名寄せのためのサンプリング

林　知己夫

(1963 年 12 月受付)

Estimation of the Size of Universe by Matching in K Samples from K Lists and Estimation of the Mean Value of That Universe

Chikio HAYASHI

In the present paper, the author treats the two kinds of estimation problem, (i) the size of universe and (ii) the mean value of that universe.

The universe consists of those elements which are put on K lists, of those elements on (K−1) lists, …… and of those elements on only one list. K samples of size, $(n_1, n_2, \cdots\cdots n_k)$ are drawn with equal probability from K lists respectively. The estimation of the size of universe is obtained by matching in K Samples. In case $k=2$, Deming and Glasser treats in [2] of references. The author treats the case $k\geqq 3$. In this case, the situations are very complicated.

The estimation in $k=3$ is given in $[E_1]$ with $[D_1]$, $[D_2]$ (cf. 53 p.∼56 p.) where N_i is the size of the i-th list, n_i is the size of the sample from the i-th list, M_{123} is the number of the elements on the lists 1, 2, 3, M_{jk} is the number of the elements on the lists j, k $(j, k=1,2,3)$ and the small letter m_{123}, m_{jk} are the number in the sample corresponding to M_{123}, M_{jk}. The variance of the estimate is given $[V_1]$ in approximate form and $[V_2]$ with table 1 and table 2 (cf. 54 p.∼58 p.).

In $k=2$, the mean value of the universe, $\bar{X}=(k_{12}\bar{X}_{12}+k_1\bar{X}_1+k_2\bar{X}_2)$, where $k_{12}=M_{12}/(N_1+N_2-M_{12})$, $k_1=(N_1-M_{12})/(N_1+N_2-M_{12})$, $k_2=(N_1-M_{12})/(N_1+N_2-M_{12})$, $\bar{X}_{12}$ is the arithmetic mean of the elements on lists 1, 2, $\bar{X}_1$ is the mean of the elements only on list 1 and $\bar{X}_2$ is the mean of the elements only on list 2. The unbiased estimate is given in $\bar{x}=\alpha k_{12}\bar{x}_{12}+k_1\bar{y}/\bar{P}_1+k_2\bar{z}/\bar{Q}_2$ in case where k_{12}, k_1 and k_2 are known. $\alpha=(n_1/N_1+n_2/N_2-1)$, $\bar{P}_1=N_2(N_1-M_{12})/(N_1N_2-M_{12}n_2)$ $\bar{Q}_2=N_1(N_2-M_{12})/(N_1N_2-M_{12}n_1)$ and $\bar{x}_{12}$ is the sample mean of the elements on m_{12}, $\bar{y}$ is the sample mean of the elements on m_1-m_{12}, $\bar{z}$ is the sample mean of the elements on n_2-m_{12}. The variance is $[V_3]$. In case where k_{12}, k_1 and k_2 are to be estimated, the variance of the estimate $[E_3]$ is given in $[V_4]$ with $[D_3]$.

Institute of Statistical Mathematics

はじめに

「名寄せ」と言うことは，L個の対象（これは，識別可能な，それぞれ異った標識をもっているものとする）が重複していくつかづつ，即ち N_1, N_2 …… N_L 個づつ含まれているリスト

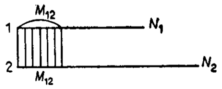
がある場合，$\sum_{i=1}^{L} N_i = N$ 個のものから，同じ標識のものをあつめ，L を見出すことを言うのである．すなわち，L人の名前を，いろいろな（でたらめな）場所・順序で，それぞれ重複して記載されているリストがあり，このリストから同じ名前をあつめて異ったL人をみつけ出すことである．Nが大きいとき，この作業は，非常に面倒なもので，われわれが把握できるN個のものから，サンプリングして，n個のものをぬき出し，ここで同じ名前のものをあつめ，つまり名寄せ，をしてLの推定としての l を求めるのが有利なことになる．

このような種類の研究は，これまで，Leo A. Goodman, W.E. Deming, G. J. Glasser 等によって行なわれている．ここでとりあげる問題は，Deming, Glasser の研究に関連の深いものである．Deming 等は，重複記載のある２つの名簿（それぞれの名簿の中では重複記載がないとする）があるとする．それぞれの名簿からサンプルを抽出し，抽出された二組のサンプルで名寄せ（matching）を行なって，異った何人が名簿にある総数であるかを推定しようとするものである．ここでは，まず，名簿の数が K 個あるとした――それぞれの名簿の中では重複記載がないとする――場合，各名簿から標本を抽出し，これらで名寄せを行なって，この結果から，全名簿に何人の別人が記載されているかを推定する問題を考える．$K=2$ の単なる拡張のように見えるが，$K=3$ では推定の様相が異ってくるので注目する必要がある．$K \geqq 4$ では $K=3$ の場合の考えが同様に通用するのである．つぎに，抽出された各人に意見調査を行なって，全別人集団に対する推定を行なおうとすることを考えることにする．ここでのべる方法は，農地被買収者の調査――K次にわたる買収があったのでK個の名簿がある，これらK個の名簿には重複記載がある――のとき考えられたもので，被買収者の総数及び被買収者の意見を推定しようとするのが目的であった．

§1.　$K=2$ の場合

まず，$K=2$ の場合についてのべてみよう．この場合の主な結果はすでに Deming 等によって得られているが，方法的に統一するためにあえてのべてみようと思う．

第１の名簿には N_1 人の名前が書いてあり，第２の名簿には N_2 人の名前が書いてある．それぞれの名簿の中では重複記載がないとする．二つの名簿 1,2 のうち重複記載あるものが M_{12} 人あるとする．知りたい人の数 L は $(N_1-M_{12})+(N_1-M_{12})+M_{12}=N_1+N_2-M_{12}=L$ である．



(N_1-M_{12}) は名簿１のみにある人 (N_2-M_{12}) は名簿２のみにある人，となる．

まず，名簿１から等確率で n_1 人を抽出する，名簿２から等確率で n_2 人を抽出するものとする．以下，とくに断わらない限り，一度抽出されたものをもとへ戻さないで抽出を行なうものとする（有限母集団の抽出と名づけておく．有限とは，母集団の大きさをいうものではなく抽出方法を規定するものとしておく．もとへもどして抽出する方式を無限母集団の抽出と名づけておく．無限とは，母集団の大きさをいうものでなく抽出方法をいうものとしておく）．

今 w_i なる確率変数を考える．$i=1,2,\cdots n_1$ としておく．この w_i は第１名簿からの i 番目サンプルにつけらるべき標識である．この標識は第２名簿からの n_2 個のサンプルとの「つけ合せ」によってきまるのである．もし，第２の名簿からの標本にのっていれば $w_i=1$，のっていなければ０とする．こうすると

$$P_r\{w_i=1\}=\frac{M_{12}}{N_1}\frac{1}{N_2}n_2$$

である．もし，第２の名簿のすべてを用いる $(n_2=N_2)$ であれば

$$P_r\{w_i=1\}=\frac{M_{12}}{N_1}$$

となるのは明らかである.
こうして

$$\sum_{i=1}^{n_1} w_i = w \quad \text{とおくと}$$

$$E(w) = n_1 \cdot \frac{M_{12}}{N_1} \frac{n_2}{N_2}$$

$$= M_{12} \frac{n_1}{N_1} \frac{n_2}{N_2}$$

となる. したがって, M_{12} に対する偏りのない推定値としては, $w\dfrac{N_1N_2}{n_1n_2}$ をとればよい.

$$w\frac{N_1N_2}{n_1n_2} = l_{12}$$

とすれば,

$$E(l_{12}) = M_{12}$$

となる.

知りたい量 L の偏りのない推定値 l としては

$$N_1 + N_2 - l_{12} = l$$

をつくればよい.

この l の分散は, m_{12} の分散に等しい. 故に $\sigma_l{}^2 = \sigma_{l_{12}}^2$ を計算すればよい.

$$\begin{aligned}
\sigma_{l_{12}}^2 &= E\left(\frac{N_1N_2}{n_1n_2} w\right)^2 - M_{12}^2 \\
&= \left(\frac{N_1N_2}{n_1n_2}\right)^2 E(w^2) - M_{12}^2 \\
&= \left(\frac{N_1N_2}{n_1n_2}\right)^2 \sum_{i,j}^{n_1} E(w_i w_j) - M_{12}^2 \\
&= \left(\frac{N_1N_2}{n_1n_2}\right)^2 \left\{\sum_i E(w_i{}^2) + \sum_{j \neq k}^{n_1} E(w_j w_k)\right\} - M_{12}^2 \\
&= \left(\frac{N_1N_2}{n_1n_2}\right)^2 \left\{M_{12}\frac{n_2}{N_1N_2} n_1 + \frac{M_{12}n_2}{N_1N_2} \cdot \frac{M_{12}-1}{N_1-1} \frac{n_2-1}{N_2-1} \cdot n_1(n_1-1)\right\} - M_{12}^2 \\
&= \left(\frac{N_1N_2}{n_1n_2}\right)^2 \left\{M_{12}\frac{n_1n_2}{N_1N_2} + M_{12}(M_{12}-1)\frac{n_1n_2}{N_1N_2} \cdot \frac{(n_1-1)(n_2-1)}{(N_1-1)(N_2-1)}\right\} - M_{12}^2 \\
&= M_{12}\frac{N_1N_2}{n_1n_2}\left\{1 + (M_{12}-1) \cdot \frac{n_1-1}{N_1-1} \cdot \frac{n_2-1}{N_2-1}\right\} - M_{12}^2 \\
&= \left(M_{12}\frac{N_1N_2}{n_1n_2} - \frac{M_{12}^2}{n_1}\right) + \left\{M_{12}(M_{12}-1)\frac{N_1}{n_1}\frac{N_2}{n_2} \cdot \frac{n_1-1}{N_1-1} \cdot \frac{n_2-1}{N_2-1} - M_{12}^2\left(1 - \frac{1}{n_1}\right)\right\}
\end{aligned}$$

第1項は w_i に関する分散のみを考慮した場合,（つまり無限母集団的抽出を行なった場合）第2項は w_j, w_k 項の共分散（負の相関をもつ）を考慮に入れた場合（有限母集団的抽出を行なった場合）である. 第2項は負でもあり, 一般のサンプリングではあまり大きな量にならないので, 安全目には第1項のみを考えても十分なのである.

このような分散は一般にあまり小でないものであるが, この計算が一つの層におけるものであり, これをつみあげて全体的な分散を考える場合であれば, 相対精度は上ってくるものである. 即ち第 r 層の推定値 ${}^r w$ を考え（この層内では2つの名簿があり上述のサンプリングが行えるものとする）$\sum_r^R {}^r w$, （R は層の数） によって, 全体の重複記載のない人の数の推定を行なうときには, この分散は $\sum_r^R {}^r\sigma_{l_{12}}^2$ となる（各層の間では, 名簿の 重複記載はないものとす

る).

数値例をあげてみよう.

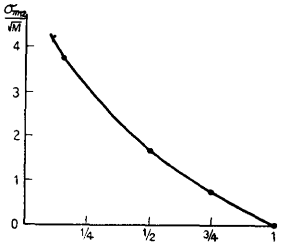
［数値例］

$$M_{12}=200,\qquad N_1=2000\qquad n_1=400$$
$$N_2=1000\qquad n_2=200$$

このとき $\sigma^2_{l_{12}}\doteqdot 4561$, $\sigma_{l_{12}}\doteqdot 67.5$ であるから，相対精度 $\dfrac{\sigma_{l_{12}}}{M_{l_{12}}}$ は大きいことが知られよう．前に示した第1項だけによる分散は 4900 であり，標準備量では 70 となり大局的にはほとんどかわりがない．今かりに $M_{12}=800$ とすれば，$\sigma^2_{l_{12}}=124.0$ 第1項だけだと $\sigma_{l_{12}}=135.6$ となり，著変はない．

つぎに $\dfrac{n_1}{N_1}=\dfrac{n_2}{N_2}=k$ とすると

$$\sigma^2_{l_{12}}\doteqdot M\frac{1}{k^2}\{1+k^2(M-1)\}-M^2$$
$$=M\left(\frac{1}{k^2}-1\right)$$



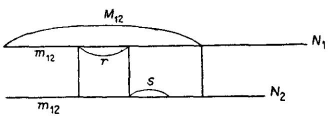
となるので，k の値に対して急激に分散が減少することがわかる．

さて，$n_1+n_2=n$ の下で，$\sigma^2_{l_{12}}$ が最小となるような n_1, n_2 はどうなるかを計算してみると，$N_1\gg n_1$，$N_2\gg n_2$ とすれば N_1, N_2 の大きさは関係なく，$n_1=n_2=n/2$ となることが直ちにわかる。前の数値例で $n_1=300$, $n_2=300$ $\sigma^2_{l_{12}}=4032$ $\sigma_{l_{12}}\doteqdot 63.6$ となり，いくらか小さくなることが示される．これは $\sigma^2_{l_{12}}+\lambda(n_1+n_2-n)$ を形式的に n_1, n_2 で微分することによって求められる．

w の分布を考える場合，無限母集団的抽出を行なったとすれば

$$w_1+\cdots\cdots\cdots+w_{n_1}=w$$

であり，$P_r\{w_1=1\}=\dfrac{M_{12}}{N_1}\dfrac{n_2}{N_2}=p$ であるから，上の分布は二項分布として与えられることがわかる．n_1 が大であれば，ガウス分布に近似されるので，われわれの考えた推定は有効なものとなり，推定の幅もガウスの時の幅づけとなり能率がよくなる．一時に N_1, N_2, M_{12} が大で，また n_1, n_2 も大で，$(N_1, N_2, M)\gg(n_1, n_2)$ であれば，まずこの形になると考えてよいのである．



抽出法，名寄せ法は前述した方法によるものとしよう．m_{12} はサンプルで共通にあらわれる数，r は本来は M_{12} の中にあり n_1 のサンプルの中にあるが n_2 の中に突き合わなかったもの，s は，本来は M_{12} の中にあり，また n_2 のサンプルの中にあるが n_1 のサンプルと突き合わなかった数をあらわす．

$$m_{12}=0,1,\cdots\cdots\min[n_1, n_2, M_{12}]=A(m)$$

ある．

$$r=0,1,\cdots\cdots\cdots\min[n_1-m_{12}, M_{12}-m_{12}]=A(r)$$
$$s=0,1,\cdots\cdots\cdots\min[n_2-m_{12}, M_{12}-m_{12}-r]=A_r(s)$$

こうすると m_{12} の分布は

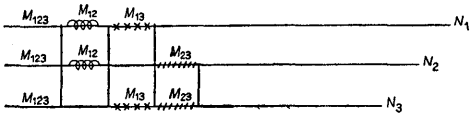
$$P_r\{m_{12}\}=\sum_{r=0}^{A(r)}\sum_{s=0}^{A\cdot(s)}\binom{M_{12}}{m_{12}}\frac{\binom{M_{12}-m_{12}}{r}\binom{N_1-M_{12}}{n_1-m_{12}-r}}{\binom{N_1}{n_1}}\cdot\frac{\binom{M_{12}-m_{12}-r}{s}\binom{N_2-M_{12}}{n_2-m_{12}-s}}{\binom{N_2}{n_2}}$$

$$=\sum_{r=0}^{A(r)}\sum_{s=0}^{A\cdot(s)}\frac{1}{\binom{N_1}{n_1}\binom{N_2}{n_2}}\binom{N_1-M_{12}}{n_1-m_{12}-r}\binom{N_2-M_{12}}{n_2-m_{12}-s}\frac{M_{12}!}{m_{12}!r!s!(M_{12}-m_{12}-r-s)!}$$

によって与えられる．$m_{12}=0,1,\cdots\cdots A(m)$ である．

§2.　$K=3$ の場合

このときは，次のように考えられる．



第 i 名簿の大きさを N_i とする．それぞれ等確率で n_i のサンプルを抽出する．M_{123} は名簿 1,2,3 に共通記載あるもの，M_{12}, M_{13}, M_{23} はそれぞれ $\{1,2\}$, $\{1,3\}$, $\{2,3\}$ に重複記載あるものの数とする．

$$N_1+N_2+N_3=N$$
$$n_1+n_2+n_3=N$$

とする．推定したい量は

$$N-2M_{123}-M_{12}-M_{13}-M_{23}=L$$

となる．われわれとしては，$2M_{123}+M_{12}+M_{13}+M_{23}=M$ の推定mと分散が目的となる．L の推定としては $N-m=l$ でよく，$\sigma_l^2=\sigma_m^2$ である．

M は標本における名寄せから，m_{123}, m_{12}, m_{13}, m_{23}（それぞれ名簿 $\{1,2,3\}$ に共通のもの，$\{1.2\}$, $\{1.3\}$, $\{2.3\}$ に共通のものの数）をもとめこれをもとにして推定しようとすることになる．名簿が一般に K 個ある場合も全く同様に $M=\sum_{j=2}^{K}(j-1)M[j]$，但し $M[j]$ は，j 個の名簿に重複記載のある数で，$M[j]$ は，$\binom{K}{j}$ 個の要素の和をあらわす．

Mを見出すサンプルとして n_1, n_2 を使いこれに名寄せによって標識を与え行なう．n_3 はただ名寄せのためにだけつかう．

$$(u_1, u_2, \cdots\cdots u_{n_1},\ v_1,\ v_2, \cdots\cdots v_{n_2})$$

u は第 i 名簿の標本の各に与えられるべき標識，v は第 2 名簿の標本に与えられるべき標識，u の標識は n_2, n_3 との突合せで，v の標識は n_3 との突合せで与えるべるものとする（但し，n_1 のものと一致しているものは m_{123} に入るので除かれる）．この標識は次のようにして与えることにする．またその時の確率を書いてみる．

[u に関して]	標識	確率
3 つ一致したとき m_{123} として計上される	a	（M_{123} から）$M_{123}\dfrac{n_2n_3}{N_1N_2N_3}$
2 つが一致したとき m_{12}, m_{13}, として計上される	b	（M_{12} から） $M_{12}\dfrac{n_2}{N_1N_2}$ （M_{123} から） $\dfrac{M_{123}}{N_1}\dfrac{n_3}{N_2}\left(1-\dfrac{n_3}{N_3}\right)$ （M_{13} から）

Table 1 クロス項における組合せ標識の生ずる確率の表（確率は $\alpha\cdot\beta$ によって求められる）

標識		2番目の標識の生ずる確率 β ／ 最初の標識の生ずる確率, α	a'	b: b_1	b: b_2
a'		$M_{123}\dfrac{n_2n_3}{N_1N_2N_3}\times$	$(M_{123}-1)\dfrac{1}{N_1-1}\dfrac{n_2-1}{N_2-1}\dfrac{n_3-1}{N_3-1}$	$(M_{123}-1)\dfrac{1}{N_1-1}\dfrac{n_2-1}{N_2-1}\left(1-\dfrac{n_3-1}{N_3-1}\right)$	$M_{12}\dfrac{1}{N_1-1}\dfrac{n_2-1}{N_2-1}$
b	b_1	$M_{123}\dfrac{1}{N_1}\dfrac{n_2}{N_2}\left(1-\dfrac{n_3}{N_3}\right)\times$	$(M_{123}-1)\dfrac{1}{N_1-1}\dfrac{n_2-1}{N_2-1}\dfrac{n_3}{N_3-1}$	$(M_{123}-1)\dfrac{1}{N_1-1}\dfrac{n_2-1}{N_2-1}\left(1-\dfrac{n_3}{N_3-1}\right)$	$M_{12}\dfrac{1}{N_1-1}\dfrac{n_2-1}{N_2-1}$
	b_2	$M_{12}\dfrac{1}{N_1}\dfrac{n_2}{N_2}\times$	$M_{123}\dfrac{1}{N_1-1}\dfrac{n_2-1}{N_2-1}\dfrac{n_3}{N_3}$	$M_{133}\dfrac{1}{N_1-1}\dfrac{n_2-1}{N_2-1}\left(1-\dfrac{n_3}{N_3}\right)$	$(M_{12}-1)\dfrac{1}{N_1-1}\dfrac{n_2-1}{N_2-1}$
c	c_1	$M_{123}\dfrac{1}{N_1}\left(1-\dfrac{n_2}{N_2}\right)\dfrac{n_3}{N_3}\times$	$(M_{123}-1)\dfrac{1}{N_1-1}\dfrac{n_2}{N_2-1}\dfrac{n_3-1}{N_3-1}$	$(M_{123}-1)\dfrac{1}{N_1-1}\dfrac{n_2}{N_2-1}\left(1-\dfrac{n_3-1}{N_3-1}\right)$	$M_{12}\dfrac{1}{N_1-1}\dfrac{n_2}{N_2-1}$
	c_2	$M_{13}\dfrac{1}{N_1}\dfrac{n_3}{N_3}\times$	$M_{123}\dfrac{1}{N_1-1}\dfrac{n_2}{N_2}\dfrac{n_3-1}{N_3-1}$	$M_{123}\dfrac{1}{N_1-1}\dfrac{n_2}{N_2}\left(1-\dfrac{n_3-1}{N_3-1}\right)$	$M_{12}\dfrac{1}{N_1-1}\dfrac{n_2}{N_2}$
d	d_1	$M_{123}\left(1-\dfrac{n_1}{N_1}\right)\dfrac{1}{N_2}\dfrac{n_3}{N_3}\times$	―――省―――	―――――――	―――略―――
	d_2	$M_{23}\dfrac{1}{N_2}\dfrac{n_3}{N_3}\times$		$*M_{123}\dfrac{1}{N_1}\dfrac{n_2-1}{N_2-1}\left(1-\dfrac{n_3-1}{N_3-1}\right)$ (term の数 n_1 個)	←― term の数 注意の事 ―→

c		d		備考
c_1	c_2	d_1	d_2	
$(M_{123}-1)\dfrac{1}{N_1-1}\left(1-\dfrac{n_2-1}{N_2-1}\right)\dfrac{n_3-1}{N_3-1}$	$M_{13}\dfrac{1}{N_1-1}\;\dfrac{n_3-1}{N_3-1}$	$(M_{123}-1)\left(1-\dfrac{n_1-1}{N_1-1}\right)\dfrac{1}{N_2-1}\;\dfrac{n_3-1}{N_3-1}$	$M_{23}\dfrac{1}{N_2-1}\;\dfrac{n_3-1}{N_3-1}$	n_2-1
$(M_{123}-1)\dfrac{1}{N_1-1}\left(1-\dfrac{n_2-1}{N_2-1}\right)\dfrac{n_3}{N_3-1}$	$M_{13}\dfrac{1}{N_1-1}\;\dfrac{n_3}{N_3-1}$	$(M_{123}-1)\left(1-\dfrac{n_1-1}{N_1-1}\right)\dfrac{1}{N_2-1}\;\dfrac{n_3}{N_3-1}$	* $M_{23}\dfrac{1}{N_2-1}\;\dfrac{n_3}{N_3-1}$	n_2-1
$M_{123}\dfrac{1}{N_1-1}\left(1-\dfrac{n_2-1}{N_2-1}\right)\dfrac{n_3}{N_3}$	$M_{13}\dfrac{1}{N_1-1}\;\dfrac{n_3}{N_3}$	$M_{123}\left(1-\dfrac{n_1-1}{N_1-1}\right)\dfrac{1}{N_2-1}\;\dfrac{n_3}{N_3}$	$M_{23}\dfrac{1}{N_2-1}\;\dfrac{n_3}{N_3}$	n_2-1
$(M_{123}-1)\dfrac{1}{N_1-1}\left(1-\dfrac{n_2}{N_2-1}\right)\dfrac{n_3-1}{N_3-1}$	$M_{13}\dfrac{1}{N_1-1}\;\dfrac{n_3-1}{N_3-1}$	$(M_{123}-1)\left(1-\dfrac{n_1-1}{N_1-1}\right)\dfrac{1}{N_2-1}\;\dfrac{n_3-1}{N_3-1}$	* $M_{23}\dfrac{1}{N_2-1}\;\dfrac{n_3-1}{N_3-1}$	n_2
$M_{123}\dfrac{1}{N_1-1}\left(1-\dfrac{n_2}{N_2}\right)\dfrac{n_3-1}{N_3-1}$	$(M_{13}-1)\dfrac{1}{N_1-1}\;\dfrac{n_3-1}{N_3-1}$	$M_{123}\left(1-\dfrac{n_1-1}{N_1-1}\right)\dfrac{1}{N_2}\;\dfrac{n_3-1}{N_3-1}$	$M_{23}\dfrac{1}{N_2}\;\dfrac{n_3-1}{N_3-1}$	n_2
——省——	——略	$(M_{123}-1)\left(1-\dfrac{n_1}{N_1-1}\right)\dfrac{1}{N_2-1}\;\dfrac{n_3-1}{N_3-1}$	$M_{23}\dfrac{1}{N_2-1}\;\dfrac{n_3-1}{N_3-1}$	n_2
* $M_{123}\dfrac{1}{N_1}\left(1-\dfrac{n_2-1}{N_2-1}\right)\dfrac{n_3-1}{N_3-1}$ (term の数 n_1 個)		$M_{123}\left(1-\dfrac{n_1}{N_1}\right)\dfrac{1}{N_2-1}\;\dfrac{n_3-1}{N_3-1}$	$(M_{23}-1)\dfrac{1}{N_2-1}\;\dfrac{n_3-1}{N_3-1}$	n_2

*註　全体つみあげれば同一のものになるが，確率表では異ったものになるので参考のためあげておく．

$[D_1]$

$$c \quad \begin{cases} M_{13}\dfrac{1}{N_1}\dfrac{n_3}{N_3} \\ (M_{123}\text{ から}) \\ \dfrac{M_{123}}{N_1}\left(1-\dfrac{n_2}{N_2}\right)\dfrac{n_3}{N_3} \end{cases}$$

[v に関して]

m_{23} として

計上される

$$d \quad \begin{cases} (M_{23}\text{ から}) \\ M_{23}\dfrac{1}{N_2}\dfrac{n_3}{N_3} \\ (M_{123}\text{ から}) \\ \left(1-\dfrac{n_1}{N_1}\right)\dfrac{M_{123}}{N_2}\dfrac{n_3}{N_3} \end{cases}$$

ここに $M[j]$ からと書いたのは，そこから生ずるという意味で M_{123} のものでも標本上 m_{12}, m_{23}, m_{13} におちることがあるのでこのようなものが出てくる．

いま，$K=2$ の場合の素朴な延長として次のように考えてみる．a として $2\dfrac{N_1N_2N_3}{n_1n_2n_3}$, b として $\dfrac{N_1N_2}{n_1n_2}$, c として $\dfrac{N_1N_3}{n_1n_3}$, d として $\dfrac{N_2N_3}{n_2n_3}$ とおいてみる．

かくして $\sum_{i=1}^{n_1}u_i+\sum_{j=1}^{n_2}v_j$ を考え，この待望値を考えると

$$\begin{aligned} E\left(\sum_{i=1}^{n_1}u_i+\sum_{j=1}^{n_2}v_j\right)&=n_1\left[a\frac{M_{123}n_2n_3}{N_1N_2N_3}+b\left\{\frac{M_{12}n_2}{N_1N_2}+\frac{M_{123}n_2}{N_1N_2}\left(1-\frac{n_3}{N_3}\right)\right\}\right. \\ &\quad\left.+c\left\{\frac{M_{13}n_3}{N_1N_3}+\frac{M_{123}}{N_1}\left(1-\frac{n_2}{N_2}\right)\frac{n_3}{N_3}\right\}\right] \\ &\quad+n_2\left[d\left\{\frac{M_{23}n_3}{N_2N_3}+\left(1-\frac{n_1}{N_1}\right)\frac{N_{123}}{N_2}\frac{n_3}{N_3}\right\}\right] \\ &=2M_{123}+M_{12}+M_{13}+M_{23} \\ &\quad+M_{123}\left(1-\frac{n_3}{N_3}\right)+M_{123}\left(1-\frac{n_2}{N_2}\right)+M_{123}\left(1-\frac{n_1}{N_1}\right) \\ &=2M_{123}+M_{12}+M_{13}+M_{23} \\ &\quad+M_{123}\left(3-\frac{n_1}{N_1}-\frac{n_2}{N_2}-\frac{n_3}{N_3}\right) \end{aligned}$$

したがって偏りのない推定を得るためには，a の代りに，a' を用いればよい．

$$[D_2]\quad\begin{cases} a'=\dfrac{N_1N_2N_3}{n_1n_2n_3}\left(\dfrac{n_1}{N_1}+\dfrac{n_2}{N_2}+\dfrac{n_3}{N_3}-1\right) \\ b=\dfrac{N_1N_2}{n_1n_2} \\ c=\dfrac{N_1N_3}{n_1n_3} \\ d=\dfrac{N_2N_3}{n_2n_3} \end{cases}$$

こうして，標識 a', b, c, d を用いれば M の偏りのない推定値を

$[E_1]$ $m=\sum_{i=1}^{n_1}u_i+\sum_{j=1}^{n_2}v_j$ として求めることができる．

次に分散であるがこれは面倒である．しかし，前節でのべたように共分散に相当する項を略す，即ち無限母集団の抽出にしたがったとし，あるいはまた M_{123} M_{12}, M_{13}, M_{23} の推定のための抽出を独立に行なったと考えると分散は簡単に求めることができる．

Table 2 クロス項標識×確率の表

標識	標識内訳	prob (表をみよ)*	積	term の数
a'^2	$a'\ a'$	$(a'a')$	a'^2 [・]	$n_1(n_1-1)$
$a'b$	$a'(b_1+b_2)$	$(a')\{(b_1),\ (b_2)\}$	$a'b$ [・+・]	$n_1(n_1-1)$
$a'c$	$a'(c_1+c_2)$	$(a')\{(c_1),\ (c_2)\}$	$a'c$ [・+・]	$n_1(n_1-1)$
b^2	$b_1b_2\ b_1b_2$ $b_2b_1\ b_2b_2$	$(b_1^2)\quad(b_1b_2)$ $(b_2b_1)\quad(b_2^2)$	b^2 [・+・+・+・]	$n_1(n_1-1)$
bc	$b_1c_1\ b_1c_2$ $b_2c_1\ b_2c_2$	$(b_1c_1)\quad(b_1c_2)$ $(b_2c_1)\quad(b_2c_2)$	bc [・+・+・+・]	$n_1(n_1-1)$
c^2	$c_1c_1\ c_1c_2$ $c_2c_1\ c_2c_2$	$(c_1^2)\quad(c_1c_2)$ $(c_2c_1)\quad(c_2^2)$	c^2 [・+・+・+・]	$n_1(n_1-1)$
d^2	$d_1d_1\ d_1d_2$ $d_2d_1\ d_2d_2$	$(d_1^2)\quad(d_1d_2)$ $(d_2d_1)\quad(d_2^2)$	d^2 [・+・+・+・]	$n_2(n_2-1)$
$a'd$	$a'(d_1+d_2)$	$(a')\{(d_1),\ (d_2)\}$	$a'd$ [・+・]	$n_1(n_2-1)$
bd	$b_1d_1\ b_1d_2$ $b_2d_1\ b_2d_2$	$(b_1d_1)\quad(b_1d_2)$ $(b_2d_1)\quad(b_2d_2)$	bd [・+・+・+・]	$n_1(n_2-1)$
cd	$c_1d_1\ c_1d_2$ $c_2d_1\ c_2d_2$	$(c_1d_1)\quad(c_1d_2)$ $(c_2d_1)\quad(c_2d_2)$	cd [・+・+・+・]	n_1n_2

*註 $(a'a')$ とあるは確率表で $(a'a')$ のところ，(ab_1) とあるは確率表で $(a'b_1)$ のところの確率の意味

この考え方は有限母集団的抽出において生ずる負の相関（この量は一般的に大きいものではない）を略すことになるので有限母集団のときの推定量の分散の推定としては大き目になり近似としては安心できるものとなる．

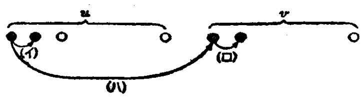
これは次のようになる．$[E_1]$ による推定の分散を $\sigma_m{}^2$ とすると，

$$[V_1]\quad\left\{\begin{aligned}
&\sigma_m{}^2\sim n_1\sigma(u)^2+n_2\sigma(v)^2\\
&\text{ここに}\\
&n_1\sigma(u)^2=a'^2\frac{n_1n_2n_3}{N_1N_2N_3}M_{123}+b^2\frac{n_1n_2}{N_1N_2}M_{12}\\
&\qquad+b^2\frac{n_1n_2(N_3-n_3)}{N_1N_2N_3}M_{123}\\
&\qquad+c^2\frac{n_1(N_2-n_2)n_3}{N_1N_2N_3}M_{123}+c^2\frac{n_1n_3}{N_1N_3}M_{13}\\
&\qquad-n_1\left[\frac{M_{123}}{n_1}\left(1+\frac{n_1}{N_1}\right)+\frac{M_{12}}{n_1}+\frac{M_{13}}{n_1}\right]^2\\
&n_2\sigma(v)^2=d^2\frac{(N_1-n_1)n_2n_3}{N_1N_2N_3}M_{123}+d^2\frac{n_2n_3}{N_2N_3}M_{23}\\
&\qquad-n_2\left\{\frac{M_{123}}{n_2}\left(1-\frac{n_1}{N_1}\right)+\frac{M_{23}}{n_2}\right\}^2
\end{aligned}\right.$$

この立場にたてば，Kが一般の場合でも推定値の分散は容易に計算することができる．

つぎに共分散の項を計算しよう．

出てくる項は，α, β, γ の三つである．



それぞれに出てくる項の標識の種類を書いてみると，（イ）では $(a'a, a'b, a'c, bb, bc, cc)$ であり（ロ）では (dd) であり（ハ）では $(a'd, bd, cd)$ である．計算を書きならべると面倒なので，ここでクロス項における確率の表及びクロス項の標識及び確率及び項の数 (c. f. Table 1 及び 2) をあげておく．なお b_1, b_2, c_1, c_2, d_1, d_2 等の区別をしたが添字 1 は M_{123} に関係のあるものをあらわす．即ち M_{123} に帰因する m_{12}, m_{13}, m_{23} に関するものをあらわすものとしておく．

このようにして分散は

$$[V_2]\quad \left\{\begin{aligned} \sigma_m{}^2 = {} & a'^2 \frac{n_1n_2n_3}{N_1N_2N_3} M_{123} \\ & + b^2 \frac{n_1n_2(N_3-n_3)}{N_1N_2N_3} M_{123} + b^2 \frac{n_1n_2}{N_1N_2} M_{12} \\ & + c^2 \frac{n_1(N_2-n_2)n_3}{N_1N_2N_3} M_{123} + c^2 \frac{n_1n_3}{N_1N_3} M_{13} \\ & + d^2 \frac{(N_1-n_1)n_2n_3}{N_1N_2N_3} M_{123} + d^2 \frac{n_2n_3}{N_2N_3} M_{23} \\ & + \text{クロスの項 (c. f. Table 1 と 2)} \\ & - (2M_{123} + M_{12} + M_{13} + M_{23})^2 \end{aligned}\right.$$

となる．

サンプルの n_1, n_2, n_3 の割当ては，$n=\sum_{1=1}^{n} n_i$ とするとき，前節の考えにしたがえるときは，$n_1=n_2=n_3$ が望ましい．抽出の便利さからは $N_i/n_i=\text{const.}$ $(i=1,2,3)$ がとられるのもよい．

3. 意見（実態）調査，$K=2$ の場合

ここでは，$K=2$ の場合について，全体の意見（実態）を推定しようとする問題をのべる．重複は計算に入れない（重複度を逆数のウエイトとする）ものとする．

M^{12}	N_1-M_{12}	N_1
M_{12}	N_2-M_{12}	N_2

各個につけらるべき数量的標識を x とする． 重複部分 M_{12} の標識に関する算術平均を $\bar{X}_{12}$, (N_1-M_{12}) 部分の標識の算術平均を $\bar{X}_1$, (N_2-M_{12}) 部分の算術平均標識を $\bar{X}_2$ とする．推定しようとする量は

$$\bar{X} = \frac{1}{N_1+N_2-M_{12}} (M_{12}\bar{X}_{12} + (N_1-M_{12})\bar{X}_1 + (N_2-M_{12})\bar{X}_2) = k_1\bar{X}_{122} + k_1\bar{X}_1 + k_2\bar{X}_2$$

$$\text{ここに}\quad k_{12} = \frac{M_{12}}{N_1+N_2-M_{12}},$$

$$k_1 = \frac{N_1-M_{12}}{N_1+N_2-M_{12}},$$

$$k_2 = \frac{N_2-M_{12}}{N_1+N_2-M_{12}},$$

まず第一に k_1, k_2, k_{12}，即ち M_{12} が既知とする．

サンプルはそれぞれの名簿から等確率で n_1, n_2 の大きさの標本抽出したとしよう．なお分散の計算では，簡単のため無限母集団的抽出によるものとする．このうち，両サンプルで重複の

みとめられた数を m_{12} とし，その標識を

$x_1, x_2, \cdots\cdots\cdots x_{m_{12}}$ とする．

$$\bar{x}_{12}=\frac{1}{m_{12}}\sum_{i=1}^{m_{12}}x_i$$

とすれば明らかに $E(\bar{x}_{12})=\bar{X}_{12}$.

$$\sigma^2_{\bar{x}_{12}} \doteqdot \frac{\sigma^2_{12}}{E(m_{12})}=\frac{\sigma^2_{12}}{\frac{M_{12}n_1n_2}{N_1N_2}}$$

ここに σ^2_{12} は M_{12} における x に関する分散をあらわす．

つぎに (n_1-m_{12})，つまり第1標本の突合せのなかった部分に着目する．

これを $y_1, y_2, \cdots y_{n_1-m_{12}}$ とする．

この y のうち本来 M_{12} に入るべきものの確率は

$$P_{12}=\frac{M_{12}}{N_1}\left(1-\frac{n_2}{N_2}\right)$$

本来 (N_1-M_{12}) に入るべきものの確率は

$$P_1=\left(1-\frac{M_{12}}{N_1}\right)$$

である．したがって n_1-m_{12} の部分にしか入らぬことを知った上でそれが M_{12} に入るべき確率は

$$\bar{P}_{12}=\frac{P_{12}}{P_{12}+P_1}=\frac{M_{12}(N_2-n_2)}{N_1N_2-M_{12}n_2}$$

それが N_1-M_{12} に入るべき確率

$$\bar{P}_1=\frac{P_1}{P_{12}+P_1}=\frac{N_2(N_1-M_{12})}{N_1N_2-M_{12}n_2}$$

となる．さて，このような y_j を代表して y とかき，この待望値 $E(y)$ は $E(y)=\bar{P}_{12}\bar{X}_{12}+\bar{P}_1\bar{X}_1$ となることは直ちに解かる．この y について算術平均値

$$\bar{y}=\frac{1}{n_1-m_{12}}\sum_{j=1}^{n_1-m_{12}}y_j$$

をつくり待望値を求めると，

$$E(\bar{y})=\bar{P}_{12}\bar{X}_{12}+\bar{P}_1\bar{X}_1$$

となる．

この分散を求めれば近似的には簡単な形となる．

ここに y^2 の待望値が

$$\begin{aligned}E(y^2)&=\bar{P}_{12}E(y'^2_{12})+\bar{P}_1E(y'^2_1)\\&=\bar{P}_{12}(\sigma^2_{12}+\bar{X}_{12})^2+\bar{P}_1(\sigma_1{}^2+\bar{X}_1)^2\end{aligned}$$

ここに y'^2_{12} は本来 M_{12} に入るものとしたときの標識

$y'_1{}^2$ は本来 N_1-M_{12} に入るものとしたときの標識

となるから

$$\begin{aligned}\sigma_y{}^2&=E(y^2)-(\bar{X}_{12}\bar{P}_{12}+\bar{X}_1\bar{P}_1)^2\\&=\sigma^2_{12}\bar{P}_{12}+\sigma_1{}^2\bar{P}_1+[(\bar{X}^2_{12}\bar{P}_{12}+\bar{X}_1{}^2\bar{P}_1)-(\bar{X}_{12}\bar{P}_{12}+\bar{X}_1\bar{P}_1)^2]\end{aligned}$$

となる．

第1項は $(M_{12}), (N_1-M_{12})$ とわけたときの所謂内分散，第2項は外分散に対応するものである．なお $\sigma_1{}^2$ は (N_1-M_{12}) における X に関する分散である．

したがって

$$\sigma_{\bar{y}}^2 \doteqdot \frac{\sigma_y^2}{E(n_1-m_{12})} \doteqdot \frac{\sigma_y^2}{n_1-M_{12}\dfrac{n_1}{N_1}\dfrac{n_2}{N_2}}$$

さてのこりの n_2-m_{12} であるが

同様に
$$Q_{12}=\frac{M_{12}}{N_2}\left(1-\frac{n_1}{N_1}\right)$$

$$Q_2=1-\frac{M_{12}}{N_2}$$

を考え，これより
$$\bar{Q}_{12}=\frac{Q_{12}}{Q_{12}+Q_2}$$

$$\bar{Q}_2=\frac{Q_2}{Q_{12}+Q_2}$$

をつくる．標識として $z_1, z_2, \cdots\cdots z_{n_2-m_{12}}$ とする．この代表を z としておく．

$$E(z)=\bar{Q}_{12}\bar{X}_{12}+\bar{Q}_2\bar{X}_2$$

となる．

したがって
$$\bar{z}=\frac{1}{n_2-m_{12}}\sum_{i=1}^{n_2-m_{12}} z_i$$

をつくると

$$E(\bar{z})=\bar{Q}_{12}\bar{X}_{12}+\bar{Q}_2\bar{X}_2$$

となる．同様にして

$$\sigma_{\bar{z}}^2 \doteqdot \frac{\sigma_z^2}{E(n_2-m_{12})}=\frac{\sigma_z^2}{n_2-M_{12}\dfrac{n_1}{N_1}\dfrac{n_2}{N_2}}$$

$$\sigma_z^2=\sigma_{12}^2\bar{Q}_{12}+\sigma_2{}^2\bar{Q}_2+[\bar{X}_{12}^2\bar{Q}_{12}+\bar{X}_2{}^2\bar{Q}_2)-(\bar{X}_{12}\bar{Q}_{12}+\bar{X}_2\bar{Q}_2)^2]$$

ここに σ_z^2 は N_2-M_{12} における X の分散をあらわすものとする．

さて $\bar{X}$ の偏りのない推定値を考えるために，［E_2］ $\bar{x}=\alpha k_{12}\bar{x}_{12}+k_1\bar{y}/\bar{P}_1+k_2\bar{z}/\bar{Q}_2$ を考える．α はこれから定めるべき常数とする．こうして，$E(\bar{x})$ をとると，

$$E(\bar{x})=\alpha k_{12}\bar{X}_{12}+k_1\bar{X}_1+k_1\frac{\bar{P}_{12}}{\bar{P}_1}\bar{X}_{12}+k_2\frac{\bar{Q}_{12}}{\bar{Q}_2}\bar{X}_{12}+k_2\bar{X}_2$$

となるので

$$\alpha=\left(1-\frac{\bar{P}_{12}}{\bar{P}_1}\frac{k_1}{k_{12}}-\frac{\bar{Q}_{12}}{\bar{Q}_2}\frac{k_2}{k_{12}}\right)=\left(\frac{n_1}{N_1}+\frac{n_2}{N_2}-1\right)$$

ととれば $\bar{x}$ は $\bar{X}$ の偏りのない推定量となる．さてこの $\bar{x}$ の分散であるが，近似的には

［V_3］
$$\sigma_{\bar{x}}^2 \doteqdot (\alpha k_{12})^2\,\sigma_{\bar{x}_{12}}^2+(k_1/\bar{P}_1)^2\sigma_{\bar{y}}^2+(k_2/\bar{Q}_2)^2\sigma_{\bar{z}}^2$$
$$-2[(\alpha k_{12})(k_1/\bar{P}_1)\bar{P}_{12}+(\alpha k_{12})(k_2/\bar{Q}_2)\bar{Q}_{12}+(k_1/\bar{P}_1)(k_2\bar{P}_2)\bar{P}_{12}\bar{Q}_{12}]\frac{\sigma_{12}^2}{M_{12}-1}$$
$$\doteqdot (\alpha k_{12})^2\sigma_{\bar{x}}^2+(k_1/\bar{P}_1)^2\sigma_{\bar{y}}^2+(k_2/\bar{Q}_2)^2\sigma_{\bar{z}}^2$$

なお上述の負の項は，無限母集団的抽出を行なっても同じサンプルを三つの部分にわける以上必然的に出てくる負の相関量であるために生ずる項である．$\bar{x}_{12}$ を知るための抽出，$\bar{y}$ を知るための抽出，$\bar{Z}$ を知るための抽出をそれぞれ独立に行なったとすればこの項は生じない．

また負の相関の項は一般には大きいものではないし負であるから安全目にみて標本抽出計画ではおとして考えても大局的には差支えないものであると考えて大過はない．即ち無限母集団的抽出，また上にのべた抽出を行なったとした場合，すべて独立となるのでこのようなものとなるのである．実用上は，この式で十分である．

次に M_{12} が未知のものであり，推定値の場合は，M_{12} のかわりに m_{12} が用いられる．即ち比率そのものも推定値となる．そのため M_{12} 既知のときの $\bar{x}$ を書きなおしてみると

$$\bar{x}=\alpha\frac{M_{12}}{N_1+N_2-M_{12}}\bar{x}_{12}+\frac{N_1N_2-M_{12}n_2}{N_2(N_1+N_2-M_{12})}\bar{y}+\frac{N_1N_2-M_{12}n_1}{N_1(N_1+N_2-M_{12})}\bar{z}$$

但し　$\alpha=\left(\frac{n_1}{N_1}+\frac{n_2}{N_2}-1\right)$

となるのでこの M_{12} のかわりに m_{12} を代入し

$$\bar{x}=\alpha\frac{m_{12}}{N_1+N_2-m_{12}}\bar{x}_{12}+\frac{N_1N_2-m_{12}n_2}{N_2(N_1+N_2-m_{12})}\bar{y}+\frac{N_1N_2-m_{12}n_1}{N_1(N_1+N_2-m_{12})}\bar{z}$$

を推定値と考えることになる．係数に比の推定が入り偏りが出てくる．ここで比推定の考えがつかわれ，平均二乗誤差が計算される．

$$[D_3]\quad\begin{cases}\dfrac{m_{12}}{N_1+N_2-m_{12}}=f_1,\\[2mm] \dfrac{N_1N_2-m_{12}n_2}{N_2(N_1+N_2-m_{12})}=f_2\\[2mm] \dfrac{N_1N_2-m_{12}n_1}{N_1(N_1+N_2-m_{12})}=f_3\end{cases}$$

F_1, F_2, F_3 は m_{12} のかわりに M_{12} を代入したものとする．

こうすると推定は

$[E_3]$　$\bar{x}=\alpha f_1\bar{x}_{12}+f_2\bar{y}+f_3\bar{z}$ とかける． 無限母集団的抽出としてサンプルの数が大であるとすれば，また負の相関の一般には小さい量を無視すれば，$\tau_{\bar{x}}{}^2=E(\bar{x}-\bar{X})^2$ から

$$[V_4]\quad \tau_{\bar{x}}^2\doteqdot\alpha^2(\sigma_{\bar{x}_{12}}{}^2+\bar{X}_{12}^2)\tau_{f_1}^2+(\sigma_{\bar{y}}{}^2+\bar{Y}^2)\tau_{f_2}^2+(\sigma_{\bar{z}}{}^2+\bar{Z}^2)\tau_{f_3}^2+\alpha^2F_1{}^2\sigma_{\bar{x}_{12}}{}^2+F_2{}^2\sigma_{\bar{y}}{}^2+F_3{}^2\sigma_{\bar{z}}{}^2$$
$$+2\alpha\bar{X}_{12}\bar{Y}\tau_{f_1f_2}+2\alpha\bar{X}_{13}\bar{Z}\tau_{f_1f_3}+2\bar{Y}\bar{Z}\tau_{f_2f_3}$$
$$+2\alpha^2F_1\sigma_{\bar{x}_{12}}{}^2E(f_1-F_1)+2F_2\sigma_{\bar{y}}^2E(f_2-F_2)+2F_3\sigma_{\bar{z}}{}^2E(f_3-F_3)$$

但し　$\bar{Y}=E(\bar{y})$

$\bar{Z}=E(\bar{z})$

$\tau_{f_if_j}=E(f_i-F_i)(f_j-F_j)$

となる．この $\tau_{f_if_j}$，$\tau_{f_i}^2$ は比推定の考えで容易に求められる．また $E(f_2-F_2)$ なる偏りも比推定のときと同様の考えで直ちに求められる．

$K\geqq 3$ のときも，同様の考えの下に近似計算をすれば，同一の考えによって推定式，分散をみちびくことができる．

統計数理研究所

文　献

［1］ Leo. A. Goodman, On the Estimation of the Number of Classes in a Population, A.M.S. Vol. 20, 1949, 572-579.; On the Analysisof Samples from K Lists, A.M.S. Vol. 23, 1952,632-639.

［2］ W.E. Deming and G.J. Glasser, On the Problem of Matching Lists by Samples, J.A.S.A., Vol. 54, 1959, 403-415.