S 145 (M 23a)

Toets voor de hypothese $P_1 = P_2 = \ldots = P_k$ met behulp van een 2 × k tabel

Stichting Mathematisch Centrum
Statistische Afdeling
1954

MATHEMATISCH CENTRUM,
2de Boerhaavestr. 49,
A m s t e r d a m - O.
Statistische Afdeling
S 145 (M 23a)

Toets voor de hypothese $p_1 = p_2 = \ldots = p_k$ met behulp van een 2 x k tabel [1]

Wij beschouwen k reeksen $R_1, R_2, \ldots, R_k$ van onafhankelijke waarnemingen, waarbij iedere waarneming als resultaat het kenmerk A of het kenmerk $\overline{A}$ (non A) kan geven. De kans op A is binnen ieder der reeksen constant en wel gelijk aan p_i voor de waarnemingen van reeks R_i $(i = 1,2,\ldots,k)$.
Laat het aantal waarnemingen van reeks R_i $(i = 1,2,\ldots,k)$ gelijk zijn aan n_i en laat hieronder het aantal met kenmerk A, m_i zijn.
Gevraagd wordt dan de hypothese H_o: $p_1=p_2= \ldots =p_k$
te toetsen op grond van deze gegevens.
De gegevens kunnen in een 2 x k -tabel worden samengevat;

	R_1	R_2		R_i		R_k	totaal
A	m_1	m_2	...	m_i	...	m_k	m
$\overline{A}$	n_1-m_1	n_2-m_2	...	n_i-m_i	...	n_k-m_k	n-m
totaal	n_1	n_2		n_i		n_k	n

waarin dus $m_1 + m_2 + \ldots + m_k = m$
en $n_1 + n_2 + \ldots + n_k = n$.
De hypothese H_o wordt getoetst met de grootheid

$$\chi_c^2 = \sum_i \frac{\left(m_i - \frac{m\,n_i}{n}\right)^2}{\frac{m\,n_i}{n}} + \sum_i \frac{\left(n_i - m_i - \frac{(n-m)n_i}{n}\right)^2}{\frac{(n-m)\,n_i}{n}} =$$

$$= \sum_i \frac{(n\,m_i - m\,n_i)^2}{m(n-m)\,n_i} = \frac{n^2}{m(n-m)} \sum_i \frac{m_i^2}{n_i} - \frac{nm}{n-m}$$

Deze grootheid $\underline{\chi_c^2}$ [2] heeft onder de hypothese H_o bij benadering een χ^2-verdeling met k-1 vrijheidsgraden (zie b.v. [1] p. 445 e.v.).

---

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.
2) Als wij grootheden als stochastische grootheden (dit zijn grootheden met een waarschijnlijkheidsverdeling) beschouwen geven wij dit door onderstreping aan. Niet onderstreepte letters geven waarden aan, die door de stochastische grootheden worden aangenomen.

Deze benadering is goed, indien $m\frac{n_i}{n} \geqq 5$ voor iedere i (zie [2]). Indien H_o onjuist is, dus als er bij verschillende reeksen verschillende kansen op A zijn, zal $\underline{\chi}^2_c$ gewoonlijk grotere waarden aannemen, dan wanneer H_o juist is.
De kritieke zone bestaat uit die waarden van $\underline{\chi}^2_c$, waarvoor geldt $\underline{\chi}^2_c \geqq \chi^2_\alpha$. Hierin is χ^2_α die waarde van $\underline{\chi}^2$, die voldoet aan

$$P[\underline{\chi}^2 \geqq \chi^2_\alpha] = \alpha$$

met α als van te voren vastgelegde onbetrouwbaarheid.
De overschrijdingskans behorende bij een bepaalde gevonden waarde χ^2_c van $\underline{\chi}^2_c$ is gedefinieerd als

$$P[\underline{\chi}^2_c \geqq \chi^2_c \mid H_o]$$

waarin "$\mid H_o$" aangeeft, dat deze kans berekend wordt op grond van H_o.
χ^2_α en de overschrijdingskans kunnen in tabellen of nomogrammen worden opgezicht (zie [3]).

Opmerking. Indien niet voldaan is aan de voorwaarde $m\frac{n_i}{n} \geqq 5$ voor iedere i, kan men een (meer bewerkelijke) exacte toets baseren op de voorwaardelijke waarschijnlijkheidsverdeling van de grootheden $\underline{m}_i$ ($i = 1,\ldots,k$), onder de voorwaarde, dat hun som de waarde m aanneemt:

$$P[\underline{m}_1 = m_1, \underline{m}_2 = m_2, \cdots, \underline{m}_K = m_K \mid \underline{m}_1 + \underline{m}_2 + \cdots + \underline{m}_K = m ; H_o] =$$

$$= \binom{n_1}{m_1}\binom{n_2}{m_2}\cdots\binom{n_K}{m_K} \Big/ \binom{n}{m}$$

De geldigheid van deze formule volgt direct uit de waarschijnlijkheidsverdelingen van de $\underline{m}_i$ en van $\underline{m}$ (onder H_o) en uit de definitie van een voorwaardelijke waarschijnlijkheid.

In dit geval definiëren wij de overschrijdingskans behorend bij een gevonden resultaat $(m_1, m_2, \ldots, m_K)$ met $m_1 + m_2 + \ldots + m_K = m$ als de som van alle waarschijnlijkheden van bovengenoemde verdeling (met de gevonden waarde van m), die hoogstens gelijk zijn aan de waarschijnlijkheid van het gevonden resultaat.

Literatuur.

[1] H.Cramér, Mathematical methods of statistics, Princeton University Press, 1946.
[2] P.G.Hoel, On indices of dispersion, Ann. Math. Stat. 14 (1943), p. 155-163.

[3] Tabellen en nomogrammen van de -verdeling.
M.G.Kendall, The advanced theory of statistics, I, 1947, p. 444-446.
H.Cramér, Mathematical methods of statistics, Princeton University Press, 1946, p. 559.
Statistica 1 (1946), p. 109.