ZASTOSOWANIA MATEMATYKI
VI (1961)

CZEN PIN i DZAN DZO-I (Wrocław)

O ESTYMACJI LICZNOŚCI POPULACJI ZA POMOCĄ METODY ŁOWIENIA I ZNAKOWANIA

Wstęp. W naukach przyrodniczych często spotykamy się z problematem estymacji liczności populacji. Przykładem może być estymacja liczności ryb określonego gatunku i wielkości w jeziorze, estymacja ilości widzów telewizyjnych w danym kraju itp. Do tego celu może służyć metoda „łowienia i znakowania". Jest ona obecnie powszechnie znana i używana w różnych dziedzinach nauk przyrodniczych. Metoda ta polega na losowaniu pewnej ilości elementów z całej populacji, znakowaniu elementów wylosowanych i zwracaniu do populacji z powrotem, ponownym losowaniu i zwracaniu itd. Z otrzymanych wyników buduje się estymator liczności całej populacji. Na ogół liczność populacji biologicznej jest zmienna w czasie. Powodem mogą być śmierć i urodzenie się nowych elementów populacji, emigracja i imigracja itd. Aby więc badać takie zagadnienia, trzeba konstruować różne modele statystyczne. Modele takie skonstruowano do różnych celów, o czym można się dowiedzieć np. z pracy Chapmana [1].

W niniejszej pracy zajmiemy się populacją o stałej liczności. W tym przypadku wygodnie jest sformułować zagadnienie korzystając z modelu urny. Wyobraźmy sobie urnę, w której znajduje się nieznana ilość kul. Oznaczmy tę ilość przez N. Chcemy ją oszacować.

H. Steinhaus zaproponował następujący sposób postępowania: Najpierw losujemy bez zwracania m_1 kul i zaopatrujemy każdą z nich w znaczek. Jest to pierwsza podpróbka. Po oznakowaniu wrzucamy te kule razem do urny. Następnie losujemy bez zwracania m_2 kul i znaczymy każdą z nich, bez względu na to, czy kula ta już była, czy nie była znakowana. Jest to druga podpróbka. Wrzucamy kule z powrotem do urny i w ten sam sposób postępujemy dalej aż do n-tej podpróbki. W ten sposób otrzymamy ciąg liczb $V_1, V_2, \ldots, V_n$, gdzie V_i oznacza ilość kul, na których jest i znaczków. Oczywiście $\sum_{i=1}^{n} iV_i = \sum_{i=1}^{n} m_i = m$. Liczbę m nazywamy *licznością próbki*. Chcemy na podstawie $V_1, \ldots, V_n$ zbudować estymator parametru N.

Oznaczmy przez V łączną ilość różnych kul wylosowanych w trakcie pobierania n podpróbek. Mamy więc $V = \sum_{i=1}^{n} V_i$. Powstaje pytanie: Jaki jest rozkład zmiennej losowej V?

1. Rozkład zmiennej V. Możemy założyć, że wszystkie kule w urnie są ponumerowane. Oznaczmy ich ilość przez N. Możemy zatem mówić o i-tej kuli ($i \leqslant N$). Oznaczmy przez R_i ilość znaczków na i-tej kuli. R_i są zmiennymi losowymi i $\sum_{i=1}^{n} R_i = m$.

Prawdopodobieństwo tego, że pewna kula zostanie wylosowana w i-tym losowaniu j-tej podpróbki, jest równe $1/(N-i+1)$. Ale wszystkie podpróbki są wylosowane niezależnie. A więc

$$(1)\quad P[V = v;\ R_{k_i} = r_{k_i}\ (i = 1, 2, \ldots, v),\ R_k = 0 \text{ dla } k \notin \{k_1, \ldots, k_v\}] =$$

$$= C(r_{k_1}, r_{k_2}, \ldots, r_{k_v}) \prod_{i=1}^{n} \left(\frac{1}{N(N-1)\ldots(N-m_i+1)} \right),$$

gdzie r_{k_i} $(i = 1, \ldots, v)$, są liczbami całkowitymi spełniającymi warunki $r_{k_i} > 0$, $\sum_{i=1}^{n} r_{k_i} = m$, a $C(r_1, r_2, \ldots, r_l)$ jest zdefiniowane jako liczba wszystkich możliwych rozmieszczeń l różnych kul w n podpróbkach tak, żeby i-ta kula powtarzała się r_i razy i żeby żadna kula nie powtarzała się w jednej i tej samej podpróbce. Oznaczmy

$$(2)\qquad G(v) = \sum_{\substack{r_1 \geqslant 1, \ldots, r_v \geqslant 1 \\ r_1 + \ldots + r_v = m}} C(r_1, r_2, \ldots, r_v).$$

Wtedy

$$(3)\quad P[V = v \mid N] =$$

$$= \binom{N}{v} \sum_{\substack{r_1 \geqslant 1, \ldots, r_v \geqslant 1 \\ r_1 + \ldots + r_v = m}} P[V = v;\ R_1 = r_1, \ldots, R_v = r_v,\ R_k = 0 \text{ dla } k > v] =$$

$$= \binom{N}{v} G(v) \prod_{i=1}^{n} \left(\frac{1}{N(N-1)\ldots(N-m_i+1)} \right).$$

W przypadku $m_1 = m_2 = \ldots = m_n = 1$, funkcja $G(v)$ daje się łatwo wyrachować. Mamy wtedy

$$G(v) = \sum_{\substack{r_1 \geqslant 1, \ldots, r_v \geqslant 1 \\ r_1 + \ldots + r_v = n}} \frac{n!}{r_1!\, r_2! \ldots r_v!}.$$

Porównując współczynniki rozwinięcia funkcji

$$(e^x-1)^v = \sum_{i=0}^{v}(-1)^i\binom{v}{i}e^{(v-i)x}$$

w szereg potęgowy, otrzymujemy

$$G(v) = \sum_{i=0}^{v}(-1)^i\binom{v}{i}(v-i)^n.$$

A więc ostatecznie

$$P[V=v \mid N] = \binom{N}{v}\left(\frac{1}{N}\right)^n \sum_{i=0}^{v}(-1)^i\binom{v}{i}(v-i)^n. \tag{4}$$

W przypadku, gdy równość $m_1 = m_2 = \ldots = m_n = 1$ nie zachodzi, znalezienie dokładnego rozkładu zmiennej losowej V jest trudne. Położmy $L = m-V$. W następnej sekcji znajdziemy asymptotyczny rozkład zmiennej L, gdy $m \to \infty$ i $m/N \to 0$.

2. Graniczny rozkład zmiennej losowej L.

Z wzoru (3) wynika

$$P[L=l \mid N] = P[V=m-l \mid N] =$$

$$= \binom{N}{m-l}G(m-l)\prod_{i=1}^{n}\left(\frac{1}{N(N-1)\ldots(N-m_i+1)}\right).$$

Udowodnimy najpierw

LEMAT 1.

(a) *Jeśli* $m/\sum_{i\neq j}^{n} m_i m_j \to 0$, *to*

$$C(\underbrace{2,2,\ldots,2}_{l\text{ razy}}) = \Big(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j\Big)^l\Big[1+O\Big(m/\sum_{i\neq j}^{n} m_i m_j\Big)\Big];$$

(b) *jeśli* $\sum_{j=1}^{k} r_j = l+k$ *i* $r_j \geqslant 2$ $(j=1,2,\ldots,k)$, *to*

$$C(r_1, r_2, \ldots, r_k) \leqslant \Big(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j\Big)^{l-1} m.$$

Dowód. (a) Z definicji $C(r_1=2,\ldots,r_l=2)$ jest ilością wszystkich możliwych rozmieszczeń l różnych liczb w n podpróbkach tak, żeby każda liczba powtarzała się dwa razy i żeby żadna liczba nie powtarzała się w jednej i tej samej podpróbce. Jest rzeczą jasną, że liczbę 1 możemy

umieścić w $\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j$ różnych sposobów. Po umieszczeniu liczby 1, możemy umieścić liczbę 2 w

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(m_i-\varepsilon_i^1)(m_j-\varepsilon_j^1)=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j-\sum_{i\neq j}^{n}\varepsilon_i^1 m_j+\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\varepsilon_i^1\varepsilon_j^1=$$

$$=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j\left(1+O\left(m\Big/\sum_{j\neq i}^{n} m_i m_j\right)\right)$$

różnych sposobów, gdzie

$$\varepsilon_i^1=\begin{cases}1, & \text{jeżeli w } i\text{-tej podpróbce znajduje się liczba 1,}\\ 0, & \text{jeżeli tak nie jest.}\end{cases}$$

Podobnie, jeśli mamy już rozmieszczone liczby $1, 2, \ldots, k-1$, to liczbę k możemy umieścić w

$$\text{(5)}\qquad \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(m_i-\varepsilon_i^1-\varepsilon_i^2-\ldots-\varepsilon_i^{k-1})(m_j-\varepsilon_j^1-\varepsilon_j^2-\ldots-\varepsilon_j^{k-1})=$$

$$=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j\left(1+O\left(m\Big/\sum_{i\neq j}^{n} m_i m_j\right)\right)$$

różnych sposobów, gdzie

$$\varepsilon_i^t=\begin{cases}1, & \text{jeżeli w } i\text{-tej podpróbce znajduje się liczba } t,\\ 0, & \text{jeżeli tak nie jest}\end{cases}$$

($t=1,2,\ldots,k-1$; $i=1,2,\ldots,n$). Z wzoru (5) wynika pierwsza część lematu.

(b) Przejdźmy do dowodu drugiej części lematu. Rozpatrzmy najpierw przypadek $l=k+1$. Wówczas z założenia, że $r_j\geqslant 2$ ($j=1,2,\ldots,k$) wynika, że $\max\limits_{1\leqslant i\leqslant k} r_i=3$. Bez naruszenia ogólności możemy założyć, że $r_1=3$, a więc $r_2=r_3=\ldots=r_k=2$. W tym przypadku możemy liczbę 1 umieścić w $\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{s=j+1}^{n} m_i m_j m_s$ różnych sposobów, a żadnej z pozostałych $k-1$ liczb nie możemy umieścić w więcej niż $\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j$ sposobów. Mamy więc

$$\text{(6)}\qquad C(3,\underbrace{2,\ldots,2}_{l-2\text{ razy}})\leqslant\Big(\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{s=j+1}^{n} m_i m_j m_s\Big)\Big(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j\Big)^{l-2}\leqslant$$

$$\leqslant m\Big(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j\Big)^{l-1}.$$

Dla dowolnego układu $(r_1, r_2, \ldots, r_k)$ spełniającego warunki $\sum_{i=1}^{k} r_i = l+k$, $r_j \geqslant 2$ $(j = 1, 2, \ldots, k)$, mamy

$$C(r_1, r_2, \ldots, r_k) \leqslant C(3, \underbrace{2, \ldots, 2}_{l-2 \text{ razy}}). \tag{7}$$

Z wzorów (6) i (7) wynika druga część lematu.

LEMAT 2. *Jeśli l jest ustalone i $m/\sum_{i\neq j}^{n} m_i m_j \to 0$, to*

$$G(m-l) = \frac{(m-l)!}{l!}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j\right)^l \left(1+O\left(m\Big/\sum_{i\neq j}^{n} m_i m_j\right)\right). \tag{8}$$

Dowód. Udowodnimy najpierw, że

$$G(v) = \begin{cases} \sum_{i=1}^{m-v} \frac{v!}{i!} \sum_{\substack{r_j\geqslant 2,\ldots,r_i\geqslant 2 \\ r_1+\ldots+r_i=m-v+i}} C(r_1, r_2, \ldots, r_i) & \text{dla } v < m, \\ v! & \text{dla } v = m. \end{cases} \tag{9}$$

Z definicji liczba $C(r_1, \ldots, r_l)$ oznacza ilość wszystkich możliwych rozmieszczeń l różnych liczb w n podpróbkach tak, żeby liczba i powtarzała się r_i razy i żeby żadna liczba nie powtarzała się w jednej i tej samej podpróbce, natomiast

$$G(v) = \sum_{\substack{r_1\geqslant 1,\ldots,r_v\geqslant 1 \\ r_1+\ldots+r_v=m}} C(r_1, r_2, \ldots, r_v).$$

Jeżeli więc $v = m$, to $G(v) = C(r_1 = 1, \ldots, r_m = 1) = m!$. Aby udowodnić równość (9) dla $v \leqslant m$, zauważmy, że

$$C(r_1, \ldots, r_{v-i}, \underbrace{1, \ldots, 1}_{i \text{ razy}}) = i!\, C(r_1, \ldots, r_{v-i}),$$

i że

$$C(r_1, \ldots, r_i, \ldots, r_j, \ldots, r_v) = C(r_1, \ldots, r_j, \ldots, r_i, \ldots, r_v) \qquad (i, j = 1, 2, \ldots, v).$$

Załóżmy teraz, że wśród liczb $r_1, r_2, \ldots, r_v$ jest $v-i$ jedynek. Pozostałych i liczb, oraz $v-i$ jedynek można rozmieścić na miejscach $1, 2, \ldots, v$

na $\binom{v}{i}$ sposobów. Stąd więc wynika, że

$$(10)\qquad u(i) \overset{\text{def}}{=} \sum C(r_1, r_2, \ldots, r_v) =$$

$$= \binom{v}{i} \sum_{\substack{s_1 \geqslant 2, \ldots, s_i \geqslant 2 \\ s_1+s_2+\ldots+s_i = m-v+i}} C(s_1, s_2, \ldots, s_i, \underbrace{1, \ldots, 1}_{v-i \text{ razy}}) =$$

$$= \binom{v}{i} \sum_{\substack{s_1 \geqslant 2, \ldots, s_i \geqslant 2 \\ s_1+\ldots+s_i = m-v+i}} (v-i)!\, C(s_1, s_2, \ldots, s_i) = \frac{v!}{i!} \sum_{\substack{s_1 \geqslant 2, \ldots, s_i \geqslant 2 \\ s_1+\ldots+s_i = m-v+i}} C(s_1, s_2, \ldots, s_i),$$

gdzie pierwsza suma rozciąga się na wszystkie te układy $(r_1, \ldots, r_v)$ $(r_j \geqslant 1, \sum_{j=1}^{n} r_j = m)$, które zawierają dokładnie $v-i$ jedynek. Korzystając z definicji funkcji $G(v)$ i z wzoru (10) otrzymujemy ostatecznie wzór (9). Podstawiając we wzorze (9) $v = m-l$ i korzystając z lematu 1 otrzymamy tezę lematu 2.

LEMAT 3. *Jeśli* $m \to \infty$ *i*

$$\text{(i)}\qquad \frac{m^2}{\sum_{i \neq j}^{n} m_i m_j} \leqslant k, \qquad\qquad \text{(ii)}\qquad \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_i m_j}{N} = \lambda,$$

to

$$\frac{N!}{(N-m+l)!} \prod_{i=1}^{n} \left(\frac{1}{N(N-1)\ldots(N-m_i+1)} \right) = \frac{1}{N^l} e^{-\lambda} \left[1 + O\left(\frac{m}{N} \right) \right].$$

Dowód. Z nierówności

$$\sum_{j=1}^{m_i} \log(N-j) \leqslant \int_{N-m_i}^{N} \log x\, dx \leqslant \sum_{j=0}^{m_i-1} \log(N-j)$$

i wzoru

$$\int_{N-m_i}^{N} \log x\, dx = \log \frac{N^N}{(N-m_i)^{N-m_i}} - m_i$$

wynika, że

$$\log\bigl(N(N-1)\ldots(N-m_i+1)\bigr) = \log \frac{N^N}{(N-m_i)^{N-m_i}} - m_i + O\left(\frac{m_i}{N} \right).$$

Mamy zatem

$$(11)\qquad \sum_{i=1}^{n} \log\bigl(N(N-1)\ldots(N-m_i+1)\bigr) =$$

$$= \log \frac{N^{nN}}{(N-m_1)^{N-m_1} \ldots (N-m_n)^{N-m_n}} - m + O\left(\frac{m}{N} \right)$$

oraz

$$\log\frac{N!}{(N-m+l)!} = \log\frac{N^N}{(N-m+l)^{N-m+l}} - m + l + O\left(\frac{m}{N}\right). \tag{12}$$

A więc na podstawie wzorów (11) i (12)

$$\log\left[\frac{N!}{(N-m+l)!}\prod_{i=1}^{n}\frac{1}{N(N-1)\dots(N-m_i+1)}\right] = \tag{13}$$

$$= -\log N^l + \sum_{i=1}^{n}(N-m_i)\log\left(1-\frac{m_i}{N}\right) - (N-m+l)\log\left(1-\frac{m-l}{N}\right) +$$

$$+ l + O\left(\frac{m}{N}\right) =$$

$$= -l\log N + \sum_{i=1}^{n}(N-m_i)\left(-\frac{m_i}{N} - \frac{m_i^2}{2N^2}\right) -$$

$$-(N-m+l)\left(-\frac{m-l}{N} - \frac{(m-l)^2}{2N^2}\right) + l + O\left(\frac{m}{N}\right) =$$

$$= -l\log N + \frac{1}{2N}\sum_{i=1}^{n} m_i^2 - \frac{1}{2N}\left(\sum_{i=1}^{n} m_i\right)^2 + O\left(\frac{m}{N}\right) =$$

$$= -l\log N - \frac{1}{N}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j + O\left(\frac{m}{N}\right),$$

c. b. d. o.

TWIERDZENIE 1. *Jeżeli* $m \to \infty$ *w ten sposób, że*

(i) $$\frac{m^2}{\sum\limits_{i\neq j}^{n} m_i m_j} \leqslant k = \text{const}, \qquad \text{(ii)} \quad \frac{\sum\limits_{i=1}^{n-1}\sum\limits_{j=i+1}^{n} m_i m_j}{N} = \lambda,$$

to

$$P[L = l \mid N] = \frac{\lambda^l}{l!}e^{-\lambda}\left[1 + O\left(\frac{m}{N}\right)\right]. \tag{14}$$

Dowód. Twierdzenie jest prostą konsekwencją lematów 2 i 3.

3. Estymator $\overline{N}$ liczności N. W niniejszym ustępie będziemy cały czas zakładali, że warunki twierdzenia 1 są spełnione. Łatwo wtedy wykazać, że

$$E\left(\frac{1}{L+1}\right)^k \to \sum_{l=0}^{\infty} \frac{1}{(l+1)^k} \cdot \frac{\lambda^l}{l!} e^{-\lambda}, \quad k > 0. \tag{15}$$

Dla $k = 1$ mamy

$$E\left(\frac{1}{L+1}\right) \approx \sum_{l=0}^{\infty} \frac{\lambda}{(l+1)!} e^{-\lambda} = \frac{1}{\lambda}(1-e^{-\lambda}) = \tag{16}$$

$$= \frac{N}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_i m_j} \left[1 - e^{-\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_i m_j\right)/N}\right].$$

Zdefiniujmy estymator $\overline{N}$:

$$\overline{N} \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_i m_j}{L+1}. \tag{17}$$

D. G. Chapman ([2]) znalazł, dla tego samego postępowania, metodą największej wiarogodności estymator

$$\overline{N}_c = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_i m_j}{L}.$$

Nie znalazł on jednak rozkładu zmiennej losowej L i nie mógł zbadać głębiej własności tego estymatora.

Zajmijmy się bliżej estymatorem $\overline{N}$. W przybliżeniu

$$E(\overline{N}) \approx N(1-e^{-\lambda}) = N\left[1 - e^{-\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_i m_j\right)/N}\right]. \tag{18}$$

Ponadto z równości

$$\frac{1}{(l+1)^2} = \sum_{r=2}^{R} \frac{(r-2)!}{(l+1)\dots(l+r)} + \frac{(R-1)!}{(l+1)^2(l+2)\dots(l+R)}$$

oraz

$$E\left[\frac{1}{(l+1)\dots(l+r)}\right] = \sum_{l=0}^{\infty} \frac{\lambda^l}{(l+r)!} e^{-\lambda} =$$

$$= \frac{1}{\lambda^r}\left[1 - e^{-\lambda}\left(1+\lambda+\dots+\frac{\lambda^r}{(r-1)!}\right)\right],$$

otrzymujemy wzór

$$(19)\qquad E(\overline{N}-N)^2 = N^2\left(\frac{1}{\lambda}+\frac{2}{\lambda^2}+\frac{6}{\lambda^3}+O\left(\frac{1}{\lambda^4}\right)\right) =$$

$$= N^2\left[\frac{N}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j}+\frac{2N^2}{\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j\right)^2}+\frac{6N^3}{\left(\sum_{i=1}^{n-1}\sum_{i=j+1}^{n} m_i m_j\right)^3}+\right.$$

$$\left.+O\left(N^4\Big/\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j\right)^4\right)\right].$$

Z wzoru (18) widzimy, że estymator $\overline{N}$ nie jest nieobciążony, ale dla dużych λ obciążenie tego estymatora jest bardzo małe. Z wzoru (19) wynika, że jeśli chcemy, żeby średni błąd kwadratowy tego estymatora był mały, musimy powiększyć liczność próbki tak, żeby λ było duże, np. rzędu kilkuset.

Łatwo również obliczyć, że dla $\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j\right)/N \geqslant \log N$ zachodzi nierówność $|E(\overline{N})-N| \leqslant 1$.

Dla przypadku $m_1 = m_2 = m_3 = \ldots = m_n = k$ wzór (17) sprowadza się do

$$(20)\qquad \overline{N}_{n,k} = \frac{k^2 n(n-1)}{2(L+1)}$$

i

$$(21)\qquad E(\overline{N}_{n,k}-N)^2 = N^2\left[\frac{2N}{k^2 n(n-1)}+\frac{8N^2}{k^4 n^2(n-1)^2}+O\left(\frac{N^3}{k^6 n^3(n-1)^3}\right)\right].$$

Z wzoru (21) widzimy, że dla ustalonej liczności próbki kn wyrażenie $E(\overline{N}_{n,k}-N)^2$ jest tym mniejsze, im mniejsza jest liczba k. Optymalny przypadek jest dla $k = 1$. Mamy wtedy

$$\overline{N}_{n,1} = \frac{n(n-1)}{2(L+1)}$$

oraz

$$(22)\qquad E(\overline{N}_{n,1}-N)^2 = N^2\left[\frac{2N}{n(n-1)}+\frac{8N^2}{n^2(n-1)^2}+O\left(\frac{N^3}{n^3(n-1)^3}\right)\right].$$

W najprostszym przypadku, gdy $n = 2$, $\overline{N} = m_1 m_2/(L+1)$. Przypuśćmy, że $m_1 = m_2 = k$. Mamy wtedy

$$\overline{N}_{2,k} = \frac{k^2}{L+1}$$

oraz

$$E(\overline{N}_{2,k}-N)^2 = N^2\left[\frac{N}{k^2}+\frac{2N^2}{k^4}+O\left(\frac{N^3}{k^6}\right)\right]. \tag{23}$$

Dla tej samej liczności próbki $2k$

$$E(\overline{N}_{2k,1}-N)^2 = N^2\left[\frac{N}{k(2k-1)}+\frac{2N^2}{k^2(2k-1)^2}+O\left(\frac{N^3}{k^3(2k-1)^3}\right)\right] \tag{24}$$

otrzymujemy zatem

$$\frac{E(\overline{N}_{2,k}-N)^2}{E(\overline{N}_{2k,1}-N)^2} \approx 2. \tag{25}$$

Z wzoru (25) widzimy, że na tym samym poziomie dokładności (tzn. tej samej dyspersji) w postępowaniu dwupróbkowym trzeba będzie wylosować 2 razy więcej kul niż w postępowaniu pojedynczym.

4. Przedziałowa estymacja liczności N. W praktyce na ogół nie chodzi nam o estymację punktową liczności populacji. Wolelibyśmy orzekać, że liczność danej populacji jest mniej więcej tyle a tyle, niż mówić, że liczność danej populacji jest np. sto, sto jeden, sto pięć, a nie inna. Innymi słowy, chcemy znaleźć taki przedział o końcach będących zmiennymi losowymi, o którym możemy z jakimś współczynnikiem ufności powiedzieć, że zawiera liczbę N.

Z twierdzenia 1 i z centralnego twierdzenia granicznego wynika, że dla $\lambda \to \infty$

$$P\left[\frac{(L-\lambda)^2}{\lambda} \leqslant a^2 \mid N\right] = P\left[\left|\frac{L-\lambda}{\sqrt{\lambda}}\right| \leqslant a \mid N\right] \to \Phi(a)-\Phi(-a), \tag{26}$$

gdzie

$$\Phi(a) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-\frac{1}{2}t^2}\,dt.$$

Z wzoru (26) wnioskujemy, że jeśli

$$\begin{aligned} A &= \tfrac{1}{2}\left[2L+a^2+a\sqrt{4L+a^2}\right], \\ B &= \tfrac{1}{2}\left[2L+a^2-a\sqrt{4L+a^2}\right], \end{aligned} \tag{27}$$

to

$$P[B \leqslant \lambda \leqslant A \mid N] \approx \Phi(a)-\Phi(-a).$$

A więc

$$P\left[\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j}{A} \leqslant N \leqslant \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j}{B}\,\middle|\, N\right] \approx \Phi(a)-\Phi(-a). \tag{28}$$

Dla $a = 2$ i dla poziomu ufności 0,95 mamy

$$P\left[\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j}{(\sqrt{L+1}+1)^2} \leqslant N \leqslant \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j}{(\sqrt{L+1}-1)^2}\right] = 0{,}95. \tag{29}$$

Wtedy długość przedziału

$$\left(\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j}{(\sqrt{L+1}+1)^2}, \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j}{(\sqrt{L+1}-1)^2}\right)$$

jest równa

$$\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j}{L^2}\, 4\sqrt{L+1} \approx \tag{30}$$

$$\approx \frac{4\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j}{L^{3/2}} = \frac{4\sum_{i=1}^{n} m_i}{L}\left(\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_i m_j}{\sum_{i=1}^{n} m_i \sqrt{L}}\right).$$

W przypadku $m_1 = m_2 = \ldots = m_n = k$

$$P\left[\frac{k^2 n(n-1)}{2(\sqrt{L+1}+1)^2} \leqslant N \leqslant \frac{k^2 n(n-1)}{2(\sqrt{L+1}-1)^2}\right]$$

i długość przedziału $\left(k^2(n-1)n/2(\sqrt{L+1}+1)^2,\ k^2(n-1)n/2(\sqrt{L+1}-1)^2\right)$ jest równa

$$\frac{2k^2 n(n-1)}{L^{3/2}} = \frac{nk}{L} \cdot \frac{2k(n-1)}{\sqrt{L}}.$$

Przykład. Jeżeli dla $k = 1$, $n = 2000$ otrzymaliśmy $L = 200$, to z warunku

$$P[8676 \leqslant N \leqslant 11515] = 0{,}95$$

długość przedziału jest $11515 - 8676 = 2839$.

Dziękujemy dr Stanisławowi Trybule za pomoc w przygotowaniu ostatecznej wersji niniejszej pracy.

Prace cytowane

[1] D. G. Chapman, *The estimation of biological populations*, Ann. Math. Statist. 125 (1954), str. 1-15.

[2] — *Inverse, multiple and sequential sample censuses*, Biometrics 8 (1952), str. 288-306.

Praca wpłynęła 12. 4. 1960

ЧЕНЬ ПИНЬ и ДЗАН ДЗО-И (Вроцлав)

ОБ ОЦЕНКЕ ЧИСЛЕННОСТИ СОВОКУПНОСТИ МЕТОДОМ ЛОВЛИ И ОТМЕТОК

РЕЗЮМЕ

В статье рассматривается следующая схема выбора: Выбирается n подвыборок численностью соответственно $m_1, m_2, \ldots, m_n$. В каждой из них выбор производится без возвращений. Перед выбором $(k+1)$-ой подвыборки отмечаются и возвращаются в совокупность все элементы предыдущей подвыборки.

Пусть $m = \sum_{i=1}^{n} m_i$ и пусть V обозначает число разных элементов в выворке (т.е. во свех n подвыборках вместе). В статье доказывается, что случайная беличина $L = m - V$ имеет в пределе распределение Пуассона. Пользуясь этой теоремой мы исследуем свойства оценки

$$\overline{N} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_i m_j}{L+1}$$

численности N совокупности. Дается также способ оценки этой численности методом доверительных пределов.

CZEN PIN and DZAN DZO-I (Wrocław)

ON ESTIMATING THE SIZE OF POPULATION BY CAPTURE-MARK METHOD

SUMMARY

The following sampling scheme is considered in this paper. We draw n subsamples of sizes $m_1, \ldots, m_n$, each one without replacement. Before drawing the $(k+1)$-st subsample we put back to the population all the elements of k-th subsample.

Let $m = \sum_{i=1}^{n} m_i$, and let V denote the number of distinct elements in the whole sample (i.e. in n subsamples altogether). It is proved here that the

random variable $L = m - V$ has the limiting Poisson distribution. Using this theorem the properties of the estimate

$$\overline{N} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_i m_j}{L+1}$$

of the population size N are investigated. The paper also gives the method of construction of confidence intervals for the population size.