

STICHTING  
MATHEMATISCH CENTRUM  
2e BOERHAAVESTRAAT 49  
AMSTERDAM  
STATISTISCHE AFDELING

Rapport S 270

De nauwkeurigheid van enkele methoden  
voor het tellen van leucocyten

door

A.R. Elcmena

en

Rina Korswagen

1960

## 1. Inleiding

Indien men het aantal cellen van een gegeven soort wil schatten, dat zich b.v. in een  $\text{mm}^3$  vloeistof bevindt, brengt men een kleine hoeveelheid van de al of niet verdunde vloeistof in een telkamer en telt onder het microscoop hoeveel cellen zich in dit volume bevinden. In de praktijk is een telkamer door een serie roosters in kleine en groot vierkantjes verdeeld; men telt dan bij leucocyten b.v. de aantallen, die voorkomen in 144 van deze vierkantjes.

In dit rapport zal de nauwkeurigheid worden besproken van de tellingsmethoden, welke voor leucocyten als routine-methoden worden gebruikt. Verder zal de nauwkeurigheid van een methode, waarbij de waarnemingsresultaten grafisch worden verwerkt, worden besproken.

Indien hier over nauwkeurigheid wordt gesproken, wil hiermede alleen aangeduid worden, welke conclusies zijn te trekken uit gegeven telresultaten, indien verdelingsfouten, pipetfouten, etc., buiten beschouwing worden gelaten.

## 2. De Poisson-verdeling

In de regel maakt men bij tellingen de onderstelling, dat het aantal cellen in één vierkantje van de telkamer een POISSON-verdeling bezit, dus dat de kans dat er  $r$  cellen in een vierkantje zullen worden geteld, gelijk is aan

$$(1) \quad P[\underline{r} = r] = e^{-\lambda} \frac{\lambda^r}{r!}, \quad r=0,1,\dots,$$

waarbij  $\lambda$  een parameter is. De verwachting van  $\underline{r}$  is  $\lambda$ . Uit de tellingsresultaten zal men een schatting van  $\lambda$  willen berekenen.

Over de onderstelling (1) bestaat een grote hoeveelheid literatuur, o.a. E. ABBÉ (1878), STUDENT (1907), W.R. BRYAN, L.L. CHASTAIN and W.E. GARREY (1935), J. BERKSON (1944), H.O. LANCASTER (1950). In het algemeen wordt in de literatuur opgegeven, dat (1) een goed model is van de werkelijkheid, en dat afwijkingen eerst dan een rol gaan spelen indien de cellen klontering vertonen (thrombocyten), dan wel dat het aantal cellen per telkamer te groot is, zodat de cellen elkaar verdringen. Voor leucocyten vindt men in de regel opgegeven, dat goed voldaan is aan (1). Aan ons werden nog 9 tellingsresultaten voorgelegd. Ieder van deze resultaten bestond uit de 144 aantallen, zijnde de aantallen getelde cellen per vierkantje.

Aan de hand van deze tellingsresultaten hebben wij getoetst in hoeverre er aanwijzingen zijn, dat aan (1) niet is voldaan. De toets die hierbij is toegepast is beschreven in het als bijlage toegevoegde

memorandum S 73 (M29), paragraaf 2. De gevonden waarden van de  $\chi^2$  toetsingsgrootheden zijn met de bijbehorende overschrijdingskansen vermeld in tabel I. Eveneens is vermeld het resultaat van het combineren van de negen toetsen door optellen van de  $\chi^2$  toetsingsgrootheden.

Tabel I: Resultaten van de toetsen voor de hypothese (1)

telling	$\chi^2$	aantal vrijheidsgraden	overschrijdingskans k
I	18,85	8	0,015
II	2,26	7	0,95
III	5,58	7	0,59
IV	9,95	8	0,27
V	6,42	6	0,38
VI	4,12	9	0,90
VII	12,65	8	0,13
VIII	10,65	7	0,16
IX	4,07	8	0,86
gecombin.	74,55	68	0,25 < k < 0,30

Uit de tabel blijkt, dat er geen duidelijke aanwijzingen zijn dat de hypothese (1) zou moeten worden verworpen.

In de volgende paragrafen zal worden ondersteld, dat aan onderstelling (1) is voldaan.

### 3. De gebruikelijke tellingsmethoden

#### 3.1 Volledig tellen

Men kan uiteraard de aantallen cellen in de 144 vierkantjes tellen, hetgeen als waarnemingen geeft:  $x_1, \dots, x_{144}$ . Het gemiddelde

$$\frac{1}{144} \sum_{i=1}^{144} x_i$$

is een schatting  $\hat{\lambda}_A$  van  $\lambda$ .

Ieder der  $x_i$  heeft een verwachting  $\lambda$  en variantie  $\lambda$ , zodat indien de waarnemingen onderling onafhankelijk worden ondersteld

$\sum_{i=1}^{144} x_i$  een verwachting  $144 \lambda$  en variantie  $144 \lambda$  bezit, en dus

$\frac{1}{144} \sum_{i=1}^{144} x_i$  een verwachting  $\lambda$  (zodat  $\hat{\lambda}_A$  een zuivere schatting is) en een variantie  $\frac{1}{144} \lambda$ . De spreiding van  $\hat{\lambda}_A$  is dus  $\frac{1}{12} \sqrt{\lambda}$ .

### 3.2 Gedeeltelijk tellen

Waar volledig tellen vrij bewerkelijk is, volstaat men in de medische praktijk in de regel met het tellen van 2 x 25 vierkantjes. Hierbij berekent men een schatting  $\hat{\lambda}_1$  uit de telresultaten van de eerste 25 vierkantjes en een schatting  $\hat{\lambda}_2$  uit die van de tweede 25 vierkantjes. Het verschil  $\hat{\lambda}_1 - \hat{\lambda}_2$  geeft tevens een indruk van de nauwkeurigheid. Wij zullen alleen de schatting

$$\hat{\lambda}_B = \frac{1}{2}(\hat{\lambda}_1 + \hat{\lambda}_2)$$

beschouwen, welke een verwachting  $\lambda$  en spreiding  $\sqrt{\frac{\lambda}{50}}$  bezit.

### 4. Andere tellingsmethoden

Een snelle en waarschijnlijk weinig nauwkeurige schatting verkrijgt men bijvoorbeeld door alleen te tellen hoeveel vierkantjes geen cellen bevatten. Is dit aantal  $r_0$ , dan is de schatting te berekenen uit

$$e^{-\hat{\lambda}} = \frac{r_0}{144},$$

dus

$$\hat{\lambda} = -\ln \frac{r_0}{144}.$$

Is  $r_0$  zeer klein, dan zal deze schatting zeer slecht zijn, maar dan kan men ertoe overgaan ook  $r_1$ , het aantal vierkantjes met 1 cel te tellen en de vergelijking

$$e^{-\hat{\lambda}} + \hat{\lambda} e^{-\hat{\lambda}} = \frac{r_0 + r_1}{144}$$

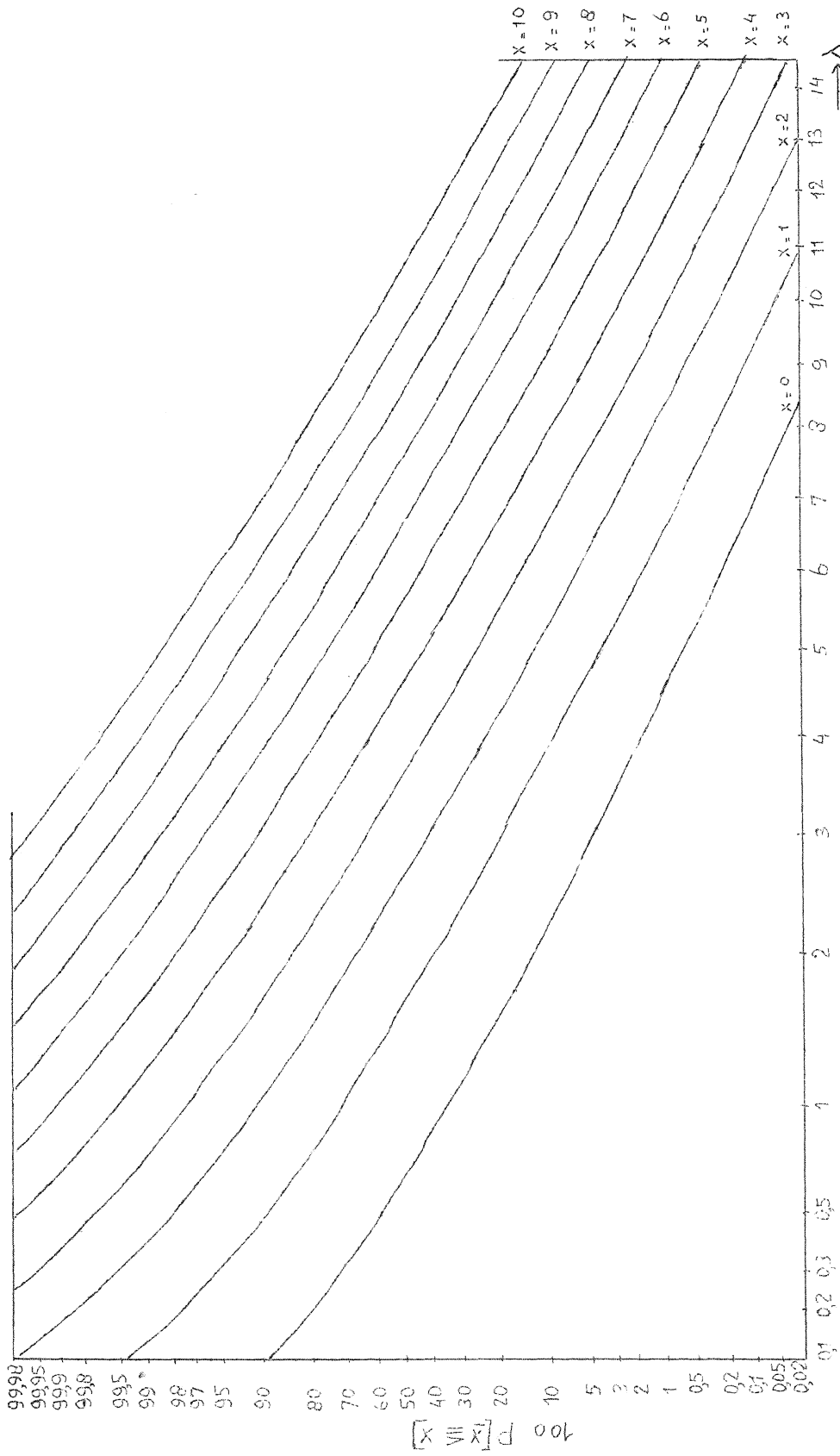
op te lossen, hetgeen grafisch op eenvoudige wijze kan geschieden.

Door G. GRIMM (1958) is een methode aangegeven om uit de tellingsresultaten op zeer vlugge wijze grafisch een schatting te bepalen. In de grafiek die men hiervoor nodig heeft (zie figuur 1) is  $\sum_{i=0}^x e^{-\lambda} \frac{\lambda^i}{i!}$  uitgezet voor  $x=0,1,\dots$  tegen  $\lambda$ .

Zijn de aantallen vierkantjes met 0,1,2,... cellen resp.  $r_0, r_1, r_2, \dots$ , dan zet men langs de verticale as uit de cumulatiefractionen

$$\frac{r_0}{144}, \frac{r_0 + r_1}{144}, \frac{r_0 + r_1 + r_2}{144}, \dots$$

en trekt voor  $j=0,1,2,\dots$  door het punt  $\frac{r_0 + \dots + r_j}{144}$  een horizontale lijn, tot waar deze de lijn voor  $x=j$  snijdt. Dit snijpunt zij  $S_j$ . Men trekt nu de verticale lijn, welke zo goed mogelijk aansluit bij  $S_0, S_1, \dots$ . Het snijpunt met de horizontale as geeft nu  $\hat{\lambda}$ .



figuur 1: Nomogram voor het schatten van  $\lambda$

Deze methode zal alleen dan een vereenvoudiging zijn, indien men niet alle aantallen  $r_0, r_1, \dots$ , zal tellen. Waar bij de gebruikelijke wijze van werken bij leucocyten  $\lambda$  ongeveer 2 is, is een mogelijke eenvoudige methodiek: alleen  $r_0, r_1$  en  $r_2$  tellen (methode  $C_2$ ). Wij hebben getracht een indruk te krijgen van de nauwkeurigheid van deze methode. Ook is hierbij nog nagegaan in hoeverre het tellen van  $r_0, \dots, r_3$  (methode  $C_3$ ) aanleiding geeft tot een essentieel nauwkeuriger methodiek. De resultaten van dit onderzoek zullen in de volgende paragraaf besproken worden.

5. Nauwkeurigheid van methoden  $C_2$  en  $C_3$

Waar het op het oog aanpassen van een rechte lijn aan een aantal punten een essentieel onderdeel is van het bepalen van  $\hat{\lambda}$  bij methoden  $C_2$  en  $C_3$ , en waar de  $\lambda$ -coördinaat van de punten  $S_0, S_1, \dots$  de oplossingen zijn van transcendente vergelijkingen, is het niet mogelijk langs analytische weg de variantie van  $\hat{\lambda}$  te berekenen. Op twee manieren hebben wij getracht een indruk te krijgen van de nauwkeurigheid van de methode.

In de eerste plaats kan men uit de voorgelegde tellingen de schattingen vergelijken, zoals deze met de verschillende methoden gevonden zouden worden. De resultaten hiervan zijn in tabel II weergegeven. De overeenstemming tussen de verschillende schattingen is redelijk te noemen.

Tabel II: Schattingen van het gemiddelde aantal leucocyten per telvierkantje

	berekend uit resultaten van 144 vierkantjes	berekend uit resultaten van 50 vierkantjes	methode $C_2$	methode $C_3$
I	2,33	2,12	2,10	2,18
II	2,01	2,08	1,99	2,09
III	1,87	1,98	1,90	1,87
IV	2,63	2,60	2,47	2,51
V	1,99	2,08	2,15	2,10
VI	3,08	3,08	3,02	3,00
VII	2,76	3,16	2,68	2,70
VIII	2,66	2,66	2,95	2,80
IX	2,64	2,68	2,71	2,69

Voor de methoden  $C_2$  en  $C_3$  werd een grotere grafiek gebruikt dan die van figuur 1 en met een fijnere schaalindeling voor  $\lambda$ , zodat de beide decimalen verantwoord zijn.

De getallen in tabel II geven evenwel geen volledig beeld van de nauwkeurigheid. Een betere indruk daarvan verkrijgt men door het uitvoeren van een steekproefexperiment.

Een waarneming van een POISSON-verdeling met parameter  $\lambda$  kan men verkrijgen door een aselekt getal,  $l$ , te kiezen met  $0 \leq l < 1$ , en die waarde van  $m$  op te zoeken, waarvoor voldaan is aan

$$\sum_{r=0}^{m-1} P[r=r] < l < \sum_{r=0}^m P[r=r].$$

De waarneming uit de POISSON-verdeling is dan  $m$ .

Op deze wijze zijn door ons 10 steekproeven van ieder 144 waarnemingen getrokken uit een POISSON-verdeling met  $\lambda = 2,0$ . Iedere reeks van 144 waarnemingen is nu beschouwd als een uitkomst van een telling. De schattingen van de gemiddelden volgens de verschillende methoden bepaald zijn vermeld in Tabel III. Daarbij is voor methoden  $C_2$  en  $C_3$  de schatting bepaald door twee personen (nr 1 en 2), onafhankelijk van elkaar.

Tabel III: Schattingen van het gemiddelde aantal witte bloedlichaampjes (Steekproefexperiment)

berekend uit alle 144 waarnemingen	berekend uit 2 x 25 waarnemingen		bepaald volgens methode GRIMM			
			$C_2$		$C_3$	
			persoon 1	persoon 2	persoon 1	persoon 2
2,18	2,16	2,16	2,12	2,16	2,19	2,16
2,08	2,16	2,32	2,00	2,21	2,07	2,16
1,90	1,80	1,96	1,99	1,96	1,89	1,88
1,99	1,80	2,16	2,03	2,07	2,00	1,99
2,17	2,28	2,24	2,09	2,17	2,08	2,13
1,91	1,80	1,96	1,93	1,92	1,94	1,93
2,17	1,72	2,16	2,05	2,11	2,08	2,14
2,12	2,24	1,44	2,12	2,16	2,08	2,11
2,09	2,28	2,08	2,05	2,05	2,04	2,08
2,08	2,24	1,68	2,01	2,03	2,00	2,01
$s_t$	-	-	0,055		0,026	
$s_s$	-	-	0,034		0,099	
$s_{\lambda}$	0,120	0,171**	0,100*		0,102*	
$\sigma_{\lambda}$	0,117	0,200	-		-	

\*\* )  $s_{\lambda}$  is de spreiding van  $\frac{\lambda_1 + \lambda_2}{2}$

\*)  $s_{\lambda} = \sqrt{s_t^2 + s_s^2}$

Met behulp van variantieanalyse kan men uit de resultaten, vermeld in de laatste vier kolommen van tabel III de invloed van de steekproeffluctuaties scheiden van die van het uitzetten in het nomogram, het trekken van de lijn en het aflezen. Daartoe stellen wij, dat de gevonden schatting  $\underline{x}_{ij}$  bij steekproef  $i$  ( $i=1, \dots, 10$ ) door persoon  $j$  ( $j=1, 2$ ) voorgesteld kan worden door

$$\underline{x}_{ij} = 2 + \underline{\varepsilon}_i + \underline{\varepsilon}_{ij} .$$

Hierbij is  $\underline{\varepsilon}_i$  de invloed op de schatting ten gevolge van het feit, dat het een steekproef betreft uit een POISSON-verdeling met  $\lambda = 2$ . Verder is  $\underline{\varepsilon}_{ij}$  de bijdrage tot  $\underline{x}_{ij}$ , die wordt veroorzaakt door de onnauwkeurigheden, geïntroduceerd door de grafische bepaling van de  $\hat{\lambda}$ . Wij onderstellen:  $E \underline{\varepsilon}_i = 0$ ;  $E \underline{\varepsilon}_i^2 = \sigma_s^2$ ;  $E \underline{\varepsilon}_{ij} = 0$ ,  $E \underline{\varepsilon}_{ij}^2 = \sigma_t^2$  voor alle  $i$  en  $j$ . Nooit onderstellen wij, dat alle  $\underline{\varepsilon}_i$  en  $\underline{\varepsilon}_{ij}$  onderling onafhankelijk zijn. Het schema van de variantie-analyse is dan:

bron van variantie	som van kwadraten	aantal vrijheidsgraden	verwachting v.d. gemiddelde som van kwadraten
tussen steekproeven	$2 \sum_i (x_i - 2)^2$	10	$2\sigma_s^2 + \sigma_t^2$
tussen personen	$\sum_{i,j} (x_{ij} - x_i)^2$	10	$\sigma_t^2$
totaal	$\sum_{i,j} (x_{ij} - 2)^2$	20	

Hierbij is  $x_i = \frac{1}{2} \sum_j x_{ij}$ . De schattingen  $s_s$  en  $s_t$  van  $\sigma_s$  en  $\sigma_t$  die op deze wijze worden berekend, zijn opgenomen in Tabel III. Men ziet dat de steekproeffluctuaties relatief groot zijn ten opzichte van de teken- en afleesnauwkeurigheid.

Gaat men er vanuit, dat bij ieder telresultaat éénmaal een  $\hat{\lambda}$  wordt bepaald, dan is een schatting van de variantie van deze  $\hat{\lambda}$

$$s_{\hat{\lambda}}^2 = s_t^2 + s_s^2 .$$

Uit Tabel III volgt dat de gevonden variantie van de  $\hat{\lambda}$ , bepaald met methoden  $C_2$  en  $C_3$  weinig verschilt van die van de  $\hat{\lambda}$  verkregen uit volledige tellingen, maar wel aanzienlijk kleiner is dan de gevonden variantie van de  $\hat{\lambda}$  verkregen bij telling van 50 vierkantjes.

Bij dit steekproefexperiment werd steeds  $\lambda = 2,0$  gebruikt. Daar men geneigd is bij het aflezen af te ronden naar een naastbijliggend geheel getal, kan dit mogelijk-erwijs de variantie van de  $\hat{\lambda}$  van methoden  $C_2$  en  $C_3$  gunstig hebben beïnvloed. Daarom hebben we een tweede steekproefexperiment verricht, ditmaal met 20 tellingen met een aantal waarden van  $\lambda$ . De uitkomsten zijn vermeld in tabel IV.



Tabel IV  
Tweede steekproefexperiment

telling	$\lambda$	methode		telling	$\lambda$	methode	
		$C_2$	$C_3$			$C_2$	$C_3$
1	2,36	2,30	2,33	11	2,03	2,15	2,04
2	2,36	2,28	2,25	12	2,03	2,26	2,26
3	2,36	2,40	2,39	13	1,91	1,87	1,84
4	2,36	2,37	2,36	14	1,91	1,87	1,83
5	2,64	2,47	2,47	15	1,91	1,82	1,85
6	2,64	2,59	2,60	16	1,91	2,08	2,11
7	2,64	2,67	2,74	17	2,15	2,22	2,18
8	2,64	2,50	2,51	18	2,15	2,14	2,26
9	2,03	2,32	2,29	19	2,15	2,21	2,24
10	2,03	2,07	2,01	20	2,15	2,07	2,08

We vinden voor methode  $C_2$

$$s_{\hat{\lambda}} = \sqrt{\frac{(\hat{\lambda}_i - \lambda_i)^2}{20}} = 0,12$$

en voor methode  $C_3$  eveneens

$$s_{\hat{\lambda}} = 0,12 ,$$

hetgeen zeer goed aansluit bij de resultaten van Tabel III.

Van enig systematisch verschil in nauwkeurigheid tussen methode  $C_2$  en methode  $C_3$  is geen sprake.

Daar hier steeds 4 waarnemingen van methoden  $C_2$  en  $C_3$  bij één waarde van  $\lambda$  zijn verricht, stelt dit ons in staat te toetsen of bijv. methode  $C_2$  even nauwkeurig is als het tellen van 50 vierkantjes. Stellen  $x_{i,1}, \dots, x_{i,4}$  de 4 waarnemingen voor bij  $\lambda = \lambda_i (i=1, \dots, 5)$ , dan heeft

$$\sum_{i=1}^5 \frac{(x_{i,1} - \lambda_i)^2}{\lambda_i} \cdot 50$$

onder deze nulhypothese een  $\chi^2$ -verdeling met 20 vrijheidsgraden. De waarde berekend uit Tabel IV is 6,4, en heeft een tweezijdige overschrijdingskans = 0,004. De nulhypothese kan hier dus verworpen worden ten gunste van de hypothese dat methode  $C_2$  nauwkeuriger is.

Vergelijken we de gevonden spreiding uit Tabel IV op dezelfde wijze met die welke men vindt bij het tellen van 144 vierkantjes, dan wordt de toetsingsgroottheid  $\chi^2 = 18,5$ , met een tweezijdige overschrijdingskans van 0,90.

Opmerkingen

1. De schattingen van het gemiddelde, die men op de gebruikelijke wijze verricht, zijn "verdelingsvrij". Dus ook als de telresultaten niet POISSON-verdeeld zijn, maar - zoals dikwijls voorkomt - een NEYMAN A-verdeling bezitten, dan blijft een gemiddelde een zinvolle grootheid. De methode van GRIMM heeft deze eigenschappen niet, daar men werkt met een nomogram, dat zeer bepaald alleen voor de POISSON-verdeling geldig is.

2. Omgekeerd geldt: wanneer de ligging van de punten in het nomogram een POISSON-verdeling onwaarschijnlijk maakt, houdt dit een waarschuwing in voor de onderzoeker, dat het monster niet goed geschud is, of dat klontervorming is opgetreden.

Literatuur:

- E. ABBÉ (1878), Ueber Blutkörper Zählung, Sitzungsbericht Jenaische Gesellschaft. Med. Naturwiss. 48.
- J. BERKSON (1944), Blood-cell count, Medical Physics, Chicago, p.110-114.
- W. RAY BRYAN, L.L. CHASTAIN and W.E. GARREY (1935), Errors of routine analysis in the counting of leucocytes, Am.Journ. of Physiology 113, p. 416-429.
- H.O. LANCASTER (1950), Statistical control in haematology. Journ. Hygiene 48, p. 402-417.
- G. GRIMM (1958), Einige Probleme der Bakterienzählung. Lezing, gehouden op de Biometrische Dag te Brussel.
- STUDENT (1907), On the error of counting with a haemocytometer, Biometrika 5, p.351-364.

S 73 (M 29)

Enkele toetsen voor de hypothese, dat  
een groep waarnemingen een steekproef  
uit een Poissonverdeling is.<sup>1)</sup>

door J. van Klinken

1. Inleiding

Een stochastische grootheid  $\underline{x}$  bezit een Poissonverdeling, indien voor  $x = 0, 1, 2, \dots$  geldt:

$$P \left[ \underline{x} = x \right] = \frac{e^{-\lambda} \lambda^x}{x!} . \quad (1)$$

Hierbij is  $\lambda$  het gemiddelde van de verdeling. Ook het kwadraat van de spreiding van  $\underline{x}$  (de variantie) is gelijk aan  $\lambda$ .

In dit memorandum worden enkele toetsen beschreven van de hypothese  $H_0$ , dat een reeks waarnemingen  $x_1, \dots, x_n$  een steekproef uit een Poissonverdeling met onbekend gemiddelde  $\lambda$  is. De verzameling van alternatieve hypothesen waarvoor de toetsen onderscheidend zijn, bevat o.a. de hypothesen  $H$ , dat de waarnemingen  $x_1, \dots, x_n$  uit verdelingen komen, waarbij het quotiënt van spreidingskwadraat en gemiddelde groter dan 1 is. Zoals uit de eerste alinea van deze paragraaf volgt, is dit quotiënt voor een Poissonverdeling juist één. Ook zijn de toetsen onderscheidend voor de alternatieve hypothesen, dat de waarnemingen uit verschillende Poissonverdelingen afkomstig zijn.

Daar de toetsen, die in de volgende paragrafen beschreven zullen worden, niet voor alle waarden van  $n$ , het aantal waarnemingen, en  $m$ , de som der waarnemingen te gebruiken zijn, zullen wij de volgende gevallen onderscheiden:

- a)  $\frac{m}{n}$  groot ( $m = x_1 + x_2 + \dots + x_n$ )  
b)  $\frac{m}{n}$  klein
- 1)  $m$  en  $n$  beide groot
  - 2)  $m$  en  $n$  of één van beiden klein.

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

In par. 2 zal het geval a besproken worden, terwijl in par. 3 en par. 4 b aan de orde komt; par. 5 geeft beknopte afleidingen van enkele der gebruikte formules; par. 6 geeft literatuurverwijzingen, waarnaar door cijfers tussen vierkante haken verwezen wordt.

## 2. Benaderingsmethoden voor $\frac{m}{n}$ groot.

Voor het geval, dat  $\frac{m}{n}$  betrekkelijk groot is kan een goede benaderingsmethode gegeven worden. Deze benadering is reeds zeer goed bruikbaar wanneer  $\frac{m}{n}$  groter dan 5 is (Vgl. [5]).

Laat  $\bar{x}$  het gemiddelde van de waarnemingen zijn en verder:

$$d = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\bar{x}} \quad (2)$$

$d$  is afhankelijk van de waarnemingen  $x_1, \dots, x_n$  en is dus, indien wij alle mogelijke resultaten beschouwen, een stochastische grootheid, hetgeen wij aangeven door onderstreping:  $\underline{d}$ ; de gevonden waarde wordt zonder onderstreping geschreven. Is  $H_0$  juist, d.w.z. komen de waarnemingen uit een Poissonverdeling, dan bezit  $\underline{d}$  bij benadering een  $\chi^2$ -verdeling met  $n-1$  vrijheidsgraden (vgl. [6]).  $H_0$  zal met onbetrouwbaarheidsdrempel  $\alpha$  verworpen worden, als voor de gevonden waarde  $d$  geldt:

$$d \geq \chi_{\alpha}^2,$$

waarin de bij gegeven  $\alpha$  en  $n-1$  behorende waarde  $\chi_{\alpha}^2$  in een tabel van de  $\chi^2$ -verdeling kan worden opgezocht (zie [11]).

## 3. De exacte methode (m of n klein).

De toetsingsmethode, die in deze paragraaf beschreven zal worden is exact, d.w.z. geen benaderingsmethode. In principe is ze uitvoerbaar voor alle waarden van  $n$  en  $m$ . In verband met het benodigde rekenwerk is ze echter alleen praktisch te gebruiken, wanneer minstens één der beide getallen  $m$  en  $n$  klein is.

De som der waarnemingen  $m = x_1 + x_2 + \dots + x_n$  is een stochastische grootheid ( $\underline{m}$ ), daar ze van de waarnemingen afhangt. Wij beschouwen nu al die waarnemingsuitkomsten waarbij  $\underline{m}$  een bepaalde vaste waarde  $m$  heeft (voor deze waarde wordt bij toepassing van de toets de bij het experiment gevonden waarde genomen). Men kan dan bewijzen dat, onder onze hypothese  $H_0$  de volgende formule geldt:

$$P \left[ \underline{x}_1 = x_1, \dots, \underline{x}_n = x_n \mid \underline{m} = m; H_0 \right]$$

$$= \frac{m!}{x_1! \dots x_n!} n^{-m} \quad (3)$$

In woorden: onder de voorwaarde, dat  $\underline{m}$  de waarde  $m$  heeft en onder hypothese  $H_0$ , bezitten de grootheden  $\underline{x}_1, \dots, \underline{x}_n$  de door (3) gegeven verdeling. Dit is een multinomiale verdeling met gelijke kansen. D.w.z. deze verdeling verkrijgt men ook als  $m$  voorwerpen over  $n$  vakken verdeeld worden, waarbij de kansen van elk voorwerp om in de verschillende vakken terecht te komen voor alle vakken even groot zijn. De grootheden  $x_1, \dots, x_n$  geven dan de aantallen voorwerpen in het eerste, tweede,  $\dots$ ,  $n^e$  vakje aan.

In (3) komt de onbekende grootheid  $\lambda$  niet meer voor.

De uitvoering van de toets die op deze waarschijnlijkheidsverdeling berust, zetten wij uiteen aan de hand van een voorbeeld. Veronderstel, dat er 6 waarnemingen zijn en dat de som der waarnemingen 8 is; dus  $n=6$  en  $m=8$ . Wij schrijven nu alle splitsingen van het getal 8 in 6 getallen  $x_1, \dots, x_6$  op, waarbij echter alleen die splitsingen genoteerd worden waarvoor geldt:

$$x_1 \geq x_2 \geq x_3 \geq \dots \geq x_n \quad (4)$$

Wij krijgen dan tabel I; de gevolgde systematiek is duidelijk.

Tabel I

splitsing no	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
1	8	0	0	0	0	0
2	7	1	0	0	0	0
3	6	2	0	0	0	0
4	6	1	1	0	0	0
5	5	3	0	0	0	0
6	5	2	1	0	0	0
7	5	1	1	1	0	0
8	4	4	0	0	0	0
9	4	3	1	0	0	0
10	4	2	2	0	0	0
11	4	2	1	1	0	0
12	4	1	1	1	1	0
13	3	3	2	0	0	0
14	3	3	1	1	0	0
15	3	2	2	1	0	0
16	3	2	1	1	1	0
17	3	1	1	1	1	1
18	2	2	2	2	0	0
19	2	2	2	1	1	0
20	2	2	1	1	1	1

Elke splitsing (d.w.z. elke regel) bevat een aantal groepen van gelijke getallen. Noem de aantallen gelijke getallen in elk van de regels  $k_1, \dots, k_l$ . In de eerste regel b.v. is  $k_1=1$  en  $k_2=5$ , in de tweede  $k_1=1$ ,  $k_2=1$ ,  $k_3=4$ . Iedere regel vertegenwoordigt, indien wij de voorwaarde (4) opheffen, een aantal splitsingen met dezelfde waarden van  $x$  als in deze regel aangegeven staan, maar nu in alle mogelijke volgorden. Het aantal verschillende volgorden (dus het aantal splitsingen), behorend bij één regel, is

$$K = \frac{n!}{k_1! \dots k_l!} .$$

Nu wordt bij elk van de splitsingen uit tabel I de daarbij behorende kans  $P$  uit (3) berekend en vermenigvuldigd met de met die regel corresponderende  $K$ . Voor iedere regel wordt op die wijze een getal

verkregen, dat, onder  $H_0$  en onder de voorwaarde  $\underline{m}=m$ , de kans voorstelt op één der door die regel vertegenwoordigde splitsingen. De zo verkregen kansen worden vervolgens geplaatst in een rij naar afdalende waarden van  $X=x_1^2+x_2^2+\dots+x_n^2$  (bij iedere regel behoort één waarde van  $S$ ). Dit is uitgevoerd in tabel II.

Tabel II

no	S	K	$10^5 \cdot P$	$10^5 \cdot K \cdot P$
1	64	6	0	0
2	50	30	0	1
3	40	30	2	51
4	38	60	3	198
5	34	30	3	99
8	32	15	4	62
6	30	120	10	1200
7	28	60	20	1200
9	26	120	17	2004
10	24	60	25	1501
13	22	60	33	1998
11	22	180	50	9004
14	20	90	67	6003
12	20	30	100	3001
15	18	180	100	18007
18	16	15	150	2251
16	16	120	200	24010
19	14	60	300	18007
17	14	6	400	2401
20	12	15	600	9005

Er dient nog opgemerkt te worden, dat bij verschillende splitsingen gelijke waarden van  $S$  kunnen behoren; in dat geval is gerangschikt naar opklimmende waarden van  $P$ .

Een kritieke zône met onbetrouwbaarheid  $\alpha$  wordt nu gevormd door de splitsingen in de volgorde van tabel II bij elkaar te nemen tot  $\alpha$  nog juist niet overschreden wordt. Als voorbeeld nemen wij

$\alpha = 0,05$ . Wij sommeren nu de kansen in de laatste kolom van tabel II tot en met de kans, waarbij de gevormde som nog juist niet groter dan 0,05 is; in ons geval behoort de laatste kans die nog gesommeerd wordt, bij de splitsing met nummer 9. De splitsingen boven de stippe lijn en de splitsingen die hieruit ontstaan door indices verwisseling, vormen nu de kritieke zône. De juiste onbetrouwbaarheid is iets kleiner dan 0,05, en wel 0,048.

De kritieke zône bestaat dus in de eerste plaats uit grote waarden van S, terwijl van twee regels met gelijke S in de eerste plaats de regel met de kleinste P bij de kritieke zône genomen wordt.

Het is verder duidelijk, dat men niet alleen een kritieke zône kan construeren, maar bij gegeven resultaat ook een overschrijdingskans kan berekenen. Deze is gelijk aan de onbetrouwbaarheid van de kleinste kritieke zône van het beschreven type, die dit resultaat nog juist bevat.

Voor de berekening van een overschrijdingskans is het in het algemeen niet nodig om een volledige tabel, zoals tabel II op te stellen, in de regel kan met een gedeelte ervan worden volstaan.

#### 4. Benaderingsmethode voor zeer grote waarden van n en m.

Het kan voorkomen, dat men bij een bepaald experiment een groot aantal waarnemingen verkrijgt, waarbij  $\frac{m}{n} < 5$  is. De in par. 2 beschreven benaderingsmethode is nu niet te gebruiken. Ook de exacte is veelal voor dit geval veel te omslachtig. Soms kan nu evenwel met voordeel gebruik gemaakt worden van een methode die berust op het aantal nullen onder de waarnemingen. Het aantal nullen is een stochastische grootte  $\underline{h}$ . Men kan aantonen, dat onder  $H_0$  en de voorwaarde  $\underline{m}=m$ , dat de som van de waarnemingen een vaste waarde m heeft, de volgende benaderingsformule geldt:

$$P \left[ \underline{h} = h \mid \underline{m} = m; H_0 \right] \approx \frac{e^{-\lambda} \lambda^h}{h!},$$

waarbij  $\lambda = n \frac{m}{n}$  is (zie [2], [8]). (5)

In woorden: Onder  $H_0$  en de conditie, dat de som der waarnemingen m is, wordt de verdeling van het aantal nullen  $\underline{h}$ , bij benadering gegeven door een Poissonverdeling waarvan de parameter

$\frac{m}{n}$  is.



Grote aantallen nullen wijzen in het algemeen op een te grote spreiding in de waarnemingen. Onze kritieke zône zal dus moeten bestaan uit grote waarden van  $h$ . Bij gegeven onbetrouwbaarheid  $\alpha$  wordt de kritieke zône gevormd door de waarden  $h \geq h_\alpha$ , waarbij  $h_\alpha$  het kleinste getal is waarvoor geldt

$$\sum_{h=h_\alpha}^{\infty} \frac{e^{-\lambda} \lambda^h}{h!} \cong \alpha \quad \left( \lambda = ne^{-\frac{m}{n}} \right).$$

$h_\alpha$  kan gemakkelijk met behulp van een tafel voor de termen van een Poissonreeks bepaald worden [7].

### 5. Enkele mathematische details

Deze paragraaf geeft enkele beknopte afleidingen van de in de vorige paragrafen voorkomende formules.

Onder  $H_0$  bezit elk der  $\underline{x}_i$  de verdeling

$$P \left[ \underline{x}_i = x_i \mid H_0 \right] = \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \quad (x_i = 0, 1, 2, \dots) \quad (6)$$

De  $\underline{x}_i$  zijn onafhankelijk verdeeld. De simultane verdeling wordt dus gegeven door

$$\begin{aligned} P \left[ \underline{x}_i = x_i, (i=1, \dots, n) \mid H_0 \right] &= \\ &= \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda} \lambda^m}{\prod_{i=1}^n x_i!} \end{aligned} \quad (7)$$

Onder  $H_0$  bezit ook  $\underline{m} = \sum_{i=1}^n \underline{x}_i$  een Poissonverdeling en wel met parameter  $n\lambda$ .

Hieruit volgt:

$$P \left[ \underline{m} = m \mid H_0 \right] = \frac{e^{-n\lambda} (n\lambda)^m}{m!} \quad (8)$$

Nu geldt:

$$\frac{P \left[ \underline{x}_i = x_i \quad (i = 1, \dots, n) \mid \underline{m} = m; H_0 \right]}{P \left[ \underline{m} = m \mid H_0 \right]} = \quad (9)$$

De teller van (9) kan vervangen worden door

$$P \left[ \underline{x}_i = x_i \quad (i = 1, \dots, n) \mid H_0 \right],$$

daar uit  $\underline{x}_i = x_i (i=1, \dots, n)$  volgt, dat  $\underline{m}=m$  is.

Derhalve is

$$\begin{aligned} P \left[ \underline{x}_i = x_i \quad (i=1, \dots, n) \mid \underline{m} = m; H_0 \right] &= \\ &= \frac{e^{-n\lambda} \lambda^m}{\prod_{i=1}^n x_i!} \frac{e^{-n\lambda} (n\lambda)^m}{m!} = \frac{m!}{\prod_{i=1}^n x_i!} n^{-m}. \end{aligned} \quad (10)$$

Dit is formule (3) uit par. 3. De exacte toets is hiermee geweest.

(10) stelt een multinomiale verdeling met gelijke kansen voor. Het is nu bekend ([1], [9]) dat in dat geval

$$\underline{d} = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\bar{x}}$$

asymptotisch een  $\chi^2$ -verdeling met  $n-1$  vrijheidsgraden bezit. Hiermee is de geldigheid van de in par. 2 uiteengezette benaderingsmethode aangetoond.

De verdeling van het aantal lege vakken,  $\underline{h}$ , wanneer  $m$  dingen over  $n$  vakken verdeeld worden, waarbij de kansen om in de verschillende vakken terecht te komen gelijk zijn, wordt gegeven door:

$$P \left[ \underline{h}=h \right] = \binom{n}{h} \sum_{\nu=0}^{n-h} (-1)^\nu \binom{n-h}{\nu} \left(1 - \frac{h+\nu}{n}\right)^m. \quad (11)$$

Nemen  $n$  en  $m$  onbegrensd toe, terwijl  $\lambda = ne^{-\frac{m}{n}}$  begrensd blijft, dan geldt:

$$P \left[ \underline{h}=h \right] - \frac{e^{-\lambda} \lambda^h}{h!} \rightarrow 0. \quad (12)$$

Hierop berust de toetsingsmethode in par. 4 beschreven. De afleiding van (11) en (12) vindt men in [2] en [8]. Voorzover bekend is deze benadering pas dan voldoende nauwkeurig, indien  $m$  en  $n$  beide zeer groot (b.v. beide groter dan 100) zijn.

Opmerking 1. De beschreven toetsen hebben het karakter van voorwaardelijke toetsen met als voorwaarde  $\underline{m} = m$ . Dit is evenwel niet essentieel; de voorwaarde kan geëlimineerd worden.

Opmerking 2. Eigenschappen zoals de bruikbaarheid voor verschillende klassen van alternatieve hypothesen worden in dit memorandum niet besproken.

6. Literatuuropgave.

- [1]. H. Cramér, Mathematical methods of statistics, Princeton University Press, 1946, p. 418.
- [2]. W. Feller, An introduction to probability theory and its applications, Vol. I, John Wiley & Sons, Inc., New York pp. 69-74.
- [3]. R.A. Fisher, The accuracy of the plating method of estimating the density of bacterial populations, Annals of Applied Biology, Vol. IX, 1922, Nos. 3 and 4, pp. 325-359.
- [4]. R.A. Fisher, The significance of deviations from expectations in a Poisson series, Biometrics, Vol. 6, 1950, pp. 17-25.
- [5]. P.G. Hoel, On indices of dispersion, The Annals of Mathematical Statistics. Vol. XIV, No 1, pp. 155-163. 1943.
- [6]. P.G. Hoel, Introduction to mathematical statistics, John Wiley & Sons, Inc. New York. pp. 195-197.
- [7]. E.C. Molina, Poisson's Exponential Binomial Limit, D.v.Nostrand Company, New York 1945.
- [8]. R. von Mises, Uber Aufteilungen und Besetzungswahrscheinlichkeiten, Revue de la Faculté des Sciences de l'Université d'Istanbul N.S. Vol. 4. 1939. pp. 1-19.
- [9]. K. Pearson, On the criterion that a given system of deviations from the Probable in the case of a Correlated System of variables is such that it can be reasonably supposed to have arisen from Random Sampling, The London Edinburgh and Dublin Philosophical Magazine and Journal of Science (1900) fifth series, vol. 50, pp. 157-175.
- [10]. S.S. Wilks, Mathematical Statistics, Princeton University Press, Princeton, 1946, p. 136.
- [11]. Tabellen van de  $\chi^2$ -verdeling kan men b.v. vinden in:
  - a) R.A. Fisher and F. Yates, Statistical tables for biological agricultural and medical research, Oliver and Boyd London 1949.
  - b) P.G. Hoel, Introduction to mathematical statistics, John Wiley & Sons, Inc., New York 1946, p. 246.
  - c) M.G. Kendall, Rank correlation methods, London, Charles Griffin & Company, Ltd., 1948, p. 153.