ON A LARGE SAMPLE METHOD OF ESTIMATING UNEMPLOYMENT IN LARGE CITIES

J. C. KOOP

(Directorate of Labour, Rangoon, Burma)

SUPPOSE we have a large city which possesses a certain number of employment exchanges serving different parts of it. If each exchange registers all unemployed persons, then the total number unemployed in the city would be completely known. However, as they do not always seek the assistance of exchanges, the numbers registered by exchanges do not represent all the unemployed. Now, how may this information be used, as supplementary information, to estimate the total number unemployed? Since the problem is also closely linked with the estimation of the proportion of persons unemployed, and, of that number, the proportion using the exchanges, they may as well be discussed together.

Assume that a plan of the city showing natural boundaries (streets, lanes, walls, etc) is available It may be divided up along natural boundaries into very small areas (grids, in Mahalanobis' terminology) for the following reasons:

(1) Small size grids are more convenient to handle than large ones

(2) The sampling theory will require the number of grids to be as large as possible. Hence the sizes of grids should be as small as possible.

Suppose the total area of the city has been divided up into N grids without paying careful attention to the numbers included in each grid. At the moment of time, to which the survey refers, let z_i be the total number of persons inhabiting the ith grid, y_i the number unemployed in the same grid and of this number let x_i be the number who have registered themselves as unemployed at an exchange; ($i = 1, 2, \ldots N$) Experience shows that the number unemployed is generally more in poor than in relatively well-to-do localities, and also persons in indigent circumstances depend more on the services of exchanges than persons in more favourable circumstances. Hence we may assume that x, y and z are mutually independent Let σ_x^2, σ_y^2 and σ_z^2 be the variances of x, y and z respectively. $X = \sum_{i=1}^{N} x_i$, which is known exactly, is the total number of unemployed persons on the books of exchanges. $Y = \sum_{i=1}^{N} y_i$ will be the total number of unemployed persons in the city. $Z = \sum_{i=1}^{N} z_i$ will be the population of the city, and generally, this figure cannot be expected to be known with precision except if an up-to-date and accurate census figure is available. Assuming that no person enters or leaves the city, at the period of time when the survey is conducted, X, Y and Z wil be fixed numbers. (Obviously $X < Y < Z$). We are attempting to estimate Y and the ratios $\frac{X}{Y}$ and $\frac{Y}{Z}$. It will be shown that the estimating equations of these statistics involve ratios and on this account it will not be possible to obtain accurate and exact expressions for their standard errors. However, it is possible to obtain their confidence limits by an application of a theorem, first stated by Fieller (1940), to the special case for large samples in the way described below.

Let n grids be chosen at random, each grid being assigned an equal probability of selection. Let the quantities x, y and z, as defined above, be observed in each grid. Then

$$\bar{x} = \sum_{i=1}^{n} x_i/n,\ \bar{y} = \sum_{i=1}^{n} y_i/n \text{ and } \bar{z} = \sum_{i=1}^{n} z_i/n$$

will be unbiassed estimates of the average number of registrations for employment, of the number unemployed, and of the number inhabiting each grid respectively. Consider the function

$$d = \bar{x} - \frac{X}{Y}\bar{y}. \qquad (1)$$

Its expectation can be shown to be zero, i.e., $E(d) = 0$, so that an estimate Y′ of Y or $\left(\frac{X}{Y}\right)'$ of $\frac{X}{Y}$ can be obtained by putting $d=0$ in (1) and this yields two estimating equations

$$Y' = \frac{\bar{y}}{\bar{x}} X \text{ and } \left(\frac{X}{Y}\right)' = \frac{\bar{x}}{\bar{y}} \qquad (2),$$

which are ratio estimates, which one could have obtained intuitively. The variance of d, $V(d)$, can be shown to be

$$V(d) = \left(\frac{\sigma_x^2}{n} + \frac{X^2}{Y^2}\frac{\sigma_y^2}{n}\right) \cdot \frac{N-n}{N-1}$$

assuming that x and y are independent. When N is infinite, or for all practical purposes when it is effectively large, then

$$V(d) = \frac{1}{n}\left(\sigma_x^2 + \frac{X^2}{Y^2}\sigma_y^2\right). \quad (3)$$

Further in large-scale sampling (*i.e.*, when $n>150$), $\bar{x}$ and $\bar{y}$ will be almost normally distributed, so that for all practical purposes d will be normally distributed with zero mean and standard deviation $\sqrt{V(d)}$. Hence, under the conditions stated above,

$$u=\left(\bar{x}-\frac{X}{Y}\bar{y}\right)\Big/\left\{\frac{1}{n}\left(\sigma_x^2+\frac{X^2}{Y^2}\sigma_y^2\right)\right\}^{\frac{1}{2}} \quad (4)$$

will be distributed normally with zero mean and unit variance. It follows that if we wish to obtain confidence limits of Y, or $\frac{X}{Y}$ corresponding to the confidence level $\frac{p}{100}$, we need only to substitute the appropriate normal deviate for u corresponding to the $p\%$ significance level, and solve the above quadratic equation for Y, or $\frac{X}{Y}$ as the case may be. When rearranged as a quadratic in Y, (4) appears as

$$Y^2\left(\bar{x}^2-\frac{u^2\sigma_x^2}{n}\right)-2\bar{x}\bar{y}XY+X^2\left(\bar{y}^2-\frac{u^2\sigma_y^2}{n}\right)=0 \quad (5),$$

the roots of which are

$$X\left(n\bar{x}\bar{y}\pm u\sqrt{n\left(\bar{x}^2\sigma_y^2+\bar{y}^2\sigma_x^2\right)-u^2\sigma_x^2\sigma_y^2}\right)\Big/\left(n\bar{x}^2-u^2\sigma_x^2\right) \quad (6)$$

the root with the positive sign before u giving the upper confidence limit and that with the negative sign before u giving the lower confidence limit of Y. Similarly the upper and lower confidence limits of $\frac{X}{Y}$ may be obtained by the solution of equation (5), treating $\frac{X}{Y}$ as the unknown.

The same arguments apply for the estimation of $\frac{Y}{Z}$. Its estimate would be $\frac{\bar{y}}{\bar{z}}$, and, its confidence limits could be obtained in a similar manner as for $\frac{X}{Y}$. If Z is accurately known, a second estimate of Y, which would be $\frac{\bar{y}}{\bar{z}}Z$ could serve as an independent check on the first. However, too much reliance cannot be placed on this estimate, except if Z is a recent census figure. For, if Z is in error by ΔZ, then the estimate of Y will also be in error by $\frac{\bar{y}}{\bar{z}}\Delta z$. It is precisely for this reason that the estimation of Y is recommended by the method first described rather than by this method.

The confidence limits discussed above presuppose that σ_x^2, σ_y^2 and σ_z^2 are known. In practice, they are not at all likely to be known. However, in large samples, their estimates may be substituted without serious danger of error or loss in accuracy. Also, it may be possible to control σ_z^2, but it will not be possible to control σ_x^2 and σ_y^2, since for any system of fixed grids, the number of unemployed cannot be expected to remain constant in each grid and therefore x and y will be varying throughout time. If every grid contained the same number of persons then $\sigma_z^2=0$; however, this cannot be realised in practice. But it may be possible to demarcate the grids so that they hold approximately the same number of persons, if some prior knowledge about the distribution of houses or population is available. Under such circumstances, σ_z^2 would assume a lower value than if grids had been demarcated otherwise

The above suggestion has a bearing on the question of obtaining the shortest possible confidence intervals, under the conditions stated, given of course the confidence coefficient. An inspection of (6) will show that the length of the confidence interval of Y is

$$2uX\sqrt{n\left(\bar{x}^2\sigma_y^2+\bar{y}^2\sigma_x^2\right)-u^2\sigma_x^2\sigma_y^2}\Big/\left(n\bar{x}^2-u^2\sigma_x^2\right)$$

and the only way to narrow it down (and for that matter, also that of $\frac{X}{Y}$) would be by increasing the sample size. For the confidence interval of $\frac{Y}{Z}$, the width of the interval is

$$2u\sqrt{n\left(\bar{z}^2\sigma_y^2+\bar{y}^2\sigma_z^2\right)-u^2\sigma_y^2\sigma_z^2}\Big/\left(n\bar{z}^2-u^2\sigma_z^2\right)$$

and, it is clear from this formula, that the narrowness of width, besides depending on n, depends also on how low the value of σ_z^2 is, and therefore, it may be narrowed down further by demarcating the grids in the way already suggested.

To sum up, ratio estimates have been suggested for three unemployment statistics, namely, (*a*) the total number unemployed, (b) the proportion using employment exchanges, (c) the proportion in the city unemployed, the first named depending on supplementary information for its estimation. Their confidence limits have also been derived on the basis of large sample theory, and on the assumption that the number of grids is very large. For the statistic representing the proportion unemployed, a method of reducing the width of its confidence interval is proposed.

---

1. Fieller, E. C., *Jour. Roy. Stat. Soc., Supplement*, 1940 7, 1–64.