SANKHYĀ

THE INDIAN JOURNAL OF STATISTICS

Edited by : P. C. MAHALANOBIS

SERIES B, VOL. 24 | JANUARY 1962 | PARTS 1 & 2

VARIANCE AND CONFIDENCE INTERVAL ESTIMATION

By M. N. MURTHY

Indian Statistical Institute

SUMMARY. In this paper some results of the investigations on the efficiencies of the different methods of estimation of variance of the estimate and of setting up of confidence intervals for the population parameter in large-scale sample surveys are given. The efficiencies of setting up confidence intervals based on different methods of estimating the variance of the estimate have been studied with respect to a number of criteria such as the expected value and the distribution of the length of the confidence interval. It is shown that the efficiency of the confidence interval based on the sub-sample estimates approaches that of the confidence interval obtained by conventional methods more rapidly for initial increases in the number of sub-samples than for further increases. The results have been shown to be valid in case of stratified sampling with a few sub-samples in each stratum.

1. INTRODUCTION

1.1. Some of the results of the investigations on (i) methods of estimation of variance of the estimate in large-scale sample surveys and (ii) methods of setting up confidence interval for the population parameter will be given in this paper. As there is an abundance of literature on these two topics, it may not be out of place here to give a brief review of the work that has already been done. This review is by no means exhaustive.

1.2. Here the aim is to study the efficiencies of methods of estimation of variance and of setting up confidence intervals which are operationally convenient. Invariably, such methods are less efficient than the conventional methods involving much calculation at the stage of analysis. Sometimes it may be possible to strike a balance between the efficiency aimed at and the labour involved.

2. METHODS OF ESTIMATION OF VARIANCE

2.1. In the case of simple random sampling from a normal population, the variance of the estimate of μ, the mean in the population, involves the parameter σ, the population standard deviation. Let a simple random sample of size N be drawn from a normal population with mean μ and standard deviation σ. Let the observations be $X_1, X_2, \ldots, X_N$. The minimum variance estimate among the class of unbiased estimates of μ is $\bar{X}$, the sample mean and its standard error is $\sigma/\sqrt{N}$. A list of estimates of σ available in statistical literature is given below.

$$s_1 = \sqrt{\sum_{i=1}^{N} \frac{(X_i - \bar{X})^2}{N}} \quad \ldots \quad (1)$$

$$s_2 = \frac{1}{c_2}\sqrt{\sum_{i=1}^{m} \frac{(x_i - \bar{x})^2}{m}} \quad \ldots \quad (2)$$

$$s_3 = \frac{1}{c_2}\sqrt{n \sum_{i=1}^{m} \frac{(\bar{x}_i - \bar{X})^2}{m}} \quad \ldots \quad (3)$$

$$s_4 = \frac{\bar{w}}{d_2} \quad \ldots \quad (4)$$

$$s_5 = \frac{\lambda_2 - \lambda_1}{u_2 - u_1} \quad \ldots \quad (5)$$

where $\bar{x}$ is the mean of a sub-sample of size m, $\bar{x}_i$ is the mean of the i-th random group with n observations such that $nm = N$, $(i = 1, 2, \ldots, m)$; $\bar{w}$ is the mean of the ranges in sub-group of n elements in each ; λ_1 and λ_2 are two numbers to be properly chosen ; u_1 and u_2 are given by

$$p = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u_1} e^{-x^2/2}\, dx \text{ and } p+q = \frac{1}{\sqrt{2\pi}} \int_{u_1}^{\infty} e^{-x^2/2}\, dx$$

where p and q are proportions of the observations less than λ_1 and between λ_1 and λ_2 respectively ;

$$c_2 = \sqrt{\frac{2}{m}}\, \frac{\Gamma\left(\frac{m}{2}\right)}{\Gamma\left(\frac{m-1}{2}\right)}$$

and

$$d_2 = \int_{-\infty}^{\infty} \{1 - \alpha_1^n - (1-\alpha_1)^n\}\, dx_1$$

where $\alpha_1 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_1} e^{-x^2/2}\, dx$ and x_1 is the smallest observation in the sample. The values of c_2 and d_2 are tabulated for different values of m and n in the manual of the American Society for Testing Materials (1951).

2.2. Of these estimates, as is to be expected, s_1 is the most efficient and the most difficult to calculate. In sampling from normal populations, the estimates s_2 and s_3 have the same efficiency. Hansen, Hurwitz and Madow (1953) have compared the variances of the estimates of the type s_2^2 and s_3^2 namely $s_2'^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})^2$ and

$s_3'^2 = \frac{n}{m-1} \sum_{i=1}^{m} (\bar{x}_i - \bar{X})^2$ and observe that $s_2'^2$ is more or less precise than $s_3'^2$ according as β is less than or greater than 3 where $\beta = \frac{\mu_4}{\mu_2^2}$. The expressions for the variance of $s_2'^2$ and $s_3'^2$ are

$$V(s_2'^2) = \left(\beta - \frac{m-3}{m-1} \right) \frac{\mu_2^2}{m} \qquad \ldots \quad (6)$$

$$V(s_3'^2) = \left(\beta_n - \frac{m-3}{m-1} \right) \frac{\mu_2^2}{m} \qquad \ldots \quad (7)$$

where $\beta_n = \frac{\beta}{n} + 3\,\frac{n-1}{n}$. Hence it follows that $V(s_2'^2) \gtreqless V(s_3'^2)$ according as $\beta \gtreqless 3$.

2.3. If the size of the sub-group is small (about 7 or 8 observations), then the loss of efficiency in using s_4 instead of s_1 as an estimate of σ is not large (Pearson and Haines, 1935). Pearson (1932) has tabulated the mean, standard deviation and percentage limits (0.5%, 1%, 5% and 10%) of range in samples from a normal population for sample sizes 2(1) 30(5) 100. Cadwell (1954) has given an asymptotic expression for the probability integral of range of samples from a symmetrical unimodal population and has studied its accuracy for the case of normal parent population and for sample sizes 20 to 100. Stevens (1948) suggested the estimate s_5 and tabulated the efficiency of this estimate as compared to that of s_1 in large samples for different values of $\frac{\lambda_1 - \mu}{\sigma}$ and $\frac{\lambda_2 - \mu}{\sigma}$, while sampling from a normal population with mean μ and standard deviation σ.

2.4. An empirical study was conducted to study the efficiency of s_5 as compared to that of s_1 for a sample of size 100 from a normal population. For this purpose the samples from a normal population with mean 0 and standard deviation 1 given by Mahalanobis and others (1934) have been used. There are 104 samples of size 100. For each of these the mean, standard deviation and frequency distribution have been given. The mean and variance of the sample standard deviations are 0.9887 and 0.0049 respectively. Taking λ_1 and λ_2 to be -0.5 and 0.5, for each sample, s_5 was calculated. The mean and variance of s_5 turned out to be 1.0009 and 0.0193 respectively. Hence s_5 can be considered to be unbiased for this sample size and the efficiency of s_5 as compared to that of s_1 is 25% which agrees with the figure given by Stevens. The efficiency of s_5 can be increased by taking the values of λ_1 and λ_2 near about the mean μ on either side of it.

3. Interpenetrating sub-samples

3.1. In a stratified sampling design where n independent and interpenetrating sub-samples are taken from each stratum, an estimate of the variance of the estimate can be obtained by using (i) the sub-sample estimates of total or (ii) the sub-sample estimates of strata totals. It may be of interest to get an expression for the loss of efficiency in using the former in preference to the latter.

3.2. Let there be k strata and n independent and interpenetrating sub-samples in each stratum. For the sake of simplicity let the sub-sample sizes within each stratum be the same. Suppose $\hat{y}_{ij}$ is an unbiased estimate of the j-th stratum total y_j from the i-th sub-sample ($i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$). The two estimates of the variance of the estimate $\hat{y}$ of the total y are

$$\text{(i)} \quad \hat{V}_1(\hat{y}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{y}_{i\cdot} - \hat{y})^2 \qquad \ldots \quad (8)$$

and

$$\text{(ii)} \quad \hat{V}_2(\hat{y}) = \frac{1}{n(n-1)} \sum_{j=1}^{k} \sum_{i=1}^{n} (\hat{y}_{ij} - \hat{\bar{y}}_{\cdot j})^2 \qquad \ldots \quad (9)$$

where

$$\hat{y}_{i\cdot} = \sum_{j=1}^{k} \hat{y}_{ij}, \quad \hat{\bar{y}}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_{ij} \quad \text{and} \quad \hat{y} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_{i\cdot} \qquad \ldots \quad (10)$$

3.3. It can be easily verified that the above two estimates of the variance are unbiased. The variances of the two estimates are given by

$$V\{\hat{V}_1(\hat{y})\} = \frac{1}{n^3(n-1)} \left[\sum_{j=1}^{k} \{(n-1)\,\mu_{4j} + (3-n)\,\mu_{2j}^2\} + 4n \sum_{j=1}^{k} \sum_{i>j} \mu_{2i}\,\mu_{2j} \right] \qquad \ldots \quad (11)$$

and

$$V\{\hat{V}_2(\hat{y})\} = \frac{1}{n^3(n-1)} \left[\sum_{j=1}^{k} \{(n-1)\mu_{4j} + (3-n)\mu_{2j}^2\} \right] \qquad \ldots \quad (12)$$

where μ_{2j} and μ_{4j} are the second and fourth moments of the estimate $\hat{y}_{ij}$. From the above expressions it follows that $V\{\hat{V}_1(\hat{y})\} > V\{\hat{V}_2(\hat{y})\}$. The loss of efficiency in using $\hat{V}_1(\hat{y})$ instead of $\hat{V}_2(\hat{y})$ as an estimate of $V(\hat{y})$ is given by

$$L = \frac{V\{\hat{V}_1(\hat{y})\} - V\{\hat{V}_2(\hat{y})\}}{V\{\hat{V}_2(\hat{y})\}} = \frac{n}{n-1} \cdot \frac{4 \sum_{j=1}^{k} \sum_{i>j} \mu_{2j}\mu_{2i}}{\sum_{j=1}^{k} \left(\beta_j - \frac{n-3}{n-1}\right) \mu_{2j}^2} \qquad \ldots \quad (13)$$

where $\beta_j = \frac{\mu_{4j}}{\mu_{2j}^2}$. If the distribution of the estimates within each stratum can be assumed to be normal, then $\beta_j = 3$ for all j. Hence L becomes

$$L = \frac{2 \sum_{j=1}^{k} \sum_{i>j} \mu_{2i}\,\mu_{2j}}{\sum_{j=1}^{k} \mu_{2j}^2} = \frac{\mu_2^2}{\sum_{j=1}^{k} \mu_{2j}^2} - 1 \qquad \ldots \quad (14)$$

where $\mu_2 = \sum_{j=1}^{k} \mu_{2j}$. If the coefficient of variation of the estimate in each of the strata can be assumed to be equal, then L is given by

$$L = \frac{2 \sum_{j=1}^{k} \sum_{i>j} y_i^2\, y_j^2}{\sum_{j=1}^{k} y_j^4}\,. \qquad \ldots \quad (15)$$

Instead, if it is assumed that the variance of the estimate in each stratum is the same, then L is equal to $k-1$. It may be noticed that the loss may be substantial if the number of strata is large.

4. Confidence interval estimation

4.1. If a sample of size N is drawn from a normal population with mean μ and standard deviation σ, then the confidence interval for μ is given by

$$P\left\{\bar{X}-t_{\alpha}\frac{\sigma}{\sqrt{N}} < \mu < \bar{X}+t_{\alpha}\frac{\sigma}{\sqrt{N}}\right\} = 1-\alpha \qquad \dots \quad (16)$$

where $1-\alpha$ is the confidence coefficient and t_{α} is the $\alpha\%$ point of the distribution of $\frac{\bar{X}-\mu}{\sigma}\sqrt{N}$. In practice one has to estimate σ from the sample itself by one of the procedures given in Section 2.

4.2. If s_1 is taken as an estimate of σ, then it is well known that the statistic

$$t = \frac{\bar{X}-\mu}{s_1}\sqrt{N-1} \qquad \dots \quad (17)$$

is distributed as Student's t with $N-1$ degrees of freedom. Of course, for large samples the above statistic is distributed normally with mean 0 and standard deviation 1. Similarly the statistics

$$t' = \lambda' \frac{(\bar{X}-\mu)}{s_2} \text{ and } t'' = \lambda'' \frac{(\bar{X}-\mu)}{s_3} \qquad \dots \quad (18)$$

are also distributed as Student's t with $(n-1)$ degrees of freedom, where n is the number of groups or sub-sample size and λ' and λ'' are constants.

4.3. Daly (1946) has proved that $\bar{x}$ and w, the mean and range of sample of N independent observations on a normally distributed variate x are statistically independent. Lord (1947) has given the 5% and 1% points of the distribution of the statistic

$$u = \frac{(\bar{X}-\mu)}{\bar{w}} d_2\sqrt{nm} \qquad \dots \quad (19)$$

where n is the sub-group size and m the number of sub-groups. Patnaik (1950) has obtained an approximation to the distribution of and making use of this has derived the distribution of u. Jackson and Ross (1955) have transformed the tables of Lord so as to provide the percentage points of the distribution of the statistic

$$G_1 = \frac{(\bar{X}-\mu)}{\bar{w}} = \frac{u}{d_2\sqrt{nm}}. \qquad \dots \quad (20)$$

Noether (1955) has considered the statistics

$$G_1 = \frac{(\bar{X}-\mu)}{\bar{w}} \quad \text{and} \quad G_2 = \frac{|\bar{X}_1-\bar{X}_2|}{\bar{w}^1},$$

where $\bar{w}^1$ is the mean of the ranges of all sub-groups of both the samples, and has given the percentage points for G_1 and G_2 so that confidence intervals for μ and $(\mu_1-\mu_2)$ can be set up in the form

$$P\{\bar{X}-g_{1\alpha}\bar{w} < \mu < \bar{X}+g_{1\alpha}\bar{w}\} = 1-\alpha \qquad \text{... (21)}$$

and

$$P\{(\bar{X}_1-\bar{X}_2)-g_{2\alpha}\bar{w}^1 < (\mu_1-\mu_2) < (\bar{X}_1-\bar{X}_2)+g_{2\alpha}\bar{w}^1\} = 1-\alpha \qquad \text{... (22)}$$

where $g_{1\alpha}$ and $g_{2\alpha}$ are the α% points of the distributions of G_1 and G_2. Further he has tabulated the values of a_N for different values of n the sub-group size and m the number of sub-group $(nm=N)$ which when multiplied by the sum of the ranges in the sub-groups provides us an unbiased estimate of σ.

4.4. Let $x_1, x_2 \ldots, x_n$ be independent observations on a variate x with some distribution function, arranged in the increasing order of magnitude. Thompson (1936) has shown that

$$P(X_k < M < X_{n-k+1}) = 1-2I_{0.5}\,(n-k+1, k) \qquad \text{... (23)}$$

where M is the median in the population and $I_x(p, q)$ is the incomplete Beta function $\frac{1}{\beta(p,q)} \int_0^x y^{p-1}\,(1-y)^{q-1}\,dy$ which has been tabulated by Karl Pearson. If the distribution of x is symmetrical, then the above expression gives us the confidence region for the population mean. Nair (1940) has tabulated the values of k which give us confidence intervals with confidence coefficient greater than or equal to 0.95 and 0.99 for values of $n = 6(1)\,81$.

4.5. In what follows the efficiencies of the confidence intervals based on s_1 and s_2 will be compared. For the sake of convenience let us redefine s_1^2 and s_2^2 as

$$s_1^2 = \frac{1}{N(N-1)} \sum_{i=1}^{N} (x_i-\bar{X})^2 \qquad \text{... (24)}$$

and

$$s_2^2 = \frac{1}{N(n-1)} \sum_{i=1}^{n} (x_i-\bar{x})^2 \qquad \text{... (25)}$$

where $x_1\, x_2 \ldots, x_N$ are the N observations drawn from a normal population with mean μ and standard deviation σ and n is the sub-sample size. It is clear that $N(N-1)\,\frac{s_1^2}{\sigma^2}$ and $N(N-1)\,\frac{s_2^2}{\sigma^2}$ are distributed as χ^2 with $N-1$ and $n-1$ degrees of freedom respectively. Hence it follows that the statistics

$$t_1 = \frac{(\bar{X}-\mu)}{s_1} \quad \text{and} \quad t_2 = \frac{(\bar{X}-\mu)}{s_2} \qquad \text{... (26)}$$

are distributed as Student's t with $N-1$ and $n-1$ degrees of freedom respectively. If $t_{1\alpha}$ and $t_{2\alpha}$ are the $\alpha\%$ limits of the distribution of t_1 and t_2, then the lengths of the confidence intervals by the two methods will be

$$L_1 = 2t_{1\alpha}s_1 \quad \text{and} \quad L_2 = 2t_{2\alpha}s_2.$$

4.6. A number of criteria can be suggested for comparing the efficiencies of L_1 and L_2. L_1 and L_2 may be said to have approximately the same efficiency if $E(L_2)$ and $V(L_2)$ are nearly equal to $E(L_1)$ and $V(L_1)$ respectively. It is to be noted that $E(L_2)$ and $V(L_2)$ tend to $E(L_1)$ and $V(L_1)$ respectively as n tends to N. But the convergence after a certain stage becomes slow in the case of the expected value. The expected values are given by

$$E(L_1) = 2t_{1\alpha}.\frac{\sigma}{\sqrt{N}} \qquad \text{if } N \text{ is large } (>25) \qquad \dots \quad (27)$$

and

$$E(L_2) = 2t_{2\alpha}\, c_2^1 \; \frac{\sigma}{\sqrt{N}} \qquad \text{where } c_2^1 = \sqrt{\frac{n}{n-1}}\; c_2. \qquad \dots \quad (28)$$

If N is fairly large (>100) then $t_{1\alpha} = 1.96$, for in that case t_1 is distributed normally with mean 0 and standard deviation unity. Table 1 gives the values of the ratio $E(L_2)/E(L_1)$ for different values of n, assuming N to be large.

TABLE 1. VALUES OF THE RATIO OF THE EXPECTED VALUE OF L_2 TO THAT OF L_1 FOR DIFFERENT VALUES OF n

n	2	3	4	5	6	7	8	9	10	15	20	25
$\frac{E(L_2)}{E(L_1)}$	5.172	1.946	1.496	1.331	1.248	1.198	1.164	1.141	1.123	1.075	1.054	1.042

4.7. The confidence interval L_1 and L_2 may be said to have approximately the same efficiency if L_2^* is nearly equal to L_1^* where L_1^* and L_2^* are given by

$$P\{L_1 \leqslant L_1^*\} = 0.95 \qquad \dots \quad (29)$$

$$P\{L_2 \leqslant L_2^*\} = 0.95. \qquad \dots \quad (30)$$

This criterion is defective in the sense that even if L_2^* is nearly equal to L_1^* at this level of confidence, this may not be true for some other level. A better approach is to compare the distribution functions of L_1 and L_2 for different values of n. Here also it may be observed that the convergence of the distribution function of L_2 to that of L_1 is likely to become very slow for values of n greater than a certain value. Table 2 gives the values of L_1^* and L_2^* for different values of N and n. Table 3 shows the distribution function of L_1 for $N = 100$ and that of L_2 for $n = 4$, 5, 10, 20 and 40.

TABLE 2. COMPARISON OF THE VALUES OF L_1^* AND L_2^* FOR DIFFERENT VALUES OF N AND n AT 95% CONFIDENCE LEVEL

N	L_1^*	values of L_2^*							
		$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=8$	$n=10$	$n=24$
96	0.223	2.559	0.750	0.514	*	0.380	0.332	*	0.261
120	0.198	2.273	0.680	0.469	0.380	0.349	0.306	0.283	0.234
200	0.150	1.761	*	0.363	0.302	*	0.237	0.210	0.181

TABLE 3. COMPARISON OF THE DISTRIBUTION FUNCTIONS OF L_1 AND L_2 FOR $N = 100$, $n = 4, 5, 10, 20$ AND 40

L	$P(L_1 < L)$	$P(L_2 < L)$				
		$n=4$	$n=5$	$n=10$	$n=20$	$n=40$
0.22	0.0000	0.1204	0.1298	0.1053	0.0599	0.0200
0.24	0.0040	0.1644	0.1734	0.1701	0.1379	0.0832
0.26	0.1128	0.1998	0.2101	0.2541	0.2570	0.2304
0.28	0.5902	0.2352	0.2709	0.3541	0.4111	0.4551
0.30	0.9522	0.2760	0.3278	0.4591	0.5772	0.6930
0.32	0.9991	0.3218	0.3833	0.5621	0.7252	0.8055
0.34	1.0000	0.3663	0.4416	0.6641	0.8401	0.9547

4.8. As it is easier to compute s_2 than s_1, the object should be to find sub-sample size required to give us L_2 which does not differ much from L_1. In other words the sub-sample size should be so chosen that the variation of L_2 about L_1 is not much keeping an eye on the labour involved. L_2 may be said to be approximately as efficient as L_1, if

$$P\left(\left|\frac{L_2}{L_1}-1\right| < \delta\right) \qquad \dots \quad (31)$$

is fairly large where δ is a small quantity. To find this probability we require the distribution of (L_2/L_1), which can be deduced from a general theorem given by Rao (1953) in the theory of least squares. The result required for our purpose is quoted in the form of a theorem and proved for completeness.

Theorem: *If $x_1, x_2, \ldots, x_N$ be N observations on a variate x which is normally distributed with mean μ and standard deviation σ, then the distribution of*

$$Z = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}{\sum_{i=1}^{N} (x_i - \bar{x}_N)^2}, \qquad \ldots \quad (32)$$

is that of a Beta variate with parameters $\frac{n-1}{2}$ and $\frac{N-n}{2}$, where $\bar{x}_n$ is the mean of the first n observations and $\bar{x}_N$ is the mean of all the N observations.

Proof: $$\sum_{i=1}^{N} (x_i - \bar{x}_N)^2 = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 + \sum_{i=n+1}^{N} (x_i - \bar{x}_{N-n})^2 + n(\bar{x}_n - \bar{x}_N)^2 + (N-n)(\bar{x}_{N-n} - \bar{x}_N)^2$$

where $$\bar{x}_{N-n} = \frac{1}{N-n} \sum_{i=n+1}^{N} x_i.$$

Since $$\bar{x}_N = \frac{n\bar{x}_n + (N-n)\bar{x}_{N-n}}{N},$$

$$\bar{x}_n - \bar{x}_N = \frac{N-n}{N} (\bar{x}_n - \bar{x}_{N-n}) \quad \text{and} \quad \bar{x}_{N-n} - \bar{x}_N = \frac{n}{N} (\bar{x}_{N-n} - \bar{x}_n).$$

Hence $$n(\bar{x}_n - \bar{x}_N)^2 + (N-n)(\bar{x}_{N-n} - \bar{x}_N)^2 = \frac{n(N-n)}{N} (\bar{x}_n - \bar{x}_{N-n})^2$$

which when divided by σ^2 is a χ^2 with one degree of freedom, for

$$\sqrt{\frac{n(N-n)}{N\sigma^2}} (\bar{x}_n - \bar{x}_{N-n}) \quad \text{is} \quad N(0, 1).$$

Further, $\sum_{i=1}^{n} (x_i - \bar{x}_n)^2$ and $\sum_{i=n+1}^{N} (x_i - \bar{x}_{N-n})^2$ are $\chi^2\sigma^2$ with $n-1$ and $N-n-1$ degrees of freedom respectively. Z can be written as

$$Z = \frac{\chi^2_{n-1}}{\chi^2_{n-1} + \chi^2_{N-n}}. \qquad \ldots \quad (33)$$

In this case χ^2_{n-1} and χ^2_{N-n} are independent. Hence the distribution of Z is a Beta distribution with parameters $\frac{n-1}{2}$ and $\frac{N-n}{2}$.

$$P\left\{\frac{L_2}{L_1} < \delta\right\} = P\left\{\frac{L_2^2}{L_1^2} < \delta^2\right\} = P\left\{Z < \frac{n-1}{N-1} \cdot \frac{t^2_{1\alpha}}{t^2_{2\alpha}} \delta^2\right\} = I_x(p,q) \quad \ldots \quad (34)$$

where $$x = \frac{n-1}{N-1} \cdot \frac{t_{1\alpha}^2}{t_{2\alpha}^2} \delta^2, \quad p = \frac{n-1}{2}, \quad q = \frac{N-n}{2}$$

and $$I_x(p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^x y^{p-1}(1-y)^{q-1}\, dy.$$

TABLE 4. GIVING THE DISTRIBUTION FUNCTION OF L_2/L_1 FOR $N = 100$ AND $n = 10, 20, 25$

δ	$P\left\{\frac{L_2}{L_1} < \delta\right\}$		
	$n = 10$	$n = 20$	$n = 25$
.2	0.0002	—	—
.4	.0011	—	—
.6	.0184	0.0013	0.0003
.8	.1053	.0498	.0325
1.0	.3198	.3579	.3692
1.2	.6184	.8190	.8703
1.4	.8593	.9870	.9968
1.6	.9630	.9909	1.0000
1.8	.9944	1.0000	
2.0	.9995		
2.2	1.0000		

5. Stratified sampling

5.1. Let us now consider a case where the population is divided into strata and from each stratum n independent and interpenetrating sub-samples have been selected. Let y_{ij} be an unbiased estimate of μ_j, the j-th stratum total from the i-th sub-sample ($j = 1, 2, \ldots, k$; $i = 1, 2, \ldots, n$). The object is to set up confidence interval for $\mu = \sum_{j=1}^{k} \mu_j$, the population total. For this two methods have been suggested and their efficiencies compared.

5.2. Let us assume that y_{ij} is distributed normally with mean μ_j and standard deviation σ_j. Then an unbiased estimate of μ is given by $y = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} y_{ij}$. In fact y is distributed normally with mean μ and variance $\frac{1}{n} \sum_{j=1}^{k} \sigma_j^2$. The following two estimates of this variance can be considered.

$$\text{(i)} \quad s_1^2 = \frac{1}{n(n-1)} \sum_{j=1}^{k} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 \qquad \ldots \quad (36)$$

and

$$\text{(ii)} \quad s_2^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} (y_{i\cdot} - y)^2 \qquad \ldots \quad (37)$$

where

$$\bar{y}_j = \frac{1}{n} \sum_{i=1}^{n} y_{ij} \quad \text{and} \quad y_{i\cdot} = \sum_{j=1}^{k} y_{ij}.$$

The variance of these estimates have been compared in Section 3. If $\sigma_j = \sigma$ for all j, then $n(n-1) \frac{s_1^2}{k\sigma^2}$ and $n(n-1) \frac{s_2^2}{k\sigma^2}$ are distributed as χ^2 with $k(n-1)$ and $(n-1)$ degrees of freedom. Hence the statistics

$$t_1 = \frac{y-\mu}{s_1} \quad \text{and} \quad t_2 = \frac{y-\mu}{s_2} \qquad \ldots \quad (38)$$

will be distributed as Student's t with $(n-1)k$ and $(n-1)$ degrees of freedom respectively. If $t_{1\alpha}$ and $t_{2\alpha}$ are the α percentage points of t distribution with $k(n-1)$ and $(n-1)$ degrees of freedom respectively, then the lengths of the confidence intervals based on s_1 and s_2 are given by

$$L_1 = 2t_{1\alpha}s_1 \qquad \text{and} \qquad L_2 = 2t_{2\alpha}s_2.$$

If k is fairly large $t_{1\alpha} = 1.96$ and the Table 1 gives the ratio of $E(L_2)/E(L_1)$ for different values of n.

References

ASTM Committee (1951): *Manual on Quality Control of Materials*, American Society for testing materials.

Cadwell, J. H. (1954): The probability integral of range from a symmetrical unimodal population. *AMS*, 25, 803-806.

Daly, J. F. (1946): The use of sample range in an analogue of Student's t test. *AMS*, 17, 71-74.

Evans, W. D. (1951): On the variance of estimates of standard deviation and variance. *J. Amer. Stat. Ass.*, 46, 220-224.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953): *Sample Survey Methods and Theory*, Vol. 1, Ch. 10, John Wiley & Sons, New York.

Jackson, J. E. and Ross, E. L. (1955): Extended tables for use with the G-test. *J. Amer. Stat. Ass.*, 50, 416-433.

Lord, E. (1947): The use of range instead of standard deviation in the t test. *Biometrika*, **34**, 41-67.

Mahalanobis, P. C. and others (1934): Tables of random samples from a normal population. *Sankhyā*, **1**, 289.

Nair, K. R. (1940): Tables of confidence interval for the median in samples from any continuing population. *Sankhyā*, **4**, 551-558.

Noether, G. E. (1955): Use of the range instead of the standard deviation. *J. Amer. Stat. Ass.*, **50**, 1040-1055.

Patnaik, P. B. (1950): The use of mean range as an estimate of variance in statistical tests. *Biometrika*, **37**, 78-87.

Pearson, E. S. (1932): The percentage limits to the distribution of range in samples from a normal population. *Biometrika*, **24**, 404-417.

Pearson, E. S. and Haines, J. (1935): The use of range in place of standard deviation in small samples. Supplement to *J. Roy. Stat. Soc.*, **2**, 83-98.

Rao, C. R. (1953): On transformations useful in the distribution problems of least squares. *Sankhyā*, **12**, 339.

Stevens, W. L. (1948): Control by Gauging. *J. Roy. Stat. Soc.*, Series B, **10**, 54-98.

Thompson, W. R. (1936): On confidence ranges for the median and other expectation distributions for populations of unknown distribution form. *AMS*, **7**, 122-128.

Paper received: October, 1960.