatting van de onvolledige methode kan men nemen,

$$(6.4.8) \qquad \overline{\underline{D}} = \frac{1}{m}\sum_{i=1}^{m} \underline{D}(i, m+i).$$

Haar variantie is:

$$(6.4.9) \qquad \operatorname{var} \overline{\underline{D}} = \frac{1}{m}\,\frac{2\sigma^2}{(m+1)^2}$$

en haar doeltreffendheid

$$(6.4.10) \qquad F_{\overline{\underline{D}}} = \frac{3}{4}\,\frac{n+1}{n} \gtrsim \frac{3}{4},$$

dus identiek met die van Wald. Hierbij zij aangetekend, dat, als n haar laagste waarde, nl. 3 (wij hebben aangenomen, dat n oneven is), aanneemt, de doeltreffendheid van $\underline{b}_W$, $\underline{b}_B$ en $\overline{\underline{D}}$ gelijk is aan 1. Hieruit blijkt, dat het middelste punt, dat door geen dezer drie methoden wordt gebruikt, niets tot de kennis van β bijdraagt; uiteraard geldt dit slechts onder de hier gebezigde veronderstelling, dat de $x\,(=\xi)$-waarde van dit punt midden tussen de $x(=\xi)$ -waarden der beide andere punten ligt.

Voor de volledige methode kan men als puntschatting nemen

$$(6.4.11) \qquad \overline{\underline{D}}' = \frac{\sum_{i<j}\sum (x_i - x_j)\,\underline{D}(ij)}{\sum\sum (x_i - x_j)} = \frac{\sum_{i<j}\sum (y_i - \underline{y}_{ij})}{\sum\sum (x_i - x_j)}.$$

Deze schatting is identiek met die van Housner en Brennan (vgl. (3.2.1)) en vereist dus geen afzonderlijke behandeling.[5])

Het blijkt dus, dat, althans in het hier beschouwde geval, de schatting van Housner en Brennan de voorkeur verdient. Aangezien zij bovendien betrekkelijk eenvoudig berekend kan worden, is zij zeer geschikt, wanneer men slechts een puntschatting van β wenst te kennen.
Wenst men daarentegen bovendien een maatstaf voor de betrouwbaarheid van deze schatting, dan moet men òf betrouwbaarheidsintervallen òf standaardfouten berekenen. Voor eerstgenoemde intervallen heeft men normaliteitsveronderstellingen nodig, behalve bij de methoden der hellingen. Behoudens bij deze methoden gaat het voordeel van het geringe rekenwerk dan verloren.

7. LITERATUUR.

In deze lijst zijn mede opgenomen de titels van enkele artikelen, waarin hier niet besproken methoden worden gegeven.

Bartlett, M.S. (1949), Fitting a straight line when both variables are subject to error. Biometrics, vol. 5, pp 207-212.

Drion, E.F. (1951),Estimation of the parameters of a straigt line and of the variances of the variables, if they are both subject to error. Proceedings Kon. Ned. Akad,vol. 54, Series A, pp. 256-260.

Hemelrijk, J, (1949 a), Construction of a confidence regine for a line. Proceedings Kon. Ned. Akad., vol. 52, pp. 995-1005

Hemelrijk, J.,(1949 b), Over de bepaling van betrouwbaarheidsintervallen en schattingen van de coëfficiënten van een rechte lijn uit een aantal onnauwkeurig waargenomen punten. Overzichtsrapport no.1 van de Statistische Afdeling van het Mathematisch Centrum.

Housner, G.W. and J.F. Brennan (1948), The estimation of linear trends. Aan Math. Statist. vol. 19, pp. 380-393.

Mood, A.M. (1950), Introduction to the Theory of Statistics. New York.

Nair, K.R. and M.P. Shrivastava (1942), On a simple method of curve fitting. Sankhya, vol. 6, pp. 121-132.

---

5) De lezer zal inzien, dat het toekennen van "gewichten" $(x_i - x_j)$ voor de volledige methode onder de veronderstellingen van deze paragraaf overbodig is.

Nair, K.R. and K.S. Banerjee (1942), A note on fitting of straight lines if both variables are subject to error. Sankhya, vol. 6, pp. 331 e.v.

Scott, E.L. (1950), Note on consistent estimates of linear structural relation between two variables. Ann. Math. Statist., vol. 21, pp. 284-288.

Theil, H. (1950), A rank-invariant method of linear and Polynnial regression analysis. Proceedings Kon. Ned. Akad., vol. 53, pp. 386-392, 521-525, 1397-1412.

Wald, A. (1940), The fitting of straight lines if both variables are subject to error. Ann. Math. Statist., vol. 11, pp. 284- 300.

Wijngaarden, A. van (1950), Table of the cumulative symmetric binomial distribution. Proceedings Kon. Ned. Akad., vol. 53 pp. 857-868.

Hieronder volgt een tabel voor de waarschijnlijkheid

$$P = 1 - 2^{1-m} \sum_{s=0}^{r-1} \binom{m}{s}$$

voor m $\leqq$ 50 (dus voor steekproeven tot de uitgebreidheid 100) en voor r zodanig, dat P $\geqq$ 0,7 is. In de rijen is m afgezet, in de kolommen r.

Getabbelleerd is: $P.10^3$

r →	1	2	3
m ↓			
3	750		
4	875		
5	937		
6	969	781	
7	984	875	
8	992	930	711
9	996	961	820
10	998	979	891

	1	2	3	4	5
11	999	988	935	773	
12	1000	994	961	854	
13		997	978	908	733
14		998	987	943	820
15		999	993	965	882

	2	3	4	5	6	7	8
16	999	996	979	923	790		
17	1000	998	987	951	857		
18		999	992	969	904	762	
19		999	996	981	936	833	
20		1000	997	988	959	885	737

	4	5	6	7	8	9	10
21	999	993	973	922	811		
22	999	996	983	948	866	714	
23	1000	997	989	965	907	790	
24		998	993	977	936	848	
25		999	996	985	957	892	770

	5	6	7	8	9	10	11	12
26	999	998	991	971	924	831		
27	1000	998	994	981	948	878	752	
28		999	996	987	964	913	815	
29		999	998	992	976	939	864	735
30		1000	999	995	984	957	901	800

	7	8	9	10	11	12	13	14
31	999	997	989	971	929	850	719	
32	999	998	993	980	950	890	785	
33	1000	999	995	986	965	920	837	704
34		999	997	991	976	942	879	771
35		999	998	994	983	959	910	825

	8	9	10	11	12	13	14	15	16	17
36	1000	999	996	989	971	935	868	757		
37		999	997	992	980	953	901	812		
38		1000	998	995	986	966	927	857	744	
39			999	997	991	976	947	892	800	
40			999	998	994	983	962	919	846	732

	10	11	12	13	14	15	16	17	18	19
41	1000	999	996	988	972	940	883	789		
42		999	997	992	980	956	912	836	720	
43		999	998	995	986	968	934	874	778	
44		1000	999	996	990	977	951	904	826	709
45			999	998	993	984	964	928	865	767

	12	13	14	15	16	17	18	19	20	21
46	999	998	995	989	974	946	896	816		
47	1000	999	997	992	981	960	921	856	757	
48		999	998	994	987	971	941	889	807	
49		1000	999	996	991	979	956	915	848	747
50			999	997	993	985	967	935	881	797

MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 59

Cursus Parametervrije Methoden

V. Parametervrije tolerantiegrenzen

door

T.J.Terpstra

October 1951

§1. Inleiding.

De theorie der tolerantiegrenzen vindt vooral toepassing in de kwaliteitscontrole. De problemen, die hier optreden, zijn meestal van de volgende aard.

1. Een grote partij goederen is volgens één of ander productieschema gefabriceerd. Van deze goederen wil men nu een bepaald kwaliteitskenmerk controleren (b.v. de diameter van een grote hoeveelheid schroefjes). Deze grootheid, die we zullen aangeven met x, neemt op de partij verschillende waarden aan.
Een methode, om de gehele partij op het kenmerk x te controleren, is, alle artikelen te meten. Dat is echter zeer tijdrovend en kostbaar.
In de kwaliteitscontrole volgt men dan ook meestal een andere weg. In plaats van elk artikel te meten, neemt men een steekproef uit de gehele partij en bepaalt voor deze steekproef de waarden x. Uit deze waarden leidt men dan een waarschijnlijkheidstheoretische uitspraak af, die geldt voor de gehele partij.

2. Een ander geval doet zich voor, indien men een fabricageproces "onder statistische controle" wil houden, daarbij bepaalde eisen aan de kwaliteit van het product stellende. Men neemt dan van tijd tot tijd een steekproef en trekt uit de gevonden waarden conclusies omtrent het procédé.

Voor de mathematische verwerking bedienen wij ons van het volgende mathematische model. Wij onderstellen dat de grootheid $\underline{x}$[1)] op een collectie (dat is b.v. de te keuren partij of de verzameling van alle, in een bepaald tijdvak geproduceerde artikelen

1) stochastische grootheden worden onderstreept aangegeven.

bij één of andere fabricagemethode) een wh-verdeling[1] bezit, terwijl het optreden van gelijke waarnemingen onmogelijk is. Uit deze collectie worden dan, met of zonder teruglegging, één of meer aselecte[2] steekproeven genomen.
Hieraan is bij bovengenoemde voorbeelden voldaan, indien bij de keuring van een partij de steekproef op aselecte wijze genomen wordt, resp. indien het productieproces inderdaad "onder controle" is.

In verband met het bovenstaande zullen we de volgende twee, elkaar aanvullende categorieën van problemen behandelen.

A. Bij de <u>eerste groep van problemen</u> wordt ondersteld, dat aan een kenmerk $\underline{x}$ van een product twee vaste grenzen L_1 en L_2 opgelegd worden. Ligt de $\underline{x}$ van een product binnen deze grenzen, dan wordt het product goedgekeurd, ligt de $\underline{x}$ buiten de grenzen, dan wordt het afgekeurd. Deze grenzen L_1 en L_2 worden in de practijk meestal tolerantiegrenzen genoemd.
Wat men nu graag zal willen weten, is de werkelijke fractie p_o van producten uit de collectie, waarvoor $\underline{x}$ tussen L_1 en L_2 ligt. Voor deze onbekende fractie p_o kan men nu uit de gevonden steekproefwaarden een <u>betrouwbaarheidsinterval</u> $[\underline{p}_1, \underline{p}_2]$ berekenen, waarin de werkelijke fractie p_o ligt, behouden een van te voren gekozen onbetrouwbaarheid α (α is b.v. 0,05). Dit probleem behoort dus tot de theorie der betrouwbaarheidsintervallen.

<u>Opmerking</u>: Het gebruik van de naam "tolerantiegrenzen" voor L_1 en L_2 is nogal verwarrend, daar deze term in de mathematische statistiek gewoonlijk voor grootheden van een ander karakter gebruikt wordt, namelijk voor de onder B te bespreken stochastische grootheden $\underline{L}_1$ en $\underline{L}_2$. Het is van belang deze laatste grootheden goed te onderscheiden van de zojuist besprokene.

B. In de <u>tweede groep van problemen</u>, die we willen beschouwen, worden niet, zoals onder A, vaste grenzen L_1 en L_2 gekozen, maar worden de grenzen gedefiniëerd als functies van de naar opklimmende grootte gerangschikte steekproefwaarden $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$.
Dus is

$$\underline{L}_1 = L_1(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n),$$
$$\underline{L}_2 = L_2(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$$

De grenzen zijn dus stochastisch.
De fractie $\underline{p}$ van de collectie met een $\underline{x}$ tussen $\underline{L}_1$ en $\underline{L}_2$ is dus eveneens stochastisch. Voor deze grootheid $\underline{p}$ kunnen we schrijven

1) wh betekent waarschijnlijkheid.
2) aselect is de door Prof. Dr D. van Dantzig voorgestelde vertaling van "random".

$$\underline{p} = F(\underline{L}_2) - F(\underline{L}_1),$$

waarin $F(x)$ de verdelingsfunctie van $\underline{x}$ is. Deze verdelingsfunctie wordt gedefiniëerd door

$$F(x) \equiv P\left[\underline{x} \leq x\right] .$$

De stochastische grootheid $\underline{p}$ bezit eveneens een verdelingsfunctie $G(p)$. Wanneerdeze verdelingsfunctie $G(p)$ bekend is, kunnen we voor $\underline{p}$, bij gegeven waarde van α, een voorspellingsinterval van de gedaante $[b,1]$ bepalen. Men kan ook voorspellingsintervallen van de gedaante $[c,d]$ aangeven, maar meestal is een interval met $d = 1$ als bovengrens voor de toepassingen alleen van belang.
Definiëren we de grenzen $\underline{L}_1$ en $\underline{L}_2$ anders, dan bezit $\underline{p}$ een verdelingsfunctie $G^*(p)$, die in het algemeen van $G(p)$ zal verschillen, zodat (bij dezelfde α) ook de ondergrens b^* van het met behulp van de functie $G^*(p)$ bepaalde voorspellingsinterval $[b^*,1]$ voor $\underline{p}$ van b zal verschillen.

Met behulp van de theorie der tolerantiegrenzen kunnen we nu de grenzen $\underline{L}_1$ en $\underline{L}_2$ zodanig definiëren, dat de onder-grens b van het voorspellingsinterval gelijk is aan een van te voren gekozen waarde β, m.a.w.: de grenzen $\underline{L}_1$ en $\underline{L}_2$ kunnen zo gedefiniëerd worden, dat de fractie $\underline{p}$ van elementen uit de collectie met een $\underline{x}$ tussen de uit de steekproef bepaalde waarden voor $\underline{L}_1$ en $\underline{L}_2$, behoudens een onbetrouwbaarheid α, groter is dan de van te voren gekozen waarde β.
Tussen de grenzen $\underline{L}_1$ en L_2, α en β bestaat dus de relatie

$$P\left[\underline{p} \geq \beta\right] = P\left[F(\underline{L}_2) - F(\underline{L}_1) \geq \beta\right] \geq 1 - \alpha . \qquad (1.1)$$

<u>Definitie</u>:

Wanneer betrekking (1.1) onafhankelijk geldt van de vorm der verdelingsfunctie $F(x)$ heten $\underline{L}_1$ en $\underline{L}_2$ <u>parametervrije tolerantiegrenzen</u>.

In plaats van twee tolerantiegrenzen wordt ook dikwijls slechts één(namelijk een onderste of een bovenste) tolerantiegrenz bepaald.
Voor deze grenzen geldt eveneens bovenstaande definitie, met $\underline{L}_2 = +\infty$ voor een onderste tolerantiegrens en $\underline{L}_1 = -\infty$ voor een bovenste tolerantiegrens.

Voor $\underline{L}_1$ en $\underline{L}_2$ worden in de practijk meestal twee uit de steekproef $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ bepaalde waarden $\underline{x}_r$ en $\underline{x}_{n-s+1}$ gekozen, waardoor (1.1) overgaat in

$$P\left[F(\underline{x}_{n-s+1}) - F(\underline{x}_r) \geq \beta\right] \geq 1 - \alpha. \qquad (1.2)$$

Met behulp van deze betrekking, welk een verband geeft tussen

n,r,s,α en β kan steeds één van de grootheden opgelost worden, wanneer de overigen gegeven zijn.
De volgende problemen kunnen dus met behulp van deze theorie opgelost worden:

a. Voor gegeven waarden van n,r, α en β kan een interval $[\underline{x}_r, \underline{x}_{n-s+1}]$ bepaald worden, met t = r + s zo groot mogelijk (d.w.z. het interval zo klein mogelijk), waarin, behoudens een onbetrouwbaarheid α , de $\underline{x}$ ligt van tenminste een fractie β van de collectie. Wanneer alleen de waarden n, α en β gegeven zijn, kunnen verschillende intervallen $[x_r, x_{n-s+1}]$ bepaald worden, waarvoor het bovenstaande geldt. Soms kan echter dan r zo gekozen worden, dat het interval $[x_r, x_{n-s+1}]$ gemiddeld ongeveer zo kort mogelijk is. (zie §2.2).

b. Voor gegeven waarden van n, r en s ,dus voor een steekproef van de uitgebreidheid n, waaruit de waarden $\underline{x}_r$ en $\underline{x}_{n-s+1}$ als tolerantiegrenzen gekozen worden, kan men

1. voor een gegeven waarde α (dat is de toegelaten onbetrouwbaarheid) de maximale waarde β berekenen, waarboven, behoudens een onbetrouwbaarheid α , de fractie p van elementen der collectie ligt met een $\underline{x}$ tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$.
2. voor gegeven waarde van β de minimale onbetrouwbaarheid α berekenen, waarvoor geldt, dat tenminste een fractie β van de collectie een $\underline{x}$ tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$ bezit.

c. Bij gegeven r,s, α en β kan men de minimale uitgebreidheid n van de steekproef berekenen, waarvoor geldt, dat, behoudens een onbetrouwbaarheid α , de waarden $\underline{x}$ van tenminste een fractie β van de collectie, tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$ liggen.
Wanneer in het bijzonder de kleinste en grootste waarde uit de steekproef als grenzen gekozen worden, zal n zo klein mogelijk zijn. Indien de mogelijkheid van het optreden van "uitschieters" aanwezig is, geeft men vaak de voorkeur aan andere waarden (dus met $r > 1$, $s > 1$). Hierdoor wordt de invloed van deze ver weggelegen waarnemingen te niet gedaan, tenzij er meer dan r-1 resp. s-1 aan één zijde liggen. Daartoe moet n n dan dus vergroot worden.

De hierboven vermelde problemen kunnen eveneens geformuleerd worden, wanneer men in plaats van één steekproef en de gehele collectie, twee steekproeven beschouwt.
Deze problemen treden in de practijk ook dikwijls op: uit een eerste ontvangen partij goederen leidt men voorspellingen af omtrent een tweede partij. Of men heeft b.v. een kleine partij goederen en wil deze controleren door middel van een steekproef hieruit, waarbij dan het restant van deze partij als de tweede

steekproef wordt beschouwd.
Wanneer de beide steekproeven van de uitgebreidheid n en N zijn en $\underline{m}$ is het aantal elementen uit de tweede steekproef waarvoor $\underline{x}$ tussen de uit de eerste steekproef bepaalde waarden $\underline{x}_r$ en $\underline{x}_{n-s+1}$ ligt, dan neemt betrekking (1.2) de volgende gedaante aan:

$$P\left[\underline{m} \geq N_o \,|\, n,N;\ r,s\right] \geq 1-\alpha . \qquad (1.3)$$

Het <u>waarschijnlijkheidsveld</u> wordt hier gevormd door de (n+N)-dimensionale ruimte, waarvan ieder punt twee speciale steekproeven voorstelt. <u>Men moet steeds een nieuwe steekproef nemen, om voor een volgende, tweede steekproef iets te kunnen voorspellen.</u>

De theorie van de tolerantiegrenzen voor twee steekproeven kunnen we als de meest algemene beschouwen, daar die voor een steekproef uit een oneindig groot onderstelde collectie, hieruit door een limiet-overgang volgt.
Stellen we n.l.

$$\lim_{N\to\infty} \frac{N_o}{N} = \beta ,$$
$$\text{en } \lim_{N\to\infty} \frac{m}{N} = \underline{p} ,$$

dan gaat betrekking (1.3) over in (1.2)

In de volgende paragraaf zullen we daarom eerst de formules afleiden voor het meer algemene probleem van twee steekproeven.
De formules voor een steekproef uit een oneindig groot onderstelde collectie, kunnen hieruit dan door bovenstaande limiet-overgang afgeleid worden.
Voor deze formules geven we bovendien nog een tweede afleiding.
Daarna zullen we verschillende methoden bespreken voor het bepalen van simultane tolerantiegrenzen voor een aantal al of niet stochastisch onafhankelijke grootheden.
In de laatste paragraaf wordt tenslotte het onder A vermelde probleem behandeld.

§2. Tolerantiegrenzen voor kleine collecties.

2.1. Een onderste tolerantiegrens.

Wanneer $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ en $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_N$ de naar opklimmende grootte gerangschikte waarden van het onderzochte kenmerk uit beide steekproeven zijn, zullen we eerst de wh berekenen, dat $\underline{y}$ voor precies m_o elementen uit de tweede

steekproef groter is dan de waarde $\underline{x}_r$ uit de eerste steekproef. Wanneer $\underline{m}$ het aantal elementen uit de tweede steekproef is, waarvoor $\underline{y} > \underline{x}_r$ is, dan willen we dus de kans $P[\underline{m} = m_o | n,N;\ r,s=0]$[1] berekenen.

Deze wh kan als volgt bepaald worden.

Daar de twee steekproeven op aselecte wijze uit eenzelfde collectie genomen zijn, zijn alle $(N+n)!$ permutaties van de waarden $\underline{x}$ en $\underline{y}$ even waarschijnlijk. Wanneer de waarnemingen naar opklimmende grootte gerangschikt worden, zijn er in totaal $\binom{N+n}{n}$ verschillende situaties mogelijk, d.w.z. de n waarden van de eerste steekproef kunnen op $\binom{N+n}{n}$ wijzen tussen de y's verspreid liggen (er zij n.l. $\binom{a}{b}$ verschillende manieren om b elementen uit a elementen te kiezen).

Daar ieder van deze situatie $N! \,.\, n!$ permutaties vertegenwoordigt, zijn ze alle even waarschijnlijk en bezitten dus ieder de wh

$$\frac{1}{\binom{N+n}{n}} .$$

Willen we de wh berekenen, dat precies m_o elemehten $\underline{y}$ groter zijn dan $\underline{x}_r$, dan kunnen we dit doen, door het aantal verschillende situaties te bepalen, waarvoor hier aan voldaan is (vgl. fig. 1).

$$\begin{array}{l|c|l} x_1, x_2, \ldots, x_{r-1}, & x_r, & x_{r+1}, \ldots, x_n. \\ y_1, y_2, \ldots, y_{N-m_o}, & & y_{N-m_o+1}, \ldots, y_N. \end{array}$$

fig. 1 ($\underline{m} = m_o$)

Beschouwen wij het linker-groepje waarnemingen apart. Dit bevat in totaal $N-m_o+r-1$ elementen en kan op $\binom{N-m_o+r-1}{r-1}$ verschillende manieren in twee groepjes van r-1 waarden van $\underline{x}$ en $N-m_o$ waarden van $\underline{y}$ worden gesplitst. Ieder van deze splitsingen geeft een situatie met $\underline{m} = m_o$.

Op analoge wijze zijn er $\binom{m_o+n-r}{n-r}$ splitsingen van het rechtergroepje van waarnemingen, die $\underline{m} = m_o$ laten. In totaal zijn er dus $\binom{N-m_o+r-1}{r-1}\binom{m_o+n-r}{n-r}$ situaties met $\underline{m} = m_o$, dus is:

$$P[\underline{m} = m_o | n,\ N;\ r, 0] = \binom{N-m_o+r-1}{r-1}\binom{m_o+n-r}{n-r} \qquad (2.1.1)$$

Voor deze wh geldt de recurrente betrekking

$$P[\underline{m} = m_o-1 | n,\ N;\ r, 0] = \frac{N-m_o+r}{N-m_o+1} \cdot \frac{m_o}{m_o+n-r} \cdot P[\underline{m} = m_o | n,N;\ r,o], \qquad (2.1.2)$$

die bij exacte berekeningen (voor kleine n en N) van nut kan zijn.

Met behulp van deze formules kunnen nu geheel analoge problemen opgelost worden, als die, vermeld in § 1.

1e. Men kan de grootste waarde N_α bepalen, waarboven,behoudens

1) s = 0 betekent, dat een enkele onderste tolerantiegrens wordt gekozen.

een onbetrouwbaarheid α , een aantal $\underline{m}$ van elementen uit de tweede steekproef ligt, waarvoor $\underline{y}$ groter dan $\underline{x}_r$ is.
Deze ondergrens N_α is het grootste gehele getal, waarvoor de betrekking

$$\sum_{m_s = N_\alpha}^{N} P\left[\underline{m}(n, N; r, 0) = m_o\right] \geq 1 - \alpha . \qquad (2.1.3)$$

geldt.
Men bepaalt dus een voorspellingsinterval voor $\underline{m}$.

2e. Omgekeerd kan men met behulp van (2.1.1.) en (2.1.3) de minimale onbetrouwbaarheid α bepalen, waarvoor geldt, dat voor tenminste een gegeven aantal N_α van elementen uit de tweede steekproef $\underline{y}$ groter dan $\underline{x}_r$ is.

3e. Men kan de grootste waarde van r bepalen ($1 \leq r \leq n$), waarvoor geldt, dat, behoudens een onbetrouwbaarheid α , tenminste een aantal N_α van elementen uit de tweede steekproef een $\underline{y}$ bezit, die groter is dan de uit de eerste steekproef bepaalde waarde $\underline{x}_r$.

4e. Men kan de kleinste waarde n bepalen, dus de minimale uitgebreidheid van de eerste steekproef, waarvoor geldt, dat, behoudens een onbetrouwbaarheid α , voor tenminste een gegeven aantal N_α van elementen uit de tweede steekproef, $\underline{y}$ groter is dan de waarde $\underline{x}_r$ uit de eerste steekproef.

5e. Eveneens kan men de minimale uitgebreidheid N van de tweede steekproef berekenen, waarvoor geldt, dat de $\underline{y}$, behoudens een onbetrouwbaarheid α , voor minstens N_α elementen uit deze steekproef groter is dan $\underline{x}_r$.

Voor kleine steekproeven kunnen bovenstaande problemen exact opgelost worden met behulp van de recurrente betrekking (2.1.2). Hierbij kan men gebruik maken van de tabel met binomiaal coefficiënten in T.C. FRY [18] [1].
Wanneer de tweede steekproef groot is, kan men gebruik maken van de in § 3.4 vermelde nomogrammen en formules, die een benaderde oplossing geven voor de problemen.

2.2 Een onderste en bovenste tolerantiegrens.

We willen nu de analoge formules afleiden voor het geval, dat men twee tolerantiegrenzen $\underline{x}_r$ en $\underline{x}_{n-s+1}$ uit de eerste steekproef kiest.
Allereerst berekenen we de wh, dat voor precies m_o elementen uit de tweede steekproef $\underline{y}$ tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$ ligt, dus de wh

$$P\left[\underline{m} = m_o \mid n, N; r, s\right] .$$

Zijn $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ en $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_N$ de naar opklimmende grootte gerangschikte waarden der twee steekproeven, dan moet dus een

1) Zie litteratuurlijst.

situatie van de in fig. 2 geschetste aard optreden(hierbij liggen m_o waarden $\underline{y}$ tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$).

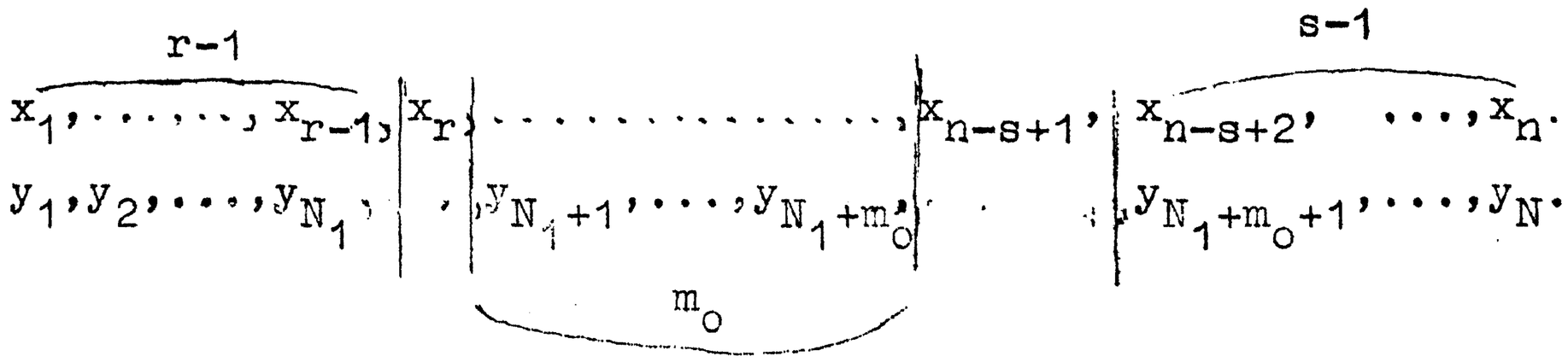


fig. 2

Zoals in § 2.1 reeds uiteengezet is, zijn er $\binom{N+n}{n}$ verschillende, even waarschijnlijke situaties mogelijk.
De wh $P\left[\underline{m} = m_o \mid n,N\ r,s\right]$ kunnen we berekenen, door het aantal verschillende situaties te bepalen, waarbij precies m_o $\underline{y}$'s tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$ liggen.
Dit aantal kunnen we gemakkelijk bepalen aan de hand van fig.2. Voor een vaste waarde N_1 kan het linker groepje waarnemingen in de figuur op $\binom{N_1+r-1}{r-1}$ verschillende manieren in twee groepjes van r-1 waarden $\underline{x}$ en N_1 waarden $\underline{y}$ gesplitst worden, zonder dat het aantal m_o van waarden $\underline{y}$ tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$ verandert; het middelste groepje op $\binom{n-(r+s)+m_o}{m_o}$ manieren en het rechter groepje op $\binom{N-(N_1+m_o)+(s-1)}{s-1}$ manieren (in twee groepjes van n-r-s en m_o resp. s-1 en $N-N_1-m_o$ waarden $\underline{x}$ en $\underline{y}$). Daar N_1 de waarden 0 tot en met $N-m_o$ kan doorlopen, zijn er in totaal

$$\sum_{N_1=0}^{N-m_o}\binom{N_1+r-1}{r-1}\binom{n-(r+s)+m_o}{m_o}\binom{N-(N_1+m_o)+(s-1)}{s-1} =$$

$$=\binom{N-m_o+r+s-1}{r+s-1}\binom{m_o+n-r-s}{n-r-s}$$

verschillende situaties, waarbij precies m_o waarden $\underline{y}$ tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$ liggen.
Daar ieder van deze situaties de wh $\frac{1}{\binom{N+n}{n}}$ bezit, vinden we tenslotte

$$P\left[\underline{m} = m_o \mid n,N;r,s\right] = \frac{\binom{N-m_o+r+s-1}{r+s-1}\binom{m_o+n-r-s}{n-r-s}}{\binom{N+n}{n}} \qquad (2.2.1)$$

Uit deze formule valt het volgende op te merken:

1e. Stellen we s=0, dan vinden we formule (2.1.1.) terug voor een enkele onderste tolerantiegrens. Dit is direct duidelijk uit fig. 2 en fig. 1, daar voor s=0 de figuren identiek zijn. Formule (2.2.1) kunnen we dus als de meest algemene beschouwen.

2e. De wh $P\left[\underline{m} = m_o \mid n,N;\ r,s\right]$ is afhankelijk van t = r+s (en

niet van r en s afzonderlijk), zodat voor ieder tweetal grenzen $\underline{x}_r$ en $\underline{x}_{n-s+1}$ met t = r + s, de wh-verdeling $P\left[\underline{m} = m_o \mid n,N;\ r,s\right]$ dezelfde is.

Evenals in de vorige paragraaf kan men met behulp van (2.2.1) de grootste gehele waarde N_α bepalen, waarboven, behoudens een onbetrouwbaarheid α, het aantal $\underline{m}$ van elementen uit de tweede steekproef ligt, met $\underline{y}$ groter dan $\underline{x}_r$.
Deze ondergrens N_α is het grootste gehele getal, waarvoor de betrekking

$$\sum_{m_o = N_\alpha}^{N} P\left[\underline{m} = m_o \mid n,N;\ r,s\right] \geqq 1 - \alpha \qquad (2.2.2)$$

vervuld is.
Met behulp van (2.2.1) en (2.2.2) kunnen weer geheel analoge problemen opgelost worden als die, vermeld onder de punten (a), (b) en (c) in §1.

Er is in dit verband echter één speciaal punt, dat nadere aandacht verdient.
Zoals hierboven reeds opgemerkt is, is de wh $P\left[\underline{m} = m_o \mid n,N;r,s\right]$ afhankelijk van t = r + s. Wanneer men dus twee grenzen $\underline{x}_r$ en $\underline{x}_{n-s+1}$ wil bepalen met t = r + s zo groot mogelijk, waartussen, behoudens een onbetrouwbaarheid α, $\underline{y}$ ligt voor tenminste een gegeven aantal N_α van elementen uit de tweede steekproef, dan vindt men geen ondubbelzinnige oplossing. Wanneer $t = t_o$ de grootste waarde van t = r + s is, waarvoor betrekking (2.2.2) voor gegeven waarden van n, N, α en N_α vervuld is, dan voldoen n.l. alle intervallen $\left[\underline{x}_r, \underline{x}_{n-s+1}\right]$ met $r+s = t_o$ aan het gevraagde.
Men zal dus een speciale keus voor r moeten maken, die echter, om de onbetrouwbaarheid niet te beïnvloeden, onafhankelijk van de gevonden steekproefwaarden moet zijn. Men kiest deze waarde r meestal zodanig, dat men, op grond van een vermoeden omtrent de vorm der verdelingsdichtheid f(x), kan verwachten, dat deze keuze gemiddeld het korste interval $\left[\underline{x}_r, \underline{x}_{n-s+1}\right]$ geeft. Indien beide steekproeven b.v. genomen zijn uit een collectie, waarop de gemeten grootheid een symmetrische verdeling met één maximum bezit, is het centrale interval $\left[\underline{x}_r, \underline{x}_{n-r+1}\right]$ het gevraagde. Het bewijs van deze eigenschap zullen we niet geven.

2.3 Keuring van kleine partijen.

Wanneer men een kleine partij goederen van de uitgebreidheid N' wil controleren door een steekproef $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$, hieruit, kan men dezelfde theorie toepassen als in § 2.1 en

en § 2.2. Indien n.l. de steekproef op aselecte wijze uit de partij gekozen is, zijn alle $\binom{N'}{n'}$ mogelijke steekproeven even waarschijnlijk, hetgeen overeenkomt met de gelijke whn van de $\binom{N'}{n'}$ mogelijke situaties uit de vorige paragrafen. Daarbij neemt dan het restant van de N'-n' niet getrokken waarden de plaats van de tweede steekproef in. Hier treedt dus precies hetzelfde model op, met behulp waarvan de formules in § 2.1 en § 2.2 zijn afgeleid. Voor twee tolerantiegrenzen $\underline{x}_{r'}$ en $\underline{x}_{n'-s'+1}$ kan men dus formule (2.2.1) toepassen, echter met

$$n = n'$$
$$N = N' - n'$$
$$r = r'$$
$$s = s'$$
$$m_o = m_o' - \{n' - (r' + s')\}$$

Voor een enkele onderste tolerantiegrens is s' = 0, voor een bovenste tolerantiegrens is r' = 0.
De in deze paragraaf gegeven formules zijn door WILKS [11] op een andere wijze afgeleid.

§3. Tolerantiegrenzen voor grote collecties.

3.1 Inleiding.

Zoals in § 1 reeds uiteengezet is, kan de theorie der tolerantiegrenzen voor een oneindig groot onderstelde collectie door een limiet-overgang afgeleid worden uit de theorie voor twee steekproeven. Laten we n.l. de uitgebreidheid van de in de vorige paragrafen beschouwde tweede steekproef onbeperkt toenemen, dan gaan wij door deze limiet-overgang over op het keuren van grote collecties door middel van een steekproef.
In plaats van het aantal $\underline{m}$ van elementen uit de tweede steekproef met een $\underline{y}$ tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$, beschouwen nu de fracties $\underline{p}$ van elementen uit de gehele, oneindig groot onderstelde collectie, met een $\underline{x}$ tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$.
De verdelingsdichtheid van deze fractie $\underline{p}$ kan door een limiet-overgang afgeleid worden uit de wh $P\left[\underline{m} = m_o \middle| n,N;r,s\right]$ voor twee steekproeven.
Door substitutie van de waarde s = 0 (zie §2.2) in deze formule, vinden we de verdelingsdichtheid voor de fractie $\underline{p}$ uit de oneindig groot onderstelde collectie, met een $\underline{x}$ groter dan de waarde $\underline{x}_r$ uit de steekproef.

3.2 Een onderste en bovenste tolerantiegrens.

Voor de kans $P\left[\underline{m} = m_o \middle| n,N;\ r,s\right]$, dat voor m_o elementen

uit de tweede steekproef de $\underline{y}$ tussen de waarden $\underline{x}_r$ en $\underline{x}_{n-s+1}$ uit de eerste steekproef ligt, hebben we in § 2 de volgende formule gevonden

$$P\left[\underline{m}=m_o \mid n,N;r,s\right] = \frac{\binom{N-m_o+r+s-1}{r+s-1}\binom{n+m_o-r-s}{n-r-s}}{\binom{N+n}{n}}.$$

Schrijven we het rechterlid volledig uit, dan is

$$P\left[\underline{m}=m_o \mid n,N;r,s\right] = \frac{n!}{(r+s-1)!(n-r-s)!} \cdot \frac{(N-m_o+r+s-1)\ldots(N-m_o+1)(n+m_o-r-s)\ldots(m_o+1)}{(N+n)\ldots\ldots\ldots(N+1)}.$$

Dus is

$$P\left[\frac{\underline{m}}{N} = \frac{m_o}{N} \mid n,N;r,s\right] = \frac{n!}{(r+s-1)!(n-r-s)!} \cdot \frac{\left(1-\frac{m_o}{N}+\frac{r+s-1}{N}\right)\ldots\left(1-\frac{m_o}{N}+\frac{1}{N}\right)\left(\frac{m_o}{N}+\frac{n-r-s}{N}\right)\ldots\left(\frac{m_o}{N}+\frac{1}{N}\right)}{\left(1+\frac{n}{N}\right)\ \ldots\ \ \left(1+\frac{1}{N}\right)\ N},$$

en

$$P\left[a \leqslant \frac{\underline{m}}{N} \leqslant b \mid n,N;r,s\right] = \frac{n!}{(r+s-1)!(n-r-s)!} \cdot \sum_{a \leqslant \frac{m_o}{N} \leqslant b} \frac{\left(1-\frac{m_o}{N}+\frac{r+s-1}{N}\right)\ldots\left(1-\frac{m_o}{N}+\frac{1}{N}\right)\left(\frac{m_o}{N}+\frac{n-r-s}{N}\right)\ldots\left(\frac{m_o}{N}+\frac{1}{N}\right)}{\left(1+\frac{n}{N}\right)\ \ldots\ \ \left(1+\frac{1}{N}\right)N}.$$

Stellen we $\underline{p} = \frac{\underline{m}}{N}$ en $\Delta p = \frac{1}{N}$, dan is

$$P\left[a \leqslant \underline{p} \leqslant b \mid n,r,s\right] = \frac{n!}{(r+s-1)!(n-r-s)!} \cdot \sum_{a \leqslant p \leqslant b} \frac{(1-p+(r+s-1)\Delta p)\ldots(1-p+\Delta p)(p+(n-r+s)\Delta p)\ldots(p+\Delta p)\Delta p}{(1+n\,\Delta p)\ \ldots\ \ (1+\Delta p)}$$

Laten we N onbeperkt toenemen, dan nadert de som in het rechterlid tot een eindige limiet. Deze limiet is de integraal $\int_a^b g(p)dp$, met

$$g(p) = \frac{n!}{(r+s-1)!(n-r-s)!}(1-p)^{r+s-1}\ p^{n-r-s}. \qquad (3.2.1)$$

Met behulp van deze verdelingsdichtheid g(p) kunnen we de grootste waarde β berekenen, waarboven, behoudens een onbetrouwbaarheid α, de fractie $\underline{p}$ van elementen uit de collectie ligt, met een $\underline{x}$ tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$. Deze waarde β voldoet aan de betrekking

$$\int_\beta^1 g(p)dp = 1 - \alpha. \qquad (3.2.2)$$

Het interval $[\beta,1]$ is een voorspellingsinterval voor $\underline{p}$ met onbetrouwbaarheidscoëfficient α.

Met behulp van (3.2.1.) en (3.2.2) kunnen de in §1 onder

(a), (b) en (c) geformuleerde problemen opgelost worden.

Opgemerkt dient echter weer te worden, dat evenals in §2.2, de verdelingsdichtheid g(p) een functie is van de parameter t=r+s. Wil men dus een interval $[\underline{x}_r, \underline{x}_{n-s+1}]$ bepalen, waartussen, behoudens een onbetrouwbaarheid α, de $\underline{x}$ ligt van tenminste een fractie β van de elementen uit de collectie, dan is het interval niet ondubbelzinnig bepaald en moet de waarde r weer (onafhankelijk van de gevonden steekproef)gekozen worden. Deze keuze kan men soms zo doen,dat het interval gemiddeld ongeveer zo klein mogelijk is (zie de analoge opmerking in § 2.2).

3.3 Een onderste tolerantiegrens.

De verdelingsdichtheid g(p) van de fractie $\underline{p}$ van elementen uit de gehele (oneindig groot onderstelde) collectie met een $\underline{x}$ groter dan de uit de steekproef bepaalde waarde $\underline{x}_r$, volgt direct uit formule (3.2.1), door hierin de waarde s = 0 te substitueren. Dus is

$$g(p) = \frac{n!}{(r-1)!(n-r)!}\,(1-p)^{r-1}\,p^{n-r} \qquad (3.3.1)$$

De grootste waarde β, waarboven, behoudens een onbetrouwbaarheid α, de fractie $\underline{p}$ van elementen uit de collectie ligt met $\underline{x}$ groter dan $\underline{x}_r$, voldoet weer aan de betrekking

$$\int_{\beta}^{1} g(p)dp = 1-\alpha. \qquad (3.3.2)$$

Met behulp van deze twee betrekkingen kan men weer analoge problemen oplossen als in § 2.

De verdelingsdichtheid g(p) is door S.S. WILKS [10]en [11] op een andere wijze bepaald.

3.4 Nomogrammen en formules.

De oplossingen van de in §3.2 en § 3.3 genoemde problemen worden gevonden met de betrekking

$$\int_{\beta}^{1} g(p)dp = 1-\alpha\,,$$

waarin

$$g(p) = \frac{n!}{(r+s-1)!(n-r-s)!}\,(1-p)^{r+s-1}\,p^{n-r-s}$$

is.

Uit deze twee vergelijkingen volgt direct

$$I_{\beta}(t, n-t+1) = \alpha\,, \qquad (t = r+s)$$

waarin I_{β} de onvolledige B-functie is.

Voor het bepalen van β (bij gegeven waarden van n, t en α)

uit bovenstaande vergelijking, zijn voor de waarden $\alpha = 0{,}005$, 0,1, 0,5, 0,9 en 0,995 nomogrammen berekend door L.E. SIMON [21]. In deze nomogrammen doorloopt n de waarden 1 tot 500 en t de waarden 0 tot 200.
Bovendien kan men de tabellen van K. PEARSON [20] gebruiken.

Voor het bepalen van de waarde van n (bij gegeven α, β en t) is door H. SCHEFFÉ en J.W. TUKEY [4] de volgende benaderingsformule afgeleid

$$n \approx \frac{1}{4}\chi_\alpha^2 \frac{1+\beta}{1-\beta} + \frac{1}{2}(t-1).$$

χ_α^2 wordt hierbij bepaald door de vergelijking

$$P[\chi^2 \geqq \chi_\alpha^2] = \alpha,$$

waarin χ^2 een verdeling met 2t vrijheidsgraden bezit.
De fout, die deze formule geeft , is zeer klein. Bovendien is ze zodanig, dat men nooit een te kleine waarde voor n vindt.

Door R.B. MURPHY [3] is de volgende benaderingsformule gegeven voor het berekenen van β , voor gegeven waarden van n, t en α.

$$\beta \approx \left[\frac{\sqrt{(\chi_\alpha^2 - 2t)^2 + 16n(n-t)} - (\chi_\alpha^2 - 2t)}{4n}\right]^2.$$

χ_α^2 wordt hierbij weer bepaald door bovenstaande betrekking.
Voor het bepalen van de waarden β zijn nomogrammen [19] berekend, voor $\alpha = 0{,}90$, $\alpha = 0{,}95$ en $\alpha = 0{,}99$. Hierbij doorloopt n de waarden 1 tot 500 en t de waarden 1 tot 100.

§4. Een tweede methode voor de afleiding van g(p).

In de voorgaande paragrafen is g(p) door een limiet-overgang bepaald uit de formules voor het geval van twee eindige steekproeven.
We zullen nu een meer directe afleiding geven, waarbij ondersteld wordt, dat de grootheid $\underline{x}$ een verdelingsdichtheid f(x) bezit.

4.1 Een onderste tolerantiegrens.

Wordt de waarde $\underline{x}_r$ uit de steekproef als onderste tolerantiegrens gekozen, dan kan op de volgende wijze de verdelingsdichtheid g(p) bepaald worden van de fractie

$$\underline{p} = \int_{\underline{x}_r}^{\infty} f(x)\, d\, x$$

van elementen uit de gehele collectie met $\underline{x}$ groter dan $\underline{x}_r$.

We maken gebruik van de eigenschap, dat de grootheid

$$\underline{u} = \int_{-\infty}^{\underline{x}} f(x)\ d\ x \qquad (4.1.1)$$

een homogene verdeling bezit tussen 0 en 1, d.w.z. een verdeling van de in figuur 3 geschetste gedaante.

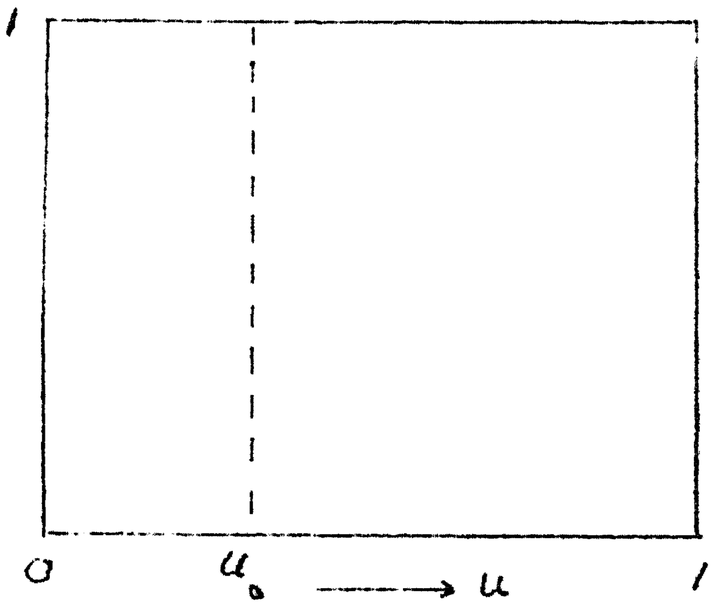


fig. 3

Dit ziet men als volgt in:

Is x_0 de volgens formule (4.1.1) met willekeurige waarde $u_0 (0 \leqslant u_0 \leqslant 1)$ corresponderende waarde van x, dan geldt

$$P\left[\underline{u} \leqslant u_0\right] = P\left[\underline{x} \leqslant x_0\right] = \int_{-\infty}^{x_0} f(x)\ d\ x = u_0,$$

hetgeen te bewijzen was.

Volgens formule (4.1.1) zijn de met de naar opklimmende grootte gerangschikte steekproefwaarden $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ corresponderende waarden $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_n$, eveneens naar opklimmende grootte gerangschikt.

We willen nu eerst de verdelingsdichtheid $h(u_r)$ bepalen van de stochastische grootheid

$$\underline{u}_r = \int_{-\infty}^{\underline{x}_r} f(x)\ d\ x.$$

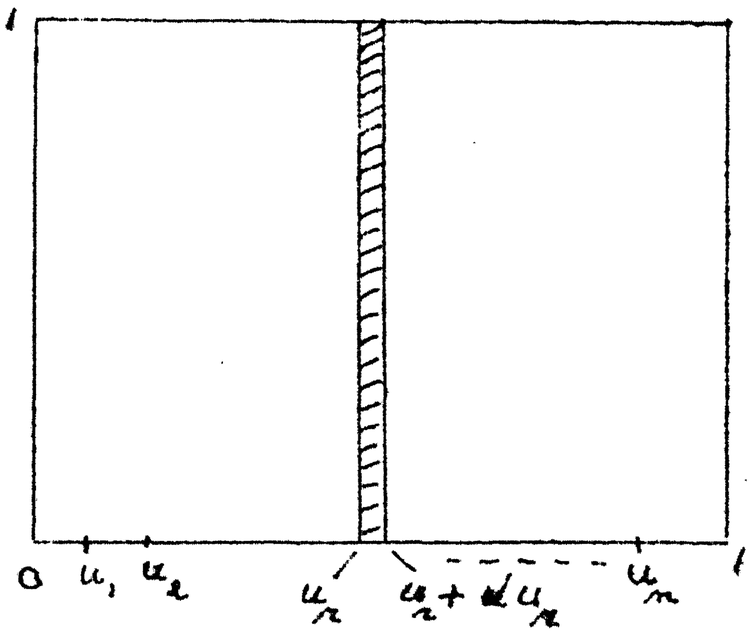


fig. 4

Nu is $h(u_r)du_r = P\left[u_r \leqslant \underline{u}_r \leqslant u_r + du_r\right]$.

Aan de hand van fig. 4 vinden we voor deze laatste wh

$$P\left[u_r \leqslant \underline{u}_r \leqslant u_r + du_r\right] = \frac{n!}{(r-1)!(n-r)!} u_r^{r-1} (1-u_r)^{n-r} du_r;$$

dus is

$$h(u_r) = \frac{n!}{(r-1)!(n-r)!}\, u_r^{\,r-1}(1-u_r)^{n-r}. \qquad (4.1.2)$$

Uit deze verdelingsdichtheid $h(u_r)$ volgt nu de gevraagde verdelingsdichtheid $g(p)$, daar

$$\underline{p} = \int_{\underline{x}_r}^{\infty} f(x)dx = 1 - \int_{-\infty}^{\underline{x}_r} f(x)dx = 1 - \underline{u}_r,$$

zodat

$$g(p) = \frac{n!}{(r-1)!(n-r)!}\,(1-p)^{r-1}\,p^{n-r}.$$

4.2 Een onderste en bovenste tolerantiegrens.

De verdelingsdichtheid $g(p)$ van de fractie $\underline{p}$ van elementen uit de collectie met een $\underline{x}$ tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$ leiden we op de volgende wijze af:

$$\underline{p} = \int_{\underline{x}_r}^{\underline{x}_{n-s+1}} f(x)dx = \underline{u}_{n-s+1} - \underline{u}_r .$$

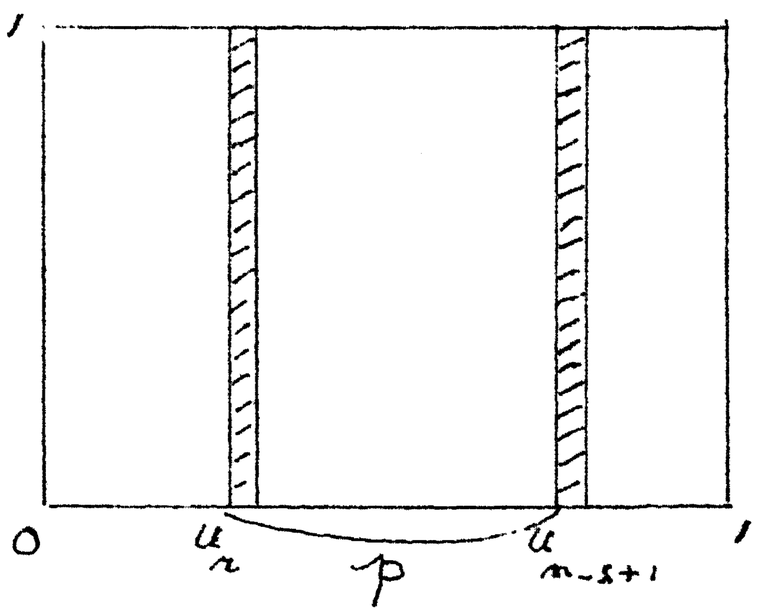


fig. 5

We bepalen eerst de verdelingsdichtheid $g(p \mid \underline{u}_r = u_r)$ van $\underline{p}$, onder de voorwaarde, dat $\underline{u}_r$ tussen u_r en $u_r + du_r$ ligt.
Uit fig. 5 is duidelijk, dat onder de voorwaarde $\underline{u}_r = u_r$, de $(n-r)$elementen $\underline{u}_{r+1}, \ldots, \underline{u}_n$ beschouwd kunnen worden als een steekproef uit de in fig. 6 geschetste homogene verdeling.

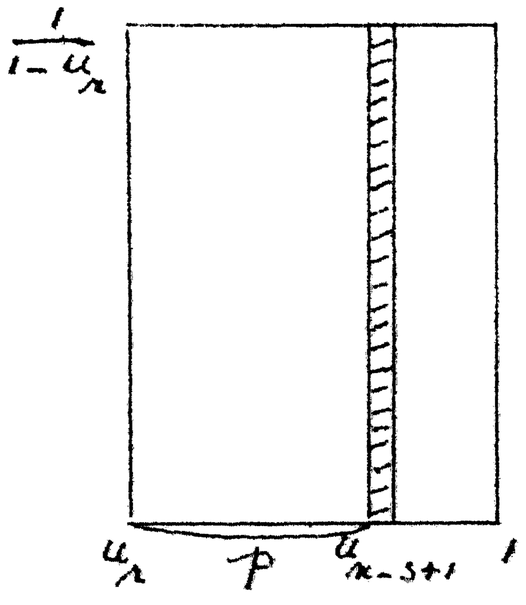


fig. 6

De verdelingsdichtheid $g(p \mid \underline{u}_r = u_r)$ wordt bepaald door

$$g(p \mid \underline{u}_r = u_r)dp = P\left[p \leq \underline{u}_{n-s+1} \leq p + dp\right] =$$

$$= \frac{(n-r)!}{(n-r-s)!(s-1)!} \left(\frac{p}{1-u_r}\right)^{n-r-s} \left(\frac{1-u_r-p}{1-u_r}\right)^{s-1} \frac{dp}{1-u_r}. \qquad (4.2.1)$$

Voor de simultane verdelingsdichtheid $g(p,u_r)$ vinden we, met behulp van (4.1.2) en (4.2.1)

$$g(p,u_r) = g(p|\underline{u}_r = u_r)\cdot h(u_r)$$

$$= \frac{n!}{(r-1)!(n-r-s)!(s-1)!} u_r^{r-1} p^{n-r-s} (1-u_r-p)^{s-1}.$$

Daar de grootheid $\underline{u}_r$ het interval $(0,1-p)$ kan doorlopen (zie fig. 5), vinden we de gevraagde verdelingsdichtheid $g(p)$ door integratie van $g(p,u_r)$ over het interval $0 \leq \underline{u}_r \leq 1-p$. Dus is

$$g(p) = \int_0^{1-p} g(p,u_r)du_r = \frac{n!}{(r-1)!(n-r-s)!(s-1)!} \cdot$$

$$\cdot p^{n-r-s} \int_0^{1-p} u^{r-1}(1-u-p)^{s-1}du.$$

Stellen we

$$u = (1-p)v,$$

dan wordt dit

$$g(p) = \frac{n!}{(r-1)!(n-r-s)!(s-1)!} p^{n-r-s}(1-p)^{r+s-1} \int_0^1 v^{r-1}(1-v)^{s-1}dv.$$

Deze laatste integraal is de bekende B-functie met parameter r en s, waarvoor geldt

$$B(r,s) = \int_0^1 v^{r-1}(1-v)^{s-1}dv = \frac{(r-1)!(s-1)!}{(r+s-1)!} .$$

Dus is

$$g(p) = \frac{n!}{(r+s-1)!(n-r-s)!} (1-p)^{r+s-1} p^{n-r-s}.$$

Opmerking:
Voor discontinue één-dimensionale verdelingen [5] gelden dezelfde formules als hierboven afgeleid, echter is, wanneer als tolerantie-interval het gesloten interval $[\underline{x}_r, x_{n-s+1}]$ genomen wordt, de werkelijke onbetrouwbaarheid kleiner dan α en voor een open interval $(\underline{x}_r, \underline{x}_{n-s+1})$ groter dan α .

§5. Simultane tolerantiegrenzen voor twee onafhankelijke stochastische grootheden.

5.1 Inleiding.

In de vorige paragrafen hebben we steeds problemen beschouwd, waarbij slechts één kenmerk $\underline{x}$ van de producten gemeten werd. In de practijk komen echter ook dikwijls problemen voor, waarbij twee of meer kenmerken van een product van belang zijn (b.v. de diameter en lengte van staven).
In deze paragraaf zullen we nagaan of de theorie voor één sto-

chastische grootheid gemakkelijk gegeneraliseerd kan worden voor twee stochastische onafhankelijke grootheden.
Allereerst zullen wij het probleem beschouwen, waarbij voor ieder der grootheden één enkele onderste tolerantiegrens gekozen wordt. Het zal blijken, dat de bij dit probleem behorende formules van zeer ingewikkelde aard worden.

5.2. Twee onderste tolerantiegrenzen.

Wanneer $\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n$ en $\underline{y}_1,\underline{y}_2,\ldots,\underline{y}_n$ de naar opklimmende grootte van $\underline{x}$ resp. $\underline{y}$ gerangschikte steekproefwaarden zijn voor een steekproef van n artikelen uit de collectie, dan zullen we eerst weer de wh berekenen, dat in een tweede steekproef van de uitgebreidheid N, m_o elementen voorkomen, waarvoor $\underline{x}$ groter is dan $\underline{x}_r$ uit de eerste steekproef en $\underline{y}$ groter is dan $\underline{y}_s$ uit de eerste steekproef.
Stellen we

$$\int_{-\infty}^{\underline{x}_r} f(x)dx = \underline{u}$$

en

$$\int^{\underline{y}_s} g(y)dy = \underline{v}, \qquad (5.2.1)$$

dan bezitten $\underline{u}$ en $\underline{v}$ verdelingsdichtheden k(u) en l(v) met (zie § 4.1):

$$k(u) = \frac{n!}{(r-1)!\,(n-r)!}\, u^{r-1}(1-u)^{n-r}$$

en

$$l(v) = \frac{n!}{(s-1)!(n-s)!}\, v^{s-1}(1-v)^{n-s}.$$

De wh, dat voor een bepaald artikel uit de tweede steekproef de twee waarden $\underline{x}$ en $\underline{y}$ groter zijn, dan twee vaste waarden x_r en y_s uit de eerste steekproef, is, volgens de onderstelde onafhankelijkheid der verdelingen van $\underline{x}$ en $\underline{y}$ gelijk aan (1-u)(1-v), waarbij u en v bepaald worden door de vergelijkingen (5.2.1).
Voor de wh, dat m_o elementen uit de tweede steekproef deze eigenschap bezitten, vinden we dus de binomiale verdeling

$$\binom{N}{m_o}\left[(1-u)(1-v)\right]^{m_o}\left[1-(1-u)(1-v)\right]^{N-m_o}.$$

De wh, dat m_o elementen uit de tweede steekproef groter zijn dan de twee stochastische grootheden $\underline{x}_r$ en $\underline{y}_s$ uit de eerste steekproef, is dus:

$$P[\underline{m}=m_o] = \int_o^1\int_o^1 k(u)l(v)\binom{N}{m_o}\left[(1-u)(1-v)\right]^{m_o}\left[1-(1-u)(1-v)\right]^{N-m_o} du\,dv$$

$$= \frac{(n!)^2}{(r-1)!(n-r)!(s-1)!(n-s)!}\int_o^1\int_o^1 u^{r-1}(1-u)^{n-r+m_o}\cdot v^{s-1}(1-v)^{n-s+m_o}\cdot\left\{1-(1-u)(1-v)\right\}^{N-m_o} du\,dv.$$

Ontwikkelen we $\{1-(1-u)(1-v)\}^{N-m_o}$ naar opklimmende machten van $(1-u)(1-v)$, dan is

$$\{1-(1-u)(1-v)\}^{N-m_o}=\sum_{h=0}^{N-m_o}(-1)^h\binom{N-m_o}{h}(1-u)^h(1-v)^h.$$

Dus is

$$P[\underline{m}=m_o]=\frac{(n!)^2}{(r-1)!(n-r)!(s-1)!(n-s)!}\sum_{h=0}^{N-m_o}(-1)^h\binom{N-m_o}{h}\cdot$$
$$\cdot\iint_{oo}^{11}u^{r-1}(1-u)^{n-r+m_o+h}v^{s-1}(1-v)^{n-s+m_o+h}\,du\,dv.$$

De integralen in het rechter-lid zijn de bekende B-functies, waarvoor het volgende geldt:

$$B(k,l)=\int_o^1 u^{k-1}(1-u)^{l-1}du=\frac{(k-1)!(l-1)!}{(k+l-1)!}$$

Dus is

$$(5.2.2)\qquad P[\underline{m}=m_o]=\frac{(n!)^2}{(r-1)!(n-r)!(s-1)!(n-s)!}\sum_{h=0}^{N-m_o}(-1)^h\binom{N-m_o}{h}\cdot$$
$$\cdot\frac{(r-1)!(n-r+m_o+h)!(s-1)!(n-s+m_o+h)!}{(n+m_o+h)!\,(n+m_o+h)!}$$

$$P[\underline{m}=m_o]=\frac{(n!)^2}{(n-r)!(n-s)!}\sum_{h=0}^{N-m_o}(-1)^h\binom{N-m_o}{h}\frac{(n-r+m_o+h)!(n-s+m_o+h)!}{(n+m_o+h)!^2}$$

Met behulp van bovenstaande formule (5.2.2) en de relatie

$$\sum_{m_o=N_\alpha}^{N}P[\underline{m}=m_o]\geq 1-\alpha,$$

waarin N_α de grootste gehele waarde is, waarvoor de betrekking nog geldt, kunnen in principe problemen opgelost worden, analoog aan de in de voorafgaande paragrafen geformuleerde.

Voor practische toepassingen is formule (5.2.2) echter veel te ingewikkeld. Deze formules worden nog ingewikkelder, wanneer men in plaats van twee onderste tolerantiegrenzen een combinatie van onderste en bovenste tolerantiegrenzen neemt voor de beide grootheden $\underline{x}$ en $\underline{y}$. Voor meer dan twee stochastische grootheden geldt hetzelfde.

In de volgende paragraaf zullen we daarom een eenvoudiger methode geven voor het bepalen van simultane tolerantiegrenzen voor k stochastische grootheden $\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_k$. Deze methode geldt echter alleen, wanneer de tweede steekproef zeer groot is en de k stochastische grootheden (dat zijn de beschouwde k kenmerken van één artikel) een continue simultane verdelingsfunctie bezitten.

De methode kan ook toegepast worden, wanneer de grootheden gecorreleerd zijn.

§6. Simultane tolerantiegrenzen voor k stochastische grootheden met een continue simultane verdelingsfunctie.

6.1 Inleiding.

De in § 4 besproken methode voor het bepalen van tolerantiegrenzen (in deze paragraaf werd de tweede steekproef oneindig groot ondersteld) is door A. WALD [9] gegeneraliseerd voor k al of niet onafhankelijke stochastische grootheden $\underline{x}_1$, $\underline{x}_2$, . . . , $\underline{x}_k$. Deze k grootheden stellen steeds k kenmerken van één artikel voor. Voor een steekproef van de uitgebreidheid n uit een collectie vinden we dus voor ieder der n elementen k getallen.
We zullen de door Wald ontwikkelde methode bespreken voor twee kenmerken $\underline{x}$ en $\underline{y}$, zonder de daarbij behorende ingewikkelde bewijzen te geven.
De methode voor meer dan twee kenmerken is geheel analoog.

6.2 Simultane tolerantiegrenzen voor 2 stochastische grootheden.

Zijn $(\underline{x}_1, \underline{y}_1)$, $(\underline{x}_2, \underline{y}_2)$, . . . , $(\underline{x}_n, \underline{y}_n)$ de gevonden waarden van $\underline{x}$ en $\underline{y}$ voor een steekproef van de uitgebreidheid n, waarbij de waarnemingen zo gerangschikt zijn, dat $\underline{x}_1 < \underline{x}_2 \; . \; . < \underline{x}_n$, dan kunnen we op volgende wijze een tolerantiegebied $\underline{T}$ bepalen voor het stochastische punt $(\underline{x}, \underline{y})$ op de collectie, waaruit de steekproef afkomstig is.
Door de twee punten met abscis $\underline{x}_{r_1}$ en $\underline{x}_{n-s_1+1}$ trekken we twee verticale lijnen $\underline{V}_{r_1}$ en $\underline{V}_{n-s_1+1}$. Hierbij moet voldaan zijn aan $r_1 + s_1 \leqslant n - 2$, zodat er nog minstens twee punten tussen de beide lijnen liggen. De waarden r_1 en s_1 worden van tevoren (onafhankelijk van de puntenwolk) vastgesteld. De wh, dat twee elementen uit de steekproef dezelfde $\underline{x}$ of dezelfde $\underline{y}$ bezitten, is nul, daar de twee-dimensionale verdeling continu is ondersteld.
Vervolgens beschouwen we de verzameling $\underline{S}$ van punten, welke binnen deze twee verticale lijnen liggen.
Noemen we de naar opklimmende grootte gerangschikte waarden van $\underline{y}$ uit deze verzameling $\underline{y}_1'$, $\underline{y}_2'$, . . . , $\underline{y}_{n-(r_1+s_1)}'$, dan trekken we door de twee punten met ordinaat $\underline{y}_{r_2}'$ en $\underline{y}'_{n-(r_1+s_1)-s_2+1}$ twee horizontale lijnen $\underline{H}'_{r_2}$ en $\underline{H}'_{n-(r_1+s_1)-s_2+1}$. De waarden r_2 en s_2 zijn weer van tevoren vastgesteld.
Het tolerantiegebied $\underline{T}$ is nu de rechthoek, bepaald door deze 4 lijnen.

Voorbeeld:

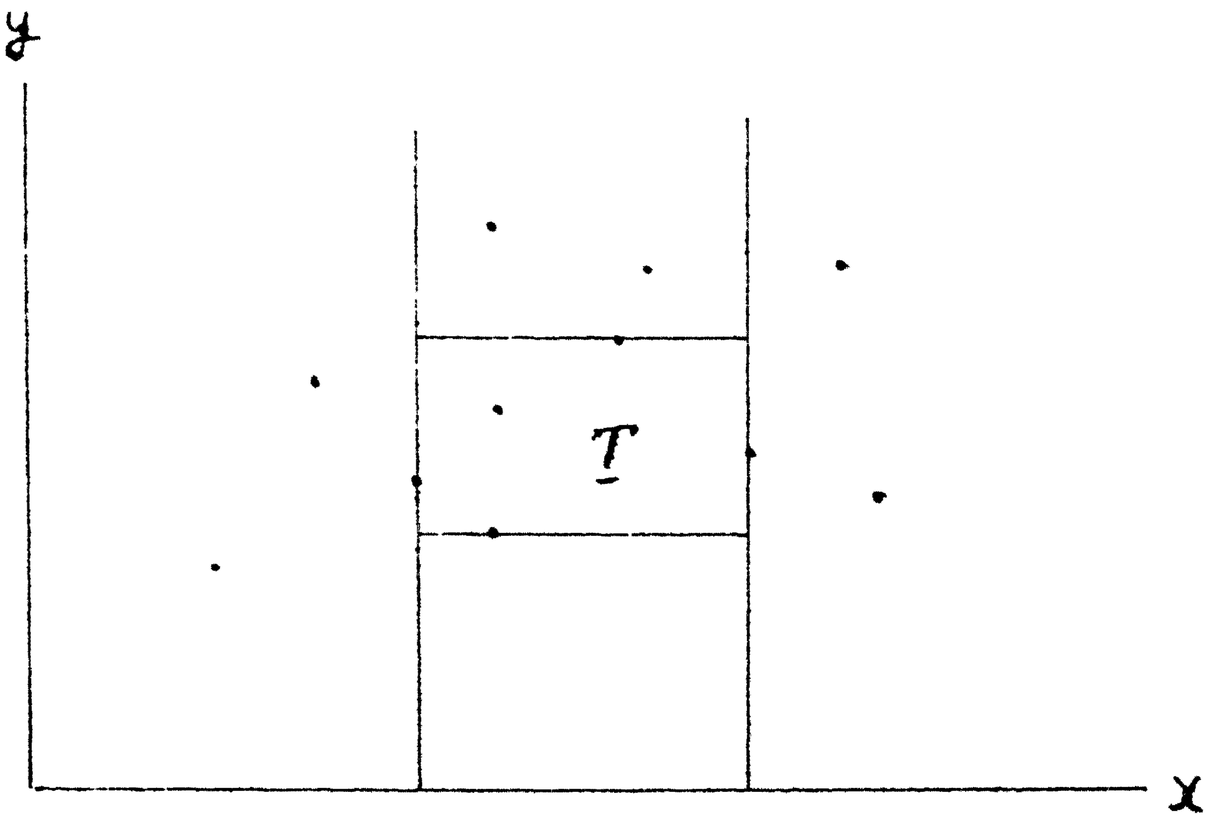


fig. 7

n = 11
r_1 = 3
s_1 = 3
r_2 = 1
s_2 = 3

Wald bewijst nu, dat de verdelingsdichtheid g(p) van de stochastische fractie p van de gehele, oneindig groot onderstelde collectie, waarvoor (x,y) binnen T ligt, gegeven wordt door

$$g(p) = \frac{n!}{(n-r_1-s_1-r_2-s_2)!(r_1+s_1+r_2+s_2-1)!}\, p^{n-r_1-s_1-r_2-s_2} \cdot$$

$$\cdot (1-p)^{r_1+s_1+r_2+s_2-1}. \qquad (6.2.1)$$

De grootheden x en y behoeven hiertoe niet onafhankelijk verdeeld te zijn.

Uit deze formule blijkt, dat g(p) alleen afhankelijk is van de uitgebreidheid n der steekproef en het aantal $t = r_1 + s_1 + r_2 + s_2$ der punten van de steekproef, die niet binnen T liggen. De verdelingsdichtheid g(p) is dus formeel dezelfde als die, afgeleid in § 3 en § 4 voor één stochastische grootheid.

Met behulp van formule (6.2.1) en de betrekking

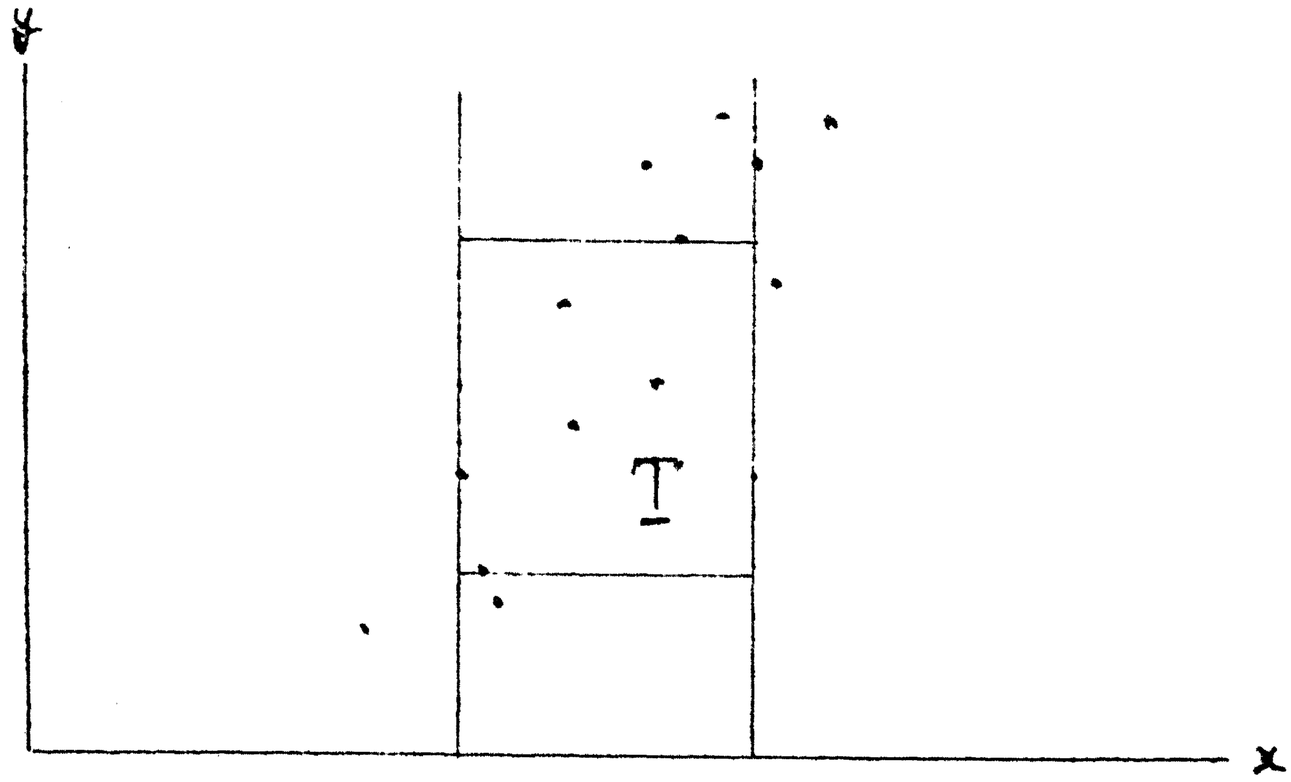
$$\int_{\beta}^{1} g(p)dp = 1 - \alpha, \qquad (6.2.2)$$

kunnen weer analoge problemen opgelost worden als die, geformuleerd voor één stochastische grootheid.

De keuze van de waarden r_1, s_1, r_2, s_2 moet weer onafhankelijk van de gevonden steekproefwaarden geschieden. Door middel van deze keuze tracht men dan, op grond van vermoedens omtrent de verdeling van x en y, het gebied T zo klein mogelijk te maken.

6.3 Twee gecorreleerde stochastische grootheden.

Wanneer de stochastische grootheden x en y sterk gecorreleerd zijn, is het nadeel van de vorige methode, dat het tolerantiegebied T zeer groot wordt, zoals b.v. uit fig. 8 blijkt.



$n = 13$
$r_1 = 2$
$s_1 = 3$
$r_2 = 2$
$s_2 = 3$

fig. 8

In In dit geval kunnen we echter een kleiner tolerantiegebied $\underline{T}$ construeren, door de volgende methode toe te passen:

De punten worden eerst gerangschikt naar opklimmende waarden van $\underline{x}$. Door een van te voren vastgesteld aantal k van deze punten worden vervolgens verticale lijnen getrokken, waardoor het vlak van de puntenwolk in k+1 stroken verdeeld wordt.
De k punten, waardoor de verticale lijnen getrokken worden, worden weer gekozen, door van te voren k rangnummers vast te stellen en die punten te nemen, die in de gerangschikte rij deze rangnummers dragen.
In ieder van de k+1 verticale stroken ligt nu een van te voren bekend aantal punten van de puntenwolk. De punten, welke op de randen van de stroken liggen worden buiten beschouwing gelaten.
Door de in de stroken punten worden vervolgens horizontale lijnen getrokken, zodat iedere strook verdeeld wordt in een van te voren bekend aantal blokken. Deze blokken denken we van boven naar beneden genummerd (zie fig. 9).

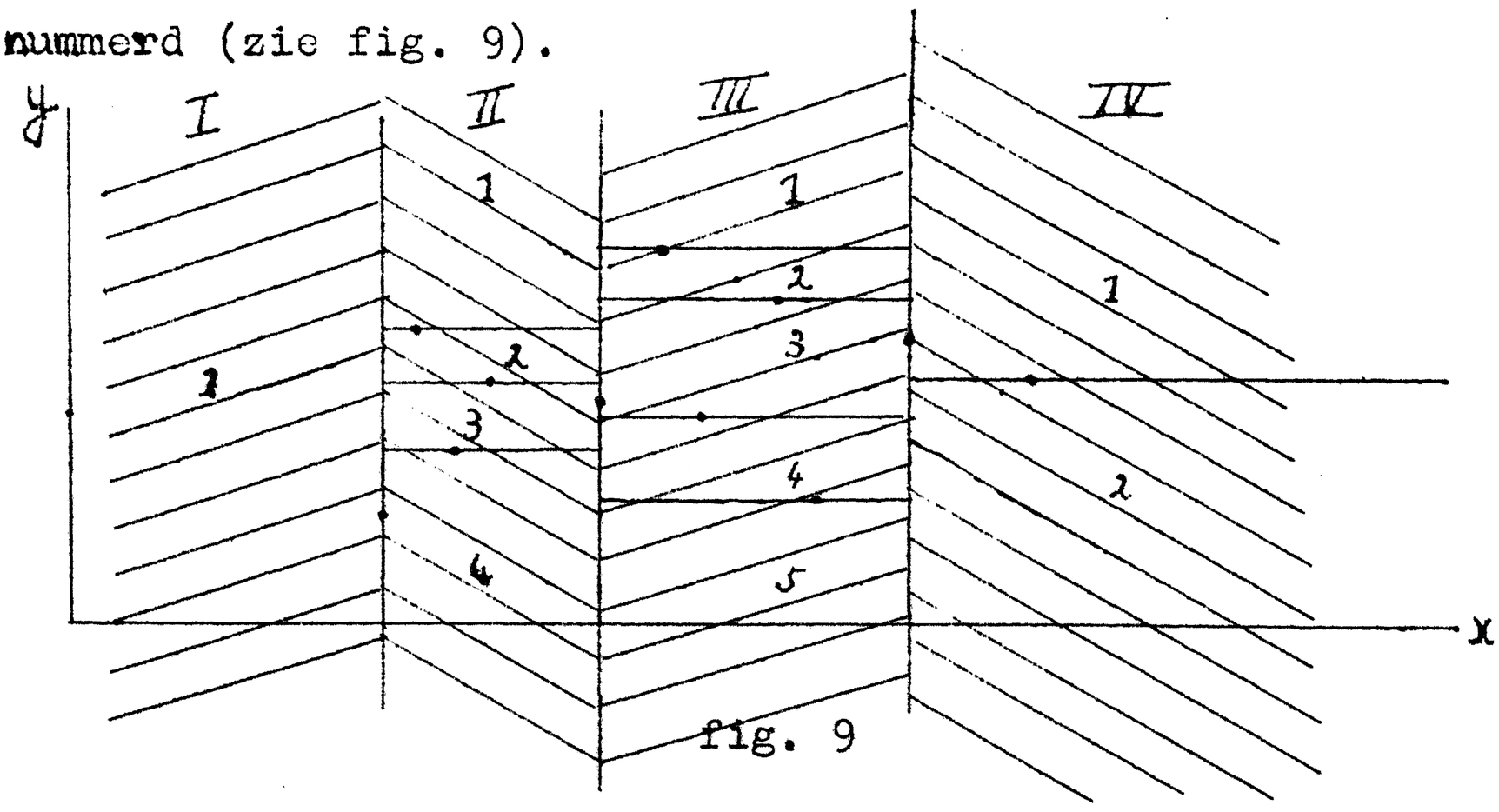


fig. 9

Elimineren we nu uit iedere strook een aantal blokken met een van te voren vastgesteld rangnummer en is t het totale aantal geëlimineerde blokken in het vlak van de puntenwolk, dan bezit, wanneer we het resterende gebied $\underline{T}$ als tolerantiegebied beschouwen, de

fractie p uit de collectie met het "punt" (x,y) in T, de verdelingsdichtheid

$$g(p) = \frac{n!}{(n-t)!(t-1)!}\, p^{n-t}(1-p)^{t-1}. \qquad (6.3.1)$$

Voorbeeld:

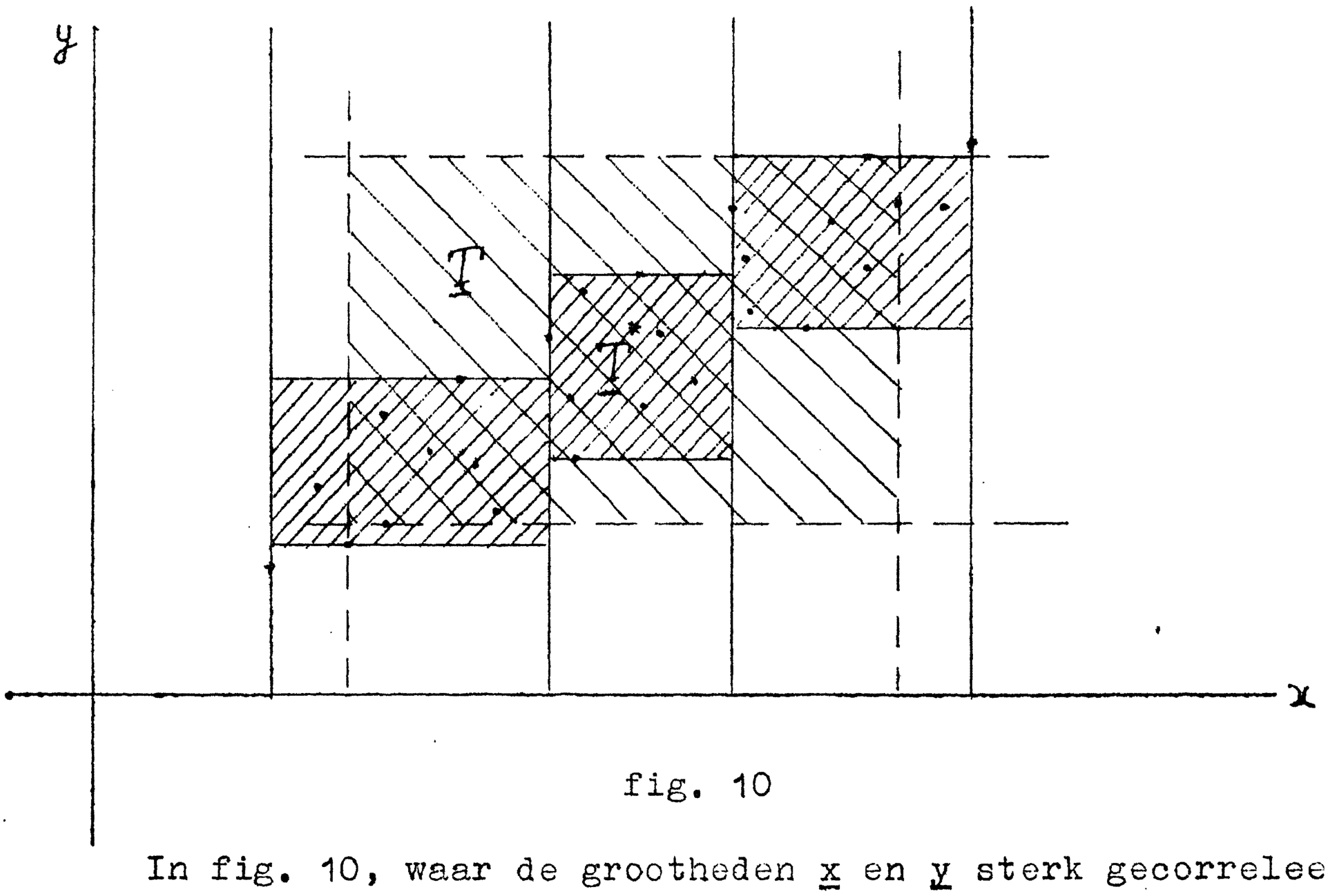


fig. 10

In fig. 10, waar de grootheden x en y sterk gecorreleerd zijn, zijn, volgens de twee zojuist besproken methoden, tolerantiegebieden T en T^* geconstrueerd. Het gebied T is het door de gestippelde rechthoek aangegeven gebied, T^* bestaat uit drie gearceerde gebieden. Het tolerantiegebied T^* is hier bepaald, door in iedere verticale strook de onderste en bovenste, aan één zijde onbegrensde blokken, te elimineren. Deze constructie geeft bij sterk gecorreleerde grootheden bevredigende resultaten. Indien het aantal punten groot genoeg is, kunnen aan onder- en bovenzijde ook meerdere blokken weggelaten worden of men kan het aantal stroken vergroten.

De bij deze twee gebieden behorende verdelingsdichtheden g(p) en $g^*(p)$ zijn nu dezelfde, daar de waarden n en t, welke de formules bepalen (zie (6.2.1) en (6.3.1)), voor de twee gebieden dezelfde zijn. Het gebied T^* is echter veel kleiner dan het gebied T.
Voor T geldt:

$$\left.\begin{aligned} n &= 27 \\ r_1 &= 3 \\ s_1 &= 3 \\ r_2 &= 1 \\ s_2 &= 1 \end{aligned}\right\} \quad \text{dus is } t = r_1+s_1+r_2+s_2 = 8$$

Voor $\underline{T}^*$ geldt:

$n = 27$
$k = 4$
$t_1 = 1$
$t_2 = 2$
$t_3 = 2$
$t_4 = 2$
$t_5 = 1$

Dus is $t = t_1+t_2+t_3+t_4+t_5 = 8$.

§7. De algemene theorie.

7.1 Inleiding.

De in de vorige paragraaf aangegeven methoden voor het construeren van tolerantiegebieden, zijn tenslotte nog gegeneraliseerd, in die zin, dat men in plaats van horizontale en verticale lijnen, van te voren gedefiniëerde continuekrommen door de "punten" van de steekproef legt. Voor deze continue krommen gebruikt men in de practijk meestal rechte lijnen met een van te voren gedefiniëerde helling.

Een eerste uitbreiding van de theorie heeft J.W. TUKEY [7] gegeven en is analoog aan de in § 6.2 behandelde. Van deze methode is in fig. 13 een voorbeeld gegeven.

De door Wald aangegeven generalisatie van de vorige paragraaf, bestaande uit de verdeling in stroken, is ook hier weer van toepassing (zie fig. 14).

Een overzicht van de beide methoden geeft R.B. MURPHY [3]. We zullen deze methoden in het kort uiteenzetten voor een tweedimensionale verdeling.

7.2 Het construeren van tolerantiegebieden met behulp van continue krommen.

Onderstellen we, dat de steekproef bestaat uit de n waarnemingen $(\underline{x}_1, \underline{y}_1)$, $(\underline{x}_2, \underline{y}_2)$, . . . , $(\underline{x}_n, \underline{y}_n)$ in het (x,y)-vlak. Verder zijn n <u>continue</u> functies $f_1(x,y)$, $f_2(x,y)$ ..., $f_n(x,y)$ gegeven speciale gedaante van deze functies wordt, zoals verderop zal blijken, weer bepaald door het vermoeden omtrent de aard van de verdeling van $\underline{x}$ en $\underline{y}$, maar de keuze van deze functies is onafhankelijk van de in de steekproef gevonden waarden. Ondersteld wordt bovendien nog, dat de simultane verdeling van $\underline{x}$ en $\underline{y}$ continu is. Ook hier behoeven $\underline{x}$ en $\underline{y}$ niet onafhankelijk verdeeld te zijn.

We beschouwen allereerst de functie $f_1(x,y)$. Deze functie neemt in de n verschillende punten $(\underline{x}_1, \underline{y}_1)$, $(\underline{x}_2, \underline{y}_2)$, . . . , $(\underline{x}_n, \underline{y}_n)$ n verschillende waarden aan. (De wh, dat twee waarden

gelijk zijn, is weer nul, daar de n functies f_1, f_2, . . . , f_n en de verdeling van $\underline{x}$ en $\underline{y}$ continu ondersteld zijn). In één van de punten zal de functie $f_1(x,y)$ een kleinste waarde aannemen. Onderstellen we, dat dit geschiedt in het punt $(\underline{x}_{i_1}, \underline{y}_{i_1})$. Noemen we deze kleinste waarde $\underline{w}_1$, dan geldt dus:

$$f_1(\underline{x}_{i_1}, \underline{y}_{i_1}) = \underline{w}_1$$

$$f_1(\underline{x}_j, \underline{y}_j) > \underline{w}_1, \text{ voor } j = 1,2,\ldots,i_1-1, i_1+1,\ldots,n. \qquad (7.2.1)$$

Door het punt $(\underline{x}_{i_1}, \underline{y}_{i_1})$ leggen we nu de <u>continu kromme</u> $f_1(x,y) = \underline{w}_1$ (zie fig. 11).

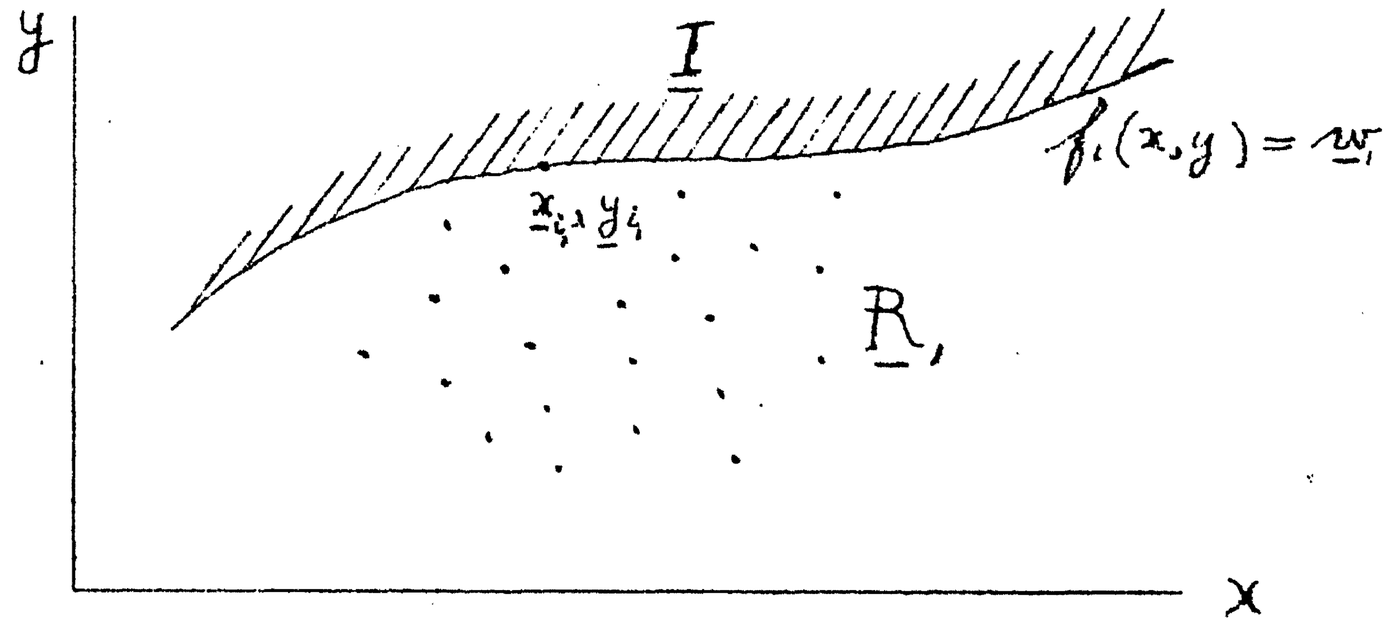


fig. 11

Daar de kromme $f_1(x,y) = \underline{w}_1$ een continue kromme is, geldt voor alle punten (x,y) aan één kant van de kromme, dat $f_1(x,y) > \underline{w}_1$ is en voor alle punten aan de andere kant van de kromme, dat $f_1(x,y) < \underline{w}_1$ is. Wegens relatie (7.2.1) liggen dus de overblijvende n-1 punten van de steekproef aan eenzelfde kant van de kromme.
Het in fig. 11 gearceerde gebied, noemen we nu het door de eerste functie bepaalde "blok", of ook wel het "eerste blok". De rest van het gebied geven we aan door $\underline{R}_1$.

Met behulp van de functie $f_2(x,y)$ kan op dezelfde wijze een kromme $f_2(x,y) = \underline{w}_2$ bepaald worden, door de minimale waarde $\underline{w}_2$ te berekenen, die de functie $f_2(x,y)$ in de (n-1) overblijvende punten aanneemt. Deze kromme gaat weer door het punt $(\underline{x}_{i_2}, \underline{y}_{i_2})$, waar $f_2(x,y)$ de minimale waarde $\underline{w}_2$ bezit. Het deel van het gebied $\underline{R}_1$, waar $f_2(x,y) < \underline{w}_2$, noemen we het "tweede blok". We krijgen dan b.v. de in fig, 12 geschtste situatie, waarbij het restant van het vlak nu $\underline{R}_2$ wordt genoemd.

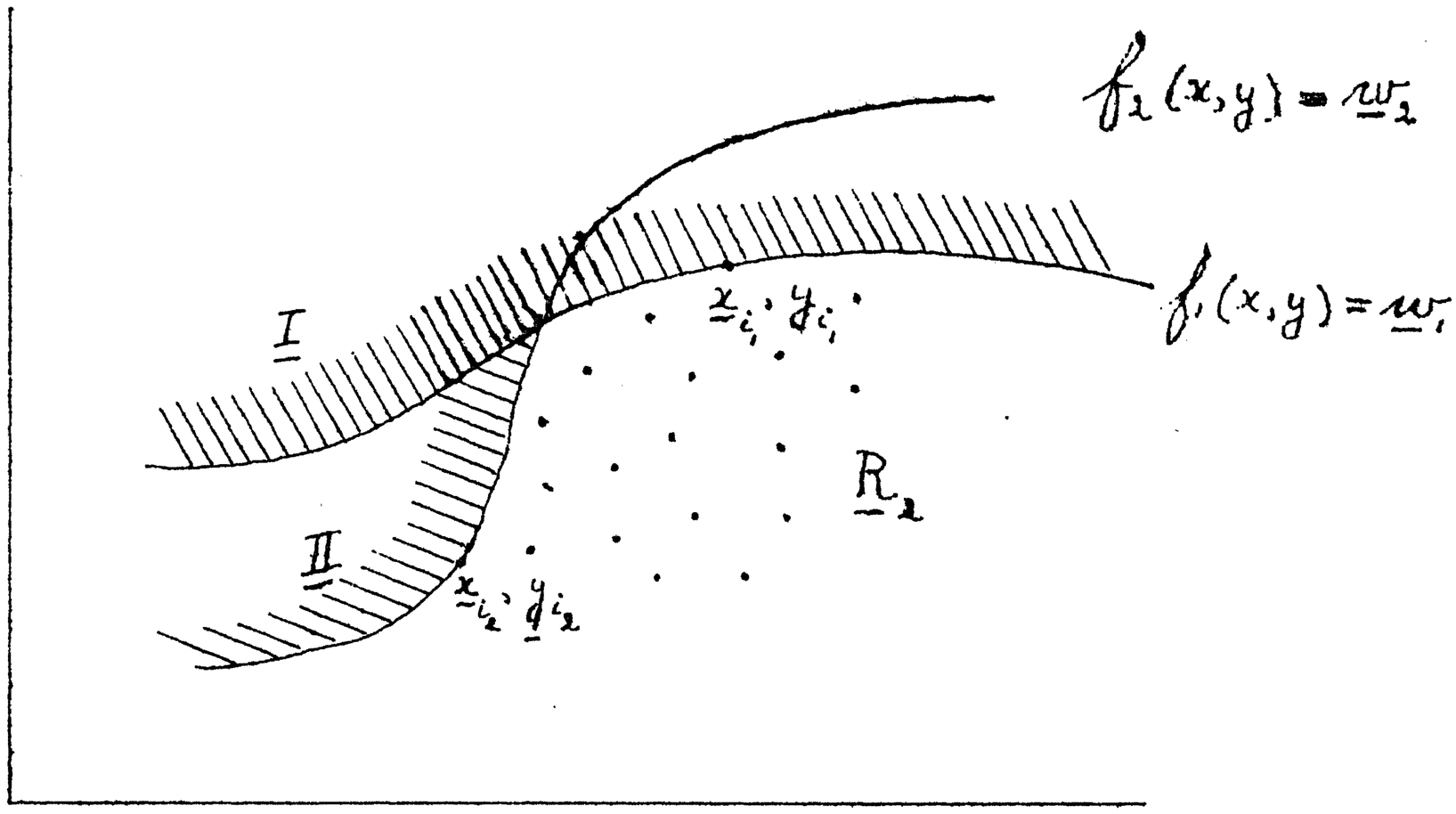


fig. 12

In het gebied $\underline{R}_2$ kan nu weer op dezelfde wijze een nieuwe kromme bepaald worden met behulp van de functie $f_3(x,y)$. Door deze kromme wordt dan in het gebied $\underline{R}_2$ een "derde blok" bepaald en een nieuw resterend gebied $\underline{R}_3$.
Bij de laatste stap hebben we nog de functie $f_n(x,y)$ en het resterende punt $(\underline{x}i_n, \underline{y}i_n)$ tot onze beschikking. Noemen we de waarde van $f_n(x,y)$ in dit punt $\underline{w}_n$, dan trekken we tenslotte door dit overblijvende punt de kromme $f_n(x,y) = \underline{w}_n$, welke kromme het n^e "blok" bepaalt en een resterend gebied $\underline{R}_n$. Dit laatste gebied noemen we nu het $(n+1)^e$ "blok".

We hebben nu tenslotte bereikt, dat met behulp van de n functies het vlak van de puntenwolk in (n+1) "blokken" is verdeeld, die onderling geen punten gemeen hebben.
Laten we nu t van deze n+1 "blokken" uit het vlak weg, waarbij de rangnummers der weggelaten "blokken" onafhankelijk zijn van de steekproef (b.v. van te voren zijn vastgelegd), dan is het restant van het vlak een tolerantiegebied $\underline{T}$, waarvoor geldt dat

$$g(p) = \frac{n!}{(n-t)!(t-1)!}\, p^{n-t}\, (1-p)^{t-1}, \qquad (7.2.2)$$

wanneer $\underline{p}$ de fractie van de collectie is, die in $\underline{T}$ ligt.

Het ligt het meest voor de hand om de eerste t blokken weg te laten, zodat men de overige niet behoeft te construeren. Daarbij kan men dan de functies $f_1, f_2, \ldots, f_t$ zodanig trachten te kiezen (b.v. op grond van vroegere ervaring omtrent de aard van de collectie), dat $\underline{T}$ een gunstige vorm krijgt.

Voorbeeld:

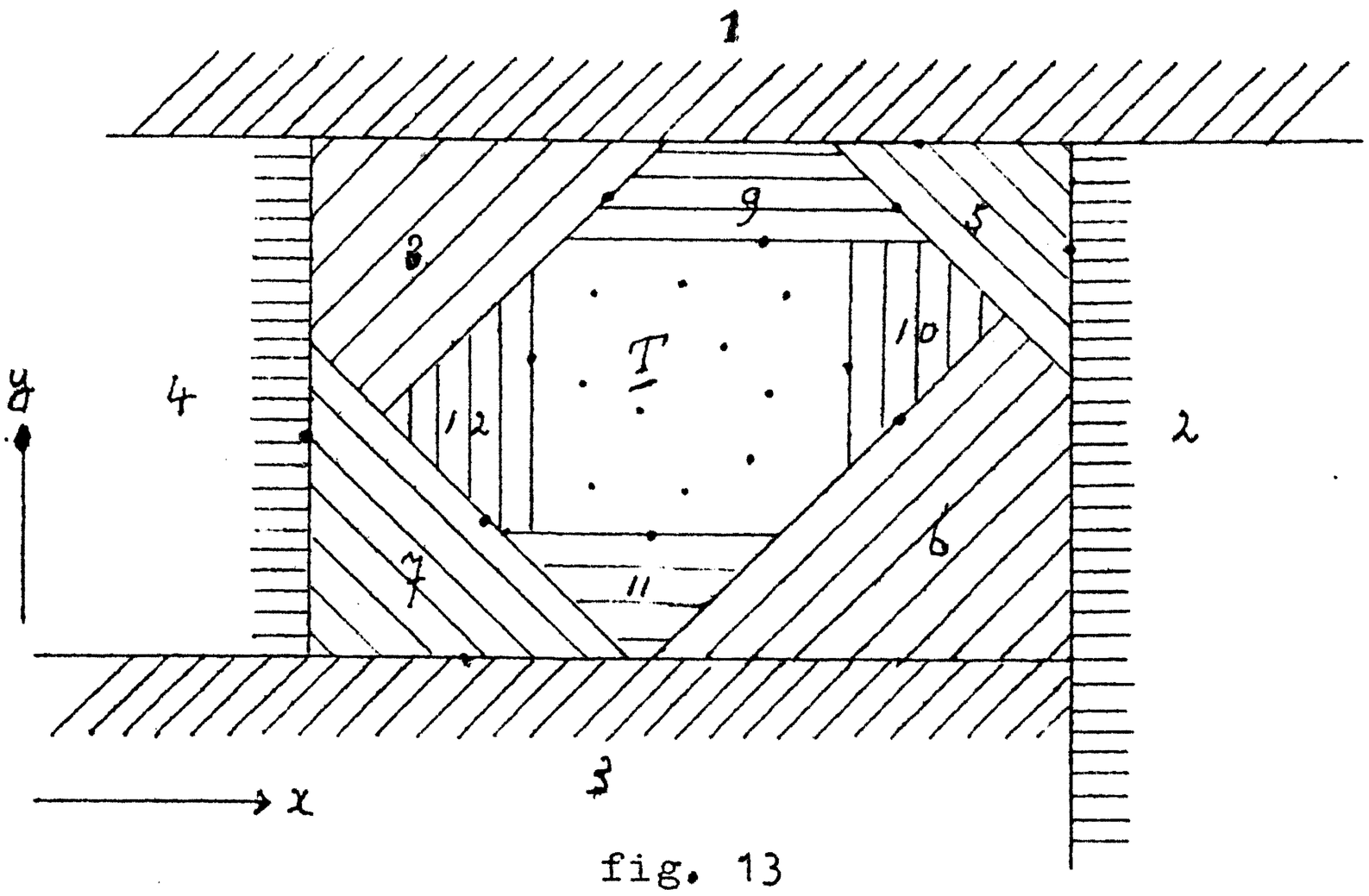


fig. 13

In fig. 13 zijn de blokken in de aangegeven volgorde bepaald door de functies y, x, -y, -x, x+y, x-y, -x-y, -x+y, y, x, -y, -x.

Wald heeft bovenstaande constructie uitgebreid, door eerst de puntenwolk in een aantal stroken te verdelen. Deze indeling in stroken geschiedt met behulp van een van tevoren gedefiniëerde continue functie f(x,y) en k "punten" $(\underline{x}_{a_1}, \underline{y}_{a_1})$, $(\underline{x}_{a_2}, \underline{y}_{a_2})$, ..., $(\underline{x}_{a_k}, \underline{y}_{a_k})$, waarvan de keuze op een van tevoren vastgestelde wijze met behulp van f(x,y) bepaald wordt. Neemt de functie f(x,y) in deze punten resp. de waarden $\underline{w}_{a_1}, \underline{w}_{a_2}, \ldots, \underline{w}_{a_k}$ aan, dan wordt het vlak van de puntenwolk door de k krommen

$$f(x,y) = \underline{w}_{a_j}, \quad (j = 1,2,\ldots,k)$$

in (k+1) stroken verdeeld.
In ieder van deze stroken wordt daarna, op de zojuist beschreven wijze (zie b.v. fig. 13) een tolerantiegebied geconstrueerd.
Onderstellen we, dat in de j^e strook n_j punten liggen, dan geldt voor deze aantallen n_j de betrekking

$$\sum_{j=1}^{k+1} n_j = n - k.$$

Elimineren we nu in de j^e strook, met behulp van t_j van tevoren gedefiniëerde functies, t_j "strookblokken", en noemen we het overblijvende gebied $\underline{T}_j$, dan beschouwen we de verzameling van al deze gebieden $\underline{T}_j (j = 1,2,\ldots,k+1)$ als het tolerantiegebied $\underline{T}$.
Dit tolerantiegebied $\underline{T}$ is dus ontstaan door eliminatie van in totaal $t = t_1 + t_2 + \ldots + t_{k+1}$ "strookblokken" en bestaat

zelf uit

$$(n_1+1-t_1) + (n_2+1-t_2) + \ldots + (n_{k+1} + 1-t_{k+1}) =$$

$$\sum_{j=1}^{k+1} n_j + (k+1) - \sum_{j=1}^{k+1} t_j = (n-k) + (k+1) - t = n+1-t$$

"strookblokken".

Deze "strookblokken" zijn dan aequivalent met de door Tukey geconstrueerde "blokken".

De fractie p van de oneindig groot onderstelde collectie, met een (x,y) in het aldus gedefiniëerde tolerantiegebied T, bezit weer de verdelingsdichtheid (7.2.2).

Voorbeeld:

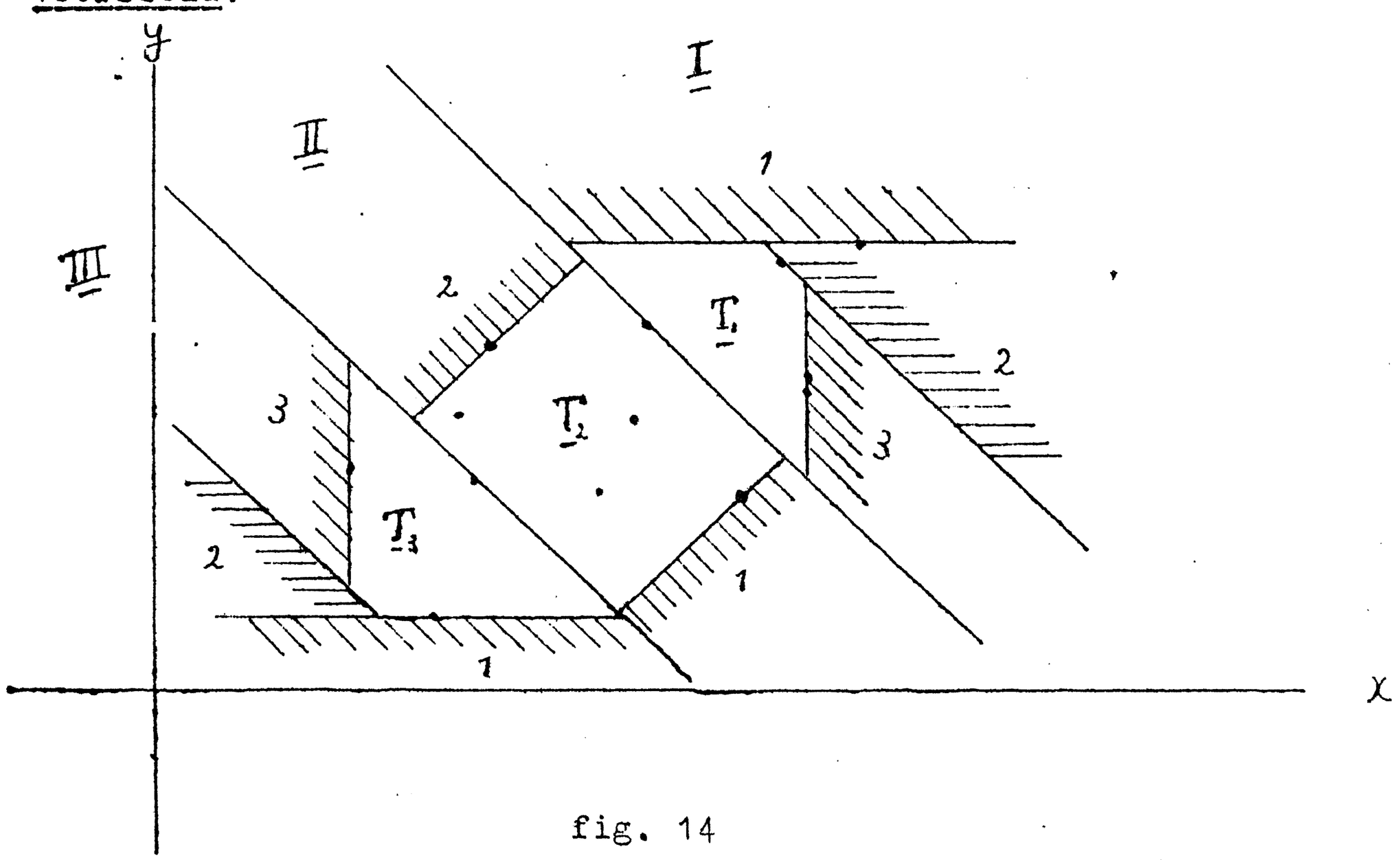


fig. 14

$f(x,y) = x + y$

$k = 2$

De drie stroken I, II en III zijn bepaald door de functie $f(x,y) = x + y$ en de twee punten van de steekproef, waar deze functie de op 4 na kleinste en op 4 na grootste waarde bezit.

In strook I is het tolerantiegebied T_1 bepaald door de functies $-y$, $-x-y$ en $-x$,

T_2 is bepaald door de functies $y-x$ en $x-y$,

T_3 is bepaald door de functies $+y$, $y+x$ en x.

Opmerking:

Tot deze algemene constructie kunnen alle in de voorgaande paragrafen behandelde constructies van tolerantiegebieden teruggebracht worden.

Voorbeeld:

In het één-dimensionale geval werd uit een steekproef $\underline{x}_1$, $\underline{x}_2$, . . . , $\underline{x}_n$ een tolerantie-interval $\left[\underline{x}_r, \underline{x}_{n-s+1}\right]$ bepaald. Stellen we r + s = t, dan bezit de fractie $\underline{p}$ van de collectie met een $\underline{x}$ tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$ precies dezelfde verdelingsdichtheid (7.2.2).

Dat is nu ook direct duidelijk, daar de n punten uit de steekproef in totaal n + 1 intervallen definiëren. Deze intervallen kunnen we opvatten als één-dimensionale "blokken", waarvan er t geëlimineerd worden.

Tenslotte valt nog op te merken, dat men de keuze van de methode van tevoren dient te bepalen, zonder daarbij de vorm van de gevonden puntenwolk in aanmerking te nemen. Dit is uiteraard een groot nadeel, vooral indien men van tevoren over deze vorm niets kan zeggen. Bovendien zal men vaak geneigd zijn tegen dit voorschrift te zondigen. Een methode ,waarbij de vorm van $\underline{T}$ automatisch aan die van de puntenwolk wordt aangepast, zou dus superieur aan de beschreven zijn.

Een eerste aanloop tot een dergelijke methode is gegeven door D.A.S. FRASER [1] .

Voor discontinue meer-dimensionale verdelingen is nog geen volledige oplossing gegeven voor het construeren van tolerantiegebieden. Enkele beschouwingen hieromtrent zijn gegeven door J.W. TUKEY [8] en D.A.S. FRASER en R. WORMLEIGHTON [2] .

§ 8. Niet-stochastische tolerantiegrenzen.

8.1 Inleiding.

In deze paragraaf zullen we het in § 1, onder punt A genoemde probleem behandelen. Dit was in het kort het volgende.

Van een grote partij goederen wordt een zeker kenmerk $\underline{x}$ gecontroleerd. Wanneer de $\underline{x}$ van een product tussen twee vaste grenzen L_1 en L_2 ligt, wordt het product goedgekeurd. Ligt $\underline{x}$ buiten deze grenzen, dan wordt het product afgekeurd.

Wat we nu willen weten, is de werkelijke fractie p_o van de collectie met een $\underline{x}$ tussen L_1 en L_2.

Voor deze onbekende fractie p_o kunnen we nu, door middel van een steekproef, een betrouwbaarheidsinterval $\left[\underline{p}_1, \underline{p}_2\right]$ bepalen. In dit interval ligt dan de werkelijke fractie p_o, behoudens een onbetrouwbaarheid α .

Alvorens we dit probleem nader uitwerken, zullen we eerst aangeven, op welke wijze betrouwbaarheidsintervallen in het algemeen bepaald worden.

Onderstellen we, dat een stochastische grootheid $\underline{x}$ een wh-verdeling bezit met een onbekende parameter Θ. Is de werkelijke waarde van deze parameter Θ_o, en we toetsen met behulp van de gevonden steekproef de hypothese H, dat $\Theta=\Theta_o$ is, dan zal, indien een kritieke zône gedefiniëerd is met onbetrouwbaarheidsdrempel α, deze steekproef, behoudens een onbetrouwbaarheid α, niet tot verwerping leiden van deze hypothese.
Beschouwen we nu de verzameling van alle mogelijke waarden van Θ, welke door deze toets en op grond van de steekproef niet verworpen worden, dan behoort derhalve de onbekende, werkelijke parameterwaarde Θ_o, behoudens een onbetrouwbaarheid α, tot de-deze verzameling. Voor een enkele stochastische grootheid $\underline{x}$, vormen deze waarden gewoonlijk een interval $[\underline{\Theta}_1, \underline{\Theta}_2]$. Een dergelijk interval heet een betrouwbaarheidsinterval. De onbekende, werkelijke parameterwaarde Θ_o ligt, behoudens een onbetrouwbaarheid α, in dit interval.

8.2. Een betrouwbaarheidsinterval voor de fractie p_o.

Onderstellen we, dat van de steekproef $x_1, x_2, \ldots, x_n$, r waarden in het interval (L_1, L_2) liggen. We vinden dan op de volgende wijze een betrouwbaarheidsinterval voor de onbekende fractie p_o van de collectie met een $\underline{x}$ tussen L_1 en L_2.
We toetsen de hypothese H, dat de onbekende fractie van de collectie met een $\underline{x}$ tussen L_1 en L_2 zekere waarde p bezit. Onder deze hypothese bezit het aantal $\underline{r}$ van elementen met een $\underline{x}$ tussen L_1 en L_2 de binomiale verdeling

$$P[\underline{r}=r|n,p]=\binom{n}{r}p^r(1-p)^{n-r}. \qquad (8.2.1)$$

Voor het toetsen van de hypothese H gebruiken we een tweezijdige kritieke zone met onbetrouwbaarheidsdrempel α.
Het gevraagde betrouwbaarheidsinterval voor de onbekende, werkelijke fractie p_o, wordt nu gevormd door al die waarden p, welke niet verworpen worden op grond van de gevonden steekproef.
Voor deze waarden p geldt

$$\sum_{k=r}^{n}\binom{n}{k}p^k(1-p)^{n-k} > \tfrac{1}{2}\alpha$$

en

$$\sum_{k=0}^{r}\binom{n}{k}p^k(1-p)^{n-k} > \tfrac{1}{2}\alpha,$$

indien r de bij de steekproef door $\underline{r}$ aangenomen waarde voorstelt.

Nu geldt:

$$\sum_{k=r}^{n} \binom{n}{k} p^k (1-p)^{n-k} = I_p(r, n-r+1),$$

waarin I de onvolledige Bêta-functie voorstelt in de notatie van K,PEARSON, [15].
Het betrouwbaarheidsinterval voor de onbekende fractie p_o, wordt dus gevormd door al die waarden p, waarvoor

$$I_p(r, n-r+1) > \tfrac{1}{2}\alpha$$
$$1-I_p(r+1, n-r) > \tfrac{1}{2}\alpha .$$

De functie I_p is een monotoon stijgende functie van p. Beschouwen wij r weer als de stochastische grootheid $\underline{r}$, dan wordt dus ook het betrouwbaarheidsinterval stochastisch. De grenzen $\underline{p}_1$ en $\underline{p}_2$ hiervan, kunnen, voor een aantal waarden van α, bepaald worden met behulp van de in de litteratuurlijst vermelde nomogrammen [12] , [15] en [16].

Een naar beneden begrensd betrouwbaarheidsinterval voor p_o, met onbetrouwbaarheidscoëfficiënt α, wordt voor $\underline{r}=r$ gevormd door al die waarden p, waarvoor

$$I_p(r, n-r+1) > \alpha,$$

en een naar boven begrensd betrouwbaarheidsinterval door de waarden p, waarvoor

$$1-I_p(r+1, n-r) > \alpha .$$

Bovenstaande methode kan eveneens toegepast worden voor meer-dimensionale verdelingen (al of niet continu), waarvoor een niet-stochastisch tolerantiegebied T gedefiniëerd is. De wh, dat een punt in dit gebied valt, is dan de onbekende fractie p_o. Het betrouwbaarheidsinterval voor deze fractie p_o wordt weer bepaald door het totale aantal punten n van de steekproef en het aantal r hiervan, dat in het gebied T ligt.

Litteratuur.

A. Tolerantiegrenzen.

1 D.A.S. Fraser, Sequentially determined statistically equivalent blocks, Ann.Math.Stat. 22 (1951), p. 372-382.

2 D.A.S. Fraser and R. Wormleighton, Nonparametric estimation IV, ann.Math.Stat. 22 (1951), p. 294-298.

3 R.B. Murphy, Non-parametric tolerance-limits, Ann.Math. Stat. 19 (1948), p. 581-589.

4 H.Scheffé and J.W. Tukey, A formula for sample-sizes for population-tolerance-limits, Ann.Math.Stat. 15 (1944)p.217.

5 H. Scheffé and J.W. Tukey, Non-parametric estimation,I. Validation of order statistics, Ann.Math.Stat. 16 (1945).

p. 187 - 192.

6 Statistical Research Group, Columbia University, Techniques of Statistical Analysis, New York 1947, p. 3 - 235 (Industrial statistics).

7 J.W. Tukey, Non-parametric estimation, II. Statistically equivalent blocks and tolerance regions, the continuous case. Ann Math Stat 18 (1947), p. 529 - 539.

8 J.W. Tukey, Non-parametric estimation,III. Statistically equivalent blocks and multivariate tolerance regions, the discontinuous case. Ann Math Stat. 19 (1948), p. 30-39.

9 A. Wald, An extension of Wilks' method for setting tolerance-limits, Ann Math Stat 14 (1943), p. 45 - 55.

10 S.S. Wilks, Determination of sample-sizes for setting tolerance limits, Ann Math Stat 12 (1941), p. 91 - 96.

11 S.S. Wilks, Statistical prediction with special reference to the problem of tolerance limits, Ann Math Stat 13 (1942), p. 400 - 409.

B. Betrouwbaarheidsintervallen.

12 M.G. Kendall, The advanced theory of statistics, London 1946, deel II, p. 62 - 84.

13 A.M. Mood, Introduction to the theory of statistics, London 1950, p. 220 - 244.

14 J. Neyman, First course in probability and statistics, New York 1950.

15 Statistical Research Group, Columbia University, Techniques of Statistical Analysis, New York 1947, p. 3 - 235(Industrial statistics).

C. Tabellen en Nomogrammen.

16 W.J. Dixon and F.J. Massey, Introduction to statistical analysis, New York 1951.

17 R.A. Fisher and F. Yates, Statistical tables, London 1949.

18 T.C. Fry, Probability and its engineering uses, New York 1928.

19 R.B. Murphy, Non-parametric tolerance-limits, Ann Math. Stat 19 (1948), p. 581 - 589.

20 K. Pearson, Tables of the incomplete B-function, London 1934.

21 L. Simon, Engineer's manual of statistical methods, New York 1945.

Enkele voorbeelden van het gebruik van tolerantiegrenzen.

I. Inleiding.

In de kwaliteitscontrole maakt men meestal gebruik van de theorie der stochastische tolerantiegrenzen, wanneer men zich een algemeen idee wil vormen omtrent de kwaliteit van een hoeveelheid producten.
Wanneer een zeker kenmerk $\underline{x}$ van het product beschouwd wordt, is de vraag, die men zich dan stelt, deze: Tussen welke twee grenzen $\underline{L}_1$ en $\underline{L}_2$ voor $\underline{x}$ ligt, behoudens een onbetrouwbaarheid α, de $\underline{x}$ van tenminste een fractie β van de elementen uit de collectie?
Om deze grenzen te bepalen, maakt men gebruik van een steekproef, die op de volgende wijze gevormd wordt. Noemen we de naar opklimmende grootte gerangschikte waarden uit de steekproef $x_1, x_2, \ldots, x_n$ en worden uit deze steekproef de waarden $\underline{x}_r$ en $\underline{x}_{n-s+1}$ als tolerantiegrenzen $\underline{L}_1$ en $\underline{L}_2$ gekozen, dan ligt, behoudens een onbetrouwbaarheid α, de $\underline{x}$ van tenminste een fractie β van de elementen tussen $\underline{x}_r$ en $\underline{x}_{n-s+1}$, indien de uitgebreidheid n van de steekproef groot genoeg is. De minimale uitgebreidheid n van de steekproef, waarbij het bovenstaande zal gelden, wordt bepaald uit de relatie (zie p. 96 van de syllabus),

$$I_\beta(t, n - t + 1) = \alpha,$$

met

$$t = r + s.$$

Daar de verdeling van $\underline{x}$ meestal niet erg "scheef" is, wordt $r = s$ genomen. In het algemeen neemt men r en $s \neq 1$ om de invloed van z.g.n. "uitbijters" te elimineren.

Wanneer men zich op bovenstaande wijze een idee gevormd heeft, omtrent het algemene gedrag van het kenmerk $\underline{x}$, kan men vervolgens de vraag stellen, of deze partij geschikt is om gebruikt te worden voor een speciaal doel. Hiertoe is, b.v. om economische redenen, noodzakelijk dat tenminste een gegeven fractie van de elementen uit de partij een $\underline{x}$ bezit tussen 2 vaste grenzen L_1 en L_2. Wanneer deze twee grenzen resp. kleiner en groter dan zijn dan de hierboven bepaalde stochastische grenzen $\underline{L}_1$ en $\underline{L}_2$, ligt, behoudens een onbetrouwbaarheid α, tenminste de $\underline{x}$ van zeker een fractie β van de elementen tussen de vaste

grenzen L_1 en L_2.
Wanneer de twee grenzen L_1 en L_2 niet kleiner resp. groter zijn dan $\underline{L_1}$ en $\underline{L_2}$, dan is het noodzakelijk, met behulp van een nieuwe steekproef $x_1, x_2, \ldots x_n$ te bepalen, welke fractie γ van elementen uit de collectie, behoudens een onbetrouwbaarheid α, ten minste een $\underline{x}$ bezit, die tussen de vaste tolerantiegrenzen L_1 en L_2 ligt.
Om deze fractie γ te bepalen, wordt een naar beneden begrensd betrouwbaarheidsinterval $[\gamma, 1]$ berekend voor de werkelijke onbekende fractie p van elementen met een $\underline{x}$ tussen L_1 en L_2. De benedengrens γ van het betrouwbaarheidsinterval wordt bepaald uit de relatie:

$$I_\gamma (r, n - r + 1) \geq \alpha ,$$

waarin n de uitgebreidheid van de steekproef is en r het aantal elementen uit de steekproef met een $\underline{x}$ tussen L_1 en L_2 (zie p. 113 en 114 van de syllabus).

Voorbeeld.

. Gevraagd wordt, twee grenzen $\underline{L_1}$ en $\underline{L_2}$ te bepalen, waartussen, behoudens een onbetrouwbaarheid $\alpha = 0{,}10$, de $\underline{x}$ ligt van tenminste een fractie $\beta = 0{,}75$ van de elementen uit de collectie.
Wanneer $r = s = 3$ gekozen wordt, volgt de grootte van de uitgebreidheid n van de steekproef uit

$$I_{0,75}(6, n - 5) = 0{,}10 .$$

Hieruit volgt $n = 17$.
De op 3 na kleinste en op 3 na grootste waarde uit een steekproef van de uitgebreidheid $n = 17$ zijn de gevraagde grenzen $\underline{L_1}$ en $\underline{L_2}$.
Wanneer deze steekproef, naar opklimmende grootte gerangschikt, de volgend waarden geeft:
24,8; 25,0; $\underline{25{,}8}$; 26,2; 26,3; 26,6; 27,1; 27,3; 27,6; 28,1; 28,2; 28,4; 28,5; 28,7; $\underline{28{,}9}$; 30,1; 30,4,
zijn de twee gevraagde grenzen resp. 25,8 en 28,9.

2. Wanneer men nu echter wil weten, welke fractie γ van elementen uit de collectie, behoudens een onbetrouwbaarheid $\alpha = 0{,}10$, tenminste een $\underline{x}$ bezit, die ligt tussen de $\underline{\text{vaste}}$ grenzen $L_1 = 25{,}0$ en $L_2 = 27{,}5$, is het noodzakelijk een nieuwe steekproef te nemen.
Stel, dat dit de volgende steekproef is:
‖25,1; 25,3; 25,8; 26,1; 26,2; 26,5; 26,7; 27,1; 27,2; 27,4‖
27,6; 28,1; 28,3; 28,7; 28,9; 29,2; 29,5; 29,9; 30,1; 30,3;

$n = 20$,

$r = 10$,

$\alpha = 0{,}10$

De gevraagde fractie γ wordt bepaald uit

$$I_\gamma\ (10,11) \geqq 0{,}10\ .$$

Hieruit volgt $\gamma = 0{,}38$.

P. 92 van de syllabus:

te bewijzen: $$\sum_{N_1=0}^{N-m_0} \binom{N_1+r-1}{r-1}\binom{N-(N_1+m_0)+(s-1)}{s-1} = \binom{N-m_0+r+s-1}{r+s-1}$$

Bewijs:

$$(1-t)^{-r} = 1+rt+\frac{r(r+1)}{2!}t^2+\frac{r(r+1)(r+2)}{3!}t^3+\ldots\ldots$$

$$= \sum_{k=0}^{\infty} \binom{r-1+k}{r-1}\ t^k$$

Evenzo:

$$(1-t)^{-s} = \sum_{l=0}^{\infty} \binom{s-1+l}{s-1}\ t^l$$

en

$$(1-t)^{-(r+s)} = \sum_{m=0}^{\infty} \binom{r+s-1+m}{r+s-1}\ t^m$$

Wegens

$$(1-t)^{-(r+s)} = (1-t)^{-r}(1-t)^{-s},$$

geldt

$$\sum_{m=0}^{\infty} \binom{r+s-1+m}{r+s-1}\ t^m = \sum_{k=0}^{\infty} \binom{r-1+k}{r-1}\ t^k \sum_{l=0}^{\infty} \binom{s-1+l}{s-1}\ t^l.$$

Hieruit volgt door gelijkstelling van coëfficiënten van dezelfde machten van t in linker-en rechterlid:

$$\binom{r+s-1+m}{r+s-1} = \sum_{k=0}^{m} \binom{r-1+k}{r-1}\binom{s-1+m-k}{s-1}$$

Stellen we hierin $m = N-m_0$

en $k = N_1$,

dan gaat bovenstaande betrekking over in

$$\binom{r+s-1+N-m_0}{r+s-1} = \sum_{N_1=0}^{N-m_0} \binom{r-1+N_1}{r-1}\binom{s-1+N-m_0-N_1}{s-1},$$

hetgeen te bewijzen was.

Aanvulling bij de syllabus van de cursus "Parametervrije methoden", hoofdstuk V.

J. Fabius

Na het tot stand komen van deze cursus is de theorie van de tolerantiegebieden verder gegeneraliseerd, voornamelijk door D.A.S. Fraser. [1], [2].
Hij heeft nodige en voldoende voorwaarden gegeven waaraan de karakteristieke functie van een gebied moet voldoen, opdat het een verdelingsvrij tolerantiegebied is, [2]
Voorts zijn door hem, en ook door J.H.B. Kemperman, [1] en [3],algemenere methoden ontwikkeld om de steekproefruimte in blokken te verdelen. Bij deze methoden hangt elke stap af van de bij de vorige stappen verkregen resultaten. In dit opzicht kunnen we dus van sequente methoden spreken. Er is ook een in ander opzicht sequente methode ontwikkeld door J.C. Saunders, [4]. Bij zijn methode, die van toepassing is op het ééndimensionale geval, begint men met een aantal waarnemingen, waarna men, op grond hiervan, beslist of er nog een waarneming gedaan zal worden of niet. Herhaling van deze procedure leidt tenslotte tot een steekproef waarvan de omvang stochastisch is, met een op deze steekproef gebaseerd tolerantiegebied.

Litteratuur:

1. D.A.S. Fraser, Non-parametric tolerance regions, Ann. Math. Stat. 24 (1953), p. 44-55.
2. D.A.S. Fraser and I Guttman, Tolerance regions, Ann. Math. 27 (1956), p. 162-179.
3. J.H.B. Kemperman, Generalized tolerance limits, Ann. Math. Stat. 27 (1956), p. 180-186.
4. J.C. Saunders, Sequential distribution-free tolerance regions, (Abstr.) Ann. Math. Stat. 27 (1956), p. 865.
5. Z.W. Birnbaum and H.S. Zuckerman, A graphical determination of sample size for Wilks' tolerance limits, Ann. Math. Stat. 20 (1949), p. 313-316.
6. D.B. Owen, Distribution-free tolerance limits (tabellen), Technical Memorandum SCTM 66A-57-51. (Verkrijgbaar bij Office of technical services, Department of Commerce, Washington 25, D.C.).

MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S ~~59~~ 76

Cursus Parametervrije Methoden

VI. De mediaantoets en de toets van Smirnov

door

Dr E.F.Drion

Januari 1952

§1. Probleemstelling. Vergelijking met de toets van Wilcoxon.

In het derde hoofdstuk van deze cursus werd de oplossing behandeld, die door WILCOXON werd gegeven voor het probleem van de twee steekproeven. Ter herinnering moge het voorbeeld van dit ptobleem herhaald worden, dat aldaar werd gegeven. Twee groepen duiven werden op een verschillend diëet gesteld en de onderzoeker wenst na te gaan of deze diëten na enige tijd geleid hebben tot verschillende vetgehalten. In de meeste gevallen waarin zich het probleem van de twee steekproeven voordoet, zoals ook in het hier behandelde voorbeeld, zal het de onderzoeker vooral interesseren of het niveau van de geanalyseerde grootheid in de ene steekproef hoger ligt dan in de andere. Het is vooral op een onderscheid in niveau, dat de toets van Wilcoxon gericht is.

Er zijn echter een aantal gevallen, waarin deze toets minder goed bruikbaar is.

In de eerste plaats kan het voorkomen, dat de onderzoeker een geheel nieuw terrein betreedt en in het geheel geen idee heeft of de verschillende condities, waaronder de twee steekproeven staan, een verschil in niveau ten gevolge zullen hebben, of b.v. een verschil in spreiding, of wel een verschil in scheefheid of exces van de frequentiekromme van het onderzochte verschijnsel. Met andere woorden, men wenst te onderzoeken of de frequentiekrommen van het onderzochte verschijnsel onder de verschillende condities enig verschil vertonen. Is dit het geval, dan kan een verder onderzoek, hetzij aan de hand van de gevonden uitkomsten of aan de hand van een nieuwe proef nadere inlichtingen verschaffen over de vraag, waarin het verschil nu eigenlijk bestaat. Voor een dergelijk onderzoek is de toets van SMIRNOV geschikt. Deze toets werd gevonden door SMIRNOV in 1939 [1]. Later gaf FELLER [2] er een nieuwe afleiding van. De formule van Smirnov is een asymptotische formule, d.w.z. zij geldt voor ein-

dige steekproeven slechts bij benadering. Wij zullen hier voor het geval beide steekproeven even groot zijn, een nieuwe exacte formule geven, die op eenvoudige wijze afgeleid kan worden. Voor oneindige grote steekproeven gaat deze formule in die van Smirnov over.

Een tweede geval, waarin de toets van Wilcoxon minder geschikt is, is het geval dat onder de waarnemingen vele gelijke uitkomsten voorkomen, hetzij doordat de verdelingen van nature discontinu zijn (aantal kelkbladen van dotterbloemen), of doordat de metingen slechts met een beperkte nauwkeurigheid verricht zijn. In het derde hoofdstuk werd reeds een kunstgreep behandeld om in dit geval toch een analogon van de toets van Wilcoxon toe te passen. In dit geval kan echter vaak beter de mediaan- (eventueel een quantile) toets toegepast worden in de vorm daaraan gegeven door HEMELRIJK. De mediaantoets is ook weer hoofdzakelijk gevoelig voor een verschil in niveau bij de beide steekproeven.

Een voordeel van de mediaantoets is verder nog, dat in aansluiting eraan een intervalschatting voor de mediaan kan worden gegeven. De exacte afleiding van de mediaantoets is in 1948 door WESTENBERG [3] gevonden voor het geval van discontinue verdelingen. Onafhankelijk van Westenberg is de mediaantoets later eveneens door A.M. MOOD [4] gevonden. HEMELRIJK [5] tenslotte gaf aan hoe, in geval dat de verdelingsfunctie niet overal continu is, de toets gewijzigd kan worden.

§2. Grafische voorstelling van het resultaat van een probleem van twee steekproeven.

Bij elke statistische toets dient men zich te realiseren, wat precies de hypothese H_o is, die getoets wordt, en welke de mogelijke alternatieve hypothesen zijn, die men dus bereid is aan te nemen, indien de getoetste hypothese verworpen wordt. Het probleem van de twee steekproeven doet zich gewoonlijk in deze vorm voor, dat de te toetsen hypothese H_o luidt: De uitkomsten $x_1 \dots x_{n_1}$ en $y_1 \dots y_{n_2}$ zijn waarnemingen van de onderling onafhankelijk verdeelde stochastische grootheden $\underline{x}_1 \dots \underline{x}_{n_1}$ en $\underline{y}_1 \dots \underline{y}_{n_2}$, die alle dezelfde verdeling bezitten.

De alternatieve hypothesen H luiden gemeenlijk: De uitkomsten $x_1 \dots x_{n_1}$ en $y_1 \dots y_{n_2}$ zijn waarnemingen van onderling onafhankelijk verdeelde stochastische grootheden $\underline{x}_1 \dots \underline{x}_{n_1}$ en $\underline{y}_1 \dots \underline{y}_{n_2}$. De stochastische grootheden $\underline{x}_1 \dots \underline{x}_{n_1}$ heb-

ben alle dezelfde verdeling, evenals de stochastische grootheden $y_1 \cdots y_{n_2}$. De verdeling van de grootheid $\underline{x}$ is verschillend van die van de grootheid $\underline{y}$.

Vaak worden dan nog de mogelijke verschillen tussen de verdeling van $\underline{x}$ en die van $\underline{y}$ onder de alternatieve hypothese H verder gespecifieerd.

(Het is mogelijk om als hypothese H_o een ruimere te kiezen, zie hiervoor b.v. HEMELRIJK (§3.2.1) [5]. Deze ruimere hypothese zullen wij hieronder nog behandelen. (blz 122)).

Indien de hypothese H_o juist is, zijn de uitkomsten $x_1 \cdots x_{n_1}$ en $y_1 \cdots y_{n_2}$ dus $n_1 + n_2$ waarnemingen van een stochastische grootheid $\underline{w}$ (er is in dit geval t.a.v. de onderzochte eigenschap, geen verschil tussen de stochastische grootheden $\underline{x}$ en $\underline{y}$, zodat wij deze, om verwarring te voorkomen, met het symbool $\underline{w}$ zullen aangeven).

Indien nu de $n_1 + n_2$ waarnemingen in volgorde van de grootte gerangschikt worden (hetzij van hoog naar laag, zoals wij in het vervolg om de gedachte te bepalen zullen veronderstellen, of van laag naar hoog) en elke waarneming, al naar gelang deze uit de eerste serie of de tweede serie afkomstig is door een x of een y vervangen wordt, dan wordt dus een bepaalde groepering van n_1 letters x en n_2 letters y verkregen. Het is duidelijk dat onder de hypothese H_o elk der mogelijke $\binom{n_1 + n_2}{n_1}$ groeperingen even waarschijnlijk is. Er zijn n.l. $n_1 + n_2$ plaatsen, waarop de n_1 letters x en de n_2 letters y geplaatst kunnen worden. Voor de eerste x zijn er dus $n_1 + n_2$ plaatsen, voor de tweede x $(n_1 + n_2 - 1)$ plaatsen enz. Wij kunnen de letters x dus op $(n_1 + n_2).(n_1 + n_2 - 1)\ldots(n_2 + 1) = \frac{(n_1 + n_2)!}{n_2!}$ manieren rangschikken. Dit zijn echter niet evenzoveel verschillende groeperingen, daar dezelfde situatie ontstaat wanneer b.v. de eerste x op plaats a en de tweede x op plaats b wordt gelegd als wanneer de eerste x op plaats b en de tweede op plaats a wordt gelegd. Wij moeten het gevonden aantal $\frac{(n_1 + n_2)!}{n_2!}$ dus nog door $n_1!$ (het aantal permutaties van n_1 letters x) delen om het aantal mogelijke groeperingen te verkrijgen. (Voorlopig zullen wij aannemen, dat de verdelingen continu zijn, zodat onder de waarnemingen, (behoudens een waarschijnlijkheid nul) geen gelijke voorkomen.

Later zullen wij de complicaties bespreken, welke optreden indien er wel gelijke onder de waarnemingen voorkomen).

Wij kunnen nu op de volgende wijze elk dezer groeperingen voorstellen door wegen in een tweedimensionaaldiagram.

Teken een rechthoekig rooster van $(n_1+ 1)$ bij $(n_2+ 1)$ punten. Elke x wordt voorgesteld door een horizontale lijn tussen twee opeenvolgende punten, elke y door een verticale lijn. Begonnen wordt bij de linker benedenhoek. Er ontstaat zo een gebroken lijn, die deze linker benedenhoek O met de rechter bovenhoek P van het rooster verbindt.

Als voorbeeld mogen dienen de uitkomsten van de proeven met duiven vermeld in hoofdstuk III §4 (Serie I en Serie III).

Gevonden werd: voor groep I 1,42/2,22/1,68/1,68/2,60/2,54/7,80/2,28/

voor groep III 3,02/3,42/6,43/2,93/2,23/7,85/

Als groep I met de letter x, groep III met de letter y aangeduid wordt, dan wordt na rangschikking in volgorde van grootte gevonden:

7,85/7,80/6,43/3,42/3,02/2,93/2,60/2,54/2,28/2,23/2,22/1,68/1,68/1.42/

y x y y y y x x x y x x x x

Ons diagram wordt nu:

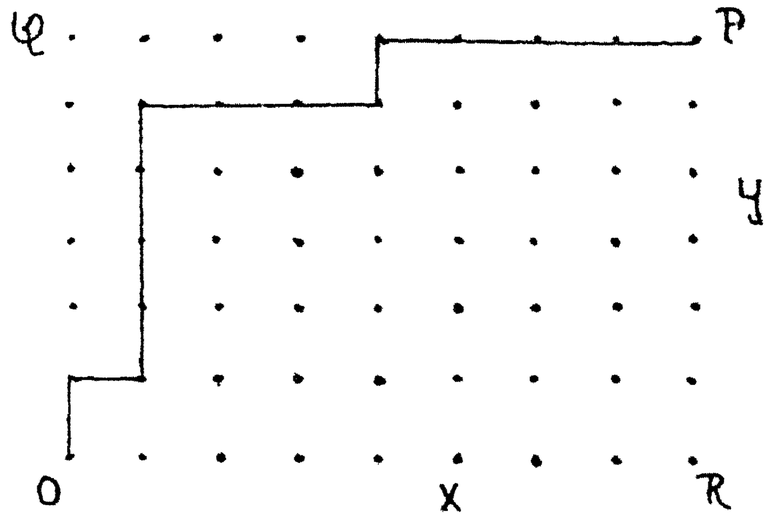


Fig. 1.

Aanschouwelijke voorstelling van een x,y-groepering.

Onder de hypothese H_o is elke weg in dit diagram even waarschijnlijk, onder de alternatieve hypothese H echter niet meer.

Is b.v. onder de hypothese H de verdelingsfunctie van $\underline{x}$, $P[\underline{x} \leqq x] = F(x)$ en die van y, $P[\underline{y} \leqq y] = G(y) = F(y+d)$, dan zullen de wegen, die in het begin hoofdzakelijk verticaal, en later hoofdzakelijk horizontaal lopen (een extreem voorbeeld van zulk een weg is OQP) waarschijnlijker zijn dan de andere wegen. Immers voor de meeste waarden van a is $P[\underline{x} \leqq a] = F(a)$

kleiner dan $P[y \leqq a] = F(a+d)$, daar $F(z)$ een monotoon niet-dalende functie van z is, dus $F(a)$ meestal (bij continue functies altijd) $\leqslant F(a+d)$ is.

In dit geval zullen wij dus als kritieke zône die wegen kiezen, die dicht langs Q lopen, d.w.z. dat wij de hypothese H_o verwerpen als de weg OP dicht langs Q verloopt. Indien d zowel positief als negatief kan zijn, zullen wij, als kritieke zône die wegen kiezen, die hetzij dicht langs Q of dicht langs R lopen.

Zoals gemakkelijk in te zien is, is de grootheid U van de toets van Wilcoxon (zie hoofdstuk III § 1.2) de oppervlakte tussen OQR en de weg OP, als de punten van het rooster op afstand 1 staan. Als de weg dicht langs Q loopt, is deze grootheid klein, als de weg dicht langs R loopt, ligt U dicht bij $n_1 \cdot n_2$. Op deze wijze wordt dus aanschouwelijk, dat de toets van Wilcoxon voornamelijk gevoelig is voor een verschil in niveau.

Wij hebben in het voorafgaande stilzwijgend aangenomen, dat de verdelingsfunctie overal continu is, zodat, behoudens een waarschijnlijkheid 0, twee verschillende waarnemingen steeds een verschillende waarde vertonen. Het is dus in dit geval steeds mogelijk alle waarnemingen $x_1 \cdots x_{n_1}$ en $y_1 \cdots y_{n_2}$ op ondubbelzinnige wijze in één serie van groot naar klein te rangschikken; de weg door het rooster ligt dan dus ook vast. Aangezien de mediaantoets (in de vorm die Hemelrijk daarvoor gegeven heeft) ook toepasbaar is, indien de verdelingsfuncties niet overal continu zijn, zullen wij nu eerst bespreken welke complicaties uit discontinuiteiten voortvloeien.

Indien de verdelingsfunctie niet overal continu is, is het niet steeds meer mogelijk de waarnemingen op één wijze van groot naar klein te rangschikken; een zelfde soort moeilijkheid troffen wij ook aan bij de toets van Wilcoxon, indien er gelijke waarnemingen voorkomen. Een gevolg hiervan is, dat het aantal mogelijke wegen door het rooster niet meer bepaald is: stel b.v. dat x_i gelijk is aan y_j. Dan loopt de weg door het rooster van het punt $(i-1, j-1)$ naar het punt (i,j), maar we kunnen niet meer zeggen, dat deze weg hetzij over het punt $(i-1, j)$ dan wel over het punt $(i, j-1)$ moet lopen.

Om in dit geval ook tot ondubbelzinnige rangschikking te komen, zodat er $\binom{n_1 + n_2}{n_1}$ wegen door het rooster zijn, kunnen verschillende methoden toegepast worden. Wij zullen er hier twee noemen, die in wezen wel gelijk zijn. De eerste methode,

die wij verder niet zullen toepassen, is in principe afkomstig van de Russische mathematicus LIAPOUNOFF [6] Hierbij wordt bij de stochastische grootheid $\underline{x}$ een stochastische grootheid $\underline{v}$ opgeteld, welke grootheid $\underline{v}$ een normale verdeling met gemiddelde nul en spreiding σ heeft. Het is duidelijk, dat de stochastische grootheid $\underline{z} = \underline{x} + \underline{v}$ een verdelingsfunctie zal hebben, die des te minder van die van $\underline{x}$ zal afwijken, naarmate σ kleiner is. Voorts heeft $\underline{z}$, als de som van een cohtinu verdeelde grootheid $\underline{v}$ en een eventueel niet continu verdeelde grootheid $\underline{x}$, zelf een continue verdeling.

Nu kan bewezen worden, dat voor $\sigma \searrow 0$, de verdeling van $\underline{z}$ voor alle punten x waar de verdeling van x continu is, tot die van x convergeert. Dit bewijs, dat wij hier niet zullen geven, staat o.a. vermeld in USPENSKY [7] , hoofdstuk 13, sect, 7.

Door dit toevoegen van een kleine toevallige (normaal verdeelde) grootheid aan alle waarnemingen wordt de verdeling, zoals boven gezegd, continu, zodat $P\left[\underline{x}_i = \underline{y}_j\right] = 0$ is voor alle combinaties van $i = 1 \ldots n_1$ en $j = 1 \ldots n_2$. Deze methode wordt in de practijk nog slechts zelden toegepast (zie echter STEVENS [8], die echter een stochastische grootheid $\underline{v}$ met een rechthoekige verdeling optelt bij de gevonden uitkomsten), doch is van groot belang voor de afleiding van verschillende limietstellingen uit de waarschijnlijkheidsrekening. Men zou de methode wellicht kunnen noemen: "randomising a posteriori".

Een tweede methode, die men "randomising a priori" zou kunnen noemen, is in principe door FISHER aangegeven (FISHER [9], zie ook LEHMAN en STEIN [10]). De methode kan in ons geval op de volgende wijze toegepast worden.

De proefobjecten worden voor het begin van de proef genummerd van 1 tot $n_1 + n_2$. Vervolgens wordt door één of andere loting bepaald welke objecten het kenmerk x en welke het kenmerk y krijgen, dus b.v. welke duiven het ene en welke het andere diëet zullen krijgen. Na de proef worden de uitkomsten in volgorde van de grootte gerangschikt; zijn er gelijke onder de uitkomsten, dan worden deze in volgorde van de oorspronkelijke nummering gerangschikt. Op deze wijze wordt bereikt, dat er steeds een ondubbelzinnige rangschikking bestaat, zodat ook het aantal wegen door het rooster wederom $\binom{n_1 + n_2}{n_1}$ is.

Er bestaat nog een tweede, vaak zeer belangrijk voordeel van het consequent toepassen van methoden bij experimenteren en wel het volgende.

Voor het mathematische model, waarvan wij zijn uitgegaan,

n.l. de hypothese H_o alsmede alle alternatieve hypothesen H, geldt, dat de n_1 stochastische grootheden $\underline{x}_1 \ldots \underline{x}_{n_1}$ alle dezelfde verdelingsfunctie hebben, en evenzo de n_2 stochastische grootheden $\underline{y}_1 \ldots \underline{y}_{n_2}$. Wij weten echter, dat deze gelijkheid der verdelingsfunctie in de practijk slechts bij benadering geldt. Indien de objecten voor het experiment volgens een of ander lotingsprincipe over de beide behandelingswijzen verdeeld zijn, dan blijven de kansen op een bepaalde uitkomst, berekend voor het geval de hypothese H_o juist is, gelden, ook als i.p.v. de hypothese H_o zou gelden de hypothese H_o':

De n_1 grootheden $x_1 \ldots x_{n_1}$ en de n_2 grootheden $y_1 \ldots y_{n_2}$ zijn waarnemingen van de $n_1 + n_2$ onderling onafhankelijke stochastische grootheden $\underline{w}_1 \ldots \underline{w}_{n_1+n_2}$, die eventueel een verschillende verdelingsfunctie kunnen hebben. Door één of andere loting, waarbij elke permutatie $\underline{w}_1 \ldots \underline{w}_{n_1+n_2}$ even waarschijnlijk is, zijn deze $n_1 + n_2$ grootheden over de beide behandelingswijzen verdeeld (zie HEMELRIJK [5], Hoofdstuk 3 . Deze opmerking geldt <u>niet</u> voor de asymptotische bruikbaarheid der toets, zie HEMELRIJK [5] 3.2.7. en vlgd.).

In sommige gevallen is het loten a priori niet mogelijk. Dan kan men ook achteraf om de volgorde van gelijke waarnemingen loten, wat op hetzelfde neerkomt als toepassing van de methode van Liapounoff in een enigszins gewijzigde vorm.

Hoewel bij toepassing van het boven aangegeven lotingsprincipe volgens Hemelrijk de parametervrije statistische methoden ook bij ongelijkheid van het experimenteel materiaal onbeperkt toegepast mogen worden, is het toch wenselijk om een dergelijke ongelijkheid zoveel mogelijk te vermijden, teneinde de kansen op fouten van de tweede soort (niet-verwerpen van de hypothese H_o, hoewel deze onjuist is) te verkleinen.

De voorafgaande beschouwingen zullen velen min of meer gekunstold lijken. Vooral de methode van Liapounoff zal menigeen afschrikken, daar verschillende statistici, als zij deze methode op hetzelfde waarnemingsmateriaal (zelfde uitkomsten $x_1 \ldots x_{n_1}$ en $y_1 \ldots y_{n_2}$) toepassen tot verschillende resultaten (overschrijdingskansen) zullen komen. Bij de "randomisation a priori" valt deze moeilijkheid niet zo zeer op, daar hier vaak de nummering al door de experimentator gedaan werd. Toch is de moeilijkheid hier in feite ook aanwezig. Het bovengenoemd bezwaar is hoofdzakelijk een kwestie van (denk)gewoonte: wij zien er niets vreemds in, dat bij herhaling van een proef een andere

overschrijdingskans gevonden wordt dan de eerste maal !

Wij zullen echter zien, dat bij de vorm, die Hemelrijk aan de mediaantoets gegeven heeft, het mogelijk is om een methode aan te geven, waarbij iedere statisticus uit een reeks uitkomsten tot dezelfde overschrijdingskans komt.

§3. Principe van de mediaantoets.

Bij de niet-parametervrije methoden kiezen wij als karakteristieke grootheid om het niveau van een stochastische grootheid aan te geven meestal het (rekenkundig) gemiddelde, bij de parametervrije methoden gewoonlijk de mediaan. Bij een oneven aantal waarnemingen is de mediaan de waarde van de middelste waarneming, als deze in volgorde van grootte gerangschikt zijn, bij een even aantal 2n enig getal tussen de n^{de} en de $(n+1)^{ste}$ waarneming. Vaak kiest men, om de onbepaaldheid van de mediaan bij een even aantal waarnemingen te omzeilen, hiervoor het gemiddelde van de n^{de} en de $(n+1)^{ste}$ waarneming. Dit is echter zuiver een conventie, zij het een practische conventie. (Opgemerkt moge nog worden, dat voor de mediaan x_M een analoge minimum eigenschap geldt als voor het gemiddelde: bij het gemiddelde $\bar{x}$ is $\sum(x_i-\bar{x})^2$ een minimum, bij de mediaan x_M is $\sum|x_i-x_M|$ een minimum).

Bij de mediaantoets wordt als toetsingsgrootheid om de gelijkheid van twee verdelingen te beoordelen het aantal waarnemingen tussen de (experimentele) medianen van de waarnemingen der beide series gebruikt, of een grootheid, die met de genoemde op eenvoudige wijze samenhangt. Indien de hypothese H_o (of H_o') juist is, zullen (meestal) tussen de medianen x_M en y_M slechts weinig waarnemingen van de beide reeksen liggen. Derhalve zullen x_M en y_M ook vrijwel samenvallen met de mediaan w_M van het gehele materiaal (dus de mediaan van de reeks, die men krijgt, als men alle waarnemingen in één serie van hoog naar laag rangschikt). Is de hypothese H_o niet juist en is een alternatieve hypothese H juist, volgens welke er een verschil in niveau is tussen de grootheid $\underline{x}$ en $\underline{y}$, dan zullen tussen de medianen x_M en y_M (meestal) veel waarnemingen van beide reeksen liggen.

Wij zullen nu eerst in ons diagram nagaan waar de medianen van elk der waarnemingsreeksen van $\underline{x}$ en $\underline{y}$ liggen, en waar de mediaan van de gehele waarnemingsreeks van $\underline{w}$ ligt.

Kiezen wij als voorbeeld $n_1 = 8$ en $n_2 = 5$.

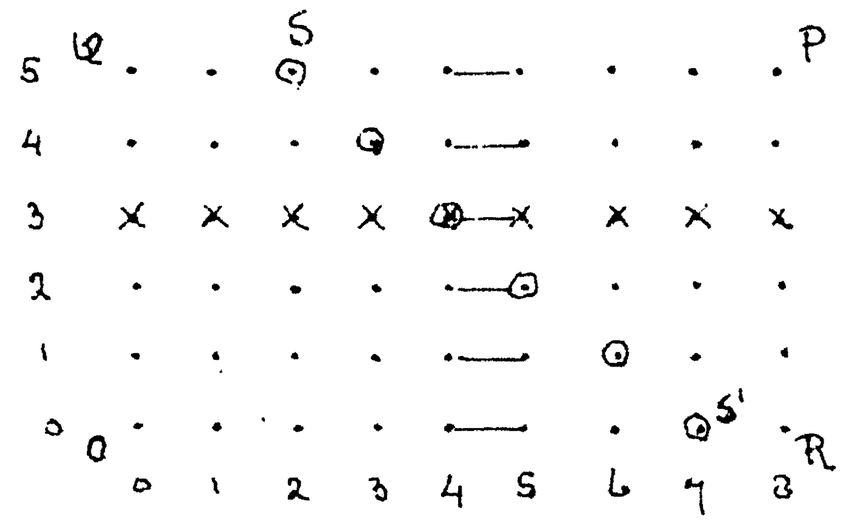


Fig. 2.
Aanschouwelijke voorstelling van de medianen in een x,y-groepering.

Aangezien de afzonderlijke waarnemingen weergegeven worden door horizontale (x-waarnemingen) of verticale (y-waarnemingen) verbindingslijnen tussen twee roosterpunten, is de mediaan van de x-waarnemingen kleiner dan de waarneming weergegeven door een lijnstuk eindigende in een roosterpunt met abscis $\xi = 4$ en groter dan de waarneming weergegeven door een lijnstuk eindigende in een roosterpunt met abscis $\xi = 5$. De mediaan van de y-waarnemingen wordt weergegeven door een lijnstuk eindigende in een roosterpunt met ordinaat $\eta = 3$ (aangegeven door $\times$), die van het gehele materiaal (w) door een horizontaal óf verticaal lijnstuk eindigende in een roosterpunt waarvan abscis ξ en ordinaat η gelijk is aan 7 (daar $n_1 + n_2 = 13$). Deze roosterpunten zijn aangegeven door $\odot$.

Wij moeten nu een kritieke zône uitkiezen. Aangezien wij een toets zoeken, die vooral gevoelig is voor verschil in niveau, zullen wij hiervoor die wegen in het rooster kiezen, die hetzij dicht langs QP of dicht langs OR lopen. Deze keus is trouwens in overeenstemming met wat wij reeds aangaven omtrent de uitvoering van de mediaantoets. Een weg dicht langs QP zal b.v. het punt S (2,5) passeren, zodat er dan tenminste vier waarnemingen tussen de mediaan van de y-waarnemingen en die van de x-waarnemingen liggen (n.l. als de weg door het rooster door de punten (2,3), (2,4), (2,5), (3,5), (4,5) gaat).

Het is duidelijk, dat elke weg door het rooster ten minste één der punten op de lijn SS' (punten gemerkt $\odot$) moet passeren. We kunnen onze kritieke zône dus karakteriseren door aan te geven welke punten op deze lijn er in vallen. De kritieke zône is dus een verzameling van wegen. Karakteristiek voor de wegen die tot de kritieke zône behoren, is, dat zij bepaalde punten op de lijn SS' moeten passeren. Wij zullen deze verzame-

ling van punten, waarvan elke weg der kritieke zône er één passeert, aanduiden met de naam "barrière".

Er blijft echter nog één moeilijkheid op te lossen. Hoe te handelen in het geval, dat er gelijke waarnemingen voorkomen. Aangezien onze kritieke zône geheel bepaald wordt door de barrière, kunnen gelijke waarnemingen alleen dan moeilijkheden opleveren, als een punt van de barrière een waarneming karakteriseert, die gelijk is aan de daaropvolgende waarneming, m.a.w. als de waarneming $w_i = w_j$ terwijl w_i door een horizontaal of verticaal lijnstuk uitkomende in de barrière, gekarakteriseerd wordt. Indien de waarnemingen vooraf door loting over het proefmateriaal verdeeld zijn, en indien de oorspronkelijke nummering opgegeven is, kunnen wij het aangegeven principe van "randomisation a priori" toepassen, en de gelijke waarnemingen in volgorde van de oorspronkelijke nummering nemen. Ook kunnen de methode van Liapounoff of loting achteraf toepassen.

Er is echter ook nog een andere manier, aangegeven door Hemelrijk, om deze moeilijkheid te boven te komen. Zoals wij reeds zagen, wordt de weg door het rooster onbepaald voor die punten, waar een waarneming der ene serie gelijk wordt aan een waarneming der tweede serie, dus waar b.v. $x_i = y_j$ ($i = 1 \ldots n_1$, $j = 1 \ldots n_2$). Welnu, dan zullen wij zorgen, dat onze barrière niet een dergelijk punt bevat. Bevat SS' wel één of meer dergelijke punten, dan kiezen wij onze barrière niet op SS', doch op een lijn S_1S_1' parallel hieraan, die er geen bevat. De vergelijking van deze lijn luidt dus $\xi + \eta = r$ (ξ en η zijn de coördinaten van de roosterpunten, dus gehele getallen). Wij zullen r zodanig kiezen, dat de lijn S_1S_1' zo dicht mogelijk bij SS' gelegen is. Indien er twee dergelijke lijnen mogelijk zijn, die even ver van SS' afliggen, wordt de lijn gekozen, die het dichtst bij P gelegen is (dus waarvoor r de grootste waarde heeft). Dit laatste voorschrift (betreffende de keus bij twee lijnen, die even ver van SS' gelegen zijn) is een conventie, die maakt, dat twee verschillende statistici bij bewerking van hetzelfde materiaal tot dezelfde uitkomst komen [1]. Wij krijgen op die manier feitelijk een voorwaardelijke toets, daar de in de plaats van SS' gekozen lijn van de waarnemingen afhankelijk is, doch deze mag, blijkens een algemene stelling, die door Hemelrijk expliciet bewezen wordt, geinterpreteerd worden als

1) In Hemelrijk [5] is een iets andere definitie van r gegeven, waarbij de lijn S_1S_1' nooit links van SS' kan liggen. Dus $r \geq \frac{n_1+n_2}{2}$. De hier gegeven definitie bezit zekere voordelen.

een onvoorwaardelijke toets (Hemelrijk [5] 3.2.6). Voor het geval van de mediaantoets zullen wij de stelling verderop bewijzen. Een noodzakelijk gevolg van de toepassing van deze methode is wel dat, indien er gelijken onder de waarnemingsuitkomsten zijn, als toetsingsgebied niet meer het aantal waarnemingen tussen de medianen der beide waarnemingsreeksen, doch het aantal waarnemingen tussen de beide q-quantilen gebruikt wordt, waarbij q een getal is, dat in het algemeen (tenzij er zeer vele gelijke waarnemingen zijn) vrij dicht bij 0,5 ligt. (Het q-quantiel is de q.N-de waarneming in een naar grootte gerangschikte reeks van N waarnemingen. Dit is echter in de practische voorkomende gevallen geen ernstig bezwaar: als een alternatieve hypothese H waar is, volgens welke er niveauverschillen zijn tussen de beide populaties, dan zullen die even goed in de q-quantilen ($q \neq 0,5$) als in de medianen tot uitdrukking komen, zeker als q weinig van 0,5 verschilt.

§ 4. Uitvoering van de mediaantoets.

Eerst wordt de ligging van de lijn SS', dus de waarde van r bepaald. Indien alle waarnemingen ongelijk zijn, is $r = \frac{1}{2}(n_1+n_2)$ als n_1+n_2 even is en $r = \frac{1}{2}(n_1+n_2) + \frac{1}{2}$ als n_1+n_2 oneven is. Indien er gelijke onder de waarnemingen voorkomen en $w_1, w_2, \ldots, w_{n_1+n_2}$ de waarnemingen voorstellen, gerangschikt in groepen van gelijken, maar overigens in volgorde van afnemende grootte, dan wordt r gedefinieerd als het getal, dat zo dicht mogelijk bij $\frac{1}{2}(n_1+n_2)$ ligt en waarvoor hetzij $w_{r-1} \neq w_r$ (als $r < \frac{n_1+n_2}{2}$) hetzij $w_r \neq w_{r+1}$ (als $r > \frac{n_1+n_2}{2}$).

De weg in het rooster, die de gevonden steekproeven voorstelt, snijdt de lijn $\xi + \eta = r$ in een punt, waarvan wij de coördinaten voorstellen door (u, r-u). Zowel r als u zijn afhankelijk van de steekproeven en zijn dus stochastische grootheden, indien wij de collectie van alle mogelijke steekproeven beschouwen.

Wij zullen nu nagaan, hoeveel mogelijke wegen van 0 naar P bij gegeven r door een punt (u, r-u) met gegeven u leiden. Het aantal wegen van 0 naar (u, r-u) is het aantal wegen in een rechthoek met zijden u en r-u dus $\binom{r}{u}$ in aantal. Het aantal wegen van (u, r-u) naar P is het aantal wegen in een rechthoek met zijden $n_1 - u$ en $n_2 - r + u$, dus $\binom{n_1+n_2-r}{n_1-u}$ in aantal. Het aantal wegen van 0 via (u, r-u) naar P is dus gelijk aan het product van deze twee aantallen.

Daar er in totaal $\binom{n_1+n_2}{n_1}$ wegen van O naar P leiden, is dus de kans, dat $\underline{u}$ de waarde u aanneemt, als $\underline{r}$ een gegeven waarde r bezit:

$$P_{r,u} = P[\underline{u} = u \mid \underline{r} = r] = \frac{\binom{r}{u}\binom{n_1+n_2-r}{n_1-u}}{\binom{n_1+n_2}{n_1}}$$

Zoals uit figuur 2 afgelezen kan worden, liggen alleen die punten op de lijn SS' waarvoor geldt, dat $\xi \geqq \text{Max}(0, r-n_2)$ en $\xi \leqq \text{Min}(n_1, r)$. De waarschijnlijkheid $P_{r,u}$ neemt aanvankelijk toe met toenemende u, om vervolgens weer af te nemen. De volgende tabel geeft een overzicht van de waarden van $P_{r,u}$ voor $n_1 = 8$, $n_2 = 12$ en $r = 11$ (Voorbeeld ontleend aan HEMELRIJK [5], blz 25).

u	$P_{r,u}$
0	0,00007
1	0,003
2	0,037
3	0,165
4	0,331
5	0,309
6	0,132
7	0,023
8	0,001

Om een kritieke zône te bepalen met onbetrouwbaarheidsdrempel $\leqq \alpha$, moeten, als barrière, die punten op de lijn SS' gekozen worden, die dicht bij OQP resp. dicht bij ORP liggen, zodanig, dat de som van de waarden van $P_{r,u}$ voor deze punten α juist niet overschrijdt. Dit kan natuurlijk nog op verschillende wijzen gebeuren. Bij een tweezijdige toets is een voor de hand liggende keuze wel die, welke bewerkstelligt, dat, als de hypothese H_o juist is, de kans op een weg door het kritieke gebied dicht langs Q ongeveer even groot is als de kans op een weg dicht langs R. Geschiedt de keus op een andere wijze, dan zal de hypothese H_o eerder verworpen worden als b.v. een alternatieve hypothese H juist is, waarbij een grotere kans is, dat de weg langs Q loopt (dus als $x_M < y_M$ is), dan wanneer een alternatieve hypothese H' juist is, waarbij $y_M < x_M$ is. Met andere woorden, het onderscheidingsvermogen van de toets is voor beide mogelijke alternatieve hypothesen ongelijk. Een andere voor de hand liggen-

de keuze is die, waarbij de barrière zoveel mogelijk punten bevat. Meestal zullen beide principes tot practisch gelijke keuze van de barrière leiden. Wij zullen als keuze-principe voor de barrière die, welke tot de grootste barrière leidt toepassen. Hiertoe zullen wij dus die waarden van $P_{r,u}$ bijeen zoeken, die het kleinste zijn en wel zoveel, dat hun som hoogstens gelijk aan α (de onbetrouwbaarheidsdrempel) is. In ons voorbeeld zijn dat dus, als wij $\alpha = 0{,}05$ stellen, $P_{11,0}$, $P_{11,8}$, $P_{11,1}$ en $P_{11,7}$. De som hiervan bedraagt 0,027. Zoals meestal bij discrete waarschijnlijkheden, is het hier onmogelijk een kritiek gebied te vinden met een onbetrouwbaarheid van precies $\alpha = 0{,}05$.

In plaats van aan te geven of het resultaat van een experiment al dan niet in een kritieke zône valt, zullen vele onderzoekers liever de overschrijdingskans aangeven. Men zou deze kunnen definiëren als de onbetrouwbaarheid van de kleinste kritieke zône, die nog juist het resultaat bevat. De gerustheid, waarmee de hypothese H_o verworpen wordt, zal immers des te groter zijn, naarmate deze in een kritieke zône met een kleinere onbetrouwbaarheid valt.

Om de (tweezijdige) overschrijdingskans te berekenen, tellen wij de waarden van $P_{r,u}$ op voor de bij het experiment gevonden waarde u van $\underline{u}$ en voor alle waarden van u, waarvoor $P_{r,u}$ kleiner is. Zouden wij in het gegeven voorbeeld ($n_1 = 8$, $n_2 = 12$, $r = 11$) voor $\underline{u}$ de waarde 1 gevonden hebben, dan is de overschrijdingskans $P_{11,1} + P_{11,8} + P_{11,0} = 0{,}004\ldots$.

Indien de vraagstelling bij het onderzoek dusdanig is, dat alleen alternatieve hypothesen van de vorm $x_M > y_M$ (resp. $x_M < y_M$) interessant of mogelijk zijn, dan kan een eenzijdige toetsing gebruikt worden. In dit geval wil men dus alleen dan de hypothese H_o verworpen zien, indien $x_M > y_M$ (resp. $x_M < y_M$) is. In ons diagram zullen dus die wegen tot verwerping van de hypothese H_o leiden, die dicht langs R (resp. dicht langs Q) lopen. Als barrière moeten dus, hetzij zoveel van de hoogste waarden van u gekozen worden, dat nog juist $\sum_{\underline{u}} P_{r,u} \leqq \alpha$ is, hetzij zoveel van de laagste waarden, dat deze voorwaarde nog juist geldt. Wij werken dan dus met een éénzijdige barrière. Het is duidelijk, dat bij een eenzijdige toetsing de kritieke zône groter wordt (meer punten van de lijn SS' bevat) voor de hoge (resp. lage) waarden van u. Dit betekent natuurlijk ook, dat het onderscheidingsvermogen van de toets groter wordt, dus dat reeds bij

een gegeven verschil tussen de beide medianen, mits liggend in de richting van de toegelaten alternatieve hypothesen, dezelfde kans op verwerping van H_o bestaat, die bij tweezijdige toetsing pas bij een groter verschil bestaat.

De mediaantoets, zoals die hier gegeven is, is, voor het geval dat de verdelingen van $\underline{x}$ en $\underline{y}$ niet continu zijn, een voorwaardelijke toets. De grootheid $\underline{r}$ is n.l. een stochastische grootheid. Bij gegeven verdelingsfunctie F(x) en F(y) (onder de hypothese H_o) bestaat er een kans P(r) dat $\underline{r}$ de waarde r bezit. Toch mag de toets blijkens de volgende redenering als een onvoorwaardelijke toets geinterpreteerd worden:

De barrière B_r bij $\underline{r} = r_1$ behorende bij de kritieke zône is zo gekozen, dat $\sum_{u \in B_r} P_{r,u} \leqq \alpha$. (De betekenis van $u \in B_r$ is: u is abscis van een punt van de barrière B_r). De (samengestelde) waarschijnlijkheid, dat $\underline{r} = r$ en $\underline{u} \in B_r$ is $P(r) . \sum_{u \in B_r} P_{r,u}$; of in woorden uitgedrukt: de waarschijnlijkheid, dat een uitkomst (r,u) gevonden wordt, waarbij $\underline{r}$ de waarde r heeft en u abscis van de barrière is, is $P(r) . \sum_{u \in B_r} P_{r,u}$.

Indien dus voor elke mogelijke waarde van $\underline{r}$ de barrière zo gekozen wordt, dat $\sum_{u \in B_r} P_{r,u} \leqq \alpha$, dan is de waarschijnlijkheid, dat $\underline{u}$ in een barrière valt $\sum_r P(r) . \sum_{u \in B_r} P_{r,u} \leqq$ $\sum_r P(r).\alpha = \alpha$, waarbij de sommatie naar r over alle mogelijke waarden van $\underline{r}$ geschiedt.

§ 5. Asymptotische formule voor de verdeling van $\underline{u}$.

De hier gevonden voorwaardelijke verdeling van $\underline{u}$ is een zgn. hypergeometrische verdeling. Wij zullen nu bewijzen dat het gemiddelde van u voor deze verdeling $\frac{rn_1}{n_1+n_2}$ en het spreidingskwadraat $\frac{r(n_1+n_2-r)n_1n_2}{(n_1+n_2)^2(n_1+n_2-1)}$ bedraagt.

Hiertoe merken wij eerst op dat $\sum \frac{\binom{r}{u}\binom{n_1+n_2-r}{n_1-u}}{\binom{n_1+n_2}{n_1}} = 1$, waarbij gesommeerd wordt over u van $\max(0,r-n_2)$ tot $\min(r,n_2)$. Immers elke weg door het rooster moet één van de punten, liggende op de schuine lijn $\xi + \eta = r$, passeren. Het gemiddelde van $\underline{u}$ bedraagt

$$\mathcal{E}(\underline{u}|\underline{r} = r;H_o) = \sum \frac{u\binom{r}{u}\binom{n_1+n_2-r}{n_1-u}}{\binom{n_1+n_2}{n_1}}$$

In elke term van de som kunnen wij de factor $u\binom{r}{u} = \frac{ur!}{u!(r-u)!}$ vervangen door $\frac{r(r-1)!}{(u-1)!(r-u)!} = r\binom{r-1}{u-1}$, behalve in de eerste term indien $u = 0$. Wij kunnen dus voor $\mathcal{E}(\underline{u})$ schrijven

$$r\sum \frac{\binom{r-1}{u-1}\binom{n_1+n_2-r}{n_1-u}}{\binom{n_1+n_2}{n_1}}$$

De teller van de breuken uit deze som krijgt eenzelfde vorm als die van de verdeling van $\underline{u}$ zelf, wanneer daarin u, r en n_1 vervangen worden door $u'=u-1$, $r'=r-1$ en $n_1'=n_1-1$.

De som van de tellers is dus gelijk aan $\binom{n_1+n_2-1}{n_1-1} = \frac{n_1}{n_1+n_2}\binom{n_1+n_2}{n_1}$

Derhalve is

$$\mathcal{E}(\underline{u}|\underline{r} = r;H_o) = \frac{rn_1}{n_1+n_2}$$

Teneinde het spreidingskwadraat σ^2 te berekenen is het, zoals meestal bij discrete verdelingen, het eenvoudigst om eerst het tweede factoriële moment, dit is $\mathcal{E}(\underline{u}(\underline{u}-1)\,|\,\underline{r} = r;H_o)$ te berekenen. De berekening verloopt geheel op analoge wijze, daar voor een factor als $u(u-1)\binom{r}{u}$ geschreven kan worden

$$\frac{u(u-1)r(r-1)(r-2)!}{u(r-u)!} = r(r-1)\binom{r-2}{u-2}$$

voor zover althans $u(u-1)$ niet nul is.

Voor $\mathcal{E}(\underline{u}(\underline{u}-1))$ wordt zo gevonden: $\frac{r(r-1)\times n_1(n_1-1)}{(n_1+n_2)(n_1+n_2-1)}$

Met behulp van de betrekkingen

$$\mathcal{E}(\underline{u}^2) = \mathcal{E}(\underline{u}^2-\underline{u}) + \mathcal{E}(\underline{u}) \quad \text{en}$$

$$\sigma^2(\underline{u}|\underline{r}=r;H_o)=\mathcal{E}(\underline{u}-\mathcal{E}(\underline{u}))^2=\mathcal{E}(\underline{u}^2)-[\mathcal{E}(\underline{u})]^2$$

kan door eenvoudige algebraische berekeningen gevonden worden:

$$\sigma^2(\underline{u}|\underline{r}=r;H_o)=\frac{r(n_1+n_2-r)n_1n_2}{(n_1+n_2)^2(n_1+n_2-1)}$$

Voor $(n_1+n_2)\to\infty$ en $\sigma^2\to\infty$ is $\underline{u}$ asymptotisch normaal verdeeld met gemiddelde $\frac{rn_1}{n_1+n_2}$ en spreidingskwadraat σ^2. Dan geldt dus

$$P(\underline{u}=u|\underline{r}=r;H_o)\approx\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\tfrac{1}{2}\frac{(u-\frac{rn_1}{n_1+n_2})^2}{\sigma^2}\right].$$

Indien de exacte waarschijnlijkheden tengevolge van de grote waarde van n_1+n_2 te omslachtig te berekenen zijn, is deze benadering zeer practisch. Voor het bewijs van de asymptotische normaliteit zie b.v. D.van Dantzig [11] blz. 90-93 en hoofdstuk 6§ 4.

Ter verduidelijking van deze benaderingsmethode geven wij het volgende recept, tevens voorbeeld. Bereken bij de gegeven waarde van n_1 en n_2 en bij de gevonden waarde r van $\underline{r}$ en u van $\underline{u}$ de grootheden:

$$\mu=\mathcal{E}(\underline{u}|\underline{r}=r;H_o)=\frac{rn_1}{n_1+n_2}$$

en

$$\sigma^2=\sigma^2(\underline{u}|\underline{r}=r;H_o)=\frac{r(n_1+n_2-r)n_1n_2}{(n_1+n_2)^2(n_1+n_2-1)}\quad .$$

Voor het in § 4 genoemde voorbeeld ($n_1=8$, $n_2=12$, $r=11$, $u=1$) wordt dit: $\mu=\frac{11.8}{20}=4{,}4$; $\sigma^2=\frac{11.9.8.12}{20^2.19}=1{,}25$, dus $\sigma=1{,}1$.
Bereken vervolgens $\frac{|u-\mu|}{\sigma}$, dus in het voorbeeld: $\frac{|1-4{,}4|}{1{,}1}=3{,}1$.
Zoek bij dit getal in een tabel van de normale verdeling de (één- of tweezijdige, naar het geval zich voordoet) overschrijdingskans op. Deze is (tweezijdig) voor de waarde 3,1:k= 0,0026. De benadering kan nog verbeterd worden door toepassing van de z.g. continuïteitscorrectie,die in dit geval bestaat uit het berekenen van

$$\frac{|u-\mu|-\frac{1}{2}}{\sigma}$$

i.p.v. de bovengenoemde grootheid. De waarde hiervan wordt bij het genoemde voorbeeld 2,6 en de daarbij behorende tweezijdige overschrijdingskans is 0,009.

De exacte berekening in § 4 gaf 0,004. Dit doet vermoeden, dat de benadering (vooral met continuïteitscorrectie) over het algemeen wel betrouwbaar zal zijn. Dit is ook inderdaad het

geval, vooral als n_1 en n_2 ongeveer gelijk, of ongeveer $\frac{1}{2}(n_1 + n_2)$ is. Is dit echter niet het geval, dan is de benadering niet zo goed, terwijl de bovengevonden zeer nauwkeurige overeenstemming met de exacte toets als een uitzonderlijk gunstig resultaat bij zulke kleine aantallen moet worden beschouwd.

§ 6. Betrouwbaarheidsinterval voor de mediaan en voor het verschil van twee medianen.

6.1. Betrouwbaarheidsinterval voor de mediaan.

In deze § zal aangenomen worden, dat de verdelingsfuncties zowel onder de hypothese H_o als onder de alternatieve hypothese H continu zijn, zodat, behoudens een waarschijnlijkheid 0, twee waarnemingsuitkomsten steeds verschillend zijn.

In hoordstuk IV blz. 76 is reeds in een ander verband het betrouwbaarheidsinterval voor de mediaan behandeld. Wij zullen de bepaling ervan hier nog eens samenvatten als inleiding voor de bepaling van het betrouwbaarheidsinterval voor het verschil tussen twee medianen. Stel dat de mediaan van de verdeling van de stochastische grootheid $\underline{x}$ de waarde $M_{\underline{x}}$ heeft. Dan is per definitie van de mediaan de waarschijnlijkheid $\mathcal{P}[\underline{x} < M_{\underline{x}}] = \frac{1}{2}$. (Feitelijk luidt de definitie van de mediaan $\mathcal{P}[\underline{x} \leq M_{\underline{x}}] = \frac{1}{2}$, maar bij continu verdeelde grootheden is, behoudens een waarschijnlijkheid 0, $\mathcal{P}[\underline{x} < \underline{a}] = \mathcal{P}[\underline{x} \leq a]$. Met oog op de symmetrie van de verdere formules kiezen wij de definitie $\mathcal{P}[\underline{x} < M_{\underline{x}}] = \frac{1}{2}$, waarvoor wij, wederom behoudens een waarschijnlijkheid 0, ook mogen schrijven $\mathcal{P}[\underline{x} > M_{\underline{x}}] = \frac{1}{2}$.)

De waarschijnlijkheid, dat van n waarnemingen de waarde van de a-de waarneming, na rangschikking van hoog naar laag, groter is dan de mediaan en de (a+1)-ste kleiner is dan de mediaan, dus $x_a > M_{\underline{x}} > x_{a+1}$ bedraagt $\binom{n}{a}(\frac{1}{2})^n$. Elke waarneming toch heeft de kans $\frac{1}{2}$ om kleiner dan de mediaan te zijn. De kans, dat a van te voren aangewezen waarnemingen (in de nog niet gerangschikte reeks) groter zijn dan de mediaan en de (n-a) overigen kleiner bedraagt dus $(\frac{1}{2})^a(\frac{1}{2})^{n-a} = (\frac{1}{2})^n$. Uit de gehele reeks kunnen op $\binom{n}{a}$ wijze a waarnemingen gekozen worden. Dus de kans, dat er juist a groter en de (n-a) overigen kleiner dan de mediaan zijn bedraagt $\binom{n}{a}(\frac{1}{2})^n$. Dezelfde formule geldt natuurlijk voor de kans, dat de (n-a)-de waarneming kleiner en de (n-a+1)ste waarneming groter dan de mediaan is.

De kans, dat de a-de waarneming (in de naar grootte gerangschikte reeks) groter en de (n-a)-de kleiner dan de mediaan $M_{\underline{x}}$ is, bedraagt dus $\sum_{i=a}^{i=n-a} \binom{n}{i}(\frac{1}{2})^n = 1 - 2^{-n+1} \sum_{i=0}^{i=a} \binom{n}{i}(\frac{1}{2})^n$. Met andere

woorden

$$\mathcal{P}\left[\underline{x}_a > M_{\underline{x}} > \underline{x}_{n-a+1}\right] = 1 - 2^{-n+1} \sum_{i=0}^{i=a} \binom{n}{i} (\tfrac{1}{2})^n .$$

Wordt bij gegeven n het rangnummer a zodanig gekozen, dat $1 - 2^{-n+1} \sum_{i=0}^{i=a} \binom{n}{i} (\tfrac{1}{2})^n \leqq \alpha$ is, dan wordt het interval van $\underline{x}_a$ tot $\underline{x}_{n-a+1}$ een betrouwbaarheidsinterval met onbetrouwbaarheid α genoemd. (Voor de keuze van a bij gegeven n kan gebruikt gemaakt worden van de tabel van A. VAN WIJNGAARDEN [12], welke in verkorte vorm als bijlage III van hoofdstuk IV is opgenomen).

De aandacht moge erop gevestigd worden, dat $\mathcal{P}[\underline{x}_a > M_{\underline{x}} > \underline{x}_{n-a+1}] \leqq \alpha$ geen waarschijnlijkheidsuitspraak over de grootte van $M_{\underline{x}}$ is. $M_{\underline{x}}$ is geen stochastische grootheid ! De onderstreping onder de index $\underline{x}$ wijst er slechts op, dat $\underline{x}$ een stochastische grootheid is, terwijl $M_{\underline{x}}$ een parameter van de verdeling van $\underline{x}$ is. Indien $M_{\underline{x}}$ stochastisch was, zou de M onderstreept zijn. Het is daarentegen een waarschijnlijkheidsuitspraak over de uitkomst van een experiment: de kans dat de a-de waarneming (van boven) groter en de (n-a+1)-ste waarneming (van boven) kleiner dan de mediaan $M_{\underline{x}}$ is, is hoogstens gelijk aan α . Indien de experimentator dus steeds aanneemt, dat de mediaan $M_{\underline{x}}$ tussen de a-de waarneming van boven en de a-de waarneming van beneden ligt, zal hij hoogstens in ongeveer in een fractie α van de gevallen, waarin hij dit toepast, een foutieve uitspraak doen.

6.2. Betrouwbaarheidsinterval voor het verschil van twee medianen.

Er bestaat nog geen methode, om een betrouwbaarheidsinterval met gegeven onbetrouwbaarheid α voor het verschil van de medianen van twee stochastische grootheden $\underline{x}$ en $\underline{y}$ met willekeurige verschillende verdelingsfuncties F(x) en G(y) te bepalen. Indien echter de (continue) verdelingsfunctie van x gelijk is aan F(x) en die van y gelijk is aan F(y+d), dan is het wel mogelijk om een betrouwbaarheidsinterval met gegeven onbetrouwbaarheid α voor het verschil der medianen $M_{\underline{x}}$ en $M_{\underline{y}}$ te vinden. Wij zullen dit dus verder aannemen en deze onderstelling met het symbool H aangeven. Uit $\mathcal{P}[\underline{x} \leqq M_{\underline{x}}] = F(M_{\underline{x}}) = \frac{1}{2}$ en $\mathcal{P}[\underline{y} \leqq M_{\underline{y}}] = F(M_{\underline{y}}+d) = \frac{1}{2}$ volgt $M_{\underline{x}} = M_{\underline{y}}+d$ of wel $M_{\underline{x}} - M_{\underline{y}} = d$, zodat d het verschil tussen de mediaan van de verdeling van $\underline{x}$ en die van de verdeling van $\underline{y}$ is.

Beschouwen wij nu de grootheden $\underline{x}' = \underline{x} - M_{\underline{x}}$ en $\underline{y}' = \underline{y} - M_{\underline{y}}$. Het is duidelijk, dat deze grootheden dezelfde verdeling $F_1(x')$ resp. $F_1(y')$ bezitten en dat hun mediaan gelijk is aan nul.

($F_1(x') = F(x'+M_{\underline{x}})$, $F_1(y') = F(y'+M_{\underline{y}}+d) = F(y'+M_{\underline{x}})$).

Aangezien $M_{\underline{x}}$ en $M_{\underline{y}}$ onbekend zijn, zijn de grootheden x'_i en y'_i, die waarnemingen zijn van de stochastische grootheden $\underline{x}'$ en $\underline{y}'$ ook onbekend; bekend zijn slechts de grootheden x_i en y_i die waarnemingen zijn van $\underline{x}$ en $\underline{y}$. Wel kan berekend worden hoe groot de waarschijnlijkheid is, dat in de naar grootte gerangschikte reeks van de verzameling der waarnemingen x'_i en y'_i de u-de waarneming van $\underline{x}'$ dus x'_u groter is dan de v-de waarneming van $\underline{y}'$ dus y_v en tegelijkertijd $x'_{u'}$ kleiner dan $y'_{v'}$, dus $\mathcal{P}\left[\underline{x}'_u > \underline{y}'_v \text{ en } \underline{x}'_{u'} < \underline{y}'_{v'} \middle| H_o\right]$, althans onder de bovengenoemde aanname H, dat $\underline{x}'$ en $\underline{y}'$ een gelijke verdeling hebben. Wij kunnen deze vergelijking ook anders schrijven, immers

$$\mathcal{P}\left[\underline{x}'_u > \underline{y}'_v \middle| H_o\right] = \mathcal{P}\left[\underline{x}_u - M_{\underline{x}} > \underline{y}_v - M_{\underline{y}} \middle| H_o\right] = \mathcal{P}\left[\underline{x}_u - \underline{y}_v > M_{\underline{x}} - M_{\underline{y}} \middle| H_o\right]$$

$$\mathcal{P}\left[\underline{x}'_{u'} < \underline{y}'_{v'} \middle| H_o\right] = \mathcal{P}\left[\underline{x}_{u'} - M_{\underline{x}} < \underline{y}_{v'} - M_{\underline{y}} \middle| H_o\right] = \mathcal{P}\left[\underline{x}_{u'} - \underline{y}_{v'} < M_{\underline{x}} - M_y \middle| H_o\right] \text{ dus}$$

$$\mathcal{P}\left[\underline{x}'_u > \underline{y}'_v \text{ en } \underline{x}'_{u'} < \underline{y}'_{v'} \middle| H_o\right] = \mathcal{P}\left[\underline{x}_u - \underline{y}_v > M_{\underline{x}} - M_{\underline{y}} > \underline{x}_{u'} - \underline{y}_{v'} \middle| H_o\right].$$

Dit betekent echter, dat als u en v zowel als u' en v' zodanig gekozen worden, dat $\mathcal{P}\left[\underline{x}'_u > \underline{y}'_v \text{ en } \underline{x}'_{u'} < \underline{y}'_{v'} \middle| H_o\right] \leqq \alpha$, de grenzen van een betrouwbaarheidsinterval voor het verschil der medianen van $\underline{x}$ en $\underline{y}$ n.l. $d = M_{\underline{x}} - M_{\underline{y}}$ gegeven wordt door de grenzen $x_u - y_v$ en $x_{u'} - y_{v'}$.

Wij zullen thans de mogelijke keuzen van u, v, u' en v' beperken. Aangezien $x_u - y_v$ een bovengrens van het betrouwbaarheidsinterval voor het verschil van twee medianen is, is het redelijk om $u < \frac{1}{2}n_1$ en v groter dan $\frac{1}{2}n_2$ te nemen. Dit betekent immers, dat wij als bovengrens van $M_{\underline{x}} - M_{\underline{y}}$ nemen het verschil tussen een waarneming groter dan de (empirische) mediaan der waarnemingsreeks x en een waarneming kleiner dan de (empirische) mediaan der waarnemingsreeks y (de reeksen zijn van hoog naar laag geordend). Om dezelfde redenen is het redelijk om $u' > \frac{1}{2}n_1$ en $v' < \frac{1}{2}n_2$ te kiezen. Uit symmetrieoverwegingen zullen voorts deze grootheden dusdanig gekozen worden, dat $u + u' = n_1 + 1$ en $v + v' = n_2 + 1$.

Voor de berekening van $\mathcal{P}\left[\underline{x}'_u > \underline{y}'_v \text{ en } \underline{x}'_{u'} < \underline{y}'_{v'}\right]$ zal wederom gebruik gemaakt worden van de grafische voorstelling, die wij nu getekend denken voor de reeksen x' en y' en waarin wij dus de bij de steekproef behorende weg niet werkelijk kunnen aangeven, daar wij wel de waarden van x en y, maar niet die van x' en y' kennen. In dit schema hebben weer alle mogelijke wegen gelijke waarschijnlijkheden, wegens de aanname H. In plaats van een verzameling wegen als kritieke zône te kiezen, zullen wij nu de wegen van een dergelijke verzameling uitsluiten, d.w.z. aannemen, dat de bij de steekproef behorende (onbekende) weg niet één van deze verzameling van wegen is. Kiezen wij deze verzameling zodanig, dat de kans op één van deze wegen $\leqq \alpha$ is, dan bezit onze bewering, dat de weg, behorend bij de steekproef, niet één van de uitgesloten verzameling is, een onbetrouwbaarheid $\leqq \alpha$. Op grond van deze redenering kan dan een be-

trouwbaarheidsinterval voor d bepaald worden.

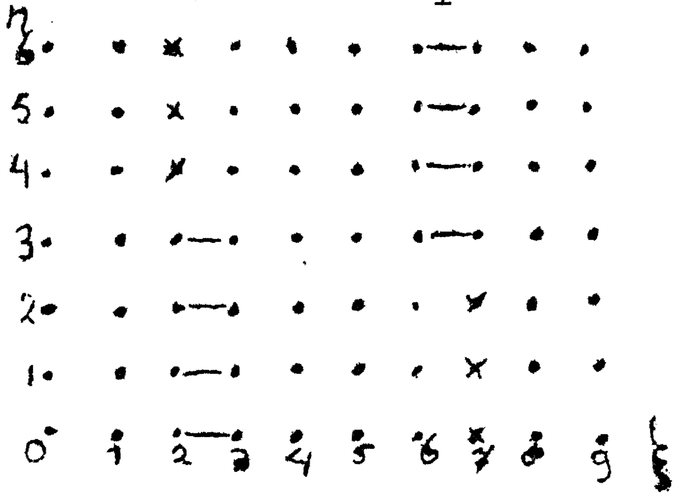


Fig. 3.

Grafische voorstelling voor de berekening van $\mathcal{P}\left[\underline{x}'_u > \underline{y}'_v \text{ en } \underline{x}'_{u'} < \underline{y}'_{v'}\right]$ met u=3 en v=4, u'=7 en v'=3 bij $n_1=9$ en $n_2=6$.

In de figuur zijn aangegeven 2 x 4 horizontale lijnen, waarvan elke weg, die een groepering van de grootheden x' en y' voorstelt, waarvoor geldt $x'_u > y'_v$ en $x'_{u'} < y'_{v'}$, er twee moet bevatten. Dit betekent, dat elke weg, die door één der punten met × aangegeven een groepering voorstelt, waarvoor niet geldt $x'_u > y'_v$ en $x'_{u'} < y'_{v'}$. Dus, m.a.w.

$$\mathcal{P}\left[\underline{x}'_u > \underline{y}'_v \text{ en } \underline{x}'_{u'} < \underline{y}'_{v'}\right] = 1 - \frac{\text{Aantal wegen door een } \times.}{\binom{n_1 + n_2}{n_1}}.$$

Uit de beperkingen, die aan de grootheden u, u', v en v' opgelegd zijn, volgt in de eerste plaats, dat een verboden weg, hetzij één (of meer) der punten × met $\xi = 2$, hetzij één (of meer) der punten met $\xi = 7$ passeert.

De eerste reeks wegen (die een punt × met $\xi = 2$ passeren) karakteriseert groeperingen, waarvoor geldt $x'_u < y'_v$, de tweede reeks (die een punt × met $\xi = 7$ passeren) groeperingen, waarvoor geldt $x'_{u'} > y'_{v'}$. Door de beperkingen, die wij aan u, u', v en v' oplegden, geldt dus:

$$\mathcal{P}\left[\underline{x}'_u > \underline{y}'_v \text{ en } \underline{x}'_{u'} < \underline{y}'_{v'}\right] = 1 - \mathcal{P}\left[\underline{x}'_u < \underline{y}'_v\right] - \mathcal{P}\left[\underline{x}'_{u'} > \underline{y}'_{v'}\right].$$

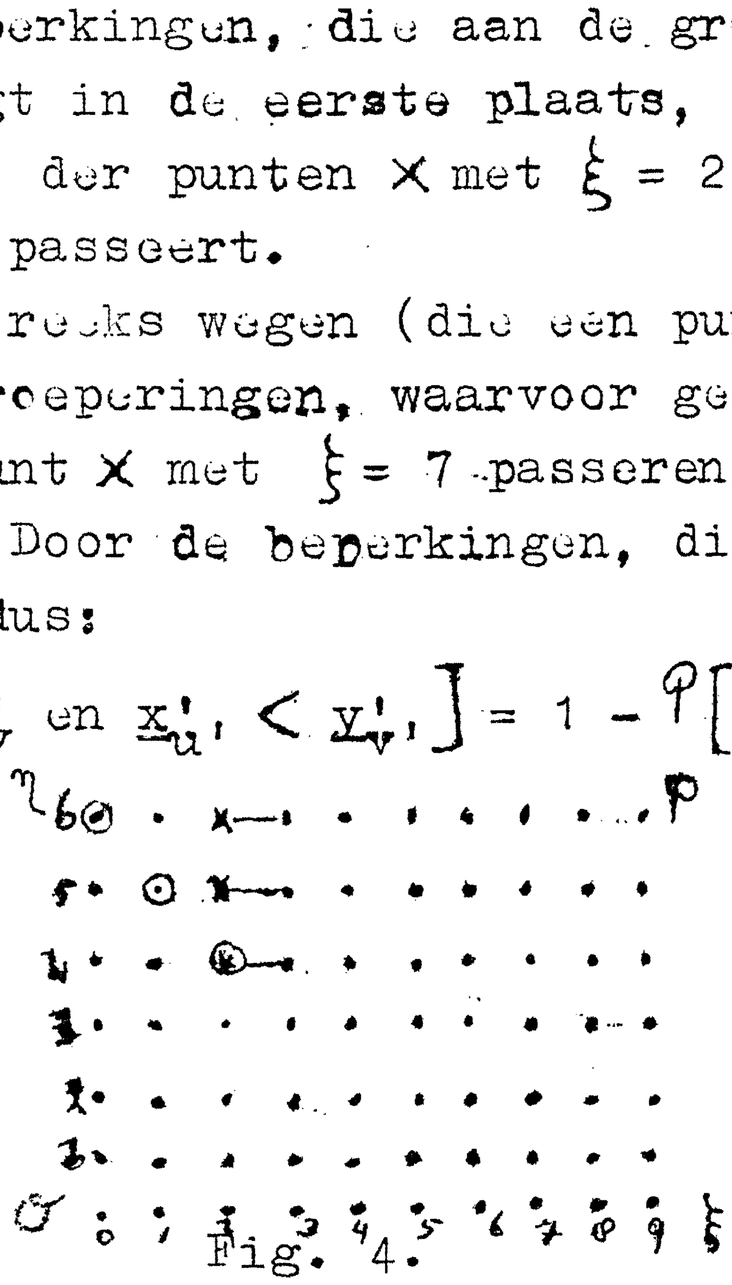


Fig. 4.

Grafische voorstelling voor de berekening van $\mathcal{P}\left[\underline{x}'_u < \underline{y}'_v\right]$ met u=3, v=4.

Alle wegen door een punt × moeten langs een der aangegeven lijnen van $\xi = 2$ naar $\xi = 3$ gaan. Het aantal A dezer wegen is:

$$A = \sum_{i=v}^{i=n_2} \binom{u-1+i}{u-1}\binom{n_1-u+n_2-i}{n_1-u}.$$

De eerste binomiaalcoëfficiënt geeft het aantal wegen aan van $\mathcal{O}$ naar een punt X met coördinaten $\xi = u-1$, $\eta = i$, de tweede het aantal wegen van dat punt X langs het getrokken horizontale lijnstuk naar $\mathcal{P}$.

Uit de grafische voorstelling volgt echter, dat voor dit aantal wegen ook nog een andere uitdrukking gevonden kan worden. Elke weg, die door een punt X gaat, gaat n.l. ook door één der punten $\odot$. Het aantal wegen A door een punt X is dus ook

$$A = \sum_{j=\alpha}^{j=u-1} \binom{u+v-1}{j}\binom{n_1+n_2-u-v+1}{n_1-j} \quad \text{waarbij} \quad \alpha = \max(0, u+v-1-n_2).$$

Wij vinden dus

$$P\left[\underline{x}'_u < \underline{y}'_v\right] = \frac{A}{\binom{n_1+n_2}{n_1}} = \frac{\sum_{i=v}^{i=n_2}\binom{u-1+i}{u}\binom{n_1+n_2-u-i}{n_1-u}}{\binom{n_1+n_2}{n_1}} = \frac{\sum_{j=\alpha}^{j=u-1}\binom{u+v-1}{j}\binom{n_1+n_2-u-v+1}{n_1-j}}{\binom{n_1+n_2}{n_1}}.$$

In MOOD [4] § 16.4 wordt de bepaling van het betrouwbaarheidsinterval van de mediaan behandeld, zij het niet met behulp van de grafische voorstelling. MOOD geeft alleen de eerste formule voor A (zie boven).

De bepaling van de rangorde cijfers u en v kan nog op verschillende wijzen geschieden. Zij zijn slechts aan de voorwaarde gebonden dat

$P[\underline{x}_u < \underline{y}_v] + P[\underline{x}_{u'} > \underline{y}_{v'}] \leqq \alpha$, waarbij $u' = n_1+1-u$ en $v' = n_2+1-v$.

Voor een bepaalde waarde van v (bij gegeven n_1 en n_2) is dus u door bovenstaande betrekking bepaald. Bij verschillende waarden van v zal u echter ook verschillende waarden aannemen. Een redelijke keus, die echter bij grotere steekproeven vermoedelijk niet het kortste betrouwbaarheidsinterval geeft, is die waarbij $u+v = \frac{n_1+n_2+1}{2}$ resp. $\frac{n_1+n_2}{2}$, dat is dus de mediaan van het gehele materiaal.

In MOOD[4] § 16.4 wordt aangegeven, dat, voor grote waarden van n_1 en n_2, waarbij dus de normale verdeling als benadering van de hypergeometrische verdeling gebruikt mag worden, de beste keus voor u en v is:

$$u \sim \frac{n_1}{2} - \frac{c\sqrt{n_1}\sqrt{n_1+n_2}}{2(\sqrt{n_1}+\sqrt{n_2})}$$

$$v \sim \frac{n_2}{2} + \frac{c\sqrt{n_2}\sqrt{n_1+n_2}}{2(\sqrt{n_1}+\sqrt{n_2})}$$

waarbij c aan de betrekking $\int_{-\infty}^{c} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \frac{1}{2}\alpha$ voldoet.

Tenslotte moge nog gewezen worden op het verband, dat bestaat tussen de mediaantoets en het betrouwbaarheidsinterval voor de mediaan. Dit verband is niet geheel toevallig. Uit een toets kan steeds een betrouwbaarheidsinterval afgeleid worden en omgekeerd. Toch is hier het verband niet zo volkomen als gewoonlijk, omdat de uitgangspunten niet geheel gelijk zijn.

Indien in de laatste vorm voor $P[\underline{x}'_u < \underline{y}'_v]$ n.l.

$$\frac{\sum_j \binom{u+v-1}{j}\binom{n_1+n_2-u-v+1}{n_1-j}}{\binom{n_1+n_2}{n_1}}$$

de grootheid u+v-1 vervangen wordt door r, dan herkennen wij hierin de uitdrukking, die bij de (eenzijdige) mediaantoets gevonden wordt voor de waarschijnlijkheid dat, onder de hypothese H_o van gelijkheid der verdelingen van $\underline{x}'$ en $\underline{y}'$ bij de r grootste waarnemingen van de gezamenlijke reeks der waarnemingen x en y, u-1 waarnemingen x of minder voorkomen. Voor $P[\underline{x}'_{u'} > \underline{y}'_{v'}]$ wordt een overeenkomstige uitdrukking gevonden, echter met in het algemeen een andere waarde voor r n.l. u'+v'-1. Wanneer u+v-1 = u'+v'-1 dan is de overeenstemming echter volkomen. Daar wij gesteld hebben dat $u+u' = n_1+1$ en $v+v' = n_2+1$, dus $u+v-1+u'+v'-1 = n_1+n_2$ wordt een volledige overeenstemming met de formule van de toets van Hemelrijk verkregen, indien $r = u+v-1 = \frac{n_1+n_2}{2}$. In dit geval is er echter ook het bovenbedoelde verband tussen toets en betrouwbaarheidsinterval:

$P[\underline{x}'_u < \underline{y}'_v]$ betekent, dat er onder de u+v-1 = r grootste waarnemingen van de gehele serie ten hoogste u-1 waarnemingen x voorkomen (zie fig. 4), analoog voor $P[\underline{x}'_{u'} > \underline{y}'_{v'}]$.

De hypothese H_o van de gelijkheid der medianen van de verdelingen van $\underline{x}'$ en $\underline{y}'$ wordt verworpen als er onder de r grootste waarnemingen der serie u-1 of minder waarnemingen x voorkomen, resp. v'-1 of minder waarnemingen y, ofwel als $\underline{x}'_u < \underline{y}'_v$ dan wel $\underline{x}'_{u'} > \underline{y}'_{v'}$, ofwel als $\underline{x}'_u - \underline{y}'_v < 0$ dan wel $\underline{x}'_{u'} - \underline{y}'_{v'} > 0$. Daar

$$\underline{x}'_u - \underline{y}'_v = \underline{x}_u - M_{\underline{x}} - (\underline{y}_v - M_y) = (\underline{x}_u - \underline{y}_v) - (M_{\underline{x}} - M_{\underline{y}}) \text{ en}$$

$$\underline{x}'_{u'} - \underline{y}'_{v'} = \underline{x}_{u'} - M_{\underline{x}} - (\underline{y}_{v'} - M_{\underline{y}}) = (\underline{x}_{u'} - \underline{y}_{v'}) - (M_{\underline{x}} - M_{\underline{y}})$$

betekent dit, dat indien $(\underline{x}_u - \underline{y}_v) < (M_{\underline{x}} - M_{\underline{y}})$ resp. $(\underline{x}_{u'} - \underline{y}_{v'}) > M_{\underline{x}} - M_{\underline{y}}$ geldt, de hypothese H_o van gelijkheid der medianen van $\underline{x}'$ en $\underline{y}'$ ofwel van een verschil $d = M_x - M_y$ tussen de medianen van $\underline{x}$ en $\underline{y}$ verworpen wordt. De grootheid u-1 kan, op de bij de bespreking van de toets van Hemelrijk aangegeven methode dusdanig gekozen worden, dat de waarschijnlijkheid dat, als de hypothese H_o juist is, deze

toch verworpen wordt $\leq \alpha$ bedraagt. Kiezen wij als betrouwbaarheidsinterval $(\underline{x}_u - \underline{y}_v) > (M_{\underline{x}} - M_{\underline{y}}) > (\underline{x}_{u'} - \underline{y}_{v'})$ met onbetrouwbaarheid $\leq \alpha$, dan bestaat er een kans $\leq \alpha$ dat het experiment dusdanig uitvalt, dat $(\underline{x}_u - \underline{y}_v) < (M_{\underline{x}} - M_{\underline{y}})$ ofwel $(\underline{x}_{u'} - \underline{y}_{v'}) > (M_{\underline{x}} - M_{\underline{y}})$. Dat wil dus zeggen dat als elke waarneming y vervangen wordt door $y'' = y-d$ en vervolgens de toets van Hemelrijk toegepast wordt, de hypothese H_o van de gelijkheid der verdelingen van $\underline{x}$ en $\underline{y}''$ niet verworpen wordt zolang d ligt tussen de grenzen $(\underline{x}_u - \underline{y}_v)$ en $(\underline{x}_{u'} - \underline{y}_{v'})$, en wel verworpen wordt zodra d buiten deze grenzen valt.

Tenslotte moge hier nog gewezen worden op een artikel van WALSH [13], waarin deze volgens een andere theorie dan de hier behandelde een betrouwbaarheidsinterval voor één mediaan geeft benevens een daarmee samenhangende toets om na te gaan of de mediaan van een verdeling verschilt van een van te voren gegeven getal. Daarbij wordt als gegeven verondersteld, dat de waarschijnlijkheidsverdelingen der stochastische grootheden symmetrisch zijn.

Ook hier moge een rekenvoorbeeld gegeven worden voor de bepaling van het betrouwbaarheidsinterval voor het verschil van twee medianen, waarbij aangesloten wordt op het getallenvoorbeeld van blz. 119. Hierbij is $n_1 = 8$, $n_2 = 6$. Voor een betrouwbaarheidsinterval met onbetrouwbaarheid 0,05 moet met behulp van een tabel van de normale verdeling c bepaald worden uit: $\int_{-c}^{c} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt = \frac{1}{2} \times 0{,}05$.

Gevonden wordt $c = 1{,}96$. Dus:

$$u' = \frac{8}{2} - \frac{1{,}96\sqrt{8}\sqrt{14}}{2(\sqrt{8}+\sqrt{6})} = 4 - 1{,}97 \sim 2 \qquad u' = 8+1-2 = 7$$

$$v = \frac{6}{2} + \frac{1{,}96\sqrt{6}\sqrt{14}}{2(\sqrt{8}+\sqrt{6})} = 3 + 1{,}70 \sim 5 \qquad v' = 6+1-5 = 2.$$

Als grenzen van het betrouwbaarheidsinterval met onbetrouwbaarheid $\sim 0{,}05$ voor het verschil $d = M_{\underline{x}} - M_{\underline{y}}$ wordt dus gevonden

$x_2 - y_5 = 2{,}60 - 2{,}93 = -0{,}33$ en

$x_7 - y_2 = 1{,}68 - 6{,}43 = -4{,}75$.

Een directe berekening van de onbetrouwbaarheid is hier nog nodig, daar wij hier de normale benadering op wel zeer kleine aantallen toepasten.

$$\alpha = \frac{\sum_{i=0}^{i=2-1} \binom{2+5-1}{i}\binom{14-2-5+1}{8-i} + \sum_{i=7}^{i=8} \binom{2+7-1}{i}\binom{14-2-7+1}{8-i}}{\binom{14}{8}} = \frac{98}{3003} = 0{,}033.$$

Wij vinden dus met onbetrouwbaarheid 0,033

$$-0{,}33 > M_{\underline{x}} - M_{\underline{y}} > -4{,}75 \text{ ofwel}$$

$$0{,}33 < M_{\underline{y}} - M_{\underline{x}} < 4{,}75.$$

Aangezien het betrouwbaarheidsinterval nul niet bevat, wordt tegelijk met overschrijdingskans 0,033 de hypothese H_o verworpem, dat serie I en serie III van pag. 119 een gelijke verdeling hebben.

§ 7. De toets van Smirnov.

Zoals uit het in § 1 opgemerkte valt af te leiden, is de toets van Smirnov in het bijzonder geschikt om de hypothese H_o te toetsen tegen alternatieve hypothesen H, die vrijwel aan geen beperkingen zijn gebonden. Alleen zullen wij zowel aan de hypothese H_o als aan de alternatieve hypothesen H de beperking opleggen, dat de verdelingsfunctie continu is.

Onder een alternatieve hypothese H zijn de uitkomsten $x_1,\ldots,x_{n_1}$ dus waarnemingen van onderling onafhankelijk verdeelde stochastische grootheden $\underline{x}_1,\ldots,\underline{x}_{n_1}$, die allen eenzelfde verdeling bezitten, en zijn de uitkomsten $y_1,\ldots,y_{n_2}$ waarnemingen van onderling onafhankelijk verdeelde stochastische grootheden $\underline{y}_1,\ldots,\underline{y}_{n_2}$, die ook allen eenzelfde verdeling bezitten, die evenwel in enig opzicht verschilt van die van $\underline{x}$. Dit verschil kan b.v. bestaan in een verschil in niveau, in welk geval dus ook de mediaantoets of de toets van Wilcoxon gebruikt kan worden, maar ook in een verschil in spreiding, of in een verschil in scheefheid. In de beide laatstgenoemde gevallen hebben de mediaantoets en de toets van Wilcoxon vermoedelijk beide een zeer gering onderscheidingsvermogen.

Als toetsingsgrootheid wordt hier gebruikt de maximumwaarde d van het verschil $|F_1(u) - F_2(u)|$ waarin $F_1(u)$ resp. $F_2(u)$ de empirische verdelingsfuncties van $\underline{x}$ en $\underline{y}$ zijn, d.w.z. $F_1(u) = \frac{i}{n_1}$ als $x_i \leqq u < x_{i+1}$ is, waarbij wij de waarnemingen in tegenstelling met het voorafgaande van klein naar groot gerangschikt denken. De definitie voor $F_2(u)$ luidt natuurlijk $F_2(u) = \frac{j}{n_2}$ als $y_j \leqq u < y_{j+1}$.

Zijn b.v. de waarden van $x_1,\ldots,x_{n_1}$ en $y_1,\ldots,y_{n_2}$ en van hun empirische verdelingsfuncties $F_1(u)$ en $F_2(u)$ respectievelijk:

	x_1	x_2	x_3			x_4		x_5	
	2,0	2,3	2,4			2,8		3,1	
$F_1(u)$	0,20	0,40	0,60			0,80		1,00	
$F_2(u)$				0,25	0,50		0,75		1,00
				2,6	2,7		3,0		3,3
				y_1	y_2		y_3		y_4

dan is $|F_1(u) - F_2(u)|$ maximum in het interval $2,4 \leqq u < 2,6$. Deze maximumwaarde bedraagt hier 0,60 onze statistische grootheid $d = 0,60$. De onderstaande figuur moge dit nog toelichten.

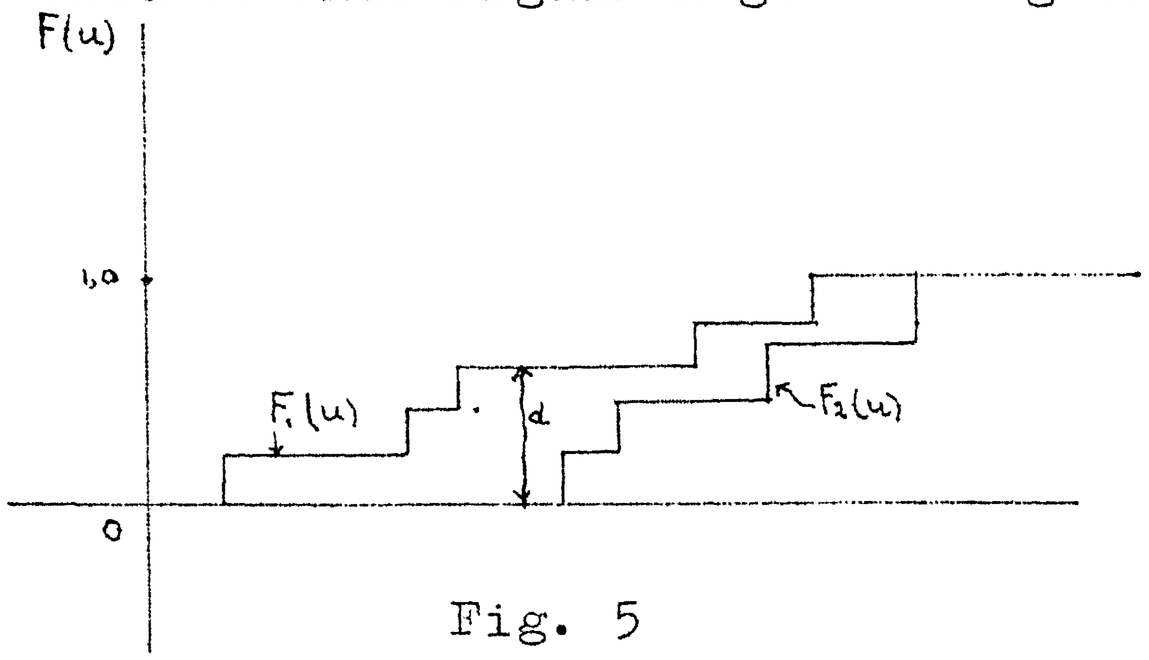


Fig. 5
Grafische voorstelling van de bepaling van
$d = \max |F_1(u) - F_2(u)|$.

Het is duidelijk, dat als de hypothese H_o juist is, de kans dat d een grote waarde zal hebben, gering is, terwijl daarentegen als de verdelingsfuncties van de stochastische grootheden $\underline{x}$ en $\underline{y}$ verschillen, $F_1(u)$ en $F_2(u)$ voor een of andere waarde van u ook een duidelijk verschil zullen vertonen.

Voorts zal d onder de hypothese H_o des te kleiner zijn naarmate de steekproeven (n_1 en n_2) groter zijn, daar dan, behoudens een kleine waarschijnlijkheid, de empirische verdelingen $F_1(u)$ en $F_2(u)$ tot de ware verdeling van $\underline{x}$ en $\underline{y}$ zullen naderen. Het zal blijken dat, onder de hypothese H_o, de grootheid d van de orde van grootte $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ is.

Wij zullen nu voor het geval de beide steekproeven even groot zijn ($n_1 = n_2 = n$) de waarde van $P[\max|F_1(u) - F_2(u)| \geqq d|H_o]$ exact berekenen. Daarna zal de asymptotische waarde van deze waarschijnlijkheid voor $n \to \infty$ afgeleid worden. Deze asymptotische formule geldt ook (met een kleine wijziging) voor ongelijke steekproeven, doch het bewijs hiervan zal hier niet gegeven worden.

§ 8. Berekening van $P[\max|F_1(u) - F_2(u)| \geqq d|H_o]$ voor twee even grote steekproeven.

Aangezien de steekproeven even groot zijn, hebben beide empirische verdelingsfuncties voor iedere u de vorm $F(u) = \frac{i}{n}$ $(i=0,1,\ldots,n)$. Het verschil $|F_1(u) - F_2(u)|$ heeft dus ook de vorm $\frac{h}{n}$ $(h=0,1,\ldots,n)$. Derhalve is $P[\max|F_1(u) - F_2(u)| \geqq d|H_o]$ constant voor $\frac{h-1}{n} < d \leqq \frac{h}{n}$. Deze waarschijnlijkheid behoeft dus alleen berekend te worden voor waarden van $d = \frac{h}{n}$, waarbij h een geheel getal

$\leqq n$ is, zodat geldt: $P\left[\max|F_1(u) - F_2(u)| \geqq d|H_o\right] =$
$= P\left[\max|nF_1(u) - nF_2(u)| \geqq h|H_o\right]$ (h geheel getal $\leqq n$). Nu is in de naar grootte gerangschikte reeks der waarnemingen $x_1, \ldots, x_n$ en $y_1, \ldots, y_n$ de waarde van $nF_1(x_i) = i$ en de waarde van $nF_2(y_j) = j$, zodat, wanneer $y_{j-1} < x_i < y_j < y_{j+k} < x_{i+1}$ ($i = 1, \ldots, n-1$, $j = 2, \ldots, n-k$, $k = 0, \ldots, n-2$), met evidente veranderingen als $i = 0$ of $i = n$ resp. $j = 1$, voor waarden van u in het interval $x_i \leqq u < y_j$ geldt:

$$nF_1(u) - nF_2(u) = i - (j-1),$$

en voor waarden van u in het interval $y_{j+h} \leqq u < y_{j+h+1}$ ($h = 0, \ldots, k-1$):

$$nF_1(u) - nF_2(u) = i - (j+h).$$

Worden de waarnemingen $x_1, \ldots, x_n$ en $y_1, \ldots, y_n$ in één reeks van klein naar groot gerangschikt, dan wordt $\max|nF_1(u) - nF_2(u)|$ dus gevonden door voor alle plaatsen, waar hetzij een x door een y gevolgd wordt, of een y door een x gevolgd wordt, de absolute waarde van het verschil der rangnummers te bepalen en onder deze getallen de grootste uit te zoeken.

In de grafische voorstelling wordt dit grootste verschil gevonden door dat punt van de weg door het rooster van O naar P op te zoeken, dat in verticale richting het verst van de diagonaal OP ligt (en dus ook in horizontale richting, daar de diagonaal OP een hoek van 45° met de zijden van het vierkant OQPR maakt). Voor alle wegen, die door een punt op één der lijnen $\eta = \xi \pm h$ gaan, geldt dat $\max|nF_1(u) - nF_2(u)| \geqq h$, zoals uit het voorafgaande volgt. (zie fig. 6).

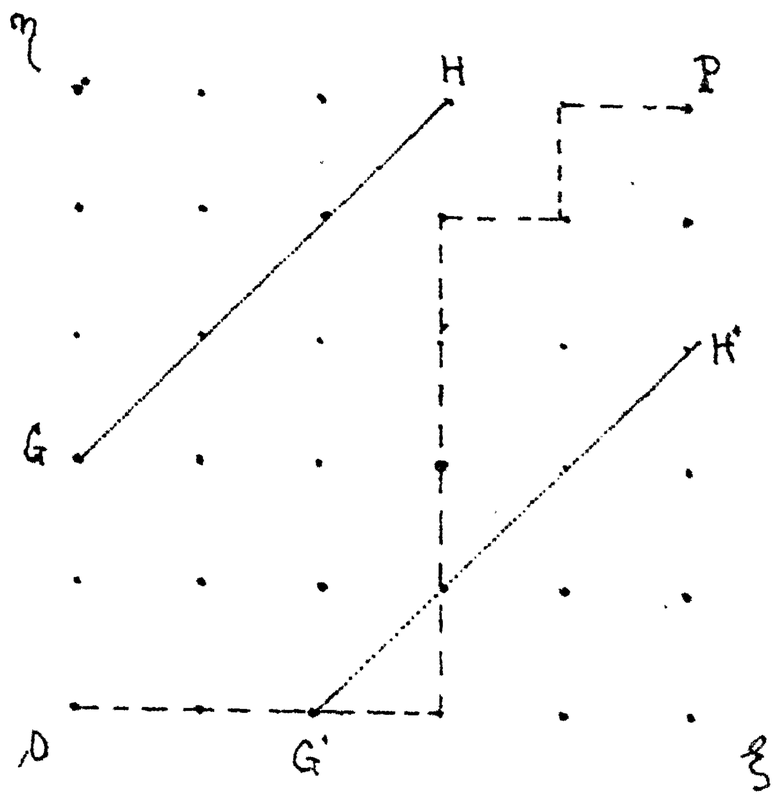


Fig. 6

Voorbeeld van een weg, die een paar stookproeven karakteriseert, waarvoor geldt $\max|nF_1(u) - nF_2(u)| \geqq h$ of wel $\max|F_1(u) - F_2(u)| \geqq d$ met $n = 5$, $h = 2$, $d = \frac{h}{n} = \frac{2}{5}$. De lijn GH heeft als vergelijking $\eta = \xi + 2$, de lijn G'H': $\eta = \xi - 2$.

Om de gevraagde waarschijnlijkheid P te vinden, moet dus bepaald worden het aantal wegen van O naar P, welke hetzij één der lijnen $\eta = \xi \pm h$, hetzij beide lijnen, bereiken of doorkruisen. P is dan gelijk aan dit aantal, gedeeld door het totaal aantal wegen van O naar P, nl. $\binom{2n}{n}$.

Wij zullen dit probleem in twee phasen oplossen. Eerst zal het algemene probleem behandeld worden van het aantal wegen van O naar P in een rechthoekig rooster, met zijden $n_1 + 1$ en $n_2 + 1$, dat de lijn GH met vergelijking $\eta = \xi + h$ bereikt of doorkruist, vervolgens zal, met behulp van de oplossing van dit probleem ons eigenlijke vraagstuk opgelost worden. De moeilijkheid hierbij blijkt te liggen in het feit, dat er, voor waarden van $h < \frac{1}{2}n$, wegen zijn, die zowel GH als G'H' bereiken of doorkruisen.

§ 9. Bepaling van het aantal wegen, dat de lijn $\eta = \xi + h$ bereikt of doorkruist.

Indien een weg van O naar P in een rooster met zijden n_1+1 en n_2+1 de lijn GH met vergelijking $\eta = \xi + h$ bereikt of doorkruist, dan betekent dit, dat ergens in de gezamenlijke reeks der naar grootte gerangschikte waarnemingen x en y het aantal letters y, geteld vanaf het begin, het aantal letters x met tenminste h overtreft. Ons probleem is nauw verwant met het probleem uit de klassieke waarschijnlijkheidsrekening van de ruïnering van spelers: indien de letters x een gewonnen partij en de letters y een verloren partij betekenen en de inzet steeds een gulden is, dan is een persoon met een (speel)kapitaal van h gulden geruïneerd, zodra in de serie van gewonnen en verloren partijen het aantal verloren partijen (letters y) het aantal gewonnen partijen (letters x) met h overtreft. Onder deze inkleding is door WHITWORTH het probleem langs algebraische weg opgelost (WHITWORTH [14], Proposition 39). Wij zullen hier de oplossing uit de figuur langs een eenvoudige weg afleiden.

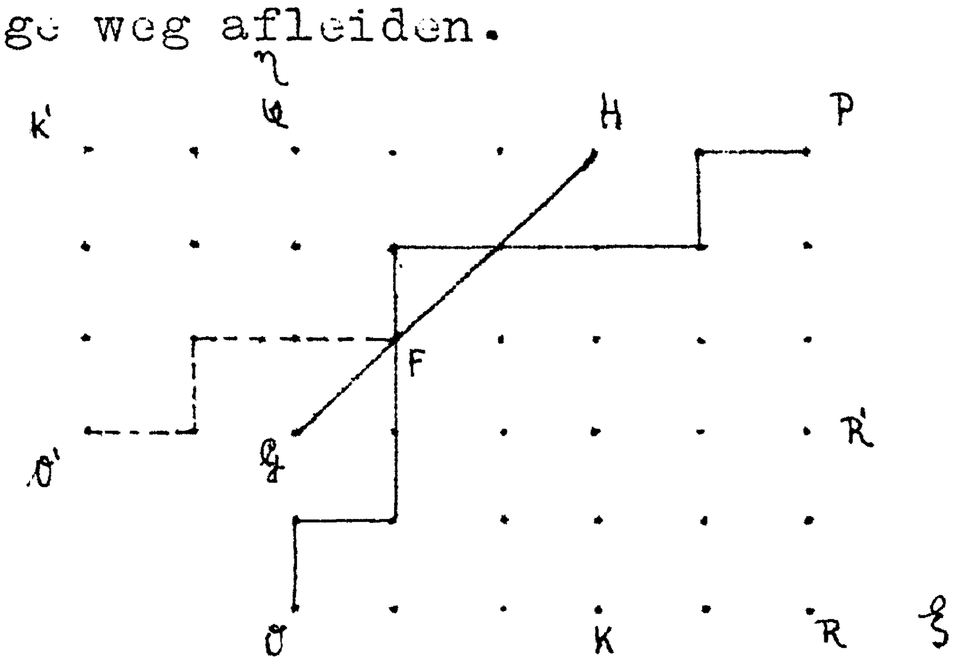


Fig. 7

Bepaling van het aantal wegen van O naar P, dat de lijn GH bereikt of doorkruist.

Elke weg, die althans GH bereikt, kan door omklappen om GH van het deel van O tot het punt F, waar hij de lijn GH voor het eerst bereikt, ondubbelzinnig veranderd worden in een weg van O' naar P. De coördinaten van het punt O' zijn $\xi = -h$, $\eta = h$, zoals uit de figuur volgt. Anderzijds kan elke weg van O' naar P, die noodzakelijkerwijs de lijn GH moet doorkruisen, ondubbelzinnig veranderd worden in een weg van O naar P door omklappen van het deel van de weg van O tot het punt F, waar hij de lijn GH voor het eerst bereikt, om dezelfde lijn GH. Hieruit volgt, dat het aantal van die wegen van O naar P, die de lijn GH ergens bereiken of doorkruisen, gelijk is aan het aantal wegen van O' naar P. Dit laatst genoemde aantal is echter het aantal wegen in een rooster met

zijden n_1+h+1 en n_2-h+1, dat is dus $\binom{n_1+h+n_2-h}{n_1+h} = \binom{n_1+n_2}{n_1+h}$. Het aantal der wegen van O naar P, die de lijn GH ergens bereiken of doorkruisen is dus ook $\binom{n_1+n_2}{n_1+h}$.

§ 10. Bepaling van het aantal der wegen, die één van beide of beide lijnen $\eta = \xi \pm h$ bereikt of doorkruist.

Indien het aantal wegen bepaald kan worden die hetzij één van beide ofwel beide lijnen $\eta = \xi \pm h$ snijden, dan is het probleem van Smirnov opgelost. Immers voor een dergelijke weg geldt, dat $\max|nF_1(u) - nF_2(u)| \geqq h$, dus $\max|F_1(u) - F_2(u)| \geqq \frac{h}{n} = d$ is.

Wij zullen bovenbedoelde wegen classificeren in vier categorieën:

A. Wegen, die uitsluitend de lijn $\eta = \xi + h$ één of meer malen bereiken of doorkruisen (deze wegen hebben dus geen punt gemeen met de lijn $\eta = \xi - h$).

B. Wegen, die uitsluitend de lijn $\eta = \xi - h$ één of meer malen bereiken of doorkruisen (deze wegen hebben dus geen punt gemeen met de lijn $\eta = \xi + h$).

C. Wegen, die eerst (voordat zij de lijn $\eta = \xi - h$ bereikt hebben) de lijn $\xi = \eta + h$ één of meer malen bereiken of doorkruisen en vervolgens de lijn $\eta = \xi - h$ bereiken of doorkruisen. Na het bereiken der lijn $\eta = \xi - h$ kunnen deze wegen de lijn $\eta = \xi + h$ al dan niet nogmaals bereiken.

D. Wegen, die eerst (voordat zij de lijn $\eta = \xi + h$ bereikt hebben) de lijn $\eta = \xi - h$ één of meer malen bereiken of doorkruisen en vervolgens de lijn $\eta = \xi + h$ bereiken of doorkruisen. Na het bereiken der lijn $\eta = \xi + h$ kunnen deze wegen de lijn $\eta = \xi - h$ al dan niet nogmaals bereiken.

Het is duidelijk, dat deze vier categorieën elkaar uitsluiten. De wegen behorende tot de categorieën A,C en D vormen samen de wegen, die ergens de lijn $\eta = \xi + h$ bereiken of doorkruisen (wij noemen deze tezamen de groep a van wegen). De wegen behorende tot de categorieën B,C en D vormen samen de wegen (groep b), die ergens de lijn $\eta = \xi - h$ bereiken of doorkruisen. De groepen a en b sluiten elkaar niet uit.

Wij zullen het aantal wegen in ieder der categorieën A,B,C en D en in ieder der groepen a en b eveneens aangeven met de letters A,B,C en D resp. a en b.

Uit symmetrieoverwegingen volgt direct, dat A = B, C = D en a = b. Wij kunnen dit o.a. zien door te spiegelen om de diagonaal OP.

Verder geldt:

$$A + C + D = a,$$

$$B + C + D = b.$$

Gezocht wordt echter $A + B + C + D$, n.l. alle wegen, die éé of beide der lijnen $\eta = \xi \pm h$ bereiken of doorkruisen. Met behulp der bovengegeven betrekkingen vinden wij nu:

$$a + b = 2a = A + B + 2(C + D),$$

dus

$$A + B + C + D = 2a - (C + D) = 2\,(a - D).$$

Het aantal wegen a is reeds in de vorige paragraaf berekend. Dit bedroeg $\binom{2n}{n+h} = \binom{2n}{n-h}$. Er blijft dus nog over het aantal wegen D te berekenen, dit zijn dus de wegen, die eerst de lijn $\eta = \xi - h$ één of meer malen bereiken of doorkruisen en vervolgens de lijn $\eta = \xi + h$ bereiken of doorkruisen. Wij zullen dit aantal D berekenen door een herhaalde toepassing van het rotatieprincipe, dat in § 9 reeds gebruikt werd om het aantal wegen a te vinden.

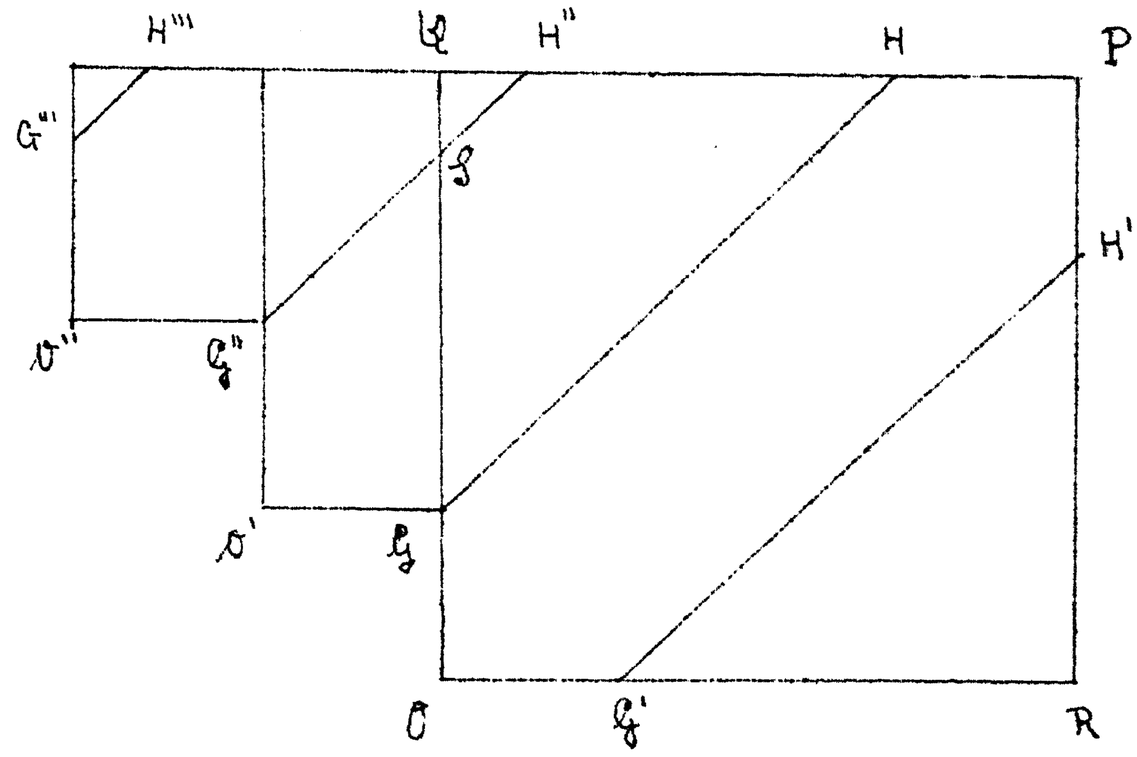


Fig. 8

Bepaling van het aantal wegen D. OR = OQ = n, G'O = OG = = GO' = O'G'' = G''O'' = h. In deze figuur is voor de duidelijkheid het rooster en de "weg" van O naar P niet getekend.

Een weg D in fig. 8 is een weg van O naar P, die eerst G'H' bereikt en eventueel doorkruist en vervolgens GH bereikt en eventueel doorkruist (daarna kan deze weg eventueel weer G'H' bereiken). Door omklappen om GH van het deel van de weg tussen O en het punt waar de weg voor het eerst GH bereikt, wordt deze ondubbelzinnig getransformeerd in een weg van O' naar P. Deze is weer terug te transformeren op ondubbelzinnige wijze, door het deel van de weg tussen O' en het punt waar GH voor het eerst bereikt wordt, om GH om te klappen. Bij de eerste transformatie (waarbij O in O'

komt) wordt G'H' getransformeerd in G"H". Het aantal wegen D is dus gelijk aan het aantal wegen van O' naar P, die G"H" bereiken of doorkruisen voordat zij voor de eerste maal GH bereiken. Een weg als O'GSQP is dus niet de transformatie van een weg D, daar deze weg de lijn G"H" pas bereikt nadat reeds de lijn GH (in het punt G) bereikt was. (Deze weg is de getransformeerde van een weg A, n.l. van OGSQP.)

Het aantal wegen van O' naar P dat de lijn G"H" bereikt of doorkruist, kunnen wij door een overeenkomstige transformatie (nu wentelen om G"H") berekenen. Het is dus gelijk aan het aantal wegen van O" naar P. Hieronder behoren die wegen tot de categorie D die, voor zij G"H" bereiken niet de getransformeerde van GH, n.l. G"'H"' bereiken. Daar de zijden van de rechthoek, waarvan O" en P de einden van een diagonaal zijn, resp. n+2h en n-2h bedragen, is het aantal wegen D gelijk aan $\binom{2n}{n-2h}$ verminderd met het aantal der wegen van O" naar P, die G"'H"' bereiken voor zij G"H" bereiken. Dit laatst genoemde aantal wegen, kan door rotatie om G"'H"' weer op analoge wijze berekend worden en bedraagt natuurlijk $\binom{2n}{n-3h}$ verminderd met het aantal der wegen van O"' naar P, die de getransformeerde van G"H" bereiken voor zij G"'H"' bereiken enz.

Het aantal wegen D is dus gelijk aan $\binom{2n}{n-2h} - \binom{2n}{n-3h} + \binom{2n}{n-4h} \ldots$, waarbij de reeks zolang voortgezet wordt als $n - ih \geqq 0$.

Het aantal wegen van O naar P, die één der beide lijnen of beide lijnen $\eta = \xi \pm h$ bereiken of doorkruisen is dus

$$2(a-D) = 2\left[\binom{2n}{n-h} - \binom{2n}{n-2h} + \binom{2n}{n-3h} - \ldots\right],$$

waarbij de reeks zolang doorgezet moet worden als $n - ih \geqq 0$ is.

Aangzien het totaal aantal wegen van O naar P gelijk is aan $\binom{2n}{n}$ volgt hieruit, dat

$$P\left[\max\left|F_1(u) - F_2(u)\right| \geqq \frac{h}{n} \middle| H_o\right] =$$

$$= \frac{2\left[\binom{2n}{n-h} - \binom{2n}{n-2h} + \binom{2n}{n-3h} - \ldots\right]}{\binom{2n}{n}} .$$

§ 11. De asymptotische verdeling van $P[\max|F_1(u) - F_2(u)| \geqq d|H_o]$ en het geval van steekproeven van ongelijke grootte.

A. De asymptotische verdeling van $P[\max|F_1(u) - F_2(u)| \geqq d|H_o]$.

Voor enigszins grote waarden van n is de berekening van $P = P[\max|F_1(u) - F_2(u)| \geqq d|H_o] = 2\dfrac{\sum_{a=1}^{a=[\frac{n}{h}]}(-1)^{a-1}\binom{2n}{n-ah}}{\binom{2n}{n}}$, waarbij $d = \frac{h}{n}$, zeer omslachtig. Het is derhalve wenselijk, om een asymptotische formule af te leiden, die blijkens een numerieke berekening voor n = 20 reeds bevredigende uitkomsten geeft.

Indien teller en noemer van de breuk, die de waarde van P aangeeft, met $(\frac{1}{2})^{2n}$ vermenigvuldigd worden, dan staat in de teller de som van een aantal termen (afwisselend met het positieve en negatieve teken) uit de ontwikkeling van het binomium $(\frac{1}{2} + \frac{1}{2})^{2n}$, terwijl de noemer door de middelste term van deze ontwikkeling weergegeven worden. Het is echter bekend (zie b.v. Neyman [15]) dat een term van dit binomium voor grote waarden van n zeer goed door een e-macht benaderd wordt (de zgn. verdelingsdichtheid van de normale verdeling) en wel is

$$\binom{2n}{n-k}\left(\tfrac{1}{2}\right)^{2n} \approx \frac{1}{\sqrt{\frac{1}{2}\pi n}}\, e^{-\frac{k^2}{n}}.$$

Wij vinden dus:

$$P = \frac{2\sum_{a=1}^{a=[\frac{n}{h}]}(-1)^{a-1}\cdot\binom{2n}{n-ah}\cdot\left(\frac{1}{2}\right)^{2n}}{\binom{2n}{n}\cdot\left(\frac{1}{2}\right)^{2n}} \approx \frac{2\sum_{a=1}^{a=[\frac{n}{h}]}(-1)^{a-1}\frac{1}{\sqrt{\frac{1}{2}\pi n}}\, e^{-\frac{(ah)^2}{n}}}{\frac{1}{\sqrt{\frac{1}{2}\pi n}}} =$$

$$= 2\sum_{a=1}^{a=[\frac{n}{h}]}(-1)^{a-1}\, e^{-\frac{(ah)^2}{n}}.$$

Voeren wij in deze formule voor h wederom d.n in, dan vinden wij dus

$$P \approx 2\sum_{a=1}^{a=[\frac{n}{h}]}(-1)^{a-1}\, e^{-a^2d^2n}.$$

Indien in de bovengegeven formule voor P niet gesommeerd wordt tot $a = \left[\frac{n}{h}\right]$, doch tot $a \to \infty$, dan worden aan de som toegevoegd termen, die in absolute waarde kleiner zijn dan $e^{-\left(\frac{n}{h}\right)^2d^2n} = e^{-\left(\frac{n}{h}\right)^2\left(\frac{h}{n}\right)^2\cdot n} = e^{-n}$,

afwisselend positief en negatief teken hebben en bovendien monotoon dalen. De som van de toegevoegde termen is dus hoogstens e^{-n}, dus voor enigszins grote waarden van n kan gesteld worden:

$$P \approx 2\sum_{a=1}^{\infty}(-1)^{a-1}\, e^{-a^2d^2n}.$$

In onderstaande tabel zijn voor enige waarden van n en d de exacte en de met de asymptotische formule berekende waarde van P gegeven.

n	d	h	P(exacte formule)	P(asymptotische formule)
20	0,25	5	0,571	0,560
20	0,45	9	0,034	0,035
20	0,50	10	0,012	0,013
50	0,16	8	0,549	0,544
50	0,28	14	0,039	0,040
50	0,32	16	0,012	0,012
100	0,12	12	0,470	0,468
100	0,19	19	0,054	0,054
100	0,23	23	0,010	0,010

Uit deze tabel volgt dat voor kleine waarden van d de asymptotische formule bij n = 20 reeds voldoende nauwkeurig is, terwijl bij n = 100 de asymptotische formule voor alle practische gevallen een voldoende benadering geeft.

B. De toets van Smirnov bij steekproeven van ongelijke grootte.

SMIRNOV heeft zijn toets rechtstreeks als asymptotische toets afgeleid voor steekproeven met $n_1 \to \infty$ resp. $n_2 \to \infty$ waarnemingen, waarbij de verhouding $\frac{n_1}{n_2} \to c$ (SMIRNOV [1]). Een eenvoudiger, doch nog zeer gecompliceerd bewijs is gegeven door FELLER [2]. Wij zullen deze afleiding hier niet geven, doch alleen het resultaat vermelden:

$$P = P\left[\max\left|F_1(u) - F_2(u)\right| \geqq d \,|\, H_0\right] = 2 \sum_{a=1}^{\infty} (-1)^{a-1} \, e^{-2a^2 z^2},$$

waarbij

$$z = d \sqrt{\frac{n_1 n_2}{n_1 + n_2}}.$$

Voor steekproeven van gelijke grootte ($n_1 = n_2 = n$) gaat deze formule in de bovengegeven asymptotische formule over.

Een tabel van $L(z) = 1 - 2 \sum_{a=1}^{\infty} (-1)^{a-1} \, e^{-2a^2 z^2}$ is o.a. te vinden in SMIRNOV [16]. Aan de hand van deze tabel zijn voor enkele typische waarden van P de bijbehorende waarden van z berekend, waarvan hieronder een opgave volgt:

z	P
0,828	0,50
1,073	0,20
1,224	0,10
1,358	0,05
1,517	0,02
1,625	0,01
1,731	0,005
1,950	0,001
2,25	0,0001
2,47	0,00001

§ 12. De toets van Kolmogorov.

In aansluiting op de toets van Smirnov moge hier nog in het kort de toets van Kolmogorov behandeld worden. Terwijl de toets van Smirnov gebruikt wordt om twee empirische verdelingsfuncties te vergelijken, dient de toets van Kolmogorov om een empirische verdelingsfunctie met een theoretische verdelingsfunctie te vergelijken. Dit betekent dus, dat de toets van Smirnov gebruikt kan worden om te toetsen of twee steekproeven uit eenzelfde verdeling afkomstig zijn, terwijl de toets van Kolmogorov dient om een aanpassing aan een volledig gegeven verdeling te toetsen. De toets van Kolmogorov kan b.v. de bekende χ^2-toets vervangen om te toetsen of een empirische verdeling met een gegeven verdeling overeenstemt, mits deze laatste continu is. Een voordeel van deze toets boven de χ^2-toets is nog, dat bij de laatste altijd het materiaal in klassen ingedeeld moet worden en de uitkomst vaak van deze indeling afhangt. Bij de toets van Kolmogorov is een indeling in klassen niet nodig en verzwakt zelfs het onderscheidingsvermogen.

Aangezien een theoretische verdelingsfunctie beschouwd zou kunnen worden als een empirische verdelingsfunctie, berustende op oneindig veel waarnemingen, zou men kunnen verwachten, dat de toets van Smirnov hier toegepast mag worden met $\frac{n_1}{n_2} \to \infty$, alhoewel hier aan één van de voorwaarden voor de toets van SMIRNOV niet voldaan is nl. als $n_1 \to \infty$ en $n_2 \to \infty$ dat dan $\frac{n_1}{n_2} \to c$. Desondanks leidt deze heuristische beschouwingswijze tot het juiste antwoord, zodat de volgende stelling van Kolmogorov geldt:

Stel dat $(x_1, \ldots, x_n)$ een aantal, naar opklimmende grootte gerangschikte, onderling onafhankelijke waarnemingen is van de stochastische variabele x met verdelingsfunctie $F(x)$ en $F_1(x_i) = \frac{i}{n}$ de experimentele cumulatieve verdelingsfunctie is, dan is, voor $n \to \infty$

$$P\left[\max \left|F_1(x) - F(x)\right| \geqq d\right] = 1 - L(d\sqrt{n}).$$

Hierbij is

$$L(z) = 1 - 2\sum^{\infty} e^{-2a^2z^2}.$$

De bovengegeven beschouwingswijze om de toets van SMIRNOV met de toets van Kolmogorov in verband te brengen, heeft alleen enige heuristische en mnemotechnische waarde, doch kan natuurlijk geenszins als poging tot bewijs gelden. Het helderste en toegankelijkste bewijs voor de stelling van Kolmogorov wordt gegeven in FELLER [2]. Wij zullen dit bewijs, dat mathematisch nog vrij ingewikkeld is, hier verder niet behandelen.

De toets van Kolmogorov is in 1933 geplubiceerd [17] en [18]. In 1941 verscheen nog een kort artikel van KOLMOGOROV in de "Annals" [19]. De toets van SMIRNOV is in 1939 geplubiceerd [1] en [20]. Een eenvoudiger afleiding van beide toetsen werd gegeven door FELLER [2], terwijl ook DOOB [21] een, zij het niet geheel strenge, afleiding gegeven heeft. Tenslotte moge nog verwezen worden naar een viertal artikelen van MASSEY [22], [23], [24] en [25], en een artikel van WALD en WOLFOWITZ [26].

Litteratuur

[1] N.Smirnov, On the estimation of the discrepancy between empirical curves of distribution for two independent samples, Bulletin Mathématique de l'Université de Moscou, Vol.2 (1939),fasc.2.

[2] W.Feller, On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions, Ann Math.Stat. 19 (1948) p.177-190.

[3] J.Westenberg, Significance test for the median and interquartile range in samples from continuous populations of any form,Proc. Kon. Ned.Ak.v.Wet. 51 (1948) p.252-261.

[4] A.M.Mood, Introduction to the theory of statistics (1950).

[5] J.Hemelrijk, Symmetrietoetsen en andere toepassingen van de theorie van Neyman en Pearson (1950)

[6] A.Liapounoff, Nouvelle forme du théorème sur la limite de probabilité, Mém.Acad.Sci.St.Petersbourg, 12,(1901)

[7] J.V.Uspensky, Introduction to Mathematical Probability (1937)

[8] W.L.Stevens, Fiducial limits of the parameter of a discontinuous distribution, Biometrika Vol.37 (1950) p.117

[9] R.A.Fisher, The Design of Experiments, (1947)

[10] E.L.Lehmann and C.Stein, On the Theory of Some Non-Parametric Hypotheses, Ann.Math.Stat. 20 (1949) p.28-45

[11] D.van Dantzig, Kadercursus mathematische statistiek, Amsterdam Math. Centrum (1947-1950)

[12] A. van Wijngaarden, Table for the cumulative symmetric binomial distribution, Proc.Kon.Ned.Ak.v.Wet. 53 (1950) p.857-868

[13] J.E.Walsh, Some significance Tests for the Median Which are Valid under Very General Conditions, Ann.Math.Stat.20 (1949) p.64

[14] W.A.Whitworth, Choice and chance with 100 exercises (1901)

[15] J.Neyman, First course in probability and statistics (1950)

[16] N.Smirnov, Table for estimating the goodness of fit of empirical distributions, Ann.Math.Stat.19 (1948) p.279-282

[17] A.Kolmogoroff, Sulla determinazione empirica di una legge di distribuzione, Inst.Ital.Attuari,Giorn.Vol.4 (1933)p.1-11

[18] A.Kolmogoroff, Uber die Grenzwertsätze der Wahrscheinlichkeitsrechnung, Bulletin (Izvestija) Academie des Sciences URSS (1933) p.363-372

[19] A.Kolmogoroff, Confidence limits for an unknown distribution function, Ann.Math.Stat. Vol.12 (1941)p.461-463

[20] N.Smirnov, Sur les écarts de la courbe de distribution empirique, Rec.Math.N.S, Vol.6 (1939)3

[21] J.L.Doob, Heuristic Approach to the Kolmogorov-Smirnov Theorems, Ann.Math.Stat.20 (1949) p.393-404

[22] F.J.Massey Jr, The Kolmogorov-Smirnov Test for Goodness of Fit American Stat.Association Vol.46 (1951)p.68-79

[23] F.J.Massey Jr, A note on the estimation of a distribution function by confidence limits, Ann.Math.Stat. Vol.21 (1950) p.116-119

[24] F.J.Massey Jr, A note on the power of a non-parametric test, Ann.Math.Stat.Vol.21 (1950) p.440-443

[25] F.J.Massey Jr, The distribution of the maximum deviation between two sample cumulative step functions

[26] A.Wald and J.Wolfowitz, Note on Confidence Limits for Continuous Distribution Functions.Ann.Math.Stat.Vol.12 (1941) p.118.

Cursus "Parametervrije Methoden"

Correcties bij hoofdstuk VI.

Bladzij	regel	staat	moet staan
116	29 v.o.	ptobleem	probleem
117	12 v.o.	getoets	getoetst
118	2 v.b.	is verschil-	is even wel verschil
120	2 en 3 v.b.	(bij continue functies altijd)	weglaten
121	12 v.b.	x continu	$\underline{x}$ continu
	13 v.b.	x	$\underline{x}$
	3 v.o.	methoden	lotingen
126	5 v.b.	toetsingsgebied	toetsingsgrootheid
	11 v.b.	waarnemingen	waarnemingen).
128	4 v.b.	leidt toepassen	leidt, toepassen
129	12 v.b.	$\underline{r} = r_1$	$\underline{r} = r$
130	5 v.o.	$u(r-u)!$	$u!(r-u)!$
131	6 v.b.	$\frac{2n_1}{n_1+n_2}$	$\frac{rn_1}{n_1+n_2}$
	17 v.o.	$\mu = E(\underline{u}\|r=r;H_0)$	$\mu = E(\underline{u}\|\underline{r}=r;H_0)$
132	5 v.o.	kleiner	groter
		groter	kleiner
	1 v.o.	$\sum_{i=0}^{i=a}\binom{n}{i}(\frac{1}{2})^n$	$\sum_{i=0}^{i=a-1}\binom{n}{i}$
133	2 v.b.	$\sum_{i=0}^{i=a}\binom{n}{i}(\frac{1}{2})^n$	$\sum_{i=0}^{i=a-1}\binom{n}{i}$
	4 v.b.	idem $\leqq \alpha$	idem $\geqq 1-\alpha$
	9 v.b.	idem	idem
134	15 v.b.	idem	idem.

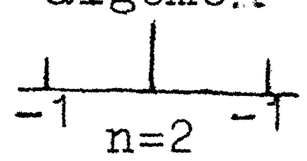
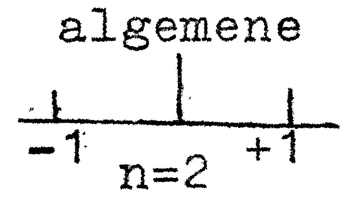
Register der cursus "Parametervrije Methoden".
(Rapport S 59-S 76 van het M.C.)

Tabellen:

ERRATA

Cursus "Parametervrije methoden".

(Rapport S 59 en S 76 van het M.C.)

Bladzijde	regel	staat	moet staan
3	16 v.b.	4^e en 5^e	5^e en 6^e
5	5 v.o.	algemen	algemene
6	1 v.b.		
8	12 v.o.	Ec^2	$\underline{Ec}^2$
10	14 v.b.	..+4+5+2+1+..	..+4+3+4+1+..
	11 v.o.	σ^2	$\sigma_{\underline{s}}^2$
	10 v.o.	σ	$\sigma_{\underline{s}}$
	9 v.o.	$\frac{25}{20,0}$	$\frac{25}{20,2}$
12	18 v.b.	geven +1	geven elk +1
13	7 v.b.	in beide	in één der beide
	15 v.o.	2,2,2,1,1	2,2,2,2,1,1,1
15	4.v.b.	$\sqrt{\sum a_{ij}^2 - \sum b_{ij}^2}$	$\sqrt{\sum a_{ij}^2 \times \sum b_{ij}^2}$
20	15 v.o.	berekeneing	berekening
21	13 v.b.	0,038	0,045
22	16 v.b.	123...n	(II 4) 123...n
	7 v.o.	$(m-mj)^2+(2m-mj)^2+.$ $...(nm-mj)^2$	$(m-m\bar{j})^2+(2m-m\bar{j})^2+..$ $...(nm-m\bar{j})^2$
	6 v.o.	$(1-j)^2+(2-j)^2+..$ $..(n-j)^2$	$(1-\bar{j})^2+(2-\bar{j})^2+...$ $..(n-\bar{j})^2$
23	3 v.b.	$\frac{1}{4}(n+1)^2$	$\frac{1}{4}n(n+1)^2$
24	2.v.b.	$..+8^2+15^2+..$	$..+8^2+14^2+15^2+..$
	8 v.o.	(II 14)	(II 15^a)
	6 v.o.	(II 15)	(II 15^b)
25	4 v.b.	rangnummers-verwisselingen	rangnummerverwisselingen.

Bladzijde	regel	staat	moet staan
25	12 v.o.	--	(II 17) toevoegen
26	18 v.b.	8	S
27	5 v.o.	m-rangschikkingen	m rangschikkingen
28	7 v.b.	36	$36\frac{1}{2}$
29	2 v.b.	$+\frac{(t+1)^2}{4}$	$-\frac{t(t+1)^2}{4}$
	3 v.b.	$\frac{1}{4}(t+1)^2$	$\frac{1}{4}t(t+1)^2$
	12 v.o.	$\sum_{j=1}^{m}$	$\sum_{j=1}^{n}$
30	2 v.b.	+ 17	$+ (17)^2$
32	8 v.b.	fig.	fig. II
	14 v.b.	verdeling	verdelingen ge-
33	10 v.b.	$\sum_{i=1}^{m}\sum_{k=1}^{i-1}$	$\sum_{i=2}^{m}\sum_{k=1}^{i-1}$
34	5 v.b.	(II.39)	(II.34)
	6 v.b.	e^{2z}	$e^{2z} = F$
35	7 v.b.	(II.35)	(II.37)
36	17 v.b.	$\sum_{i=1}^{m}\sum_{k=1}^{i-1}$	$\sum_{i=2}^{m}\sum_{k=1}^{i-1}$
	18 v.b.	$\sum_{i=1}^{m}\sum\sum$	$\sum_{i=2}^{m}\sum\sum$
37	3 v.o.	(geen	(geen gelijke rangnummers)
39	3 v.b.	$W' = \frac{-1}{max^{+2}} = \frac{-1}{\frac{1}{12}m^2n(n^2-1)+2}$	$W' = \frac{S-1}{S_{max}^{+2}} = \frac{S-1}{\frac{1}{12}m^2n(n^2-1)+2}$
	3 v.o.	$W' = \frac{-1}{\frac{1}{12}m^2(n^2-1)+2}$	$W' = \frac{S-1}{\frac{1}{12}m^2(n^2-1)+2}$
40	4 v.b.	$W = \frac{12}{nm^2(n^2-1)}$	$W = \frac{12\ S}{m^2n(n^2-1)}$
41	5 v.b.	$P[->]$	$P[\underline{S} \geqslant S]$
43	3 v.o.	bloedlcihaampjes	bloedlichaampjes
44	14 v.o.	erythroo ten	erythrocyten
	12 v.o.	de van	de hand van
26	5 v.b.	$8 \ldots \frac{2}{5} \ldots \frac{1}{6}$	$8 \ldots \frac{1}{6} \ldots \frac{1}{6}$

Bladzijde	regel	staat	moet staan
47	5 v.o.	gezegd: H'_o	gezegd: H_o
49	12 v.b.	kleine	kleinere
50	4 v.o.	a	a
	3 v.o.	b	b
53	tussen 13 en 14 v.b. invoegen:		4 3 $\frac{3}{20}$ 0,50
	10 v.o.	$\leqslant \alpha \leqslant$	$\leqslant \alpha =$
	3 v.o.	Deze berekeningen	deze
54	3 v.b.	$i(\underline{x}_i - \underline{y}_j)$	$\iota(\underline{x}_i - \underline{y}_j)$
55	7 v.b.	$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{v} e^{-\frac{}{2}} t^2 dt$	$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{v} e^{-\frac{1}{2}} t^2 dt$
	2 v.o.	§ 3	§ 1.3
56	16 v.o.	is er	heeft men
	15 v.o.	gevonden kunnen worden	kunnen vinden
	6 v.o.	die is immers	die immers
57	1 v.b.	$\sqrt{(x_i-\bar{x})^2+(y_j-\bar{y})^2}$	$\sqrt{\sum_{i=1}^{m}(x_i-\bar{x})^2+\sum_{j=1}^{n}(y_j-\bar{y})^2}$
59	2 v.b.	van 6	van
	3 v.o.	$P\left[\underline{t} \geqslant \lvert t\rvert \,\middle\vert\, H_o\right]$	$P\left[\lvert\underline{t}\rvert \geqslant \lvert t\rvert \mid H_o\right]$
60	5 v.b.	I'-III 0,025	I'-III 7 6 3 0,0082 0,025
62	4 v.b.	"gelijkheid	"gelijkheid"
65	9 v.b.	2u-5	2u+5
		2m-5	2m+5
		2n-5	2n+5
	12 v.o.	daarin vervangen	daarin σ vervangen
66[a]	3 v.o.	ver	over
67	19 v.o.	van	van ξ
	18 v.o.	van gemeten	van η gemeten
69	13 v.b.	(i=1,2...n)	i=(1,2...n) [1]
	12 v.o.	fun ties	functies

Bladzijde	regel	staat	moet staan
69	12 v.o.	waarnemeingen	waarnemingen
	4 v.o.	$\frac{1}{n}\ \frac{\sigma^2}{\text{var } x}$ 1)	$\frac{1}{n}\ \frac{\sigma^2}{\text{var } x}$
70	11 v.b.	ongecorelleerd	ongecorreleerd
	20 v.b.	a en b	a en b
	5 v.o.	Brenna	Brennan
71	4 v.b.	verg	ver
	7 v.b.	B)	b)
	13 v.o.	bij	hij
	6 v.o.	schattting	schatting
	1 v.o.	$(\eta_{m+1}+\omega_{m+1})+..$	$(\eta_{m+1}-\omega_{m+1})+..$
		$..+(\eta_n+\omega_n)$	$..+(\eta_n-\omega_n)$
72	15 v.b.	zijn	is
73	17 v.b.	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1}(y_i-y_j)}{\sum_{i=1}^{n}\sum_{j=1}^{i-1}(x_i-x_j)}$	$\frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1}(y_i-y_j)}{\sum_{i=2}^{n}\sum_{j=1}^{i-1}(x_i-x_j)}$
74	8 v.b.	$\underline{z}=\beta^2\text{var }\underline{u}+$	$\underline{z}=\beta^2\text{var }\underline{u}+$
		$+\text{ var}(\underline{v}+\underline{w})$	$+\text{ var}(\underline{v}+\underline{w})$
75	4 v.b.	zij	wij
	14 v.b.	$P\left[\frac{z_{m+i}-z_i}{x_{m+i}-x_i}<o\right]$	$P\left[\frac{\underline{z}_{m+i}-\underline{z}_i}{\underline{x}_{m+i}-\underline{x}_i}<o\right]$
76	6 v.b.	Q en Q	Q en Q_1
	9 v.b.	elk tot	elk β tot
	12 v.o.	$(\underline{D}_r,\underline{D}_{m+r-1})$	$(\underline{D}_r,\underline{D}_{m-r+1})$
77	2 v.o.	dezer	dezes
78	6 v.b.	$(i=1,..,j-1;j=1,..,n)$	$(i=1,..,j-1;j=2,..,n)$
	9 v.o.	kleiner	groter
	7 v.o.	geval	getal
79	10 v.b.	im ers	immers
	19 v.o.	opklim ende	opklimmende
80	1 v.b.	meet fouten	meetfouten

Bladzijde	regel	staat	moet staan
80	6 v.b.	de ξ volgorde	de ξ -volgorde
	21 v.o.	a	af
81	15 v.b.	§ 13	§ 1.3
82	2 v.b.	stellen K	stellen geeft K
	11 v.o.	heb en	hebben
83	16 v.o.	regine	region
	8 v.o.	Aan	Ann.
87	4 v.b.	$P\left[\underline{x} = \underline{x}\right]$	$P\left[\underline{x} = x\right]$
	7 v.b.	v n	van
	17 v.o.	$\underline{L}_1$ en L_2	$\underline{L}_1$ en $\underline{L}_2$
	9 v.o.	grenz	grens
88	10 v.b.	$[x_r, x_{n-s+1}]$	$[\underline{x}_r, \underline{x}_{n-s+1}]$
	12 v.b.	$[x_r, x_{n-s+1}]$	$[\underline{x}_r, \underline{x}_{n-s+1}]$
	10 v.o.	moet n n dan	moet n dan
90	12 v.b.	zij	zijn
	14 v.b.	situatie	situaties
	9 v.o.	$\binom{N-m_o+r-1}{r-1}\binom{m_o+n-r}{n-r}$	$\binom{N-m_o+r-1}{r-1}\binom{m_o+n-r}{n-r} \Big/ \binom{N+n}{n}$
94	12 v.o.	beschouwen nu	beschouwen we nu
95	12 v.b.	$P\left[a \leqslant \frac{m}{N} \leqslant b \middle\| n,N;r,s\right]$	$P\left[a \leqslant \frac{\underline{m}}{N} \leqslant b \mid n,N;r,s\right]$
	15 v.o.	$\underline{p} = \frac{m}{N}$	$\underline{p} = \frac{\underline{m}}{N}$
	13 v.o.	..(1-p+Δp)(p+ +(n-r+s)Δ p)..	..(1-p+Δp)(p+ +(n-r-s)Δ p)..
96	3 v.o.	$I_\beta(t, n-t+1) = \alpha$	$I_\beta(n-t+1, t) = \alpha$
99	7 v.o.	$g(p) \mid \underline{u}_r = u_r)$	$g(p \mid \underline{u}_r = u_r)$
	1 v.o.	$P[p \leqslant \underline{u}_{n-s+1} \leqslant p+dp]$	$P[u_r+p < \underline{u}_{n-s+1} \leqslant u_r+p+dp]$
100	12 v.o.	$[\underline{x}_r, x_{n-s+1}]$	$[\underline{x}_r, \underline{x}_{n-s+1}]$
105	1 v.b.	In In dit	In dit
107	9 v.o.	speciale	De speciale

Bladzijde	regel	staat	moet staan
108	8 v.b.	continu	continue
	2 v.o.	geschtste	geschetste
110	2 v.b.	alle tekens veranderen!	
112	18 v.b.	beschreven	beschrevene
Enkele voorbeelden van het gebruik van tolerantiegrenzen.	blz. 1		
	9 v.b.	$\underline{I}$ en $\underline{L}_2$	$\underline{L}_1$ en $\underline{L}_2$
	16 v.o.	$I_\beta(t,n-t+1)-\alpha$	$I_\beta(n-t+1,t)-\alpha$
	blz. 3		
	5 v.b.	$I_\gamma(10,11) \geqslant 0,10$	$1-I_\gamma(11,10) > 0,10$
	6 v.b.	$\gamma = 0,38$	$\gamma = 0,66$
	9 v.b.	$= + rt +$	$= 1 + rt +$
116	9 v.b.	ptobleem	probleem
117	12 v.o.	getoets	getoetst
118	2 v.b.	is verschillend	is evenwel verschillend
120	2 v.b.	bij continue functie altijd	weg laten!
	12 v.b.	OQR	OQP
121	12 v.b.	x	$\underline{x}$
	13 v.b.	x	$\underline{x}$
	3 v.o.	methoden	lotingen
124	17 v.b.	wa rneming	waarneming
125	15 v.b.	kunnen de	kunnen we de
126	5 v.b.	toetsingsgebied	toetsingsgrootheid
	11 v.b.	waarnemingen	waarnemingen)
129	12 v.b.	$\underline{r} = r_T$	$\underline{r} = r$
130	6 v.b.	,14 v.b. $\sum$	$\sum_u$
	7 v.b.	$\min(r,n_2)$	$\min(r,n_1)$
	5 v.o.	$u(r-u)!$	$u!(r-u)!$

Bladzijde	regel	staat	moet staan
132	13 v.b.	hoordstuk	hoofdstuk
	5 v.o.	kleiner	groter
		groter	kleiner
	1 v.o.	$-2^{-n+1}\sum_{i=0}^{a}\binom{n}{i}(\frac{1}{2})^n$	$-2^{-n+1}\sum_{i=0}^{a-1}\binom{n}{i}$
133	1 v.b.	$2^{-n+1}\sum_{i=0}^{a}\binom{n}{i}(\frac{1}{2})^n$	$2^{-n+1}\sum_{i=0}^{a-1}\binom{n}{i}$
	3 v.b.	$1-2^{-n+1}\sum_{i=0}^{a}\binom{n}{i}(\frac{1}{2})^n \leqslant \alpha$	$1-2^{-n+1}\sum_{i=0}^{a-1}\binom{n}{i} \geqslant 1-\alpha$
	9 v.b.	$\leqslant \alpha$	$> 1-\alpha$
	18 v.b.	hoogstens gelijk aan α	minstens gelijk aan $1-\alpha$
134	15 v.b.	$< \alpha$	$> 1-\alpha$
	17 v.b.	wordt	worden
136	13 v.o.	$P\left[\underline{x}_u < \underline{y}_v\right] + P\left[\underline{x}_u > \underline{y}_v\right] < \alpha$	$P\left[\underline{x}'_u < \underline{y}'_v\right] + P\left[\underline{x}'_u > \underline{y}'_v\right] < \alpha$
140	14 v.o.	geval de	geval dat de
141A	6 v.b.	k=0,...,n-2	k=1,...,n-2
141B	2 v.o.	van $< \frac{1}{2}n$	van $h < \frac{1}{2}n$
144	4 v.b.	éé	één
145	4 v.o.	aanggien	aangezien
146	13 v.b.	worden	wordt
149	8 v.b.	geplubiceerd	gepubliceerd

Tevens dienen de volgende pagina's vervangen te worden door de bijgevoegde verbeterde exemplaren:

bijlage blz 30
31
38
101
102
113
114
131

Enkele voorbeelden van het gebruik van tolerantiegrenzen.

blz 2.