STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Rapport S 276

Enige opmerkingen over het "best friend" probleem

door

Dr Constance van Eeden

juli 1960

1. Inleiding

In dit rapport wordt het volgende probleem beschouwd: Gegeven zijn N boeren, verdeeld in twee groepen A en B ter grootte m en n ($m+n=N$). Groep A bestaat bijv. uit vooruitstrevende en groep B uit conservatieve boeren.
Ieder van deze boeren geeft twee andere boeren op waar hij veel contact mee heeft. Gevraagd wordt te onderzoeken of de boeren overwegend contact hebben met boeren uit hun eigen groep en wel

I op grond van het eerste door de boeren opgegeven contact,
II op grond van beide door de boeren opgegeven contacten.

Dit probleem is in de literatuur bekend als het "best friend" probleem. Een uitvoerige bespreking en literatuurlijst zijn bijv. te vinden bij A. RAPOPORT (1957).

2. Beschrijving van de toets

Als toetsingsgrootheid in geval I kiezen we het aantal boeren dat een boer uit eigen groep kiest. Dus als voor de i^e boer

$$(2;1) \qquad x_i = \begin{cases} 1 & \text{als de } i^e \text{ boer een boer uit zijn eigen groep kiest} \\ 0 & \text{als de } i^e \text{ boer een boer uit de andere groep kiest} \end{cases}$$

dan is $x = \sum_{i=1}^{N} x_i$ de toetsingsgrootheid.

In geval II wordt de toetsingsgrootheid als volgt gedefinieerd: Stel voor de i^e boer

$$(2;2) \qquad y_i = \begin{cases} 2 & \text{als de } i^e \text{ boer 2 boeren uit zijn eigen groep kiest} \\ 1 & \text{als de } i^e \text{ boer 1 boer uit zijn eigen en 1 boer uit de andere groep kiest,} \\ 0 & \text{als de } i^e \text{ boer 2 boeren uit de andere groep kiest} \end{cases}$$

dan is $y = \sum_{i=1}^{N} y_i$ de toetsingsgrootheid.

De getoetste hypothese H_0 houdt in beide gevallen in dat iedere boer aselect kiest onafhankelijk van de keuze van de andere boeren.

In beide gevallen zal de toetsingsgrootheid onder H_0, voor grote waarden van N en als m en n onderling niet te veel in grootte verschillen, bij benadering normaal verdeeld zijn, zodat het voldoende is verwachting en variantie te bepalen. Hiervoor geldt (zie voor de bewijzen par. 3)

$$(2;3) \qquad \left.\begin{aligned} \mathcal{E}(\underline{x}|H_0) &= \frac{m(m-1)+n(n-1)}{N-1} \\ \sigma^2(\underline{x}|H_0) &= \frac{mn(N-2)}{(N-1)^2} \end{aligned}\right\} \quad \text{geval I}$$

en

$$(2;4) \qquad \left.\begin{aligned} \mathcal{E}(\underline{y}|H_0) &= 2\,\frac{m(m-1)+n(n-1)}{N-1} \\ \sigma^2(\underline{y}|H_0) &= \frac{2mn(N-3)}{(N-1)^2} \end{aligned}\right\} \quad \text{geval II}$$

De kritieke zone bestaat uit grote waarden van x (resp. y), daar H_0 getoetst wordt tegen de alternatieve hypothese dat de boeren bij voorkeur uit hun eigen groep kiezen.

In werkelijkheid zullen de boeren, ook als zij onafhankelijk van de groepen A en B kiezen, niet onderling onafhankelijk kiezen. Als boer i boer j gekozen heeft, zal de kans dat boer j boer i kiest groter zijn dan $\frac{1}{N-1}$ (de waarde onder H_0). Om de invloed van deze afhankelijkheid op de spreiding na te gaan beschouwen we ook de volgende situatie:

<u>in geval I</u>: gegeven dat boer i boer j gekozen heeft is de kans dat boer j boer i kiest gelijk aan 1. Wil iedere boer een andere kunnen kiezen dan moet dus in dit geval N even zijn. De waarnemingen bestaan nu dus uit paren boeren, waarbij ieder paar een bijdrage 0 of 2 tot x geeft en wel

$$(2;5) \qquad \begin{cases} 2 & \text{als de twee boeren van het paar tot dezelfde groep behoren} \\ 0 & \text{als de twee boeren van het paar tot verschillende groepen behoren.} \end{cases}$$

De te toetsen hypothese H_0' luidt nu dat deze paren zijn ontstaan door het aselect zonder teruglegging kiezen van $\frac{N}{2}$ paren uit de N boeren;

<u>in geval II</u>: gegeven dat boer i de boeren j en h gekozen heeft, is de kans dat boer j de boeren i en h kiest en boer h de boeren i en j kiest gelijk aan 1. Hier moet dus N een drievoud zijn en de waarnemingen bestaan uit drietallen boeren, waarbij ieder drietal een bijdrage 2 of 6 tot y geeft en wel

$$(2;6) \qquad \begin{cases} 6 \text{ als de drie boeren van het drietal tot één groep behoren,} \\ 2 \text{ als twee van de boeren tot één groep behoren en de derde tot de andere groep.} \end{cases}$$

De te toetsen hypothese H_0' luidt nu dat deze drietallen zijn ontstaan door het aselect en zonder teruglegging kiezen van $\frac{N}{3}$ drietallen uit de N boeren.

Ook onder de hypothese H_0' zullen de grootheden $\underline{x}$ en $\underline{y}$, voor grote waarden van N en als m en n onderling niet te veel in grootte verschillen, bij benadering normaal verdeeld zijn. Voor verwachting en variantie geldt (zie par. 3)

$$(2;7) \qquad \left.\begin{aligned} \mathcal{E}(\underline{x}|H_0') &= \mathcal{E}(\underline{x}|H_0) \\ \sigma^2(\underline{x}|H_0') &= \frac{8mn(m-1)(n-1)}{(N-1)^2(N-3)} \end{aligned}\right\} \qquad \text{geval I}$$

en

$$(2;8) \qquad \left.\begin{aligned} \mathcal{E}(\underline{y}|H_0') &= \mathcal{E}(\underline{y}|H_0) \\ \sigma^2(\underline{y}|H_0') &= \frac{16mn(m-1)(n-1)}{(N-1)^2(N-2)} \end{aligned}\right\} \qquad \text{geval II.}$$

Uit het bovenstaande blijkt dat

$$(2;9) \qquad \frac{\sigma^2(\underline{x}|H_0')}{\sigma^2(\underline{x}|H_0)} = \frac{\sigma^2(\underline{y}|H_0')}{\sigma^2(\underline{y}|H_0)} = 8\,\frac{(m-1)(n-1)}{(N-2)(N-3)} .$$

Voor grote waarden van m en n is deze verhouding dus hoogstens gelijk aan 2.

3. Bewijzen van de resultaten uit par. 2

3.1. De momenten van $\underline{x}$ en $\underline{y}$ onder H_0.

De grootheden x en y kunnen geschreven worden in de vorm

$$(3.1;1) \qquad x = \sum_{i=1}^{N} x_i, \qquad y = \sum_{i=1}^{N} y_i,$$

waarbij x_i resp. y_i $(i = 1,\ldots,N)$ gedefinieerd worden door (2;1) resp. (2;2). Zowel de grootheden $\underline{x}_1,\ldots,\underline{x}_N$ als de grootheden $\underline{y}_1,\ldots,\underline{y}_N$ zijn onderling onafhankelijk, zodat het probleem neerkomt op het bepalen van verwachting en variantie van $\underline{x}_i$ resp. $\underline{y}_i (i = 1,\ldots,N)$.

Zijn de boeren genummerd van 1 tot m in groep A en van m+1 tot N in groep B dan is (zie (2;1))

$$(3.1;2) \qquad \begin{cases} P[\underline{x}_i=1|H_0] = \begin{cases} \frac{m-1}{N-1} & i = 1,\ldots,m \\ \frac{n-1}{N-1} & i = m+1,\ldots,N \end{cases} \\ P[\underline{x}_i=0|H_0] = \begin{cases} \frac{n}{N-1} & i = 1,\ldots,m \\ \frac{m}{N-1} & i = m+1,\ldots,N. \end{cases} \end{cases}$$

Dus

$$(3.1;3) \qquad \mathcal{E}(\underline{x}_i|H_0) = \mathcal{E}(\underline{x}_i^2|H_0) = \begin{cases} \frac{m-1}{N-1} & i=1,\ldots,m \\ \frac{n-1}{N-1} & i=m+1,\ldots,N \end{cases}$$

en

$$(3.1;4) \qquad \sigma^2(\underline{x}_i| H_0) = \begin{cases} \frac{n(m-1)}{(N-1)^2} & i=1,\ldots,m \\ \frac{m(n-1)}{(N-1)^2} & i=m+1,\ldots,N. \end{cases}$$

Uit (3.1;3) en (3.1;4) volgt dan

$$(3.1;5) \qquad \begin{cases} \mathcal{E}(\underline{x}|H_0) = \sum_{i=1}^{N} \mathcal{E}(\underline{x}_i| H_0) = \frac{m(m-1)+n(n-1)}{N-1} \\ \sigma^2(\underline{x}|H_0) = \sum_{i=1}^{N} \sigma^2(\underline{x}_i| H_0) = \frac{mn(m-1)+nm(n-1)}{(N-1)^2} = \frac{mn(N-2)}{(N-1)^2}. \end{cases}$$

Voor de grootheden $\underline{y}_1, \ldots, \underline{y}_N$ geldt (zie (2;2))

$$
(3.1;6) \quad \begin{cases}
P[\underline{y}_i = 2 \mid H_0] = \begin{cases} \dfrac{(m-1)(m-2)}{(N-1)(N-2)} & i=1,\ldots,m \\[2ex] \dfrac{(n-1)(n-2)}{(N-1)(N-2)} & i=m+1,\ldots,N \end{cases} \\[6ex]
P[\underline{y}_i = 1 \mid H_0] = \begin{cases} 2\,\dfrac{n(m-1)}{(N-1)(N-2)} & i=1,\ldots,m \\[2ex] 2\,\dfrac{m(n-1)}{(N-1)(N-2)} & i=m+1,\ldots,N \end{cases} \\[6ex]
P[\underline{y}_i = 0 \mid H_0] = \begin{cases} \dfrac{n(n-1)}{(N-1)(N-2)} & i=1,\ldots,m \\[2ex] \dfrac{m(m-1)}{(N-1)(N-2)} & i=m+1,\ldots,N\ . \end{cases}
\end{cases}
$$

Uit (3.1;6) volgt

$$
(3.1;7) \quad \begin{cases}
\mathcal{E}(\underline{y}_i \mid H_0) = \begin{cases} 2\,\dfrac{m-1}{N-1} & i=1,\ldots,m \\[2ex] 2\,\dfrac{n-1}{N-1} & i=m+1,\ldots,N \end{cases} \\[6ex]
\sigma^2(\underline{y}_i \mid H_0) = \begin{cases} \dfrac{2n(m-1)(N-3)}{(N-1)^2(N-2)} & i=1,\ldots,m \\[2ex] \dfrac{2m(n-1)(N-3)}{(N-1)^2(N-2)} & i=m+1,\ldots,N \end{cases}
\end{cases}
$$

dus

$$
(3.1;8) \quad \begin{cases}
\mathcal{E}(\underline{y} \mid H_0) = \displaystyle\sum_{i=1}^{N} \mathcal{E}(\underline{y}_i \mid H_0) = 2\,\frac{m(m-1)+n(n-1)}{N-1} \\[4ex]
\sigma^2(\underline{y} \mid H_0) = \displaystyle\sum_{i=1}^{N} \sigma^2(\underline{y}_i \mid H_0) = \frac{2nm(m-1)(N-3)}{(N-1)^2(N-2)} + \frac{2mn(m-1)(N-3)}{(N-1)^2(N-2)} = \\[4ex]
\qquad = \dfrac{2mn(N-3)}{(N-1)^2}\ .
\end{cases}
$$

3.2. De momenten van $\underline{x}$ en $\underline{y}$ onder H_0'.

Onder de hypothese H_0' zijn de paren (resp. drietallen) aselect en zonder teruglegging getrokken uit de N boeren. Als deze paren (resp. drietallen) in volgorde van trekking genummerd worden van 1 tot $\frac{N}{2}$ (resp. van 1 tot $\frac{N}{3}$) en als $\underline{u}_\nu$ (resp. $\underline{v}_\nu$) de bijdrage tot $\underline{x}$ (resp. $\underline{y}$) van het ν^e paar (resp. drietal) voorstelt, dan is

$$(3.2;1) \qquad \underline{x} = \sum_{\nu=1}^{\frac{N}{2}} \underline{u}_\nu \ , \quad \underline{y} = \sum_{\nu=1}^{\frac{N}{3}} \underline{v}_\nu ,$$

waarbij (zie (2;5) en (2;6))

$$(3.2;2) \qquad \begin{cases} P\left[\underline{u}_\nu=2 \,|\, H_0'\right] = \dfrac{m(m-1)+n(n-1)}{N(N-1)} \ , \quad P\left[\underline{v}_\nu=6\right] = \dfrac{m(m-1)(m-2)+n(n-1)(n-2)}{N(N-1)(N-2)} \\ P\left[\underline{u}_\nu=0 \,|\, H_0'\right] = 2\,\dfrac{mn}{N(N-1)} \ , \quad P\left[\underline{v}_\nu=2\right] = \dfrac{3mn}{N(N-1)} \ . \end{cases}$$

Uit (3.2;2) volgt

$$(3.2;3) \qquad \begin{cases} \mathcal{E}(\underline{x}\,|\,H_0') = \displaystyle\sum_{\nu=1}^{\frac{N}{2}} \mathcal{E}(\underline{u}_\nu\,|\,H_0') = \frac{N}{2}.2\,\frac{m(m-1)+n(n-1)}{N(N-1)} = \frac{m(m-1)+n(n-1)}{N-1} \\ \mathcal{E}(\underline{y}\,|\,H_0') = \displaystyle\sum_{\nu=1}^{\frac{N}{3}} \mathcal{E}(\underline{v}_\nu\,|\,H_0') = \\ \qquad = \dfrac{N}{3}.6\left[\dfrac{m(m-1)(m-2)+n(n-1)(n-2)}{N(N-1)(N-2)} + \dfrac{mn}{N(N-1)}\right] = \\ \qquad = 2.\dfrac{m(m-1)+n(n-1)}{N-1} \ . \end{cases}$$

Hieruit blijkt dus dat

$$(3.2;4) \qquad \mathcal{E}(\underline{x}\,|\,H_0') = \mathcal{E}(\underline{x}\,|\,H_0) \text{ en } \mathcal{E}(\underline{y}\,|\,H_0') = \mathcal{E}(\underline{y}\,|\,H_0) .$$

Dit was ook direct in te zien. Ieder der grootheden $\underline{x}_i$ (resp. $\underline{y}_i$) heeft onder H_0' dezelfde (marginale) verdeling als onder H_0. Onder H_0 zijn zij onafhankelijk, onder H_0' afhankelijk, wat op de verwachting van $\underline{x}$ (resp. $\underline{y}$) geen invloed heeft.

Om de variantie van $\underline{x}$ en $\underline{y}$ te berekenen zijn de volgende momenten nodig

$$(3.2;5) \qquad \mathcal{E}(\underline{u}_\nu^2|H_0'),\ \mathcal{E}(\underline{v}_\nu^2|H_0'),\ \mathcal{E}(\underline{u}_\nu\underline{u}_\kappa|H_0') \text{ en } \mathcal{E}(\underline{v}_\nu\underline{v}_\kappa|H_0')$$

voor $\nu \neq \kappa$.

Uit (3.2;2) volgt

$$(3.2;6) \qquad \begin{cases} \mathcal{E}(\underline{u}_\nu^2|H_0') = 4\,\dfrac{m(m-1)+n(n-1)}{N(N-1)} \\ \mathcal{E}(\underline{v}_\nu^2|H_0') = 36\,\dfrac{m(m-1)(m-2)+n(n-1)(n-2)}{N(N-1)(N-2)} + 12\,\dfrac{mn}{N(N-1)} \end{cases} .$$

We berekenen eerst $\mathcal{E}(\underline{u}_\nu\underline{u}_\kappa|H_0')$ voor $\nu \neq \kappa$. Daar de grootheden $\underline{u}_\nu$ slechts de waarden 0 en 2 aannemen, is

$$(3.2;7) \qquad \mathcal{E}(\underline{u}_\nu\underline{u}_\kappa|H_0') = 4P[\underline{u}_\nu=2,\ \underline{u}_\kappa=2|H_0'] .$$

Nu is (zie ook (3.2;2))

$$(3.2;8) \qquad P[\underline{u}_\kappa=2|\underline{u}_\nu=2;\ H_0'] = \begin{cases} \dfrac{(m-2)(m-3)+n(n-1)}{(N-2)(N-3)} & \text{als het } \nu^e \text{ paar tot A behoort} \\ \dfrac{m(m-1)+(n-2)(n-3)}{(N-2)(N-3)} & \text{als het } \nu^e \text{ paar tot B behoort} \end{cases}$$

dus

$$(3.2;9) \qquad P[\underline{u}_\nu=2,\ \underline{u}_\kappa=2|H_0'] =$$

$$= \frac{m(m-1)}{N(N-1)} \cdot \frac{(m-2)(m-3)+n(n-1)}{(N-2)(N-3)} + \frac{n(n-1)}{N(N-1)} \cdot \frac{m(m-1)+(n-2)(n-3)}{(N-2)(N-3)} .$$

Uit (3.2;6), (3.2;7) en (3.2;9) volgt dan

$$(3.2;10) \quad \mathcal{E}(\underline{x}^2|H_0') = \sum_\nu \mathcal{E}(\underline{u}_\nu^2|H_0') + \sum_{\nu\neq\kappa}\sum \mathcal{E}(\underline{u}_\nu\underline{u}_\kappa|H_0') =$$

$$= \frac{N}{2} \cdot 4\,\frac{m(m-1)+n(n-1)}{N(N-1)} +$$

$$+ \frac{N}{2}\left(\frac{N}{2}-1\right) . 4\left[\frac{m(m-1)}{N(N-1)} \cdot \frac{(m-2)(m-3)+n(n-1)}{(N-2)(N-3)} + \frac{n(n-1)}{N(N-1)}\,\frac{m(m-1)+(n-2)(n-3)}{(N-2)(N-3)}\right] .$$

Uit $\sigma^2(\underline{x}) = \mathcal{E}\underline{x}^2 - (\mathcal{E}\underline{x})^2$ volgt dan na enige herleiding

$$(3.2;11) \qquad \sigma^2(\underline{x}|H_0') = \frac{8mn(m-1)(n-1)}{(N-1)^2(N-3)} .$$

We berekenen nu $\mathcal{E}(\underline{v}_\nu \underline{v}_\kappa | H_0')$ voor $\nu \neq \kappa$. Daar de grootheden $\underline{v}_\nu$ slechts de waarden 2 en 6 aannemen, is

(3.2;12) $\mathcal{E}(\underline{v}_\nu \underline{v}_\kappa | H_0') =$

$$= 4P[\underline{v}_\nu=2, \underline{v}_\kappa=2 | H_0'] + 24P[\underline{v}_\nu=2, \underline{v}_\kappa=6 | H_0'] + 36P[\underline{v}_\nu=6, \underline{v}_\kappa=6 | H_0'].$$

Nu is (zie ook (3.2;2))

(3.2;13)

$$P[\underline{v}_\kappa=2 | \underline{v}_\nu=2; H_0'] = \begin{cases} \dfrac{3(m-1)(n-2)}{(N-3)(N-4)} & \text{als precies één van het } \nu^e \text{ drietal tot A behoort} \\[2ex] \dfrac{3(n-1)(m-2)}{(N-3)(N-4)} & \text{als precies één van het } \nu^e \text{ drietal tot B behoort} \end{cases}$$

$$P[\underline{v}_\kappa=6 | \underline{v}_\nu=2; H_0'] = \begin{cases} \dfrac{(m-1)(m-2)(m-3)+(n-2)(n-3)(n-4)}{(N-3)(N-4)(N-5)} & \text{als precies één van het } \nu^e \text{ drietal tot A behoort} \\[2ex] \dfrac{(m-2)(m-3)(m-4)+(n-1)(n-2)(n-3)}{(N-3)(N-4)(N-5)} & \text{als precies één van het } \nu^e \text{ drietal tot B behoort} \end{cases}$$

$$P[\underline{v}_\kappa=6 | \underline{v}_\nu=6; H_0'] = \begin{cases} \dfrac{(m-3)(m-4)(m-5)+n(n-1)(n-2)}{(N-3)(N-4)(N-5)} & \text{als het } \nu^e \text{ drietal geheel tot A behoort} \\[2ex] \dfrac{m(m-1)(m-2)+(n-3)(n-4)(n-5)}{(N-3)(N-4)(N-5)} & \text{als het } \nu^e \text{ drietal geheel tot B behoort.} \end{cases}$$

Uit (3.2;13) volgt

(3.2;14) $$P[\underline{v}_\nu=2, \underline{v}_\kappa=2 | H_0'] = \frac{3mn(n-1)}{N(N-1)(N-2)} \cdot \frac{3(m-1)(n-2)}{(N-3)(N-4)} +$$

$$+ \frac{3mn(m-1)}{N(N-1)(N-2)} \cdot \frac{3(m-2)(n-1)}{(N-3)(N-4)} = \frac{9mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)}$$

en

(3.2;15) $$P\left[\underline{v}_\nu=2,\ \underline{v}_\kappa=6 \,\middle|\, H_0'\right] =$$

$$= \frac{3mn(n-1)}{N(N-1)(N-2)} \cdot \frac{(m-1)(m-2)(m-3)+(n-2)(n-3)(n-4)}{(N-3)(N-4)(N-5)}$$

$$+ \frac{3mn(m-1)}{N(N-1)(N-2)} \cdot \frac{(m-2)(m-3)(m-4)+(n-1)(n-2)(n-3)}{(N-3)(N-4)(N-5)} =$$

$$= \frac{3mn(m-1)(m-2)(m-3)}{N(N-1)(N-2)(N-3)(N-4)} + \frac{3mn(n-1)(n-2)(n-3)}{N(N-1)(N-2)(N-3)(N-4)}$$

en

(3.2;16) $$P\left[\underline{v}_\nu=6,\ \underline{v}_\kappa=6 \,\middle|\, H_0'\right] =$$

$$= \frac{m(m-1)(m-2)}{N(N-1)(N-2)} \cdot \frac{(m-3)(m-4)(m-5)+n(n-1)(n-2)}{(N-3)(N-4)(N-5)} +$$

$$+ \frac{n(n-1)(n-2)}{N(N-1)(N-2)} \cdot \frac{m(m-1)(m-2)+(n-3)(n-4)(n-5)}{(N-3)(N-4)(N-5)} .$$

Uit (3.2;12), (3.2;14), (3.2;15) en (3.2;16) volgt dan na enige herleiding

(3.2;17) $$\mathcal{E}\left(\underline{v}_\nu \underline{v}_\kappa \,\middle|\, H_0'\right) =$$

$$= \frac{72mn(m-1)(n-1)+36n(n-1)(n-2)(n-3)+36m(m-1)(m-2)(m-3)}{N(N-1)(N-2)(N-3)} .$$

Verder is

(3.2;18) $$\mathcal{E}\left(\underline{y}^2 \,\middle|\, H_0'\right) = \sum_\nu \mathcal{E}\left(\underline{v}_\nu^2 \,\middle|\, H_0'\right) + \sum_{\nu\neq\kappa}\sum \mathcal{E}\left(\underline{v}_\nu \underline{v}_\kappa \,\middle|\, H_0'\right) =$$

$$= \frac{N}{3}\mathcal{E}\left(\underline{v}_\nu^2 \,\middle|\, H_0'\right) + \frac{N}{3}\left(\frac{N}{3}-1\right)\mathcal{E}\left(\underline{v}_\nu \underline{v}_\kappa \,\middle|\, H_0'\right) .$$

Uit (3.2;6) en (3.2;17) volgt dan na enige herleiding

(3.2;19) $$\sigma^2\left(\underline{y} \,\middle|\, H_0'\right) = \frac{16mn(m-1)(n-1)}{(N-1)^2(N-2)} .$$

Literatuur

A. RAPOPORT, Contribution to the theory of random and biased nets, Bull. of Math. Biophysics 19 (1957) 257-277.