STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 73(M 29)

Enkele toetsen voor de hypothese, dat een groep waarnemingen een steekproef uit een Poisson-verdeling is.

J. van Klinken

1952

MATHEMATISCH CENTRUM
2de Boerhaavestr. 49,
A m s t e r d a m (O)
Statistische Afdeling
S 73 (M 29).

Februari 1952

Enkele toetsen voor de hypothese, dat een groep waarnemingen een steekproef uit een Poissonverdeling is.[1)]

door J. van Klinken

§1. Inleiding.

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

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

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

In dit memorandum worden enkele toetsen beschreven van de hypothese H_0, dat een reeks waarnemingen $x_1 \ldots x_n$ een steekproef uit een Poissonverdeling met onbekend gemiddelde λ is. De verzameling van alternatieve hypothesen waarvoor de toetsen onderscheidend zijn, bevatten o.a. de hypothesen H_1, dat de waarnemingen $x_1 \ldots x_n$ uit verdelingen komen, waarbij het quotient van spreidingskwadraat en gemiddelde groter dan 1 is. Zoals uit de eerste alinea van deze paragraaf volgt, is dit quotient 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 $\quad (m = x_1 + x_2 + \ldots + x_n)$.

b) $\frac{m}{n}$ klein
   1) m en n beide groot.
   2) m en n of één van beiden klein.

In § 2 zal het geval a besproken worden, terwijl in § 3 en § 4 b aan de orde komt; § 5 geeft beknopte afleidingen van

---

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

enkele der gebruikte formules; § 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:

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

d is afhankelijk van de waarnemingen $x_1 \ldots x_n$ 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_o juist, d.w.z. komen de waarnemingen uit een poissonverdeling dan bezit $\underline{d}$ bij benadering een χ^2 verdeling met n-1 vrijheidsgraden (vgl.[6]). H_o zal met onbetrouwbaarheidsdrempel α verworpen worden, als voor de gevonden waarde d geldt:

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

waarin de bij gegeven α en n-1 behorende waarde χ_α^2 in een tabel van de χ^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 practisch te gebruiken, wanneer minstens één der beide getallen m en n klein is.

De som der waarnemingen $m = x_1 + x_2 \ldots + x_n$ 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_o de volgende formule geldt:

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

$$= \frac{m!}{x_1! \ldots x_n!}\, n^{-m} . \qquad (3)$$

In woorden, onder de voorwaarde, dat $\underline{m}$ de waarde m heeft en onder hypothese H_o, bezitten de grootheden $\underline{x}_1, \ldots \underline{x}_n$ de door (3) gegeven verdeling. Dit is een multinoniale verdeling

met gelijke kansen. D.w.z. deze verdeling verkrijgt men ook al 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 $x_1,\cdots,x_n$ geven dan de aantallen voorwerpen in het eerste, tweede ,$\cdots$,n^e vakje aan.

In (3) komt de onbekende grootheid λ 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 $x_1,\cdots,x_6$ op, waarbij echter alleen die splitsingen genoteerd worden waarvoor geldt:

$$x_1 \geqslant x_2 \geqslant x_3 \cdots \geqslant x_n. \qquad (4)$$

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 $k_1,\cdots,k_l$. 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

$$K = \frac{n!}{k_1! \ldots k_\ell!} .$$

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_o 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 $S=x_1^2+x_2^2+\ldots+x_n^2$. (bij iedere regel behoort één waarde van S). Dit is uitgevoerd in tabel II.

Tabel II.

no	S	K	10^5.P	10^5.K.P.
1	64	6	0	0
2	5C	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 α wordt nu ge-

vormd door de splitsingen in de volgorde van tabel II bij elkaar te nemen tot α 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 stippellijn 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. <u>Benaderingsmethode voor zeer grote waarden van n en m</u>.

Het kan voorkomen, dat men bij een bepaald experiment een groot aantal waarnemingen verkrijgt, waarbij $\frac{m}{n} < 5$ is. De in § 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 grootheid $\underline{h}$. Men kan aantonen, dat onder H_o en de voorwaarde $\underline{m} = m$, dat de som van de waarnemingen een vaste waarde m heeft, de volgende benaderingsformule geldt:

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

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

In woorden: Onder H_o 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 $ne^{-\frac{m}{n}}$ 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 α wordt de kritieke zône gevormd door de waarden $h \geq h_\alpha$, waarbij h_α het kleinste getal is waarvoor geldt

$$\sum_{h=h_\alpha}^{\infty} \frac{e^{-\lambda}\,\lambda^h}{h!} \leq \alpha \qquad (\lambda = ne^{-\frac{m}{n}})$$

h_α 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 afleiding van de in de vorige paragrafen voorkomende formules.

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

$$P\,[\underline{x}_i = x_i \mid H_o] = \frac{e^{-\lambda}\,\lambda^{x_i}}{x_i!} \qquad (x_i = 0,1,2\ldots) \quad (6)$$

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

$$P\,[\underline{x}_i = x_i, (i = 1,\ldots,n) \mid H_o] =$$

$$= \prod_{i=1}^{n} \frac{e^{-\lambda}\,\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\,\lambda^m}{\prod_{i=1}^{n} x_i!} \,. \qquad (7)$$

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

Hieruit volgt:

$$P\,[\underline{m} = m \mid H_o] = \frac{e^{-n\lambda}(n\lambda)^m}{m!} \,. \qquad (8)$$

Nu geldt:

$$P\,[\underline{x}_i = x_i,\ (i = 1,\ldots,n) \mid \underline{m} = m;\ H_o] =$$

$$\frac{P\,[\underline{x}_i = x_i \quad (i = 1,\ldots,n),\quad \underline{m} = m \mid H_o]}{P\,[\underline{m} = m \mid H_o]} \,. \qquad (9)$$

De teller van (9) kan vervangen worden door

$$P\,[\underline{x}_i = x_i\,,\ (i = 1,\ldots,n) \mid H_o]\,,$$

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

Derhalve is:

$$P\,[\underline{x}_i = x_i\,,\ (i = 1,\ldots,n) \mid \underline{m} = m;\ H_o] =$$

$$= \frac{e^{-n\lambda}\,\lambda^m}{\prod_{i=1}^{n} x_i!} \Big/ \frac{e^{-n\lambda}(n\lambda)^m}{m!} = \frac{m!}{\prod_{i=1}^{n} x_i!}\, n^{-m} . \qquad (10)$$

Dit is formule (3) uit § 3. De exacte toets is hiermee bewezen.

(10) stelt een multinomiale verdeling met gelijke kansen voor. Het is nu bekend ([7] , [9]) dat in dat geval

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

asymptotisch een χ^2-verdeling met n-1 vrijheidsgraden bezit. Hiermee is de geldigheid van de in § 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:

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

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

$$P[\underline{h}=h] - \frac{e^{-\lambda}\lambda^h}{h!} \longrightarrow 0. \qquad (12)$$

Hierop berust de toetsingsmethode in § 4 beschreven. De afleiding van (11) en (12) vindt men in [2] en [8] . Voor zover bekend is deze benadering pas dan voldoende nauwkeurig, indien m en n beide zeer groot (b.v. beide groter van 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 Limit, D.v. Nostrand Company, New York 1945.

[8]. R.von Mises, Uber Aufteilungs 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 haven 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 χ^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.