Erroneous Formulae for International Comparison of Variability of Income

by

R. C. Geary

A suggestive and painstaking paper by F. G. Williamson*, dealing incidentally with international comparisons of income variability, has attracted some attention in ESRI. Two formulae for assessing variability are considered -

$$(1) \quad V_w = \sqrt{\left[\sum_{i=1}^{N} (\bar{y}_i - \bar{y})^2 f_i\right]} \,/\, \sqrt{n}.\bar{y} \,,$$

$$(2) \quad V_{uw} = \sqrt{\left[\sum_{i=1}^{N} (\bar{y}_i - \bar{y})^2\right]} \,/\, \sqrt{N}.\bar{y} \,,$$

where f_i = population of the ith region, n = national population, $\bar{y}_i$ = income per capita of the ith region, $\bar{y}$ = national income per capita, N = number of regions. We use the author's notation, with one small change. Though preferring (1), the author uses both formulae for comparing variability in 24 countries, including Ireland, for which he apparently used Geary's county income data for 1960. Hence in this case N = 26. Number of regions N varied greatly, ranging from 6 for Australia to 76 for Puerto Rico.

---

* "Regional Development in Particular Countries" by J. G. Williamson in Regional Analysis (ed. L. Needleman), Penguin Modern Economics, 1968.

In this note we show that the author's formulae are invalid. We also show that it is not possible to institute valid comparisons of income variability from the kind of data available to the author.

First we make the point that valid comparisons of variability can be made only on the basis of individual incomes of persons or households. Such a concept should be invariant to regional grouping. The inadequacy of the concepts enshrined in formulae (1) and (2) will be evident from regarding the country as one region, i.e. $N = 1$. Then V_w and V_{uw} are zero. The income variability in all countries is zero, which, as Euclid said, is absurd.

While income per head of population is a proper concept on a welfare basis or comparisons between regions and between countries, it is unsuitable for comparisons of income variability on an individual basis. Here, we suggest, the individual should be the income earner unless, of course, we used household incomes when the unit would be the household. To fix ideas, we take the former. The earlier notation stands, except that n and the f_i are now numbers of income earners.

$$(3) \qquad (n-1)s^2 = \sum_{i=1}^{N} \sum_{j=1}^{f_i} (y_{ij} - \bar{y})^2$$

and where y_{ij} is the income of individual j in region i. Now, as is well known, the right side of (3) may be written in two terms, to give

$$(4) \qquad \Sigma_i \, \Sigma_j (y_{ij} - \bar{y})^2 = \Sigma_i \, \Sigma_j (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{N} f_i (\bar{y}_i - \bar{y})^2$$

Effectively, Williamson takes into account only the second term on the right. His formulae ordinarily underestimate the right coefficient of variation $c = s/\bar{y}$, and only a fantastic set of coincidental similarities would render his formulae suitable for international comparison or any concept of variability whatsoever.

From elementary analysis of variance considerations it is known that an unbiased estimate of the population variance can be produced from the second term on the right of (4). But this can happen only in the null-hypothesis case which here assumes that the measures y_{ij} are those of a random sample of n from an infinite population with variance σ^2. One would then have to envisage the sample of n divided at random into N groups with f_i in the ith group.

Then in (4), the last term, say T, has (N-1) degrees of freedom so that T/(N-1) is an estimate s^2 of the variance. But, according to this concept, e.g. Dublin and Leitrim would have the same expected (i.e. operator E) average income which, to repeat, is absurd.

For international comparisons knowledge of the frequency distribution of individual incomes is necessary. Knowledge of numbers and regional averages is insufficient.

R. C. Geary

3 May 1973