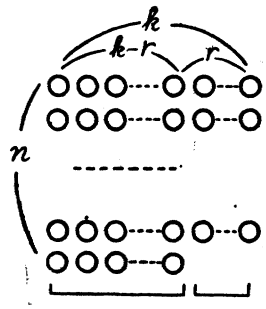
On Practical Systematic Sampling

By Hirojiro AOYAMA

When we take sample of size n from the population of size N, we adopt frequently systematic sampling method. In this paper we will make some remarks on the practical procedure.

§1. Errors in the Case $N \neq kn$[1]

Ordinarily we discuss only about the case when the size N of population is a multiple of sample size n, and use approximately the same estimate for $N=kn-r$, $0<r<k$ when $n>50$. But in the latter case the estimate of mean is not unbiased.



first group second group

Put $1 \leqq i \leqq k$ and let i be selected for a random start number. Then we have the sample mean $\bar{x}_i$:

$$\bar{x}_i = \frac{x_i + x_{i+k} + \cdots\cdots + x_{i+(n'-1)k}}{n'} \tag{1}$$

where $n' = n$, for $i \leqq k-r$

and $n' = n-1$, for $i > k-r$

So we have

$$E(\bar{x}_i) = \frac{1}{nk}\sum_{j=1}^{k-r}(X_j + X_{j+k} + \cdots\cdots + X_{j+(n-1)k}) + \frac{1}{(n-1)k}\sum_{j=1}^{r}(X_{k-r+j} + X_{k-r+j+k} + \cdots\cdots + X_{k-r+j+(n-2)k})$$

$$= \left(1-\frac{r}{k}\right)\bar{X}_1 + \frac{r}{k}\bar{X}_2 \tag{2}$$

where the mean of elements of size $n(k-r)$ (now these elements are to be

called the first group) is $\bar{X}_1$, and that of size $r(n-1)$ (these are to be called the second group) is $\bar{X}_2$.

On the other hand, as the population mean we have

$$\bar{X}=\frac{\bar{X}_1 n(k-r)+{}_2r(n-1)}{nk-r}=\bar{X}_1\frac{k-r}{k-\frac{r}{n}}+\bar{X}_2\frac{r\left(1-\frac{1}{n}\right)}{k-\frac{r}{n}} \tag{3}$$

Therefore we can get, when $n>50$

$$E(\bar{x}_t)\fallingdotseq\bar{X} \tag{4}$$

$$V(\bar{x}_t)=\left(1-\frac{r}{k}\right)(\bar{X}_1{}^2+\sigma_1{}^2)+\frac{r}{k}(\bar{X}_2{}^2+\sigma_2{}^2)-\frac{1}{k}\sum_{i=1}^{k}\overset{*}{\sigma}_i{}^2$$
$$-\left\{X_1\left(1-\frac{r}{k}\right)+\bar{X}_2\frac{r}{k}\right\}^2 \tag{5}$$

where $\sigma_1{}^2$ and $\sigma_2{}^2$ are population variances of the first and second groups respectively and $\overset{*}{\sigma}_i{}^2$ is population variance of the i-th column.

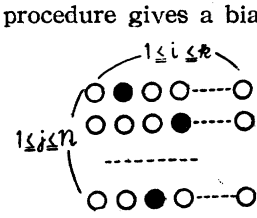
So we have as the mean square error, σ^2 being the total population variance,

$$M.S.E.\ (\bar{x}_t)\fallingdotseq\sigma^2-\frac{1}{k}\sum_{i=1}^{k}\overset{*}{\sigma}_i{}^2 \tag{6}$$

§ 2. Case when the Frame Includes the Non Universe Elements and the Size $N=kn$

We use frequently the lists, for example, the voters' lists as the frame in sampling. But at the time of sampling it often includes the non-universe elements that have changed their residences or that died already. If we can exclude these members previously, we can use ordinary systematic sampling. But when we have no time to do so, and use simple sampling interval, say 10 and multiples of 10, we always take samples in this scheme and remove the non-universe elements, if they were selected.

It is clear that this procedure gives a bias for estimate as follows.



● *non universe element*

Let the population be divided into k sets. As for the element of i-th column and j-th row

$f(i,j)=1$, when it is universe element

$=0$, when it is non-universe element.

Suppose that the number of elements of the i-th column except the non-universe elements is N_i' and

$$\sum_{i=1}^{k} N_i' = N'.$$

When we have selected a random start number i, we have as the sample mean

$$\bar{x}_i' = \frac{x_i f(i,1) + x_{i+} f(i,2) + \cdots\cdots + x_{i+(n-1)k} f(i,n)}{\sum_{j=1}^{k} f(i,j)} \tag{7}$$

We can easily prove that

$$E(\bar{x}_i') = \frac{1}{k} \sum_{i=1}^{k} \bar{X}_i' \tag{8}$$

But for the population mean,

$$\bar{X}' = \frac{1}{N'} \sum_{i=1}^{k} N_i' \bar{X}_i'. \tag{9}$$

So we have

$$E(\bar{x}_i') \neq \bar{X}'$$

that is, $\bar{x}_i'$ is not unbiased.

And we get

$$V(\bar{x}_i') = E\left(\bar{x}_i' - \frac{1}{k}\sum_{j=1}^{k} \bar{X}_j\right)^2 = \frac{1}{k}\sum_{i=1}^{k} (\bar{X}_i' - \bar{X}')^2 - \left(\bar{X}' - \frac{1}{k}\sum_{i=1}^{k} \bar{X}_i'\right)^2, \tag{10}$$

therefore

$$M.S.E.\ (\bar{x}_i') = \frac{1}{k}\sum_{i=1}^{k} (\bar{X}_i' - \bar{X}')^2. \tag{11}$$

In order to evaluate the order of this bias we assume that the N' elements are divided into k sets at random, then, taking expectation of $\bar{X}_i'$,

$$E(\bar{X}_i') = \frac{E(X_{i1} + \cdots\cdots + X_{iN_i'})}{E(N_i')} + E(R_i) = \bar{X}' + E(R_i) \tag{12}$$

because $E(N_i') = N'/k = \bar{N}'$ (which is proved in the next section), where R_i denotes error term of the i-th mean.

Thus we have

$$E\left(\frac{1}{k}\sum_{i=1}^{k} \bar{X}_i'\right) = \bar{X}' + E\left(\frac{1}{k}\sum_{i=1}^{k} R_i\right) \tag{13}$$

and

$$E(R_i) \fallingdotseq \frac{E(X_{i1} + \cdots\cdots + X_{iN_i'})\ V(N_i')}{\{E(N_i')\}^3} = \frac{\bar{X}'}{\bar{N}'^2}\tau^2 \tag{14}$$

where

$$V(N_i') = \tau^2.$$

Therefore as the mean of $E(\bar{x}_i')$, we have

$$\bar{E}(\bar{x}_i') = \bar{X}' + \frac{\bar{X}'}{\bar{N}'^2}\tau^2. \tag{15}$$

Also we have

$$V(\bar{X}_i') = \frac{V(X_{i1}+\cdots\cdots+X_{iN_i'})}{\{E(N_i')\}^2} + \frac{\bar{X}'^2\tau^2}{\{E(N_i')\}^2} - 2\rho_i\tau\bar{X}'\frac{\sqrt{V(X_{i1}+\cdots\cdots+X_{iN_i'})}}{\{E(N_i')\}^2}$$

where ρ_i is the correlation coefficient between N_i' and $X_i' = \sum_{j=1}^{N_i'} X_{ij}$. After some computation, σ^2 being the population variance,

$$V(X_{i1}+\cdots\cdots+X_{iN_i'}) = E\left(\sum_{j=1}^{N_i'} X_{ij} - \bar{N}'\bar{X}'\right)^2 = \frac{\bar{N}'^2(k-1)-\tau^2}{N'-1}\sigma^2 + \tau^2\bar{X}'^2,$$

so as the mean of $V(\bar{x}_i')$, we have

$$\bar{V}(\bar{x}_i') = EV(\bar{x}_i') = \frac{1}{k}\left(\frac{\bar{N}'^2(k-1)-\tau^2}{N'-1}\frac{\sigma^2}{\bar{N}'^2} + 2\tau^2\frac{\bar{X}'^2}{\bar{N}'^2}\right)$$
$$-2\tau\frac{\bar{X}'(k-1)}{k^2\bar{N}'^2}\sqrt{\frac{\bar{N}'^2(k-1)-\tau^2}{N'-1}\sigma^2+\tau^2\bar{X}'^2}\sum_{i=1}^{k}\rho_i$$
$$+ (\text{terms including the correlations between } \bar{X}_i' \text{ and } \bar{X}_j') \quad (16)$$

When we can put approximately $\tau \fallingdotseq 0$, $\rho = 0$, etc., we have

$$\bar{V}(\bar{x}_i') \fallingdotseq \frac{(k-1)^2}{k}\frac{\sigma^2}{N'-1} \fallingdotseq \frac{\sigma^2}{\bar{N}'} \quad (17)$$

Example:

In the voters' lists of 4-chome Ikebukuro, Toshima-ku, Tokyo,

$$N=2580,\quad N'=2476,\quad k=60,\quad n=43,\quad \bar{N}'=41.3,$$

where we neglected the fractions for convenience. In this case the numbers of non-universe elements were as follows:

1, 4, 4, 2, 4, 6, 9, 11, 8, 0, 1, 2, 1, 1, 0,
2, 0, 1, 1, 2, 2, 1, 1, 0, 1, 3, 1, 1, 1, 2,
0, 1, 0, 0, 0, 4, 0, 1, 0, 2, 1, 0, 0, 2, 0,
1, 2, 1, 0, 4, 1, 1, 1, 2, 1, 2, 0, 3, 0, 1.

Therefore we obtain 4.773 as the estimate of $V(N_i') = \tau^2$ and the bias of $\bar{E}(\bar{x}_i')$ in regard to $\bar{X}'$ is

$$\frac{\tau^2}{\bar{N}'^2} \times 100\,\% = 0.284\,\%.$$

§ 3. Numbers of Partition of Natural Numbers

In the preceding paragraph we have got $\bar{E}(\bar{x}_i')$ assuming that the methods of dividing N' elements into k sets should occur equally likely. In this paragraph we will prove $E(N_i') = \bar{N}'$ and introduce the evaluation formula of τ^2.

The number of methods of dividing N' elements into k sets which are allowed to include no element, is $p_k(N'+k)$ where $p_k(m)$ denotes the number of dividing natural number m into exact k sets of natural numbers, satisfying,[3]

$$\sum_{m,k=1}^{\infty} p_k(m)\, x^m z^k = \frac{1}{(1-xz)\,(1-x^2z)\cdots\cdots} \tag{18}$$

$$p_k(m) = \sum_{l=1}^{\infty} p_l(m-k)\,. \tag{19}$$

We may assume that we take only one element from among elements of k sets, and so we may compute $E(N_i')$ and $V(N_i')$ for these $kp_k(N'+k)$ elements.

Now let the number of 0 be $F_{N',k}(0)$, that of 1 be $F_{N',k}(1)$, ..., then we have

$$\sum_{x=0}^{N'} F_{N',k}(x) = kp_k(N'+k) \tag{20}$$

$$F_{N',k}(0) = (k-1)\,p_1(N') + (k-2)\,p_2(N') + \cdots\cdots + p_{k-1}(N')\,. \tag{21}$$

Also by means of mathematical induction,

$$\begin{aligned} F_{N',k}(x) =& p_1(N'-x) + p_2(N'-x) + \cdots\cdots + p_{k-1}(N'-x) \\ &+ p_1(N'-2x) + p_2(N'-2x) + \cdots\cdots + p_{k-2}(N'-2x) \\ &+ \cdots\cdots\cdots\cdots \\ &+ p_1(N'-(k-1)x) + \delta_{N'-kx} \\ =& p_{k-1}(N'+k-(x+1)) + p_{k-2}(N'+k-2(x+1)) + \cdots\cdots \\ &+ p_1(N'+k-(k-1)(x+1)) + \delta_{N'-kx} \end{aligned} \tag{23}$$

where $p_k(m)=0$ for $m<k$ and

$$\begin{aligned} \delta_{N'-kx} &= 1, \quad \text{for} \quad N' = kx \\ &= 0, \quad \text{for} \quad N' \neq kx \end{aligned}$$

From these relations we can prove by double induction

$$\sum_{x=0}^{N'} xF_{N',k}(x) = N'p_k(N'+k) \tag{24}$$

For, assuming, for $(N', k+1)$ and $(N'+1, k)$, the equation (24) holds, we can prove

$$\begin{aligned} \sum_{x=0}^{N'+1} xF_{N'+1,k+1}(x) &= \sum_{x=0}^{N'+1} x\Big(F_{N'+1,k}(x) + F_{N'-k,k+1}(x-1)\Big) \\ &= (N'+1)\,p_k(N'+k+1) + (N'+1)\,p_{k+1}(N'+1) \\ &= (N'+1)\,p_{k+1}(N'+k+2) \end{aligned}$$

Therefore we have for unrestricted N_i'

$$E(N_i') = \frac{\sum_{x=0}^{N'} xF_{N',k}(x)}{kp_k(N'+k)} = \frac{N'}{k}\,. \tag{25}$$

But according to our procedure $N_i' \leqq n$ for every i. So the number of sets for $N_i' > n$ being A, we have the same formula for the mean of N_i' as (25)

$$E(N_i') = \frac{\sum_{x=0}^{n} xF^*_{N',k}(x)}{kp_k(N'+k) - kA} = \frac{N'}{k} \tag{26}$$

where F^* denotes the F which excludes the number of cases when $N_i' > n$.

On the other hand,

$$\sum_{x=1}^{n} x^2 F^*_{N',k}(x) < n^2 \sum_{x=1}^{n} F^*_{N',k}(x)$$

$$\therefore \quad \tau^2 = \frac{\sum x^2 F^*}{\sum F^*} - \left(\frac{N'}{k}\right)^2 < \frac{N'^2}{k^2}\left(\frac{N^2}{N'^2} - 1\right) \tag{27}$$

Therefore the relative measure of bias of $\bar{E}(\bar{x}_i')$ in regard to $\bar{X}'$ is

$$\frac{\tau^2}{\bar{N}'^2} < \frac{N^2}{N'^2} - 1 \tag{28}$$

For the preceding example of voters' lists of 4-chome Ikebukuro,

$$N' = 2476 = 0.96\,N$$

$$\therefore \quad \frac{\tau^2}{\bar{N}'^2} < 0.085.$$

§ 4. Case when we should take just n Samples

In § 2, we estimated the mean removing the non-universe elements, but we frequently are obliged to take just n samples. In this case we will fill up at random the missing elements from other sets, that is, when the i-th set is selected, we select at random $n - N_i'$ samples from other $N' - N_i'$ elements. Then the estimate of mean $\bar{x}_i''$ is

$$\bar{x}_i'' = \frac{1}{n}\left(x_i f(i,1) + x_{i+k} f(i,2) + \cdots\cdots + x_{i+(n-1)k} f(i,n)\right) + \frac{1}{n}\left(x_{j1} + x_{j2} + \cdots\cdots + x_{j,\,n-N_i'}\right) \tag{29}$$

and

$$E(\bar{x}_i'') = \frac{1}{kn}\left(N'\bar{X}' + \sum_i \frac{n - N_i'}{N' - N_i'}\left(N'\bar{X}' - N_i'\bar{X}_i'\right)\right) \tag{30}$$

But except when $\bar{X}' = \bar{X}_i'$ or $N_i' = \bar{N}'$, $\bar{x}_i''$ is biased estimate. Let this bias be B_2 and that of $\bar{x}_i'$ in § 2 be B_1. Then,

$$B_1 = \frac{1}{k}\sum_i \bar{X}_i' - \bar{X}' = \sum_i \bar{X}_i'\left(\frac{1}{k} - \frac{N_i'}{N'}\right),$$

$$B_2 = \frac{N'}{N}\bar{X}'\left(1 + \sum_i \frac{n - N_i'}{N' - N_i'}\right) - \frac{1}{N}\sum \frac{n - N_i'}{N' - N_i'} N_i' \bar{X}_i' - \bar{X}'.$$

If we can put $N_i' \fallingdotseq N'k$, then we have

$$B_2 \fallingdotseq \frac{N - N'}{N(k-1)}\sum_i \bar{X}_i'\left(\frac{N_i'}{N'} - \frac{1}{k}\right)$$

so we get

$$|B_2| \leqq |B_1|.$$

And let the correlation between $N_i'\bar{x}_i'$ and $(n - N_i')\bar{x}'_{n-N_i'}$ be ρ and the variance of elements besides i-th set be $\bar{\sigma}_i^2$, we have

$$V(\bar{x}_i'') = \frac{1}{nN}\sum_i N_i'^2\bar{X}_i'^2 - \frac{N'^2}{N^2}\bar{X}'^2 + \frac{1}{n^2}\left\{\frac{1}{k}\sum_i\left(\frac{n-N_i'}{N'-N_i'}(N'\bar{X}'-N_i'\bar{X}_i')\right)^2\right.$$
$$\left.-\left(\frac{1}{k}\sum_i\frac{n-N_i'}{N'-N_i'}(N'\bar{X}'-N_i'\bar{X}_i')\right)^2\right\} + \frac{N'-n}{nN}\sum_i\left(\frac{n-N_i'}{N'-N_i'}\right)\bar{\sigma}_i^2$$
$$+2\rho\sqrt{\frac{1}{nN}\sum N_i'^2\bar{X}_i'^2 - \frac{N'^2}{N^2}\bar{X}'^2}\sqrt{\frac{N'-n}{nN}\sum_i\frac{n-N_i'}{N'-N_i'}\bar{\sigma}_i^2}$$
$$= \frac{1}{n^2}\left\{V(N_i'\bar{X}_i') + V\left(\frac{n-N_i'}{N'-N_i'}(N'\bar{X}'-N_i'\bar{X}_i')\right)\right.$$
$$\left.+\frac{N'-n}{k}\sum_i\frac{n-N_i'}{N'-N_i'}\bar{\sigma}_i^2 + 2\rho\sqrt{V(N_i'\bar{X}_i')}\sqrt{\frac{N'-n}{k}\sum_i\frac{n-N_i'}{N'-N_i'}\bar{\sigma}_i^2}\right\} \tag{31}$$

If we can also put $N_i' \fallingdotseq N'/k$, we have

$$V(\bar{x}_i'') = \frac{N'^2(k-1)^2-(N-N')^2}{N^2(k-1)^2}V(\bar{x}_i') + \frac{(N-N')(N'-n)}{nNN'(k-1)}\sum_i\bar{\sigma}_i^2$$
$$+ (\text{term including } \rho) \tag{32}$$

and when we can neglect the second and third terms in camparison with the first term, we get

$$V(\bar{x}_i'') < V(\bar{x}_i').$$

Therefore we have in this case

$$M.S.E.\ (\bar{x}_i'') < M.S.E.\ (\bar{x}_i')$$

that is, $\bar{x}_i''$ is better estimate than $\bar{x}_i'$.

Reference

(1) W. G. COCHRAN: Sample Survey Techniques, 1948.
(2) K. FUSHIMI: Theory of Probabilities and Statistics, *Kawade Shobō*, Japan, 1942.

Institute of Statistical Mathematics.