Statistics Canada Statistique Canada

Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Canada

[redacted]

WORKING PAPER
METHODOLOGY BRANCH

RESAMPLING PROCEDURES APPLIED TO VARIANCE ESTIMATION OF THE GINI COEFFICIENT ESTIMATOR

SSMD - 94 - 004 E

Nadine Labrèche

April 1994

Report produced at the Social Survey Methods Division, Statistics Canada, as partial requirement for the statistical internship program of Carleton University.

Resampling Procedures Applied to Variance Estimation of the Gini Coefficient Estimator

Abstract
Resampling methods are often considered for variance estimation of complex statistics estimated from stratified multistage designs. Three techniques, the balanced repeated replication, the jackknife and the bootstrap are considered here for the Gini coefficient. Using income data from the Survey of Consumer Finances, the jackknife and bootstrap methods are implemented and compared.

Key words: Gini coefficient, variance estimation, confidence intervals, balanced repeated replications, jackknife, bootstrap.

Application de méthodes de rééchantillonnage à l'estimation de la variance du coefficient Gini

Résumé
Les méthodes de rééchantillonnage sont souvent employées pour estimer la variance de statistiques complexes estimées à partir d'un plan de sondage stratifié et à plusieurs degrés. Nous étudions l'application des méthodes des répliques équilibrées répétées, de Quenouille-Tukey et d'auto-amorçage pour le coefficient Gini. Les deux dernières techniques sont mises en pratique en utilisant des données provenant de l'enquête sur les finances des consommateurs.

Mots clefs: coefficient Gini, estimation de la variance, intervalles de confiance, répliques équilibrées répétées, Quenouille-Tukey, auto-amorçage.

1 Introduction

Income inequality measures are used to study the shape and evolution over time of the size distribution of income. These quantities can be estimated from sample surveys. However, since these estimates are complex statistics, conventional variance estimation techniques fail and one has to rely on methods such as resampling procedures to provide information about sampling variability.

This paper presents three resampling methods for evaluating the variance of the Gini coefficient estimate: balanced repeated replication, jackknife and bootstrap. Using 1988 data from the Canadian Survey of Consumer Finance (SCF), the last two methods are implemented to obtain variance estimates and confidence intervals for the Gini coefficient.

2 Lorenz Curve and the Gini Coefficient

The distribution of income among the population can be depicted by observing the share of income received by the poorest p percent of the population. The Lorenz curve, which consists of plotting the cumulated percentage of the population (displayed from poorest to richest) against the percentage of total wealth held by that group, is a graphical representation of that quantity.

Perfect income equality is attained when the poorest p percent of the population receives p percent of the total income. As a result, the closer the Lorenz curve is to the diagonal in the Lorenz diagram, the lesser is the inequality in the distribution. This distance is measured by the area between the diagonal and the Lorenz curve: it is called the Lorenz area (LA). The Gini coefficient is defined as the ratio between the LA and the largest possible LA: $G = 2\,\mathrm{LA}$, and hence $0 \leq G \leq 1$.

Let $F(y)$ denote the distribution function of a variable Y (e.g. household income) with finite mean $\mu = \int_{-\infty}^{\infty} y dF(y) \neq 0$. The share of the poorest $p = F(y)$ percent of the population can then be expressed as $L(p) = \mu^{-1} \int_{-\infty}^{y} t dF(t)$. If we define the inverse of the distribution function as

$$F^{-1}(p) = \begin{cases} \inf_y\{y|F(y) > 0\} & \text{if } p = 0 \\ \inf_y\{y|F(y) \geq p\} & \text{if } 0 < p \leq 1 \end{cases}$$

the Lorenz curve ordinate with abscissa p can be written as $L(p) = \int_0^p F^{-1}(t)dt/\mu$. The LA is then given by $\int_0^1 (p - L(p))dp$ and the Gini coefficient by $2\int_0^1 (p - L(p))dp$. The Gini coefficient can also be written as

$$G = \frac{\int_0^1 [2F(y) - 1] y dF(y)}{\mu}$$

(see Nygård and Sandström, 1985b).

Computation of the Gini coefficient in a finite population is done as follows. The finite population distribution function F_N is defined as

$$F_N(y) = N^{-1} \sum_{i=1}^{N} I\{y_i \leq y\}$$

where $I\{\cdot\}$ is the indicator function which takes the value 1 if $\{\cdot\}$ is true, and 0 otherwise.

Suppose there are $N^* \leq N$ distinct values of y. We define the probability function at $y_{(i)}$, where $y_{(1)} < y_{(2)} < \cdots < y_{(N^*)}$, as

$$f_N(y_{(i)}) = F_N(y_{(i)}) - F_N(y_{(i-1)}).$$

For unordered distinct values y_i, the Gini coefficient becomes

$$G_N = \frac{\sum_{i=1}^{N^*} [2F_N(y_i) - 1 - f_N(y_i)]\, y_i f_N(y_i)}{\mu_N} \tag{2.1}$$

where $\mu_N = \sum_{i=1}^{N^*} y_i f_N(y_i)$ and the term $-f_N(y_i)$ appearing within the brackets is the Gini finite population correction (Gfpc). In the case where no tied values are present, $f_N(y_i) = 1/N$, $F_N(y_{(i)}) = i/N$, $\mu_N = \sum_{i=1}^N y_i/N$ and we can write

$$G_N = \frac{\sum_{i=1}^N \left[2\frac{i}{N} - 1 - \frac{1}{N}\right] y_{(i)}}{\mu_N N} = \frac{\sum_{i=1}^N \frac{2i-1}{N} y_{(i)}}{\sum_{i=1}^N y_i} - 1.$$

In the case of survey samples, we have a sample s of n observations $y_1, \ldots, y_n$ to which are attached the corresponding weights $w_1, \ldots, w_n$. Define $\tilde{w}_i = w_i/\hat{N}$ where $\hat{N} = \sum_{i\in s} w_i$. An estimator of the finite population distribution function F_N is

$$\hat{F}_N(y) = \sum_{j\in s} \frac{w_j I\{y_j \leq y\}}{\hat{N}} = \sum_{j\in s} \tilde{w}_j I\{y_j \leq y\}.$$

Also $\hat{\mu}_N = \sum_{i\in s} \tilde{w}_i y_i$ and $\hat{f}_N(y_{(i)}) = \hat{F}_N(y_{(i)}) - \hat{F}_N(y_{(i-1)}) = \tilde{w}_i$, i.e., the Gfpc for the i-th observation is $-\tilde{w}_i$. Hence, from (2.1),

$$\begin{aligned} \hat{G}_N &= \hat{\mu}_N^{-1} \sum_{i\in s} [2\hat{F}_N(y_i) - 1 - \tilde{w}_i]\tilde{w}_i y_i \\ &= \frac{2\sum_{i\in s} \hat{F}_N(y_i)\tilde{w}_i y_i}{\sum_{i\in s} \tilde{w}_i y_i} - 1 - \frac{\sum_{i\in s} \tilde{w}_i^2 y_i}{\sum_{i\in s} \tilde{w}_i y_i}. \end{aligned} \tag{2.2}$$

Note that if we omit the Gfpc, the last term of (2.2) vanishes.

3 Variance Estimation

The Gini coefficient is defined in terms of the finite population distribution function F_N. It is therefore a complex statistic whose variance cannot be expressed by a simple formula nor can it be easily estimated by conventional means. Its variance may be approximated by variance estimation techniques such as balanced repeated replication, jackknife and bootstrap.

The main quality of these methods is that they use a single variance formula for all statistics. Hence, though the following description involves θ's, it applies to the Gini coefficient $G_N = G(F_N)$. Also, to reflect their application to the SCF, the three methods are described for stratified multistage designs with unequal number of clusters selected in each stratum. (An overview of the SCF design is given in the next section.) In order that the results for the balanced repeated replication and jackknife methods be valid, we assume that the clusters are selected with replacement and that independent subsamples are selected within clusters selected more than once.

3.1 Balanced Repeated Replication

Many surveys employ stratification to the extent that only a few primary units are selected from each stratum. For the case $n_h = 2$ clusters per stratum, the balanced repeated replication (BRR) method is commonly used for variance estimation of the parameter of interest θ (see Wolter, 1985, chapter 3).

The BRR method can be extended to the case $n_h \geq 2$ clusters per stratum. Furthermore, the case of unequal n_h in each stratum must be considered for surveys such as the SCF. Wu (1991) proposed the following method. A set of R replicates are formed by selecting one sample unit from each stratum. This set is defined by a $R \times L$ design matrix (δ_h^r), $r = 1, \ldots, R$, $h = 1, \ldots, L$ with $\delta_h^r = 1, \ldots, n_h$, say, depending on whether the first, second, ..., or n_h-th sample cluster is in the h-th stratum of the r-th replicate. Ideally, all columns of the matrix should be mutually orthogonal, i.e., each combination of selected clusters should appear equally often. This orthogonality condition results however in a large number of required replicates R, for general n_h. Furthermore, orthogonal matrices do not exist for all combinations of n_h's. One solution consists of using mixed orthogonal arrays of strength 2.

A mixed orthogonal array of strength d, $(R, n_1 \times \cdots \times n_L, d)$ is an $R \times L$ matrix whose h-th column has n_h symbols (say $1, \ldots, n_h$) arranged such that for any d columns, each possible combination of symbols appears equally often. Only tables of strength 2 are considered since no major gain in efficiency is obtained by considering $d \geq 3$, while the number of required replications, R, increases considerably.

The variability between the R replicate estimates approximates the sampling variance of the estimator $\hat{\theta}$ (e.g. $\hat{G}_N = G(\hat{F}_N)$). Let $\hat{\theta}^{(r)}$ be the estimator of θ obtained from the r-th replicate. The estimator $\hat{\theta}^{(r)}$ is calculated by using the weight adjustments obtained from formula (6) in Wu (1991). The weight of the i-th element of the c-th cluster of stratum h, w_{hci}, is transformed at the r-th replication to $A_{hci}^{(r)} w_{hci}$ where

$$A_{hci}^{(r)} = \begin{cases} 1 + \sqrt{n_h - 1} & \text{if the } (hci)\text{-th element is selected in replicate } r \\ 1 - \frac{1}{\sqrt{n_h - 1}} & \text{otherwise.} \end{cases}$$

The computation of $\hat{\theta}^{(r)}$ is then performed by using the modified weights in the formula for $\hat{\theta}$. A BRR variance estimator of $\hat{\theta}$ is given by

$$v_{BRR}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2. \tag{3.3}$$

If $\hat{\theta}^{(\cdot)} = \sum_r \hat{\theta}^{(r)}/R$ is substituted for $\hat{\theta}$ in (3.3), another variant of the BRR variance estimator is obtained.

The mixed orthogonal array approach may not be applicable to all cases. The mixed-orthogonality condition leads to the same problems as above. In order to find an economical mixed orthogonal array, the n_h clusters in stratum h can be grouped into two to four groups of clusters. The BRR method can then be applied to the groups by treating them as units. Given the large number of clusters in some strata of the SCF (e.g. $n_h = 16$), the grouped BRR method seems suitable for this survey.

3.2 Jackknife

The jackknife method was first developed to approximate the variance of smooth functions θ of independent, identically distributed (i.i.d.) observations. The method roughly consists of computing $\hat{\theta}_{(i)}$, the estimate obtained from omitting the i-th observation ($i = 1, \ldots, n = \sum n_h$) and then estimating the variance by the variability among these replicate statistics.

Kovar, Rao and Wu (1988) showed the inconsistency of the jackknife variance estimator for non-smooth statistics. A generalized version of this method, the delete-one cluster jackknife, has been shown to perform adequately (see Shao and Wu, 1989; Rao, Wu and Yue, 1992). In particular, Shao (1993) showed that under weak conditions, the asymptotic variance of the Gini coefficient can be consistently estimated by delete-one cluster jackknifing.

The Gini coefficient estimator can be expressed as $\hat{\theta} = G(\hat{F}_N)$ where $\hat{F}_N$ is the estimated distribution function. Let $\hat{F}_{(gj)}$ be the estimator of the distribution function based on the subsample obtained by removing the j-th cluster of the g-th stratum, ($j = 1, \ldots, n_g; g = 1, \ldots, L$):

$$\hat{F}_{(gj)}(y) = \sum_{\substack{(hci) \in s \\ (hc) \neq (gj)}} \tilde{w'}_{hci} I\{y_{hci} \leq y\}$$

where $\tilde{w'}_{hci} = w'_{hci} / \sum_s w'_{hci}$ are the normalized weights modified for the jackknife procedure. To compensate for the removal, the weights are adjusted to

$$w'_{hci} = \begin{cases} w_{hci} & h \neq g \\ A_{(gj)} w_{hci} & h = g,\, c \neq j \\ 0 & h = g,\, c = j. \end{cases}$$

Wishing to keep $\sum_s w'_{hci} = \sum_s w_{hci}$, Kovačević and Pandher (1993) define

$$A_{(gj)} = \frac{\hat{N}_g}{\hat{N}_g - \hat{N}_{gj}} \tag{3.4}$$

where $\hat{N}_g = \sum_c \sum_i w_{gci}$ and $\hat{N}_{gj} = \sum_i w_{gji}$. The usual adjustment factor is

$$A_{(gj)} = \frac{n_g}{n_g - 1}. \tag{3.5}$$

When the primary sampling units (clusters) are selected with probability proportional to cluster size, however, the two are equal. Normalization of the modified weights is attained by dividing w'_{hci} by $\hat{N} = \sum_s w_{hci}$ or $\hat{N} + \frac{1}{n_g - 1}[\hat{N}_g - n_g \hat{N}_{gj}]$ depending on which of the adjustments (3.4) or (3.5) is applied.

In both cases, a delete-one cluster jackknife estimator of the variance of $\hat{\theta}$ is given by

$$v_{J1}(\hat{\theta}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\theta}_{(gj)} - \hat{\theta})^2 \tag{3.6}$$

where $\hat{\theta}_{(gj)} = G(\bar{F}_{(gj)})$. A variation of (3.6) is obtained by changing $\hat{\theta}$ to $\hat{\theta}_{(..)} = \sum_g \sum_j \hat{\theta}_{(gj)}/n$:

$$v_{J2}(\hat{\theta}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\theta}_{(gj)} - \hat{\theta}_{(..)})^2. \tag{3.7}$$

3.3 Bootstrap

The principle of the bootstrap method is to select with replacement a large number of samples from the original sample. The variability between the estimates θ^* of these bootstrap samples approximates the variance of the estimator $\hat{\theta}$.

Efron (1982) gave a Monte Carlo algorithm for estimating the variance of the estimator $\hat{\theta}$ in the i.i.d. case. Rao and Wu (1988) extended this method to stratified multistage designs, covering smooth as well as non-smooth statistics. Their method is the following.

Independently for each stratum h, a simple random sample of m_h clusters is drawn with replacement from the n_h sample clusters. In order to ensure consistency of the bootstrap variance estimator, the survey weights, w_{hci}, are rescaled to the bootstrap weights $w^*_{hci} = A_{hci} w_{hci}$ where

$$A_{hci} = 1 - \sqrt{\frac{m_h}{n_h - 1}} + \sqrt{\frac{m_h}{n_h - 1} \frac{n_h}{m_h}} m^*_{hc} \tag{3.8}$$

and m^*_{hc} counts the number of times the (hc)-th sample cluster is selected $(\sum_c m^*_{hc} = m_h)$. Note that if $m^*_{hc} = 0$, the last term of (3.8) disappears. The choice of $m_h \leq n_h - 1$ ensures that the bootstrap weights w^*_{hci} are all positive if $w_{hci} > 0$ for all $(hci) \in s$. Also, if $m_h = n_h - 1$, the adjustment factor reduces to $\left(\frac{n_h}{n_h - 1}\right) m^*_{hc}$. The bootstrap estimate θ^* is obtained by using the bootstrap weights w^*_{hci} in the formula for $\hat{\theta}$.

This procedure is repeated independently a large number, B, of times and yields the bootstrap estimates $\theta^*_1, \ldots, \theta^*_B$. The estimator

$$v_{B1}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} [\theta^*_b - \hat{\theta}]^2 \tag{3.9}$$

approximates the bootstrap variance estimator $E_*(\theta^* - E_*\theta^*)^2$, where E_* denotes the expectation with respect to bootstrap sampling. Another variance estimator is obtained by substituting $\hat{\theta}$ by $\bar{\theta}^* = \sum_b \theta^*_b / B$:

$$v_{B2}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} [\theta^*_b - \bar{\theta}^*]^2. \tag{3.10}$$

This estimator yields a lower estimate value than $v_{B1}(\hat{\theta})$ since $\sum_{b=1}^{B} [\theta^*_b - x]^2$ is minimized at $x = \bar{\theta}^*$. Finally, note that a Monte Carlo error affects the variance estimates since different estimate values are obtained for different bootstrap samples.

4 Canadian Survey of Consumer Finance (SCF)

The Canadian Survey of Consumer Finance (SCF) is a special survey conducted by Statistics Canada every two years to collect information about the financial situation of households. The SCF uses the sample frame and sampling procedures of the Canadian Labour Force Survey (LFS), whose framework is based on a stratified, multistage design. A detailed description of the LFS design is given in Singh et al. (1990).

Two characteristics of this design are particularly relevant to the application of repeated sampling methods to variance estimation of the Gini coefficient estimate. First, clusters in the LFS are selected with probability proportional to sizes without replacement. At the variance estimation stage however, clusters are treated as though they were selected with replacement, and subsampling done independently each time a cluster is selected, to simplify the calculations. Second, the final weight attached to each record is the result of complex operations. The basic weight (inverse of sampling ratio) is first corrected for factors such as nonresponse. A generalized regression estimator is then used to ensure consistency of the sample with known totals of some relevant post-stratification variables.

5 Application of Jackknife and Bootstrap to SCF

The application of jackknife and bootstrap resampling methods to obtain variance estimates for $\hat{G}_N$ is illustrated by using family income data collected in the SCF in 1988 (SCF-88). The file on the disposable income of economic families obtained for the province of Ontario was used. Disposable income is defined as the total income reduced by the tax reported in the survey.

The SCF-88 Ontario sample comprised 7474 households grouped into 525 clusters allocated in 91 strata. The number of clusters in each stratum varies from 2 to 16 as shown in this table:

# clusters in stratum	2	3	4	5	6	8	10	12	14	16
Frequency	36	2	12	1	5	18	10	4	4	1

5.1 Gini Coefficient Estimate

Figure 1 shows the empirical distribution function of family income obtained from the SCF-88 sample in Ontario. The corresponding Lorenz curve appears in Figure 2.

Given the complexity of the survey weights in the SCF, the Gfpc is dropped and the Gini coefficient estimator (2.2) becomes

$$\hat{G}_N = 2\frac{\sum_{(hci)\in s} \hat{F}_N(y_{hci})\tilde{w}_{hci}y_{hci}}{\sum_{(hci)\in s} \tilde{w}_{hci}y_{hci}} - 1.$$

The new subscript notation indicates the multiple stages of the data. The index hci refers to the i-th unit in the c-th cluster of stratum h. The Gini coefficient estimate obtained from the SCF-88 Ontario data is $\hat{G}_N = 0.34836$.

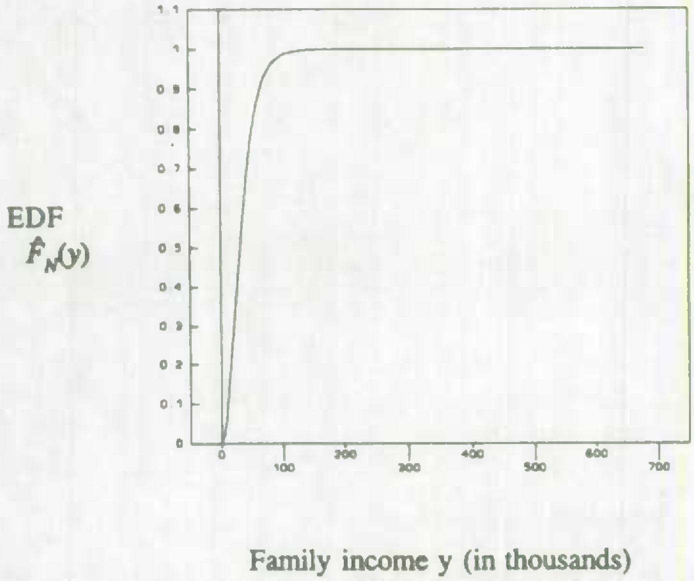

Figure 1: Empirical distribution function of family income in Ontario – SCF-88.

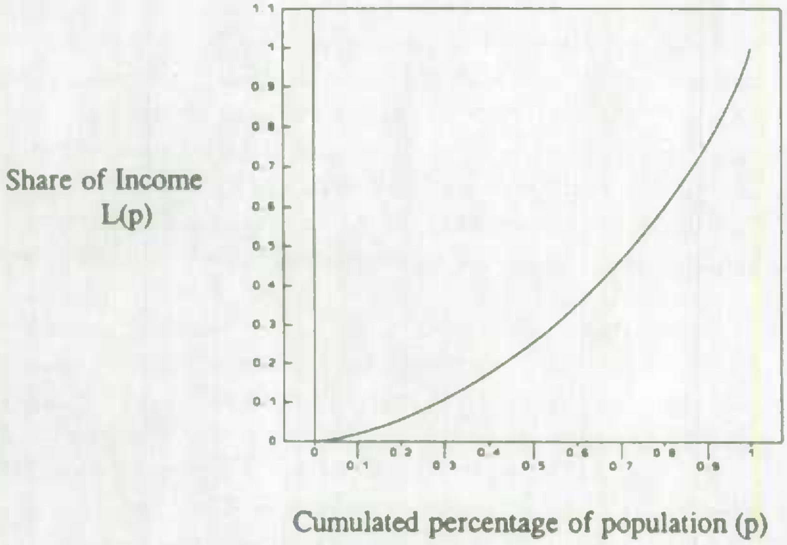

Figure 2: Lorenz curve for income distribution in Ontario – SCF-88.

Method	Estimated Variance	
	$v_{\dagger 1}$	$v_{\dagger 2}$
Jackknife-1 $(\dagger = J)$	.000024010	.000024009
Jackknife-2 $(\dagger = J)$	.000022694	.000022694
Bootstrap $(\dagger = B)$	.000029846	.000023816

NOTE: v_{J1}, v_{J2}, v_{B1} and v_{B2} refer respectively to Equations (3.6), (3.7), (3.9) and (3.10).

Table 1: Estimated variance for the Gini coefficient estimate obtained from the SCF-88 sample for Ontario.

5.2 Variance Estimates

Three techniques were applied to the SCF-88 Ontario data to estimate the variance of the Gini coefficient estimator: two variants of the delete-one cluster jackknife and the bootstrap. The delete-one cluster jackknife procedure was implemented for both weight adjustments (3.4) and (3.5). Call these variants jackknife-1 and jackknife-2 respectively. In each case, the variance of the Gini coefficient was estimated with estimators (3.6) and (3.7). The validity of these estimators is based on the assumption that the clusters are selected independently, which is approximately true for the LFS design. In that case, the jackknife variance estimator tends to slightly overestimate the variance and, therefore, is conservative.

The bootstrap method was implemented with $B = 525$ iterations so that the number of subsamples selected would be the same as in the delete-one cluster jackknife applications (there are 525 clusters). The size of the subsamples was fixed to $m_h = n_h - 1$. This choice simplifies the weight adjustment factor (3.8). The bootstrap variance of the Gini coefficient estimate was approximated by (3.9) and (3.10).

Table 1 displays the variance estimates for the estimated Gini coefficient. The jackknife-1, jackknife-2 and bootstrap SAS programs created to obtain these values are given respectively on pages A1 to A3.

We observe that the two variants $v_{\dagger 1}$ and $v_{\dagger 2}$ are much closer in the case of the jackknife than for the bootstrap. The bootstrap yields a higher variance estimate when the first estimator is used. The second estimated variances ($v_{\dagger 2}$) are similar. Note that the jackknife-1 and jackknife-2 procedures do not give the same variance estimates. Hence the condition for the equality of the two weight adjustment factors (3.4) and (3.5) does not hold. This may be explained by the complexity of the final weight.

To pursue the comparison of the three procedures, we consider confidence intervals.

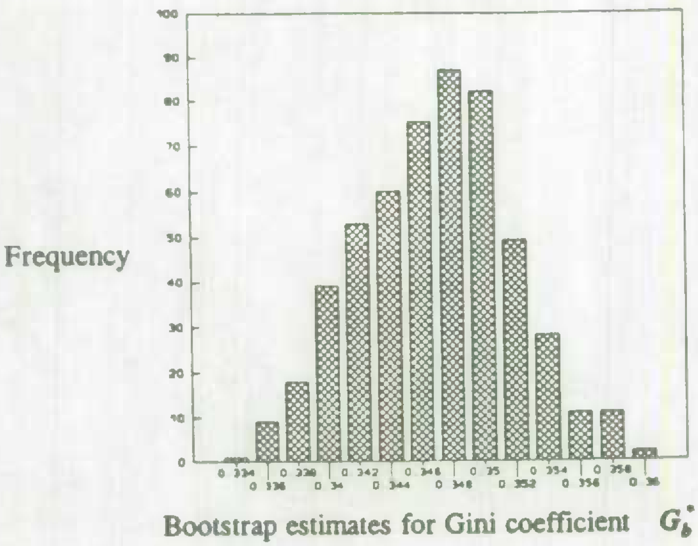

Figure 3: Frequency histogram of Gini coefficient bootstrap estimates G_b^* - SCF-88, Ontario.

5.3 Confidence Intervals

Figure 3 presents the frequency histogram of the Gini coefficient bootstrap estimates G_b^*, $b = 1, \ldots, 525$, obtained from the SCF-88 Ontario sample.

An approximate $100(1-\alpha)\%$ confidence interval for G_N may be derived from the bootstrap histogram by using the percentile method which is described as follows. Let

$$\begin{aligned} G^*_{LOW}(\alpha/2) &= \inf\{G_b^* | F_*(G_b^*) \geq \alpha/2\} \\ \text{and} \qquad G^*_{UP}(\alpha/2) &= \inf\{G_b^* | F_*(G_b^*) \geq 1-\alpha/2\} \end{aligned} \tag{5.11}$$

where $F_*(t) = \#\{G_b^* \leq t; b = 1, \ldots, B\}/B$. Then the interval $[G^*_{LOW}(\alpha/2), G^*_{UP}(\alpha/2)]$, which consists of the central $1-\alpha$ proportion of the bootstrap distribution, is an approximate $100(1-\alpha)\%$ confidence interval for G_N.

Visual inspection of Figure 3 reveals that the bootstrap estimates tend to underestimate $\hat{G}_N$. In fact, $\text{Prob.}_*\{G_b^* \leq \hat{G}_N\} = \#\{G_b^* \leq .34836; b = 1, \ldots, B\}/B = 68.8\%$. Efron (1982, p.82) suggested a percentile method which corrects the bias when $\text{Prob.}_*\{\hat{\theta}^* \leq \hat{\theta}\} \neq .50$, for general estimator $\hat{\theta}$. Taking $\hat{\theta} = \hat{G}_N$, define

$$\begin{aligned} G^*_L(\alpha/2) &= \inf\{G_b^* | F_*(G_b^*) \geq \Phi(z_0 - z_{\alpha/2})\} \\ \text{and} \qquad G^*_U(\alpha/2) &= \inf\{G_b^* | F_*(G_b^*) \geq \Phi(z_0 + z_{\alpha/2})\} \end{aligned} \tag{5.12}$$

where $z_0 = \Phi^{-1}(F_*(\hat{G}_N))$, $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$, Φ is the cumulative distribution function of a standard normal and F_* is defined as above. The bias corrected percentile method consists of taking

$$[G^*_L(\alpha/2), G^*_U(\alpha/2)]$$

as an approximate $100(1-\alpha)\%$ confidence interval for G_N. Note that if $\text{Prob.}_*\{G_b^* \leq \hat{G}_N\} = .50$, then $z_0 = 0$, $\Phi(z_0 - z_{\alpha/2}) = \alpha/2$ and $\Phi(z_0 + z_{\alpha/2}) = 1 - \alpha/2$, i.e., (5.12) reduces to (5.11) and the two confidence intervals are the same.

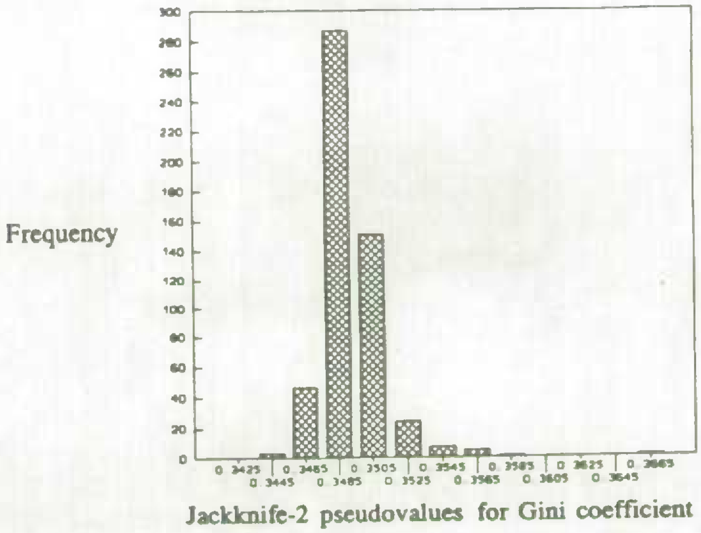

Figure 4: Frequency histogram of jackknife-2 pseudovalues – SCF- 88, Ontario.

The 95% percentile method and bias-corrected percentile method intervals obtained from the bootstrap histogram in Figure 3 are given in Table 2.

Figure 4 presents the frequency histogram of the Gini coefficient jackknife-2 estimates. (The histogram of the jackknife-1 estimates is similar and hence is not shown.) The estimate $\tilde{G}_{gj}$, called pseudovalue, is obtained from the delete-one cluster estimate $\hat{G}_{(gj)}$ and the original sample estimate $\hat{G}_N$ by the formula

$$\tilde{G}_{gj} = n_g\hat{G}_N - (n_g - 1)\hat{G}_{(gj)}.$$

The pseudovalues are highly concentrated around the estimate $\hat{G}_N = .34836$ and, moreover, symmetrically distributed about that value. This explains why normality is usually assumed to produce confidence intervals from jackknife variance estimates. A $100(1-\alpha)\%$ confidence interval based on the normal approximation is given by

$$[\hat{G}_N - z_{\alpha/2}\sqrt{v_J}, \hat{G}_N + z_{\alpha/2}\sqrt{v_J}].$$

The 95% confidence intervals for jackknife-1 and jackknife-2 appear in Table 2. Since $v_{J1} \doteq v_{J2}$ in both cases, only one confidence interval is given for each method.

The two percentile method bootstrap intervals are asymmetric about $\hat{G}_N = .34836$ but to a different extent. In fact, the bounds of the bias-corrected interval are similar to those of the jackknife symmetric intervals. The percentile method interval reflects the tendency of the bootstrap to underestimate $\hat{G}_N$. Interval lengths are similar for all methods.

Note that an additional variance estimator can be derived from the boostrap intervals (5.11) and (5.12) by equating these intervals to the normal theory interval for $\hat{G}_N$:

$$v_\alpha(\hat{G}_N) = \left[\frac{L_*(\alpha)}{2z_{\alpha/2}}\right]^2 \tag{5.13}$$

METHOD		INTERVAL LENGTH
BOOTSTRAP		
- percentile method	[.33706, .35599]	.01893
- bias corrected percentile method	[.33873, .35722]	.01849
JACKKNIFE-1	[.33876, .35796]	.01920
JACKKNIFE-2	[.33902, .35770]	.01868

Table 2: (Approximate) 95% confidence intervals for the Gini coefficient – SCF-88, Ontario.

Bootstrap Intervals	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	$\alpha = .50$
Percentile method interval (5.11)	.000019950	.000023321	.000022516	.000024595
Bias corrected percentile method (5.12)	.000018805	.000022249	.000026049	.000023720

Table 3: Variance estimator $v_\alpha(\hat{G}_N)$ for different choices of α.

where $L_*(\alpha)$ is the length of the bootstrap interval of size $1 - \alpha$. Values of (5.13) are given in Table 3 for different choices of α.

Clearly, $v_\alpha(\hat{G}_N)$ depends on α and hence the optimal choice of α has to be found. Note, however, that except for values corresponding to $\alpha = .01$, the variance estimates are similar to those in Table 1.

6 Conclusion

The variance estimates and confidence intervals obtained result from the application of the bootstrap and jackknife procedures to *one* sample. Thus no major conclusion may be drawn from this work. The values in Tables 1 and 2 seem to indicate nonetheless that the bootstrap and jackknife techniques lead to similar results. The "best" method should then be the one which is the simplest to apply.

The bootstrap method requires that a new sample be drawn independently at each iteration, and thus is much more computer-intensive than the jackknife, where each subsample is predetermined (delete one cluster). The latter method therefore seems preferable. Also, since there is no apparent reason to use the weight adjustment factor (3.4) in the delete-one jackknife, the usual adjustment factor (3.5) should be used, i.e., we recommend jackknife-2.

7 References

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Foster, James E., and Wolfson, Michael C. (1992). Polarisation and the Decline of the Middle Class: Canada and the U.S. Preprint.

Kovačević, M.S., and Pandher, G.S. (1993). Estimating Sampling Variance for Measures of Income Inequality and Polarization. Mimeo, Methodology Branch, Statistics Canada, Ottawa.

Kovar, J.G., Rao, J.N.K., and Wu, C.F.J. (1988). Bootstrap and Other Methods to Measure Errors in Survey Estimates. *The Canadian Journal of Statistics*, 16, Supplement, 25-45.

Nygård, F., and Sandström, A. (1985a). Income Inequality Measures Based on Sample Surveys. Invited paper, 45th session of the International Statistical Institute, Amsterdam.

Nygård, F., and Sandström, A. (1985b). The Estimation of the Gini and the Entropy Inequality Parameters in Finite Populations. *Journal of Official Statistics*, 1, 399-412.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling Inference With Complex Survey Data. *Journal of the American Statistical Association*, 83, 231-241.

Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, 209-217.

Sandström, A., Wretman, J.H., and Waldén, B. (1988). Variance Estimators of the Gini Coefficient – Probability Sampling. *Journal of Business & Economic Statistics*, 6, 113-119.

Shao, J., and Wu, C.F.J. (1989). A General Theory for Jackknife Variance Estimation. *Annals of Statistics*, 17, 1176-1197.

Shao, J. (1993). Inferences Based on L-Statistics in Survey Problems: Lorenz Curve, Gini's Family and Poverty Proportion. *Proceedings of the Workshop on Statistical Issues in Public Policy Analysis*, Carleton University and University of Ottawa.

Singh, M.P., Drew, J.D., Gambino, J.G., and Mayda, F. (1990). *Methodology of the Canadian Labour Force Survey*, Ottawa, Statistics Canada, Catalogue No. 71-526.

Wolter, K. (1985). *Introduction to Variance Estimation.* Springer-Verlag, New York.

Wu, C.F.J. (1991). Balanced Repeated Replications Based on Mixed Orthogonal Arrays. *Biometrika*, 78, 181-188.

```
LIBNAME in 'f:\sasuser';
OPTIONS PAGESIZE=55;
%LET numclust=525;


*** --------------------------------------------------------- ***;
***                 To print the results.                     ***;
*** --------------------------------------------------------- ***;


%MACRO printout(file);
  proc print data=&file;
       title "&file";
  run;
%MEND printout;


*** --------------------------------------------------------- ***;
***      This macro estimates the Gini coefficient            ***;
***                from the (sub)sample.                      ***;
*** --------------------------------------------------------- ***;

%MACRO estimate(file,sizen);
PROC SORT DATA=sample;
  BY y;
  RUN;

PROC SUMMARY DATA=sample;
  VAR wgt wgtedy;
  OUTPUT OUT=est(DROP= _freq_ _type_)
         SUM=nhat yhat;
  RUN;           *population size and total income estimates;

DATA edf;
  SET sample;
  IF _N_=1 THEN SET est;
  topedf + wgt;
  fhat = (topedf/nhat);      *empirical distribution function;
  gc = 2 * fhat * wgtedy/yhat;
  RUN;

PROC SUMMARY DATA=edf;
  VAR gc;
  OUTPUT OUT=gindex
         SUM=sumgc;
  RUN;

DATA gindex(KEEP=ginicoef ng);
  SET gindex;
  ginicoef=sumgc - 1;                         *Gini coefficient;
  ng=&sizen;
  RUN;

PROC APPEND BASE = &file
            DATA = gindex;
    RUN;
%MEND estimate;


*** --------------------------------------------------------- ***;
***                  J A C K K N I F I N G                    ***;
***  At k-th iteration (k=1,...,525) we remove observa-       ***;
***  tions from the k-th cluster in the sample and            ***;
***  adjust the weights of the remaining records in the       ***;
***  stratum by N-hatg/(N-hatg - N-hatgj) where               ***;
***  N-hatg  = sum of the weights in stratum g                ***;
***  N-hatgj = sum of the weights in cluster k.               ***;
```

```
*** --------------------------------------------------------- ***;

%MACRO jackit;
%DO i=1 %TO &numclust;
DATA _NULL_;
  SET clusters;
  IF _N_ = &i THEN DO;
    CALL SYMPUT('strat',strata);
    CALL SYMPUT('repnow',rep);
    CALL SYMPUT('smalln',ng);
    CALL SYMPUT('weight',rwt);
    STOP;
  END;
  RUN;              *information about cluster to be deleted;

%put &strat; %put &repnow; %put &smalln; %put &weight;

DATA sample(KEEP=wgt y wgtedy rep);
  SET ontario;
  IF rep=&repnow THEN jackind = 0;
  ELSE IF strata = &strat THEN jackind=nhatg/(nhatg-&weight);
  ELSE jackind=1;
  wgt = wt*jackind;
  wgtedy = wgt*y;
  RUN;                                        *weight adjustment;

%estimate(ginijack,&smalln);
%END;
%MEND jackit;



*****************************************************************;
***                 M A I N   P R O G R A M                   ***;
*****************************************************************;
PROC SORT DATA=in.ont_rep
        OUT=clusters(KEEP=strata rep ng rwt);
  BY rep;
  RUN;

PROC SORT DATA=in.ont_new
        OUT=ontario;
  BY rep;
  RUN;

DATA sample(KEEP=strata wgt y wgtedy);
  SET ontario;
  wgt = wt;
  wgtedy = wgt * y;
  RUN;

PROC SUMMARY DATA=sample NWAY;
  BY strata;
  VAR wgt;
  OUTPUT OUT=estN
         SUM=nhatg;
  RUN;        *to determine the sum of the weights in cluster;
                                                *to be removed;

DATA ontario(KEEP=strata rep wt y nhatg);
  MERGE ontario estN;
  BY strata;
  RUN;

%estimate(giniest,0); *Gini index estimate from original sample;
```

```
%jackit;

DATA giniest;
  SET giniest;
  RENAME ginicoef = gchat;
  DROP ng;
  RUN;

PROC SUMMARY DATA=ginijack;
  VAR ginicoef;
  OUTPUT OUT=meanjack(KEEP=gcmean)
         MEAN=gcmean;
  RUN;

* We compute two variance estimators using:                ;
*   1) gchat  = Gini index estimate from original sample    ;
*   2) gcmean = mean of the Gini index estimates from       ;
*               bootstrap samples                           ;
DATA varest
     in.jackres;
  SET ginijack;
  IF _N_=1 THEN DO;
     SET giniest;
     SET meanjack;
  END;
  vjack1=(ng-1)/ng * ( (ginicoef-gchat) **2);
  vjack2=(ng-1)/ng * ( (ginicoef-gcmean) **2);
  RUN;

%printout(varest);

PROC MEANS DATA=varest SUM;
  VAR vjack1 vjack2;
  TITLE 'Jackknife Variance Estimators for the Gini Coefficient';
  RUN;

PROC DATASETS;
  DELETE giniest ginijack;
  RUN;
```

```
LIBNAME in 'f:\sasuser';
OPTIONS PAGESIZE=55;
%LET numclust=525;

*** ---------------------------------------------------- ***;
***                 To print the results.                ***;
*** ---------------------------------------------------- ***;

%MACRO printout(file);
  proc print data=&file;
       title "&file";
  run;
%MEND printout;


*** ---------------------------------------------------- ***;
***      This macro estimates the Gini coefficient       ***;
***                from the (sub)sample.                 ***;
*** ---------------------------------------------------- ***;

%MACRO estimate(file,sizen);
PROC SORT DATA=sample;
  BY y;
  RUN;

PROC SUMMARY DATA=sample;
  VAR wgt wgtedy;
  OUTPUT OUT=est(DROP= _freq_ _type_)
         SUM=nhat yhat;
  RUN;            *population size and total income estimates;

DATA edf;
  SET sample;
  IF _N_=1 THEN SET est;
  topedf + wgt;
  fhat = (topedf/nhat);      *empirical distribution function;
  gc = 2 * fhat * wgtedy/yhat;
  RUN;

PROC SUMMARY DATA=edf;
  VAR gc;
  OUTPUT OUT=gindex
         SUM=sumgc;
  RUN;

DATA gindex(KEEP=ginicoef ng);
  SET gindex;
  ginicoef=sumgc - 1;                         *Gini coefficient;
  ng=&sizen;
  RUN;

PROC APPEND BASE = &file
            DATA = gindex;
     RUN;
%MEND estimate;


*** ---------------------------------------------------- ***;
***                 J A C K K N I F I N G                ***;
***  At k-th iteration (k=1,...,525) we remove observa-  ***;
***  tions from the k-th cluster in the sample and       ***;
***  adjust the weights of the remaining records in the  ***;
***  stratum by ng/(ng-1).                               ***;
*** ---------------------------------------------------- ***;
```

```
%MACRO jackit;
%DO i=1 %TO &numclust;
DATA _NULL_;
  SET clusters;
  IF _N_ = &i THEN DO;
    CALL SYMPUT('strat',strata);
    CALL SYMPUT('repnow',rep);
    CALL SYMPUT('smalln',ng);
    STOP;
  END;
  RUN;                 *information about cluster to be deleted;

%put &strat; %put &repnow; %put &smalln;

DATA sample(KEEP=wgt y wgtedy rep);
  SET ontario;
  IF rep=&repnow THEN jackind = 0;
  ELSE IF strata = &strat THEN jackind=&smalln/(&smalln-1);
  ELSE jackind=1;
  wgt = wt*jackind;
  wgtedy = wgt*y;
  RUN;                                         *weight adjustment;

%estimate(ginijack,&smalln);
%END;
%MEND jackit;



**************************************************************;
***                  M A I N   P R O G R A M             ***;
**************************************************************;
PROC SORT DATA=in.ont_rep
          OUT=clusters(KEEP=strata rep ng);
  BY rep;
  RUN;

PROC SORT DATA=in.ont_new
          OUT=ontario(KEEP=strata rep wt y);
  BY rep;
  RUN;

DATA sample(KEEP=strata wgt y wgtedy);
  SET ontario;
  wgt = wt;
  wgtedy = wgt * y;
  RUN;

%estimate(giniest,0);  *Gini index est. from original sample;

%jackit;

DATA giniest;
  SET giniest;
  RENAME ginicoef = gchat;
  DROP ng;
  RUN;

PROC SUMMARY DATA=ginijack;
  VAR ginicoef;
  OUTPUT OUT=meanjack(KEEP=gcmean)
         MEAN=gcmean;
  RUN;
```

```
* We compute two variance estimators using:                ;
*   1) gchat  = Gini index estimate from original sample    ;
*   2) gcmean = mean of the Gini index estimates from       ;
*                bootstrap samples                          ;

DATA varest
     in.jack2res;
  SET ginijack;
  IF _N_=1 THEN DO;
     SET giniest;
     SET meanjack;
  END;
  vjack1=(ng-1)/ng * ( (ginicoef-gchat) **2);
  vjack2=(ng-1)/ng * ( (ginicoef-gcmean) **2);
  RUN;

%printout(varest);

PROC MEANS DATA=varest SUM;
  VAR vjack1 vjack2;
  TITLE 'Jackknife Variance Estimators for the Gini Coefficient';
  RUN;

PROC DATASETS;
  DELETE giniest ginijack;
  RUN;
```

```
LIBNAME in 'f:\sasuser';
OPTIONS PAGESIZE=55;
%LET numtimes =525;

*** ----------------------------------------------------------- ***;
***                To print the results.                        ***;
*** ----------------------------------------------------------- ***;


%MACRO printout(file);
  proc print data=&file;
       title "&file";
  run;
%MEND printout;


*** ----------------------------------------------------------- ***;
***      This macro estimates the Gini coefficient              ***;
***                from the (sub)sample.                        ***;
*** ----------------------------------------------------------- ***;

%MACRO estimate(file);
PROC SORT DATA=sample;
  BY y;
  RUN;

PROC SUMMARY DATA=sample;
  VAR wgt wgtedy;
  OUTPUT OUT=est(DROP= _freq_ _type_)
         SUM=nhat yhat;
  RUN;           *population size and total income estimates;

DATA edf;
  SET sample;
  IF _N_=1 THEN SET est;
  topedf + wgt;
  fhat = (topedf/nhat);      *empirical distribution function;
  gc = 2 * fhat * wgtedy/yhat;
  RUN;

PROC SUMMARY DATA=edf;
  VAR gc;
  OUTPUT OUT=gindex
         SUM=sumgc;
  RUN;

DATA gindex(KEEP=ginicoef);
  SET gindex;
  ginicoef=sumgc - 1;                            *Gini coefficient;
  RUN;

PROC APPEND BASE = &file
            DATA = gindex;
     RUN;
%MEND estimate;


*** ----------------------------------------------------------- ***;
***  B O O T S T R A P   M E T H O D   with mh = nh-1           ***;
***  In the following macro we:                                 ***;
***      * draw a SRS with replacement of nh-1 clusters         ***;
***        from each stratum h                                  ***;
***      * calculate the bootstrap weights.                     ***;
*** ----------------------------------------------------------- ***;

%MACRO bootit;
```

```
DATA bootsamp(KEEP= strata bsind);
  SET info;
  ARRAY cnt(&maxng);
  DO i=1 TO ng;
     cnt(i)=0;
  END;
  DO i=1 TO (ng-1);
     ranum = ROUND( (ng-1) * RANUNI(TIME()) ) + 1;
     cnt(ranum)+1;
  END;
  DO i=1 TO ng;
     bsind=cnt(i);
     OUTPUT bootsamp;
  END;
  RUN;    *determines how many times each cluster is selected;


DATA sample(KEEP= rep bsind ng);
  MERGE clusters bootsamp;
  BY strata;
  RUN;

PROC SORT DATA=sample;
  BY rep;
  RUN;

DATA sample(KEEP=wgt y wgtedy);
  MERGE ontario sample;
  BY rep;
  wgt = ng * bsind * wt /(ng-1);              *bootstrap weights;
  wgtedy = wgt*y;
  RUN;

%estimate(giniboot);
%MEND bootit;



*** ----------------------------------------------------------- ***;
***    This macro is invoked to repeat the process of           ***;
***           selecting a bootstrap sample                      ***;
***             from the original sample.                       ***;
*** ----------------------------------------------------------- ***;

%MACRO justdoit;
  %DO j=1 %TO &numtimes;
      %bootit;
  %END;
%MEND justdoit;



*****************************************************************;
***                 M A I N   P R O G R A M                   ***;
*****************************************************************;
DATA clusters
     info(KEEP = strata ng);
  SET in.ont_rep;
  BY strata;
  OUTPUT clusters;
  IF FIRST.strata THEN OUTPUT info;
  RUN;                                      * cluster information;


PROC SUMMARY DATA=info;
```

```
  VAR ng;
  OUTPUT OUT=temp
         MAX=mng;   *determines the maximum sample size among;
  RUN;         *all strata (to fix array size in macro bootit);

DATA _NULL_;
  SET temp;
  CALL symput('maxng',mng);
  RUN;

PROC SORT DATA=in.ont_new
         OUT=ontario;
  BY rep;
  RUN;                                            *original data set;

DATA sample(KEEP=wgt y wgtedy);
  SET ontario;
  wgt = wt;
  wgtedy =  wgt * y;
  RUN;

%estimate(giniest);  *Gini index estimate from original sample;

%justdoit;

DATA giniest;
  SET giniest;
  RENAME ginicoef = gchat;
  RUN;

PROC SUMMARY DATA=giniboot;
  VAR ginicoef;
  OUTPUT OUT=meanboot(KEEP=gcmean)
         MEAN=gcmean;
  RUN;

* We compute two variance estimators using:                 ;
*    1) gchat  = Gini index estimate from original sample    ;
*    2) gcmean = mean of the Gini index estimates from       ;
*                bootstrap samples                           ;
DATA varest
     in.bootres;
  SET giniboot;
  IF _N_=1 THEN DO;
     SET giniest;
     SET meanboot;
  END;
  vboot1 = (ginicoef-gchat)**2/&numtimes;
  vboot2 = (ginicoef-gcmean)**2/&numtimes;
  RUN;


%printout(varest);

PROC MEANS DATA=varest SUM;
  VAR vboot1 vboot2;
  TITLE 'Bootstrap Variance Estimators for the Gini Coefficient';
  RUN;

PROC DATASETS;
  DELETE giniest giniboot;
  RUN;
```