MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

S 160 (M 60)

Schattingsmethoden voor overlevingskansen

III Een schattingsmethode met behulp van regressie.

door

Gerda Klerk - Grobben

1955

1. Inleiding.[1)]

De in dit memorandum beschreven schattingsmethode wordt voorgesteld door D.G.CHAPMAN: The estimation of biological populations (Annals of Math.Stat.,25 (1954), paragraaf 4).

Evenals in memorandum S 160 (M 57) zullen we ook nu een samenvoeging gebruiken van het waarnemingsmateriaal dat in het eerste memorandum in deze serie, S 160 (M 54), wordt beschreven. De samenvoeging is nu echter een andere dan in S 160 (M 57) werd behandeld: inplaats van de aantallen m_{ij} (het aantal van de in jaar j gevangen dieren, die in jaar i voor het laatst gevangen werden) zullen we nu uitgaan van de aantallen n_{ij} , die als volgt worden gedefinieerd:

$n_{ij} \overset{\text{def}}{=}$ het aantal van de in jaar j gevangen dieren, die in jaar i voor het eerst gevangen werden.

Bij de bepaling van n_{ij} letten we er dus niet op hoevaak en wanneer de betreffende dieren nog tussen de jaren i en j gevangen werden. Uitgedrukt in de oorspronkelijke aantallen $z_{j_1 \cdots j_k j}$ (het aantal in jaar j gevangen dieren dat ook in de jaren $j_1 \cdots j_k$ gevangen werd) in het waarnemingsmateriaal is dus b.v.

$$n_{25} = z_{25} + z_{235} + z_{245} + z_{2345}.$$

De onderstellingen over de populatie en de vangmethode zijn ook nu weer:

1) de overleveringskans, θ , is in ieder jaar dezelfde en voor alle dieren gelijk,
2) er verdwijnen geen dieren door emigratie, en
3) de vangkans is per jaar voor ieder dier gelijk (q_j), maar mag van jaar tot jaar variëren.

2. Relaties tussen θ en de gevonden aantallen n_{ij} .

Uit de onderstellingen volgt, dat voor een dier uit de beschouwde populatie de kans om de periode van jaar i tot en met jaar j te overleven gelijk aan θ^{j-i} zal zijn. Daar de vangkans in jaar j q_j is zal de kans dat een dier, dat in jaar i werd gevangen, ook in jaar j tot de vangst zal behoren gelijk zijn aan $q_j \theta^{j-i}$.

We beperken ons tot die dieren die in jaar i voor het

[1)] Dit memorandum dient slechts ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.

eerst gevangen werden om te voorkomen dat de dieren in de volgende vangjaren elk tot meerdere groepen per jaar bijdragen en zo althans een deel van de mogelijke afhankelijkheden in het waarnemingsmateriaal te vermijden.

Uit het voorgaande volgt nu dus dat het aantal n_{ij} een trekking uit een binomiale (n, p)-verdeling zal zijn met parameters $n = u_i$ (alle in jaar i voor het eerst gevangen dieren) en $p = g_j \theta^{j-i}$. In plaats van g_j kunnen we ook gebruiken R_j/N_j (waarin R_j de totale vangst en N_j de populatiegrootte in jaar j voorstelt). Zowel g_j als N_j zijn onbekend. Gebruiken we $g_j \theta^{j-i}$ dan vinden we schattingen voor θ en g_j ; gebruiken we $\frac{\theta^{j-i} R_i}{N_j}$ dan schattingen voor θ en N_j (en eventueel hieruit $g_j = R_j/N_j$). We zullen dit laatste toepassen. Voor de verwachting van het aantal $\underline{n}_{ij}$ [2] (onder voorwaarde van de totale vangst R_j) geldt nu:

$$(1) \qquad \mathcal{E}\, \underline{n}_{ij} = \frac{u_i \theta^{j-i} R_j}{N_j} .$$

Hierin zijn dus θ en N_j de onbekende parameters, waarvoor uit de gevonden aantallen n_{ij} schattingen zullen worden afgeleid.

3. Kleinste kwadraten schattingen voor $\log \theta$ en $\log N_j$.

D.G.CHAPMAN gaat nu, om een regressie lijn aan te passen, uit van de onderstelling:

$$(2) \qquad \mathcal{E}\{\log(\underline{n}_{ij} + 1)\} = (j-i)\log\theta + \log u_i + \log R_j - \log N_j .$$

op grond van de volgende overwegingen.

---

[2] Dat een grootheid een stochastisch karakter bezit (een waarschijnlijkheidsverdeling bezit) geven we aan door onderstreping van het symbool. Een door de stochastische grootheid aangenomen waarde wordt met hetzelfde symbool zonder onderstreping aangegeven.

1) De overgang op logarithmen geeft een lineair regressieprobleem. Theoretisch kan dit bovendien enigszins gerechtvaardigd worden doordat
2) de variantie van de logarithme van een Poisson verdeelde grootheid $\underline{x}$ constant is tot op termen van de orde $(\mathcal{E}\underline{x})^{-1}$ en een binomiale verdeling (waarvan het totale aantal vrij groot en de kans vrij klein is) met een Poisson verdeling benaderd kan worden zodat de overgang op logarithmen ook in ons geval wellicht een in eerste benadering constante variantie zal geven (zoals bij regressie analyse zeer gewenst is).
3) het gebruik van $(\underline{n}_{ij}+1)$ in plaats van $\underline{n}_{ij}$ heeft het voordeel dat de eventueel optredende aantallen $n_{ij}=0$ bij het nemen van de logarithme geen moeilijkheden geven. Theoretisch kan dit verdedigd worden (hoewel zwak) doordat
4) voor de logarithme geldt:

$$\mathcal{E}\log \underline{x} < \log \mathcal{E}\underline{x}$$

met als gevolg dat de overschatting van $\mathcal{E}\underline{x}$ door $\underline{x}+1$ in plaats van $\underline{x}$ te nemen enigszins gecompenseerd kan worden door de onderschatting bij het nemen van de logarithme. (Het resultaat blijft een overschatting).

In het volgende zal blijken dat het eenvoudiger is in (2) $j-i$ te vervangen door k. (2) gaat dan over in

(3) $$\mathcal{E}\log(\underline{n}_{j-k,j}+1) - \log u_{j-k} - \log R_j = k\log\theta - \log N_j$$

waarin $k=1,\cdots,j$ en $j=1,\cdots,t$.
We vereenvoudigen deze betrekkingen nog door te substitueren:

$$y_{kj} \overset{\text{def}}{=} \log(\underline{n}_{j-k,j}+1) - \log u_{j-k} - \log R_j,$$
$$\beta \overset{\text{def}}{=} \log\theta \quad \text{en}$$
$$\alpha_j \overset{\text{def}}{=} \log N_j.$$

waardoor ontstaat

(4) $$\mathcal{E}\,\underline{y}_{kj} = k\beta - \alpha_j \quad ; k=1,\cdots,j \; ; j=1,\cdots,t$$

met als onbekende parameters β en α_j. Bv. als $t=3$ dan

vinden we de betrekkingen:

$\mathcal{E}\, \underline{y}_{11} = \beta - \alpha_1$; $\underline{y}_{11}$ bepaald door $\underline{n}_{01}$, u_0 en R_1,
$\mathcal{E}\, \underline{y}_{12} = \beta - \alpha_2$; $\underline{y}_{12}$ bepaald door $\underline{n}_{12}$, u_1 en R_2,
$\mathcal{E}\, \underline{y}_{22} = 2\beta - \alpha_2$; $\underline{y}_{22}$ bepaald door $\underline{n}_{02}$, u_0 en R_2,

$\mathcal{E}\, \underline{y}_{13} = \beta - \alpha_3$; $\underline{y}_{13}$ bepaald door $\underline{n}_{23}$, u_2 en R_3,
$\mathcal{E}\, \underline{y}_{23} = 2\beta - \alpha_3$; $\underline{y}_{23}$ bepaald door $\underline{n}_{13}$, u_1 en R_3,
$\mathcal{E}\, \underline{y}_{33} = 3\beta - \alpha_3$; $\underline{y}_{33}$ bepaald door $\underline{n}_{03}$, u_0 en R_3.

Om uit de betrekkingen (4) de onbekende parameters te schatten passen we de methode der kleinste kwadraten toe, d.w.z. we bepalen die schattingen $\hat{\beta}, \hat{\alpha}_1 \cdots \hat{\alpha}_t$ voor de parameters die de kwadratische vorm:

$$\sum_{j=1}^{t} \sum_{k=1}^{j} (y_{kj} - k\beta + \alpha_j)^2 = \varphi$$

minimaal maken. Voor deze schattingen geldt dan:

$$\left[\frac{\partial \varphi}{\partial \alpha_j}\right]_{\substack{\alpha_j = \hat{\alpha}_j \\ \beta = \hat{\beta}}} = 0 \quad \text{en} \quad \left[\frac{\partial \varphi}{\partial \beta}\right]_{\substack{\alpha_j = \hat{\alpha}_j \\ \beta = \hat{\beta}}} = 0 ,$$

dus:

$$\sum_{k=1}^{j} (y_{kj} + \hat{\alpha}_j - k\hat{\beta}) = 0 \quad , (j = 1, \cdots, t),$$

en:

$$\sum_{j=1}^{t} \sum_{k=1}^{j} k(y_{kj} + \hat{\alpha}_j - k\hat{\beta}) = 0 .$$

Lossen we deze vergelijkingen op dan ontstaat:

(5)
$$\hat{\alpha}_j = \tfrac{1}{2}(j+1)\hat{\beta} - \frac{1}{j} \sum_{k=1}^{j} y_{kj} \quad , (j = 1, \cdots, t),$$

$$\hat{\beta} = \frac{\sum_{j=2}^{t} \sum_{k=1}^{j} k\, y_{kj} - \sum_{j=2}^{t} \frac{1}{2}(j+1) \sum_{k=1}^{j} y_{kj}}{\frac{1}{48} t(t^2-1)(t+2)}$$

In de uitdrukking voor $\hat{\beta}$ beginnen de sommaties over j bij $j = 2$ omdat voor $j = 1$ geldt:

$$1 \cdot y_{11} - \tfrac{2}{2} y_{11} = 0 .$$

Met de betrekkingen

$$\beta = \log \theta \quad \text{en} \quad \alpha_j = \log N_j$$

zijn dan schattingen $\hat{\theta}$ en $\hat{N}_j$ voor θ en N_j te vinden. Over de zuiverheid van deze schattingen is niet veel te zeggen. Gezien de opmerkingen onder 4) in paragraaf 3 zullen ze zeker niet beide zuiver zijn.

4. Rekenschema en voorbeelden.

De berekeningen die nodig zijn voor de bepaling van $\hat{\beta}$ en $\hat{\alpha}_j$ zijn ondergebracht in het rekenschema I, op blz. 6. Het eerste gedeelte hiervan zijn de uit de waarnemingen gevonden getallen n_{ij}, het tweede deel geeft de bepaling van y_{jk}, terwijl het derde deel de voor $\hat{\beta}$ en $\hat{\alpha}_j$ benodigde sommen geeft.

Als voorbeeld geven we de bepaling van $\hat{\theta}$ en $\hat{N}_j$ uit de resultaten van twee steekproefexperimenten uitgevoerd door onze afdeling. Hierbij wordt ondersteld dat in 5 achtereenvolgende jaren ($t=4$) tellingen werden verricht aan een populatie (bestaande uit $\sigma^{\prime}\sigma^{\prime}$ en $\varphi\varphi$) met verschillende overlevingskansen voor $\sigma^{\prime}\sigma^{\prime}$ en $\varphi\varphi$. Deze waren respectievelijk voor $\sigma^{\prime}\sigma^{\prime}$: $\theta = 0{,}8$ en voor $\varphi\varphi$: $\theta = 0{,}7$. De ware populatiegroottes waren:

voor $\sigma^{\prime}\sigma^{\prime}$: $N_0 = 600$, $N_1 = 594$, $N_2 = 590$, $N_3 = 573$, $N_4 = 591$
en voor $\varphi\varphi$: $N_0 = 400$, $N_1 = 397$, $N_2 = 403$, $N_3 = 395$, $N_4 = 381$.

(De populaties werden door de "geboortes" weer op peil gebracht.)
In de rekenschema's I en II, welke opgebouwd zijn als Schema I zijn de schattingen $\hat{\theta}$ en $\hat{N}_j$ berekend.

Schema I. Berekeningen nodig bij de bepaling van $\hat{\beta}$ en $\hat{\alpha}_j$

jaar j	0	1	2	3	4	
1° vangst totale vangst	u_0	u_1 R_1	u_2 R_2	u_3 R_3	u_4 R_4	
		n_{01}	n_{12} n_{02}	n_{23} n_{13} n_{03}	n_{34} n_{24} n_{14} n_{04}	
$k\downarrow$	$\log u_0$	$\log u_1$ $\log R_1$	$\log u_2$ $\log R_2$	$\log u_3$ $\log R_3$	$\log u_4$ $\log R_4$	
1 2 3 4		$\log(n_{01}+1)$	$\log(n_{12}+1)$ $\log(n_{02}+1)$	$\log(n_{23}+1)$ $\log(n_{13}+1)$ $\log(n_{03}+1)$	$\log(n_{34}+1)$ $\log(n_{24}+1)$ $\log(n_{14}+1)$ $\log(n_{04}+1)$	
1 2 3 4		$y_{11} = \log(n_{01}+1) - \log u_0 - \log R_1$	$y_{12} = \log(n_{12}+1) - \log u_1 - \log R_2$ $y_{22} = \log(n_{02}+1) - \log u_0 - \log R_2$	y_{13} y_{23} y_{33}	y_{14} y_{24} y_{34} y_{44}	
			$\sum_{k=1}^{2} y_{k2}$ $\sum_{k=1}^{2} k\, y_{k2}$	$\sum_k y_{k3}$ $\sum_k k y_{k3}$	$\sum_k y_{k4}$ $\sum_k k y_{k4}$	$\sum_{j=2}^{t} \frac{j+1}{2} \sum_k y_{kj}$ $\sum_{j=2}^{t} \sum_k k y_{kj}$

Schema II. Berekening van $\hat{\theta}$ en $\hat{N}_j$ bij ♂♂.

jaar(j)		0	1	2	3	4
k	u_j / R_j	116 116	109 128	96 127	61 100	88 125
1	$n_{j-k,j}$		19	17	14	8
2				14	14	8
3					11	13
4						8
	$\log u_j$ $\log R_j$	2,0645 2,0645	2,0374 2,1072	1,9823 2,1038	1,7853 2,0000	1,9445 2,0969
1	$\log(n_{j-k,j}+1)$		1,3010	1,2553	1,1761	0,9542
2				1,1761	1,1761	0,9542
3					1,0792	1,1461
4						0,9542
1	$y_{k,j}$		0,1293-3	0,1141-3	0,1938-3	0,0720-3
2				0,0078-3	0,1387-3	0,8750-4
3					0,0147-3	0,0113-3
4						0,7928-4
	Σy_{kj}		0,1293-3	0,1219-6	0,3472-9	0,7516-13
	$\Sigma k\, y_{kj}$		0,1293-3	0,1297-9	0,5153-18	0,0286-31

$t=4 \rightarrow \frac{1}{48}t(t^2-1)(t+2) = 7,5$

$$\sum_{j=2}^{4} \frac{j+1}{2} \Sigma y_{kj} = 0,2563\text{-}57$$

$$\sum_{j=2}^{4} \Sigma k\, y_{kj} = 0,6736\text{-}58$$

$$\hat{\beta} = \frac{0,6736\text{-}58-(0,2563\text{-}57)}{7,5} = 0,9250\text{-}1\,, \qquad \hat{\theta} = 0,84 \quad (\theta = 0,8)$$

$\hat{\alpha}_1 = 0,9222\text{-}1-(0,1293\text{-}3) = 2,7929,$ $\hat{N}_1 \simeq 620$ $(N_1 = 594)$

$\hat{\alpha}_2 = \frac{3}{2}(0,9222\text{-}1)-\frac{1}{2}(0,1219\text{-}6) = 2,8223,$ $\hat{N}_2 \simeq 660$ $(N_2 = 590)$

$\hat{\alpha}_3 = \frac{4}{2}(0,9222\text{-}1)-\frac{1}{3}(0,3472\text{-}9) = 2,7287,$ $\hat{N}_3 \simeq 540$ $(N_3 = 573)$

$\hat{\alpha}_4 = \frac{5}{2}(0,9222\text{-}1)-\frac{1}{4}(0,7516\text{-}13) = 2,8676,$ $\hat{N}_4 \simeq 740$ $(N_4 = 591)$

Schema III. Berekening $\hat{\theta}$ en $\hat{N}_j$ bij ♀♀.

jaar (j)		0	1	2	3	4
k ↓	u_j	89	54	58	61	61
	R_j	89	64	64	76	85
1	$n_{j-k,j}$		10	1	10	10
2				5	3	4
3					2	6
4						4
	$\log u_j$	1,9494	1,7324	1,7634	1,7853	1,7853
	$\log R_j$	1,9494	1,8062	1,8062	1,8808	1,9294
1	$\log(n_{j-k,j}+1)$		1,0414	0,3010	1,0414	1,0414
2				0,7782	0,6021	0,6990
3					0,4771	0,8451
4						0,6990
1	y_{kj}		0,2858-3	0,7624-4	0,3972-3	0,3267-3
2				0,0226-3	0,9889-4	0,0062-3
3					0,6469-4	0,1833-3
4						0,8202-4
	Σy_{kj}		0,2858-3	0,7850-7	0,0330-9	0,3364-12
	$\Sigma k y_{kj}$		0,2858-3	0,8076-10	0,3157-19	0,1698-30

$t=4 \rightarrow \frac{1}{48}t(t^2-1)(t+2) = 7,5$

$\sum_{j=2}^{4} \frac{j+1}{2} \Sigma y_{kj} = 0,5845-57$

$\sum_{j=2}^{4} \Sigma k\, y_{kj} = 0,2931-58$

$\hat{\beta} = \frac{0,2931-58-(0,5845-57)}{7,5} = 0,8278-1, \quad \hat{\theta} = 0,67 \quad (\theta = 0,7)$

$\hat{\alpha}_1 = 0,8278-1-(0,2858-3) = 2,5420, \quad \hat{N}_1 \simeq 350 \quad (N_1 = 397)$

$\hat{\alpha}_2 = \frac{3}{2}(0,8278-1)-\frac{1}{2}(0,7850-7) = 2,8492, \quad \hat{N}_2 \simeq 710 \quad (N_2 = 403)$

$\hat{\alpha}_3 = \frac{4}{2}(0,8278-1)-\frac{1}{3}(0,0330-9) = 2,6446, \quad \hat{N}_3 \simeq 440 \quad (N_3 = 395)$

$\hat{\alpha}_4 = \frac{5}{2}(0,8278-1)-\frac{1}{4}(0,3364-12) = 2,4854 \quad \hat{N}_4 \simeq 310 \quad (N_4 = 381)$