

Variance Estimation of Relative Poverty Rate Estimate in Two-stage Sampling

*La Stima della Varianza della Stima dell'Incidenza di Povertà Relativa
in Disegni Campionari a Due Stadi*

Claudia De Vitiis, Loredana Di Consiglio, Stefano Falorsi
ISTAT - devitiis@istat.it, diconsig@istat.it, stfalors@istat.it

Riassunto: La stima dell'incidenza di povertà relativa è funzione implicita della stima della spesa nazionale pro-capite per consumi (linea di povertà relativa), in base alla quale una famiglia è definita povera o meno. Per la determinazione corretta della variabilità della stima è, quindi, necessario tener conto anche della variabilità campionaria connessa con la stima della linea. Il presente lavoro propone uno stimatore della varianza ottenuto in base ad una generalizzazione di un risultato di Zheng (2001).

Keywords: Decomposable poverty measure, calibration estimator; Woodruff transform

1. Introduction

The Italian Statistical Institute disseminates relative poverty estimates based on the Consumer Expenditure Sample Survey. The relative poverty rate for the d -th domain ($d=1, \dots, D$) is defined as

$$\bar{Y}_d^{\bar{Q}} = \frac{1}{M_d} \sum_{j=1}^{M_d} I(q_j^e; \bar{Q}) \quad (1)$$

where M_d is the number of the households in the d -th domain, $I(q_j^e; \bar{Q})$ is an indicator variable equal to one if the j -th household's equivalised expenditure, q_j^e , is lower than the *poverty line* and zero otherwise. The poverty line, $\bar{Q}$, is the average per capita expenditure at national level. The parameter (1) is currently estimated by a calibration estimator (Deville e Särndal, 1992), which can be expressed as

$${}_c\tilde{Y}_d^{\tilde{Q}} = \frac{1}{M_d} \sum_{j=1}^{m_d} I(q_j^e; \tilde{Q}) W_j \quad (2)$$

where $I(q_j^e; \tilde{Q})$ is the indicator variable defined on the basis of the sample estimates of the poverty line, $\tilde{Q}$. The final weights of the j -th household, W_j , are obtained by minimizing the distance with the initial weights and calibrating for the known totals, $\mathbf{X}$ (in particular, in this case, the auxiliary variables are age and sex).

An unbiased estimator for the sample variance of the estimator (2) has to take into account the sampling variability of the poverty line. In fact the relative poverty is not an

attribute of the households, but depends on the sample estimate of the line; therefore the usual methods of variance estimation, that consider the line as fixed, may be biased. In this work an appropriate method of variance estimation for poverty measures with relative poverty lines is developed for two-stage sampling design, which is the sampling design currently used for the Consumer Expenditure Survey.

2. Methodology for the variance estimation of a poverty measure

2.1 Decomposable poverty measures

An analytic expression for the estimation of the variance of the relative poverty estimate can be obtained on the basis of the result developed by Zheng (2001). This result can be applied for the class of *decomposable poverty measures* (DPM). A DPM $P(F; \omega)$ can be expressed as

$$P(F; \omega) = \int_0^\infty p(u, \omega) dF(u) \quad (3)$$

where u is a continuous variable with cumulative function $F(u)$ continuous on $(0, \infty)$; ω is a general *poverty line*, e.g. $\omega = c \mu_u$, with $c > 0$ and μ_u mean of the variable u ; $p(u, \omega)$ is a *poverty deprivation function*; it is a continuous function with respect to u and ω with the following properties: $p(u, \omega) = 0$ if $u > \omega$ $p_u(u, \omega) \equiv \frac{\delta p(u, \omega)}{\delta u} \leq 0$,

$\frac{\delta^2 p(u, \omega)}{\delta u^2} \geq 0$ on $(0, \omega)$; moreover $p_u(u, \omega)$ is bounded and its derivative

$p_\omega(u, \omega) \equiv \frac{\delta p(u, \omega)}{\delta \omega}$ exists and is uniformly continuous over $(0, \infty)$. Besides,

$a \equiv \int_0^\infty p_\omega(u, \omega) dF(u) = \frac{\delta P(F, \omega)}{\delta \omega}$ exists and is finite.

A consistent estimator of a DPM can be expressed as a sum of two terms: the first term is the estimator of the poverty rate with respect to the *true* poverty line, hence not affected by the sampling variability of the line; the second term is a quantity depending on the difference between the line estimator and the true value. In the case of the relative poverty rate used in ISTAT, the conditions on the distribution of u are sufficient for the proof of the result.

An extension of the above result applied to the calibration estimator in two-stage sampling design enables to write the estimator as

$${}_C \tilde{\tilde{Y}}_d^{\tilde{\tilde{Q}}} = {}_C \tilde{\tilde{Y}}_d^{\tilde{\tilde{Q}}} + (\tilde{\tilde{Q}} - \bar{Q}) \tilde{\tilde{\Gamma}}_d^{\tilde{\tilde{Q}}} + o(1/\sqrt{n}) \quad (4)$$

where ${}_C \tilde{\tilde{Y}}_d^{\tilde{\tilde{Q}}}$ is analogous to (2) with $\bar{Q}$ in place of $\tilde{\tilde{Q}}$, and $\tilde{\tilde{\Gamma}}_d^{\tilde{\tilde{Q}}}$ is a weighted average of the density functions of the variable u (in our case the equivalised expenditure of the household) evaluated at $\bar{Q}$. The density functions may be different among strata and

primary units. The weights are the final weights of the calibration estimator. This approximation of the estimator allows to obtain an approximately unbiased variance estimator which takes explicitly into account the variance of the poverty line estimate. It can be outlined that a “standard” variance estimation considering only the first component of right term of the (3) will be the more biased the higher is the relative importance of the variability of the estimator of the poverty line.

2.2 The proposed variance estimation method

The expression (3) is not a linear function of sample data. As a matter of fact it depends on the product $\tilde{Q}\tilde{F}_d^{\tilde{Q}}$ and on $\tilde{C}\tilde{Y}_d^{\tilde{Q}}$, which is itself a non linear function of the data. A consistent variance estimator may be obtained by means of Woodruff transform (Woodruff, 1971), which is based on a Taylor linearization of the non linear term. The linearized variable of the estimator (3) is

$$C_{,d}E_j^{\bar{Q}} \cong \frac{\xi_{1j}}{M_d}\delta_d + f_d(\bar{Q})\frac{\xi_{2j}}{N} \quad (5)$$

where δ_d is the indicator variable of the domain d , ξ_{1j} e ξ_{2j} are expressed as

$$\xi_{1j} = I(q_j^e; \bar{Q}) - \mathbf{X}_j' \boldsymbol{\beta} \quad \text{and} \quad \xi_{2j} = q_j - \mathbf{X}_j' \boldsymbol{\beta} \quad (6)$$

being $\mathbf{X}_j$ the B -dimensional vector of the auxiliary variables and $\boldsymbol{\beta}$ the unknown vector of regression coefficients in the population. For the case under study the density function in (5) has been assumed to be a log-normal density. The operative expression for the Woodruff transform can be obtained by replacing the unknown quantities in (6) with the corresponding calibration estimators and the parameter of the log-normal distribution with their sample estimates.

3. Empirical Analysis

3.1 The simulation set-up

To evaluate the performance of the proposed method, a Monte Carlo study has been carried out. In the simulation, 1000 replicates of the Consumer Expenditure Survey sample design have been drawn from a pseudo-population, defined according to the data on five years of the Consumer Expenditure Survey. For each replicate the poverty line and the poverty rate have been estimated; on the latter estimate the proposed variance method has been applied, as well as the method of variance estimation generally used for ISTAT surveys, which does not consider the variability of the line. For computational reasons the analysis has been conducted using the expansion estimator (the simplest case of calibration estimator) for the estimation of the relative poverty rate. The comparison between the two variance estimators has been carried out on the basis of the biases of the coefficients of variation. The CV has been chosen as target since it is the statistic usually disseminated together with survey estimates. The bias has been

obtained as difference between the average over the 1000 replicates of the sampling CV and the *true* CV, evaluated as CV over the 1000 estimates of the poverty rate estimates.

3.2 Analysis of the results

The following table 1 contains a synthesis of the results of the simulation study; for each region the poverty rate in the pseudo-population, the average CV for the method which assumes fixed line (standard) and for the proposed method (estimated line), the *true* CV, and the respective biases are reported. The outcome outlines that for the Italian regions the two methods produce nearly the same results, but the standard one underestimates the true CV in a greater number of cases and with higher discrepancy. In conclusion, the proposed method is preferable under a conservative approach.

Table 1. Simulation results

Region	Poverty rate of the target population	Average CV (standard) (a)	Average CV (est. line) (b)	“true” CV (c)	Bias (standard) (a) -(c)	Bias (est. line) (b) -(c)
Piemonte	6.1	15.5	16.5	17.6	-2.05	-1.09
Valle d’Aosta	8.9	17.4	18.0	17.5	-0.11	0.48
Lombardia	4.6	14.8	16.5	13.9	0.86	2.59
Trentino-Alto Adige	5.9	15.6	16.8	17.3	-1.70	-0.44
Veneto	5.3	20.5	21.3	22.3	-1.77	-1.00
Friuli Venezia Giulia	9.7	17.7	18.1	19.4	-1.74	-1.33
Liguria	7.7	17.7	18.3	17.8	-0.10	0.49
Emilia Romagna	4.4	18.8	20.4	19.5	-0.72	0.92
Toscana	6.3	17.8	18.5	19.1	-1.32	-0.58
Umbria	10.2	19.0	19.3	21.3	-2.31	-2.00
Marche	6.0	19.5	20.5	19.4	0.11	1.07
Lazio	9.4	12.7	12.8	12.8	-0.05	0.06
Abruzzo	15.7	13.5	13.6	14.6	-1.07	-0.93
Molise	24.5	16.3	16.3	16.3	-0.07	-0.03
Campania	23.4	8.0	7.8	8.1	-0.12	-0.34
Puglia	23.0	10.5	10.2	10.1	0.33	0.07
Basilicata	26.9	13.2	13.2	13.0	0.19	0.22
Calabria	27.4	11.0	10.9	11.3	-0.30	-0.39
Sicilia	25.8	8.8	8.7	8.4	0.37	0.22
Sardegna	20.4	11.3	11.4	11.8	-0.44	-0.42

References

- SÄRNDAL C-E., SWENSSON B., WRETMAN J., 1992, *Model assisted survey sampling*, New York: Spriger-Verlag.
- ZHENG B., 2001, Statistical inference for poverty measures with relative poverty lines, *Journal of Econometrics*, 101, 337-356.
- WOODRUFF R.S., 1971, A Simple method for approximating the variance of a complicate estimate, *Journal of the American Statistical Association*, 66, pp 411-414.