Jour. Ind. Soc. Ag. Statistics
Vol. XL, No. 2 (1988), pp. 154-163

APPROXIMATION OF FIRST TWO MOMENTS AND SAMPLING DISTRIBUTION OF GINI'S COEFFICIENT

MURARI SINGH and RAJENDRA P. SINGH
International Crops Research Institute for the Semi-Arid Tropics,
Patancheru, A.P. 502 324

(Received : December, 1987)

SUMMARY

A procedure for obtaining the approximation of the first two moments and the probability distribution of the sample Gini's coefficient has been discussed using some results on order statistics.

Keywords : Order statistics; Asymptotic distribution; Log normal; Exponential distribution; Pareto distribution.

Introduction

The measurement of inequality of income is often presented in terms of the Gini's coefficients or Lorenz's concentration ratio (Kendall and Stuart [4]; Kakwani [5]) estimated from the random sample from the concerned population. The sampling distribution and the moments of the sample Gini's coefficient will be useful for the comparison of income inequalities across several populations. The moments of Gini's coefficient have been considered by Iyengar [6] for log normal population. This paper attempts to provide results for any population in general.

Let $f(x)$ and $F(x)$ denote respectively the probability density function and distribution function of a random variable X (for example, income) on a population Π. Gini's coefficient G for the population (distribution) Π is defined as the ratio of absolute mean difference Δ and the mean μ

of the population. (Kendall and Stuart [4], p. 48.)

$$G = \Delta/(2\mu) \tag{1}$$

$$\Delta = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} | x - y | f(x) f(y) \, dx \, dy$$

$$\mu = \int_{-\infty}^{\infty} x f(x) \, dx$$

For estimating G consider a random sample $x_1, x_2, \ldots, x_n$ of size n from the population Π. The estimate g of G is given by the ratio of unbiased estimates δ of Δ and $\bar{x}$ of μ.

$$g = \delta/(2\bar{x}) \tag{2}$$

where

$$\delta = (1/n.(n - 1)) \sum_i \sum_j | x_i - x_j |$$

$$\bar{x} = (1/n) \sum_i x_i$$

It has been established that

$$E(\delta) = \Delta$$

and $\text{Var}(\delta) = (1/n(n - 1)) (4\sigma^2 + 4(n - 2) J - 2(2n - 3) \Delta^2)$

where $E(\cdot)$ and Var $(\cdot)$ stand for expectation and variance respectively with respect to the population Π, and

$$J = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} | x - y | \; | x - z | f(x) f(y) f(z) \, dx \, dy \, dz$$

$$= 4 \int_{-\infty}^{\infty} \{(xF - \psi_1)^2 + (\mu - x)(xF - \psi_1)\} \, dF(x) + \sigma^2$$

$$\psi_1(x) = \int_{-\infty}^{x} u f(u) \, du$$

σ^2 is the variance of x in the population Π.

It is evident that the evaluation of the variance of δ involves an integral which could be complicated and we will still require the information on covariance of δ and $\bar{x}$ for obtaining the variance of g.

In Section 2 results are obtained for the variance and probability distribution function of g using an alternative expression for variance of δ and the properties of order statistics.

It is easy to note the algebraic identity (David [2], p. 216)

$$(1/4) \sum_{i}^{n} \sum_{j}^{n} | x_i - x_j | = \sum_{i=1}^{n} (i - (n+1)/2) \, x_{(i)}$$

where $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ represents the ordering of the sample $x_1, x_2, \ldots, x_n$ in increasing order. Thus

$$\begin{aligned} g &= (1/n(n-1)) \, 4 \sum_{i=1}^{n} (i - (n+1)/2) \, x_{(i)}/(2\bar{x}) \\ &= 2S/(n-1) - (n+1)/(n-1) \end{aligned} \qquad (3)$$

where

$$S = \sum_{i=1}^{n} i \, x_{(i)} / \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} i \, x_{(i)} / \sum_{i=1}^{n} x_{(i)}$$

Simplification of the results for some specific distributions such as Normal, Log normal, Uniform and Pareto distributions has been suggested in Section 3.

2. Main Results

We shall need the following lemmas.

LEMMA 1 (Kendall and Stuart [4], pp. 246-247). *For random variables x_1 and x_2, an approximation for the expectation and variance of the ratio x_1/x_2 are*

$$E(x_1/x_2) = E(x_1)/E(x_2)$$

$$\begin{aligned} \text{Var}\,(x_1/x_2) = {} & (E(x_1)/E(x_2))^2 \, (\text{Var}\,(x_1)/(E(x_1))^2 \\ & + \text{Var}\,(x_2)/(E(x_2))^2 - 2\text{Cov}\,(x_1, x_2)/(E(x_1)\,E(x_2))) \end{aligned} \qquad (4)$$

LEMMA 2 (David [2], p. 81). *If $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ represent the order statistics for the random sample $x_1, x_2, \ldots, x_n$ of size n drawn from population with distribution funetion $F(x)$ and probability dcnsity function $f(x)$, then upto order $1/n^2$, we have the foliowing*

$$E(x_{(r)}) = Q_r^1 + p_r q_r\, Q_r^{11}/(2\,(n+2))$$
$$+\, p_r q_r\, [(q_r - p_r)\; Q_r^{111}/3 + p_r q_r\, Q_r^{1111}/8]/(n+2)^2 \qquad (5)$$

$$\mathrm{Var}\,(x_{(r)}) = p_r q_r\, (Q_r^1)^2/(n+2)$$
$$+\, p_r q_r\, [2(q_r - p_r)\, Q_r^1\, Q_r^{11} + p_r q_r\, (Q_r^1\, Q_r^{111} + (Q_r^{11})^2/2)]$$
$$/(n+2)^2 \qquad (6)$$

$$\mathrm{Cov}\,(x_{(r)}, x_{(s)}) = p_r q_s\, Q_r^1\, Q_s^1\, /(n+2)$$
$$+\, p_r q_s\, [(q_r - p_r)\, Q_r^{11}\, Q_s^1 + (q_s - p_s)\, Q_r^1\, Q_s^{11}$$
$$+\, p_r q_r\;\; Q_r^{111}\, Q_s^1\, /2 + p_s q_s\, Q_r^1\;\; Q_s^{111}/2 + p_r q_s\, Q_r^{11}\, Q_s^{11}/2]$$
$$/(n+2)^2 \qquad (7)$$

where

$$P_r = r/(n+1), \;\; q_r = 1 - p_r$$
$$Q_r = Q(p_r)$$
$$F(Q_r) = F(Q(p_r)) = p_r$$
$$Q_r^1 = dQ_r/dp_r,\; Q_r^{11} = d^2\, Q_r/dp_r^2\,,\, Q_r^{111} = d^3\, Q_r/dp_r^3$$
$$Q_r^{1111} = d^4 Q_r/dp^4$$

These results upto order $(n+2)^{-2}$ have been presented by David and Johnson [1].

We now have the main theorem.

THEOREM 1. *Against above background,*

(i) *the expected value and variance of the sample Gini's coefficient g are given by*

$$E(g) = 2 \sum_{i=1}^{n} i\; E(x_{(i)})/(n\,(n-1)\;\mu) - (n+1)/(n-1) \qquad (8)$$

$$\text{Var}(g) = (4/(n-1)^2)\left(\sum_{i=1}^{n} i\, E(x_{(i)})/(n\mu)\right)^2 \left(\sum_i \sum_j ij \,\text{Cov}\,(x_{(i)}, x_{(j)})\right)$$

$$/(\Sigma\, i\, E(x_{(i)}))^2 + \sigma^2/(n\mu^2)$$

$$- 2 \sum_{i=1}^{n} i \sum_{j=1}^{n} \text{Cov}\;(x_{(i)}, x_{(j)})\Big/\left(n\mu\left(\sum_i i\, E(x_{(i)})\right)\right) \qquad (9)$$

(ii) *the probability distribution function*

$H(t) = \text{Prob}\,[g \leqslant t]$, *is given by*

$$H(t) = \text{Prob}\left[\sum_{i=1}^{n} (2i - (n+1) - (n-1)\,t)\, x_{(i)} \leqslant 0\right] \qquad (10)$$

Proof. (i) follows from the expression (3) of g and Lemma 1; and (ii) from (3) and simplification.

The above theorem can be simplified in terms of the standardized variables $Z_{(r)}$ derived from $x_{(r)}$ using $Z_{(r)} = (x_{(r)} - \mu)/\sigma$ and noticing

$$\sigma^2 \,\text{Var}\,(Z_{(r)}) = \text{Var}\,(x_{(r)}).$$

THEOREM 2 (*Theorem* 1 *in terms of* $Z_{(r)}$'s)

(i) $\quad E(g) = 2C \sum_{i=1}^{n} i\, E(z_{(i)})/(n\,(n-1))$

$$\text{Var}\,(g) = (2C/(n\,(n-1)))^2\, (\sum_i \sum_j ij\, \text{Cov}\,(z_{(i)}, z_{(j)})$$

$$+ n\lambda^2 - 2\lambda \sum_i \sum_j i\, \text{Cov}\,(z_{(i)}, z_{(j)}))$$

where C = *Coefficient of Variation* (σ/μ) *and*

$$\lambda = (n+1)/2 + C \sum_i i\, E(z_{(i)})/n$$

(ii) $H(t) = \text{Prob}\,[g \leqslant t]$ *tends to*

$$\Phi((n\,(n-1)\,t/C - w_1)/w_2^{1/2})$$

where

$$\Phi\,(x) = \int_{-\infty}^{x} (1/(2\Pi)^{1/2}\, e^{-u^2/2}\, du$$

(*normal probability integral*),

$$w_1 = \sum_{i=1}^{n} (2i - (n - 1)\, t - (n + 1))\, E(z_{(i)})$$

$$w_2 = \sum_i \sum_j (2i - (n - 1)\, t - (n + 1)\, (2j - (n - 1)\, t - (n + 1)$$

$$\text{Cov}\,(z_{(i)}, z_{(j)}).$$

Proof. Using Lemma 1, and expression (3) we find

$$\begin{aligned} E(g) &= 2\, \Sigma\, i\, (\mu + \sigma E(z_{(i)}))/(n\,(n - 1)\, \mu) - (n + 1)/(n - 1) \\ &= (n\,(n + 1)\, \mu + 2\sigma\, \Sigma\, i\, E(z_{(i)}/(n\,(n - 1)\, \mu) - (n + 1)/(n - 1) \\ &= 2C\, \Sigma\, i\, E(z_{(i)})/(n\,(n - 1)); \quad C = \sigma/\mu \end{aligned}$$

Let us evaluate various terms in Theorem 1, expression (9)

$$\Sigma\, i\, E(x_{(i)})/(n\mu) = (n + 1)/2 + C\, \Sigma\, i\, E(z_{(i)})/n = \lambda \quad \text{say}$$

$$\Sigma\, i\, E(x_{(i)}) = n\lambda\mu$$

$$\begin{aligned} \sum_i \sum_j ij\, \text{Cov}\,(x_{(i)}, x_{(j)}) &= \sum_i \sum_j ij\, \text{Cov}\,(\mu + \sigma z_{(i)}, \mu + \sigma\, z_{(j)}) \\ &= \sigma^2 \sum_i \sum_j ij\, \text{Cov}\,(z_{(i)}, z_{(j)}) \end{aligned}$$

$$\sum_i \sum_j i\, \text{Cov}\,(x_{(i)}, x_{(j)}) = \sigma^2 \sum_i \sum_j i\, \text{Cov}\,(z_{(i)}, z_{(j)})$$

Thus the proof of (i) of Theorem 2 follows from Lemma 1, and the above expressions after substituting in Theorem 1.

To prove (ii) of Theorem 2, consider (ii) of Theorem 1

$$\begin{aligned} H(t) &= \text{Prob}\, [g \leqslant t] \\ &= \text{Prob}\, [\sum_{i=1} (2i - (n + 1) - (n - 1)\, t)\, x_{(i)} \leqslant 0] \\ &= \text{Prob}\, [\sum_{i=1} (2i - (n + 1) - (n - 1)\, t)\, (\mu + \sigma z_{(i)}) < 0] \\ &= \text{Prob}\, [w \leqslant n\,(n - 1)\, t/C] \end{aligned}$$

where

$$w = \sum_{i=1}^{n} (2i - (n + 1) - (n - 1)\, t)\, z_{(i)}$$

Since w is a linear function of order statistics, the asymptotic distribution of w may be approximated using normal distribution (see David [4], p. 273; Hoeffding [3]). Hence

$$H(t) \text{ tends to } \Phi\left((n(n-1)\, t/C - w_1)/w_2^{1/2}\right)$$

where w_1 and w_2 are the expectation and the variance of w respectively.

It is obvious from Theorem 2 that the moments and distribution of g can be obtained if we know the expectation, variances and covariances of $Z_{(i)}$'s. In the following section we shall evaluate the expectations, variances and covariances of $Z_{(i)}$'s in terms of the distribution functions, especially for four chosen populations (Normal, Log normal, Uniform ane Pareto distributions). This will require the representation of Q and its derivatives with respect to p_r in terms of the distribution functions.

3. Derivatives of Q_r for Some Chosen Distributions

1. *Normal* distribution (David [2], p. 81)

$$F(x) = \int_{-\infty}^{x} (1/(2\Pi)^{1/2}\,\sigma)\, e^{-((t-\mu)/\sigma)^2/2}\, dt = \Phi((x-\mu)/\sigma))$$

$$= \Phi(z), \quad z = (x-\mu)/\sigma$$

At $F(x) = p_r$

$$\therefore p_r = \Phi(z_{(r)}) \text{ and } Z_{(r)} = \Phi^{-1}(p_r)$$

Thus

$$Q_r = Q(p_r) = z_{(r)} = \Phi^{-1}(p^r)$$

Derivatives :

Note $\Phi(Q(p_r)) = p_r$

differentiating above with respect to p_r .

$$\phi(Q(p_r)\, dQ(p_r)/dp_r = 1$$

$$Q_r' = dQ(p_r)/dp_r = 1/\phi(Q(p_r)) = 1/\phi(Q_r)$$

where

$$\phi\,(Q_r) = (1/(2\Pi)^{1/2}\, e^{-Q_r^2/2}$$

$$Q_r^{11} = dQ_r^1/dp_r = -\,(\phi\,(Q_r))^{-2}\,(-\,Q_r\ \phi\,(Q_r))\cdot dQ_r/dp_r$$

$$= (\phi\,(Q_r))^{-2} Q_r\,\phi\,(Q^r)\cdot(1/\phi(Q^r)) = Q_r/(\phi\,(Q_r))^2$$

$$Q^{111} = d(Q_r/(\phi\,(Q_r))^2)/dp_r$$

$$= Q_r^1/(\phi\,(Q^r))^2 + Q_r\,((-\,2)/(\phi\,(Q_r))^3)\,(d\phi\,(Q_r)/dQ_r)\cdot dQ^r/dp^r$$

$$= (1 + 2Q_r^2)/(\phi\,(Q^r))^3$$

$$Q^{1111} = Q_r\,(7 + 6\,Q_r^2)/(\phi\,(Q_r))^4$$

2. *Log normal distribution*

$$F(x) = \int_0^x (1/(u\sigma\,(2\Pi)^{1/2}))\; e^{-((\log\,x-\mu)/\sigma)2/2}\,du$$

$$= \int_0^{\log x} (1/\sigma\,(2\Pi)^{1/2}))\; e^{-(u-\mu)/\sigma)2/2}\,du$$

$$= \Phi\,((\log x - \mu)/\sigma) = \Phi\,(z),$$

where $z = (\log x - \mu)/\sigma$.

Thus the case of log normal distribution can be derived on lines similar to normal distribution case.

3. *Uniform distribution* $U(0, 1)$

$$F(x) = x \qquad\qquad 0 \leqslant x \leqslant 1$$

At $F(x_{(r)}) = p_r$

$$x_{(r)} = p_r = Q(p_r)$$

Derivatives : $Q_r^1 = 1,\ Q_r^{11} = Q_r^{111} = Q_r^{1111} = 0$

4. *Exponential distribution*

$$F(x) = 1 - e^{-x/\theta} \qquad x, \theta > 0$$

At $F(x_{(r)} = p_r$

We get

$$x_{(r)} = -\theta \log (1 - p^r) = Q(p^r)$$

Derivatives :

$$Q_r^{1} = \theta (1 - p^r)^{-1}$$

$$Q_r^{11} = -\theta (1 - p^r)^{-2}$$

$$Q_r^{111} = 2\theta (1 - p^r)^{-3}$$

$$Q_r^{1111} = -6\theta (1 - p^r)^{-4}$$

5. *Pareto distribution*

$$f(x) = u\, a^u\, x^{-u-1} \qquad (a, u > 0, x \geqslant a)$$

$$F(x) = \int_a f(t)\, dt = 1 - (a/x)^u$$

For $F(x_{(r)}) = p_r$, we get

$$x_{(r)} = a(1 - p^r)^{-(1/u)} = Q(p^r)$$

Derivatives :

$$Q_r^{1} = (a/u)(1 - p^r)^{-(1+u)/u}$$

$$Q_r^{11} = -(a(1 + u)/u^2)(1 - p_r)^{-(1+2u)/u}$$

$$Q_r^{111} = (a(1 + u)(1 + 2u)/u^3)(1 - p^r)^{-(1+3u)/u}$$

$$Q_r^{1111} = -(a(1 + u)(1 + 2u)(1 + 3u)/u^4)(1 - p_r)^{-(1+4u)/u}$$

REFERENCES

[1] David, F. N. and Johnson, M. L. (1954) : Statistical treatment of censored data. I. Fundamental formulae. *Biometrika* 41 : 228-240.

[2] David, H. A. (1981) : *Order Statistics.* Second edition. New York, John Wiley and Sons Inc.

[3] Hoeffding, W. (1948) : A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* 19 : 293-325.

[4] Kendall, M. G. and Stuart, A. (1977) : *The Advanced Theory of Staiistics,* Volume I, Fourth edition. London, Charles Griffin and Company Limited.

[5] Kakwani, N. C. (1980) : *Income Inequality and Poverty.* New York, World Bank/ Oxford University Press.

[6] Iyengar, N. S. (1960) : On the standard error of the Lorenz concentration ratio. *Sankhya* : *The Indian Journal of Statistics* 22 : 371-378.