STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Rapport S 268

Het aselect kiezen van knopen in een knoopsel

door

Ir A.R. Bloemena

Maart 1960

1. De probleemstelling.

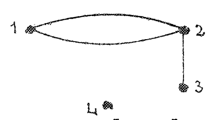
Gegeven n knopen, genummerd van 1,...,n en positieve gehele getallen m_{ij} $(i,j = 1,...,n)$. Men kan m_{ij} interpreteren als het aantal draden tussen knoop i en knoop j. Ondersteld zal worden dat

(1.1) $$m_{ij} = m_{ji}.$$

Voorbeeld: $n = 4$

$m_{11} = 0$ $m_{22} = 0$ $m_{33} = 0$ $m_{44} = 0$
$m_{12} = 2$ $m_{23} = 1$ $m_{34} = 0$
$m_{13} = 0$ $m_{24} = 0$
$m_{14} = 0$

welke situatie afgebeeld kan worden als



Een dergelijke combinatie van draden en knopen zullen wij een knoopsel*) noemen. Voorbeelden van knoopsels zijn: veelhoeken, stambomen, een kaart met verkeerswegen, schakelschema's van electrische installaties.

Ondersteld zal worden dat voor $i = 1,..,n$ $m_{ii} = 0$, dus dat er geen lussen optreden, en dat $m_{ij} < \infty$.

Men kiest in het knoopsel aselect r knopen $(r \leq n)$. Beschouw nu de kansgrootheden $\underline{x}_{ij}$ $(i,j = 1,...,n)$, gedefinieerd door

$$\underline{x}_{ii} = 0 \qquad \text{spr } 0,$$

$i \neq j$

$$\underline{x}_{ij} = \begin{cases} 1 \text{ als knoop } i \text{ en } j \text{ beide gekozen zijn} \\ 0 \text{ als dit niet zo is,} \end{cases}$$

en definieer

(1.2) $$\underline{x} = \sum_{i,j} m_{ij}\underline{x}_{ij}.$$

$\underline{x}$ is tweemaal het aantal draden tussen gekozen knopen.

----

*) E: graph. De vertaling is van Prof. Dr J. Hemelrijk.

Het probleem is de kansverdeling van $\underline{x}$ te berekenen. Eerst zullen twee voorbeelden van situaties besproken worden, waar deze kansverdeling van toepassing is.

2. Een toets tegen besmettelijkheid van een kenmerk op een populatie.

Op een proefveld zal een gewas, bij voorbeeld aardappelen, worden verbouwd. Volgens een regelmatig patroon, bij voorbeeld op de hoekpunten van een rechthoekig rooster, worden de pootaardappelen in de grond gebracht. Na verloop van enige tijd, als de planten boven de grond verschijnen, bemerkt men dat van de n planten er r ziek zijn. Onderstellen wij dat men de ziekte niet direct kan determineren, dan zal de vraag rijzen, wat men aan de hand van de geografische verspreiding van de zieke planten kan concluderen ten aanzien van het mechanisme van de ziekte. Met andere woorden: zijn de zieke planten aselect verdeeld over het veld of niet? Zijn de planten niet aselect verdeeld over het veld, maar staan de zieke planten bij voorbeeld dicht bij elkaar, dan zal men dit als een aanwijzing opvatten dat de ziekte besmettelijk is, dan wel (mede) veroorzaakt kan zijn door zeer locale verschillen in bodemcondities.

Een oordeel over deze kwestie kan men vormen door bij elke zieke plant te tellen hoeveel van de acht planten er omheen eveneens ziek zijn. Op deze wijze krijgt men voor iedere zieke plant een getal. De som, x, van deze getallen kan nu als toetsingsgrootheid worden gebruikt voor de hypothese, dat de zieke planten aselect verdeeld zijn over het proefveld. Indien deze hypothese waar is, is de toetsingsgrootheid x een waarneming van de kansgrootheid $\underline{x}$, gedefinieerd in (1.2), als men voor m_{ij} de waarde 1 kiest, als plant i en plant j buren zijn, en de waarde 0, als dit niet zo is. Is de nulhypothese niet vervuld, en is de ziekte bij voorbeeld besmettelijk, dan zal x systematisch grotere waarden aannemen dan in het geval de nulhypothese wel vervuld is. Een grote waarde van x zal dus leiden tot verwerping van de nulhypothese.

Bij besmettelijke ziekten worden evenwel niet alleen de naaste buren van een zieke plant besmet, maar in mindere mate ook de planten die verder weg staan. Op grond van deze overwegingen zou men dus ook een andere toetsingsgrootheid kunnen kiezen, welke bestaat uit de som van de bijdragen van ieder paar zieke planten. De bijdrage van een paar zieke planten zal men groter kiezen, naarmate de afstand van de planten op het veld kleiner is. Door een geschikte keuze van de getallen m_{ij} in (1.2) kan men dan weer zorgen dat, indien de nulhypothese vervuld is, de toetsingsgrootheid weer een waarneming is van $\underline{x}$.

3. Orde-wanorde problemen.

In de statistische mechanica kent men een groep van problemen, die men in de regel de "orde-wanorde" problemen noemt. Prof. Dr D. van Dantzig heeft er indertijd op gewezen, dat de bij dit onderzoek beschouwde problemen hiermee nauw samenhangen.

Een van de orde-wanorde problemen betreft een model voor ferromagnetisme, dat door E. ISING in 1925 is opgesteld. Volgens dit model wordt het ferromagnetisch materiaal ondersteld te bestaan uit atomen, die zich op de hoekpunten van een rooster bevinden. Elk atoom bezit een magnetisch moment μ_0, dat of wel dezelfde richting heeft als een uitwendig veld H, of wel daaraan tegengesteld gericht is. De energie van een dergelijk stelsel bestaat uit twee delen.

1) De energie tengevolge van de wisselwerking van de magnetische momenten en het uitwendige veld. Elk atoom, waarvan het magnetisch moment tegengesteld gericht is aan het veld H (een +atoom) geeft een bijdrage $+\mu_0 H$; elk atoom, waarvan het magnetisch moment gelijk gericht is aan een H (een -atoom) geeft een bijdrage $-\mu_0 H$.
2) De energie tengevolge van de wederzijdse beïnvloeding van de magnetische momenten onderling. Men voert op dit punt een aantal onderstellingen in, o.a. dat de totale interactie de som is van die tussen de paren atomen, en voorts dat alleen de interacties tussen die atomen van belang zijn, die directe buren zijn in het rooster. Zijn twee buur-atomen beide - of beide +atomen, dan geeft dit een energie bijdrage -J; is één een + en de ander een -atoom, dan geeft dit paar een bijdrage +J.

Zijn er n atomen in het rooster, en heeft ieder atoom m buren, dan zijn er dus $\frac{1}{2}mn$ paren, die van belang zijn voor dit deel van de energie van het stelsel. Stel dat er n_1 +atomen en dus $(n-n_1)$ -atomen zijn, dan noemt men

$$s = \frac{n_1-(n-n_1)}{n}$$

de afstandsorde-parameter.

Bij een gegeven verdeling van de n_1 +atomen en de $n-n_1$ -atomen zijn er nu van de $\frac{1}{2}mn$ paren naburige atomen x_1 paren atomen, die beide +atomen zijn, x_2 paren atomen, die beide -atomen zijn, en y paren atomen, die een ongelijk teken bezitten. Uiteraard is

$$2x_1 + y = n_1 m \qquad 2x_2 + y = (n-n_1)m$$

Men noemt

$$\sigma = \frac{x_1+x_2-y}{\frac{1}{2}mn}$$

de nabuurorde-parameter. De energie bij gegeven waarden van s en σ is dus

$$-\mu_0 Hs - \tfrac{1}{2}nmJ\sigma$$

Het probleem is hier het aantal mogelijke rangschikkingen te bepalen, waarop resp. n_1 + en $(n-n_1)$-atomen kunnen worden verdeeld over n roosterpunten, zodanig dat de parameter σ de gegeven waarde bezit. In de literatuur (zie bijv. J. de BOER (1947), G.F. NEWELL and E.W. MONTROLL (1953)) over dit onderwerp wordt aangegeven dat het probleem exact is opgelost voor een tweedimensionaal rooster. Voor driedimensionale roosters is een aantal benaderingsmethodes bekend.

Onderstelt men dat de +atomen en de -atomen aselect verdeeld zijn over de n roosterpunten, dan is bij een gegeven waarde van n_1 (dus bij gegeven s) de $\underline{\sigma}$ een kansgrootheid. Zoals in paragraaf 5 zal worden aangegeven is te bewijzen dat de kansverdeling van $\underline{\sigma}$ asymptotisch normaal is met verwachting s^2 en variantie $\frac{2(1-s^2)^2}{nm}$.

4. De momenten van $\underline{x}$.

De kansgrootheden $\underline{x}_{ij}$ $(i \neq j)$ hebben een verwachting

(4.1) $\quad \mathcal{E}\underline{x}_{ij} = P[\underline{x}_{ij}=1] = P[\text{knoop } i \text{ en } j \text{ beide gekozen}] =$

$= P[\text{knoop } i \text{ gekozen}]\,.\,P[\text{knoop } j \text{ gekozen} \mid \text{knoop } i \text{ gekozen}] =$

$\frac{r}{n}\cdot\frac{r-1}{n-1}.$

Uit (1.2) volgt dan

(4.2) $\quad \mathcal{E}\underline{x} = \frac{r(r-1)}{n(n-1)}\sum_{ij} m_{ij}\,.$

Kwadrateert men (1.2), dan volgt bij het nemen van verwachtingen

(4.3) $$\mathcal{E}\underline{x}^2 = \sum_{(i,j,k,l)\neq} m_{ij}m_{kl}\,\mathcal{E}\underline{x}_{ij}\underline{x}_{kl} + 4\sum_{(i,j,k)\neq} m_{ij}m_{ik}\,\mathcal{E}\underline{x}_{ij}\underline{x}_{ik} + 2\sum_{(i,j)\neq} m_{ij}^2\,\mathcal{E}\underline{x}_{ij}^2,$$

waar de indicatie $\neq$ bij het sommatieteken inhoudt, dat gesommeerd wordt over onderling ongelijke waarden van de indices.

Nu is voor $i \neq j$

$$\mathcal{E}\underline{x}_{ij}^2 = \mathcal{E}\underline{x}_{ij} = \frac{r(r-1)}{n(n-1)}\ ,$$

terwijl voor i, j en k ongelijk op analoge wijze als voor (4.1) volgt

$$\mathcal{E}\underline{x}_{ij}\underline{x}_{ik} = \frac{r(r-1)(r-2)}{n(n-1)(n-2)}$$

en voor i, j, k en l ongelijk:

$$\mathcal{E}\underline{x}_{ij}\underline{x}_{kl} = \frac{r(r-1)(r-2)(r-3)}{n(n-1)(n-2)(n-3)}\ .$$

Na enige herleiding vindt men dan uit (4.3):

$$\mathcal{E}\underline{x}^2 = \frac{r(r-1)}{n(n-1)(n-2)(n-3)}\Big[(r-2)(r-3)\Big(\sum_{ij} m_{ij}\Big)^2 + 4(r-2)(n-r)\sum_{ijk} m_{ij}m_{ik} + 2(n-r)(n-r-1)\sum_{ij} m_{ij}^2\Big].$$

en

(4.4) $$\sigma^2 = 4\,\frac{r(r-1)(r-2)(n-r)}{n(n-1)(n-2)(n-3)}\sum_i\Big(\sum_j m_{ij} - \frac{\sum_{ij} m_{ij}}{n}\Big)^2 + \frac{2r(r-1)(n-r)(n-r-1)}{n^2(n-1)^2(n-2)(n-3)}\Big\{n(n-1)\sum_{ij} m_{ij}^2 - \Big(\sum_{ij} m_{ij}\Big)^2\Big\}.$$

De eerste term van σ^2 stelt dus de bijdrage voor van de variantie van het aantal verbindingslijnen, uitgaande van een punt. Is $\sum_j m_{ij}$ constant voor alle i, hetgeen wil zeggen dat in elke knoop evenveel draden samenkomen, dan is deze bijdrage tot σ^2 gelijk aan nul.

Is r=n, dan worden alle knopen gekozen en er is niets stochastisch meer over: $\sigma^2=0$.

Als $r = n-1$ wordt de tweede term van σ^2 gelijk aan nul. Is ook $\sum_j m_{ij}$ constant voor $i = 1,\ldots,n$, dan is ook $\sigma^2=0$. Met zekerheid is $\underline{x}$ gelijk aan $(1 - \frac{2}{n}) \sum_{ij} m_{ij}$.

Het is niet moeilijk, maar alleen moeizaam, om hogere momenten te berekenen.

5. Asymptotisch gedrag van de kansverdeling van $\underline{x}$.

Wat het asymptotisch gedrag van de kansverdeling van $\underline{x}$ betreft, indien r en n beide naar oneindig gaan, dient men een aantal situaties te beschouwen. Wij zullen slechts de drie mogelijkheden bespreken, die voor practische toepassingen interessant zijn, nl.

I : $\lim \frac{r}{n} = 0$ en $\lim \frac{r^2}{n^2} \sum_{ij} m_{ij} = 0$

II : $\lim \frac{r}{n} = 0$ en $\lim \frac{r^2}{n^2} \sum_{ij} m_{ij} = 2\lambda$ $0 < \lambda < \infty$

III: $\lim \frac{r}{n} = \delta$ $0 < \delta < 1$.

In situatie I is $\underline{x}$ in de limiet met zekerheid gelijk aan nul. Voor de beide andere gevallen kunnen de volgende stellingen bewezen worden:

Stelling 1

Als voor $k = 1,\ldots,$ $\sum_j m_{ij}^k = m_k$, waar m_k niet van i en n afhangt; als $m_1 < \infty$ en als r en n beide naar oneindig gaan zodanig dat

$$\lim \frac{r}{n} = \delta \text{ , met } 0 < \delta < 1$$

dan gaat de verdeling van $[\sigma\{\underline{x}\}]^{-1} (\underline{x} - E\,\underline{x})$ naar een $N(0,1)$-verdeling. $E\,\underline{x}$ en $\sigma\{\underline{x}\}$ zijn gegeven in (4.2) en (4.4).

Stelling 2

Als $\sum_j m_{ij} = m_1$, onafhankelijk van i en als

$$(5.1)\quad \begin{cases} \lim \frac{r}{n} = 0 \\ \lim \frac{r}{n} m_1 = 0 \\ \lim \frac{r^2}{n} m_1 = 2\lambda, \qquad 0 < \lambda < \infty, \end{cases}$$

en voorts voor k = 1,...

$$\lim \frac{\sum_{ij} m_{ij}{}^k}{\sum_{ij} m_{ij}} = m_k^*,\ m_k^* < \infty \qquad (m_1^* = 1),$$

en tenslotte

$$m_{ij} < M \text{ voor alle i en j,}$$

waar M niet afhangt van n, dan heeft $\frac{1}{2}\ \underline{x}$ asymptotisch een samengestelde POISSON-verdeling met momentengenererende functie

$$\lim \mathcal{E} e^{\frac{1}{2} z \underline{x}} = \exp\{\lambda \sum_{i=1}^{\infty} m_i^* \frac{z^i}{i!}\}.$$

Is m_{ij} = 0 of 1, en geldt 5.1, dan heeft $\frac{1}{2}\ \underline{x}$ een POISSON-verdeling met parameter λ.

Bij het kiezen van punten in een rooster (zie bijv. van EEDEN en BLOEMENA (1960)), waarbij meervoudige verbindingen tussen twee punten niet voorkomen, volgt direct uit deze stellingen, dat $\frac{1}{2}\ \underline{x}$ asymptotisch normaal verdeeld is (als bijv. $\lim \frac{r}{n} = \delta$, $0 < \delta < 1$), of $\frac{1}{2}\ \underline{x}$ asymptotisch een POISSON-verdeling bezit (bijv. als $\lim \frac{r}{n} = 0$, $\lim \frac{r^2}{n} m < \infty$), welke voor grote waarden van de parameter weer te benaderen is door een normale verdeling.
Stelling 2 geldt uiteraard ook als $\frac{r}{n} \to 1$, maar dan voor het aantal verbindingslijnen tussen niet-gekozen punten.

6. Verder onderzoek.

Het onderzoek betreffende het kiezen van knopen in een knoopsel wordt nog voortgezet.

Kiest men na de r knopen nogmaals s knopen ($r+s \leq n$) op aselecte wijze en kleurt men de r knopen wit, de s knopen zwart en de n-r-s niet gekozen knopen rood, dan is dus $\frac{1}{2}\ \underline{x}$ het aantal wit-wit verbindingen. Dezelfde resultaten als voor $\frac{1}{2}\ \underline{x}$ gelden ook voor het aantal zwart-zwart verbindingen. Ook voor het aantal wit-zwart verbindingen zijn soortgelijke resultaten gevonden.

Men kan natuurlijk de knopen ook anders kiezen. Bij voorbeeld kan men voor elke knoop de kleur laten bepalen door een lotingsproces, dat met kans p_1 de kleur wit, met kans p_2 de kleur zwart en met kans $1-p_1-p_2$ de kleur rood voorschrijft. De lotingen geschieden onderling onafhankelijk voor iedere knoop. Ook voor deze situatie zijn weer analoge resultaten verkregen als die, welke reeds genoemd zijn.

Stelling 2 is een speciaal geval van een veel algemenere stelling, die geldt voor beide manieren om de knopen te kiezen en ook voor het geval, dat $\sum_j m_{ij}$ niet constant is voor $i=1,\ldots,n$.

Al deze resultaten zullen t.z.t. worden gepubliceerd.

7. Literatuur.

J. de BOER (1947): Ordeningsverschijnselen in ferromagnetica en binaire legeringen, Ned.T.v.Natuurk. 13, 29-50, 57-74.

CONSTANCE van EEDEN and A.R. BLOEMENA (1959): On probability distributions arising from points on a lattice. Rapport S 257, Statistische Afdeling Mathematisch Centrum.

D.J. FINNEY (1947): The significance of associations in a square point lattice, Journ.Roy.Stat.Soc. Suppl.(9),99.

P.V. KRISHNA IYER (1947): Random association of points on a lattice, Nature, 160, 714.

P.V. KRISHNA IYER (1948): The theory of probability distributions of points on a line, JISAS 1, 173-195.

P.V. KRISHNA IYER (1949): The first and second moments of some probability distributions arising from points on a lattice and their applications, Biometrika, 36, 135-141.

P.V. KRISHNA IYER (1950a): The theory of probability distributions of points on a lattice. Ann.Math.Stat. 21, 198-217 & 22, 310.

P.V. KRISHNA IYER (1950b,1951,1952): Further contributions to the theory of probability distributions of points on a line, JISAS 2, 141-160; 3, 80-93; 4, 50-71.

P.A.P. MORAN (1947): Random associations on a lattice, Proc. Cambr.Phil.Soc., 43, 321-328.

P.A.P. MORAN (1948): The interpretation of statistical maps. Journ.Roy.Stat.Soc. B 10, 243-251.

G.F. NEWELL and E.W. MONTROLL (1953): On the theory of the Ising model of ferro magnetism. Rev.Mod.Phys. 25, 353-389.

H. TODD (1940): Note on the random association in a square point lattice, Journ.Roy.Stat.Soc.Suppl. 7, 78.

JISAS = Journ. of the Indian Society for Agricultural Statistics.