STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Rapport S 288

Over het al of niet samenhangen van beweringen in een attitude-studie

door

P. v.d. Laan

juli
1961

1. Korte omschrijving van de proefopzet en het hiermee samenhangend probleem

In een attitude-studie werden aan 640 schoolkinderen 21 beweringen over de mensen van India voorgelegd. Hieronder waren sympathieke, onsympathieke en "neutrale" beweringen.

Enkele voorbeelden hiervan: "De mensen uit India zijn de aardigste mensen ter wereld", "De mensen uit India zijn het meest verachtelijke volk ter wereld" en "De mensen uit India zijn groot".

Uit deze 21 beweringen moest ieder kind die 6 beweringen kiezen, die het meest met zijn eigen mening overeenkwamen. Ze hoefden het er dus niet noodzakelijkerwijs volledig mee eens te zijn.

Volgens een extern (jury-)criterium zouden de 21 beweringen te verdelen zijn in 7 groepjes van ieder 3 beweringen. De 3 beweringen in een groepje vertoonden dan samenhang, terwijl de 7 groepjes onderling slechts waren te rangschikken. Om dit te controleren worden met deze studie paren van ite s, beweringen gecontroleerd op onafhankelijkheid resp. afhankelijkheid. Afhankelijkheid van twee beweringen houdt hier in dat ze enige samenhang vertonen, zodat bij de proef deze beweringen bij kinderen, die 1 van deze 2 items hebben aangestreept, vaker samen zullen voorkomen dan volgens toeval verwacht zou worden.

Voor een uitvoerige bespreking van de achtergrond van het probleem en literatuurlijst, zie: M. BROUWER en ANNIE VAN BERGEN, Communication effects on attitudes toward India. A scale comparison experiment. o.a. Appendix XIV T-I. Amsterdam 1960. Sponsored by the Unesco, Social Science Department.

Algemeen probleem

Indien van een verzameling van N elementen bekend is, dat ieder element precies k verschillende attributen van n attributen bezit, terwijl bovendien gegeven zijn: de frequentie waarmee ieder van de n attributen voorkomt en de frequenties waarmee paren van attributen voorkomen, wordt gevraagd te toetsen of een bepaalde combinatie van attributen meer of minder voorkomt dan volgens toeval verwacht zou mogen worden.

2. Inleidende opmerkingen

Het algemene probleem is voorlopig niet exact oplosbaar gebleken. Voor de eenvoud beschouwen we paren van beweringen. Ook is het mogelijk h-tallen ($2 \leqslant h \leqslant k$) te beschouwen; wellicht zou het dan aanbeveling verdienen een over-all toets te gebruiken (matrix met N rijen, n kolommen, met op iedere rij k aanstrepingen). Een dergelijke toets is echter (nog) niet beschikbaar.

Diverse overwegingen zullen bij het aanstrepen een rol spelen. Misschien zullen er verhoudingsgewijs meer sympathieke dan onsympathieke beweringen aangestreept worden, daar de kinderen wellicht menen dat onderstreping van sympathieke items "verwacht" wordt. Ook is het mogelijk dat door de laatste aanstrepingen (wegens vermoeidheid, verveling) meer "neutrale" items getroffen zullen worden, zodat deze wellicht een relatief hoog aantal keren voorkomen. Hieruit volgt zonder meer al dat men als modelveronderstelling niet kan nemen dat de kansen op aanstreping der items gelijk zijn.

De moeilijkheid is in dit geval dat door de verplichting tot het kiezen van k beweringen (in het onderhavige geval 6), we niet zonder meer met behulp van een 2 x 2-tabel op onafhankelijkheid van ieder paar items kunnen toetsen. De keuze van de j^{de} bewering A_j is daardoor niet onafhankelijk van het wel of niet gekozen zijn van de i^{de} bewering A_i. Immers in het eerste geval zijn er nog 5 aanstrepingen over en in het tweede geval 6.

De moeilijkheid is dat de verwachting van het samen voorkomen van de beweringen A_i en A_j (onder de nulhypothese dat A_i en A_j geen samenhang vertonen, onafhankelijk zijn) verschilt met die in het geval waarbij ieder item kan worden aanvaard of verworpen, zonder dat het aantal aanstrepingen voorgeschreven is. We kunnen dit als volgt aantonen.

De beweringen $A_1, A_2, \dots, A_n$ worden aangestreept met een zekere kans. Na iedere aanstreping veranderen de kansen, omdat reeds aangestreepte beweringen niet nogmaals aangestreept kunnen worden.
Stel bij de eerste aanstreping heeft bewering A_i een kans p_i om gekozen te worden $(i=1,2,\dots,n)$ $\sum_{i=1}^{n} p_i = 1$ $q_i = 1-p_i$. Als we aannemen dat bij de eerste aanstreping A_ν gekozen is, dan zijn bij de tweede aanstreping de kansen van de overgebleven n-1 beweringen:

$$\frac{p_i}{q_\nu} \; (i=1,2,\dots,n;\; i \neq \nu).$$

Deze gang van zaken herhaalt zich tot en met de k^e aanstreping. Voor de eenvoud nemen we: k=2 en n=3. Voor willekeurige k en n gaat de berekening eender.
We definiëren:

$\{A_1\}$: bij de 1ste aanstreping A_1 of bij de 2e aanstreping A_1

$\{A_2\}$: " " " " A_2 " " " " A_2

$\{A_1A_2\}$: " " beide aanstrepingen A_1 en A_2 gekozen.

Onafhankelijkheid in ons model wil dus zeggen dat de kans op aanstreping van een bepaalde bewering onafhankelijk is van de vorige aanstrepingen. Dan geldt:

$$P(\{A_1\}) = p_1+p_2\,\frac{p_1}{q_2} + p_3\,\frac{p_1}{q_3} = p_1\left(1+\frac{p_2}{q_2}+\frac{p_3}{q_3}\right) = p_1\left(\frac{1}{q_2}+\frac{p_3}{q_3}\right) \text{ daar}$$

$$p_2+q_2 = 1.$$

$$P(\{A_2\}) = p_2+p_1\,\frac{p_2}{q_1} + p_3\,\frac{p_2}{q_3} = p_2\left(1+\frac{p_1}{q_1}+\frac{p_3}{q_3}\right) = p_2\left(\frac{1}{q_1}+\frac{p_3}{q_3}\right).$$

$$P(\{A_1,A_2\}) = p_1\,\frac{p_2}{q_1} + p_2\,\frac{p_1}{q_2} = p_1p_2\left(\frac{1}{q_1}+\frac{1}{q_2}\right).$$

Vermenigvuldigen we $P(\{A_1\})$ met q_2q_3 en $P(\{A_2\})$ met q_1q_3 dan geeft dit: $q_1q_2q_3{}^2P(\{A_1\})P(\{A_2\}) = p_1p_2(q_3+p_3q_2)(q_3+q_1p_3)$. En $q_1q_2q_3{}^2\ P(\{A_1A_2\}) = p_1p_2q_3{}^2(q_1+q_2)$.

Om aan te tonen dat $P(\{A_1\})P(\{A_2\}) \geq P(\{A_1A_2\})$, is het voldoende, daar de p_i en q_i niet negatief zijn, te laten zien dat:

$(q_3+p_3q_2)(q_3+q_1p_3)-(q_1+q_2)q_3{}^2 \geq 0$.

Het bewijs gaat als volgt (N.B. $q_i=1-p_i$ en $\sum_{j=1}^{3} q_j = 2;\ i=1,2,3$)

$$(q_3+p_3q_2)(q_3+q_1p_3)-(q_1+q_2)q_3{}^2 =$$

$$(1-p_2p_3)(1-p_1p_3)-(1+p_3)(1-p_3)^2 =$$

$$1-(p_1+p_2)p_3+p_1p_2p_3{}^2-1+2p_3-p_3{}^2-p_3+2p_3{}^2-p_3{}^3 =$$

$$p_3{}^2(p_1p_2+2-p_3) =$$

$$p_3{}^2(p_1p_2+1+p_1+p_2+p_3-p_3) =$$

$$p_3{}^2(1+p_1+p_2+p_1p_2) \geq 0 \text{ daar } p_i \geq 0.\ (i=1,2,3)$$

Ook voor willekeurige n kan bewezen worden dat

$$P(\{A_1\})P(\{A_2\}) \geq P(\{A_1,A_2\}).$$

Als voorbeeld nemen we in het geval k=2 en n=3,

$p_1=p_2=0,1$ $\quad p_3=0,8$

$q_1=q_2=0,9$ $\quad q_3=0,2.$

Na bere ening krijgen we: $P(\{A_1\})P(\{A_2\}) = p_1p_2.26,1$

$P(\{A_1,A_2\}) = p_1p_2.2,2.$

We zien in dit voorbeeld duidelijk dat

$$P(\{A_1\})P(\{A_2\}) > P(\{A_1A_2\}).$$

Bij onafhankelijkheid is dus de verwachting van het aantal keren dat twee beweringen samen voorkomen kleiner dan het product van de marginale verwachtingen.

Tot slot nog een opmerking met een meer algemene strekking. Vinden we een, bijvoorbeeld positieve, correlatie tussen twee beweringen, dan wil dit niet zonder meer zeggen dat de beide beweringen direct afhankelijk van elkaar zijn. Mogelijk is dat ze via een andere bewering gecorreleerd zijn. Dit is niet te ontlopen, alleen van psychologische kant is dit nader uit te werken.

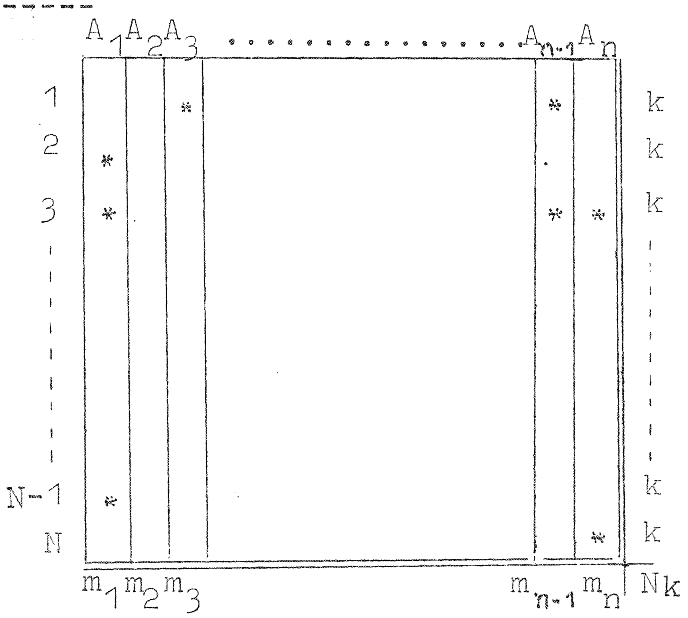
3. Model en toetsingsgrootheid

Men stellen zich voor dat de N elementen (formulieren in dit geval) een steekproef vormen uit een uitgebreide populatie.

De beweringen noemen we: $A_1, A_2, \ldots, A_n$ (attributen, items).
Ieder element bevat dus k attributen ($2 \leqslant k < n$).

Laat m_i het aantal keren zijn dat bewering A_i is aangestreept. ($i=1,2,\ldots,n$).

Schema:



* : aangestreepte bewering
Rijtotalen zijn constant en gelijk aan k
Kolomtotalen zijn:
$m_1, m_2, \ldots\ldots, m_n$.
In het beschouwde geval geldt: N=640, k=6 en n=21.

Kies, indien mogelijk, bij iedere bewering de ontkenning daarvan. Stel, 2 tegengestelde beweringen zijn A_1 en A_2. We willen nagaan de afhankelijkheid van A_3 ten opzichte van A_1 (of A_2).

Betreffende het aanstrepen door de kinderen zijn er 3 mogelijkheden:

1. Het kind heeft zowel A_1 als A_2 aangestreept. Daar dit zeer onwaarschijnlijk is, wordt dit verder buiten beschouwing gelaten.

2. Het kind heeft geen van beide items aangestreept. Het heeft dus voor geen van beide beweringen een duidelijke voorkeur. De gegevens van dit kind worden niet gebruikt om de afhankelijkheid van A_3 t.o.v. A_1 en A_2 (dus eigenlijk $\overline{A_1}$) te toetsen. Stel het aantal van zulke kinderen is f.

3. Het kind heeft één van beide items aangestreept. De corresponderende formulieren worden gebruikt om de afhankelijkheid van A_3 t.o.v. A_1 te toetsen. Aantal is N-f.

Het bepalen van de polaire items kan wellicht op twee manieren geschieden en wel:

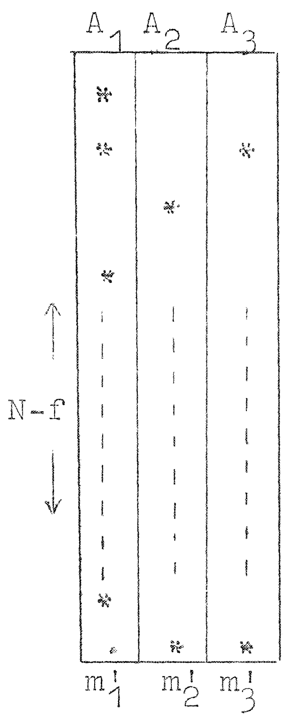
<u>a</u> Door een jury van psychologen.

<u>b</u> Met behulp van het volgende werkschema, voor afhankelijkheid van A_p en A_q.

1. Neem die beweringen welke niet samen met A_p voorkomen. Uit deze beweringen nemen A_r, met de eigenschap dat deze het vaakst gekozen is.

2. Neem die beweringen welke niet samen met A_q voorkomen. Uit deze nemen A_s, met de eigenschap dat deze het vaakst gekozen is.

3. Van A_r en A_s die nemen welke het meest aangestreept is. Stel A_r. Dan kunnen A_p en A_r dienst doen als de twee polaire beweringen.



Nu geldt dat, voor A_1 en A_2 samen, op iedere rij precies één aangestreepte bewering staat. Dus nu hebben we voor A_3 steeds k-1 mogelijkheden voor aanstreping over. We toetsen onder de voorwaarde:

òf A_1 bezet

òf A_2 bezet.

De ene dichotomie is dus: A_1 of A_2

en de andere : A_3 of $\bar{A}_3$.

In een 2 × 2-tabel neergezet krijgen we:

	A_3	$\bar{A}_3$	
A_1	a	c	m
A_2	b	d	n
	r	s	N-f

Nulhypothese H_0: A_3 onafhankelijk van A_1 en A_2. Voor het samen voorkomen van A_3 en A_1 kunnen we als model nemen: aselecte trekking zonder teruglegging met omvang m_3' en $\frac{m_1'}{N-f}$ als kans op succes.

De kansverdeling voor het aantal malen samen voorkomen van A_3 en A_1 onder de voorwaarde dat A_1 m keer gekozen is, wordt gegeven door:

$$P(\underline{a}=a/\underline{m}=m;\ H_0) = \frac{\binom{m}{a}\binom{n}{b}}{\binom{N-f}{r}}$$

$$\text{met: } m-\min(n,m) \leq a \leq \min(r,m).$$

Normale benadering:

$$\begin{array}{l} N \to \infty \\ m \approx r \end{array} \quad \text{dan is } \frac{|\underline{a}-\mu|-\frac{1}{2}}{\sigma}$$ normaal verdeeld met gemiddelde 0 en spreiding 1. Hierin is:

$$\mu = \frac{mr}{N-f}$$

$$\sigma^2 = \frac{mnrs}{(N-f)^2(N-f-1)}\ .$$

Ook is het mogelijk als benadering een POISSON verdeling of χ^2-verdeling te nemen.

De benadering met een POISSON-verdeling met parameter $\mu = \frac{mr}{N-f}$ kan men nemen, indien p (of 1-p) klein is en m veel kleiner (of veel groter) is dan n.

Dan geldt:

$$P(\underline{a}=a|\underline{m}=m;H_0) = \frac{e^{-\mu}\mu^a}{a!}\ .$$

Voor χ_1^2-benadering: $\frac{\{a(N-f)-mr\}^2}{mnrs/(N-f-1)}$ is verdeeld als χ^2 met 1 vrijheidsgraad.

4. Mogelijke uitbreiding

In plaats van twee beweringen, welke tegengesteld zijn, kunnen we ook groepen van beweringen nemen en de afhankelijkheid van een bewering A_p ten opzichte van deze groepen toetsen.

Kies hiertoe m beweringen ($1 < m < n$): $A_1, A_2, \ldots, A_m$. Neem alle kinderen, die een vast aantal h van deze m beweringen hebben aangestreept ($h \leq m$ en $< k$), zodat voor de rest nog k-h mogelijke aanstrepingen overblijven. (Beter is h oneven te kiezen, voor de nu volgende

Voor de eerste dichotomie nemen we:

Merendeel der aanstrepingen in de ene groep C
of " " " " " " andere groep D.

Bij ieder kind is na te gaan of het bewering A_p aangestreept heeft of niet. Dit wordt dan de tweede dicotomie.
Dit is zonder meer in een 2 x 2-tabel te zetten en te toetsen op onafhankelijkheid.

	A_p	$\bar{A}_p$	
Grootste aantal in C	a	c	m
Grootste aantal in D	b	d	n
	r	s	N'

Voor de tweede dichotomie kunnen we in plaats van één attribuut ook een groep van beweringen nemen en vervolgens tellen hoeveel malen een bewering aangestreept is in deze groep bij ieder van de in aanmerking komende kinderen. Om te toetsen kunnen we hier bijvoorbeeld de twee-steekproeventoets van WILCOXON toepassen.

De mogelijkheid ligt open om, nadat bovenstaande is gedaan voor verschillende h, de diverse toetsingen te combineren tot één toets.

Het combineren van 2 x 2-tabellen van de vorm:

	A	$\bar{A}$	$i=1,2,\ldots,M$
Ene groep	a_i	c_i	m_i
Andere groep	b_i	d_i	n_i
	r_i	s_i	N_i

Stel dat we, bij de i^{de} opzet, bij de eerste groep kans p_i op A hebben en bij de tweede groep kans p_i'. De te toetsen hypothese is dat voor iedere i geldt: $p_i=p_i'$. De p_i mogen dus onderling wel verschillen. De alternatieve hypothese is dat voor minstens één i geldt: $p_i \neq p_i'$.

Onder de nulhypothese heeft:

$$\chi^2 = \sum_{i=1}^{M} \frac{(a_i N_i - m_i r_i)^2}{\frac{m_i n_i r_i s_i}{N_i - 1}} = \sum_{i=1}^{M} (N_i - 1) \frac{(a_i d_i - b_i c_i)^2}{m_i n_i r_i s_i}$$

bij benadering een χ^2-verdeling met M vrijheidsgraden.
Voor andere mogelijkheden zie: C. van Eeden, Stat.Neerl.7 (1953) p. 141-162.

Het combineren van toetsingen van WILCOXON:
Laat $\underline{x}_{i,r}$ (i=1,2,...,M, dus M toetsingen; r=1,2,...,m_i) zijn de r^{de} waarneming van het aantal aanstrepingen in de ene reeks en $\underline{y}_{i,s}$ (i=1,2,...,k en s=1,2,...,n_i) voor de andere reeks.

Als toetsingsgrootheid berekenen we:

$$\underline{W} \stackrel{def}{=} \sum_{i=1}^{M} c_i \underline{w}_i \text{ met } \underline{w}_i = \sum_{r=1}^{m_i} \sum_{s=1}^{n_i} \text{signum}(\underline{x}_{ir} - \underline{y}_{is})$$

en c_i (zgn. gewichten) nader te bepalen.

Zie eventueel: PH. VAN ELTEREN (1960), Rapport SP 68 van de Stat.Afd. v.h. Mathematisch Centrum.

5. <u>Verbeterde proefopzet</u>

Daar in het practische geval de 21 beweringen gerangschikt zijn in 7 groepjes van 3, is er de moeilijkheid, bij voorafgaande methode van oplossing, om bij het middelste groepje tegengestelde beweringen te vinden.

Om deze reden kan men beter door de kinderen hun aanstrepingen in volgorde laten nummeren. Dan kan namelijk wel een geïsoleerd 2-tal van kolommen (beweringen) onderzocht worden. We nemen dan, als we bijvoorbeeld A_p en A_q willen vergelijken:

1^e dichotomie: A_p wel of A_p niet aangestreept.

2^e " A_q " " A_q " " " , maar waarbij we voor die regels, waar A_p niet werd aangestreept, de laatste (k^{de}) van de aanstrepingen, bij het beschouwen van deze dichotomie, weglaten. In beide gevallen blijven dan k-1 mogelijkheden voor aanstreping van A_q over.

Ook zou een eenvoudiger analyse mogelijk zijn, indien bij iedere bewering de ontkenning ervan was opgenomen. (Zie 3)

Eenvoudiger zou de analyse ook worden als de beweringen in ja- en nee-vragen geponeerd werden, daar dan zonder meer de 2 × 2-tabel methode gebruikt kan worden.

Literatuur:

M. BROUWER & A. VAN BERGEN, Communication effects on attitudes toward India, A scale comparison experiment, Amsterdam, 1960.

W.J. DIXON & F.J. MASSEY, Introduction to statistical analysis, Mc Graw-Hill Book Company, 1951.

C. VAN EEDEN, Methoden voor het vergelijken, toetsen en schatten van onbekende kansen, Stat.Neerl.7 (1953) 141-162. Rapport S 115 (M47) Stat.Afd.v.h. Math. Centrum.

C. VAN EEDEN & CHR.L. RUMKE, Statistiek voor medici, Korte inleiding. Stafleu, Leiden 1961.

C. VAN EEDEN & D. WABEKE, Toets van Wilcoxon, Rapport 176 Stat.Afd.v.h. Math.Centrum.

PH. VAN ELTEREN, On the combination of independent two sample tests of Wilcoxon.(1960) Rapport SP 68 Stat.Afd.v.h. Math. Centrum.

T.C. FRY, Probability and its engineering uses, D. van Nostrand Company, New York 1928.

H.O. HARTLEY and E.S. PEARSON, Tables of the χ^2-integral and of the comulative Poisson distribution, Biometrika 37 (1950) 313-325.

A. VAN WIJNGAARDEN, Table of the cumulative symmetric binomial distribution, Proc. K.N.A.v.W. 53 (1950).