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 193

vertrouwelijk

Analyse van subjectieve keuringsresultaten

door

Ir A.R. Bloemena

Februari 1956

The Mathematical Centre at Amsterdam, founded the 11th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications, and is sponsored by the Netherlands Government through the Netherlands Organization for Pure Research (Z.W.O.) and the Central National Council for Applied Scientific Research in the Netherlands (T.N.O.), by the Municipality of Amsterdam and by several industries.

Inhoud

		pagina
1.	Inleiding	2
2.	Gestelde vragen	2
3.	De statistische bewerking van het waarnemingsmateriaal	4
	3.1. De mate van gestrengheid van de beoordelaars	4
	3.2. De middelmatige eenheden	9
	3.3. Toetsing van de invloed van voorafgaande waarnemingen	10
4.	Paarsgewijs samengevoegde beoordelingen	13
	4.1. De mate van gestrengheid van de beoordelaars	13
	4.2. Beïnvloeding door voorafgaande waarnemingen	15
	4.3 Vergelijking van de keuring met G-S ten opzichte van 1,2,3,en 4	16
5.	Conclusies	16
6.	Memoranda: S 47 (M 6)	na 16
	S 47 (M 7)	
	S 47 (M 14)	
	S 102(M 17b)	
	S 53 (M 23)	

1. Inleiding

Een duur product wordt visueel gekeurd en wordt niet verkocht, indien het wordt afgekeurd. Het is voor de fabrikant van belang, deze subjectieve keuring zo nauwkeurig mogelijk te normaliseren om te voorkomen, dat producten onnodig afgekeurd worden.

Het vermoeden bestaat evenwel, dat het resultaat van de keuring afhangt van de voorafgaande keuringsresultaten, en wel in die zin, dat een controleur een middelmatige eenheid eerder zal goedkeuren, wanneer de voorafgaande eenheden slecht waren, dan wanneer zij goed waren.

Teneinde na te gaan, of dit effect inderdaad optreedt, werden dezelfde 25 eenheden in willekeurige volgorde gekeurd door 11 beoordelaars. Bij deze keuringen werden de volgende cijfers toegekend:

1 = goed
2 = voldoende
3 = onvoldoende
4 = slecht.

Een enkele maal werd ook een + of een - toegekend, of ook halve eenheden.

De waarnemingen, die bij dit experiment verkregen werden, zijn vermeld in Tabel I.

Bij een gedeelte van het experiment was de volgorde, waarin de beoordelaars de eenheden keurden, niet aselect. Daar de statistische analyse van deze groep waarnemingen van minder belang is dan de ontwikkeling van een methode, waarmee men in het algemeen dit type van keuringsresultaten kan analyseren, zijn voor dit onderzoek de waarnemingen opgevat als de keuringscijfers van in aselecte volgorde voorgelegde eenheden.

2. Gestelde vragen

De volgende vragen werden, in verband met de in de inleiding beschreven keuring, voorgelegd:

a) In hoeverre leveren de verkregen waarnemingen een bewijs, dat de beoordeling van een eenheid inderdaad afhankelijk is van de voorafgaande beoordelingen? Het gaat hierbij niet alleen om het toetsen van het effect, doch ook om het vinden van een numerieke maat voor de graad van afhankelijkheid.

b) Bestaan er verschillen tussen de beoordelaars? Zo ja, welk effect geeft de grootste mate van onzekerheid in de beoordelingen: de verschillen tussen de beoordelaars, of de correlatie tussen naburige beoordelingen?

Tabel I

Punten, toegekend bij de keuring van 25 eenheden door 11 beoordelaars

Persoon 1		2		3		4		5		6		7		8		9		10		11	
Prod No.	Oordeel	Prod No.	Oordeel	Prod No.	Oordeel	Prod No.	Oordeel	Prod No.	Oordeel	Prod No.	Oordeel	Prod No.	Oordeel	Prod No.	Oordeel	Prod No.	Oordeel	Prod No.	Oordeel	Prod No.	Oordeel
1	3	24	4	18	2	1	1	13	1	13	2	19	4	21	1	21	4	13	3	11	2
2	1	5	3^5	9	2	6	3	12	1	2	2	5	3	19	3	5	4	11	1	12	3.
3	2	12	3^5	17	3^+	10	1	16	1	15	3	4	2	7	3	25	3	15	1	16	2
4	1	2	2^+	16	2	24	3	10	2	10	1	23	4	24	3	24	4	16	1	14	1
5	3^+	20	2-	25	4	14	1	6	4	17	3	1	2	6	3	18	2	1	3	17	4
6	3	22	3	3	1^5	9	2	3	2	9	2	24	4	17	3	20	1	3	2	6	3
7	3^5	16	1	13	1	4	2	11	3	19	3	22	4	12	1	2	1	12	1	22	4
8	1	25	4	7	4	19	3	2	2	22	3	21	3	11	1	11	1	2	1	21	3
9	2	9	3	8	2	21	3^5	15	2	23	3	25	3	14	1	14	1	14	1	25	3
10	2^-	6	3^5	4	3^-	3	1	14	1	7	4	7	3^5	16	1	15	1	10	3	7	3
11	2	3	2^-	6	3	7	4	8	2	1	2	20	1	18	3	17	4	8	1	9	2
12	2	13	1	21	3	5	3^5	4	3	18	3	17	4	8	2	9	2	20	2	4	2
13	2^+	21	3	10	1	11	1	20	3	8	1	18	3	1	2	6	4	9	3	18	3
14	2	4	2	23	4	8	1	19	4	21	3	12	1	13	2	19	4	22	4	2	1
15	1^+	10	1	15	2	23	4	21	4	5	2	2	1	2	1	22	4	5	4	13	1
16	1	1	3	14	3	12	1^+	24	4	25	4	14	1	15	1	23	3	25	4	10	2
17	3	8	1	12	1	13	1^5	7	4	24	4	10	1	10	1	7	4	23	4	15	1
18	3	17	3	24	4	20	2	17	4	4	2	6	1	3	2	4	1	18	3	3	3
19	3	7	4	11	2	25	3	18	4	6	3	3	1	4	1	1	2	6	3	20	2^5
20	2^+	23	4	5	4	15	1	22	4	12	1	13	1	5	4	13	1	19	4	19	4
21	3^+	19	3^5	20	3	2	1	5	4	11	2	11	1	22	3	12	1	21	4	5	2
22	3^5	18	2^5	2	2	16	1	25	4	14	1	15	1	23	4	16	1	24	4	24	4
23	3^+	15	1	1	3	18	2	23	4	16	2	16	1	25	3	10	1	7	4	23	3
24	3^+	14	1	22	4	22	3	9	3	20	3	8	1	20	2	3	2	4	1	8	2
25	3	11	1.	19	4	17	3	1	3	3	3	9	1	9	1	8	1	17	4	1	1

c) In hoeverre is een beoordeling met 4 cijfers (1,2,3 of 4) van waarde? Zou essentiële informatie verloren gaan indien men alleen een oordeel goed of slecht had uitgesproken? dus als men de cijfers 1 en 2, en 3 en 4 paarsgewijze in één oordeel bijeen voegt.

3. De statistische bewerking van het waarnemingsmateriaal

3.1. De mate van gestrengheid van de beoordelaars

Allereerst is voor het gegeven waarnemingsmateriaal nagegaan, of er verschillen bestaan tussen de beoordelaars, wat betreft de gestrengheid, waarmee zij de hen voorgelegde eenheden keuren.

Een methode, waarmee dit onderzocht kan worden, is die der m rangschikkingen (zie memorandum S 47 (M 14). Alle in dit rapport genoemde memoranda zijn als bijlage toegevoegd). Deze methode kan worden toegepast om in het geval, dat m beoordelaars n voorwerpen rangschikken volgens opklimmende waarden van een bepaald criterium, de nulhypothese te toetsen, dat er geen overeenstemming bestaat tussen de rangschikkingen van de waarnemers.

In het beschouwde geval zal men de methode toepassen, alsof de beoordelaars gerangschikt worden door de gekeurde eenheden. Zijn enkele beoordelaars systematisch strenger, dan zullen de gekeurde eenheden hen een hoog rangnummer toekennen. De getoetste hypothese H_0: er is geen overeenstemming tussen de rangschikkingen, houdt dan in dat alle beoordelaars even streng zijn.

Een eventueel aanwezige afhankelijkheid tussen opeenvolgende waarnemingen zal niet storend werken bij de uitvoering van de toets; slechts het onderscheidingsvermogen zal iets kleiner worden. Door de aselecte volgorde, waarin de eenheden gekeurd zijn, zal evenwel van een systematische invloed door deze afhankelijkheden geen sprake zijn.

Daar er vele gelijke waarnemingen voorkomen, moet de berekeningswijze van de toetsingsgrootheid een kleine verandering ondergaan, teneinde met deze gelijken rekening te houden (zie de in het memorandum genoemde literatuur, o.a. [2]).

Stel het aantal gerangschikte voorwerpen (in dit geval de beoordelaars) n , en het aantal malen, dat deze voorwerpen gerangschikt worden m . Komen in de i-de rij van het schema met m rijen en n kolommen groepen van $t_1, t_2, \ldots, t_k$ gelijken voor, dan berekent men voor $i = 1, \ldots, m$

$$n\tilde{u}_{2i} = \frac{1}{12} n(n^2-1) - \frac{1}{12} \sum_{\nu=1}^{k} t_\nu (t_\nu^2 - 1) .$$

De in het memorandum gedefinieerde grootheid W wordt, indien gelijke waarnemingen voorkomen, gevonden uit de betrekking:

$$W = \frac{S}{m \sum_{i=1}^{n} n_i^3 - mT},$$

waarin S dezelfde betekenis heeft als in het memorandum. De grootheid

$$V = (m-1)\frac{W}{1-W}$$

heeft onder H_o bij benadering een F-verdeling met (als m en n vrij groot zijn)

$$\nu_1 = n-1$$

en

$$\nu_2 = (m-1)\nu_1$$

vrijheidsgraden.

De uitvoering van de toets kan gevolgd worden in Schema I. Hierbij werd een overschrijdingskans $k < 0{,}001$ gevonden, zodat de nulhypothese met grote stelligheid verworpen kan worden, ten gunste van de alternatieve hypothese, dat er systematische verschillen bestaan tussen de beoordelaars, wat de mate van hun gestrengheid betreft.

De methode der m rangschikkingen levert niet alleen een toets voor de overeenstemming tussen de rangschikkingen in het schema, doch ook een schatting van de objectieve volgorde van - in dit geval - de beoordelaars, gerangschikt naar hun gestrengheid. Deze schatting verkrijgt men, door hen te rangschikken volgens opklimmende kolomtotalen. In Tabel II is deze schatting van de volgorde van de beoordelaars vermeld in opklimmende mate van gestrengheid, tezamen met het gemiddelde van de puntenbeoordelingen, die door ieder van deze beoordelaars is toegekend aan de 25 eenheden.

Tabel II

Schatting van de volgorde der beoordelaars naar opklimmende gestrengheid.

beoordelaars	8	4	7	9	1	2	6	11	10	3	5
gemiddelde puntenbeoordeling	2,1	2,1	2,1	2,3	2,4	2,5	2,5	2,5	2,6	2,7	2,9

Er wordt met nadruk op gewezen, dat het rangschikken der beoordelaars geen uitspraak inhoudt over systematische verschillen van elk paar van hen.

Schema I

Uitvoering van de toets voor de gelijkheid der beoordelaars, wat hun gestrengheid van oordeel betreft.

eenheid \ beoordelaar	1	2	3	4	5	6	7	8	9	10	11	$n\tilde{\mu}_{2i}$
1	9	9	9	1½	9	4½	4½	4½	4½	9	1½	94,5
2	4	11	9	4	9	9	4	4	4	4	4	80
3	7	4	3	1½	7	10½	1½	7	7	7	10½	99
4	2½	7	10	7	11	7	7	2½	2½	2½	7	95
5	4	5½	9	5½	9	1½	3	9	9	9	1½	99
6	5	9	5	5	10½	5	1	5	10½	5	5	81,5
7	3½	8	8	8	8	8	3½	1½	8	8	1½	81
8	4	4	9½	4	9½	4	4	9½	4	4	9½	77
9	5½	10	5½	5½	10	5½	1½	1½	5½	10	5½	90
10	8	4	4	4	9½	4	4	4	4	11	9½	81,5
11	8½	3½	8½	3½	11	8½	3½	3½	3½	3½	8½	87,5
12	9	11	4	8	4	4	4	4	4	4	10	82
13	10	3½	3½	7	3½	8½	3½	8½	3½	11	3½	92
14	10	5	11	5	5	5	5	5	5	5	5	50
15	8	4	9½	4	9½	11	4	4	4	4	4	81,5
16	4½	4½	10	4½	4½	10	4½	4½	4½	4½	10	66
17	3	3	6	3	9	3	9	3	9	9	9	90
18	7½	4	2	2	11	7½	7½	7½	2	7½	7½	90,5
19	2½	5	8½	2½	8½	2½	8½	2½	8½	8½	8½	87,5
20	7	3	10	5	10	10	1½	5	1½	5	8	105,5
21	7	4	4	8	10	4	4	1	10	10	4	98
22	5	2½	8½	2½	8½	2½	8½	2½	8½	8½	8½	87,5
23	4	8	8	8	8	2	8	8	2	8	2	80
24	3	7½	7½	1½	7½	7½	7½	1½	7½	7½	7½	67,5
25	3½	9	9	3½	9	9	3½	3½	3½	9	3½	82,5
kolomtotalen	145	149	182	114	211,5	154	116,5	112,5	136	174,5	155	
gemiddelde	150	150	150	150	150	150	150	150	150	150	150	
afwijkingen van het gemiddelde	- 5	- 1	+32	-36	+61,5	+ 4	-33,5	-37,5	-14	+24,5	+ 5	

$\sum_{i=1}^{25} n\mu_{2i} = 2126,5$

$S = 9494$

$$W = \frac{9494}{25 \times 2126,5} = 0,1785$$

$$V = 24 \cdot \frac{0,1785}{0,8215} = 5,21 \quad \therefore k < 0,001$$

$\nu_1 = 10, \qquad \nu_2 = 240$

Ten einde verdere vergelijkingen van de keuringsresultaten mogelijk te maken, moet ten gevolge van dit effect een correctie voor de beoordelaars worden aangebracht. Aangezien de waarnemingen slechts enkele waarden aan kunnen nemen, is het niet zinvol een zeer verfijnde correctie toe te passen. Om deze reden zijn de correcties beperkt tot het optellen of aftrekken van $\frac{1}{2}$ of $\frac{1}{4}$ eenheid van de waarnemingsresultaten van bepaalde beoordelaars. Rekening houdend met het feit, dat beoordelingen boven de 4, resp. onder de 1, niet voor kunnen komen, moet voor het geval van een beoordelaar, die systematisch te lage, resp. te hoge, cijfers toekent, slechts op die waarnemingen, die ongelijk aan 4, resp 1, zijn een correctie worden aangebracht. Tenslotte werden, gezien de gemiddelden in Tabel II de volgende correcties toegepast:

beoordelaars	correcties
8	+1/4
4	+1/4
7	+1/4
9	+1/4
10	-1/4
3	-1/4
5	-1/2

De op deze wijze gecorrigeerde waarnemingen zijn vermeld in Tabel III.

Op de gecorrigeerde waarnemingen werd, ter controle, opnieuw de methode van de m rangschikkingen toegepast. De hierbij verkregen resultaten zijn vermeld in Schema II.

Schema II

Toepassing van de methode der m -rangschikkingen op de gecorrigeerde waarnemingen.

beoordelaar	1	2	3	4	5	6	7	8	9	10	11
kolomtotalen	$131\frac{1}{2}$	$145\frac{1}{2}$	$153\frac{1}{2}$	148	165	$147\frac{1}{2}$	151	$145\frac{1}{2}$	$170\frac{1}{2}$	$141\frac{1}{2}$	$150\frac{1}{2}$
gemiddelde	150	150	150	150	150	150	150	150	150	150	150
afwijkingen van het gemiddelde	$-18\frac{1}{2}$	$-4\frac{1}{2}$	$+3\frac{1}{2}$	-2	+15	$-2\frac{1}{2}$	+1	$-4\frac{1}{2}$	$+20\frac{1}{2}$	$-8\frac{1}{2}$	$+\frac{1}{2}$

$S = 1124$ $\qquad \sum_{i=1}^{25} n\tilde{\mu}_{2i} = 2592$

$$W = \frac{1124}{25 \times 2592} = 0,0173$$

$$V = 24 \times \frac{0,0173}{0,9827} = 0,42, \qquad \nu_1 = 10,\ \nu_2 = 240 \therefore k > 0,90$$

Tabel III

Gecorrigeerde waarnemingen

beoordelaar	1		2		3		4		5		6		7		8		9		10		11	
1	3	24	4	18	$1\frac{3}{4}$	1	$1\frac{1}{4}$	13	1	13	2	19	4	2	$1\frac{1}{4}$	21	4	13	$2\frac{3}{4}$	11	2
2	1	5	3	9	$1\frac{3}{4}$	6	$3\frac{1}{4}$	12	1	2	2	5	$3\frac{1}{4}$	19	$3\frac{1}{4}$	5	4	11	1	12	3
3	2	12	$3\frac{1}{2}$	17	3	10	$1\frac{1}{4}$	16	1	15	3	4	$2\frac{1}{4}$	7	$3\frac{1}{4}$	25	$3\frac{1}{4}$	15	1	16	2
4	1	2	$2\frac{1}{2}$	16	$1\frac{3}{4}$	24	$3\frac{1}{4}$	10	$1\frac{1}{2}$	10	1	23	4	24	$3\frac{1}{4}$	24	4	16	1	14	1
5	$3\frac{1}{4}$	20	$1\frac{3}{4}$	25	$3\frac{3}{4}$	14	$1\frac{1}{4}$	6	$3\frac{1}{2}$	17	3	1	$2\frac{1}{4}$	6	$3\frac{1}{4}$	18	$2\frac{1}{4}$	1	$2\frac{3}{4}$	17	4
6	3	22	3	3	$1\frac{1}{4}$	9	$2\frac{1}{4}$	3	$1\frac{1}{2}$	9	2	24	4	17	$3\frac{1}{4}$	20	$1\frac{1}{4}$	3	$1\frac{3}{4}$	6	3
7	$3\frac{1}{2}$	16	1	13	1	4	$2\frac{1}{4}$	11	$2\frac{1}{2}$	19	3	22	4	12	$1\frac{1}{4}$	2	$1\frac{1}{4}$	12	1	22	4
8	1	25	4	7	$3\frac{3}{4}$	19	$3\frac{1}{4}$	2	$1\frac{1}{2}$	22	3	21	$3\frac{1}{4}$	11	$1\frac{1}{4}$	11	$1\frac{1}{4}$	2	1	21	3
9	2	9	3	8	$1\frac{3}{4}$	21	$3\frac{3}{4}$	15	$1\frac{1}{2}$	23	3	25	$3\frac{1}{4}$	14	$1\frac{1}{4}$	14	$1\frac{1}{4}$	14	1	25	3
10	$1\frac{3}{4}$	6	$3\frac{1}{2}$	4	$2\frac{1}{2}$	3	$1\frac{1}{4}$	14	1	7	4	7	$3\frac{1}{4}$	16	$1\frac{1}{4}$	15	$1\frac{1}{4}$	10	$2\frac{3}{4}$	7	3
11	2	3	$1\frac{3}{4}$	6	$2\frac{3}{4}$	7	4	8	$1\frac{1}{2}$	1	2	20	$1\frac{1}{4}$	18	$3\frac{1}{4}$	17	4	8	1	9	2
12	2	13	1	21	$2\frac{3}{4}$	5	$3\frac{3}{4}$	4	$2\frac{1}{2}$	18	3	17	4	8	$2\frac{1}{4}$	9	$2\frac{1}{4}$	20	$1\frac{3}{4}$	4	2
13	$2\frac{1}{4}$	21	3	10	1	11	$1\frac{1}{4}$	20	$2\frac{1}{2}$	8	1	18	$3\frac{1}{4}$	1	$2\frac{1}{4}$	6	4	9	$2\frac{3}{4}$	18	3
14	2	4	2	23	$3\frac{3}{4}$	8	$1\frac{1}{4}$	19	$3\frac{1}{2}$	21	3	12	$1\frac{1}{4}$	13	$2\frac{1}{4}$	19	4	22	$3\frac{3}{4}$	2	1
15	$1\frac{1}{4}$	10	1	15	$1\frac{3}{4}$	23	4	21	$3\frac{1}{2}$	5	2	2	$1\frac{1}{4}$	2	$1\frac{1}{4}$	22	4	5	$3\frac{3}{4}$	13	1
16	1	1	3	14	$2\frac{3}{4}$	12	$1\frac{1}{2}$	24	$3\frac{1}{2}$	25	4	14	$1\frac{1}{4}$	15	$1\frac{1}{4}$	23	$3\frac{1}{4}$	25	$3\frac{3}{4}$	10	2
17	3	8	1	12	1	13	$1\frac{3}{4}$	7	$3\frac{1}{2}$	24	4	10	$1\frac{1}{4}$	10	$1\frac{1}{4}$	7	4	23	$3\frac{3}{4}$	15	1
18	3	17	3	24	$3\frac{3}{4}$	20	$2\frac{1}{4}$	17	$3\frac{1}{2}$	4	2	6	$1\frac{1}{4}$	3	$2\frac{1}{4}$	4	$1\frac{1}{4}$	18	$2\frac{3}{4}$	3	3
19	3	7	4	11	$1\frac{3}{4}$	25	$3\frac{1}{4}$	18	$3\frac{1}{2}$	6	3	3	$1\frac{1}{4}$	4	$1\frac{1}{4}$	1	$2\frac{1}{4}$	6	$2\frac{3}{4}$	20	$2\frac{1}{2}$
20	$2\frac{1}{4}$	23	4	5	$3\frac{3}{4}$	15	$1\frac{1}{4}$	22	$3\frac{1}{2}$	12	1	13	$1\frac{1}{4}$	5	4	13	$1\frac{1}{4}$	19	$3\frac{3}{4}$	19	4
21	$3\frac{1}{4}$	19	$3\frac{1}{2}$	20	$2\frac{3}{4}$	2	$1\frac{1}{4}$	5	$3\frac{1}{2}$	11	2	11	$1\frac{1}{4}$	22	$3\frac{1}{4}$	12	$1\frac{1}{4}$	21	$3\frac{3}{4}$	5	2
22	$3\frac{1}{2}$	18	$2\frac{1}{2}$	2	$2\frac{3}{4}$	16	$1\frac{1}{4}$	25	$3\frac{1}{2}$	14	1	15	$1\frac{1}{4}$	23	4	16	$1\frac{1}{4}$	24	$3\frac{3}{4}$	24	4
23	$3\frac{1}{4}$	15	1	1	$2\frac{3}{4}$	18	$2\frac{1}{4}$	23	$3\frac{1}{2}$	16	2	16	$1\frac{1}{4}$	25	$3\frac{1}{4}$	10	$1\frac{1}{4}$	7	$3\frac{3}{4}$	23	3
24	$3\frac{1}{4}$	14	1	22	$3\frac{3}{4}$	22	$3\frac{1}{4}$	9	$2\frac{1}{2}$	20	3	8	$1\frac{1}{4}$	20	$2\frac{1}{4}$	3	$2\frac{1}{4}$	4	1	8	2
25	3	11	1	19	$3\frac{3}{4}$	17	$3\frac{1}{4}$	1	$2\frac{1}{2}$	3	3	9	$1\frac{1}{4}$	9	$1\frac{1}{4}$	8	$1\frac{1}{4}$	17	$3\frac{3}{4}$	1	1

Als overschrijdingskans werd gevonden $k > 0{,}90$, zodat de correcties tot een bevredigend resultaat hebben geleid; voor het verdere verloop van de statistische bewerking zullen wij de gecorrigeerde waarnemingen beschouwen, als de keuringsresultaten van elf - voor practische doeleinden - even strenge beoordelaars.

Uit het bovenstaande blijkt tevens van welke orde van grootte de verschillen tussen de beoordelaars zijn.

3.2. De middelmatige eenheden

De vraagstelling hield in, dat slechts bij de waarnemingen van middelmatige eenheden beïnvloeding verwacht mocht worden door de voorafgaande waarnemingen.

In Tabel IV zijn vermeld de nummers van de eenheden en de daaraan toegekende (gecorrigeerde) puntentotalen, gerangschikt naar deze laatste grootheid. De in deze tabel voorkomende volgorde behoeft evenwel niet overeen te komen met de "werkelijke"

Tabel IV

Nr van eenheden, gerangschikt naar toegekende puntentotalen.

eenheid nr.	puntentotaal	eenheid nr.	puntentotaal	eenheid nr.	puntentotaal
14	14 3/4	4	20	17	37 3/4
16	14 3/4	3	21 1/4	25	38
8	15 1/4	9	23	19	39
2	15 1/2	20	23 1/2	22	39
15	15 1/2	1	25	23	39 1/2
10	16	18	30 1/2	7	40 1/2
11	17 1/4	6	33 1/4	24	40 3/4
13	17 1/2	21	34 1/2		
12	17 3/4	5	36 3/4		

volgorde. Het is uiteraard heel goed mogelijk, dat bijvoorbeeld 20 een slechtere eenheid is dan 1, ook al omdat het verschil tussen de puntentotalen van deze eenheden slechts zeer klein is. Maar naarmate het verschil tussen de puntentotalen a en b van twee eenheden A en B groter is, zal de uitspraak:" A en B komen in de tabel in de goede volgorde voor" een kleinere onbetrouwbaarheid bezitten. Voor het verdere verloop van het onderzoek zal gehandeld worden alsof de eenheden in tabel IV in de goede volgorde staan, althans voor zoverre hun puntentotalen in deze tabel 5 of meer verschillen. Hieraan zal wellicht niet steeds in werkelijkheid voldaan zijn, doch dit is geen bezwaar, daar afwijkingen hiervan bij de verdere statistische bewerking niet

vaker tot foute uitspraken zullen leiden dan de opgegeven onbetrouwbaarheden aangeven. Alleen kan het onderscheidingsvermogen van toetsen door eventuele verkeerde volgorden enigszins verlaagd worden.

Bij elke eenheid kan nu uit Tabel IV worden bepaald, welke eenheden beter en welke slechter zijn. Eenheden, die beter zijn dan minstens één van de 24 andere eenheden en tevens slechter zijn dan minstens één (andere) eenheid, zullen middelmatige eenheden genoemd worden (zie Tabel V).

Tabel V

Nummers van de middelmatige eenheden en van de eenheden, die beter, resp. slechter, zijn.

1 nummer middelmatige eenheid	2 nummers van eenheden, die beter zijn dan de overeenkomstige eenheden in kolom 1	3 nummers van eenheden, die slechter zijn dan de overeenkomstige eenheden in kolom 1
4	14-6	1-18-6-21-5-17-25-19-22-23-7-24
3	14-16-8-2-15-10	24-7-23-22-18-6-21-5-17-25-19
9	14-16-8-2-15-10-11-13-12	18-6-21-5-17-25-19-24-7-23-22
20	14-16-8-2-15-10-11-13-12	18-6-21-5-17-25-19-24-7-23-22
1	14-16-8-2-15-10-11-13-12-4	18-6-21-5-17-25-19-24-7-23-22
18	14-16-8-2-15-10-11-13-12-4 3-9-20-1	5-17-25-19-24-7-23-22
6	14-16-8-2-15-10-11-13-12-4 3-9-20-1	19-22-23-7-24
21	14-16-8-2-15-10-11-13-12-4 3-9-20-1	23-7-24

De nummers 4 en 21 zullen niet bij het verdere onderzoek betrokken worden, daar geen waarnemingen zijn verricht van eenheid nr. 4 voorafgegaan door een betere, en evenzo van nr. 21 voorafgegaan door een slechtere eenheid.

3.3. Toetsing van de invloed van voorafgaande waarnemingen

In deze paragraaf zal een methode van toetsen worden uitgewerkt, waarmee een aanwijzing kan worden gevonden van een invloed van de voorafgaande eenheid op een waarneming van een middelmatig exemplaar. Aan het einde van deze paragraaf zal worden aangegeven hoe men op analoge wijze een eventuele invloed van een reeks van goede of van slechte voorafgaande exemplaren kan onderzoeken.

Bij toepassing van beide methoden op één serie gegevens houde men er rekening mee, dat de verkregen uitkomsten sterk afhankelijk zijn.

Definiëert men:

gebeurtenis A: het voorafgaande exemplaar was beter,

en

gebeurtenis B: het voorafgaande exemplaar was slechter,

en verstaat men onder $F_i(x)$ de verdelingsfunctie van de beoordelingen van de i-de eenheid (i doorloopt de nummers van de middelmatige eenheden), dan kan voor elke i getoetst worden

$$H_0: \quad F_i(x|A) = F_i(x|B)$$

tegen de alternatieve hypothese

$$H_1: \quad F_i(x+d_i|A) = F_i(x-d_i|B), \quad d_i > 0.$$

De nulhypothese: er is geen beïnvloeding van de beoordelingen van middelmatige eenheden door voorafgaande slechtere en betere eenheden, wordt dus getoets tegen de alternatieve hypothese, dat men na een goed exemplaar een middelmatige eenheid eerder zal afkeuren dan wanneer een slecht exemplaar voorafgaat.

Een redelijke veronderstelling is nog, dat d_i, de mate van beïnvloeding niet afhangt van een bepaald exemplaar, maar dat (althans in eerste benadering) geldt:

$$d_i = d \quad \text{voor} \quad i = 3,9,20,1,18 \text{ en } 6$$

Uit de gecorrigeerde waarnemingen kan men met behulp van Tabel V voor iedere i een aantal waarnemingen bijeen zoeken van $\underline{x}$, onder de voorwaarde A en van $\underline{x}$, onder de voorwaarde B. Voor iedere i kan men de nulhypothese toetsen met behulp van de toets van WILCOXON (zie memorandum S 47 (M 7)). De resultaten van deze toetsingen kunnen worden gecombineerd tot één resultaat, en wel op de wijze als vermeld in memorandum S 102 (M 17b). Men kiest $c_i = \frac{1}{n_i m_i}$, waarbij n_i en m_i de omvang van de steekproeven zijn, behorende bij het i-de product.

De gang van de berekeningen is opgenomen in schema III, waarin de beschreven procedure gevolgd is, gebruik makend van de bekende formules van de toets van WILCOXON, die in het memorandum zijn vermeld.

De overschrijdingskans, die werd gevonden uit de voorgelegde gegevens, is 0,13, zodat de nulhypothese niet kan worden verworpen. Uit het gegeven materiaal blijkt dus niet dat een systematische beïnvloeding van de beoordelingen door de voorafgaande eenheid bestaat.

Schema III

Uitvoering van de toets van WILCOXON als toets voor de afwezigheid van beïnvloeding door de voorafgaande eenheid op de beoordeling van een middelmatige eenheid

i	voorkomende steekproefwaarden	aantal malen dat steekproefwaarden optreden onder de voorwaarden		$W_i - n_i m_i$	$\frac{W_i - n_i m_i}{n_i m_i}$	$\frac{\sigma_i^2}{(n_i m_i)^2}$
		A	B			
3	1¼		3			
	1½		1			
	1¾		1			
	2	2				
	2¼	1				
	3	1				
		4	5	-20	-1,00	0,1597
9	1¼	1				
	1¾		1			
	2	1	2			
	2¼	1	1			
	2½		1			
	3		1			
		3	6	+ 7	+0,39	0,1775
20	1¼		2			
	1¾	2				
	2¼	1	2			
	2¾		1			
	3	1				
		4	5	- 4	-0,20	0,1560
1	1	1				
	2		1			
	2¼	2	1			
	2¾	2				
	3	1				
		6	2	- 6	-0,50	0,2351
18	2¼	1	1			
	2½		1			
	2¾		1			
	3	2	1			
	3¼	1	1			
	3½		1			
		4	6	0	0	0,1472
6	1¼	1				
	2¾	1				
	3	1				
	3¼	1	1			
	3½	2				
	4	1				
		7	1	0	0	0,4183

$$\tilde{W} = \sum_i \frac{W_i - n_i m_i}{n_i m_i} = -1,31$$

$$\sigma^2\{\tilde{W}\} = 1,2938$$

$$\sigma\{\tilde{W}\} = 1,137$$

$$\frac{\tilde{W}}{\sigma} = -1,15 \quad k = 0,13 \text{ (eenzijdig)}$$

(kleine, d.w.z. grote negatieve waarden van $\frac{\tilde{W}}{\sigma}$, wijzen op beïnvloeding).

Wordt voor iedere i het verschil bepaald tussen de gemiddelden van de waarnemingen uit de beide steekproeven, en neemt men het gemiddelde van de gevonden verschillen, dan komt men tot een schatting van d. Deze is:

$$\hat{d} \approx 0{,}05$$

Vergelijkt men deze schatting met de verschillen, die tussen de beoordelaars bestaan, (zie 3,1) dan lijkt het in dit geval zeer aannemelijk, dat de verschillen in gestrengheid tussen de beoordelaars verre de beïnvloeding door de voorafgaande waarneming overheersen.

Een betrouwbaarheidsinterval voor d kan worden gevonden, echter blijkt dit bij uitrekenen te ontaarden tot $d \geq 0$, hetgeen reeds was ondersteld.

Indien men wil onderzoeken of een voorafgaande reeks van r opeenvolgende goede of slechte eenheden de beoordeling van een middelmatige eenheid beïnvloedt, kan dit op soortgelijke wijze geschieden. Definiëert men

gebeurtenis A' = de voorafgaande r exemplaren waren beter
gebeurtenis B' = de voorafgaande r exemplaren waren slechter,

dan kan men de hierboven voor $r=1$ beschreven methode eveneens toepassen. Men heeft dan echter kleinere steekproeven en dus minder materiaal voor de toepassing van deze toets.

4. Paarsgewijs samengevoegde beoordelingen

Uiteraard treedt meestal een verlies aan onderscheidingsvermogen op, als men de waarnemingen in grovere eenheden samen vat.

Om een indruk te krijgen van dit verlies van onderscheidingsvermogen zijn de beoordelingen paarsgewijs samengevoegd, d.w.z. 1 en 2 werden G = goed en 3 en 4 werden S = slecht genoemd. Indien een $2\frac{1}{2}$ voorkwam, werd met gelijke kansen geloot, of deze door een G of een S voorgesteld zou worden.

4.1. De mate van gestrengheid der beoordelaars

Daar in dit geval slechts twee waarnemingen voorkomen (G en S) zal nog meer dan in 3.1 rekening gehouden moeten worden met het optreden van gelijke waarnemingen.

Een goede eenheid zal vrij zeker door alle beoordelaars een G toegekend krijgen, evenzo zal een werkelijk slechte eenheid alle elf malen een S krijgen. Wil men dus de methode der m rangschikkingen toepassen als in 3.1, dan krijgt men rijen, waarvan alle elementen gelijk zijn aan G resp. S. Deze rijen zou men

nutteloze rijen kunnen noemen; gaat men na welke invloed zij op de methode der m rangschikkingen hebben, dan blijkt deze alleen daarin te bestaan, dat zij het onderscheidingsvermogen sterk verminderen. Op dezelfde wijze zijn rijen met tienmaal een S (resp. G) en eenmaal een G(resp. S) zeer schadelijk voor het onderscheidingsvermogen. Wij zullen om deze reden deze nutteloze rijen weglaten uit het schema. De grens tussen "nuttige" en "nutteloze" rijen is niet scherp te trekken en hierin schuilt dus een min of meer arbitrair element van deze methode.

De verschillen in gestrengheid tussen de beoordelaars kunnen nu, rekening houdend met bovenstaande wijzigingen, worden onderzocht met de methode der m rangschikkingen. Zie Schema IV.

Schema IV

Uitvoering van de toets voor de gelijkheid van de gestrengheid van beoordelaars.

beoordeelaars / eenheid	1	2	3	4	5	6	7	8	9	10	11	$n\tilde{\mu}_{2i}$
5	7	7	7	7	7	1½	7	7	7	7	1½	49,5
18	7½	7½	2	2	7½	7½	7½	7½	2	7½	7½	66
1	9	9	9	3½	9	3½	3½	3½	3½	9	3½	82,5
20	4½	4½	10	4½	10	10	4½	4½	4½	4½	4½	66
9	4½	10	4½	4½	10	4½	4½	4½	4½	10	4½	66
3	5	5	5	5	5	10½	5	5	5	5	10½	49,5
4	5	5	10½	5	10½	5	5	5	5	5	5	49,5
12	5	10½	5	5	5	5	5	5	5	5	10½	49,5
kolomtotalen	47½	58½	53	36½	64	47½	42	42	36½	53	47½	
gemiddelde	48	48	48	48	48	48	48	48	48	48	48	
afwijkingen van het gemiddelde	-½	+10½	+5	-11½	+16	-½	-6	-6	-11½	+5	-½	

$$\sum_i n\hat{\mu}_{2i} = 478,3 \qquad S = 753,5$$

$$W = \frac{753,5}{8.478,5} = 0,1968$$

$$V = 7.\frac{0,1968}{0,8032} = 1,715 \qquad \nu_1 = 10 \qquad \nu_2 = 70$$

$$\therefore k = 0,09$$

Op grond van de gevonden overschrijdingskans $k = 0,09$, mag de hypothese H_0 dus niet verworpen worden: er zijn geen verschillen in gestrengheid tussen de beoordelaars aan te tonen.

4.2. Beïnvloeding door voorafgaande eenheden.

Tabel VI is de tabel, waarin de nummers van de eenheden zijn opgenomen, in de volgorde van het aantal malen, dat hen een G is toegekend door de elf beoordelaars.

Tabel VI

Nummers van de eenheden in volgorde van het aantal malen dat een G is toegekend door de beoordelaars.

eenheid nr.	aantal malen G	eenheid nr.	aantal malen G	eenheid nr.	aantal malen G
24		6	1	12	9
7		5	2	13	
23		18	3	11	
22	0	1	6	10	10
19		20	8	15	
25		9	8	14	
17		3	9	2	
21	1	4	9	8	11
				16	

De eenheden nrs. 18,1,20 en 9 zijn voor het verdere onderzoek beschouwd als de middelmatige eenheden. Wij handelen nu weer, alsof een eenheid beter of slechter is dan een andere eenheid wanneer het aantal malen, dat de eerstgenoemde eenheid met G beoordeeld werd, meer dan één verschilt van dat van de ander.

Op dezelfde wijze als in 3.3 kan men voor iedere i de nulhypothese toetsen, dat de verdelingsfunctie van de beoordelingen, verricht aan de i-de eenheid onafhankelijk is van het feit, of de voorafgaande eenheid beter of slechter is. Daar X slechts twee waarden aan kan nemen: G en S, komt dit overeen met het toetsen van

$$H_0: \quad P[G|A] = P[G|B]$$

terwijl onder de alternatieve hypothese:

$$H_1: \quad P[G|A] < P[G|B]$$

De nulhypothese kan behalve met de toets van WILCOXON, ook getoetst worden met behulp van de methode der dubbele dichotomie (2x2-tabel; zie memorandum S 53 (M 23)), die in dit geval eveneens aequivalent is met de eerstgenoemde toets.

Door combineren van de toetsen werd op de in 3.1 beschreven wijze een toetsingsgrootheid gevonden, die gelijk is aan 0,00, zodat de overschrijdingskans k = 0,50 is. De nulhypothese mag dus nu zeker niet verworpen worden: men kan dus niet concluderen tot het bestaan van een beïnvloeding van opvolgende beoordelingen,

4.3. Vergelijking van de keuring met G-S t.o.v. 1,2,3 en 4.

Uit de toetsingsresultaten in 5.1 en 5.2 blijkt, dat het verlies aan onderscheidingsvermogen bij omschakeling op een G-S keuring inderdaad aanzienlijk genoemd mag worden. Het is dan ook aan te bevelen voor verder experimenteel werk, de beoordeling uit te drukken in de cijfers 1,2,3 en 4.

In hoeverre de goedgekeurde eenheden na een G-S keuring voldoen aan de eisen, die door het verdere fabricage proces of de vrijgeef-controle gesteld worden, is een vraag, die in dit rapport niet behandeld kan worden. Een vergelijking van Tabel IV met Tabel VI kan hierbij evenwel misschien reeds van nut zijn.

5. Conclusies

Kort samengevat kunnen de volgende conclusies geformuleerd worden:

1. Er treden systematische verschillen op tussen de beoordelaars wat de mate van gestrengheid betreft, waarmee zij de hun voorgelegde producten beoordelen.
2. De aanwezigheid van een invloed op de beoordeling van een middelmatig exemplaar door een voorafgaande goede of slechte eenheid kon niet aangetoond worden.
3. Een beoordeling in goed en slecht zal het onderscheidingsvermogen van de toetsen zo sterk beïnvloeden, dat het voor verder experimenteel onderzoek aanbeveling verdient, de keuringsresultaten te blijven uitdrukken in 1,2,3 of 4.

MATHEMATISCH CENTRUM
Amsterdam
Statistische Afdeling.

S47(M6)

Algemene gang van zaken bij het toetsen van een hypothese. [1]

De toetsing van een hypothese H_0 berust steeds op een aantal waarnemingen $x_1, x_2, \ldots, x_n$ van één of meer stochastische grootheden[2], of op enige groepen van waarnemingen (bv. twee steekproeven).

Bij een toets behoort een <u>toetsingsgrootheid</u> u (soms meer dan één), die een functie is van bovengenoemde stochastische grootheden en die, voor de waargenomen waarden $x_1, x_2, \ldots, x_n$ een waarde aanneemt, die berekend kan worden (bv.: het gemiddelde der waarnemingen, of de spreiding, of het verschil van de gemiddelden van twee waarnemingen).

De toetsingsgrootheid wordt steeds zo gekozen, dat men, op grond van de onderstelling, dat H_0 juist is, de waarschijnlijkheidsverdeling van deze grootheid kan berekenen.

Vervolgens kiest men een verzameling Z van mogelijke uitkomsten van u, en wel op zodanige wijze, dat de kans, dat u een in Z gelegen waarde aanneemt, onder de hypothese H_0, gelijk is aan een gegeven getal α, zodat Z dus van α afhankelijk is[3]. Z heet de <u>kritieke zône</u> van de toets, α de <u>onbetrouwbaarheidsdrempel</u> (Engels: level of significance). Voor α neemt men veelal de waarde 0,05 of 0,01.

Men verwerpt nu H_0 op grond van de waarnemingen x_1, x_2, ..., x_n, indien de bij deze waarnemingen behorende waarde van u in Z ligt. Dit wordt vaak uitgedrukt door te zeggen, dat het resultaat van het experiment "<u>significant</u>" is. De waarde van α moet dan echter worden vermeld. De kans, dat dit zal gebeuren, is, indien H_0 juist is, gelijk aan α. Derhalve is α de <u>kans op ten onrechte verwerping van de juiste hypothese</u>, ook de kans op een fout van de eerste soort genoemd. Indien men deze methode toepast, met $\alpha = 0{,}05$ resp. 0,01, zal men in gemiddeld ongeveer één op 20 resp. op 100 van de gevallen, <u>waarin de hypothese die men toetst juist is</u>, deze toch verwerpen.

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.

2) Een stochastische grootheid is een grootheid, die een waarschijnlijkheidsverdeling bezit, of, anders gezegd, een grootheid, die voor de elementen van een <u>collectie</u>(<u>universum</u>,<u>populatie</u>) gedefinieerd is en daarop allerlei waarden aanneemt. Stochastische grootheden worden aangegeven door <u>onderstreepte</u> letters.

3) Soms kan men slechts bereiken, dat deze kans $\leq \alpha$ is.

De toetsingstheorie biedt in het algemeen geen mogelijkheid om tot aanvaarding van een hypothese te komen. Indien een bepaalde hypothese H_0 niet verworpen kan worden, is dit gewoonlijk met een hele verzameling van hypothesen tegelijk het geval. Niet-verwerpen staat dus niet gelijk met aanvaarden.

Wel zal men vaak in de loop van een statistische analyse bepaalde onderstellingen, die plausibel schijnen en voor de verdere analyse van nut zijn, toetsen, alvorens ze bij de verdere bewerking van het materiaal te gebruiken. Worden zij dan op grond van de toets niet verworpen, dan houdt dit in zo verre een rechtvaardiging van die onderstellingen in, dat een grote afwijking door de toets veelal wel zou zijn ontdekt. Indien men dan verder de onderstellingen gebruikt, verwaarloost men eventueel aanwezige afwijkingen van onbekende grootte, die echter niet zo groot zijn, dat zij door de toets zijn ontdekt.

Vele toetsen gelden zelf alleen onder bepaalde onderstellingen omtrent de waarschijnlijkheidsverdelingen der stochastische grootheden, waarvan waarnemingen zijn verricht. Deze nevenvoorwaarden dienen steeds uitdrukkelijk te worden vermeld en, zo mogelijk, zelf te worden getoetst.

In plaats van de onbetrouwbaarheidsdrempel α wordt vaak bij de uitslag van een toetsing de overschrijdingskans k opgegeven; dit is de kleinste waarde van α, waarbij in het betrokken geval, nog tot verwerping van H_0 zou zijn overgegaan; anders gezegd: de kleinste α, waarvoor de gevonden waarde der toetsingsgrootheid nog juist in de (bij α behorende) kritieke zône Z ligt. Wordt dus de waarde k opgegeven en werkt men met onbetrouwbaarheidsdrempel α, dan wordt H_0 verworpen, indien $k \leq \alpha$ is.

Voor het onderscheid tussen één- en tweezijdige toetsing en de keuze tussen deze twee mogelijkheden vergelijke men bv. de tweede hieronder gegeven litteratuurplaats. Wij moeten hier volstaan met de opmerking, dat éénzijdige toetsing veelal eerder tot verwerping van H_0 leidt, maar dat deze slechts onder bijzondere omstandigheden kan worden toegepast.

Litteratuur:

J. Neyman, First course in probability and statistics, New York, 1950, Chapter 5.

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

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O.
Statistische Afdeling
Rapport S 47 (M 7)

De toets van Wilcoxon [1]

Deze methode dient tot het toetsen van de hypothese H_o, inhoudende, dat twee steekproeven $x_1, x_2, \ldots, x_m$ en $y_1, y_2, \ldots, y_n$ afkomstig zijn uit één collectie (ook wel populatie of universum genaamd).

De gegeven waarnemingsreeks kan dan worden samengevat als aangegeven in tabel I.

Tabel I

Voorkomende steekproef-waarden	Aantal malen dat z_i optreedt bij		totaal
	$\underline{y}$	$\underline{x}$	
z_1	$\underline{b}_1$	$\underline{a}_1$	$\underline{t}_1$
z_2	$\underline{b}_2$	$\underline{a}_2$	$\underline{t}_2$
.	.	.	.
.	.	.	.
.	.	.	.
z_k	$\underline{b}_k$	$\underline{a}_k$	$\underline{t}_k$
totaal	n	m	N

Hierin zijn $z_1, z_2, \ldots, z_k$ de voorkomende steekproefwaarden, gerangschikt naar opklimmende grootte, n en m de steekproefgrootten, $t_1, t_2, \ldots, t_k$ de grootten der groepen gelijke waarnemingen en $\underline{a}_i$ [2] (resp. $\underline{b}_i$) het aantal malen dat de waarde z_i in de eerste (resp. tweede) steekproef optreedt; $\underline{a}_i$ kan dus de waarden $0, 1, \ldots, t_i$ aannemen ($i = 1, 2, \ldots, k$).

Voor het toetsen van de hypothese H_o wordt gebruik gemaakt van een toetsingsgrootheid $\underline{W}$ [2], die gelijk is aan twee maal

---

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.
2) Een stochastische grootheid is een grootheid, die een waarschijnlijkheidsverdeling bezit, hetgeen gewoonlijk blijkt uit het feit, dat bij herhaalde waarneming onder dezelfde omstandigheden verschillende uitkomsten verkregen worden. Stochastische grootheden worden door onderstreepte letters aangegeven; dezelfde letters, niet onderstreept, worden vaak gebruikt voor waarden, die zij aan kunnen nemen of aangenomen hebben.

het aantal paren waarnemingen (x_i, y_j), waarvoor $x_i > y_j$ is, vermeerderd met het aantal paren (x_k, y_ℓ) waarvoor $x_k = y_\ell$ is.[3)]

Het is duidelijk, dat W een kleine waarde zal aannemen, indien de x-waarden overwegend kleiner zijn dan de y-waarden, een grote, indien het omgekeerde het geval is en een intermediaire waarde, indien geen van deze beide situaties zich voordoet.

Indien de hypothese H_0 juist is, dus als de twee steekproeven uit dezelfde verdeling afkomstig zijn, kan de verdeling van de toetsingsgrootheid W berekend worden. Deze verdeling heeft de volgende eigenschappen:

1. De verdeling van W is discreet, want W kan alleen gehele waarden aannemen.
2. De verdeling van W is in het algemeen niet symmetrisch, behalve,

a) als $m = n$,

b) als $t_1 = t_k$, $t_2 = t_{k-1}$ enz.

3. De verwachting (het theoretische gemiddelde) μ, van de verdeling van W is gelijk aan het product der steekproefgrootten

$$\mu = mn. \tag{1}$$

4. De variantie van de verdeling van W wordt gegeven door:

$$\sigma^2 = \frac{mn\,(N^3 - D)}{3N(N-1)}, \tag{2}$$

waarin $D = t_1^3 + t_2^3 + \ldots + t_k^3$.

Bevatten de beide steekproeven geen gelijke waarnemingen, dan wordt $D = N$ en dan is dus

$$N^3 - D = N^3 - N = N(N-1)(N+1)$$

en in dit geval wordt de variantie

$$\sigma^2 = \tfrac{1}{3}\,mn(N+1) \quad \text{(geen gelijke waarnemingen)}. \tag{3}$$

Wij zagen reeds dat de toetsingsgrootheid W een kleine waarde aan zal nemen, indien de x-waarden overwegend kleiner zijn dan de y-waarden een een grote waarde, indien het omgekeerde het geval is. Als tweezijdige kritieke zone nemen wij dus grote waarden van $|W - \mu|$.

---

3) Gewoonlijk wordt voor de toets van WILCOXON de toetsingsgrootheid $U = \frac{1}{2}W$ gebruikt. Om gebroken getallen te vermijden, wordt hier W ingevoerd (zie [4], [6]).

Als m en n groot zijn en onderling niet te veel verschillen en als bovendien de groepen gelijke waarnemingen niet te veel in omvang verschillen, kan men gebruik maken van het feit, dat de grootheid $\underline{W}$, als de hypothese H_0 juist is, bij benadering normaal verdeeld is met gemiddelde en variantie volgens (1) en (2).

De tweezijdige overschrijdingskans k, behorende bij een waarde W van $\underline{W}$ is dus

$$k = \frac{2}{\sqrt{2\pi}} \int_{\frac{|W-\mu|-1}{\sigma}}^{\infty} e^{-\frac{1}{2}u^2} \, du \qquad 4)$$

en kan in een tabel van de normale verdeling worden gevonden.

Voor $m \leq n \leq 10$, resp. voor $m+n \leq 40$, $n \geq 11$ en $m \leq n$ zijn tabellen van de eenzijdige overschrijdingskansen, resp. linker kritieke waarden beschikbaar (zie [6] tabel I en II). Deze tabellen gelden strikt genomen alleen als er geen gelijke waarnemingen zijn, maar zij geven voor het geval van weinig gelijke waarnemingen veelal een goede benadering.

Indien m en n klein zijn en bovendien onder de waarnemingen grote groepen gelijken voorkomen, kan men de overschrijdingskansen noch met behulp van de normale benadering van de verdeling van $\underline{W}$, noch met behulp van de genoemde tabellen bepalen. Hoe men in zulke gevallen dient te handelen, staat beschreven in [1].

Opmerking: Als er gelijke waarnemingen zijn, dan geldt

$$\frac{1}{3} m n (N+1) > \frac{m n (N^3 - D)}{3 N (N-1)} .$$

Dus als wij, in de formule voor de variantie van $\underline{W}$, de gelijke waarnemingen niet in rekening brengen, vinden wij een te grote waarde voor deze variantie en dus een te grote waarde voor de overschrijdingskans. Is deze te grote waarde voor de overschrijdingskans $\leq \alpha$ dan wordt H_0 dus verworpen en hoeft men de berekening der derde machten niet uit te voeren.

---

4) De term -1 in de teller van $\frac{|W-\mu|-1}{\sigma}$ is de z.g. continuïteitscorrectie (zie [6], pag. 16).

Literatuur

[1] van Eeden, Constance en Ir Doraline Wabeke, Handleiding voor de toets van WILCOXON (vervolg): Exacte behandeling als er gelijke waarnemingen zijn. Rapport S 176 (M 65A) van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam, 1955.

[2] Hemelrijk, J., Note on WILCOXON's two sample test when ties are present, Ann.Math.Stat. 23 (1952), 133-135.

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

[4] van der Vaart, H.R., Gebruiksaanwijzing voor de toets van WILCOXON, Rapport S 32 (M 4) van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam, 1950.

[5] Wilcoxon, F., Individual comparisons by ranking methods, Biometrics 1 (1945), 80-83.

[6] Wabeke, Ir Doraline en Constance van Eeden, Handleiding voor de toets van WILCOXON, Rapport S 176 (M 65) van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam 1955.

MATHEMATISCH CENTRUM
2de Boerhaavestr.49
A M S T E R D A M O.
Statistische Afdeling
S47 (M14)

Methode der m rangschikkingen [1]

Een duidelijke voorstelling van deze toetsingsmethode verkrijgt men door n elementen te beschouwen, die een bepaald kenmerk, eventueel in verschillende mate, bezitten. dit kenmerk wordt door m waarnemers beoordeeld en ieder van deze waarnemers rangschikt deze n elementen volgens zijn beoordeling naar opklimmende waardering. Op deze wijze ontstaan m rijen van rangschikkingen. We willen nu een maat aangeven voor de overeenstemming tussen deze rangschikkingen, m.a.w. een maat voor de overeenstemming tussen de m beoordelingen. De hypothese H_0, die met deze methode getoetst kan worden, houdt in dat er geen overeenstemming tussen de waarnemers bestaat; precieser gezegd, dat alle rangschikkingen onafhankelijk van elkaar op toevallige wijze zijn ontstaan. Dit is b.v. het geval, als het betrokken kenmerk in werkelijkheid voor alle elementen dezelfde waarde bezit.

We kunnen de afleiding voor de maat van overeenstemming het eenvoudigst geven aan de hand van een voorbeeld.

elementen	A	B	C	D	E	F
rangnummers toegekend door waarnemer						
a	5	4	1	6	3	2
b	2	3	1	5	6	4
c	4	1	6	3	2	5
d	4	3	2	5	1	6
	15	11	10	19	12	17

De som van alle rangnummers is $\frac{1}{2}nm(n+1)$. Onder de hypothese H_0 is het theoretische gemiddelde van iedere kolom: $\frac{1}{2}m(n+1)$

We beschouwen nu de afwijkingen van dit gemiddelde. In ons voorbeeld is het theoretisch kolomgemiddelde gelijk aan 14. De afwijkingen daarvan zijn

1	−3	−4	5	−2	3

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid

De som der kwadraten van deze afwijkingen noemen wij S.
In ons voorbeeld is S = 64.
Als alle m rangschikkingen gelijk zijn wordt het maximum van S bereikt.
Dit maximum is $\frac{1}{12}\, m^2(n^3-n)$.
We definiëren nu als coëfficiënt van overeenstemming

$$W = \frac{12\ S}{m^2(n^3-n)}$$

In ons voorbeeld is $W = \frac{12 \times 64}{16 \times 210} = 0{,}229$.

W varieert dus tussen 0 en 1.

De verdeling van $\underline{S}$ onder de hypothese H_o is exact berekend voor een aantal waarden van n en m [1] , terwijl voor grote m en n benaderingen bekend zijn.

De meeste gebruikelijke benaderingen zijn de volgende.

1°. De χ^2-benadering:

$\chi_r^2 = m(n-1)\underline{W} = \frac{12\ \underline{S}}{mn(n+1)}$ heeft voor $m \to \infty$ een χ^2-verdeling met n-1 vrijheidsgraden ([1] pg. 84 [2] pg. 36-37).

2°. De z-benadering:

$\underline{V} = (m-1)\ \frac{\underline{W}}{1-\underline{W}}$ is bij benadering verdeeld als $\underline{F} = e^{2z}$

($\underline{F}$ is de $\underline{F}$ van Snedecor, $\underline{z}$ de $\underline{z}$ van Fisher) met

$\nu_1 = n-1-\frac{2}{m}$

$\nu_2 = (m-1)\,\nu_1$ vrijheidsgraden ([1] pg. 84 [2] pg.33-36

Met behulp van de verdelingen van $\underline{S}$ of $\underline{W}$ onder de hypothese H_o, kan deze hypothese getoetst worden, waarbij H_o verworpen wordt als $\underline{W}$ waarden dichtbij 1 (resp. $\underline{S}$ dichtbij $\frac{1}{12}\, m^2(n^3-n)$) aanneemt, de kritieke zône is dus van de vorm $W \geqq W_o$ (resp. $S \geqq S_o$).

Het kan voorkomen dat de waarnemers geen onderscheid ontdekken in de mate waarin verschillende elementen het kenmerk bezitten. Ze geven deze elementen dan gelijke rangnummers.

Veronderstel, dat door een waarnemer geen onderscheid wordt gemaakt tussen de elementen, die de rangnummers 3 t/m 6 moeten dragen. Dan wordt als rangnummer van iedere van deze elementen het gemiddelde van de rangnummers $\frac{1}{4}\,(3 + 4 + 5 + 6) = 4\frac{1}{2}$ gebruikt.

Daar het maximum van $\underline{S}$ nu verandert, moeten wij een correctie op de formule voor $\underline{W}$ toepassen. Deze vindt men in [1] (pg. en [2] (pg. 28-30). Eveneens veranderen dan de formules voor χ^2-benadering ([1] pg. 86, [2] pg. 37) en voor de z-benade ([1] pg. 86 [2] pg. 34), doch deze correcties zijn van wei betekenis, tenzij het aantal gelijken groot is.

Literatuur: [1] M.G.Kendall, Rank correlation methods, London 1948, Hoofdstuk 6, pag. 80.
Tabel van de verdelingsfunctie van S voor:

n = 3	m = 2 t/m 10
n = 4	m = 2 t/m 6
n = 5	m = 3

op pag. 146-149.
Tabel van de waarden van S, waarvan de overschrijdingskansen onder de hypothese H_o gelijk zijn aan 0,05 of 0,01, berekend met behulp van de z-benadering voor:

n = 3	m = 8,9,10,12,14,15,16,18,20
n = 4	m = 4,5,6,8,10,15,20
n = 5 t/m 7	m = 3,4,5,6,8,10,15,20

op pag. 150.

[2] Ph.van Elteren, Methode der m rangschikkingen, Cursus "Parametervrije Methoden", Hoofdstuk II, Rapport S 59, Mathematisch Centrum (1951), Blz. 18-45.

MATHEMATISCH CENTRUM,
2de Boerhaavestr. 49,
A m s t e r d a m - O.

Statistische Afdeling

Rapport S 102 (M 17b)

Het combineren van onafhankelijke toetsen (aanvulling) [1].

In memorandum S 73 (M 17a) wordt een methode voor combinatie van onafhankelijke toetsen behandeld, waarbij het nodig is de overschrijdingskans van iedere toets te bepalen. In vele gevallen kan men de combinatie ook direct op de afzonderlijke toetsingsgrootheden baseren en dit verdient zelfs de voorkeur.

Wij beschouwen hier het geval, dat een bepaalde toets moet worden toegepast op een heterogeen materiaal. Dit materiaal wordt dan eerst verdeeld in h homogeen geachte groepen. Het aantal waarnemingen van de i^e groep zij n_i en de toetsingsgrootheid $\underline{t}_i$ [2]. Laat verder gegeven zijn, dat de verdeling van $\underline{t}_i$ onder de getoetste hypothese (voor de i^e groep aangeduid door H_i) voor grote n_i asymptotisch normaal [3] is, met bekende verwachting μ_i en bekende spreiding σ_i. Aan deze voorwaarden is o.a. voldaan, indien wij te doen hebben met toetsen van WILCOXON, rangcorrelatietoetsen van KENDALL of SPEARMAN, tekentoetsen enz.

Wij toetsen met al de hier te behandelen gecombineerde methoden de hypothese H, dat voor iedere groep de desbetreffende hypothese H_i geldt, terwijl de groepen onderling onafhankelijk zijn. De toetsen verschillen echter ten aanzien van de alternatieve (van H afwijkende) hypothesen waarvoor zij gevoelig [4] zijn.

De meest gebruikelijke toetsingsgrootheden van gecombineerde toetsen zijn van de gedaante:

$$\underline{T} = \sum_{i=1}^{h} c_i (\underline{t}_i - \mu_i)$$

waarin de letters c_i ($i = 1, 2, \ldots, h$) constanten voorstellen, die voor ieder van de combinatiemethoden op een bepaalde wijze

---

1) Dit memorandum dient slechts ter oriëntatie en streeft niet naar volledigheid of volledige exactheid. Het is bedoeld als een aanvulling op Rapport S 73 (M 17a).

2) De onderstreping geeft aan dat een toetsingsgrootheid stochastisch is, d.w.z. een waarschijnlijkheidsverdeling bezit.

3) Dit houdt in dat $\underline{t}_i$ een waarschijnlijkheidsverdeling heeft, die als n_i toeneemt, steeds minder van een normale verdeling (verdeling van Gauss) afwijkt.

4) Een toets van hypothese H is gevoelig ten opzichte van een alternatieve hypothese H', als de kans dat H verworpen wordt, indien H' juist is, groot is.

gekozen worden. Onder de hypothese H zal $\underline{T}$ asymptotisch (voor grote h en/of grote n_i) normaal verdeeld zijn met verwachting 0 en spreiding $\sigma = \sqrt{\sum_{i=1}^{h} c_i^2 \sigma_i^2}$. De dubbele overschrijdingskans van een gevonden waarde T van $\underline{T}$ is dus bij benadering gelijk aan:

$$\frac{2}{\sqrt{2\pi}} \int_{|\frac{T}{\sigma}|}^{\infty} e^{-\frac{1}{2}x^2} \, dx$$

en kan bepaald worden met behulp van een tabel van de normale verdeling. Indien de dubbele overschrijdingskans kleiner is dan de onbetrouwbaarheidsdrempel α , zal men H verwerpen.

Wij geven hier 3 combinatiemethoden van dit type:

<u>Methode 1</u>: $c_1 = c_2 = \ldots\ldots = c_h = 1$

dus: $\underline{T} = \sum_{i=1}^{h} \underline{t}_i - \sum_{i=1}^{h} \mu_i$; $\sigma = \sqrt{\sum_{i=1}^{h} \sigma_i^2}$

<u>Methode 2</u>: $c_1 = \frac{1}{\sigma_1}$, $c_2 = \frac{1}{\sigma_2}$, ... , $c_h = \frac{1}{\sigma_h}$

dus: $\underline{T} = \sum_{i=1}^{h} \frac{\underline{t}_i - \mu_i}{\sigma_i}$; $\sigma = \sqrt{h}$

<u>Methode 3</u>: $c_1 = \frac{1}{n_1}$, $c_2 = \frac{1}{n_2}$, ... , $c_h = \frac{1}{n_h}$

dus: $\underline{T} = \sum_{i=1}^{h} \frac{\underline{t}_i - \mu_i}{n_i}$; $\sigma = \sqrt{\sum_{i=1}^{h} \left(\frac{\sigma_i}{n_i}\right)^2}$

Deze methoden zijn alleen gevoelig ten aanzien van alternatieve hypothesen volgens welke de grootheden $\underline{t}_i$ verdelingen hebben die, voor zover zij afwijken van de verdelingen onder de corresponderende hypothesen H_i, dit over het algemeen in dezelfde richting doen. Men zal dan <u>methode 1</u> bij voorkeur toepassen als men aan de $\underline{t}_i$ met een kleine spreiding (in de regel zullen dat de $\underline{t}_i$ van kleine groepen zijn) een geringer gewicht wil toekennen dan aan de $\underline{t}_i$ met een grote spreiding De <u>methoden 2 en 3</u> zijn te gebruiken als men aan de verschillende groepen waarnemingen, ongeacht hun grootte, een ongeveer gelijke invloed op het resultaat wil toekennen. De keuze tussen deze twee methoden hangt verder van hier niet te behandelen theoretische overwegingen af (zie literatuur [1]).

Indien men verwacht dat mogelijke verschuivingen van de verdelingen der $\underline{t}_i$ in beide richtingen kunnen liggen, verdient het de voorkeur om gebruik te maken van de volgende toetsingsgrootheid:

$$\sum_{i=1}^{h} \left(\frac{\underline{t}_i - \mu_i}{\sigma_i}\right)^2 \qquad (\underline{\text{methode 4}})$$

gekozen worden. Onder de hypothese H zal $\underline{T}$ asymptotisch (voor grote h en/of grote n_i) normaal verdeeld zijn met verwachting 0 en spreiding $\sigma = \sqrt{\sum_{i=1}^{h} c_i^2 \sigma_i^2}$. De dubbele overschrijdingskans van een gevonden waarde T van $\underline{T}$ is dus bij benadering gelijk aan:

$$\frac{2}{\sqrt{2\pi}} \int_{\left|\frac{T}{\sigma}\right|}^{\infty} e^{-\frac{1}{2}x^2} \, dx$$

en kan bepaald worden met behulp van een tabel van de normale verdeling. Indien de dubbele overschrijdingskans kleiner is dan de onbetrouwbaarheidsdrempel α, zal men H verwerpen.

Wij geven hier 3 combinatiemethoden van dit type:

Methode 1: $c_1 = c_2 = \cdots\cdots = c_h = 1$

dus: $\underline{T} = \sum_{i=1}^{h} \underline{t}_i - \sum_{i=1}^{h} \mu_i \quad ; \quad \sigma = \sqrt{\sum_{i=1}^{h} \sigma_i^2}$

Methode 2: $c_1 = \frac{1}{\sigma_1}$, $c_2 = \frac{1}{\sigma_2}$, ..., $c_h = \frac{1}{\sigma_h}$

dus: $\underline{T} = \sum_{i=1}^{h} \frac{\underline{t}_i - \mu_i}{\sigma_i} \quad ; \quad \sigma = \sqrt{h}$

Methode 3: $c_1 = \frac{1}{n_1}$, $c_2 = \frac{1}{n_2}$, ..., $c_h = \frac{1}{n_h}$

dus: $\underline{T} = \sum_{i=1}^{h} \frac{\underline{t}_i - \mu_i}{n_i} \quad ; \quad \sigma = \sqrt{\sum_{i=1}^{h} \left(\frac{\sigma_i}{n_i}\right)^2}$

Deze methoden zijn alleen gevoelig ten aanzien van alternatieve hypothesen volgens welke de grootheden $\underline{t}_i$ verdelingen hebben die, voor zover zij afwijken van de verdelingen onder de corresponderende hypothesen H_i, dit over het algemeen in dezelfde richting doen. Men zal dan methode 1 bij voorkeur toepassen als men aan de $\underline{t}_i$ met een kleine spreiding (in de regel zullen dat de $\underline{t}_i$ van kleine groepen zijn) een geringer gewicht wil toekennen dan aan de $\underline{t}_i$ met een grote spreiding. De methoden 2 en 3 zijn te gebruiken als men aan de verschillende groepen waarnemingen, ongeacht hun grootte, een ongeveer gelijke invloed op het resultaat wil toekennen. De keuze tussen deze twee methoden hangt verder van hier niet te behandelen theoretische overwegingen af (zie literatuur [1]).

Indien men verwacht dat mogelijke verschuivingen van de verdelingen der $\underline{t}_i$ in beide richtingen kunnen liggen, verdient het de voorkeur om gebruik te maken van de volgende toetsingsgrootheid:

$$\sum_{i=1}^{h} \left(\frac{\underline{t}_i - \mu_i}{\sigma_i}\right)^2 \qquad \text{(methode 4)}$$

Deze grootheid is onder de hypothese H asymptotisch verdeeld volgens een χ^2-verdeling met h vrijheidsgraden. De overschrijdingskans van een gevonden waarde van deze grootheid kan dus met behulp van een tabel van de χ^2-verdeling bepaald worden.

De toets, behandeld in memorandum S 73 (M 17a) par. 1, waarbij men het product van linkszijdige en product van alle rechtszijdige overschrijdingskansen bepaalt en het kleinste van deze twee producten gebruikt, heeft betrekking op dezelfde gevallen als de hier behandelde methoden 2 of 3, terwijl de methode, behandeld in S 73 (M 17a) par. 2, berustend op het product van de tweezijdige overschrijdingskansen, meer overeenkomt met methode 4. Men mag echter verwachten, dat, zo aan de asymptotische normaliteit der t_i voldaan is, de in dit memorandum behandelde methodenscherper zijn dan de toetsen behandeld in S 73 (M 17a).

Literatuur:

1 C.van Eeden, Methoden voor het vergelijken, toetsen en schatten van onbekende kansen, Rapport S 115 (M 45) van het Mathematisch Centrum (1953).

2 ----------, Trendtoets met behulp van rangcorrelatie, Memorandum S 73 (M 13a). (Voorbeeld van toepassing van methode 1.)

3 Dr J.Hemelrijk, Het combineren van onafhankelijke toetsen, Memorandum S 73 (M 17a).

ATHEMATISCH CENTRUM,
de Boerhaavestraat 49,
m s t e r d a m - O.

tatistische Afdeling
S 53 (M 23).

Toetsing van de hypothese $p_1 = p_2$ met behulp van een 2 x 2-tabel [1)]

Wij beschouwen twee reeksen van onafhankelijke experimenten, waarbij ieder experiment van de ene reeks één van de twee resultaten A of $\overline{A}$ (non-A) heeft en ieder experiment van de tweede reeks één van de beide resultaten B of $\overline{B}$ (hierbij kan A=B zijn). Daarbij wordt ondersteld, dat bij ieder der experimenten van de ene reeks de kans op A gelijk aan p_1 (en dus de kans op $\overline{A}$ gelijk aan $1-p_1$) is en bij ieder der experimenten van de tweede reeks de kans op B gelijk aan p_2 (en dus de kans op $\overline{B}$ gelijk aan $1-p_2$). De te toetsen hypothese luidt nu:

$$H_o : p_1 = p_2.$$

Indien de eerste reeks uit n en de tweede reeks uit m waarnemingen bestaat, waaronder a (resp. b) maal A (resp. B) voorkomt, kunnen deze gegevens in de volgende 2 x 2-tabel worden samengevat:

	A resp. B	$\overline{A}$ resp. $\overline{B}$	totaal
eerste reeks	a	c	n
tweede reeks	b	d	m
totaal	r	s	N

Als toetsingsgrootheid wordt a, het aantal malen A in de eerste reeks waarnemingen, gebruikt. Indien H_o, juist is bezit deze grootheid onder de voorwaarde, dat r de bij het experiment gevonden waarde aanneemt, de volgende waarschijnlijkheidsverdeling: de kans, dat een bepaalde waarde a aangenomen wordt, is gelijk aan:

$$\frac{\binom{n}{a}\binom{m}{b}}{\binom{N}{r}}$$

Als kritieke zone worden de waarden van a met de kleinste waarschijnlijkheden bijeengezocht, tot de gekozen onbetrouwbaarheidsdrempel het toevoegen van een nieuwe waarde verhindert (bij éénzijdige toetsing bestaat de kritieke zone uitsluitend uit grote of uitsluitend uit kleine waarden van a). De overschrijdingskans, behorende bij de gevonden waarde van a, is gedefinieerd als de som van alle waarschijnlijkheden van bovenstaande verdeling, die hoogstens gelijk aan de waarschijn-

---

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.

lijkheid van de gevonden waarde zijn (bij éénzijdige toetsing echter gelijk aan de som van de waarschijnlijkheden van alle waarden die groter of gelijk aan de gevondene, of van alle waarden, die kleiner of gelijk aan de gevondene zijn). Deze exacte toetsingsmethode voor H_o is afkomstig van R.A.FISHER.

Indien n en m zo groot zijn, dat deze exacte berekening te omslachtig wordt, maakt men gebruik van de volgende benadering:

Gemiddelde en spreiding van de grootheid a zijn (indien H_o juist is):

$$\frac{nr}{N} \text{ resp. } \sqrt{\frac{nmrs}{N^2(N-1)}} .$$

Men gebruikt dan in plaats van de exacte waarschijnlijkheidsverdeling van a de normale verdeling met hetzelfde gemiddelde en dezelfde spreiding en in plaats van de gevonden waarde van a neemt men het getal dat $\frac{1}{2}$ dichter bij het gemiddelde ligt dan deze gevonden waarde (dit laatste is de z.g. "continuïteitscorrectie", die bij toenemende n en m weldra verwaarloosd kan worden). Bij positieve $a - \frac{nr}{N}$ berekent men dus:

$$a^* = \frac{a - \frac{nr}{N} - \frac{1}{2}}{\sqrt{\frac{nmrs}{N^2(N-1)}}} = \frac{aN - nr - \frac{1}{2}N}{\sqrt{\frac{nmrs}{N-1}}} = \frac{ad - bc - \frac{1}{2}N}{\sqrt{\frac{nmrs}{N-1}}}$$

en bij negatieve $a - \frac{nr}{N}$ berekent men:

$$a^* = \frac{a - \frac{nr}{N} + \frac{1}{2}}{\sqrt{\frac{nmrs}{N^2(N-1)}}} = \frac{ad - bc + \frac{1}{2}N}{\sqrt{\frac{nmrs}{N-1}}} .$$

De overschrijdingskans wordt nu opgezocht in een tabel der normale verdeling met gemiddelde 0 en spreiding 1. De rechts-éénzijdige (resp. links-éénzijdige) overschrijdingskans is het oppervlak rechts (resp. links) gelegen van a^*. De tweezijdige overschrijdingskans is twee maal het oppervlak der normale verdeling dat rechts van $\frac{|ad - bc| - \frac{1}{2}N}{\sqrt{\frac{nmrs}{N-1}}}$ ligt.

Literatuur.

R.A.Fisher, Statistical Methods for Research Workers, London 1948, p. 96. Opmerking: Fisher gebruikt hier de éénzijdige overschrijdingskans.

J.Hemelrijk, Waarschijnlijkheidsrekening en Statistiek, Vacantiecursus Mathematisch Centrum, Amsterdam 1950, § 4.