S 190 (M 77)

Enige toetsingsmethoden voor de aselectheid van rijen getallen

Stichting Mathematisch Centrum
Statistische Afdeling
1956

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O.

Statistische Afdeling.
Rapport S 190 (M 77).

Enige toetsingsmethoden voor de aselectheid van rijen getallen [1]

1. Inleiding

Bij de toepassingen van de statistiek op praktische problemen maakt men veelvuldig gebruik van rijen aselecte getallen [2]. Een rij aselecte getallen wordt verkregen door een aantal getallen te trekken uit een gegeven verzameling en wel zo, dat ieder getal bij iedere trekking dezelfde kans heeft om gekozen te worden. Gewoonlijk gaat men uit van de verzameling der getallen 0,1,2,..,9; wij zullen ons hier tot dat geval beperken. Uit deze definitie volgt, dat in een rij aselecte getallen geen systematisch verband mag bestaan tussen de op elkaar volgende getallen.

Wij zullen hier een aantal toetsen bespreken, waarmee men na kan gaan of een beschikbare rij getallen op aselecte wijze tot stand gekomen kan zijn. Hierop zal een overzicht van de belangrijkste literatuur volgen.

Notatie:

H_0: Hypothese dat de onderzochte getallen aselect zijn.

H_1, H_2,... alternatieve hypothesen.

x = een willekeurig cijfer uit de rij. N = 0,1,...,9 = aantal getallen van een gegeven rij.

Het stochastische karakter van een variabele wordt aangegeven door onderstreping van het desbetreffende symbool.

Voor de algemene gang van zaken bij het toetsen van een hypothese zie men memorandum S 47 (M 6).

---

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

2) Engels: "random numbers".

veel gecompliceerder. Daar het in de regel slechts van belang is, om te weten of een gegeven rij getallen als aselect kan worden beschouwd, en niet in de eerste plaats welke afwijkingen van aselectheid aanwezig zijn, zullen de hier te behandelen toetsen voor de praktijk voldoende zijn. Bij het trekken van de conclusies dient men echter met dit feit rekening te houden.

2.2. Kettingcorrelatietoets (ontleend aan [1])

Men telt het aantal keren $\underline{m}$ dat twee opeenvolgende getallen in de rij gelijk zijn. In de rij 8 5 7 7 3 0 9 9 9 4 is b.v. $\underline{m} = 3$. Onder H_o heeft $\underline{m}$ de binomiale verdeling van het aantal successen bij een reeks van N-1 experimenten met kans 1/10 op succes; voor voldoende grote N heeft $\underline{m}$ dus een normale verdeling met $\mu = \frac{1}{10}(N-1)$ en $\sigma = \frac{3}{10}\sqrt{N-1}$.

Wij definiëren hier de overschrijdingskans onder H_o dat er een waarde van $\underline{m}$ optreedt die minstens even ver van de gemiddelde waarde $\frac{1}{10}(N-1)$ afwijkt als de gevonden waarde. Men verwerpt H_o als de overschrijdingskans kleiner is dan de onbetrouwbaarheidsdrempel; deze toets reageert vooral op alternatieven, waarbij de kans dat reeksen achtereenvolgende cijfers hetzelfde zijn, kleiner of groter is dan de hypothese H_o.

2.3. Sommen-toets

Men verdeelt de te onderzoeken rij getallen in groepjes van k cijfers. Binnen elk groepje bepaalt men de som $\underline{s}_k$ van deze cijfers. De verdeling van s_k onder H_o is voor k=1 gegeven door $p(a) = P[\underline{s}_1 = a] = \frac{1}{10}$ voor a=0,1,...,9, voor k=2 door

a	0 18	1 17	2 16	3 15	4 14	5 13	6 12	7 11	8 10	9
100 p(a)	1	2	3	4	5	6	7	8	9	10

en voor k=3 door

a	0 27	1 26	2 25	3 24	4 23	5 22	6 21	7 20	8 19	9 18	10 17	11 16	12 15	13 14
1000p(a)	1	3	6	10	15	21	28	35	45	55	63	69	73	75

Door samenstellen van de verdelingen van $\underline{s}_3$ en $\underline{s}_1$ kan men de verdeling van $\underline{s}_4$ afleiden, enz. Men vergelijkt nu de frequentieverdeling van $\underline{s}_k$ gevonden uit de steekproef met de waarschijnlijkheidsverdeling onder de hypothese H_o, met behulp van de χ^2-toets van aanpassing. Voor k=1 komt dit dus neer op het toepassen van de frequentietoets zoals deze onder 2.1 is beschreven.

Deze toets reageert speciaal op alternatieven waarbij de verdeling van de partiële sommen afwijkt van de verdeling onder H_o. Dit zou b.v. het geval kunnen zijn indien de reeks getallen periodiciteiten vertoont die met k samenhangen.
Literatuur: zie [2].

2.4. Openingen-toets (ontleend aan [1])

Men beschouwt in de te onderzoeken rij getallen alle cijfers x en telt bij ieder paar opeenvolgende cijfers x het aantal y der andere cijfers welke die twee cijfers x scheiden. In de rij 0 3 2 0 0 9 7 5 0 4 4 3 vindt men bij x=0 achtereenvolgens y=2, 0 en 3. Men bepaalt nu de frequentieverdeling van y in de steekproef, dus voor iedere i het aantal keren a_i dat y=i is. Wij vergelijken de frequenties a_i met hun verwachtingswaarden onder de hypothese H_o: $\mathcal{E}\,\underline{a}_i = (0{,}9)^i \cdot \frac{N}{10}$, met behulp van de χ^2-toets van toepassing.

Deze toets reageert speciaal op alternatieven die inhouden dat de cijfers x niet aselect over de gehele rij getallen zijn verdeeld. Dit zou b.v. eveneens het gevolg kunnen zijn van periodiciteiten in de rij getallen. Bovendien reageert de toets op het speciale alternatief van de kettingcorrelatietoets (zie 2.2).

2.5. Poker-toets (ontleend aan [1])

Men verdeelt de te onderzoeken rij getallen in groepjes van 4 cijfers, zodat het aantal beschouwde groepjes dus N/4 wordt. In al deze groepjes telt men nu het aantal keren dat alle 4 elementen van het groepje gelijk zijn (a a a a); dat er 3 gelijken onder voorkomen (a a a b); dat er 2 paar gelijken voorkomt (a a b c) en dat alle 4 elementen verschillend zijn (a b c d). Deze gevonden waarden vergelijkt men met behulp van de χ^2-toets van aanpassing met de bijbehorende theoretische aantallen, welke gegeven worden in de volgende tabel:

groepje	verwachting van het aantal groepjes
a a a a	$10^{-3} . N/4$
a a a b	$36.10^{-3} . N/4$
a a b b	$27.10^{-3} . N/4$
a a b c	$432.10^{-3} . N/4$
a a c d	$504.10^{-3} . N/4$
totaal	$N/4$

Indien de kleinste van deze verwachtingswaarden, doordat het aantal N te klein is, kleiner wordt dan 5, combineert men a a a a met een der andere groepjes teneinde dit bezwaar te vermijden.

De toets reageert speciaal op alternatieven, die inhouden dat te zelden of te vaak dezelfde cijfers achter elkaar voorkomen.

2.6 Op en neer-toets

Men splitst de rij getallen in groepjes van 2. Zijn de twee cijfers van een groepje gelijk, dan laat men dit groepje weg. Men houdt dus b.v. N_1 groepjes over, waarbij het eerste cijfer hetzij kleiner is dan het tweede (op-groepjes), hetzij groter (neer-groepjes). Men telt nu het aantal $\underline{n}_1$ der op-groepjes; en toetst met de tekentoets of dit systematisch afwijkt van $\frac{1}{2}N_1$ $(= \mathcal{E}(\underline{n}_1 | H_o))$. In de praktijk kan men de normale benadering gebruiken voor de verdeling van $\underline{n}_1$ (gemiddelde $\frac{1}{2}N_1$, spreiding $\frac{1}{2}\sqrt{N_1}$).

Deze toets reageert op alternatieven waarbij de kans op op-groepjes voldoende van $\frac{1}{2}$ afwijkt. Een toets van dit type is behandeld in [5].

2.7. Run-toets

Men verdeelt de in de rij voorkomende cijfers 0,1,...,9 op een of andere wijze in twee groepen met verschillend kenmerk. B.v. geeft men alle even cijfers het kenmerk a en alle oneven cijfers het kenmerk b. Een andere mogelijkheid is, alleen het cijfer 5 het kenmerk a te geven en alle andere cijfers het kenmerk b. Men krijgt zo n keer a en m keer b (n+m=N). Onder een run van a verstaat men nu een groepje elementen die alle het kenmerk a dragen, terwijl de elementen ter weerszijden hiervan (zo deze er zijn) het kenmerk b dragen. De toetsingsgrootheid r is nu het aantal runs van a dat in de reeks optreedt. Is $P\{a\} = p$ en $P\{b\} = q$ (en zijn n en m dus binomiaal verdeeld) dan is de verwachting van r onder H_0 gelijk aan

$$\mu = N p q + p^2$$

en de variantie

$$\sigma^2 = Npq\ (1-3pq) + p^2\ (3-8p+5p^2)\ .$$

Voor voldoend grote N is r bij benadering normaal verdeeld.

Door keuze van de kenmerken a en b kan men runs van elke gewenste combinatie van kenmerken op deze wijze onderzoeken. De toets reageert op alternatieve hypothesen waarbij de gekozen kenmerken a en b niet aselect over de rij getallen zijn verdeeld.

Literatuur: zie [3] en [4].

Literatuur

[1] M.G. KENDALL and B. BABINGTON SMITH: Tables of random sampling numbers. Tracts for computers no.24, Cambridge University Press 1946, p.8 .

[2] G.M. YULE: A test of Tippett's random sampling numbers. Journ. Roy. Statist. Soc. 101 (1938) 167.

[3] A.M. MOOD: The distribution theory of runs. Ann. Math. Stats. 11 (1940) 367-392 .

[4] W. GONTCHAROFF: Sur la succession des événements dans une série d'épreuves indépendantes repondant au schème de Bernoulli. C.R.Acad.Sci. URSS 38 (1943) 283-285 .

[5] E.L. DODD: Certain tests for randomness applied to data grouped into small sets. Econometrica 10 (1942) 249 .