MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S219 (Ov 7)

Overzichtsrapport: Over de verdeling van het aantal "runs" in reeksen van alternatieven.

door

Hilda A. Kuipers.

1957

Inhoud blz.

Inleiding en probleemstelling

1. Definitie van runs

In dit rapport beschouwen wij reeksen van elementen, die alle hetzij kenmerk a, hetzij kenmerk b bezitten. Een dergelijke reeks wordt een "reeks van alternatieven" genoemd. Wij nemen aan dat voor de elementen op een of andere wijze een volgorde is gedefinieerd en dat de elementen in deze volgorde zijn geplaatst; in een dergelijke reeks zullen nu groepen elementen met hetzelfde kenmerk optreden; zo'n groep elementen, aan weerszijden begrensd door elementen met een ander kenmerk of door het begin of einde van de reeks van alternatieven noemen wij een "run". Naar de elementen die in de run voorkomen, onderscheiden wij "runs van a" en "runs van b". Het aantal runs van a, dat in een reeks van alternatieven optreedt, duiden wij aan met de letter r; het aantal runs van b met de letter s; het totaal aantal runs met de letter $t = r + s$. Verder geven we het aantal elementen met kenmerk a in de reeks aan met n, het aantal elementen met kenmerk b zij m; het totale aantal elementen zij $N = m + n$.

Voorbeeld:

Wij kunnen een reeks van alternatieven verkrijgen door het werpen met een munt. De elementen zijn dan de worpen; de kenmerken a en b corresponderen met kruis en munt. Als volgorde in de reeks nemen wij b.v. de tijdsvolgorde der worpen; indien wij dan vinden

a a b b b a b b a a a

is $m = 6$; $n = 5$; $N = 11$, $r = 3$ en $s = 2$.

Het is gemakkelijk in te zien dat r en s ten hoogste één kunnen verschillen.

1.2 Inhoud van het rapport

Dit rapport bevat een overzicht van de literatuur betreffende de verdeling van het aantal runs, onder verschillende hypothesen omtrent de wijze waarop de reeks van alternatieven tot stand gekomen is, en hiermee verband houdende toetsings methoden. Wij beperken ons uitdrukkelijk tot de aantallen

runs die in de reeks voorkomen, en behandelen dus niet de theorie van de lengten van de runs, d.w.z. de aantallen elementen die in een run optreden. De inhoud van dit rapport berust hoofdzakelijk op de meer recente literatuur (vanaf ± 1940). Uit de oudere literatuur zijn alleen de belangrijkste punten overgenomen. Een volledige lijst van de geraadpleegde literatuur vindt men aan het einde van dit rapport. Bewijzen worden in het algemeen niet of slechts schematisch gegeven, met vermelding van de plaats waar zij volledig.te vinden zijn.

1.3 Notatie

Als symbolen voor stochastische variabelen (dit zijn variabelen die een kansverdeling bezitten) worden steeds onderstreepte latijnse letters gebruikt. Voor willekeurige door een stochastische grootheid aangenomen waarden wordt hetzelfde symbool gebruikt als voor die grootheid, doch zonder onderstreping.

Voor de kans dat de stochastische variabele $\underline{x}$ een waarde x aanneemt, gebruiken wij het symbool $P\{\underline{x} = x\}$, of korter $P\{x\}$. Op overeenkomstige wijze worden de symbolen $P\{\underline{x} \leq x\}$, $P\{\underline{x} \geq x\}$ enz. gedefinieerd. Het symbool $P\{\underline{x} = x \,|\, \underline{y} = y\}$ of korter $P\{x|y\}$ stelt voor de kans dat $\underline{x}$ de waarde x aanneemt, onder de voorwaarde dat $\underline{y}$ de waarde y aanneemt. Men gebruikt ook het symbool $P\{\underline{x} = x \,|\, H_0\}$ voor de kans dat $\underline{x}$ de waarde x aanneemt als de hypothese H_0 juist is, terwijl men achter de streep in de notatie $P\{\underline{x} = x \,|\ldots\}$ ook voorwaarden betreffende niet stochastische grootheden kan vermelden.

De elementen van deze reeks van alternatieven worden in de voorgeschreven volgorde aangegeven met $E_1, E_2, \ldots, E_N$. Tevens voeren wij de grootheid $\underline{x}_i$ in, welke als volgt gedefinieerd wordt:

$$\underline{x}_i = \begin{cases} 1 & \text{indien } E_i \text{ het kenmerk a bezit} \\ 0 & \text{" } E_i \text{ " " b "} \end{cases} \qquad (i = 1,2,\ldots,N)$$

Wij gebruiken verder de afkorting:

$$P\{\underline{x}_1 = x_1; \underline{x}_2 = x_2 \ldots; \underline{x}_N = x_N\} = P\{x_1 x_2 \ldots x_N\} .$$

1.4 Onderzochte hypothesen

H_0: De reeks letters a en b vormt het resultaat van een reeks onderling onafhankelijke experimenten, waarbij steeds met kans p een element met kenmerk a, en met kans q = 1 - p een element met kenmerk b optreedt. De reeks correspondeert dus met een serie van N onderling onafhankelijke betrekkingen met terug-legging uit een vaas met ballen waarvan een fractie p het kenmerk a en een fractie q het kenmerk b bezit. Dit wordt uit-gedrukt door de formule

$$(1.4;1) \quad P\{x_1, x_2, \ldots, x_N\} = P\{x_1\} \cdot P\{x_2\} \ldots P\{x_N\} ,$$

waarin iedere $\underline{x}_i$ dezelfde verdeling bezit ($P\{\underline{x}_i = 1\} = p$ en $P\{\underline{x}_i = 0\} = q$).

Wij merken hierbij op, dat het aantal $\underline{n} = \sum_{i=1}^{N} \underline{x}_i$ der elementen met kenmerk a dat in een dergelijke reeks van N elementen voorkomt, een binomiale verdeling bezit, dus dat geldt:

$$(1.4;2) \quad P\{\underline{n} = n \mid H_0'\} = \binom{N}{n} p^n q^m = \frac{N!}{n!\,m!} p^n q^m .$$

H_0': De reeks letters a en b vormt een aselecte permutatie van n letters a en m letters b. Zij correspondeert dus met een serie van N onderling onafhankelijke trekkingen zonder terug-legging uit een vaas die n ballen met kenmerk a en m ballen met kenmerk b bevat.

H_1: De reeks van alternatieven is een enkelvoudige Markoff-keten. Dit wil zeggen dat de waarde die elk element aanneemt alleen afhankelijk is van de waarde van het direct daaraan voorafgaande element. In formule:

$$(1.4;3) \quad P\{x_1, x_2, \ldots, x_N \mid H_1\} = P\{x_1\} \cdot P\{x_2 / x_1\} \ldots P\{x_N / x_{N-1}\} ,$$

waarbij de verdeling van x_i onder de voorwaarde dat x_{i-1} gelijk

is aan 1 resp. 0, niet van i afhangt. Deze verdeling is dus bepaald door de kansen p_{00}, p_{01}, p_{10} en p_{11}, de z.g. overgangswaarchijnlijkheden, welke als volgt worden gedefinieerd:

$$p_{00} = P\{\underline{x}_i = 0 \mid \underline{x}_{i-1} = 0\} \qquad p_{10} = P\{\underline{x}_i = 1 \mid \underline{x}_{i-1} = 0\}$$

(1.4;4)

$$p_{01} = P\{\underline{x}_i = 0 \mid \underline{x}_{i-1} = 1\} \qquad p_{11} = P\{\underline{x}_i = 1 \mid \underline{x}_{i-1} = 1\}$$

Hierbij is $p_{0j} + p_{1j} = 1 \quad (j = 0,1)$.
Om de kans $P\{x_1, x_2, \ldots, x_N \mid H_1\}$ te kunnen vinden dient men bovendien nog de verdeling van x_1, dus $p \overset{def}{=} P\{\underline{x}_1 = 1\}$ te kennen.

Opmerking 1: De in 1.1 gegeven definitie van runs is, hoewel de meest gebruikte, niet de enige definitie. Sommige auteurs gebruiken andere definities; wij beperken ons echter tot de bovengegevene.
Opmerking 2: Indien niet anders wordt aangegeven, beschouwen wij uitsluitend rangschikkingen van elementen a en b langs een rechte lijn. Formules betreffende rangschikkingen langs een cirkel zullen worden aangegeven door een index c onder het betreffende symbool. Wij zullen dan spreken van een circulaire reeks.
Opmerking 3: Aangezien over de verdeling onder H_0' het meeste bekend is, terwijl tevens de behandeling van H_0 beter aansluit bij de behandeling van H_1, behandelen wij eerst de hypothese H_0' en daarna pas H_0.

2. Verschillende methoden (hypothesen H_0 en H_0')

2.1 De verdeling van $\underline{r}$ onder H_0'

Het probleem wordt op een uiterst overzichtelijke wijze besproken in een artikel van W.L. STEVENS [17]. Latere auteurs verwijzen herhaaldelijk hiernaar en de door hen genoemde resultaten maken alle van die van Stevens gebruik.

Stevens gaat uit van een circulaire reeks van alternatieven. Bij een dergelijke reeks is steeds $\underline{r} = \underline{s}$. De kans op juist r groepjes a's is nu

$$(2.1;1) \qquad P_c\left\{\underline{r} = r \mid n, m, H_0'\right\} = \frac{\binom{n}{r}\binom{m-1}{r-1}}{\binom{N-1}{m}} \quad .$$

Het bewijs hiervan verloopt in grote trekken als volgt: Er zijn (n-1)! manieren om n (onderscheiden gedachte) elementen a op de cirkel te rangschikken en er zijn $\binom{n}{r}$ manieren om r tussenruimten tussen de a's te maken, dus $(n-1)!\binom{n}{r}$ manieren om de n elementen a met r tussenruimten op de cirkel te rangschikken. Evenzo zijn er voor de m elementen b $(m-1)!\binom{m}{r}$ manieren om deze te rangschikken met r tussenruimten. Men kan nu nog op r manieren de b's tussen de a's plaatsen, en aangezien het totaal aantal rangschikkingen van n elementen a en m elementen b langs een cirkel gelijk is aan (N-1)! wordt de gevraagde kans

$$P_c\left\{\underline{r} = r \mid n, m, H_0'\right\} = \frac{r(n-1)!\binom{n}{r}(m-1)!\binom{m}{r}}{(N-1)!} = \frac{\binom{n}{r}\binom{m-1}{r-1}}{\binom{N-1}{m}} \quad .$$

De verwachting van $\underline{r}$ wordt volgens O.LUDWIG [10]:

$$(2.1;2) \qquad \mathcal{E}(\underline{r} \mid n, m, H_0') = \frac{mn}{N-1}$$

en de variantie

$$(2.1;3) \qquad \sigma_c^2(\underline{r} \mid n, m, H_0') = \frac{n(n-1)m(m-1)}{(N-1)^2(N-2)} \quad .$$

Om de overeenkomstige formules voor rangschikking langs een rechte lijn te verkrijgen, dient men slechts m door m+1 te vervangen. Dat dit zo is, ziet men eenvoudiger in door van een circulaire reeks uit te gaan, één b willekeurig hierbij te plaatsen, en vervolgens op één der m+1 elementen b door te knippen, waarbij men dan meteen dit element weglaat. Door deze handeling ondergaan de aantallen n en m geen wijziging, terwijl de rangschikking nu evenveel runs van a bevat als in het circulaire geval met aantallen n en m+1.

Opm.: Bij een rangschikking langs een rechte lijn kunnen zoals men gemakkelijk inziet, r en s ten hoogste 1 verschillen, en dus zal t ten hoogste verschillen van 2r.

Voor rangschikking van n elementen a en m elementen b langs een rechte lijn worden de formules dus nu:

$$(2.1;4) \qquad P\{\underline{r} = r / n, m, H_0'\} = \frac{\binom{n}{r}\binom{m}{r-1}}{\binom{N}{m+1}}$$

voor de kans op juist r runs van a; (deze formule wordt ook vermeld door A.M. MOOD [11], S.S. WILKS [20] en W. FELLER [5]). Men vindt verder voor gemiddelde en variantie van r:

$$(2.1;5) \qquad \mathcal{E}(\underline{r}/n,m,H_0') = \frac{(m+1)n}{N} \; ; \; \sigma^2(r/n,m,H_0') = \frac{(m+1)m.n(n-1)}{N^2(N-1)} \; .$$

(vgl. ook MOOD. [11]).

Volgens (2.1;4) is de verdeling van r hypergeometrisch en dus asymptotisch normaal voor $N \to \infty$, $m \to \infty$ en $n \to \infty$, met gemiddelde en variantie gegeven door (2.1;5)

2.2 Toetsingsmethoden voor H_0'

Wij willen vervolgens voor een bepaalde reeks de hypothese H_0' toetsen, en gebruiken daarbij r als toetsingsgrootheid. Voor kleine waarden van N kan men de overschrijdingskans van een gevonden waarde van r exact bepalen: Indien men rechtséénzijdig wil toetsen, sommeert men alle $P[r_i]$ met $r_i \geq r$; bij tweezijdige toetsing sommeert men alle $P[r_i]$ met $P[r_i] \leq P[r]$ - Het kritieke gebied (onbetrouwbaarheidsdrempel α) wordt bij rechtseenzijdige toetsing gevormd door alle waarden van r, waarvoor de zo juist gedefinieerde overschrijdingskans resp. voor rechtséénzijdige en tweezijdige toetsen) niet groter is dan α.

Voor grote waarden van N maakt men gebruik van de asymptotische normaliteit van de verdeling van r.

2.3 De verdeling van $\underline{t}$ onder H_0'

In het geval van rangschikking langs een rechte lijn wordt de kans dat het totale aantal runs $\underline{t}$ in de reeks een bepaalde waarde aanneemt, gegeven door:

(2.3;1)
$$\text{voor } t = 2\nu \quad P[\underline{t} = 2\nu \mid H_0'] = \frac{2\binom{n-1}{\nu-1}\binom{m-1}{\nu-1}}{\binom{N}{n}},$$

$$\text{voor } t = 2\nu+1 \quad P[\underline{t} = 2\nu+1 \mid H_0'] = \frac{\binom{n-1}{\nu}\binom{m-1}{\nu-1} + \binom{n-1}{\nu-1}\binom{m-1}{\nu}}{\binom{N}{n}}.$$

Deze formules vinden wij onder meer bij S.S. WILKS [20], F.N. DAVID [4], E. ISING [8] en W.L. STEVENS [17]. F.C. SWED en C. EISENHART [18] hebben de verdeling getabelleerd, evenals M. OLEKIEWICZ [15].

A. WALD en J. WOLFOWITZ [19] geven gemiddelde en variantie:

(2.3;2)
$$\mathcal{E}(\underline{t} \mid H_0') = \frac{2nm}{N} + 1;$$

(2.3;3)
$$\sigma^2(\underline{t} \mid H_0') = \frac{2nm(2nm-N)}{N^2(N-1)}.$$

WALD en WOLFOWITZ [19] bewijzen dat voor m en n zeer groot, de verdeling bij benadering normaal is met gemiddelde en spreiding volgens (2.3;2) en (2.3;3). Het bewijs maakt gebruik van de formule van STIRLING en geldt alleen voor het geval dat $\frac{m}{n} = \alpha$ constant gehouden wordt. Zowel MOORE [12] en MORICE [14] bewijzen hetzelfde, op enigszins andere wijze en met iets minder beperking t.a.v. $\frac{m}{n}$.

Evenals bij $\underline{r}$ besproken is kan men op de verdeling van $\underline{t}$ een toets voor H_0' baseren.

Als de elementen a en b langs een cirkel gerangschikt zijn, is dus $\underline{r} = \underline{s}$ en wordt

(2.3;4)
$$P_c[\underline{t} = 2\nu \mid H_0'] = \frac{\binom{n-1}{\nu-1}\binom{m}{\nu}}{\binom{N-1}{n}},$$

(2.3;5)
$$P_c[\underline{t} = 2\nu+1 \mid H_0'] = 0.$$

STEVENS leidt af, dat hiervoor

$$(2.3;6) \qquad \mathcal{E}\left(\underline{t}/n,m,H_0'\right) = \frac{2mn}{N-1} \quad .$$

$$(2.3;7) \qquad \sigma_c^2\left(\underline{t}/n,m,H_0'\right) = \frac{4n(n-1)m(m-1)}{(N-1)^2(N-2)} \quad .$$

Tabellen: OLEKIEWICZ [15] geeft voor n en $m \leq 20$ tabellen van $\mu(\underline{t}/n,m,H_0')$ en $\sigma^2(\underline{t}/n,m,H_0')$; ook van $P[\underline{t} > \mu(t)]$. SWED en EISENHART [18] geven voor m en $n \leq 20$ tabellen van $P[\underline{t} \leq t/n,m,H_0']$ in zeven decimalen.

2.4 Toets van WALD en WOLOWITZ.

WALD en WOLFOWITZ [19] hebben een verdelingsvrije toets voor twee steekproeven opgesteld, welke op de voorafgaande theorie gebaseerd is.

Laten $y_1, \ldots, y_n$ resp. $z_1, \ldots, z_m$ de steekproeven zijn, genomen uit (continu onderstelde) verdelingen met verdelingsfuncties $F(x)$ resp. $G(x)$. Vervolgens rangschikt men de waarnemingen van beide steekproeven gezamenlijk naar opklimmende grootte. Men beschouwt de zo verkregen rij als een reeks alternatieven waarin het kenmerk a overeenkomt met een waarneming uit de steekproef der y's, en het kenmerk b met een waarneming uit de steekproef der z's, en definieert verder de $\underline{x}_i$ als in (1.3). De toetsingsgrootheid is het totale aantal runs $\underline{t}$. Indien $F \equiv G$, is voldaan aan de hypothese H_0', zodat gemiddelde en variantie van $\underline{t}$ voor dat geval gegeven worden door (2.3;2) en (2.3;3). In het artikel wordt het onderscheidingsvermogen van de toets niet behandeld. Wel wordt bewezen dat de toets onder zekere voorwaarden bruikbaar (E: consistent) is. Voor een omschrijving van deze voorwaarden verwijzen wij naar het artikel van WALD en WOLFOWITZ; wij vermelden slechts dat deze in de praktijk meestal wel vervuld zijn.

2.5 De verdeling van $\underline{r}$ onder H_0

Deze verdeling wordt besproken in een artikel van W. GONTCHAROFF [6]. Alle in dit artikel voorkomende formules worden niet bewezen; de bewijzen zijn verschenen in een ander

artikel van GONTCHAROFF [7].

De mathematische verwachting van $\underline{r}$ is

$$(2.5;1) \qquad \mathcal{E}(\underline{r} \mid H_0, p, q, N) = Npq + p^2$$

en de variantie

$$(2.5;2) \qquad \sigma^2(\underline{r} \mid H_0, p, q, N) = pq[(N-2)p^2 - (N-3)pq + Nq^2] .$$

De kans dat de reeks juist r runs van a bevat, is gelijk aan de coëfficiënt van x^r in het polynoom P_N, dat gedefinieerd wordt door de volgende recursierelatie:

$$(2.5;3) \qquad \begin{aligned} &P_N(x) = (p+q)P_{N-1}(x) - pq(1-x)P_{N-2} \qquad P_0(x) = 1 \\ &P_1(x) = q + px . \end{aligned}$$

Voor grote N is $\underline{r}$ asymptotisch normaal verdeeld met boven genoemde gemiddelde en spreiding, hetgeen eveneens in het tweede artikel van GONTCHAROFF bewezen wordt.

Evenals in het voorgaande kan men dus voor grote N de hypothese H_0 toetsen met $\underline{r}$ als toetsingsgrootheid, hierbij gebruikmakende van de asymptotische normaliteit.

2.6 De verdeling van t onder H_0

W. GONTCHAROFF [6] geeft voor de verdeling van $\underline{t}$ onder H_0 als gemiddelde:

$$(2.6;1) \qquad \mathcal{E}(\underline{t} \mid H_0, p, q, N) = 2Npq + (p^2 + q^2)$$

en als variantie:

$$(2.6;2) \qquad \sigma^2(\underline{t} \mid H_0, p, q, N) = 2pq[2(p^2 - pq + q^2)N - (3p^2 - 4pq + 3q^2)] \qquad \text{voor } N \geq 2 .$$

Deze verdeling vinden wij ook bij VON BORTKIEWICZ [2] en LUDWIG [10].

In het bijzondere geval dat $p = q = \frac{1}{2}$ wordt dit de binomiale verdeling:

$$(2.6;3) \qquad P[t \mid H_0, p = \tfrac{1}{2}, N] = \binom{N-1}{t-1}\left(\tfrac{1}{2}\right)^{N-1} .$$

G. SCHULTZ [16] leidt langs andere weg de formules (2.6;1) en (2.6;2) eveneens af, en geeft bovendien nog de asymptotische formules voor gemiddelde en variantie:

$$(2.6;4) \qquad \mathcal{E}(\underline{t} \mid H_0, p, q, N) \cong 2Npq$$

en

$$(2.6;5) \qquad \sigma^2(\underline{t} \mid H_0, p, q, N) \cong N[4pq(1-3pq)] ,$$

welke eenvoudig uit (2.6;1) en (2.6;2) kunnen worden afgeleid door de termen waarin N niet voorkomt weg te laten.

2.7 Uitbreiding tot meer dan twee kenmerken

Zowel LUDWIG [12] en MOOD [11] als WISHART en HIRCHFELD [21] behandelen het geval van k kenmerken ($k > 2$). Wij zullen enkele der voornaamste resultaten hieronder laten volgen.

Wij beschouwen nu dus een serie van N elementen waarvan elk een der kenmerken $a_1, \ldots, a_k$ bezit. Wij behandelen hier alleen de verdeling van het totaal aantal runs $\underline{t}$ onder een hypothese analoog aan H_0 in par 1.4, waarbij dus het kenmerk a_i met een kans p_i optreedt ($i = 1, 2, \ldots, k$; $\Sigma p_i = 1$).

Voor wij hierop nader ingaan, willen wij het geval beschouwen dat $p_1 = \ldots = p_k = p = \frac{1}{k}$. Onder H_0 is dan de kans dat $E_n = a_i$ is, gelijk aan p. De kans dat men dus na het optreden van een a_i direct weer een a_i krijgt, is gelijk aan p; de kans dat na a_i nu a_j komt ($i \neq j$) is $1-p$. Heeft men N elementen, dan zijn er dus N-1 van dergelijke overgangen; Is het aantal overgangen van a_i op a_j nu $t-1$ dan zijn er dus t runs in de reeks, en het aantal runs $\underline{t}$ is dus binomiaal verdeeld volgens

$$(2.7;1) \qquad P[\underline{t} = t \mid k, N, H_0] = \binom{N-1}{t-1}\left(\frac{1}{k}\right)^{t-1}\left(1-\frac{1}{k}\right)^{N-t}$$

(verg. 2.6;3).

Gemiddelde en variantie zijn dus die van de binoiale verdeling.

In het geval dat de kansen p_i ongelijk zijn en het element a_i dus met kans p_i kan optreden, is de kans dat er precies t runs optreden en dat het eerste element het kenmerk a_i bezit, gelijk aan

$$(2.7;2)\quad P[\underline{t}=t \text{ en } E_1=a_i \mid k,N,H_0]=\sum_{j=1}^{N} p_i^j\, P[\underline{t}=t-1 \text{ en } E_1=\overline{a_i} \mid k,N-j,H_0]$$

waarin $\overline{a_i}$ een der kenmerken a_j is, $(j \neq i$, en de kans op 0 runs bij 0 elementen gelijk aan 1 wordt verondersteld. Het systeem (2.7;2) is een systeem van k partiele lineaire differentie-vergelijkingen in de P's.

Als men het eerste element niet vastlegt, is de kans op t runs gelijk aan

$$(2.7;3)\quad P[\underline{t}=t \mid k,N,H_0]=\sum_{i=1}^{k} P[\underline{t}=t \text{ en } E_1=a_i \mid k,N,H_0].$$

Het bewijs der formules (2.7;2) en (2.7;3) is vrijwel triviaal en wordt hier niet gegeven.

SCHULTZ [16] geeft voor zeer grote N benaderingsformules voor het gemiddelde en de variantie van de verdeling van t. Deze luiden

$$(2.7;4)\qquad \mathcal{E}(\underline{t} \mid k,N,H_0) \cong N\left(1-\sum_{i=1}^{k} p_i^2\right)$$

en

$$(2.7;5)\quad \sigma^2(\underline{t} \mid k,N,H_0) \cong N\left\{\sum_{i=1}^{k} p_i^2\left(1-\sum_{i=1}^{k} p_i^2\right)+2\sum_{i=1}^{k} p_i^3-2\left(\sum_{i=1}^{k} p_i^2\right)^2\right\}.$$

(verg. (2.6;4) en (2.6;5)).

Voor kleine N kan men dus een exacte toets voor H_0 toepassen door gebruik te maken van de formules (2.7;2) en (2.7;3); voor grote N kan men gebruik maken van de asymptotische normaliteit van de verdeling, waarbij men dus (2.7;4) en (2.7;5) toepast.

2.8 Uitbreiding tot twee dimensies

Wij beschouwen een rechthoekig systeem van vierkantjes. Zij het aantal vierkantjes = m n (m op één zijde van de rechthoek en n op de andere), de kans op een zwart vierkantje = p en op een wit = q. Zij verder $\underline{x}$ het aantal keren dat een zwart vierkantje met tenminste één zijde aan een wit vierkantje grenst. Laat $\underline{y}$ de overeenkomstige grootheid voorstellen indien de 1e rij grenzend aan de me en de 1e kolom grenzend aan de ne gedacht wordt, zodat men een gesloten (toroïdaal) systeem krijgt. P.A.P. MORAN [13] bewijst dat de eerste 3 momenten van $\underline{y}$ zijn

$$(2.8;1) \qquad \mathcal{E}(\underline{y}\,|\,H_0,m,n,p,q) = 4mnpq.$$

$$(2.8;2) \qquad \mathcal{E}(\underline{y}^2\,|\,H_0,m,n,p,q) = 16mnpq + 8mnp^2q^2(2mn-7)$$

$$(2.8;3) \qquad \mathcal{E}(\underline{y}^3\,|\,H_0,m,n,p,q) = 64m^3n^3p^3q^3 + (192p^2q^2 - 672p^3q^3)m^2n^2 + (64pq - 720p^2q^2 + 1856p^3q^3)mn.$$

Berekent men deze momenten t.o.v. het gemiddelde (z.g. gereduceerde momenten) dan vindt men:

$$(2.8;4) \qquad \mathcal{E}(\tilde{\underline{y}}^2\,|\,H_0,m,n,p,q) = 16mnpq - 56mnp^2q^2.$$

$$(2.8;5) \qquad \mathcal{E}(\tilde{\underline{y}}^3\,|\,H_0,m,n,p,q) = mn(64pq - 720p^2q^2 + 1856p^3q^3)$$

waar $\tilde{\underline{y}} = \underline{y} - \mathcal{E}\underline{y}$.

Uit deze formules kan men de overeenkomstige momenten voor $\underline{x}$ afleiden. Deze worden echter vrij ingewikkeld, en wij volstaan daarom met het geven van

$$(2.8;6) \qquad \mathcal{E}(\underline{x}\,|\,H_0,m,n,p,q) = 2pq(2mn - m - n)$$

$$(2.8;7) \qquad \sigma^2(\underline{x}\,|\,H_0,m,n,p,q) = 2pq[8mn - 7m - 7n + 4] + 4p^2q^2[13m + 13n - 14mn - 8].$$

Vervolgens bewijst MORAN dat voor $m,n \to \infty$ $\underline{x}$ asymptotisch normaal verdeeld is met gemiddelde en spreiding gegeven door (2.8;6) en (2.8;7).

MORAN geeft als besluit een generalisatie hiervan tot 3 dimensies.

Opmerking 1: In dit artikel heeft MORAN ten onrechte niet vermeld, dat de formules (2.8;1) t/m (2.8;5) niet gelden voor $n = 1$ en $n = 2$, en dat (2.8;6) en (2.8;7) niet goed zijn voor $n = 1$. Het geval $n = 1$ stemt geheel overeen met het in 2.6 behandelde geval en dus gelden hiervoor de formules (2.6;1) en (2.6;2) in plaats van (2.8;6) en (2.8;7).

Opmerking 2: Voor het geval dat $m = n$ (dus een vierkant met zijde n) zijn door BOSE [3] en LEVENE [9] enige formules afgeleid. De uittreksels in resp. de Mathematical Reviews en de Annals of Math. Statistics (Zie [3] resp [9]) zijn niet correct; de oorspronkelijke artikelen zijn door ons niet geraadpleegd omdat zij te ver buiten het bestek van dit rapport vallen.

3. Verschillende methoden. Hypothese H_1

3.1 De verdeling van t onder H_1

De verwachting van het aantal runs onder de hypothese H_1 (enkelvoudige Markoff-keten, zie par. 1.4) wordt o.m. behandeld door G. SCHULTZ [16]. Deze voert in $q_i^{(N)} = P\{x_N = i\}$, (zie par. 1.4), en verder $q_i = \lim_{N\to\infty} q_i^{(N)}$, Hier is het kenmerk i gelijk aan 0 of 1. De determinant van de matrix der overgangswaarschijnlijkheden (zie par. 1.4) wordt nu (3.1;1)

$$\Delta = p_{00}p_{11} - p_{01}p_{10} = 1 - p_{01} - p_{10} \qquad (-1 < \Delta < +1) \tag{3.1;2}$$

hetgeen volgt uit de in 1.4 gegeven betrekkingen tussen deze grootheden.

Tevens kan worden afgeleid

$$p_{01} = (1-\Delta)q_0 \quad \text{en} \quad p_{10} = (1-\Delta)q_1 . \tag{3.1;3}$$

SCHULTZ voert verder in de grootheid

$$(3.1;4) \qquad \alpha \overset{def}{=} 2(1-\Delta)q_0q_1 ,$$

en bewijst dan, dat asymptotisch (voor $N\to\infty$) gemiddelde en variantie van de verdeling van het totaal aantal runs $\underline{t}$ gegeven worden door:

$$(3.1;5) \qquad E(\underline{t} \mid H_1, N) \cong N\alpha$$

en

$$(3.1;6) \qquad \sigma^2(\underline{t} \mid H_1, N) \cong N[\alpha(1-\alpha)+\alpha(1-4q_0q_1)] .$$

SCHULTZ bewijst tevens dat de verdeling van $\underline{t}$ onder H_1 asymptotisch normaal is. Indien men dus de hypothese H_1 wenst te toetsen met $\underline{t}$ als toetsingsgrootheid, kan men gebruik maken van (3.1;5) en (3.1;6).

Opmerking 1: Het geval van de hypothese H_0 (zie par. 1.4) is een bijzonder geval van een enkelvoudige Markoff-keten, n.l. met $p_{01} = p_{11} = p$ en $p_{00} = p_{10} = q$, en dus tevens $q_1^{(N)} = q_1 = p$ en $q_0^{(N)} = q_0 = q$.

Opmerking 2: Voor $P\{\underline{t} = t / H_1, N\}$ zijn door DAVID [4] enige formules gegeven. Aangezien deze vrij ingewikkeld zijn en van weinig praktische betekenis, zullen wij deze hier niet vermelden.

Opmerking 3: Voor de overeenkomstige grootheden onder de hypothese H_2, die inhoudt dat de waarnemingen een dubbele Markoff-keten vormen (d.w.z. dat elke waarneming slechts afhankelijk is van de twee direct daaraan voorafgaande waarnemingen) zijn te weinig gegevens uit de literatuur beschikbaar, zodat wij ons beperken tot de behandeling van H_1. BATEMAN [1] geeft formules voor $P\{\underline{t} = t / H_2, N\}$; wij vermelden deze echter niet.

3.2 Uitbreiding tot meer dan twee kenmerken

G. SCHULTZ [16] berekent de asymptotische verdeling van het totaal aantal runs, voor het meer algemene geval van k

Literatuur

[1] G. BATEMAN, On the power function of the longest run as a test for randomness in a sequence of alternatives, Biometrika 35 (1948) 97-112.

[2] L. von BORTKIEWICZ, Die Iterationen, Berlin 1917.

[3] R.C. BOSE, The patch number problem, Science and Culture 12 (1946) 199; vgl. ook Math. Reviews 8 (1946) 389.

[4] F.N. DAVID, A power function for tests of randomness in a sequence of alternatives, Biometrika 34 (1947) 335-339.

[5] W. FELLER, Probability theory and its applications I, New York 1950.

[6] W. GONTCHAROFF, Sur la succession des évenements dans une série d'épreuves indépendances répondant au schême de Bernoulli, C R Acad Sci URSS (NS) 38 (1943) 283-285.

[7] W. GONTCHAROFF, Du domaine d l'analyse combinatoire, Bull. Acad. Sci. URSS, Sér. Math. 8 (1944) 3.

[8] E. ISING, Beitrag zur Theorie des Ferromagnetismus, Zeitschr. für Physik 31 (1925) 253-258.

[9] H. LEVENE, A test of randomness in two dimensions, Abstr. Bull. Amer. Math. Soc. 52 (1946) 621 en Ann. Math. Stats, 17 (1946) 500.

[10] O. LUDWIG, Ueber die stochastische Theorie der Merkmalsiterationen, MitteilungsBlatt für Mathem. Statistik, 8 (1956) 49-82.

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

[12] P.G. MOORE, A test for randomness in a sequence of two alternatives involving a 2 x 2 table, Biometrika 36 (1949) 305-316.

[13] P.A.P. MORAN, Random associations on a lattice, Proc. Cambr. Philos. Soc. 43 (1947) 321-328.

[14] E. MORICE, Quelques tests non paramétriques, Revue de Stat. Appliq. 4 (1956) no. 4, p. 75-107.

[15] M. OLEKIEWICZ, Tables of expected values and variances of numbers of runs in random sequences with probabilities of exceeding expected values, Am. Univ. Marie Curie-Sklodowska, Vol. 5 (1951) 147-159.

kenmerken. Hiertoe stelt hij, analoog aan (3.1;1),

$$q_\iota^{(N)} = P\{\underset{N}{\leq} = \iota\} \text{ en } q_\iota = \lim_{N\to\infty} q_\iota^{(N)}.$$

Hier is $\iota = 1,2,\ldots,k$. De matrix der overgangswaarschijnlijkheden geven wij aan met $\{p_{\iota\kappa}\} = P$; de N^e macht hiervan met $P^N = \{p_{\iota\kappa}^{(N)}\}$. De matrix P moet regulier zijn in de zin van Frechet, d.w.z. dat de elementen van P^N voor voldoende grote N alle positief zijn. Verder is $\lim_{N\to\infty} P^N = P^\infty$. Van deze P^∞ kunnen de elementen eenvoudig berekend worden uit $PP^\infty = P^\infty$ of $(P-E)P^\infty = 0$, waarin $E = \{e_{\iota\kappa}\}$ de eenheidsmatrix voorstelt. Is P regulier, dan bevat de j^e rij van P^∞ dezelfde elementen, als gelijk aan q_j. Hiermee zijn dan de q_ι bepaald.

Tenslotte definieert SCHULTZ nog een matrix $\{f_{\iota\kappa}\} = F$, waarvan de elementen berekend kunnen worden uit

$$(3.2;1) \qquad (E - P)F = E - P^\infty,$$

en de grootheid

$$(3.2;2) \qquad a \overset{\text{def}}{=} \sum_{\kappa=1}^{k} p_{\kappa\kappa} q_\kappa .$$

In het artikel wordt nu bewezen, dat het gemiddelde en de variantie van het totaal aantal runs asymptotisch (van $N\to\infty$) gegeven worden door

$$(3.2;3) \qquad \mathcal{E}\{\underline{t} \mid H_1, k, N\} \cong N(1-a)$$

en

$$(3.2;4) \qquad \sigma^2(\underline{t} \mid H_1, k, N) \cong N\left[a(1-a) + 2\sum_{\kappa=1}^{k}\sum_{\rho=1}^{k} p_{\kappa\kappa} f_{\kappa\rho} p_{\rho\rho} q_\rho\right]$$

en dat de verdeling van $\underline{t}$ asymptotisch normaal is.

[16] G. SCHULTZ, Ueber die Häufigkeit der Iterationen in einer Beobachtungsfolge, Deutsche Math. 7 (1942) 22-38.

[17] W.L. STEVENS, Distribution of groups in a sequence of alternatives, Annals of Eugenics 9 (1939) 10-17.

[18] F.C. SWED and C. EISENHART, Tables for testing randomness of grouping in a sequence of alternatives, Ann. Math. Stats. 14 (1943) 66-87.

[19] A. WALD and J. WOLFOWITZ, On a test whether two samples are from the same population, Ann. Math. Stats. 11 (1940) 147-162.

[20] S.S. WILKS, Mathematical Statistics, Princeton 1943

[21] J. WISHART and H.O. HIRSCHFELD, A theorem concerning the distribution of joins between line segments, Journ. London Math. Soc. 11 (1936) 227-235.