S 145 (M 7a)

Enige opmerkingen over de verdeling van de toetsingsgrootheid van WILCOXON

Stichting Mathematisch Centrum
Statistische Afdeling
1954

MATHEMATISCH CENTRUM
2e Boerhaavestraat 49
A m s t e r d a m - O.
Statistische Afdeling

S 145 (M 7a).

Enige opmerkingen over de verdeling van de toetsingsgrootheid van WILCOXON.

In memorandum S 47 (M 7) is vermeld dat de verdeling van de toetsingsgrootheid $\underline{U}$ van WILCOXON, onder de hypothese H_0, dat de beide steekproeven afkomstig zijn uit dezelfde waarschijnlijkheidsverdeling, voor grote n en m bij benadering normaal is. Deze benadering geldt over het algemeen als er weinig of geen gelijke waarnemingen in beide steekproeven voorkomen. Als n en m klein zijn, kan men, zo er weinig gelijke waarnemingen voorkomen gebruik maken van exacte tabellen van de verdeling van $\underline{U}$ (litt. [1], [2]). Als dan echter de gegeven waarnemingsreeksen grote groepen van gelijke waarnemingen bevatten, zijn de exacte tabellen en de normale benadering niet altijd voldoende nauwkeurig. We zullen laten zien hoe men voor dergelijke gevallen de verdeling van $\underline{U}$ exact kan berekenen.

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
	$\underline{x}$	$\underline{y}$	
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	n	m	N

Hierin zijn $z_1, \ldots, z_k$ de voorkomende steekproefwaarden, gerangschikt naar opklimmende grootte; n en m de steekproef-grootten; $t_1, t_2, \ldots, t_k$ de grootten der groepen gelijke waarnemingen en $\underline{a}_i$ (resp. $\underline{b}_i$) het aantal malen dat de waarde z_i in de eerste (resp. tweede) steekproef optreedt.

De kans dat de combinatie, gegeven in tabel I, optreedt onder de hypothese H_0 , en onder de voorwaarden $\underline{t}_1 = t_1$, $\underline{t}_2 = t_2, \ldots, \underline{t}_K = t_K$ wordt nu:

$$(1)\quad P\left[\underline{a}_1 = a_1, \underline{a}_2 = a_2, \ldots, \underline{a}_K = a_K \mid t_1, t_2, \ldots, t_K ; H_0\right] = \frac{\binom{t_1}{a_1}\cdot\binom{t_2}{a_2}\cdot\ldots\cdot\binom{t_K}{a_K}}{\binom{N}{n}} \qquad 1)$$

Nu worden voor de gegeven waarde van n , m en $t_1, t_2, \ldots, t_K$ alle andere mogelijke combinaties van de a 's en de b's opgeschreven en voor ieder hiervan de $\tilde{u}$ berekend volgens

$$(2)\qquad \tilde{u} = \tfrac{1}{2} \sum_{i<j} \sum (\underline{a}_i \underline{b}_j - \underline{a}_j \underline{b}_i).$$

Voorbeeld: Gegeven 4 waarnemingen van een stochastische **grootheid** x en 11 waarnemingen van een stochastische grootheid y . De aangenomen steekproefwaarden zijn gerangschikt naar opklimmende grootte z_1 , z_2 , z_3 en z_4 en in de twee steekproeven tezamen genomen treden deze waarden resp. 1, 1, 2 en 11 maal op. Dus (zie tabel I): $n = 4$, $m = 11$, $t_1 = t_2 = 1$, $t_3 = 2$, $t_4 = 11$, $N = 15$.

Wij geven hieronder alle mogelijke combinaties met de corresponderende waarde van $\tilde{u}$.

	1)	2)	3)	4)	5)	6)
	0 1	0 1	0 1	1 0	0 1	0 1
	0 1	0 1	1 0	0 1	0 1	1 0
	0 2	1 1	0 2	0 2	2 0	1 1
	4 7	3 8	3 8	3 8	2 9	2 9
$\tilde{u} =$	− 8	− 1½	0	+ 1	+ 5	+ 6½

	7)	8)	9)	10)	11)	12)
	1 0	1 0	0 1	1 0	1 0	1 0
	0 1	1 0	1 0	0 1	1 0	1 0
	1 1	0 2	2 0	2 0	1 1	2 0
	2 9	2 9	1 10	1 10	1 10	1 10
$\tilde{u} =$	+ 7½	+ 9	+ 13	+ 14	+ 15½	+ 22

---

1) $\binom{t}{a} = \dfrac{t!}{a!\,(t-a)!}$ Tabellen voor deze binomiaal coëfficiënten vindt men o.a. in [3].

We berekenen vervolgens de waarschijnlijkheden van al deze combinaties volgens formule (1) (zie tabel II, kolom 5). Deze waarschijnlijkheidsverdeling is in figuur 1 getekend.

De linkseenzijdige overschrijdingskans van een gevonden waarde $\tilde{u}$ van $\underline{\tilde{u}}$ wordt gedefiniëerd als de som van de waarschijnlijkheden van die waarden van $\underline{\tilde{u}}$ die kleiner of gelijk zijn aan de gevonden waarde. In tabel II zijn in kolom 6 deze exacte linkseenzijdige overschrijdingskansen gegeven en in kolom 4 de linkseenzijdige overschrijdingskansen die men vindt met behulp van de normale benadering (zie ook figuur 2).

De rechtseenzijdige overschrijdingskans wordt (analoog aan de linkseenzijdige) gedefiniëerd als de som van de waarschijnlijkheden van die waarden van $\underline{\tilde{u}}$ die groter of gelijk zijn aan de gevonden waarde van $\underline{\tilde{u}}$.

Tabel II.

Vergelijking van de linkseenzijdige overschrijdingskans berekend volgens de benaderde en de exacte methode.

1	2	3	4	5	6
Combinatie	$\tilde{u}$	$\frac{\tilde{u}}{\sigma}$	linkereenzijdige overschrijdingskans m.b.v. de normale benadering	exacte wh van de combinaties	exacte linker overschrijdingskans
1	− 8	−1,34	0,090	0,242	0,242
2	− 1½	−0,25	0,401	0,242	0,484
3	0	0	0,500	0,121	0,605
4	+ 1	+0,17	0,568	0,121	0,726
5	+ 5	+0,83	0,800	0,040	0,766
6	+ 6½	+1,09	0,862	0,080	0,847
7	+ 7½	+1,26	0,896	0,080	0,927
8	+ 9	+1,51	0,935	0,040	0,967
9	+13	+2,18	0,985	0,008	0,975
10	+14	+2,35	0,991	0,008	0,983
11	+15½	+2,60	0,995	0,016	0,999
12	+22	+3,69	1,000	0,001	1,000

Zoals duidelijk blijkt is voor enkele van de genoemde combinaties de overschrijdingskans berekend volgens de methode van de normale benadering belangrijk verschillend van die gevonden met de exacte methode.

Een tweezijdige kritieke zone met onbetrouwbaarheidsdrempel α kan hier op verschillende manieren gedefiniëerd worden:

B.v. door een linkseenzijdige en een rechtseenzijdige kritieke zone te kiezen, ieder met onbetrouwbaarheidsdrempel $\frac{1}{2}\alpha$. Een bezwaar van deze methode is dat hij vaak leidt tot een groot verschil tussen α en de werkelijke onbetrouwbaarheid. Dit is vooral dan in sterke mate het geval, als de kleinste of de grootste waarde, die $\tilde{u}$ aan kan nemen, een waarschijnlijkheid bezit, die $> \frac{1}{2}\alpha$ is. Dan wordt de kritieke zone bovendien éénzijdig. Als we b.v. in ons voorbeeld nemen $\alpha = 0{,}20$ dan bestaat de op deze wijze gedefiniëerde "tweezijdige" kritieke zone uit de waarden $\tilde{u} = 9,\ 13,\ 14,\ 15\frac{1}{2}$ en 22 en de onbetrouwbaarheid is $0{,}040 + 0{,}008 + 0{,}008 + 0{,}016 + 0{,}001 = 0{,}073$.

Een andere methode is, dat men de waarden van $\tilde{u}$ met de kleinste waarschijnlijkheden bij elkaar zoekt totdat de gekozen onbetrouwbaarheidsdrempel het toevoegen van een nieuwe waarde verhindert. Deze methode heeft het voordeel boven de vorige dat in het algemeen de onbetrouwbaarheid dichterbij α ligt, maar bij verdelingen met meer dan één top het nadeel dat de kritieke zone geen aaneengesloten geheel is. Nemen we in ons voorbeeld weer $\alpha = 0{,}20$ dan bestaat deze tweezijdige kritieke zone uit de waarden $\tilde{u} = 5,\ 9,\ 13,\ 14,\ 15\frac{1}{2}$ en 22 en de onbetrouwbaarheid is $0{,}040 + 0{,}040 + 0{,}008 + 0{,}008 + 0{,}016 + 0{,}001 = 0{,}113$. De waarde $\tilde{u} = 5$ behoort dus wel tot de kritieke zone maar $\tilde{u} = 6\frac{1}{2}$ en $\tilde{u} = 7\frac{1}{2}$ niet en men kan deze kritieke zone bezwaarlijk tweezijdig noemen.

We definiëren nu de tweezijdige kritieke zone als volgt: We beginnen met van de kleinste en de grootste waarde van $\tilde{u}$ diegene te nemen met de kleinste waarschijnlijkheid. In ons voorbeeld dus $\tilde{u} = 22$. Dan kiezen we één van de waarden $\tilde{u} = -8$ en $\tilde{u} = 15\frac{1}{2}$ en wel zodanig dat het verschil tussen de som van de waarschijnlijkheden links en de som van de waarschijnlijkheden rechts in absolute waarde minimaal is. Nemen we $\tilde{u} = -8$ dan is dit absolute verschil $0{,}242 - 0{,}001 = 0{,}241$ en nemen we $\tilde{u} = 15\frac{1}{2}$ dan is dit absolute verschil $0{,}016 + 0{,}001 = 0{,}017$. We nemen dus $\tilde{u} = 15\frac{1}{2}$.

Vervolgens kiezen we één van de waarden $\tilde{u} = 14$ en $\tilde{u} = -8$. Nemen we $\tilde{u} = -8$ dan wordt het genoemde verschil tussen links en rechts in absolute waarde: $0{,}242 - (0{,}016 + 0{,}001) = 0{,}225$ en bij $\tilde{u} = 14$; $0{,}008 + 0{,}016 + 0{,}001 = 0{,}025$. We kiezen weer die waarde van $\tilde{u}$ die het kleinste absolute verschil tussen

links en rechts geeft, dus $\tilde{u} = 14$. Zo gaan we door met telkens links of rechts een waarde van $\tilde{u}$ bij de kritieke zone te nemen totdat de gekozen onbetrouwbaarheidsdrempel het toevoegen van een nieuwe waarde verhindert. Nemen we weer $\alpha = 0,2$ dan vinden we in ons voorbeeld voor deze kritieke zone de waarden $\tilde{u} = 7\frac{1}{2}$, 9, 13, 14, $15\frac{1}{2}$ en 22 en de onbetrouwbaarheid is 0,080 + 0,040 + 0,008 + 0,008 + 0,016 + 0,001 = 0,153. Op deze wijze bereikt men dus dat de onbetrouwbaarheid in het algemeen dichterbij α ligt dan volgens de eerste methode; bovendien verkrijgt men een aaneengesloten geheel voor het linker- en rechter deel der kritieke zone, terwijl de onbetrouwbaarheid zo goed als mogelijk is in gelijke mate over de twee delen verdeeld wordt.

De tweezijdige overschrijdingskans wordt nu gedefiniëerd als de onbetrouwbaarheid van de kleinste kritieke zone van deze aard, die de gevonden waarde van $\underline{\tilde{u}}$ bevat. B.v. de kleinste kritieke zone die het punt $\tilde{u} = 6\frac{1}{2}$ bevat bestaat uit de waarden $\tilde{u} = 22$, $15\frac{1}{2}$, 14, 13, 9, $7\frac{1}{2}$, -8 en $6\frac{1}{2}$ en de tweezijdige overschrijdingskans van $\tilde{u} = 6\frac{1}{2}$ is dus in ons voorbeeld de som van de waarschijnlijkheden van deze waarden van $\underline{\tilde{u}}$ en dus 0,476.

In tabel II zijn voor ieder der mogelijke waarden van $\underline{\tilde{u}}$ deze exacte en tevens de benaderde tweezijdige overschrijdingskansen gegeven.

Tabel III.

Exacte en benaderde tweezijdige overschrijdingskansen.

$\tilde{u}$	tweezijdige overschrijdingskans	
	exact	benaderd
-8 (7)*	0,395	0,180
$-1\frac{1}{2}$ (11)	0,879	0,802
0 (12)	1,000	1,000
1 (10)	0,637	0,864
5 (9)	0,516	0,400
$6\frac{1}{2}$ (8)	0,476	0,276
$7\frac{1}{2}$ (6)	0,153	0,208
9 (5)	0,073	0,130
13 (4)	0,033	0,030
14 (3)	0,025	0,018
$15\frac{1}{2}$ (2)	0,017	0,010
22 (1)	0,001	<0,001

* De cijfers tussen haakjes geven de volgorde aan, waarin de waarden van $\tilde{u}$ bij de kritieke zones getrokken worden.

Om te illustreren welke invloed de waarden van n , m , t_1 , t_2 ,..., t_k op het resultaat hebben geven we hier nog een tweetal andere voorbeelden met de tweezijdige overschrijdingskansen.

A.

54	32	86
3	5	8
0	2	2
57	39	96

$\tilde{u}$ = + 139
exact: k = 0,06.
benaderd k = 0,04

B.

13	19	32
1	4	5
0	2	2
14	25	39

$\tilde{u}$ = + 30½
exact k = 0,23
benaderd k = 0,18

In het algemeen geldt dat de exacte verdeling van $\tilde{u}$ symmetrisch is als $n = m$ en ook als $t_1 = t_k$, $t_2 = t_{k-1}$, $t_3 = t_{k-2}$, enz. (zie Tabel I). In deze gevallen zal voor niet te kleine n en m , de normale benadering meestal vrij goed zijn.

<u>Literatuur</u>:

[1] H.R.VAN DER VAART: Gebruiksaanwijzing voor de toets van Wilcoxon, Rapport S 32 (M 4) van het Mathematisch Centrum, Statistische Afdeling.

[2] C.WHITE: The use of ranks in a test of significance for comparing two treatments Biometrics, Vol.8, number 1., March 1952.

[3] T.C.FRY: Probability and its engineering uses D.van Nostrand Company, New York 1928.

[4] Rapport S 47 (M7) van de Statistische Afdeling van het Mathematisch Centrum.

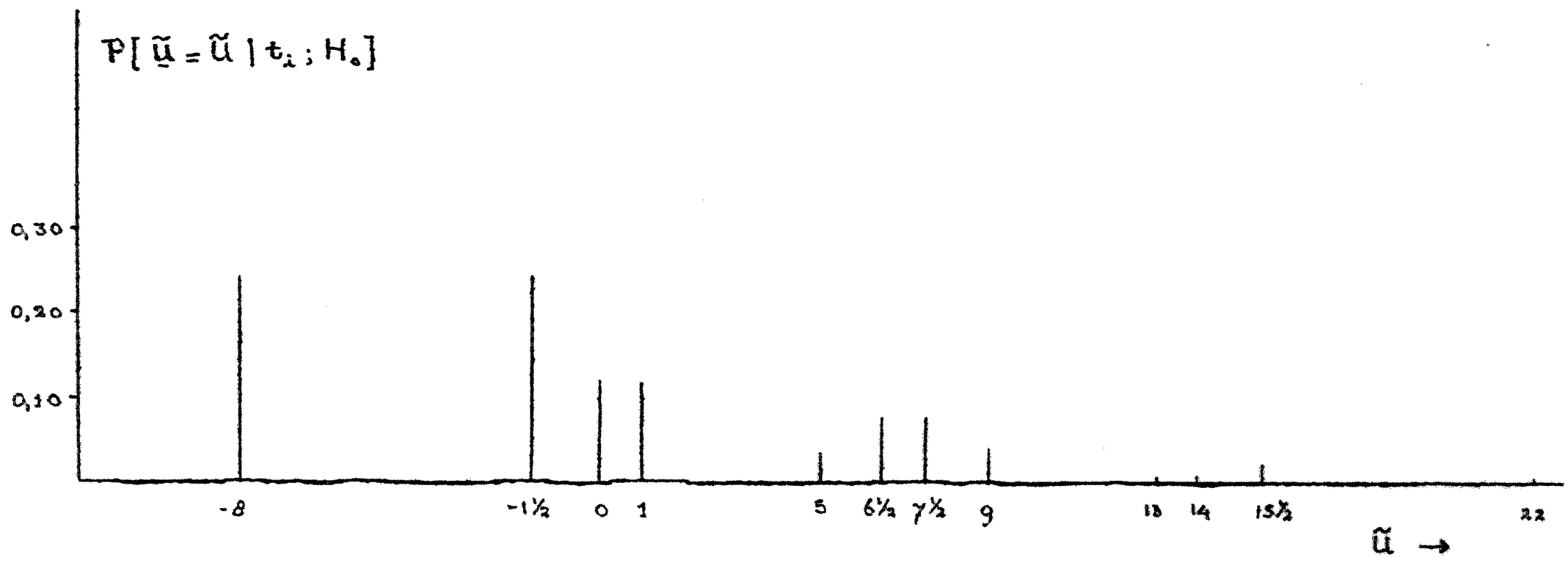

Fig. 1

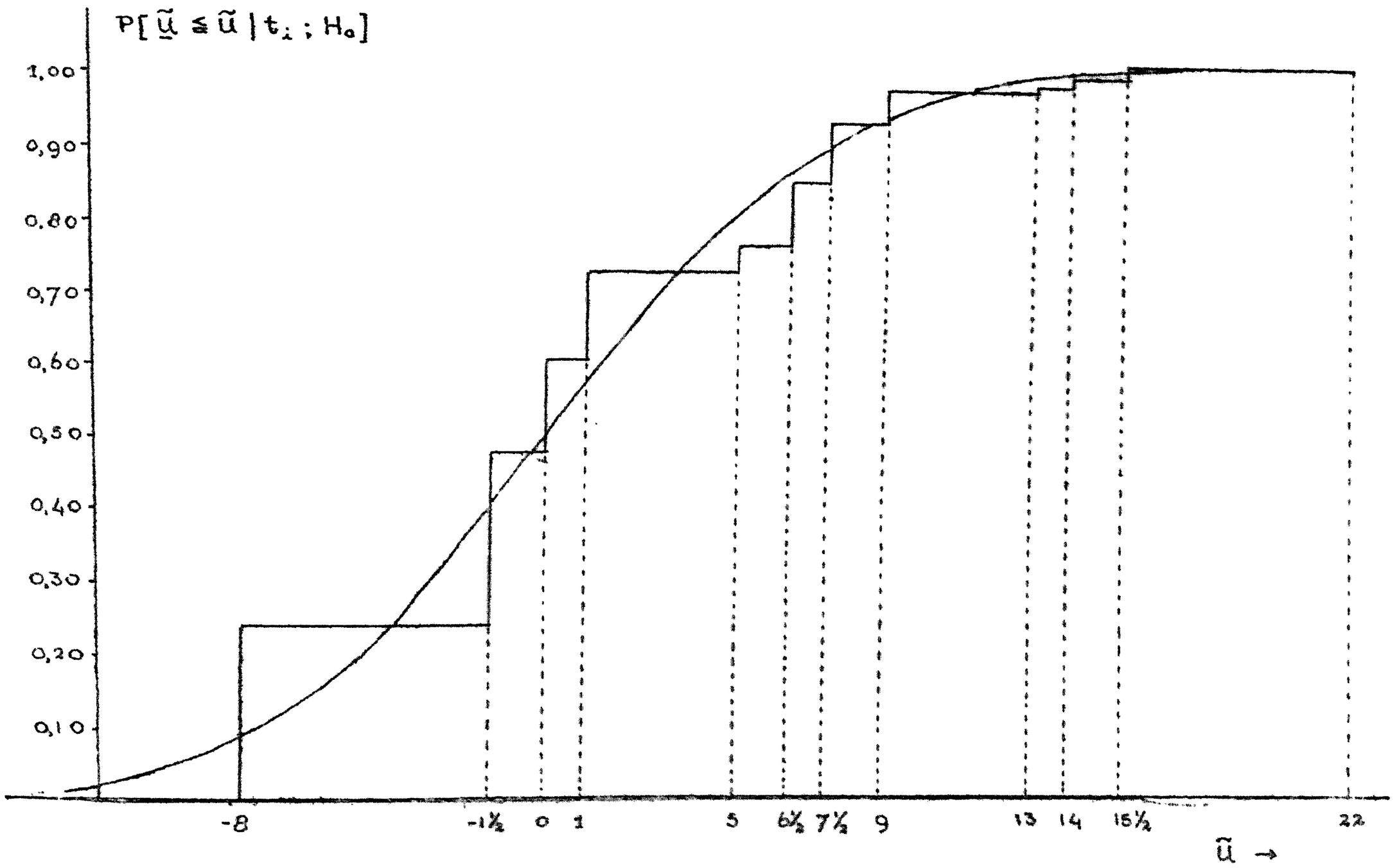

Fig. 2

$P[\tilde{u} = \tilde{u} \mid t_i; H_o]$ en $P[\tilde{u} \leq \tilde{u} \mid t_i; H_o]$, $n = 4$, $m = 11$, $t_1 = t_2 = 1$, $t_3 = 2$, $t_4 = 11$.