Een Monte-Carlo-bepaling van overschrijdingskansen, in verband met een keuzetest

door C. A. G. Nass *)

UDC 519 : 21

Summary

A Monte-Carlo method for a test of significance, applied to points on a lattice, in connection with a vocational preference test, by C. A. G. Nass.
Appendix by Constance van Eeden.

A periodical rectangular lattice, with a period of k.m, is considered. Thus there are $N = k.m$ points on the lattice, repeated in the two perpendicular directions. Two points are said to be ,,connected" if they are adjacent in a straight or diagonal way. Thus, if k and $m \geqq 3$, every point is connected with 8 other points. Out of the N points of the lattice, n points are selected and the total number of connections, x, of all possible pairs of those n points is considered for a vocational preference test with $k = m = 9$, $N = 81$, $n = 10$. The problem is to test whether the sum $y = x_1 + \ldots + x_h$, from a sample of h values of x, is significantly small, under the hypothesis that in the h cases the n points are selected at random with equal chance. A Monte-Carlo sample of 100 values of $\underline{x}$ was taken, using random numbers. For $h = 1$, the problem was solved by the determination of $P(\underline{y} \leqq x_1)$, assuming that y is taken at random from the 101 values of x, supplied by the Monte-Carlo sample and x_1 for fixed values of x_1. For $h = 2$, a similar solution is given. For greater values of h, Student's two-sample test, with correction for continuity is suggested. For $h = 2$ the results of Student's test are compared with those of the solution mentioned above.

In the appendix a summary is given of results found by P. A. P. Moran and P. V. Krishna Iyer for some closely related problems. Further some results concerning exact distributions, moments and asymptotic distributions for C. A. G. Nass' problem are given. The proofs of these results may be found in a paper by C. van Eeden and A. R. Bloemena (1959).

1. De Monte-Carlo-methode

De ,,Monte-Carlo-methode" is een verzamelnaam voor het gebruik van aselecte getallen (random numbers) voor statistische toetsen en andere statistische procedures. Het gebruik van zulke getallen voor de schatting van parameters van een natuurlijke populatie ressorteert onder de steekproeftechniek

*) Inst. voor Praeventieve Geneeskunde, Leiden.

en wordt meestal niet tot de Monte-Carlo-methode gerekend. De naam berust op de overeenkomst van een tabel met aselecte getallen en een reeks uitkomsten van het apparaat waar de betreffende stad om bekend is.

Een statistisch model is in wezen een beschrijving van een kansapparaat, waarmee men volgens een bepaald voorschrift in principe onbeperkt vele uitkomsten kan verkrijgen. ,,In principe" betekent hier ,,als men over een onbeperkt quantum tijd en geld zou beschikken." ,,In principe" zou men, zonder enige kennis van waarschijnlijkheidsrekening en met iedere gewenste graad van nauwkeurigheid behalve de absolute, ieder statistisch probleem op kunnen lossen, alleen door met het kansapparaat het in het probleem aanwezige voorschrift nauwgezet op te volgen. Meestal kan men zich een grote hoeveelheid tijd en geld besparen en bovendien een absoluut nauwkeurig antwoord vinden, door wèl waarschijnlijkheidsrekening toe te passen en het kansapparaat geheel met rust te laten. Dit is zelfs zo vaak het geval dat men zou vergeten dat het kansapparaat wel eens uitkomst kan bieden waar de waarschijnlijkheidsrekening te kort schiet. Het is daarom leerzaam om af en toe eens een uitzonderingsgeval te bestuderen waarin een deel van de functie van de waarschijnlijkheidrekening door het kansapparaat moet worden overgenomen.

2. De keuzetest

Voor psychotechnische doeleinden wil men bij scholieren voor ieder van een groot aantal bezigheden een zekere voorkeursgraad bepalen. Men heeft een lijst van 81 bezigheden, zoals: landkaarttekenen, elektrische schel aanleggen, enz. Men wil bepalen welke bezigheden de scholier hiervan het prettigst vindt. Het laten kiezen uit een zo groot aantal mogelijkheden tegelijk is bezwaarlijk. Het is beter om de bezigheden in groepjes van 4 aan te bieden en daaruit telkens de aantrekkelijkste te laten kiezen. Daartoe heeft men de collectie van 81 bezigheden aselect gerangschikt in een vierkant van 9 bij 9, dat dus door kruisende verticale en horizontale lijnen verdeeld is in 81 vakjes. Bij ieder van de 49 inwendig gelegen kruispunten kan men nu een keuze laten doen uit de 4 werkzaamheden, die in dat punt contact maken. De 28 marginaal gelegen werkzaamheden verkeren dan echter in het nadeel tegenover hun meer centraal gelegen concurrenten, omdat zij maximaal slechts twee maal verkozen kunnen worden in plaats van viermaal. En de vier bezigheden bij de hoekpunten van het grote vierkant, kunnen zelfs niet meer dan eenmaal verkozen worden. Om dit bezwaar te ondervangen beschouwt men het grote vierkant als een juist compleet monster van een behangselpatroon, dat zich in beide richtingen onbeperkt voortzet. Men herhaalt dus de onderste rij van het vierkant aan de bovenkant, het links-onderste vakje in een rechts-boven gelegen vakje, enz. Er zijn nu 81 gelijkwaardige kruispunten (keuzen) en iedere werkzaamheid kan

0, 1, 2, 3 of 4 punten behalen. Met dit patroon werd een groot aantal leerlingen van een bepaald schooljaar en een bepaald schooltype getest en de 10 werkzaamheden die de meeste punten hadden behaald, werden genoteerd. Dit lijstje is min of meer typerend voor het schooljaar en het schooltype. Een soortgelijk lijstje werd verkregen van een ander schooljaar en een ander schooltype, met 81 andere bezigheden en een andere aselecte rangschikking, onafhankelijk van de vorige.

Er bleef enige twijfel bestaan of de kans van iedere bezigheid om tot de 10 uitverkorene te behoren bij een gegeven rangschikking wel uitsluitend van zijn eigen merites afhangt. Men kan vermoeden dat de bezigheden, die in contact komen met een bijzonder favoriete concurrent, daardoor in een ongunstige positie verkeren. Een aanwijzing hiervoor meende men te zien in de omstandigheid dat in beide schooltypen de vakjes van de tien uitverkoren bezigheden weinig contacten vertonen via een kruispunt. Hoe kan men statistisch toetsen of de beide aantallen contacten zo exceptioneel laag zijn, dat men kan besluiten tot een reëel bestaande neiging tot weinig contacten tussen uitverkoren werkzaamheden?

3. Het statistisch probleem

Trekt men aselect, met gelijke kansen en zonder terugleggen 10 vakjes uit de 81 vakjes van het patroon, dan kan het aantal contacten tussen die 10 vakjes variëren van 0 tot 23. Er zijn meer dan twee biljoen verschillende keuzen van 10 uit 81, die alle dezelfde kans hebben. Heeft men geen geschikt electronisch apparaat, dan is het onmogelijk om te tellen hoeveel daarvan respectievelijk 0,1,...,23 contacten hebben. Als men niet over het vernuft beschikt om een wiskundige methode te vinden – als die bestaat – om dit astronomische telwerk voldoende te bekorten, dan is de Monte-Carlo-methode de enig toegankelijke. Men kan de kansverdeling van het aantal contacten bij benadering vinden door bijvoorbeeld 100 aselecte keuzen van 10 uit 81 te nemen en bij ieder daarvan het aantal contacten te tellen.

Stel nu dat we beschikken over 100 getallen $(x) = (x_1,...x_{100})$, zijnde de gevonden aantallen contacten in de steekproef en daarnaast over een getal a, zijnde het gevonden aantal contacten van het tiental bezigheden, gekozen door een zeker schooltype. Voor het getal a nemen we aan dat er slechts twee mogelijkheden bestaan:

1. Het is getrokken uit dezelfde kansverdeling waarvan ook de 100 getallen (x) afkomstig zijn (nulhypothese).
2. Het is getrokken uit een andere kansverdeling, met een *kleiner* gemiddelde. (tegenhypothese).

We zouden dan de gevonden frequentieverdeling van x gelijk kunnen stellen aan de kansverdeling en de *linkse* overschrijdingskans nemen: $P(\underline{x} \leqq a)$.

De gelijkstelling van frequentieverdeling en kansverdeling kan vermeden worden door de volgende hypothesen, die gelijkwaardig zijn met de vorige:

1. Het getal a is uit de 101 getallen $(y) = (x)$ Ua aselect getrokken, met gelijke kansen (nulhypothese).
2. Het getal a is getrokken uit dezelfde 101 getallen, met kansen die groter zijn naarmate de getallen kleiner zijn (tegenhypothese).

We nemen dan de linkse overschrijdingskans $P(\underline{y} \leqq a)$.

Hebben we in plaats van één getal a, een aantal van zulke getallen, $(a) = (a_1,...,a_h)$, die als het resultaat van stochastisch onafhankelijke trekkingen uit dezelfde kansverdeling beschouwd kunnen worden, dan neemt men bij analoge alternatieve hypothesen de linkse overschrijdingskans,

$$P(\underline{y}_1 + ... + \underline{y}_h \leqq a_1 + ... + a_h),$$

waarbij dus de h getallen $(\underline{y})$ aselecte trekkingen met gelijke kansen en zonder terugleggen uit $100 + h$ gegeven getallen (y) voorstellen.

4. De Monte-Carlo-steekproef

Om de genoemde methoden toe te passen, moet men eerst beschikken over de 100 getallen (x).
Deze kan men in twee stappen verkrijgen:

1. Bepaal, met behulp van een tabel van tweecijferige aselecte getallen, 100 keuzen van 10 uit 81 zonder terugleggen.
2. Bepaal voor ieder van deze keuzen het aantal contacten x.

Voert men stap 1 zo uit, dat iedere keuze bestaat uit 10 ongelijke getallen van 1 tot en met 81, dan wordt stap 2 zeer tijdrovend. Daarom kan men de 81 werkzaamheden beter niet doornummeren, maar hun positie in het grote vierkant aangeven door twee coördinaten, lopende van 0 tot en met 8. Van ieder getallenpaar uit een gegeven tiental, kan dan gemakkelijk bepaald worden of zij contact maken. Dit is het geval als zowel de eerste als de tweede coördinaat hoogstens één punt verschilt, waarbij ook 0 en 8 geacht worden één punt te verschillen.

Stap 1 bestaat dus uit het noteren van telkens 10 tweecijferige getallen, met overslaan van alle getallen die een 9 bevatten en van de getallen die reeds voor hetzelfde tiental genoteerd zijn.

Stap 2 bestaat uit het zetten van boogjes tussen ieder getallenpaar van elk tiental, waarvan beide coördinaten hoogstens een punt verschillen. Het aantal boogjes per tiental is het aantal contacten x. Als voorbeeld volgen hier een drietal keuzen met x-waarden.

Keuze										x
18	70	86	62	66	26	64	31	44	46	0
22	50	33	68	41	00	42	06	61	20	6
07	38	40	28	17	18	57	15	30	25	11

Uit 100 keuzen werd de volgende frequentieverdeling van x gevonden:

x	0	1	2	3	4	5	6	7	8	9	10	11
Aantal	1	4	8	18	23	19	12	10	3	1	0	1
Cumulatief	1	5	13	31	54	73	85	95	98	99	99	100

5. Exacte statistische toetsen

Stel dat men één getal a heeft en wil bepalen of dit significant laag is. Is men bereid het verschil tussen kansverdeling en frequentieverdeling van x te verwaarlozen, dan kan men zonder meer de cumulatieve frequenties van x als overschrijdingskansen gebruiken. Dus $a = 0$ zou significant zijn bij 1%, $a = 1$ significant bij 5% en $a \geqq 2$ niet significant bij 5%.

Wil men die verwaarlozing vermijden, dan stelt men:

$$P(\underline{y} \leqq a) = \frac{1}{101} + \frac{100}{101} P(\underline{x} \leqq a)$$

waarbij dan $P(\underline{x} \leqq a)$ de cumulatieve frequentie voorstelt van $x = a$.

Bij de verschillende waarden van a vindt men dus de volgende overschrijdingskansen:

a	0	1	2	3	4
$P(\underline{y} \leqq a)$	0,020	0,059	0,139	0,317	0,545

Met deze exacte methode blijkt dat significantie bij 1% niet mogelijk is en dat $a = 1$ reeds niet meer significant is bij 5%.

De exacte methode is zonder electronische apparatuur ook nog uitvoerbaar voor twee getallen a_1 en a_2. Daarvoor is nodig de kansverdeling van de som $x_1 + x_2$ van twee getallen zonder terugleggen getrokken uit de 100 getallen (x). Deze kan men vinden door de frequentieverdeling van x te quadrateren, waarbij men in de diagnonaal telkens invult $f(x)$ $(f(x) - 1)$ in plaats van $f^2(x)$:

x_2 \ x_1	0	1	2	3	4	5	6	7	8	9	10	11	Totaal
0	0	4	8	18	23	19	12	10	3	1	0	1	
1	4	12	32	72	92	76	48	40	12	4	0	4	
2	8	32	56	144	184	152	96	80	24	8	0	8	
3	18	72	144	306	414	342	216	180	54	18	0	18	
4	23	92	184	414	506	437	276	230	69	23	0	23	
5	19	76	152	342	437	342	228	190	57	19	0	19	
6	12	48	96	216	276	228	132	120	36	12	0	12	
7	10	40	80	180	230	190	120	90	30	10	0	10	
8	3	12	24	54	69	57	36	30	6	3	0	3	
9	1	4	8	18	23	19	12	10	3	0	0	1	
10	0	0	0	0	0	0	0	0	0	0	0	0	
11	1	4	8	18	23	19	12	10	3	1	0	0	
Totaal													9900

Door diagonaalsgewijs optellen vindt men de kansverdeling van $x_1 + x_2$:

$x_1 + x_2$	1	2	3	4	5	6	7
$P(\underline{x}_1 + \underline{x}_2 \leqq x_1 + x_2)$	0,0008	0036	0137	0386	0901	1760	3020 .....

De exacte overschrijdingskans vindt men als volgt:

Uit 102 getallen y, bestaande uit 100 getallen x en 2 getallen a, worden aselect 2 getallen getrokken. De kans dat dit de getallen a zijn is $2/(102 \times 101)$. De kans op a_1 en een van de getallen x is $200/(102 \times 101)$. De kans op a_2 en een van de getallen x is even groot. De kans op twee van de getallen x is $100 \times 99/(102 \times 101)$. In het eerste geval is de som van de twee getrokken getallen al vanzelf gelijk aan $a_1 + a_2$, dus ook hoogstens gelijk aan $a_1 + a_2$. In het tweede geval mag de getrokken x hoogstens gelijk zijn aan a_2 en in het derde geval hoogstens gelijk aan a_1. In het vierde geval mag de som van de twee getrokken getallen x hoogstens gelijk zijn aan $a_1 + a_2$.

Wij vinden dus:

$$P(\underline{y}_1 + \underline{y}_2 \leqq a_1 + a_2) = \frac{2}{101 \times 102} + \frac{200}{101 \times 102}(P(\underline{x} \leqq a_1) + P(\underline{x} \leqq a_2)) +$$
$$+ \frac{99 \times 100}{101 \times 102} P(\underline{x}_1 + \underline{x}_2 \leqq a_1 + a_2) =$$
$$= 0{,}0002 + 0{,}0194\,(P(\underline{x} \leqq a_1) + P(\underline{x} \leqq a_2)) +$$
$$+ 0{,}9610\, P(\underline{x}_1 + \underline{x}_2 \leqq a_1 + a_2).$$

Hier volgen enige aldus berekende exacte overschrijdingskansen voor verschillende waarden van a_1 en a_2, met daaronder tussen haakjes de benaderde

overschrijdingskansen volgens de Student-test op het verschil van twee gemiddelden, met correctie voor continuïteit:

a_1	0	1	2	3	4	5	6
a_2							
0	0,001	002	006	020	048	101	186
	(0,001)	(003)	(009)	(023)	(053)	(107)	(191)
1		006	017	044	098	184	
		(009)	(023)	(052)	(104)	(187)	
2			042	095	182	$P(\underline{y}_1 + \underline{y}_2 \leqq a_1 + a_2)$	
			(051)	(103)	(185)		
3				181		Exact	
				(184)		(Benaderd).	

Volgens de exacte overschrijdingskansen mogen er voor significantie hoogstens 2 contacten zijn bij een 1%-drempel en hoogstens 4 contacten bij een 5%-drempel.

6. **Benaderende statistische toetsen**

Heeft men meer dan 2 getallen a, dan is men practisch gedwongen om tot een benadering over te gaan. Uit de Monte-Carlo-steekproef vindt men:

$$\bar{x} = 4{,}47 \text{ en } \Sigma(x - \bar{x})^2 = 358{,}9.$$

Uit de h getallen (a) berekent men $\bar{a}$ en $\Sigma(a - \bar{a})^2$. Zij $\underline{\bar{a}}$ het gemiddelde van h getallen $(\underline{a})$ met gelijke kansen en zonder terugleggen getrokken uit de $100 + h$ getallen (y) en $\underline{\bar{x}}$ het gemiddelde van de overige getallen $(\underline{x})$, dan volgt de volgens Student berekende $\underline{t}$ met een zekere benadering een standaardnormale verdeling:

$$\underline{t} = \left(\underline{\bar{a}} - \underline{\bar{x}} + \frac{1}{2h} + \frac{1}{200}\right) \sqrt{\frac{100h}{\Sigma(\underline{x} - \underline{\bar{x}})^2 + \Sigma(\underline{a} - \underline{\bar{a}})^2} \frac{98 + h}{100 + h}}.$$

Omdat men als nulhypothese stelt dat de h getallen (a) als het resultaat van een trekking uit de getallen (y) zijn te beschouwen, is de volgende t dan ook te beschouwen als een trekking uit de kansverdeling van $\underline{t}$:

$$t = \left(\bar{a} - 4{,}465 + \frac{1}{2h}\right) \sqrt{\frac{100h}{358{,}9 + \Sigma(a - \bar{a})^2} \frac{98 + h}{100 + h}}.$$

Het teken van de continuïteitscorrectie $+ 1/2h + 0{,}005$ moet altijd positief zijn, omdat men met de linkse overschrijdingskans werkt.

Omdat de frequentieverdeling van (x) behoorlijk gecentreerd is, kan men verwachten, dat reeds bij $h = 3$ een goede benadering verkregen wordt. Daar staat tegenover dat de frequentieverdeling van (x) wat scheef is, hetgeen bij een eenzijdige toets afbreuk doet aan de qualiteit van de benadering. Bovendien wordt deze qualiteit o.a. in het geval van de Student-toets door vele statistici onderschat. Daarom werden bij $h = 2$ naast de exacte overschrijdingskansen ook de benaderde volgens Student berekend. Het blijkt dat de overeenstemming reeds tamelijk bevredigend is, zodat men er wel op kan vertrouwen dat bij $h \geqq 3$ de benadering bruikbaar zal zijn.

Het verdient de aandacht dat bij de bovenstaande toepassing van de toets van Student geen normaliteitsonderstelling werd gemaakt. In plaats daarvan werd de voorwaardelijke verdeling van $\bar{a}$ beschouwd, onder de voorwaarde dat de getallen y de bij het experiment gevonden waarden hebben aangenomen. Deze verdeling is, onder zeer algemene voorwaarden asymptotisch normaal. Deze toepassingswijze van de toets van Student is voor het eerst gepubliceerd door R. A. Fisher in 1935, namelijk bij zijn analyse van een proef van Charles Darwin, die daarbij reeds alle thans nodig geachte statistische voorzorgen in acht bleek te hebben genomen, behalve de verloting.

Referentie:

Fisher, R. A. 1935: The design of experiments.

APPENDIX *)

door Constance van Eeden

Het door C.A.G. Nass in het bovenstaande artikel behandelde probleem kan als volgt algemeen geformuleerd worden.

Gegeven zijn $N = k.\ m$ vakjes gerangschikt in een rechthoek van k bij m $(k \leqq m)$. Uit deze N vakjes worden aselect en zonder teruglegging n vakjes gekozen. De gekozen vakjes worden in het volgende aangeduid als zwarte vakjes, de niet-gekozen vakjes als witte vakjes.

Het gaat nu om het aantal contacten tussen de zwarte vakjes. Hierbij worden twee vakjes beschouwd contact te hebben als zij met een zijde of een hoekpunt aan elkaar raken.

Nu zijn in een rechthoek in het algemeen niet alle vakjes gelijkwaardig. Voor $k = 1$ en $m > 2$ b.v. heeft een vakje, dat aan één der uiteinden ligt, slechts contact met één ander vakje, terwijl de overige vakjes contact hebben met twee andere vakjes. Voor $k = 1$ en $m > 2$ kunnen de vakjes dus gelijkwaardig gemaakt worden door de twee korte zijden van de rechthoek tegen

*) Rapport S 271 van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam.

elkaar te leggen; ieder vakje heeft dan contact met precies twee andere vakjes. Voor $k = 2$ en $m > 2$ kunnen de vakjes eveneens gelijkwaardig gemaakt worden door de twee korte zijden van de rechthoek tegen elkaar te leggen. Ieder vakje heeft dan contact met precies 5 andere vakjes.

Voor $k \geqq 3$ moeten, om de vakjes gelijkwaardig te maken, ook de twee lange zijden van de rechthoek tegen elkaar gelegd worden. Dan ontstaat een torus (ring) en ieder vakje op de torus heeft contact met precies 8 andere vakjes.

N a s s beschouwt nu de grootheid $\underline{x}$: het aantal paren zwarte vakjes, dat contact heeft.

Voor $k = 1$ is dit probleem volledig opgelost, d.w.z. de exacte verdeling van $\underline{x}$ is bekend voor iedere m $(= N)$ en n. De grootheid $\underline{x}$ bezit dan nl. voor $0 < n < N$ een hypergeometrische verdeling

$$P[\underline{x} = x | k = 1] = \frac{\binom{n}{x}\binom{N-n-1}{n-x-1}}{\binom{N-1}{n-1}} \qquad \max(0, 2n-N) \leqq x \leqq n-1. \qquad (1)$$

Deze verdeling kan b.v. gevonden worden uit de verdeling van het aantal runs van zwarte vakjes. Is $\underline{r}$ dit aantal runs dan is (zie b.v. W. L. S t e v e n s (1939) en H. A. K u i p e r s (1957))

$$P[\underline{r} = r] = \frac{\binom{n}{r}\binom{N-n-1}{r-1}}{\binom{N-1}{n-1}} \text{ [1]}. \qquad (2)$$

De verdeling van $\underline{x}$ volgt dan uit (2) en $\underline{r} + \underline{x} = n$.

Uit (1) volgt voor de asymptotische eigenschappen van de verdeling van $\underline{x}$ voor n en N beide $\to \infty$ (zie b.v. C. v a n E e d e n (1959)):

a. $\underline{x}$ bezit asymptotisch een Poisson-verdeling als $\lim_{N\to\infty} \frac{n}{N} = 0$ en $\lim_{N\to\infty} \frac{n^2}{N}$ bestaat,

[1]) De verdeling van $\underline{r}$ kan b.v. ook worden afgeleid uit de verdeling van het aantal runs in een rij vakjes ter lengte N', waarbij de uiteinden niet tegen elkaar gelegd worden. Is r' dit aantal runs dan is (zie b.v. W. F e l l e r (1950), p. 59, opgave 15)

$$P[\underline{r}' = r] = \frac{\binom{n-1}{r-1}\binom{N'-n+1}{r}}{\binom{N'}{n}}.$$

De verdeling van $\underline{r}$ wordt hieruit gevonden door N' te vervangen door $N-1$. (Zie hiervoor ook H. A. K u i p e r s (1957), p.5).

b. $\underline{x} + N - 2n$ (het aantal paren witte vakjes dat contact heeft) bezit asymptotisch een Poisson-verdeling als $\lim_{N\to\infty} \frac{n}{N} = 1$ en $\lim_{N\to\infty} \frac{(N-n)^2}{N}$ bestaat,

c. $\frac{\underline{x} - \mathscr{E}\underline{x}}{\sigma(\underline{x})}$ bezit asymptotisch een normale verdeling met gemiddelde 0 en spreiding 1 als $0 < \lim_{N\to\infty} \frac{n}{N} < 1$.

Door P. V. Krishna Iyer (1949 en 1950) worden een aantal met het bovenstaande zeer nauw verwante problemen behandeld. In beide artikelen beschouwt hij een rechthoek van k bij m, waarbij echter de zijden niet tegen elkaar gelegd worden. De vakjes zijn dan dus niet gelijkwaardig, wat echter op de asymptotische verdeling voor k en m *beide* $\to\infty$ geen invloed heeft. Wat betreft de keuze van de vakjes beschouwt hij twee gevallen:

1. het geval waarbij voor ieder vakje met een kans p geloot wordt of dat vakje gekozen zal worden. Het aantal gekozen vakjes is dan een stochastische grootheid;
2. het door Nass beschouwde geval, waarbij een vast aantal vakjes n aselect en zonder teruglegging gekozen wordt.

Wat betreft de contacten tussen de vakjes beschouwt hij ook twee gevallen nl.

a. in zijn artikel van 1949 het ook door Nass beschouwde geval, waarbij twee vakjes contact hebben als zij met een zijde of een hoekpunt aan elkaar raken;

b. in zijn artikel van 1950 het geval, waarbij twee vakjes contact hebben als zij met een zijde aan elkaar raken. [1])

Krishna Iyer geeft voor alle door hem beschouwde gevallen het eerste en tweede moment van de verdeling; in sommige gevallen geeft hij ook het derde en vierde moment. Verder vermeldt hij, zonder strikt bewijs, dat alle verdelingen asymptotisch normaal zijn als k en m beide $\to\infty$ gaan. Hierbij onderstelt hij echter, zonder dit duidelijk te vermelden, dat $0 < \lim_{N\to\infty} \frac{n}{N} < 1$. Daar het al of niet gelijkwaardig zijn van alle vakjes geen invloed heeft op de asymptotische verdeling zou hieruit de asymptotische normaliteit van de door Nass beschouwde grootheid $\underline{x}$ volgen onder de voorwaarden k en m beide $\to\infty$ en $0 < \lim_{N\to\infty} \frac{n}{N} < 1$.

P. A. P. Moran (1948) beschouwt het geval van een willekeurig rooster.

[1]) Krishna Iyer geeft in beide artikelen ook generalisaties voor meer dan twee dimensies en meer dan twee kleuren.

Wat betreft de keuze van de vakjes beschouwt hij, evenals K r i s h n a I y e r zowel het geval van een vast aantal gekozen vakjes, als van een stochastisch aantal, waarbij voor ieder vakje met een kans p geloot wordt of dit vakje gekozen zal worden.

Voor beide gevallen geeft hij het eerste en tweede moment van de verdeling; voor het geval van stochastische $\underline{n}$ ook het derde en vierde moment. Verder bewijst hij de asymptotische normaliteit voor een rechthoekig rooster en een stochastische $\underline{n}$. (Zie hiervoor ook P. A. P. M o r a n (1947)).

De bovengenoemde artikelen vormen niet de enige literatuur betreffende dit onderwerp. K r i s h n a I y e r publiceerde nog een aantal artikelen in de Journal of the Indian Society of Agricultural Statistics. Verder maakte Prof. Dr. D. v a n D a n t z i g mij erop attent dat het gestelde probleem samenhangt met het „orde-wanorde" probleem uit de physica (Zie ook P. A. P. M o r a n (1947)). Ook in de literatuur over dit probleem zijn dus wellicht nog resultaten te vinden die van belang zijn voor het door N a s s gestelde probleem.

Door ons werden, voor de door N a s s beschouwde grootheid $\underline{x}$, de volgende resultaten gevonden (de bewijzen hiervan zijn te vinden in het rapport „On probability distributions arising from points on a lattice" door C. v a n E e d e n en A. R. B l o e m e n a (1959)).

1) de exacte verdeling van $\underline{x}$ voor

$$k = 2 \text{ met } n = 2, 3 \text{ en } 4,$$
$$k \geqq 3 \text{ met } n = 2 \text{ en } 3.$$

2) een algemene uitdrukking voor $\mathscr{E}\underline{x}$ en $\sigma^2(\underline{x})$

$$\mathscr{E}\underline{x} = \begin{cases} \dfrac{5n\,(n-1)}{2\,(N-1)} & \text{voor } k = 2 \text{ en } m > 2, \\ \dfrac{4n\,(n-1)}{N-1} & \text{voor } k \geqq 3, \end{cases}$$

$$\sigma^2(\underline{x}) = \begin{cases} \dfrac{5n\,(n-1)\,(N-6)\,(N-n)\,(N-n-1)}{2\,(N-1)^2\,(N-2)\,(N-3)} & \text{voor } k = 2 \text{ en } m > 2, \\ \dfrac{4n\,(n-1)\,(N-9)\,(N-n)\,(N-n-1)}{(N-1)^2\,(N-2)\,(N-3)} & \text{voor } k \geqq 3, \end{cases}$$

3) een algemene uitdrukking voor $P[\underline{x} = 0]$ en $P[\underline{x} = 1]$ zowel voor $k = 2$ als voor $k = 3$

$$P[\underline{x}=0]=\begin{cases}2^{n-1}\dfrac{\binom{\frac{N}{2}-n-1}{n-1}}{\binom{N-1}{n-1}} & \text{voor } k=2 \text{ en } m>2,\\[2ex] 3^{n-1}\dfrac{\binom{\frac{N}{3}-n-1}{n-1}}{\binom{N-1}{n-1}} & \text{voor } k=3,\end{cases}$$

$$P[\underline{x}=1]=\begin{cases}\dfrac{\binom{n}{2}2^{n-2}}{(N-1)\binom{N-2}{n-2}}\left[\binom{\frac{N}{2}-n}{n-2}+4\binom{\frac{N}{2}-n-1}{n-2}\right] & \text{voor } k=2 \text{ en } m>2,\\[2ex] \dfrac{\binom{n}{2}3^{n-2}}{(N-1)\binom{N-2}{n-2}}\left[2\binom{\frac{N}{3}-n}{n-2}+6\binom{\frac{N}{3}-n-1}{n-2}\right] & \text{voor } k=3.\end{cases}$$

Verder kan bewezen worden dat de grootheid $\underline{x}$ asymptotisch, voor n en N beide $\to\infty$, als $\lim_{N\to\infty}\frac{n}{N}=0$ en $\lim_{N\to\infty}\mathscr{E}\underline{x}$ bestaat, een Poissonverdeling bezit met parameter $\lim_{N\to\infty}\mathscr{E}\underline{x}$, terwijl de grootheid $\underline{x}'$ (het aantal paren witte vakjes dat contact heeft) als $\lim_{N\to\infty}\frac{n}{N}=1$ en $\lim_{N\to\infty}\mathscr{E}\underline{x}'$ bestaat asymptotisch een Poissonverdeling bezit met parameter $\lim_{N\to\infty}\mathscr{E}\underline{x}'$. Hierbij is

$$\begin{cases}x'=x-5n+\frac{5}{2}N & \text{voor } k=2\\ x'=x-8n+4N & \text{voor } k\geqq 3.\end{cases}$$

Passen we de formules voor $\mathscr{E}\underline{x}$ en $\sigma^2(\underline{x})$ toe op het door N a s s beschouwde geval ($k=m=9$ en $n=10$) dan vinden we

$$\mu=\mathscr{E}\underline{x}=\frac{4n\,(n-1)}{N-1}=4{,}5$$

$$\sigma^2=\sigma^2(\underline{x})=\frac{4n\,(n-1)\,(N-9)\,(N-n)\,(N-n-1)}{(N-1)^2\,(N-2)\,(N-3)}=3{,}267,$$

terwijl N a s s bij zijn steekproefexperiment vond

$$\bar{x}=4{,}47,\qquad s^2=3{,}625.$$

In de onderstaande tabel staan de linkeroverschrijdingskansen vermeld, gevonden met behulp van de normale benadering (met continuïteitscorrectie); daarnaast zijn opgenomen de door N a s s met zijn steekproefexperiment gevonden linkeroverschrijdingskansen.

Linkeroverschrijdingskansen van $\underline{x}$ gevonden met de normale benadering en met het steekproefexperiment

x	normale benadering	steekproefexperiment
0	0,0136	0,0198
1	0,0485	0,0594
2	0,1335	0,1386
3	0,2912	0,3168
4	0,5000	0,5446

Hieruit blijkt dat er een zeer goede overeenstemming bestaat tussen de normale benadering en de resultaten van het steekproefexperiment.

Literatuur

V a n E e d e n, C. and A. R. B l o e m e n a, (1959), On probability distributions arising from points on a lattice, Rapport S 257 van de Statistische afdeling van het Mathematisch Centrum, Amsterdam.

V a n E e d e n, C. (1959), De asymptotische eigenschappen van de hypergeometrische verdeling, Rapport S 254 (M 80) van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam.

F e l l e r, W. (1950), An introduction to probability theory and its applications, John Wiley and Sons, Inc., New York.

K r i s h n a I y e r, P. V. (1949), The first and second moments of some probability distributions arising from points on a lattice and their application, Biometrika **36**, 135–141.

K r i s h n a I y e r, P. V. (1950), The theory of probability distributions of points on a lattice, Ann. Math. Stat. **21**, 198–217. (Errata Ann. Math. Stat. **22** (1951) 310).

K u i p e r s, H. A. (1957), Over de verdeling van het aantal runs in reeksen van alternatieven, Rapport S 219 (Ov 7) van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam.

M o r a n, P. A. P. (1947), Random associations on a lattice, Proc. Cambr. Phil. Soc. **43**, 321–328.

M o r a n, P. A. P. (1948), The interpretation of statistical maps, Journal Royal Stat. Soc. B **10**, 243–251.

S t e v e n s, W. L. (1939), Distribution of groups in a sequence of alternatives, Ann. of Eug. **9**, 10–17.

Het examen Statistisch Analist

Uitbreiding van de examen-mogelijkheden tot het economische toepassingsgebied.

Het bestuur van de Vereniging voor Statistiek heeft besloten om voor het tweede gedeelte van het examen Statistisch Analist de mogelijkheid te openen, het economische toepassingsgebied als studie-richting te kiezen.

Dit examen zal betrekking hebben op de toepassing van de statistiek in het economisch onderzoek bij bedrijf en overheid. Het zal in het bijzonder zijn bestemd voor hen, die hun werkkring hebben gevonden of willen vinden in op praktische vraagstukken gerichte research, zoals b.v. marktonderzoek, intern-bedrijfseconomisch onderzoek, beleggingsonderzoek, bedrijfstaksgewijs onderzoek, accountantscontrole en nationale of regionale economische planning.

De examen-commissie Statistisch Analist van de Vereniging voor Statistiek is voor dit doel uitgebreid met een sub-commissie, waarin de volgende personen zitting hebben: H. J. R. Th. Claus ec. drs., H. Emanuel ec. drs (voorzitter), J. A. Hartog ec. drs., Prof. Dr. L. H. Klaassen, G. J. A. Mensink ec. drs. (secretaris) en Prof. Dr. H. Rijken van Olst.

Deze commissie legt thans de laatste hand aan het examenprogramma dat, naast de exameneisen, literatuurverwijzingen zal bevatten en een aantal uitgewerkte model-examenopgaven.

Er zal naar worden gestreefd, het examen voor het economische toepassingsgebied voor de eerste maal in de herfst van 1961 af te nemen.