MEDEDELINGEN VAN DE LANDBOUWHOGESCHOOL TE WAGENINGEN/NEDERLAND
53 (2) 25–30 (1953)

OPEN PLAATSEN IN VARIANTIESCHEMA'S

(WITH A SUMMARY)

door

N. H. KUIPER en L. C. A. CORSTEN

(Afd. Wiskunde der Landbouwhogeschool)

(Ontvangen/Received 3.3.'53)

Gewone orthogonale schema's

Een experiment zij zodanig, dat men het te vinden experimentele resultaat hoopt uit te kunnen drukken in een schema van $m\ n$ getallen, dat bestaat uit m rijen en n kolommen. Elk getal uit het schema wordt opgevat als een steekproef uit een normale kansverdeling, met constante (maar niet tevoren bekende) variantie σ^2. Verder neemt men aan, dat het schema μ van verwachtingswaarden op de $m\ n$ plaatsen de som is van een gemiddeld niveau $\bar{\mu}$, een effect $\mu_A - \bar{\mu}$ van een invloed A, die binnen elke rij constant is, en een effect $\mu_B - \bar{\mu}$ van een invloed B, die binnen elke kolom constant is:

$$\mu = \bar{\mu} + (\mu_A - \bar{\mu}) + (\mu_B - \bar{\mu}).$$

($\bar{\mu}$ is het schema dat uit μ ontstaat door elk getal te vervangen door het gemiddelde van alle getallen, μ_A is het schema dat uit μ ontstaat door elk getal te vervangen door het gemiddelde in zijn rij, μ_B is het schema dat uit μ ontstaat door elk getal te vervangen door het gemiddelde in zijn kolom.)

$$\mu = \begin{pmatrix} \mu_{11} & \cdot & \cdot & \cdot & \mu_{1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \mu_{m1} & \cdot & \cdot & \cdot & \mu_{mn} \end{pmatrix}$$

Bij een experiment, dat goed gelukt, verkrijgt men een schema (genaamd x) van $m\ n$ getallen.

Goede schattingen over niveau en effecten zijn dan:

$$\bar{x},\ x_A - \bar{x},\ x_B - \bar{x}$$

terwijl x_T, in de vectorvergelijking:

$$x = \bar{x} + (x_A - \bar{x}) + (x_B - \bar{x}) + x_T,$$

onverwarde (not confounded) informatie over σ^2 kan geven: $x_T^2/(mn - m - n + 1)$ is een zuivere schatting over σ^2, terwijl x_T^2/σ^2 de kansverdeling heeft van χ^2 bij $(mn - m - n + 1)$ vrijheidsgraden of dimensies. De hypothese $\mu_A - \bar{\mu} = 0$ kan getoetst worden met de F-toets door het getal

$$\frac{(x_A - \bar{x})^2 / (m-1)}{x_T^2 / (mn - m - n + 1)}$$

te beschouwen, dat onder deze nulhypothese een steekproef uit de F- kansverdeling is bij $(m-1)$ en $(mn - m - n + 1)$ dimensies.

Open plaatsen (missing plots).

Wij beschouwen thans het geval dat het experiment op enige (weinige) plaatsen (die we voortaan de *open plaatsen* zullen noemen) geen getal geleverd heeft. Wij spreken af dat een schema van getallen, waarbij geen getallen aan de „open plaatsen" zijn toegevoegd, door een *hoofdletter* zal worden weergegeven. Een schema van *mn* getallen wordt door een *kleine letter* voorgesteld. Laat men in zo'n schema de getallen op de open plaatsen weg, dan zal het resultaat met dezelfde letter maar als hoofdletter genoteerd worden weergegeven.

We noemen het gevonden experimentele resultaat X. Het schema dat hieruit ontstaat door getallen nul op de open plaatsen in te vullen, noemen we x. Een schema, dat getallen nul heeft op de gevulde plaatsen, en voorlopig willekeurige getallen op de open plaatsen, zij y. Indien er r open plaatsen zijn, dan is y een willekeurige vector in een r-dimensionale ruimte (vergelijk het voorbeeld, waarin $r = 3$).

Wij beweren nu, dat bij het gegeven experimentele resultaat X (of x, wat hetzelfde betekent) de beste schatting u voor μ gegeven wordt door die keuze van y en u, waarvoor het vectorkwadraat

$$(x + y - u)^2 \qquad (1)$$

minimaal is, onder de nevenvoorwaarde van additiviteit:

$$u = \bar{u} + (u_A - \bar{u}) + (u_B - \bar{u}).$$

Immers: heeft men een willekeurige keuze voor u gedaan, dan zal (1) zo klein mogelijk zijn, indien men op de open plaatsen voor y juist de getallen uit het schema u kiest. De bijdrage in (1) van de met y gevulde plaatsen (die steeds niet-negatief is) is dan nul, en telt dus niet mee. (1) met y en u variabel (variabel binnen het kader aan deze variabelen toegestaan) heeft hetzelfde minimum als (1) minus de bijdragen van de met y gevulde plaatsen. Dit laatste minimum levert die schatting U voor M, waarbij de waarschijnlijkheidsdichtheid ter plaatse X zo groot mogelijk is („maximum likelihood estimate"; „kleinste kwadraten").

De optimale keuze voor u in (1) is, indien y bekend is, en $z = x + y$, gelijk aan (vergelijk de eerste paragraaf):

$$u = \bar{z} + (z_A - \bar{z}) + (z_B - \bar{z}).$$

De vector $x + y - u = z - u = (z - z_A - z_B + \bar{z})$ is bij variabele y zo klein mogelijk, indien de vector z – u op de open plaatsen nullen levert (zoals wij reeds zagen). Is e_τ een vector, die het getal 1 heeft op een bepaalde open plaats, en elders nullen, dan moet dus gelden (inwendig product):

$$e_\tau (z - z_A - z_B + \bar{z}) = 0, \text{ en, omdat } z = x + y,:$$

$$e_\tau (y - y_A - y_B + \bar{y}) = - e_\tau (x - x_A - x_B + \bar{x})$$

$$e_\tau (y - y_A - y_B + \bar{y}) = e_\tau (x_A + x_B - \bar{x}) \qquad (2)$$

Bij r open plaatsen vindt men zo r vergelijkingen ($\tau = 1, \ldots r$) in de onbekende getallen van het schema y. De oplossingen hiervan worden in de open plaatsen ingevuld.

(2) uitgeschreven voor de open plaats *jk* levert (vergelijk [1]):

$$y_{jk} - \frac{1}{n}\Sigma y_{js} - \frac{1}{m}\Sigma y_{tk} + \frac{1}{mn}\Sigma\Sigma y_{ts} = \frac{1}{n}\Sigma x_{js} + \frac{1}{m}\Sigma x_{tk} - \frac{1}{mn}\Sigma\Sigma x_{ts} \qquad (2')$$

(Sommaties over t en/of s), ook in de volgende vergelijkingen.)
Bij vermenigvuldiging met mn levert dit:

$$(m-1)(n-1)y_{jk} - (m-1)\sum_{s\neq k} y_{js} - (n-1)\sum_{t\neq j} y_{tk} + \sum_{t\neq j}\sum_{s\neq k} y_{ts} =$$
$$m\,\Sigma x_{js} + n\,\Sigma x_{tk} - \Sigma\Sigma x_{ts} \qquad (2'')$$

N.B. Indien de open plekken alleen in de j-de rij voorkomen, dan worden deze vergelijkingen:

$$(m-1)(n-1)y_{jk} - (m-1)\sum_{s\neq k} y_{js} = m\,\Sigma x_{js} + n\,\Sigma x_{tk} - \Sigma\Sigma x_{ts} \qquad (3)$$

Indien er alleen een open plek is op de plaats jk, dan is

$$(m-1)(n-1)y_{jk} = m\,\Sigma x_{js} + n\,\Sigma x_{tk} - \Sigma\Sigma x_{ts}. \qquad (4)$$

Bij het schema z, dat ontstaat door de gevonden getallen y op de open plaatsen in te vullen, kunnen $\bar{z}$, z_A en z_B gevonden worden, met behulp waarvan $(z-u)^2 = (z - z_A - z_B + \bar{z})^2 = z^2 - z_A^2 - z_B^2 + \bar{z}^2$ berekend kan worden.
$(z-u)^2$ is behalve het kwadraat van de lengte van de component in de toevalsruimte van z ook het minimum van (1), dat de „maximum likelihood estimate" U voor M bij het resultaat X levert. Anders gezegd: De vector U, die ligt in de $(m+n-1)$-dimensionale ruimte C opgespannen door de ruimten van de A- en de B- effecten, is zodanig dat de restvector X – U een minimale lengte heeft; m.a.w.: U is de loodrechte projectie van X op C. Bovendien is $(x-u)^2$ het kwadraat van de lengte van de component $X_T = X - U$ van X in de loodrecht op C staande toevalsruimte van dimensie $mn - r - m - n + 1$. Hieruit volgt de (onverdachte) schatting over σ^2: $X_T^2/(mn - m - n + 1 - r)$.
Omdat de ruimten der zuivere hoofdeffecten A^* en B^* bij een schema met „open plaatsen" niet onderling loodrecht zijn, verloopt de toetsing, b.v. van de nulhypothese: „invloed A heeft geen effect", anders dan gewoonlijk.

Onder deze hypothese geldt voor de verwachtings-vector M : $M = M_B$. (Als steeds stelt de index (B) de loodrechte projectie voor, die verkregen wordt door elk getal van het schema te vervangen door het gemiddelde in zijn klasse volgens klasse-indeling B). Onder de genoemde hypothese wordt een tweede zuivere schatting over σ^2 gegeven door:

$$(m-1)\,S(\sigma^2) = (U - U_B)^2 = U^2 - U_B^2 = (X^2 - X_T^2) - X_B^2 =$$
$$X^2 - (z^2 - z_A^2 - z_B^2 + \bar{z}^2) - X_B^2.$$

Hierin is X_B een vector die uit X ontstaat door in elke kolom de getallen door het gemiddelde in die kolom te vervangen. Noemen wij de som der n_s getallen in de s-de kolom: B_s, dan is $X_B^2 = \sum_s n_s \left(\frac{B_s}{n_s}\right)^2 = \Sigma\left(\frac{B_s^2}{n_s}\right)$

Het toetsen van het effect van invloed A geschiedt nu door in de F-tabel te vergelijken het getal (dimensies $m-1$, $mn-n-n+1-r$):

$$F = \frac{(X^2 - [z^2 - z_A^2 - z_B^2 + \bar{z}^2] - X_B^2)\,/\,m-1}{(z^2 - z_A^2 - z_B^2 + \bar{z}^2)\,/\,mn-r-m-n+1} \qquad (5)$$

In plaats van deze werkwijze gaat men in de practijk vaak als volgt te werk: Men gaat uit van het schema z, waarin op de open plaatsen op de boven aan-

gegeven wijze getallen (y) ingevuld zijn. Men vult nu in plaats van de teller in (5) het volgende getal in: $(z_A^2 - \bar{z}^2)/(m-1)$. Deze werkwijze, die op een analogie redenering berust en exact fout is, kan soms een voldoend goede benadering zijn. Ten einde de aard van het verschil te onderzoeken berekenen we het verschil van de twee genoemde tellers [een factor $(m-1)$ weglatend]. Dit verschil is:

$$z_A^2 - \bar{z}^2 - [X^2 - (z^2 - z_A^2 - z_B^2 + \bar{z}^2) - X_B^2] = z^2 - z_B^2 - X^2 + X_B^2 = (z - z_B)^2 - (X - X_B)^2.$$

De som van kwadraten die door het eerste vectorkwadraat wordt voorgesteld kan gesplitst worden in een deel dat de bijdrage der niet-open plaatsen is, en in de bijdrage der open plaatsen. Beide delen zijn niet-negatief. Het eerste deel kan worden voorgesteld door $(X - X_{*_B})^2$. Hierin is X_{*_B} het schema, dat uit X ontstaat door de getallen in een kolom van X te vervangen door het gemiddelde van de getallen in de overeenkomstige kolom van z; het is ook het schema, dat uit z_B ontstaat door de getallen op de open plaatsen weg te laten.

Zowel de vector X_B als de vector X_{*_B} liggen in de ruimte der hoofdeffecten B. X_B is echter de *loodrechte* projectie van X op deze ruimte, en dus is

$$(z - z_B)^2 - (X - X_B)^2 \geq (X - X_{*_B})^2 - (X - X_B)^2 \geq 0$$

Bij de foutieve werkwijze is de teller in (5), dus ook het voor F gevonden getal, niet kleiner dan bij de goede werkwijze. Dit heeft tengevolge dat, bij gegeven onbetrouwbaarheidsdrempel, indien de goede werkwijze tot de conclusie „de nul-hypothese wordt verworpen" leidt, de foutieve werkwijze zeker ook tot deze conclusie zal leiden. Leidt de foutieve werkwijze tot verwerpen, dan behoeft de goede werkwijze dit nog niet te doen (zie het voorbeeld). Leidt de foutieve werkwijze niet tot verwerpen, dan de goede werkwijze zeker niet.

Voorbeeld ter illustratie.

	b_1	b_2	b_3	b_4	$m = 3,\ n = 4.$
a_1	460	518	524	498	
a_2	y_{21}	363	y_{23}	377	
a_3	y_{31}	349	356	355	

Vergelijkingen voor het invullen der open plaatsen:

Plaats 21 : $6y_{21} - 2y_{23} - 3y_{31} = 3(363 + 377) + 4(460) - 3800$
Plaats 23 : $6y_{23} - 2y_{21} + y_{31} = 3(363 + 377) + 4(524 + 356) - 3800$
Plaats 31 : $6y_{31} - 3y_{21} + y_{23} = 3(349 + 356 + 355) + 4(460) - 3800$

$$\begin{aligned} 6y_{21} - 2y_{23} - 3y_{31} &= 260 \\ -2y_{21} + 6y_{23} + y_{31} &= 1940 \\ -3y_{21} + y_{23} + 6y_{31} &= 1220 \end{aligned}$$

$$y_{21} = 320, \quad y_{23} = 380, \quad y_{31} = 300.$$

Ingevuld:

460	518	524	498	2000
320	363	380	377	1440 = 740 + 700
300	349	356	355	1360 = 1060 + 300
1080	1230	1260	1230	4800
= 460 + 620		= 880 + 380		

Men vindt:

$$\bar{z} = \begin{pmatrix} 400 \\ \end{pmatrix} ; \; z_A = \begin{pmatrix} 500 \\ 360 \\ 340 \end{pmatrix} ; \; z_A - \bar{z} = \begin{pmatrix} 100 \\ -40 \\ -60 \end{pmatrix};$$

$$z_B - \bar{z} = \begin{pmatrix} -40 & 10 & 20 & 10 \end{pmatrix}$$

$$z - z_A - z_B + \bar{z} = \begin{pmatrix} 0 & 8 & 4 & -12 \\ 0 & -7 & 0 & 7 \\ 0 & -1 & -4 & 5 \end{pmatrix}$$

$$X_T^2 = z^2 - z_A^2 - z_B^2 + \bar{z}^2 = (z - z_A - z_B + \bar{z})^2 = 364$$

$$X^2 = 1650964$$

$$X_A^2 = \frac{2000^2}{4} + \frac{740^2}{2} + \frac{1060^2}{3} = 1648333$$

$$X_B^2 = \frac{460^2}{1} + \frac{1230^2}{3} + \frac{880^2}{2} + \frac{1230^2}{2} = 1607400.$$

Toetsing van de nulhypothese $M_{B*} = M_B - \overline{M} = 0$ op de goede wijze levert een steekproef van F bij 3 en 3 dimensies:

$$F = \frac{(X^2 - X_T^2 - X_A^2)/3}{X_T^2/3} = \frac{2267}{364} = 6{,}23.$$

Hadden wij dezelfde hypothese formeel getoetst aan het opgevulde schema, dan zouden wij een „steekproef van F" eveneens bij 3 en 3 dimensies verkregen hebben: $F = \frac{(z_B - \bar{z})^2/3}{X_T^2/3} = \frac{6600}{364} = 20{,}4$ terwijl $P\,(F > 9{,}28/3;3) = 0{,}05$.

Goed toetsen leidt hier niet tot verwerpen der nul-hypothese, foutief toetsen wel.

Wij vermelden nog de verdeling der dimensies voor dit geval:

niveau	1
A-effect (zuiver)	2
B-effect (zuiver)	3
toeval	3
aantal gevulde plaatsen	9

SUMMARY

A well-known missing-plot technique for two-way classifications is considered.

LITERATUURVERWIJZINGEN

1) Voor de formules (2*) en (3) vergelijk:

a. COCHRAN, W. G. and G. M. COX: Experimental Designs, New York, 1950.
b. PATERSON, D. D.: Statistical Technique in Agricultural Research, New York, London, 1939, p. 182.
c. SNEDECOR, G. W.: Statistical Methods, Ames Iowa, U.S.A., 1938, p. 223.
d. YATES, F.: *The Empire Journal of Experimental Agriculture*, **1**, 129 (1933).
e. FEDERER, W. T.: Iowa Agricultural Experiment Station Research Bulletin no. 380 (1951) p. 277.

2) Voor de toepassing van vectoren in de variantie-analyse, vergelijk:

a. KUIPER, N. H.: Variantie-analyse, *Statistica* **6** (3) 149 – 194 (1952)
b. ————————: Analyse van vectoren en variantie. *Biometrisch Contact* **5**, 22 – 35 (1953).
c. ————————: On 4^3 – factorial designs, *Netherlands Journal of Agricultural Science*. **1**, 11–14 (1953).