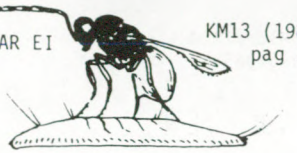
KWANTITATIEVE METHODEN BIJ: DE SLUIPWESP EN HAAR EI
(pps-trekking zonder teruglegging)

KM13 (1984)
pag 186-190



M.A.J. van Montfort *

0. Inleiding

Bij statistisch advies werd onderstaand probleem ontmoet, dat net iets moeilijker is dan simpel. Het probleem en de KM-bijdrage worden hier behandeld.

Een sluipwesp beweegt zich in een omgeving van larven van twee stadia. Van het ene stadium zijn er N_0 stuks ieder met grootte s_0, en van het andere stadium zijn er N_1 stuks ieder met grootte s_1. Bekend is dat de ontmoetingskans evenredig is met de grootte. De sluipwesp kan in een larve al/niet precies 1 ei leggen (nooit meer). Na het leggen van in totaal n eieren telt men het aantal (k) larven uit N_0 met één ei. De vraag is nu of de verdeling van de n eieren over de twee stadia slechts bepaald is door de trefkans evenredig met de groottes, óf dat er daarnaast nog sprake is van voorkeur voor een van beide stadia. De voorkeur kan een combinatie zijn van zich gericht bewegen of bij aankomst voorkeur tonen met betrekking tot al/niet eileggen.

Het toetsingsprobleem kan statistisch als volgt geformuleerd worden: de trefkansen zijn evenredig met s_0 en s_1 onder H_0 (geen voorkeur) en evenredig met s_0 en θs_1 met $\theta \neq 1$ onder H_a (wel voorkeur).

1. De hypergeometrische verdeling

De hypergeometrische kansvariabele k ontstaat door uit een populatie met N_0 0-elementen en N_1 1-elementen achtereenvolgens n stuks te trekken zonder teruglegging (zt), waarbij bij iedere trekking alle nog niet getrokken elementen gelijke kans hebben om getrokken te worden, en het aantal 0-elementen in de steekproef te tellen (en te noteren als k).

Aantallen-tabel (en notatie):

soort	0	1	som
steekproef	k	n-k	n
niet-steekproef	N_0-k	N_1-n+k	N-n
beginpopulatie	N_0	N_1	N

* Vakgroep Wiskunde LH, De Dreijen 8, 6703 BC Wageningen
Tel. 08370-82384

Dan geldt

$$P\{\underline{k} = k\} = \binom{n}{k}\binom{N-n}{N_0-k} \Big/ \binom{N}{N_0} \quad \text{voor } k = \alpha, \alpha+1, \ldots, \omega \qquad (1)$$

met $\alpha = \max(0, n-N_1)$

$\omega = \min(n, N_0)$

$E(\underline{k}) = n\, N_0/N$

$\text{var}(\underline{k}) = N_0 N_1\, n(N-n)/\{N^2(N-1)\}$

2. Verdeling bij pps-zt-trekking

Als variant op bovenvermeld trekkingsschema kan het onderstreepte vervangen worden door "een kans hebben om getrokken te worden evenredig aan hun grootte" (pps: proportional to size).
Omdat dan slechts de verhouding van de groottes van de 0- en de 1-elementen van belang is, wordt de grootte van de 0-elementen op 1 gesteld ($s_0 = 1$); de grootte van de 1-elementen wordt genoteerd als s_1.

soort	grootte	aantal in begin pop.	steekproef
0	1	N_0	k
1	s_1	N_1	
		$N_1 + N_2$	n

Bij iedere trekkingsvolgorde ($\pi_{n,k}$) is direct de bijbehorende kans op te schrijven; bijv.

$$P\{\overbrace{0 \ldots 0}^{k}\ \overbrace{1 \ldots 1}^{n-k}\} = P\{\underline{k} = k \mid \pi_{n,k} = 0\ldots 01\ldots 1\} =$$

$$= \frac{N_0}{N_0 + N_1\,{}_1} \cdot \frac{N_0 - 1}{(N_0 - 1) + N_1\,{}_1} \cdot \ldots \cdot \frac{(N_0 + 1 - k)}{(N_0 + 1 - k) + N_1\,{}_1} \cdot$$

$$\cdot \frac{N_1\,{}_1}{N_1\,{}_1 + (N_0 - k)} \cdot \frac{(N_1 - 1)\,{}_1}{(N_1 - 1)\,{}_1 + (N_0 - k)} \cdot \ldots \cdot \frac{(N_1 + 1 - n + k)\,{}_1}{(N_1 + 1 - n + k)\,{}_1 + (N_0 - k)}$$

Merk op dat de teller volgorde-ongevoelig is, en dat de noemer volgorde-gevoelig is.

Iedere trekkingsvolgorde $\pi_{n,k}$ is een permutatie van k nullen en (n-k) enen. Er zijn $\binom{n}{k}$ verschillende permutaties, tesamen vormend de verzameling $\Pi_{n,k}$. Dan geldt

$$P\{\underline{k} = k\} = \sum_{\pi_{n,k} \in \Pi_{n,k}} P\{\underline{k} = k | \pi_{n,k}\} \quad \text{voor} \quad \alpha \leq k \leq \omega \qquad (2)$$

Berekening vereist dan een algorithme, dat alle $\binom{n}{k}$ permutaties geeft, zie bijv. Ausgewählte Operations Research-Algorithmen in FORTRAN, door H. Späth (R. Oldenbourg Verlag München Wien 1975). Dit boek bevat op pag. 58,59 de gewenste subroutine genaamd COMB.

Enige opmerkingen terzijde:

*1) $E(\underline{k}) \neq \frac{N_0}{N_0 + N_1 s_1} n$ mits $s_1 \neq 1$

hetgeen direct duidelijk is, door $N = N_1 + N_2$ te kiezen als waarde van n.

*2) Ook is het mogelijk s_1 te schatten bij gegeven realisatie k m.b.v. de Maximum Likelihood methode (ML). De likelihood(L) wordt door (2) gegeven, met als speciaal geval (1) voor $s_1 = 1$.

*3) Bij toepassing van (2) kan het aantal permutaties erg groot worden. Merk op dat $\binom{n}{k}$ maximaal $\binom{n}{n/2}$ is, en dat voor grote n met de formule van Stirling voor $\binom{n}{n/2}$ gevonden wordt $\frac{1}{\sqrt{2\pi}} \frac{2^{n+1}}{\sqrt{n}}$.

3. Een snel algorithme

Een andere (recursieve) berekeningswijze (met rekenwerk evenredig met n^2) is de volgende.

Laat $P_n(k)$ de kans zijn op k nullen in een steekproef van omvang n bij pps-zt-trekking bij een begintoestand met N_0, N_1, S_1.

Dan geldt voor n = 1

$$\begin{cases} P_1(1) = \frac{N_0}{N_0 + N_1 s_1} & (3, 1a) \\ P_1(0) = 1 - P_1(1) & (3, 1b) \end{cases}$$

Bij uitbreiding van de steekproef met 1 trekking kan k constant blijven (bij trekking van een 1) of 1 toenemen (bij trekking van een 0), uiteraard mits nog een resp. 1 en 0 resteert. Dit leidt tot

$$P_n(k) = P_{n-1}(k-1)\ Pr(0) + P_{n-1}(k).Pr(1), \quad \alpha \leq k \leq \omega \qquad (3,2)$$

In de volgende tabel wordt Pr(i) afgeleid

soort	grootte	aantal begin	aantal na n - 1 trekkingen	aantal restant	volgend resultaat	
0	1	N_0	k-1	N_0-k+1	1	$Pr(0) = \frac{N_0-k+1}{N_0-k+1+(N_1+k-n)\,s_1}$
1	s_1	N_1	n-k	N_1+k-n	0	
0	1	N_0	k	N_0-k	0	$Pr(1) = \frac{(N_1+k-n+1)s_1}{(N_1+k-n+1)s_1+(N_0-k)}$
1	s_1	N_1	n-k-1	N_1+k-n+1	1	

Randproblemen in (3.2) worden vermeden door te definiëren

$$P_n(-1) = 0 \quad \text{met} \quad n = 1,\ldots,N_1-1$$

$$P_n(n+1) = 0 \quad \text{met} \quad n = 1,\ldots,N_0-1$$

Bij gegeven $P_n(k)$ berekent men de verwachting (μ) en de variantie (σ^2) van $\underline{k}$ met behulp van

$$\mu = \sum_{k=\alpha}^{\omega} k \,.\, P_n(k)$$

$$\sigma^2 = -\mu^2 + \sum_{k=\alpha}^{\omega} k^2 \,.\, P_n(k)$$

Als een voorbeeld worden de (afgeronde) kansen voor $N_0 = 4$, $N_1 = 6$, $s_1 = 2$ gegeven voor alle mogelijke steekproefomvangen n.

k	n=1	2	3	4	5	6	7	8	9	10
0	.750	.536	.357	.214	.107	.035				
1	.250	.414	.497	.504	.443	.325	.166			
2		.050	.138	.251	.369	.467	.506	.419		
3			.007	.030	.077	.161	.293	.487	.744	
4				.001	.003	.012	.035	.093	.256	1
μ	.250	.514	.796	1.10	1.43	1.79	2.20	2.67	3.25	4
σ	.433	.591	.695	.761	.795	.794	.748	.638	.437	0

De eerste kansen-kolom is berekend met formule (3.1), en ieder volgende kolom is berekend uit de voorgaande met formule (3.2). Let op de lege plaatsen en de verdeling bij n = 10 (= $N_0 + N_1$).

4. Combineren van de resultaten van verschillende proeven

Notatie: $\underline{k}_j$: waar te nemen aantal in proef j (j = 1,...,m)

μ_j: $E_0(\underline{k}_j)$, d.i. de verwachting onder de nulhypothese (H_0)

σ_j: $\sqrt{var_0(\underline{k}_j)}$, d.i. de standaardafwijking onder H_0

De gecombineerde toetsingsgrootheid (voor één en dezelfde H_0 bij één en hetzelfde alternatief) kan er als volgt uitzien:

$$\underline{T} = \sum_{j=1}^{m} g_j(\underline{k}_j - \mu_j)/\sigma_j$$

met als verdeling onder H_0 bij benadering (CLS): $N(0, \Sigma\, g_j^2)$; hierin moeten de gewichten g nog gekozen worden. Als de drager van $\underline{k}_j$ weinig waarden bevat, is $\omega_j - \alpha_j$ en ook σ_j klein; in deze situatie bevat het steekproefresultaat nauwelijks enige informatie over voorkeur. Het ligt dan ook voor de hand g_j evenredig met ($\omega_j - \alpha_j$ of) σ_j te kiezen. Met g_j evenredig met σ_j vindt men dat onder H_0

$$(\Sigma_j\, \underline{k}_j - \Sigma_j\, \mu_j)/\sqrt{\Sigma_j\, \sigma_j^2}$$

bij benadering (CLS) en N(0,1) verdeling heeft.

Ontvangen: 12-12-1983