A NOTE ON THE ESTIMATION OF SOME MEAN VALUES FOR A BIVARIATE DISTRIBUTION

By Edward Paulson[1]

Columbia University

In this paper two problems are discussed which were suggested by the theory of representative sampling [1], but which also occur in several other fields. The first problem is to set up confidence limits for $\frac{m_x}{m_y}$, the ratio of the mean values of the variates x and y. This comes up in the following situation. Let a population π consist of N units $x_1, x_2, \cdots x_N$ and suppose we wish to set up confidence limits for the mean $X = \frac{\sum_{i=1}^{N} x_i}{N}$. Also assume the population π has been divided into M groups, let v_j be the number of individuals in the j^{th} group and u_j be the sum of the values of x for the v_j individuals in the j^{th} group, so $X = \frac{u_1 + u_2 \cdots u_M}{v_1 + v_2 \cdots v_M} = \frac{Mm_u}{Mm_v}$. Now if a random sample of n out of the M groups is taken, yielding observations $(u_1, v_1), (u_2, v_2) \cdots (u_n, v_n)$ and N is unknown, the determination of confidence limits for X clearly becomes a special case of the first problem. The distribution of a ratio, discussed by Geary [2], does not seem to be well adapted for this purpose.

The second problem, which is of greater practical interest, arises when we again have a random sample $(u_1, v_1) \cdots (u_n, v_n)$ of n out of M groups and N and M are known. The standard estimate of X that has usually been made is $\hat{X} = \frac{M\bar{u}}{N}$, where $\bar{u} = \frac{\sum_{j=1}^{n} u_j}{n}$. This estimate does not utilize the fact that the n observations on v can be used to increase the precision of the estimate of the numerator of X. This is a special case of problem 2, which we can now formulate as how to best estimate m_x (the mean value of a trait x) both by a point and by an interval, when for each unit in the sample observations both on x and on a correlated variate y are obtainable, and m_y is known a priori. Situations of this type occur fairly often. It is possible to reduce the second problem to the first by using $\frac{\bar{x}}{\bar{y}} \cdot m_y$ as the estimate of m_x, and by multiplying the confidence limits for $\frac{m_x}{m_y}$ by m_y to secure limits for m_x, but this will not usually be the most efficient procedure.

In both problems two cases will be distinguished: (a) when σ_x^2, σ_y^2 and ρ are known a priori, and (b) when they are unknown. To determine confidence

[1] Work done under a grant-in-aid from the Carnegie Corporation of New York.

limits for $\frac{m_x}{m_y}$, it will first be assumed that the probability density $f(xy)$ of x and y is

$$f(x,y) = \frac{\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-m_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-m_x}{\sigma_x}\right)\left(\frac{y-m_y}{\sigma_y}\right)+\left(\frac{y-m_y}{\sigma_y}\right)^2\right\}\right]}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}. \tag{1.1}$$

Denote the ratio $\frac{m_x}{m_y}$ by K (assuming $m_y \neq 0$), and suppose it is desired to test the hypothesis that $K = K_0$ on the basis of a sample of n independent observations $(x_1, y_1) \cdots (x_n, y_n)$.

Let $z_i = x_i - Ky_i$ and $\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n}$. Since z is a linear function of x and y it must be normally distributed, and its mean value is obviously zero. Therefore

$$u = \frac{\sqrt{n}\,\bar{z}}{\sigma_z} = \frac{\sqrt{n}\,(\bar{x} - K\bar{y})}{\sqrt{\sigma_x^2 - 2K\rho\sigma_x\sigma_y + K^2\sigma_y^2}} \tag{1.2}$$

will be normally distributed about zero with unit variance, and the hypothesis is rejected if $|u(K_0)| > u_\alpha$, where $\frac{1}{\sqrt{2\pi}}\int_{u_\alpha}^{\infty} e^{-\frac{1}{2}t^2}\,dt = \frac{1}{2}\alpha$. It is easy to show that this test is equivalent to that based on the likelihood-ratio.

Confidence limits for K would now be given by values of K satisfying the inequality $\left|\frac{\sqrt{n}\,\bar{z}}{\sigma_z}\right| \leq u_\alpha$, provided they always constituted a closed non-empty interval. This is equivalent here to the requirement that K be a real valued monotonic function of u in the interval $-\infty < u < \infty$; this requirement is unfortunately never exactly fulfilled, as can be seen from the graph of (1.2) (in the u, K plane), for the curve has two horizontal asymptotes $u = \pm\frac{\sqrt{n}\,\bar{y}}{\sigma_y}$, and one maximum or minimum point $\left(\text{unless } \frac{\bar{x}}{\bar{y}} = \rho\frac{\sigma_x}{\sigma_y}\right)$. However, K will always be a monotonic function of u in the interval $-u_\alpha < u < u_\alpha$ provided $\left|\frac{\sqrt{n}\,\bar{y}}{\sigma_y}\right| > u_\alpha$. Since $m_y \neq 0$, by taking n sufficiently large the probability that $\left|\frac{\sqrt{n}\,\bar{y}}{\sigma_y}\right| < u_\alpha$ can be made arbitrarily small. Moreover, for values of α ordinarily used, in most practical problems the value of $\frac{m_y}{\sigma_y}$ will be such that even for quite small samples the probability $\left|\frac{\sqrt{n}\,\bar{y}}{\sigma_y}\right| < u_\alpha$ (that is, the proba-

bility of getting a sample for which the values of K that are accepted will not form a real interval) will be quite negligible. For example, let α have the conventional value .05, and suppose $\frac{m_y}{\sigma_y} = 2$; then for $n = 9$, Prob. $\left\{\left|\frac{\sqrt{n}\,\bar{y}}{\sigma_y}\right| < 1.96\right\} < 10^{-4}$ and for $n = 16$, Prob. $\left\{\left|\frac{\sqrt{n}\,\bar{y}}{\sigma_y}\right| < 1.96\right\} < 10^{-9}$. Subject to these rather weak restrictions on the order of magnitude of n and $\frac{m_y}{\sigma_y}$, the confidence limits for K are

$$(1.3)\qquad \frac{(n\bar{x}\bar{y} - u_\alpha^2\rho\sigma_x\sigma_y) \pm \sqrt{(n\bar{x}\bar{y} - u_\alpha^2\rho\sigma_x\sigma_y)^2 - (n\bar{y}^2 - u_\alpha^2\sigma_y^2)(n\bar{x}^2 - u_\alpha^2\sigma_x^2)}}{n\bar{y}^2 - u_\alpha^2\sigma_y^2}.$$

In case (b) when σ_x^2, σ_y^2, and ρ are unknown, each $z_i = x_i - Ky_i$ is still normally and independently distributed with zero mean and a common variance. It follows that

$$(1.4)\qquad t = \frac{\sqrt{n}\,\bar{z}}{\sqrt{\frac{\Sigma(z-\bar{z})^2}{n-1}}} = \frac{\sqrt{n}\,(\bar{x} - K\bar{y})}{\sqrt{s_x^2 - 2rs_x s_y K + s_y^2 K^2}}$$

will have Students' distribution with $n - 1$ degrees of freedom. Subject to practically the same restriction as before, the confidence limits for K as determined from (1.4) are

$$(1.5)\qquad \frac{(n\bar{x}\bar{y} - t_\alpha^2 rs_x s_y) \pm \sqrt{(n\bar{x}\bar{y} - t_\alpha^2 rs_x s_y)^2 - (n\bar{y}^2 - t_\alpha^2 s_y^2)(n\bar{x}^2 - t_\alpha^2 s_x^2)}}{n\bar{y}^2 - t_\alpha^2 s_y^2}$$

where t_α is the critical value of Students' distribution (for $n - 1$ degrees of freedom) and $s_x^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$, $s_y^2 = \frac{\Sigma(y_i - \bar{y})^2}{n-1}$, and r is the sample correlation between x and y.

When the distribution of x and y deviates considerably from a bivariate normal one, it would still appear that as a practical matter much the same methods could be used. The basis for this is the fact that there is considerable experimental evidence [3], [4] to show that the distribution of the mean of a sample drawn from any population likely to be encountered in practise will approach normality very rapidly even for n quite small. Hence $\bar{z}$ and u can be regarded as normally distributed for n say >25, and the confidence limits for $\frac{m_x}{m_y}$ will then be given by (1.3); in case (b) a somewhat larger sample is required to diminish the error in estimating σ_z. But for n say >50, t will have a distribution close to normal and the confidence limits for K are given by (1.5) (with t_α replaced by u_α). The statements for the non-normal case appear as a practical matter to also hold when the sample is drawn from a finite population of N

units without replacement if $N - n$ is not too small, provided n is replaced by $n\left(\frac{N-1}{N-n}\right)$, for now $\sigma^2_{(\bar{x}-K\bar{y})} = \frac{1}{n}\left(\frac{N-n}{N-1}\right)[\sigma_x^2 - 2\rho\sigma_x\sigma_y K + \sigma_y^2 K^2]$.

In the second problem we again start by assuming the distribution of x and y is given by (1.1). For case (a), m_x is the only unknown parameter. If $P = \prod_{i=1}^{n} f(x_i y_i \mid m_x)$ and $\phi = \frac{\partial \log P}{\partial m_x}$, then

$$\phi = \frac{1}{2(1-\rho^2)}\left\{\frac{2\Sigma(x_i - m_x)}{\sigma_x^2} - \frac{2\rho}{\sigma_x\sigma_y}\Sigma(y_i - m_y)\right\},$$

and the maximum likelihood estimate $\hat{m}_1$ of m_x is

$$\hat{m}_1 = \bar{x} - \frac{\sigma_{xy}}{\sigma_y^2}(\bar{y} - m_y), \tag{1.6}$$

where $\sigma_{xy} = \rho\sigma_x\sigma_y$. Also $\hat{m}_1$ is a sufficient statistic, and the confidence interval given by the set of values of m_x satisfying

$$\left|\frac{\sqrt{n}\left[\bar{x} - \frac{\sigma_{xy}}{\sigma_y^2}(\bar{y} - m_y) - m_x\right]}{\sigma_x\sqrt{1-\rho^2}}\right| \leq u_\alpha$$

will be a "shortest unbiased confidence interval" in the sense of Neyman.

Case (b) will be more important, since the exact values of the variances and covariance will usually be unknown. By analogy with (1.6), a similar estimate of m_x for this case is

$$\hat{m}_2 = \bar{x} - \frac{s_{xy}}{s_y^2}(\bar{y} - m_y). \tag{1.7}$$

This is precisely the least square estimate of x_i corresponding to $y_i = m_y$, and has been used for this problem before; for example, it is discussed by Cochran [5]. We shall discuss some additional aspects of the problem, and also mention the application to the special case of representative sampling by groups.

When the bivariate distribution of x and y is such that the conditional distribution of each x_i is normal with mean $A + By_i$ and a common variance, then Professor Wald has suggested that exact confidence limits for m_x for small samples can be secured by using the standard methods of the theory of least squares. The resulting confidence limits are easily seen to be

$$\hat{m}_2 \pm \frac{t_\alpha}{\sqrt{n-2}}\sqrt{\lambda},$$

where

$$\lambda = (1 - r^2)\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\frac{1}{n} + \frac{(m_y - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right],$$

and t_α is the critical value of Students' distribution with $n - 2$ degrees of freedom at a level of significance $= \alpha$.

The requirement that the regression of x on y be linear is rather stringent, although it may often be fulfilled, especially in the case of representative sampling mentioned in the opening paragraph. When the regression of x on y is non-linear, the estimate given by (1.7) requires some further justification. Let $U_{ij} = E(x^i y^j)$, where E denotes the mean value, and assume that we have n independent pairs of observations and that the moments U_{10}, U_{01}, U_{11}, U_{20}, U_{02}, U_{40}, U_{04} and U_{22} are all finite. It then follows from a theorem of Doob [6] that $\sqrt{n}(\hat{m}_2 - m_x)$ tends to a limiting distribution with increasing n which is normal with zero mean and variance equal to $\sigma_x^2(1 - \rho^2)$.

The estimate $\bar{x}$ is clearly always less efficient than $\hat{m}_2$ unless $\rho = 0$. The estimate $\frac{\bar{x}}{\bar{y}} \cdot m_y$ is known to have a large sample variance

$$V = \frac{1}{n}\left[\sigma_x^2 - 2\left(\frac{m_x}{m_y}\right)\sigma_{xy} + \left(\frac{m_x}{m_y}\right)^2 \sigma_y^2\right].$$

So $\frac{\bar{x}}{\bar{y}} \cdot m_y$ is always less efficient than $\hat{m}_2$ unless $m_x = \frac{\sigma_{xy}}{\sigma_y^2} m_y$, at which point V attains its minimum value $\frac{\sigma_x^2(1 - \rho^2)}{n}$. In fact $\hat{m}_2$ can be easily shown to have an efficiency $\geq$ any other statistic of the class Q, $\left(\text{which includes } \bar{x} \text{ and } \frac{\bar{x}}{\bar{y}} m_y\right)$ consisting of all statistics q satisfying two conditions: (1) that $\sqrt{n}(q - m_x)$ have a distribution approaching normality with zero mean and finite variance σ_q^2 and (2) σ_q^2 be independent of the joint density function of x and y, involving only certain of the moments u_{ij}. A rather artificial member of the class Q is $q = \frac{\bar{x} \log \bar{y}}{\log m_y} - \frac{s_x^2}{s_y^2}(\sqrt[3]{\bar{y}} - \sqrt[3]{m_y})$. The proof consists merely in observing that if for any bivariate distribution $\sigma_{\hat{m}_2}^2 = \sigma_x^2(1 - \rho^2) > \sigma_q^2$, this would also have to be true when the distribution of x and y is a bivariate normal one, which is impossible, since $\sigma_x^2(1 - \rho^2)$ is then the variance of $\sqrt{n}(\hat{m}_1 - m_x)$, $\hat{m}_1$ being the maximum likelihood statistic.

For moderate values of n, say $n > 100$, fairly exact confidence limits for m_x will be given by $\hat{m}_2 \pm \frac{u_\alpha}{\sqrt{n}} \sqrt{s^2(1 - r^2)}$. When the sample is drawn from a finite population of N units without replacement, the confidence limits for $n > 100$ are $\hat{m}_2 \pm \frac{u_\alpha}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}} \sqrt{s_x^2(1 - r^2)}$.

In the problem of estimating $m_x = X$ for the population II, discussed in the opening paragraph, which consists of N individuals divided into M groups, on the basis of a random sample $(u_1, v_1), (u_2, v_2) \cdots (u_n, v_n)$ of n out of the N

groups, an efficient estimate will be $m' = \dfrac{M\left[\bar{u} - \dfrac{s_{uv}}{s_v^2}\left(\bar{v} - \dfrac{N}{M}\right)\right]}{N}$. The efficiency of m' relative to the conventional estimate $\dfrac{M\bar{u}}{N}$ is $(1 - \rho_{uv}^2)^{-1}$, which ordinarily would seem to be quite large. This is easily extended to the case II is divided into l strata with M_i groups comprising N_i individuals in the i^{th} stratum, when a random sample of m_i out of the M_i groups in each stratum is taken. Let v_{ij} be the number of individuals in the j^{th} group of the i^{th} stratum and u_{ij} denote the sum of the values of x for these v_{ij} individuals. The estimate of m_x becomes

$$m'' = \frac{\sum_{i=1}^{l} M_i\left[\bar{u}_i - \dfrac{s_{u_iv_i}}{s_{v_i}^2}\left(\bar{v}_i - \dfrac{N_i}{M_i}\right)\right]}{N}$$

If $\sum_{i=1}^{l} m_i = m$ is fixed, the large sample variance of m'' will be a minimum if m_i is proportional to $M_i\sigma_{u_i}\sqrt{1 - \rho_i^2}$, where ρ_i is the correlation between u and v in the i^{th} stratum.

In conclusion, the writer wishes to thank Professor A. Wald for his advice and encouragement, and Mr. Henry Goldberg for several suggestions.

REFERENCES

[1] J. Neyman, "On the two different aspects of the representative method," Journal of the Royal Statistical Society, Vol. 97 (1934), pp. 558–606.

[2] R. C. Geary, "The frequency distribution of the quotient of two normal variates," *Roy. Stat. Soc. Jour.*, Vol. 93 (1930), pp. 442–446.

[3] W. A. Shewhart, *Economic Control of Quality of Manufactured Product*, New York, (1931), pp. 182–183.

[4] H. C. Carver, "Fundamentals in the theory of sampling," *Annals of Math. Stat.*, Vol. 1 (1930), pp. 110–112.

[5] W. G. Cochran, "The use of the analysis of variance in enumeration by sampling," *Jour. Amer. Stat. Assoc.*, Vol. 34 (1939), pp. 492–510.

[6] J. L. Doob, "The limiting distribution of certain statistics," *Annals of Math. Stat.*, Vol. 6 (1935), p. 166.

SIGNIFICANCE LEVELS FOR THE RATIO OF THE MEAN SQUARE SUCCESSIVE DIFFERENCE TO THE VARIANCE

By B. I. Hart

Ballistic Research Laboratory, Aberdeen Proving Ground

For purposes of practical application in connection with significance tests a tabulation of the argument corresponding to certain percentage points of the probability integral is usually more convenient than that of the probability integral for equal intervals of the argument. A table of probabilities for the