Statistics Division

Working Paper Series

NO: ESS / ESSG/011e

Reliability of estimates of inequality in:

Distributions derived from sample survey data

April 2007

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS

منظمة الأغذية والزراعة للأمم المتحدة | 联合国粮食及农业组织 | Food and Agriculture Organization of the United Nations | Organisation des Nations Unies pour l'alimentation et l'agriculture | Organización de las Naciones Unidas para la Agricultura y la Alimentación

On reliability of estimates of inequality in distributions derived from sample survey data

Arun Kumar Srivastava, Anil Rai and V. Ramasubramanian
Indian Agricultural Statistics Research Institute
New Delhi, India

Introduction

For studying inequalities in populations, the interest often lies in estimating frequency distributions. Several factors of interest, such as income or expenditure, depict highly skewed distributions. The study of economic inequality with respect to such distributions is of common interest, and estimation of frequency distributions is more meaningful in such cases. The sample surveys planned for estimating population parameters such as the mean or total may not adequately capture the distributional properties of the populations needed for estimation of frequency distributions. Among several measures of inequalities, the Lorenz ratio or Gini concentration coefficient (G) is considered as one of the most important measures. Estimation of G involves estimation of frequency distributions as well. It may not be reasonable to estimate these parameters from data collected through sample surveys planned for the estimation of the mean or total. The reliability of such sampling designs as well as estimates of G needs to be investigated. In this paper, the problem of choice of sampling designs for estimation of G as well as frequency distributions is attempted. Since studying the distributional properties of estimators of such complex parameters is not simple, the problem is approached empirically on simulated data.

Simulation of population

For empirical studies, populations with characteristics such as income, expenditure, holding size, etc. having highly skewed distributions have been kept in view. In particular, we consider a population of land holdings in Tamil Nadu State from Agricultural sciences. The population is highly skewed in nature. The parameters of this distribution are taken into account for generating simulated populations. Consider a population of n villages. Let there be M_i households in the ith village. The village size M_i is generated randomly following normal distribution with the mean and variance taken from a real population (for example, mean = 100, variance = 900, i.e. CV= 30 percent). Within each village, M_i values are generated with Gamma (l, m); l and m are determined such that the mean = lm and the variance = lm^2. The mean and variance of the Gamma distribution are based on real populations (Murthy, 1977). In fact, the mean = 1.25525, and the variance = 2.47430; thus, l = 0.636811 and m = 1.971149. Although populations of various sizes were simulated, for the present illustrations we present the case of n = 100. We also present here only the case of simulation using Gamma distribution. We illustrate the problems in estimation of inequality measures through the simulations of populations based on landholdings, but the approach and the results are likely to hold when looking at income distribution as well, as both are similar in terms of skewness and other distributional properties. Moreover, in rural areas, landholdings are invariably directly related to the household income. The populations generated through this approach have been used for further sampling and investigations relating to calculation of G as well as for estimation of frequency distributions.

Gini Coefficient as a measure of economic inequality

One of the most widely used measures for the extent of inequality is the Gini coefficient. An important feature of this measure is its association with the Lorenz curve in which the proportion of the population arranged from the poorest to the richest are represented on the horizontal *x*-axis, and the proportion of income held by the bottom *x* proportion of the population is depicted on the vertical *y*-axis. The mathematical formulation is as follows: Let the income y (³ 0) have continuous type distributions with density function *f(y)* with mean:

$$\mu = \int_0^{\infty} yf(y)dy$$

.

Define

$$F(x) = \int_0^{x} f(y)dy$$

and

$$F_1(x) = \int_0^{x} yf(y)dy / \mu$$

F(x) is the proportion of persons with in- come, £ *x* and $F_1(x)$ is the proportionate share of these persons in the aggregate income of all persons. *F(x)* and $F_1(x)$ both lie between 0 and 1 for x ranging from 0 to ¥, and F_1 is a monotone increasing function of *F*. The graph of F_1 against *F* is called the Lorenz curve or the concentration curve of the given distribution of income.

In general, the Lorenz curve must satisfy the following conditions:

(1) If $F = 0$, $F_1 = 0$.
(2) $F = 1$, $F_1 = 1$.
(3) $F_1 < F$.
(4) The slope of the curve increases monotonically.

The area between the Lorenz curve and the egalitarian line is called the area of concentration. The Lorenz ratio, also known as the Gini Coefficient is defined as:

G = 2 × area of concentration

= $2 \times \left(\frac{1}{2} - \text{area below the Lorenz curve} \right)$

= $1 - 2\int_0^1 F_1 dF$.

The Gini Coefficient (G) may also be represented in several alternative ways. Some of the representations and corresponding interpretations in terms of welfare economics are available in the literature (see e.g. Sen, 1973).

Estimation of Lorenz Ratio or Gini Coeffi cient (G)

For estimation of the Lorenz ratio through a numerical approach, the following procedure is followed.

Let there be K class intervals, p_k the percentage of persons in the kth class (k = 1, 2 ..., K), $\bar{x}_k$ the average of character x in the kth class, q_k the percentage share of the kth group in the aggregate expenditure, P_k the cumulative p_k (with P_K = 100) and Q_k the cumulative q_k (with Q_0 = 0 and Q_K = 100)

$$\text{Lorenz ratio (G)} = 1 - \frac{\sum_{k=1}^{k} p_k (Q_k + Q_{k-1})}{10000}$$

.

For details, please refer Bhattacharya and Coondoo (1992).

Estimation of G, in this method, is essentially based on estimates of p_k and $\bar{x}_k$ (k = 1, ... K). The role of the sampling design also appears in the estimation of these parameters.

One of the limitations of estimating p_k and $\bar{x}_k$ in skewed populations could be due to extremes in the distributions, thereby leading towards the proper sampling designs for estimation of frequency distributions. Estimation of the Lorenz curve as well as the fractile graphical analysis (Mahalanobis, 1960) is an attempt in this direction.

In the present investigation, the performance of estimators of G has been examined empirically. The sampling designs considered are as follows:

1. direct sampling of ultimate units:

1.1 simple random sampling without replacement (srswor);

2. sampling of clusters of ultimate units (villages):

2.1 sampling clusters (villages) with srswor;

2.2 sampling clusters (villages with probability proportional to size with replacement (ppswr), size being the number of ultimate units;

3. two-stage sampling:

3.1 both the stages by srswor;

3.2 first stage ppswr and srswor in second stage;

4. stratified sampling:

4.1 selection of clusters by ppswr.

The empirical investigation indicates that the estimates of G have comparatively higher biases in the case of sampling for clusters of villages and also in the case of two-stage sampling as compared with selection of ultimate units by simple random sampling. The selection of villages by probability proportional to size has resulted in larger coefficients of variation for the estimators. The larger biases are perhaps due to the skewed nature of the parent population.

The performance of various sampling designs was also examined for estimating the frequency distribution on the basis of a criterion considered by Murthy (1977). Since estimation of G through grouped data requires estimation of group proportions, it was expected that sampling designs that perform better in estimating frequency distributions should perform well for estimation of G as well. The results were in broad conformity with this expectation.

References

Bhattacharya, N. & Coondoo, D. 1992. *Collection and analysis of survey data on income and expenditure, training handbook*. Tokyo, SIAP.

Mahalanobis, P.C. 1960. A Method of fractile graphical analysis. *Econometrica*, 28: 325-351.

Murthy, M.N. 1977. *Use of empirical studies in evaluating sample designs for estimating frequency distributions.* Proceedings of the meeting of the International Statistical Institute held at New Delhi, India.

Sen, A. 1973. *On economic inequality*. Oxford, UK, Clarendon Press.