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 176 (M 65A)

Handleiding voor de toets van WILCOXON (vervolg): Exacte behandeling als er gelijke waarnemingen zijn

door

Constance van Eeden

en

Ir Doraline Wabeke

Juli 1955

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.

1. Inleiding

In rapport S 176(M 65) (par. 7 en 8) werd aangegeven hoe men voor de toets van WILCOXON, met behulp van tabellen en benaderingen, de kritieke zones en overschrijdingskansen kan vinden. In par. 9 werd opgemerkt, dat men deze tabellen en benaderingen niet kan gebruiken, als m en n klein zijn en bovendien de groepen gelijke waarnemingen veel in grootte verschillen; in deze gevallen zal men de verdeling van $\underline{W}$ onder de getoetste hypothese dus exact moeten berekenen.

In dit rapport zullen wij aangeven hoe men deze exacte verdeling kan berekenen en tevens hoe men in deze gevallen de kritieke zones en overschrijdingskansen definiëert.

De berekeningen vergen aanzienlijk veel meer tijd dan de niet-exacte en zullen dus in de regel alleen in uitzonderingsgevallen uitgevoerd worden.

2. Berekening van de exacte verdeling van $\underline{W}$ onder de getoetste hypothese H_o

De twee steekproeven $x_1, x_2, \ldots, x_m$ en $y_1, y_2, \ldots, y_n$ kan men als volgt samenvatten:

Tabel 1

Waarnemingsschema

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

Hierin zijn $z_1, z_2, \ldots, z_k$ de voorkomende steekproefwaarden gerangschikt naar opklimmende grootte; m en n de steekproefgrootten; $\underline{t}_1, \underline{t}_2, \ldots, \underline{t}_k$ de grootten der groepen gelijke waarnemingen en $\underline{a}_i$ (resp. $\underline{b}_i$) het aantal malen, dat x_i in de eerste (resp. tweede) steekproef optreedt.

De kans dat $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_k$ de waarden $a_1, a_2, \ldots, a_k$ aannemen, onder de hypothese H_o en bij de gevonden waarden voor $\underline{t}_1, \underline{t}_2, \ldots, \underline{t}_k$ wordt gegeven door de $(k-1)$-dimensionale hypergeometrische verdeling:

$$(1) \qquad P[\underline{a}_1 = a_1, \underline{a}_2 = a_2, \ldots, \underline{a}_k = a_k \mid t_1, t_2, \ldots, t_k; H_o] = \frac{\binom{t_1}{a_1}\cdot\binom{t_2}{a_2}\cdots\binom{t_k}{a_k}}{\binom{N}{m}} \quad 1)$$

Nu worden, bij de gegeven waarden van m, n en $t_1, t_2, \ldots, t_k$ alle mogelijke combinaties van $a_1, a_2, \ldots, a_k$ opgeschreven en voor iedere combinatie de waarde van de toetsingsgrootheid $\underline{W}$ berekend, zoals aangegeven is in rapport S 176(M 65).
Met behulp van formule (1) wordt dan, voor ieder der mogelijke combinaties, de kans berekend, dat die combinatie, bij de gegeven waarden van m, n en $t_1, t_2, \ldots, t_k$ optreedt als de getoetste hypothese H_o juist is.

Voorbeeld 1

Bij het in rapport S 176(M 65) beschreven voorbeeld 6 (par. 9) vindt men het volgende waarnemingsschema:

Tabel 2

Verbeteringsgraad van patiënten, behandeld met de geneesmiddelen A en B

Verbeteringsgraad	Aantal patiënten met de betreffende verbeteringsgraad bij geneesmiddel		totaal
	A	B	
1	1	0	1
2	3	5	8
3	1	2	3
4	0	3	3
totaal	5	10	15

---

1) $\binom{t}{a} = \frac{t!}{a!\,(t-a)!}$; tabellen voor deze binomiaalcoëfficiënten vindt men o.a. in [2] .

Hieronder volgen alle mogelijke combinaties van $a_1, a_2, \ldots, a_k$ (bij de gegeven waarden van n, m en $t_1, t_2, \ldots, t_k$) met de bijbehorende waarde van W.

1	0	1
4	4	8
0	3	3
0	3	3
5	10	15

W = 16

0	1	1
5	3	8
0	3	3
0	3	3
5	10	15

W = 25

1	0	1
3	5	8
1	2	3
0	3	3
5	10	15

W = 27

1	0	1
3	5	8
0	3	3
1	2	3
5	10	15

W = 33

0	1	1
4	4	8
1	2	3
0	3	3
5	10	15

W = 36

1	0	1
2	6	8
2	1	3
0	3	3
5	10	15

W = 38

0	1	1
4	4	8
0	3	3
1	2	3
5	10	15

W = 42

1	0	1
2	6	8
1	2	3
1	2	3
5	10	15

W = 44

0	1	1
3	5	8
2	1	3
0	3	3
5	10	15

W = 47

1	0	1
1	7	8
3	0	3
0	3	3
5	10	15

W = 49

1	0	1
2	6	8
0	3	3
2	1	3
5	10	15

W = 50

0	1	1
3	5	8
1	2	3
1	2	3
5	10	15

W = 53

1	0	1
1	7	8
2	1	3
1	2	3
5	10	15

W = 55

0	1	1
2	6	8
3	0	3
0	3	3
5	10	15

W = 58

0	1	1
3	5	8
0	3	3
2	1	3
5	10	15

W = 59

1	0	1
1	7	8
1	2	3
2	1	3
5	10	15

W = 61

0	1	1
2	6	8
2	1	3
1	2	3
5	10	15

W = 64

1	0	1
0	8	8
3	0	3
1	2	3
5	10	15

W = 66

1	0	1
1	7	8
0	3	3
3	0	3
5	10	15

W = 67

0	1	1
2	6	8
1	2	3
2	1	3
5	10	15

W = 70

1	0	1
0	8	8
2	1	3
2	1	3
5	10	15

W = 72

0	1	1
1	7	8
3	0	3
1	2	3
5	10	15

W = 75

0	1	1
2	6	8
0	3	3
3	0	3
5	10	15

W = 76

1	0	1
0	8	8
1	2	3
3	0	3
5	10	15

W = 78

0	1	1
1	7	8
2	1	3
2	1	3
5	10	15

W = 81

0	1	1
1	7	8
1	2	3
3	0	3
5	10	15

W = 87

0	1	1
0	8	8
3	0	3
2	1	3
5	10	15

W = 92

0	0	1
0	8	8
2	1	3
3	0	3
5	10	15

W = 98

Met behulp van formule (1) vindt men nu:

Tabel 3

Verdeling van $\underline{W}$ onder de hypothese H_o voor $m = 5$, $n = 10$, $t_1 = 1$, $t_2 = 8$, $t_3 = t_4 = 3$.

W	$P[W \mid t_1, \ldots, t_4; H_o]$	W	$P[W \mid t_1, \ldots, t_4; H_o]$
16	0,0233	59	0,0559
25	0,0187	61	0,0240
27	0,0559	64	0,0839
33	0,0559	66	0,0010
36	0,0699	67	0,0027
38	0,0280	70	0,0839
42	0,0699	72	0,0030
44	0,0839	75	0,0080
47	0,0559	76	0,0093
49	0,0027	78	0,0010
50	0,0280	81	0,0240
53	0,1679	87	0,0080
55	0,0240	92	0,0010
58	0,0093	98	0,0010

Voorbeeld 2

Stel dat de stochastische grootheden $\underline{x}$ en $\underline{y}$ slechts de twee waarden x_1 en x_2 kunnen aannemen en dat men, bij 8 waarnemingen van $\underline{x}$ en 6 van $\underline{y}$ het volgende waarnemingsschema vindt:

Tabel 4

Waarnemingsschema

Voorkomende steekproef-waarden	Aantal malen, dat z_i optreedt bij		totaal
	$\underline{x}$	$\underline{y}$	
z_1	1	4	5
z_2	7	2	9
totaal	8	6	14

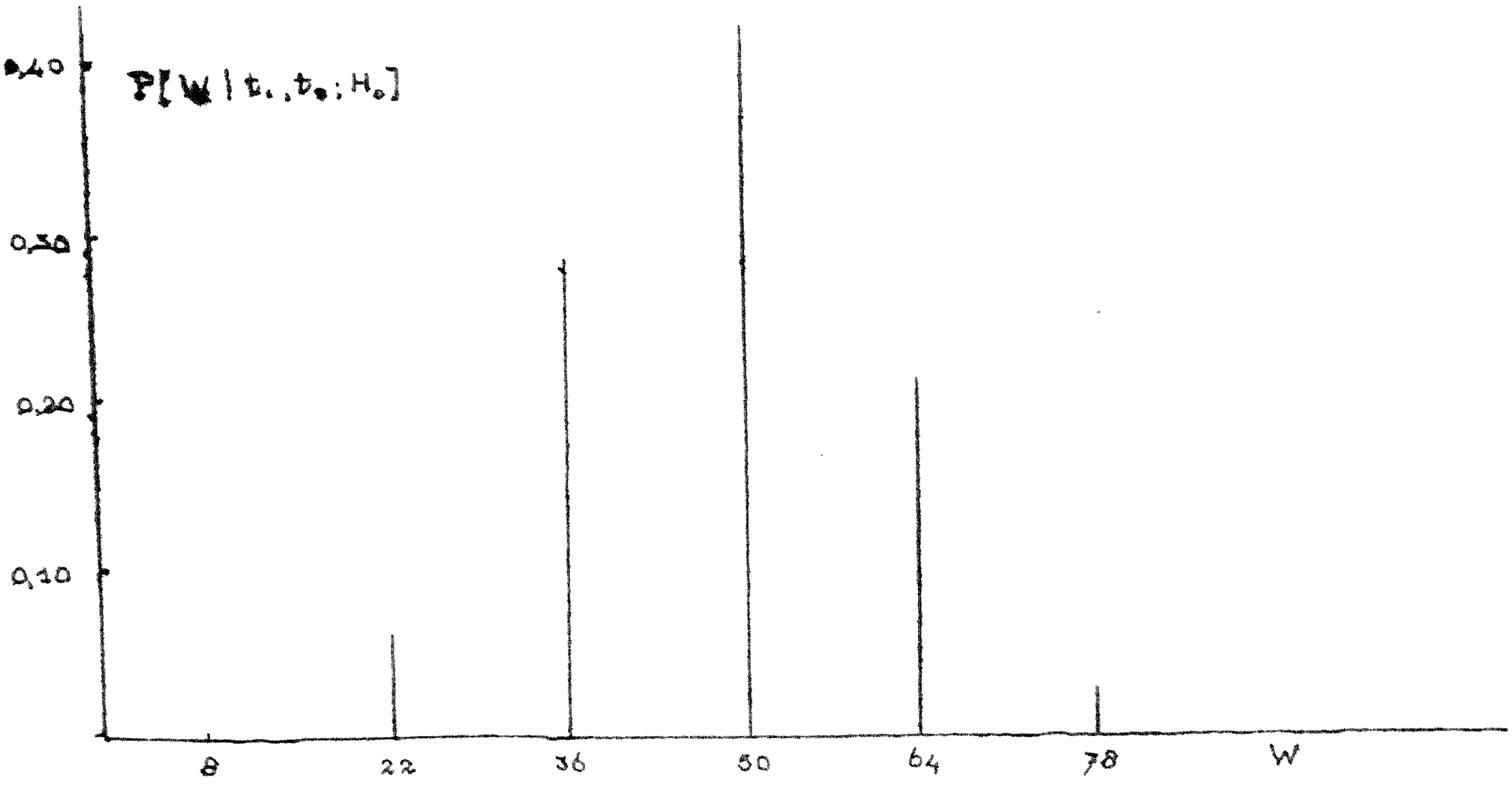
Op de bovenaangegeven wijze vindt men hier:

Tabel 5

Exacte verdeling van W onder de hypothese H_o voor m = 8, n = 6, t_1 = 5 en t_2 = 9

W	$P[W \mid t_1, t_2; H_o]$
8	0,0030
22	0,0599
36	0,2797
50	0,4196
64	0,2098
78	0,0280

Deze verdeling is in figuur 1 getekend.



Figuur 1: Exacte verdeling van W onder de hypothese H_o voor m = 8, n = 6, t_1 = 5 en t_2 = 9

3. Definitie van kritieke zones en overschrijdingskansen

De éénzijdige kritieke zones en overschrijdingskansen worden op de gebruikelijke wijze gedefinieerd. De rechts- (resp links-) éénzijdige kritieke zone bestaat dus uit grote (resp. kleine) waarden van W en de rechts- (resp. links-) éénzijdige overschrijdingskans van een waarde W van W is de kans dat W niet kleiner (resp niet groter) is dan W

Bij voorbeeld 1 is $W = 27$; de linkséénzijdige overschrijdingskans hiervan is (zie tabel 3):

$$P\,[\underline{W} \leqq 27] = 0{,}0233 + 0{,}0186 + 0{,}0559 = 0{,}0978.$$

De rechtséénzijdige overschrijdingskans van $W = 27$ is

$$P\,[\underline{W} \geqq 27] = 1 - P\,[\underline{W} < 27] = 1 - (0{,}0233 + 0{,}0186) = 0{,}9581.$$

Bij voorbeeld 2 is $W = 22$; voor de linkséénzijdige overschrijdingskans vindt men hier 0,0629; voor de rechtséénzijdige 0,9970.

Bij symmetrische verdelingen wordt als tweezijdige kritieke zone met onbetrouwbaarheidsdrempel α gewoonlijk een links en een rechtséénzijdige genomen, ieder met onbetrouwbaarheidsdrempel $\frac{1}{2}\alpha$. Deze tweezijdige kritieke zone wordt aangegeven door Z. De tweezijdige overschrijdingskans is dan gelijk aan tweemaal de kleinste éénzijdige. Bij asymmetrische verdelingen heeft deze methode echter het bezwaar, dat er een groot verschil kan ontstaan tussen α en de werkelijke onbetrouwbaarheid. Als de kleinste of de grootste waarde, die $\underline{W}$ aan kan nemen, een waarschijnlijkheid $> \frac{1}{2}\alpha$ heeft, wordt de zo gedefinieerde kritieke zone bovendien eenzijdig. Het verschil tussen α en de werkelijke onbetrouwbaarheid is dan $\geqq \frac{1}{2}\alpha$.

Bij voorbeeld 1 bestaat de kritieke zone Z met onbetrouwbaarheidsdrempel 0,05 uit de waarden $W = 16$; 87; 92 en 98 (zie tabel 3). De werkelijke onbetrouwbaarheid is 0,0333. Bij voorbeeld 2 vindt men voor de kritieke zone Z met $\alpha = 0{,}05$ de waarde $W = 8$; de werkelijke onbetrouwbaarheid is: 0,0030)(hier is de kritieke zone Z dus eenzijdig).

Bij asymmetrische verdelingen komen nu verschillende andere methoden in aanmerking. Bij voorbeeld 2 (waar men te doen heeft met een asymmetrische verdeling met één top) kan men een tweezijdige kritieke zone vormen door de waarden van W met de kleinste waarschijnlijkheden bijeen te zoeken totdat de gekozen onbetrouwbaarheidsdrempel het toevoegen van een nieuwe waarde verhindert. Deze kritieke zone wordt aangegeven door Z_1. De overschrijdingskans wordt hier de som van de waarschijnlijkheden die niet groter zijn dan de waarschijnlijkheid van de gevonden waarde van $\underline{W}$.

Bij voorbeeld 2 bestaat de kritieke zone Z_1 met onbetrouwbaarheidsdrempel 0,05 uit de waarden $W = 8$ en $W = 78$; de werkelijke onbetrouwbaarheid is 0,0030 + 0,0280 = 0,0310. De overschrijdingskans van de gevonden waarde $W = 22$ is:

0,0599 + 0,0030 + 0,0280 = 0,0909.
De kritieke zone Z_1 heeft echter bij asymmetrische verdelingen met meer dan één top (dus gevallen als die van voorbeeld 1) het bezwaar dat het linker- en rechterdeel der kritieke zone niet ieder een aaneengesloten geheel vormen. Bij voorbeeld 1 bestaat de kritieke zone Z_1 met onbetrouwbaarheidsdrempel 0,05 uit de waarden W= 49; 58; 66; 67; 72; 75; 76; 78; 87; 92 en 98; de werkelijke onbetrouwbaarheid is 0,0470.

Om te komen tot een algemene methode voor het stapsgewijs opbouwen van een tweezijdige kritieke zone, die de genoemde nadelen niet bezit en die voor symmetrische verdelingen met de gebruikelijke overeenstemt, stellen wij eerst een aantal principes vast, die wij bij deze opbouw in de aangegeven volgorde willen toepassen. Deze principes zijn:

a) het linker- en rechterdeel der kritieke zone vormen bij iedere stap ieder een aaneengesloten geheel,
b) alle waarden, die tot een kritieke zone behoren, behoren ook tot alle kritieke zones met grotere onbetrouwbaarheidsdrempel,
c) bij iedere stap wordt het verschil tussen de onbetrouwbaarheden van het linker- en rechterdeel van de kritieke zone in absolute waarde zo klein mogelijk gehouden,
d) de onbetrouwbaarheid van de kritieke zone wordt, voor iedere onbetrouwbaarheidsdrempel α, zo dicht mogelijk bij α gehouden,
e) indien de principes a,...,d bij een bepaalde stap niet één, doch twee mogelijke nieuwe waarden aanwijzen als volgende bouwsteen voor de kritieke zone, worden deze tegelijk aan de kritieke zone toegevoegd.

Deze principes sluiten dus de bovenbeschreven methoden (Z en Z_1) als algemene methoden uit en leiden in symmetrische gevallen tot $Z = Z_1$. Verder wordt een ondubbelzinnige opbouw verkregen, die op aanschouwelijke wijze, als volgt beschreven kan worden:

Volgens principe c kiezen wij van de grootste en de kleinste waarde die $\underline{W}$ aan kan nemen degene met de kleinste waarschijnlijkheid; zijn de waarschijnlijkheden van deze twee waarden aan elkaar gelijk dan nemen wij beide waarden (principe e). De volgende waarde van $\underline{W}$, die wij bij de kritieke zone nemen, is een waarde links of een waarde rechts zodat het linker- en rechterdeel derkritieke zone ieder een aaneengesloten geheel vormen

(principe a) en dat het verschil tussen de onbetrouwbaarheden links en rechts in absolute waarde zo klein mogelijk is (principe c). Wij vervolgen deze opbouw door iedere keer links of rechts een waarde van $\underline{W}$ bij de kritieke zone te nemen totdat de gekozen onbetrouwbaarheidsdrempel het toevoegen van een nieuwe waarde verhindert.

Als het toevoegen van een waarde links in absolute waarde hetzelfde verschil tussen links en rechts geeft als het toevoegen van een waarde rechts dan wordt van deze twee waarden degene met de kleinste waarschijnlijkheid toegevoegd (principe b en d). Tenslotte: is het verschil tussen links en rechts in een bepaald stadium gelijk aan 0 en hebben de volgende linker- en rechterwaarde van $\underline{W}$ dezelfde waarschijnlijkheid dan worden beide waarden toegevoegd (of als men daarmee α overschrijdt, geen van beide (principe e).

Wij zullen de zo gedefiniëerde kritieke zone aanduiden als Z_2.

De tweezijdige overschrijdingskans wordt gedefiniëerd als de onbetrouwbaarheid van de kleinste kritieke zone Z_2, die de gevonden waarde van $\underline{W}$ bevat.

Bij voorbeeld 1 bestaat de tweezijdige kritieke zone Z_2 met onbetrouwbaarheidsdrempel 0,05 uit de waarden W= 16; 87; 92 en 98; de werkelijke onbetrouwbaarheid wordt hier 0,0333.

In tabel 6 staan voor ieder der mogelijke waarden van $\underline{W}$ bij voorbeeld 1 de volgens Z_2 gedefiniëerde tweezijdige overschrijdingskansen vermeld (kolom (2)). In de kolommen (3) en (4) staan de tweezijdige overschrijdingskansen, die men vindt met behulp van de normale benadering (zie rapport S 176 (M65), par. 8); bij kolom (3) is geen continuïteitscorrectie toegepast; bij kolom (4) een continuïteitscorrectie ter grootte van 1.

De cijfers tussen haakjes in kolom (2) geven de volgorde aan waarin de waarden van $\underline{W}$ bij de kritieke zone genomen zijn.

Tabel 6

Exacte en benaderde tweezijdige overschrijdingskansen bij voorbeeld 1

(1)	(2)	(3)	(4)
	Tweezijdige overschrijdingskans		
w	exact	benaderd	
		zonder continuïteitscorrectie	met continuïteitscorrectie
16	0,0333 (4)	0,0226	0,0272
25	0,0760 (6)	0,0950	0,1074
27	0,1532 (11)	0,1236	0,1416
33	0,2930 (13)	0,2542	0,2846
36	0,4505 (17)	0,3472	0,3844
38	0,4785 (18)	0,4238	0,4592
42	0,6283 (21)	0,5892	0,6384
44	0,7455 (24)	0,6892	0,7414
47	0,9693 (26)	0,8414	0,8966
49	0,9720 (27)	0,9442	1
50	1 (28)	1	0,9442
53	0,9134 (25)	0,8414	0,8966
55	0,6616 (23)	0,7414	0,7872
58	0,6376 (22)	0,5392	0,6384
59	0,5584 (20)	0,5486	0,5892
61	0,5025 (19)	0,4592	0,5028
64	0,3806 (16)	0,3472	0,3844
66	0,2967 (15)	0,2846	0,3174
67	0,2957 (14)	0,2542	0,2846
70	0,2371 (12)	0,1802	0,2040
72	0,0973 (10)	0,1416	0,1586
75	0,0943 (9)	0,0950	0,1074
76	0,0863 (8)	0,0818	0,0950
78	0,0770 (7)	0,0602	0,0702
81	0,0573 (5)	0,0376	0,0444
87	0,0100 (3)	0,0132	0,0160
92	0,0020 (2)	0,0050	0,0060
98	0,0010 (1)	0,0012	0,0016

Opmerkingen

1. Als de kans op de gevonden waarde voor $\underline{W}$ (berekend volgens formule (1)) groter is dan de onbetrouwbaarheidsdrempel α , dan zijn zowel de één- als de tweezijdige overschrijdingskans $> \alpha$, zodat men in dit geval niet de gehele exacte verdeling behoeft uit te rekenen.

2. Heeft men de gehele exacte verdeling berekend en is de kleinste éénzijdige overschrijdingskans van de gevonden waarde van $\underline{W} > \alpha$, dan is de tweezijdige overschrijdingskans ook $> \alpha$, zodat men in dit geval de tweezijdige kritieke zone niet behoeft te bepalen.

3. De bovenbeschreven methode om de exacte verdeling van $\underline{W}$ te berekenen, kan men ook toepassen als er geen gelijke waarnemingen zijn, dus als $t_i = 1$ voor iedere i . Formule (1) gaat dan over in

$$P[\underline{a}_1 = a_1, \underline{a}_2 = a_2, \ldots, \underline{a}_k = a_k \mid H_0] = \binom{N}{m}^{-1}.$$

In dit geval is het echter veel handiger om de exacte verdeling van $\underline{W}$ onder de hypothese H_0 te berekenen zoals aangegeven is in [1] , d.w.z. te werken met de recursieformule

$$P[W \mid m,n; H_0] = \frac{m}{N} P[W-2n \mid m-1,n; H_0] + \frac{n}{N} P[W \mid m,n-1; H_0],$$

waarin $P[W \mid m,n; H_0]$ de kans op W voorstelt onder de hypothese H_0 bij m waarnemingen van $\underline{x}$ en n waarnemingen van y .

Een dergelijke recursieformule is niet bekend voor het in dit rapport behandelde algemene geval.

Literatuur

[1] 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.

[2] Fry, T.C., Probability and its engineering uses, D. van Noshand Company, New York, 1928.

[3] 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.