PRACTICAL NOTE ON ESTIMATION OF POPULATION MEANS BASED ON SAMPLES STRATIFIED BY MEANS OF ORDERING

KOITI TAKAHASI

(Received Sept. 8, 1970)

1. Introduction

In a previous paper [1] we proposed a method of estimation of population means based on a two-phase sampling: First we draw m^2h elements from a population and classify them into mh groups each of which consists of m elements. Concerning the first h groups we observe only the smallest element for each group. Concerning the next h groups we observe the second smallest element for each group and so on. Concerning the last h groups we observe the largest element for each group. We use the arithmetic mean of these mh observed values as an estimate of the population mean. In order that this method is of practical use it is necessary that we can easily choose the elements having given rank from each group.

In the paper [1] we mainly considered the efficiency of this estimate by comparing the variance of this estimate with that of the sample mean of a simple random sample of size mh. In the paper [2] we considered the effects on this estimate, of the correlation which might exist within groups.

In this paper we treat two problems which seem to be necessary or useful for practical application of this method of estimation. In Section 2 it will be shown that this sampling procedure could be simplified with just a little loss of efficiency in estimation. In the procedure described above we must measure the quantity of the element with a specified rank in advance in each group. On the other hand, in the simplified procedure we may measure the quantity of the element drawn at random from each group and then we record the rank of the element in the group.

Our procedures are of practical use when the identification of the order of magnitude is much easier than the measurement of magnitude. In Section 3 we consider in what cases the identification of the order of magnitude is difficult and how difficult the identification is for some distributions.

2. A simplified procedure

2.1. *Preliminaries*

Let $F_1(x), \cdots, F_m(x)$ and $F(x)$ be cumulative distribution functions and

$$F(x)=\sum_{j=1}^{m} w_j F_j(x) ,$$

where each $w_j>0$ and $\sum_{j=1}^{m} w_j=1$. Let (I, Y) be a random vector having the joint distribution defined by

$$\mathrm{P}[I=j]=w_j$$

and

$$\mathrm{P}[Y\leqq y \mid I=j]=F_j(y) \qquad (j=1, 2, \cdots, m) .$$

Let $Z_1, \cdots, Z_m$ be independent random variables, where Z_j has the distribution $F_j(y)$ $(j=1, \cdots, m)$. Let (I_i, Y_i) $(i=1, \cdots, h)$ be independent random vectors and let each (I_i, Y_i) have the same distribution as (I, Y).

We now consider the problem of estimation of μ, the mean of F, on the basis of observed values on (I_i, Y_i) $(i=1, \cdots, h)$ and Z_j $(j=1, \cdots, m)$. It is obvious that the statistic

$$T=\sum_{j=1}^{m} \frac{w_j}{\sum_{i=1}^{h} \chi_j(I_i)+1}\left(\sum_{i=1}^{h} \chi_j(I_i) Y_i+Z_j\right)$$

is an unbiased estimate of μ, where $\chi_j(I)=1$ if $I=j$ and 0, otherwise. It should be noted that we introduced the Z's only to avoid a trouble which arises when some of $\sum_{i=1}^{h} \chi_j(I_i)$ happen to be zero. Denote $\sum_{i=1}^{h} \chi_j(I_i)$ by n_j $(j=1, \cdots, m)$. Then $(n_1, \cdots, n_m)$ has a multinomial distribution, that is,

$$\mathrm{P}[(n_1, \cdots, n_m)=(k_1, \cdots, k_m)]=\frac{h!}{\prod_{j=1}^{m} k_j!} \prod_{j=1}^{m} w_j^{k_j} .$$

The variance of T can easily be obtained;

$$\begin{aligned}
\mathrm{Var}\, T &= \mathrm{E}(T-\mu)^2 \\
&= \mathrm{E}(\mathrm{E}((T-\mu)^2 \mid (n_1, \cdots, n_m)=(k_1, \cdots, k_m))) \\
&= \mathrm{E}\left(\sum_{j=1}^{m} w_j^2 \frac{\sigma_j^2}{k_j+1}\right) \\
&= \sum_{j=1}^{m} w_j^2 \sigma_j^2 \left[\sum_{k_j=0}^{h} \binom{h}{k_j} w_j^{k_j}(1-w_j)^{h-k_j} \frac{1}{k_j+1}\right]
\end{aligned}$$

$$=\frac{1}{h+1}\sum_{j=1}^{m} w_j\sigma_j^2(1-(1-w_j)^{h+1}) ,$$

where σ_j^2 is the variance of F_j $(j=1,\cdots, m)$.

2.2. *Application to the problem under consideration*

Let $F_{m,k}(x)$ be the cdf of the kth least order statistic of a sample of size m from $F(x)$. Then we have

$$F(x)=\sum_{j=1}^{m}\frac{1}{m}F_{m,j}(x) .$$

Let $X_{11}, X_{12}, \cdots, X_{1m}, X_{21}, \cdots, X_{2m}, \cdots, X_{h1}, \cdots, X_{hm}$ be independent random variables having the same distribution F. Denote by I_i the rank of X_{i1} in $\{X_{i1}, \cdots, X_{im}\}$, that is, the number of X_{ij}'s such that $X_{ij}\leqq X_{i1}$. The joint distribution of (I_i, X_{i1}) satisfies

$$\mathrm{P}\{I_i=j\}=\frac{1}{m} \qquad (j=1, 2, \cdots, m)$$

and

$$\mathrm{P}\{X_{i1}\leqq x \mid I_i=j\}=F_{m,j}(x) .$$

Let $Z_1, \cdots, Z_m$ be independent random variables, where Z_j has the distributioh $F_{m,j}$ $(j=1, \cdots, m)$.

From the results described in 2.1, we can obtain an unbiased estimate T of μ on the basis of $\{(I_i, X_{i1}); i=1, \cdots, h\}$ and $\{Z_j; j=1, \cdots, m\}$;

$$T=\frac{1}{m}\sum_{j=1}^{m}\frac{1}{\sum_{i=1}^{h}\chi_j(I_i)+1}\left(\sum_{i=1}^{h}\chi_j(I_i)X_{i1}+Z_j\right) .$$

The variance of T is given by

$$\mathrm{Var}\, T=\frac{1}{(h+1)m}\sum_{j=1}^{m}\sigma_{m,j}^2\left\{1-\left(1-\frac{1}{m}\right)^{h+1}\right\} .$$

In practice m will be two or three. In these cases $\mathrm{Var}\, T$ becomes to be almost equal to

$$\frac{1}{hm}\sum_{j=1}^{m}\sigma_{m,j}^2 \tag{1}$$

as h becomes large. (1) is the variance of the estimate given by (3.6) of [1].

3. Degrees of difficulty in ordering

It is essential for our estimation procedure that we can easily, in other words, at a glance, find the elements which have the given ranks among several elements. It is difficult to find an appropriate model which describes under what condition we can identify the order of magnitudes of several elements at a glance. In this section we shall consider a simple model. Suppose that we have two elements and we can identify the smaller or larger element when and only when $|x_1-x_2|\geqq\delta$, where x_1 and x_2 denote the magnitudes of two elements and δ is a positive number.

Let X_1 and X_2 be independent random variables having the same absolutely continuous distribution with cdf $F(x)$ and pdf $f(x)$. Let us consider the probability that X_1 and X_2 satisfy $|X_1-X_2|\geqq\delta$. Denote the random variable $|X_1-X_2|$ by D, the cdf of D by $G(x)$ and its density function by $g(x)$. The random variable D is the sample range of a sample of size 2 from $F(x)$. The following results are well known (for example, see [3], [4]):

$$G(x)=2\int_{-\infty}^{\infty}\{F(t+x)-F(t)\}f(t)dt \tag{2}$$

and

$$g(x)=2\int_{-\infty}^{\infty}f(t+x)f(t)dt\,. \tag{3}$$

It is almost obvious that

$$\sup \mathrm{P}(D<\delta)=1 \qquad \text{for each } \delta>0\,, \tag{4}$$

where sup is over all continuous distributions. Furthermore, (4) holds even if sup is over all continuous distributions with variance 1. In fact, let

$$f(x)=\begin{cases}\dfrac{1}{\alpha}-1 & \text{if } 0<x<\alpha<1\,,\\[2ex] 1 & \text{if } \dfrac{12-\alpha^2}{12\alpha(1-\alpha)}<x<\dfrac{12-\alpha^2}{12\alpha(1-\alpha)}+\alpha\,,\\[2ex] 0 & \text{otherwise}\,.\end{cases} \tag{5}$$

It is easily seen that the variance of this distribution is 1 and

$$\mathrm{P}(D<\delta)\geqq(1-\alpha)^2+\alpha^2\,, \qquad \text{if } \alpha<\delta\,.$$

Therefore

$$\sup \mathrm{P}(D<\delta)\geqq\lim_{\alpha\to 0}((1-\alpha)^2+\alpha^2)=1\,,$$

where sup is over all continuous distributions with variance 1.

For small $\delta>0$, we can, however, give an upper bound for $P(D<\delta)$, smaller than 1 over the family of distributions which have bounded density functions. Assume that $f(x)\leqq K$. Then $F(t+x)-F(t)\leqq Kx$ for every t and x. Therefore, from (2) we have

$$G(x)\leqq 2Kx\ .$$

If we further assume that the distribution with cdf $F(x)$ is unimodal and the mode is m, then we have

$$\begin{aligned} G(x) &\leqq 2x\int_{-\infty}^{\infty}\left\{\sup_{0<t<x} f(t+y)\right\} f(y)dy \\ &\leqq 2x\left\{\int_{-\infty}^{m-x} f(y+x)f(y)dy+\int_{m+x}^{\infty} f(y-x)f(y)dy+2xf^2(m)\right\} \\ &\leqq 2x\left\{\int_{-\infty}^{m} f^2(y)dy+\int_{m}^{\infty} f^2(y)dy+2xf^2(m)\right\}\ . \end{aligned}$$

Therefore, for sufficiently small x we have approximately

$$G(x)\leqq 2x\int_{-\infty}^{\infty} f^2(y)dy\ .$$

In general, for sufficiently small x we have approximately

$$G(x)\doteqdot g(0)x\ .$$

Thus from (3) we have

$$G(x)\doteqdot 2x\int_{-\infty}^{\infty} f^2(t)dt\ .$$

Therefore, we are interested in the values of $\int_{-\infty}^{\infty} f^2(t)dt$.

Considering the distribution defined by (5) we have

$$\sup\int_{-\infty}^{\infty} f^2(t)dt=\infty\ ,$$

where sup is over all continuous distributions with variance 1.

On the other hand, we can obtain the minimum of $\int_{-\infty}^{\infty} f^2(t)dt$ for the family of all absolutely continuous distributions with a given bounded carrier. Let $f(x)$ be a density function such that $f(x)=0$ if $x\notin[a,b]$ and let $f_0(x)$ be the pdf of the rectangular distribution on $[a,b]$:

$$f(x)=\begin{cases} \dfrac{1}{b-a} & \text{if } x\in[a,b] \\ 0 & \text{otherwise}\ . \end{cases}$$

Then we have

$$\begin{aligned}\int_a^b f^2(x)dx &= \int_a^b \{f_0+(f-f_0)\}^2 dx \\ &= \int_a^b f_0^2(x)dx + \int_a^b (f-f_0)^2 dx + 2f_0(1-1) \\ &= \int_a^b f_0^2 dx = \frac{1}{b-a}\,.\end{aligned}$$

Example 1. (i) (cf. [4]) For normal distributions with variance 1 we have

$$g(x)=\begin{cases} \dfrac{1}{\sqrt{\pi}}e^{-x^2/4} & \text{if } x\geqq 0 \\ 0 & \text{otherwise}\end{cases}$$

and

$$\begin{aligned}G(x) &= 2\Phi\left(\frac{x}{\sqrt{2}}\right)-1 \\ &= \operatorname{erf}\left(\frac{x}{2}\right),\end{aligned}$$

where $\Phi(x)$ is the cdf of the standard normal distribution and $\operatorname{erf}(x)$ is the error function, that is, $\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$.

(ii) For rectangular distributions with variance 1 we have

$$g(x)=\begin{cases} -\dfrac{x}{6}+\dfrac{1}{\sqrt{3}} & \text{if } 0<x<\sqrt{12} \\ 0 & \text{otherwise}\,,\end{cases}$$

and

$$G(x)=\frac{x}{\sqrt{12}}\left(2-\frac{x}{\sqrt{12}}\right).$$

(iii) For the exponential distribution with variance 1, we have

$$g(x)=e^{-x}$$

and

$$G(x)=1-e^{-x} \quad \text{if } x>0\,.$$

(iv) For the Gamma distribution with the density function

$$f(x)=\begin{cases} 2xe^{-\sqrt{2}x} & \text{if } x>0 \\ 0 & \text{otherwise}\,,\end{cases}$$

we have

$$g(x)=\frac{1}{\sqrt{2}}e^{-\sqrt{2}x}+xe^{-\sqrt{2}x}\,.$$

Example 2. For the distributions of Example 1 (i), (ii) and (iii) the values of $\int_{-\infty}^{\infty} f^2(t)dt$ are given by $\frac{1}{2\sqrt{\pi}}$, $\frac{1}{2\sqrt{3}}$ and $\frac{1}{2}$, respectively. For the Gamma distributions with variance 1 we have

$$f(x)=\frac{\alpha^p}{\Gamma(p)}x^{p-1}e^{-\alpha x}\,,\qquad x>0\,,$$

where $\alpha=p^{1/2}$, and

$$\int_0^{\infty} f^2(t)dt=\frac{\Gamma(2p-1)}{\Gamma^2(p)}p^{1/2}\left(\frac{1}{2}\right)^{2p-1}\,,\qquad \text{for } p>\frac{1}{2}\,.$$

For the Weibull distributions with variance 1 we have

$$f(x)=\frac{b}{\theta}x^{b-1}e^{-x^b/\theta}\,,\qquad x>0\,,$$

where $\theta=\left(\Gamma\left(1+\frac{2}{b}\right)-\Gamma\left(1+\frac{1}{b}\right)^2\right)^{-b/2}$, and

$$\int_0^{\infty} f^2(t)dt=\frac{b}{2^{2-(1/b)}}\Gamma\left(2-\frac{1}{b}\right)\left(\Gamma\left(1+\frac{2}{b}\right)-\Gamma^2\left(1+\frac{1}{b}\right)\right)^{1/2}\,,\qquad \text{for } b>\frac{1}{2}\,.$$

For triangular distributions with variance 1 we have

$$f(x)=\begin{cases}\dfrac{h}{a}x+h & \text{if } -a\leqq x\leqq 0\\[2ex] -\dfrac{h}{b}x+h & \text{if } 0\leqq x\leqq b\,,\end{cases}$$

where $b=(\sqrt{72-3a^2}-a)/2$, $h=2/(a+b)$ and $0\leqq a\leqq\sqrt{18}$, and

$$\int_{-a}^{b} f^2(x)dx=\frac{8}{3(a+\sqrt{72-3a^2})}\,.$$

In Table 1 we show numerical values of $G(x)$ and in Table 2 the values of $g(0)$ for some distributions.

Table 1. Values of $G(x)$

Distrib. \ x	0.02	0.04	0.06	0.08	0.1	0.2	0.3	0.4	0.5
Normal	.0113	.0226	.0338	.0451	.0564	.1125	.1680	.2227	.2763
Expon.	.0198	.0392	.0582	.0769	.0952	.1813	.2592	.3297	.3935
Rectan.	.0115	.0230	.0343	.0457	.0569	.1121	.1657	.2176	.2678

Table 2. Values of $g(0)$

Distrib.	Normal	Expon.	Rectang.	Gamma ($p=2$)	Weibull ($b=2$)	Triang. ($a=0$)	($a=1$)	($a=\sqrt{6}$)	($a=2$)
$g(0)$	.5642	1.0000	.5774	.7071	.5806	.6285	.5731	.5443	.5472

4. Concluding remarks

(a) From the consideration in Section 2, for large h, we may estimate the population means with a little loss of efficiency in comparison with the original procedure described in [1] as follows:

Draw a random sample $(x_1, \cdots, x_h)$ of size h from a population. For each x_i draw a sample $(y_{i2}, \cdots, y_{im})$ of size $m-1$ from the population and find the rank of x_i in $(x_i, y_{i1}, \cdots, y_{im})$. Let n_j be the number of x_i's which have the rank j. Estimate the population mean by

$$\frac{1}{m}\sum_{j=1}^{m}\frac{1}{n_j}(\Sigma^{(j)}x_i)\ ,$$

where $\Sigma^{(j)}$ means the summation over the x_i's which have the rank j.

(b) In the distributions considered above, the normal, exponential, rectangular, triangular and the gamma and Weibull with the shape parameter larger than 1, we can say that the values of $G(x)$ are between $0.5x$ and x for $x<0.5$.

(c) Even if the consideration is restricted to the family of distributions having the same variance, yet the distributions whose density functions are highly concentrated somewhere will be unfavorable for the selection of an element with a given rank in several elements at a glance.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Takahasi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering, *Ann. Inst. Statist. Math.*, **20**, 1-31.

[2] Takahasi, K. (1969). On the estimation of the population mean based on ordered samples from an equicorrelated multivariate distribution, *Ann. Inst. Statist. Math.*, **21**, 249-255.

[3] Kendall, M. G. and Stuart, A. (1958). *The advanced theory of statistics*, Vol. 1, Charles Griffin, London.

[4] Gumbel, E. J. (1958). *Statistics of Extremes*, Columbia University Press, New York.