

STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 47 (M 15)

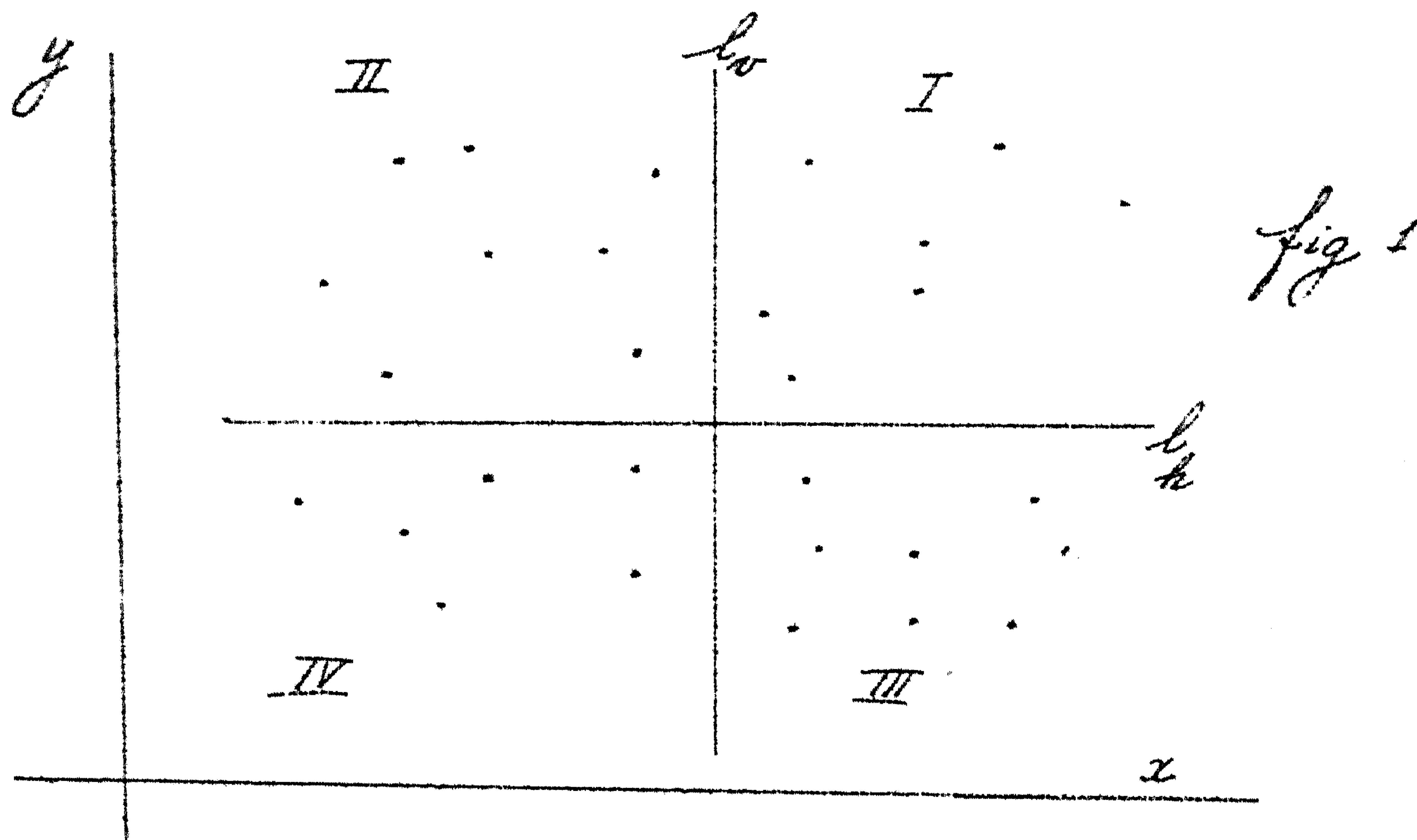
Toets voor onafhankelijkheid van 2 grootheden
met behulp van de methode der
Dubbele Dichotomie



S 47 (M 15).

Toets voor onafhankelijkheid van 2 grootheden
met behulp van de methode der
Dubbele Dichotomie ')

Deze methode dient o.a. om de onafhankelijkheid te toetsen van twee continue grootheden. $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$ zijn waarnemingsparen van de stochastische grootheden x en y . Deze getallenparen worden als punten in een vlak getekend.



We verdelen de puntenwolk door een verticale en een horizontale rechte (l_v resp. l_h) zo, dat links en rechts van l_v , en ook boven en onder l_h evenveel punten liggen. Het vlak is nu verdeeld in vier gebieden. Is er geen afhankelijkheid dan verwachten we dat in alle vier gebieden ongeveer evenveel punten zullen liggen. Een exacte afleiding, (geschikt voor kleine aantallen waarnemingen) verkrijgen we als volgt:

We beschouwen de punten in ons vlak als waarnemingen van elementen die ieder tegelijk twee kenmerken bezitten; A of $\bar{A}$ (rechts resp. links van verticale streep) en B of $\bar{B}$ (boven resp. onder horizontale streep). Er ontstaan dus vier groepen elementen met de kenmerken

AB	$A\bar{B}$	$\bar{A}B$	$\bar{A}\bar{B}$
vak I	vak III	vak II	vak IV

We tekenen dit als volgt

in een 2 x 2 tabel: (Zie blz 2)

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

' ') $\bar{A}$ betekent: "non-A"

*) Zouden hierbij punten op één der lijnen te liggen, dan laten wij deze verder buiten beschouwing.

	A	$\bar{A}$	
B	a	$n-a$	n
$\bar{B}$	b	$m-b$	m
	n	s	N

Het totale aantal elementen is N . Het aantal elementen dat kenmerk B bezit, is n , het aantal dat kenmerk $\bar{B}$ bezit, is m . Hetzelfde geldt voor de kenmerken A en $\bar{A}$ met de aantallen n resp. s . In het bovenstaande voorbeeld werd de puntenwolk zo verdeeld, dat $n = m = n = s = \frac{1}{2}N$

Dit principe wordt ^{vaak} toegepast, maar is niet noodzakelijk. De toets kan ook worden gebruikt met andere waarden van m, n, n, s , mits deze steeds op grond van de proefopstelling, niet naar aanleiding van de uitkomst van het experiment worden gekozen.

Onafhankelijkheid van de kenmerken wil zeggen, dat de kans dat een element het kenmerk B bezit even groot is, indien het tevens het kenmerk A bezit, als het tevens het kenmerk $\bar{A}$ bezit. Wij toetsen nu deze hypothese van onafhankelijkheid.

Daar de rand-totalen gegeven zijn, is er nog één vrijheidsgraad, d.w.z. indien één van de waarden in één der vier binnenvakjes bepaald is, volgen daaruit de andere waarden. Wij beschouwen het getal a in het linker-boven vakje. Dit kan verschillende waarden aannemen en is dus een stochastische grootheid a , waarvan de waarschijnlijkheidsverdeling, onder aanname van de te toetsen hypothese, een z.g. hypergeometrische verdeling is, die gegeven wordt door

$$P[a=a] = \frac{\binom{n}{a} \binom{m}{n-a}}{\binom{m+n}{n}} = \frac{n! m! n! s!}{a! b! (n-a)! (m-b)! N!}$$

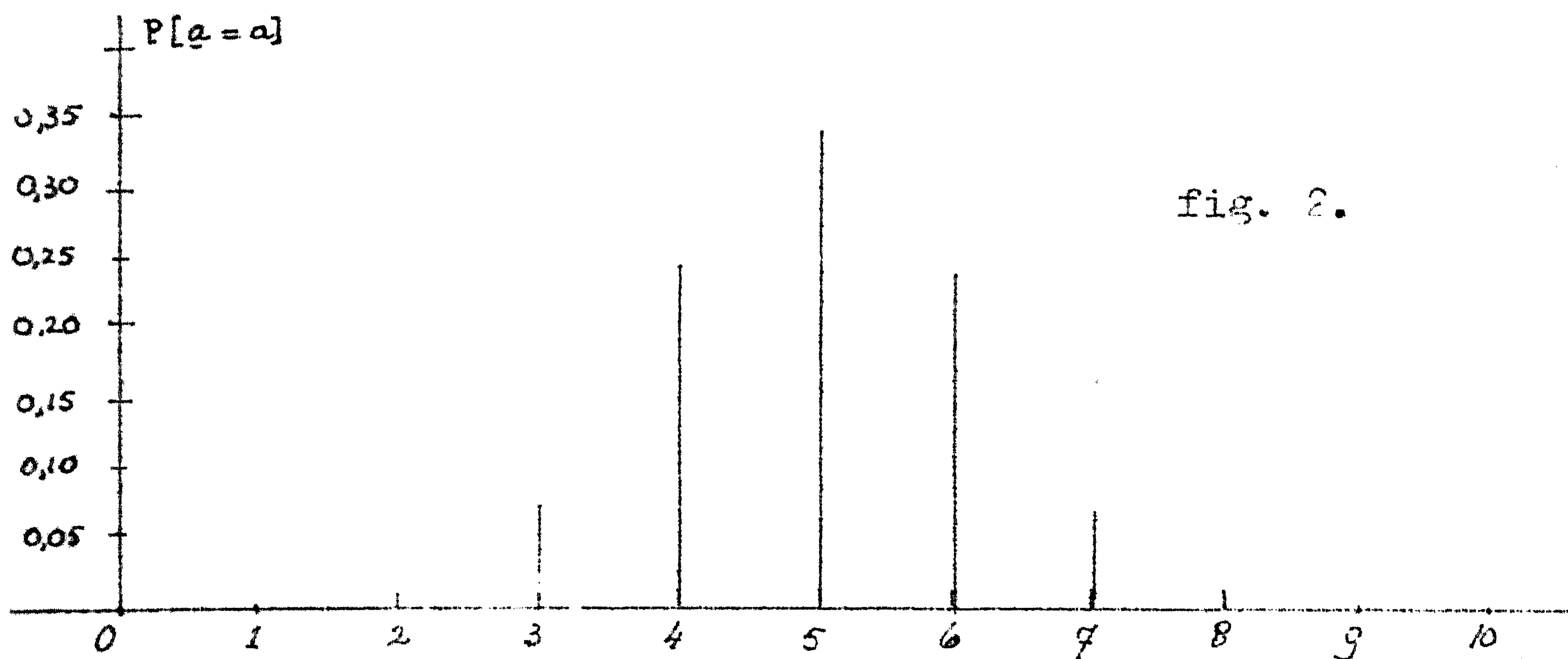
Het gemiddelde van deze verdeling is $E a = \frac{mn}{N}$ en het spreidings kwadraat is $\sigma_a^2 = \frac{n m n s}{N^2(N-1)}$

Als voorbeeld nemen wij $N=20$, $n=s=n=m=10$

De waarschijnlijkheidsverdeling van a is nu in fig. 2 en tabel I weergegeven.

In het bijzonder nemen wij de horizontale en de verticale dwarslijnen in aanmerking, dat er geen punten op liggen en dat het aantal punten ter voorzijde van ieder van deze lijnen op weinig na gelijk van elkaar verschillen. Dit is vooral dan aan te merken, indien het aantal punten klein is, of er door de regelen van 4 punten op 4 dwarslijnen te veel punten buiten berekening worden laten.

getallen v.b.



mogelijkheden:

0 10	1 9	2 8	3 7	4 6	5 5	6 4	7 3	8 2	9 1	10 0
10 0	9 1	8 2	7 3	6 4	5 5	4 6	3 7	2 8	1 9	0 10

Tabel I

a	$P[a=a]$
0	0,00001
1	0,00054
2	0,01096
3	0,07794
4	0,24356
5	0,34372
6	0,24356
7	0,07794
8	0,01096
9	0,00054
10	0,00001

Stel nu dat de volgende waarden gevonden zijn:

	A	$\bar{A}$	
B	2	8	10
$\bar{B}$	8	2	10
	10	10	20

We berekenen

$$P[\underline{a} = \underline{2}] = \frac{(10!)^4}{20!(8!)^2(2!)^2} = 0,01096$$

en zoeken verder alle waarden van a bijeen, waarvoor $P[\underline{a} = a] \leq P[\underline{a} = 2]$ is

Vervolgens tellen we de daarbij behorende waarschijnlijkheden

$P[\underline{a} = a]$ op. (In ons voorbeeld dus voor $\underline{a} = 0, 1, 2, 8, 9, 10$).

Deze som is per definitie de overschrijdingskans, behorende bij het gevonden resultaat, (in ons voorbeeld 0,02302) ¹⁾

¹⁾ De bij deze definitie van de overschrijdingskans behorende kritieke zone bestaat uit alle waarden voor $\underline{a}$ met overschrijdingskans $\leq \alpha$, in ons voorbeeld 0, 1, 2, 8, 9 en 10.

Is deze α (~~α heet~~ de onbetrouwbaarheidsdrempel), dan wordt de hypothese van onafhankelijkheid verworpen (In ons voorbeeld treedt dus, als men $\alpha = 0,05$ neemt, verwerping op).

Voor grote aantallen waarnemingen maken we gebruik van het feit dat $\frac{a - \bar{a}}{\sigma_a}$ bij benadering normaal verdeeld is, met gemiddelde 0 en spreiding 1.

Continuïteitscorrectie

Deze behoeft alleen bij kleine aantallen toegepast te worden en bestaat daarin, dat men alle getallen $a, b, n-a, m-b$ met $\frac{1}{2}$ vermeerderd of vermindert, zodanig dat de randtotalen dezelfde blijven $|a - \bar{a}|$ kleiner wordt.

Literatuur:

- M.G.Kendall, The advanced theory of Statistics, Vol.I,
 . . . (London 1947), p.303;
 E.S.Pearson, The choice of statistical tests illustrated on
 the interpretation of data classed in a 2 x 2
 table, Biometrika 34 (1949) p.139-167.