Receptuur

Uitschietende kansen *)

UDC 311.13

We beschouwen r reeksen $R_1, R_2, \ldots, R_r$ van onafhankelijke waarnemingen, waarbij elk der waarnemingen het kenmerk A wel of niet kan bezitten. We willen hier nagaan of een extreem groot dan wel een extreem klein aantal waarnemingen van het kenmerk A in de reeks R_k in vergelijking met het aantal waarnemingen van het kenmerk A in de andere reeksen, moet worden toegeschreven aan een afwijkende kans op het optreden van A binnen die reeks R_k. Uiteraard veronderstellen we dat de kans op het optreden van A binnen elke reeks apart, constant is.

De te toetsen hypothese H_0 is nu: de kans op het optreden van A is voor de verschillende R_i ($i = 1, 2, \ldots, r$) gelijk. Wordt de nulhypothese verworpen, dan beschouwen wij het aantal waarnemingen in de betreffende reeks als een uitschieter.

De gegevens kunnen worden samengevat in een $2 \times r$-tabel:

kenmerk \ reeks	R_1	R_2	...	R_k	...	R_r	totaal
A	m_1	m_2		m_k		m_r	m
non A	$n_1 - m_1$	$n_2 - m_2$		$n_k - m_k$		$n_r - m_r$	$n - m$
totaal	n_1	n_2		n_k		n_r	n

waarin dus: $m_1 + m_2 + \ldots + m_r = m$

$n_1 + n_2 + \ldots + n_r = n.$

Stel dat m_k extreem groot uitvalt, dan willen we de kans berekenen dat m_k, of een waarde die minstens even extreem groot als m_k is, voorkomt in resp. de reeks R_k of de andere reeksen R_i ($i \neq k$). We redeneren als volgt:

Beschouw naast n en $n_1, n_2, \ldots, n_r$ ook m als gegeven. De samenstelling van de reeks R_i kan dan worden opgevat als het resultaat van n_i aselecte trekkingen zonder teruglegging uit n elementen waarvan m het kenmerk A vertonen. De kans op m_i elementen met kenmerk A in reeks R_i wordt dan gegeven door

$$(1) \quad P[m_i] = \frac{\binom{m}{m_i} \cdot \binom{n-m}{n_i-m_i}}{\binom{n}{n_i}}, \quad \max(0, m-n+n_i) \leqslant m_i \leqslant \min(m, n_i),$$

de hypergeometrische verdeling. Dit is dus een voorwaardelijke verdeling van m_i

*) Rapport S 289 van het Mathematisch Centrum, Amsterdam.

onder de voorwaarde $\sum_{i=1} \underline{m}_i = m$. Indien hierbij voor alle i $m_i << n_i$, dat wil zeggen indien het kenmerk A in elk der reeksen R_i $(i = 1, 2, \ldots, r)$ zelden optreedt, dan geldt bij benadering:

$$P[m_i] = \binom{m}{m_i} \cdot \frac{(n-m)!}{(n_i - m_i)!\,(n - m - n_i + m_i)!} \cdot \frac{n_i!\,(n-n_i)!}{n!} =$$

$$= \binom{m}{m_i} \frac{n_i!}{(n_i - m_i)!} \cdot \frac{(n-n_i)!}{(n - m - n_i + m_i)!} \cdot \frac{1}{\frac{n!}{(n-m)!}} =$$

$$= \binom{m}{m_i} \frac{n_i(n_i-1)\ldots(n_i - m_i + 1)}{n(n-1)\ldots(n - m_i + 1)} \cdot$$

$$\frac{(n-n_i)\,(n-n_i-1)\ldots(n-n_i-m+m_i+1)}{(n-m_i)\,(n-m_i-1)\ldots(n-m_i-m+m_i+1)} \approx$$

$$\text{(2)} \qquad \approx \binom{m}{m_i} \left(\frac{n_i}{n}\right)^{m_i} \cdot \left(\frac{n-n_i}{n}\right)^{m-m_i} = \binom{m}{m_i} p_i^{m_i} (1-p_i)^{m-m_i},$$

waarbij $p_i = n_i/n$.

Ook deze benaderde verdeling heeft een duidelijke interpretatie: Onder de hierboven genoemde voorwaarden kan de verdeling van het (in verhouding kleine) m-tal elementen met het kenmerk A over de r reeksen R_i $(i = 1, 2, \ldots, r)$ worden beschouwd als een trekking met (bij benadering) constante kansen $p_i = n_i/n$ voor de verschillende reeksen. Hieruit volgt echter dat $\underline{m}_i$ bij benadering een binomiale verdeling (2) bezit.

De overschrijdingskansen $s_i = \sum_{j=m_i}^{m} P[j]$ kunnen zowel voor de exacte als voor de benaderde verdeling in een tabel worden gevonden (bijv. [3] resp. [4]). Is nu

$$s_k = \min_{1 \leqslant i \leqslant r} s_i \leqslant \frac{\alpha}{r},$$

waarbij α de gekozen onbetrouwbaarheidsdrempel voorstelt, dan geldt voor de kans s op een minstens even extreem resultaat als de gevonden m_k in de reeks R_k en/of één of meer der andere reeksen R_i

$$s \leqslant r \cdot s_k \leqslant \alpha \qquad \text{(zie [1]).}$$

De hypothese H_o kan dan verworpen worden en wij beschouwen de waarde m_k als een uitschieter naar boven.

Indien er meer zulke extreem hoge waarden voorkomen en we gevonden hebben dat er al een uitschieter is, dan kunnen wij dit procédé herhalen voor de $k - 1$ overgebleven reeksen met totaal aantal $n - n_i$ en totaal aantal van het kenmerk A: $m - m_i$, enz.

Willen wij echter nagaan of er uitschieters naar beneden voorkomen (d.w.z. extreem lage waarden van één of meer der m_i), dan dienen wij links éénzijdig te toetsen. Wij beschouwen dan de cumulatieve kansen

$$d_i = \sum_{j=1}^{m_i} P\,[j]$$

en besluiten tot verwerping van H_0 als

$$d_l = \min_{1 \leqslant i \leqslant r} d_i \leqslant \frac{\alpha}{r}.$$

Bij meerdere extreem kleine waarden kunnen we weer te werk gaan als in het andere geval indien we bij de voorgaande toetsing(en) uitschieter(s) gevonden hebben.

Tenslotte kan ook tweezijdig worden getoetst op uitschieters naar boven zowel als naar beneden. H_0 wordt dan verworpen indien

$$\min\,(s_k,\, d_l) \leqslant \frac{\alpha}{2r}.$$

Voorbeeld (ontleend aan [2], blz. 164)

In een artikel van W. D. R o s s e.a. "The association of certain vegetative disturbances with various psychoses", Psychosom. Med. **12** (1950), p. 170—178, komt de volgende tabel voor:

diagnostische groep	verwachting v. h. aantal diabetici	aantal diabetici	aantal niet diabetici	totale aantal
1	2,373	4	208	212
2	2,664	0	238	238
3	2,496	2	221	223
4	0,851	1	75	76
5	0,571	0	51	51
6	0,504	0	45	45
7	0,358	0	32	32
8	0,179	2	14	16
9	1,690	0	151	151
10	0,672	1	59	60
11	1,601	3	140	143
12	0,604	0	54	54
13	0,504	1	44	45
14	0,369	0	33	33
15	0,347	0	31	31
16	0,549	4	45	49
17	0,504	0	45	45
18	0,369	0	33	33
19	0,795	0	71	71
totaal	18	18	1590	1608

In de tweede kolom is de verwachting van het aantal diabetici neergeschreven; die verwachting wordt als volgt berekend:

$$\text{verwachting van het aantal diabetici in de } i^e \text{ rij} = \frac{(\text{aantal patiënten } i^e \text{ rij}) \times (\text{totale aantal diabetici})}{\text{totale aantal patiënten}}$$

dus bijvoorbeeld: $2{,}373 = \frac{212 \times 18}{1608}$.

Om nu te toetsen of de aantallen diabetici in de verschillende diagnostische groepen met gelijke waarschijnlijkheden optreden paste R o s s de χ^2-toets toe. Dit is hier echter niet raadzaam daar in de meeste rijen de verwachting van het aantal diabetici te klein is voor toepassing van de χ^2-toets. R o s s vindt:

$$\chi^2 = 51{,}8 \text{ met } 18 \text{ vrijheidsgraden}$$

wat een overschrijdingskans $< 0{,}01$ geeft en hij zegt dat dit te danken is aan het aantal diabetici in de diagnostische groep 16. Dit bedraagt 4 bij een verwachting van 0,55. Om na te gaan of dit aantal van 4 als een uitschieter naar boven beschouwd kan worden passen wij bovenstaande toets toe:

$$n = 1608,\ m = 18,\ n_k = n_{16} = 49,\ m_k = m_{16} = 4.$$

$$p_k = p_{16} = \frac{n_k}{n} = \frac{49}{1608} = 0{,}03,$$

$$s_k = \sum_{j=4}^{18} P\,[j] \approx 0{,}0018 \quad \text{(benaderd volgens (2))};$$

daar

$$s_k = \min_{1 \leqslant i \leqslant 19} s_i \approx 0{,}0018 < \frac{0{,}05}{19} = 0{,}0026,$$

kan de diagnostische groep 16 dus inderdaad als uitschieter naar boven beschouwd worden indien we $\alpha = 0{,}05$ kiezen. We schrappen deze groep nu en willen nagaan of in de overblijvende groepen de groep 8 die 2 diabetici bevat met een verwachting van 0,18 nog als uitschieter naar boven beschouwd kan worden; we vinden:

$$n' = 1559,\ m' = 14,\ n_k' = n_8 = 16,\ m_k' = m_8 = 2$$

$$p_k' = p_8 = \frac{n_k'}{n'} = \frac{16}{1559} = 0{,}01,$$

$$s_k' = \sum_{j=2}^{14} P\,[j] \approx 0{,}0084 \text{ (benaderd volgens (2))},$$

daar

$$s_k' = \min_{1 \leqslant i \leqslant 18} s_i' \approx 0{,}0084 > \frac{0{,}05}{18} = 0{,}0028$$

kan deze groep dus *niet* als een uitschieter naar boven beschouwd worden.

Indien men daarentegen wenst na te gaan of in de tabel van Ross uitschietende waarden naar beneden voorkomen, zoeken we die d_l die de kleinste waarde heeft, dit is

$$d_l = d_2 = \min_{1 \leqslant i \leqslant 19} d_i \approx 0{,}054.$$

Bij een keuze van $\alpha = 0{,}05$ kan R_2 dus niet als uitschieter naar beneden worden beschouwd.

Literatuur:

[1] Doornbos, R. en H. J. Prins (1958): "On slippage tests, II. Slippage tests for discrete variables", Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A), **61** ≡ Indagationes Mathematicae, **20**, 47—55.

[2] Ladée, G. A. (1961); "Hypochondrische syndromen", proefschrift Amsterdam.

[3] Lieberman, G. J. and D. B. Owen (1961): "Tables of the hypergeometric distribution", Stanford University Press.

[4] National Bureau of Standards (1950): "Tables of the Binomial probability distribution", Applied Mathematics, Series 6.

Catharina Korswagen,
Statistische Afdeling van het Mathematisch Centrum.