STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 10, Number 3, December 2009

ISSN 1234-7655

Address for correspondence
GUS, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tele/fax: 00 48 22 – 825 03 95

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 353—356

EDITOR'S NOTE AND ACKNOWLEDGEMENTS

Since this issue of the journal *Statistics in Transition new series* is the last one within the past 2009 year printing cycle, I would like to take this opportunity to express, also on behalf of the Editorial Office, our gratefulness to all the journal's patrons and supporters – with the Editorial Board chaired by Professor Józef Oleński, President of the Central Statistical Office, and Associate Editors – and to thank all our collaborators, authors and referees, as well as readers, who jointly contributed to successful continuation of the journal's mission. Altogether, 44 articles by 65 authors were printed in the journal during the past year (compared to 37 articles by 63 authors in the previous year).

We would especially warmly like to thank the people who acted as referees of the papers submitted for publication during the past year – their names are listed below, following the brief presentation of the contents of this issue.

***

Each of the five articles included in first part of this issue of the journal addresses some kind of the estimation problems. In ***Nonparametric Bootstrap Test for Autoregressive Additive Models*** Luca Bagnato and Antonio Punzo ask how to evaluate and decide whether an additive autoregressive model of a type that are commonly used to describe and simplify the behaviour of a nonlinear time series is really suitable to describe the observed data. Given that additivity represents a strong assumption and that there are few methods to test additivity, the authors propose a procedure for testing additivity in nonlinear time series analysis. It is based on Generalized Likelihood Ratio, Volterra expansion, and nonparametric conditional bootstrap, while its performance in terms of empirical size and power, and comparisons with other additivity tests are made with help of the Monte Carlo simulations.

The problem of defining improved variance estimators for the ordinary ratio estimator under the Midzuno-Sen sampling scheme is discussed by Jigna Patel and P. A. Patel in the ***On Non-Negative and Improved Variance Estimation for the Ratio Estimator under the Midzuno-Sen Sampling Scheme***. Starting from observation that, according to various studies, it is hard to single out the best and non-negative variance estimator in finite population, they propose a Monte Carlo comparison and suggest estimator that performs well, taking non-negative values (with probability 1).

The paper ***Estimation Of Life-Tables under Right-Censoring*** by Agnieszka Rossa deals with a class of non-parametric estimators of conditional probabilities of failure prior $x+y$ given survival to x under the random and observable right-

censorship model. For the proposed estimators based on a specific sequential sampling scheme some application in life-table analysis is presented.

In the paper ***Full Information Efficient Estimator of Finite Population Variance,*** devoted to estimating quadratic or higher order finite population functions, Manoj Kumar Srivastava, Namita Srivastava, and Housila P. Singh suggest an efficient design based full-information estimator of finite population variance. Together with providing the exact expression of the estimator variance and its relative efficiency, they also show that the proposed estimator is, according to the employed criteria of its performance in an empirical context, superior to its competitors.

Another aspect of estimation is discussed by Diwakar Shukla, Narendra Singh Thakur, Sharad Pathak and Dilip Singh Rajput in ***Estimation of Mean Under Imputation of Missing Data Using Factor-Type Estimator in Two-Phase Sampling***. As the problem of non-response is one of the most important in sample surveys, several imputation methods tend to compensate for the missing observations using the available ones. This paper presents some way to deal with the problem of unit non-response in the context of two-phase sampling. Two different strategies of such sampling – sub-sample and independent sample – are compared under imputed data setup, using Factor-Type (F-T) estimators. The results of simulation performed over multiple samples show that the first imputation strategy is found better than the second (but the second design is better than first).

In ***A Nonparametric Confidence Interval for At-Risk-of-Poverty-Rate,*** Wojciech Zieliński, refers to his earlier paper in which he proposed a distribution-free confidence interval for the *at-risk-of-poverty rate* (ARPR) that was defined as the percentage of population with income smaller than 60% of population median of the adult-equivalent disposable income. An example of application of the constructed confidence interval is given in this paper.

A set of five articles in the second part of this issue represent an array of topics. In ***Monitoring Workers' Remittances and Benefits in Uganda: The Statistical Issues,*** E.S.K. Muwanga-Zake presents the efforts that are underway by the Central Bank and Central Statistics Office in Uganda to improve the regulatory and monitoring environment in the country in order to provide credible information on remittances that are increasingly growing in terms of scope and importance. The approach used includes the enactment of a new law and regulations, improving administrative reporting and carrying out surveys in the major remitting countries, and in Uganda – some issues of collecting accurate and timely data are discussed.

The problem of poverty and income dynamics is analyzed by Natalia Nehrebecka in ***Temporal Aspects of Poverty in Poland Between 1997–2000 by Hazard Models*** using panel data from CHER (Consortium of Household Panels for European Socio-Economic Research). A tendency to persistent poverty along with generally low household income dynamics in that period was shown based

on the rate of exit from and entry to poverty while accounting for both observed and unobserved heterogeneity of individuals.

In ***Kurtosis of a Random Vector Special Types of Distibutions***, Katarzyna Budny and Jan Tatar attempt to generalize definition of kurtosis for the multidimensional case and prove its essential properties. The generalized characteristic applied in the single-dimension case has the same properties as kurtosis, that is known in the literature on single-dimensional random variables. The basis of conducted considerations is the definition of *the power of a vector in space with the scalar product.*

A software reliability growth model to study the combined effect of increasing error detection and decreasing error introduction rate under imperfect debugging is proposed by S. Chatterjee, L.N. Upadhyaya, J.B. Singh and S. Nigam in ***Combined Effect of Fault Detection and Fault Introduction Rate on Software Reliability Modelling***. The model is developed based on non homogeneous Poisson process (NHPP). It can be used to estimate and predict the reliability as well as the cost of a software product – some real life data has been used to validate the proposed model and to show its usefulness.

Some unavoidable drawbacks in data arrangements – such as overlapping among groups of observations (e.g., characterized by social, demographic or income sources categories) – that may create problem with decomposition of Gini and re-ranking indices to analyse potential redistribution effects and the unfairness of a tax systems is discussed by Achille Vernizzi in ***Applying the Hadamard Product to Decompose Gini, Concentration, Redistribution and Re-ranking Indexes***. Employing the so called matrix Hadamard product and showing how *within* group, *across* and *between* groups, and t*ransvariation* components can be written in matrix compact forms, author also demonstrates how the signs of Atkinson-Plotnick-Kakwani re-ranking index components can be analysed and split.

Włodzimierz OKRASA
Editor-in-Chief

ACKNOWLEDGEMENTS

TO REFEREES FOR 2009

The Editorial Board wishes to thank the following referees who have given their time and skills to the ***Statistics in Transition – new series*** during the period of the year 2009

- Aleksandra Baszczyńska, University of Łódź, Poland
- Jacek Białek, University of Łódź, Poland
- Katarzyna Bolonek- Lasoń, University of Łódź, Poland
- Denis Conniffe, University of Ireland - Maynooth, Ireland
- Czesław Domański, University of Łódź, Poland
- Alina Jędrzejczak, University of Łódź, Poland
- Cem Kedilar, Hacettepe University of Ankara, Turkey
- Jerzy Korzeniowski, University of Łódź, Poland
- Jerzy T.Kowaleski, University of Łódź, Poland
- Nicholas T.Longford, Pompeu Fabra University, Spain
- George Menexes, Aristotle University of Thessaloniki, Greece
- Andrzej Młodak, Statistical Office Poznań
- Amjad D.Al-Nasser, Yarmouk University of Irbid, Jordan
- Andrzej Ochocki, University of Cardinal Stefan Wyszyński in Warsaw, Poland
- Włodzimierz Okrasa, Central Statistical Office of Poland and University of Cardinal Stefan Wyszynski in Warsaw, Poland
- Walenty Ostasiewicz, Wrocław University of Economics, Poland
- Iannis Papadimitriou, Aristotle University of Thessaloniki, Greece
- Dorota Pekasiewicz, University of Łódź, Poland
- Waldemar Popiński, Central Statistical Office, Poland
- Agnieszka Rossa, University of Łódź, Poland
- Divakar Shukla, Dr.H.S.Gaur University of Sagar, India
- Meenakshi Srivastava, Dr.B.R. Ambedkar University (formerly Agra University),India
- Grażyna Trzpiot, Academy of Economics, Katowice, Poland
- Janusz Żądło, Academy of Economics, Katowice, Poland

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 357

SUBMISSION INFORMATION FOR AUTHORS

Statistics in Transition – new series (SiT) is an international journal published jointly by the Polish Statistical Association (PTS) and the Central Statistical Office of Poland, on a quarterly basis (during 1993–2006 it was issued twice and since 2006 three times a year). Also, it has extended its scope of interest beyond its originally primary focus on statistical issues pertinent to transition from centrally planned to a market-oriented economy through embracing questions related to systemic transformations of and within the national statistical systems, world-wide.

The SiT-*ns* seeks contributors that address the full range of problems involved in data production, data dissemination and utilization, providing international community of statisticians and users – including researchers, teachers, policy makers and the general public – with a platform for exchange of ideas and for sharing best practices in all areas of the development of statistics.

Accordingly, articles dealing with any topics of statistics and its advancement – as either a scientific domain (new research and data analysis methods) or as a domain of informational infrastructure of the economy, society and the state – are appropriate for *Statistics in Transition new series*.

Demonstration of the role played by statistical research and data in economic growth and social progress (both locally and globally), including better-informed decisions and greater participation of citizens, are of particular interest.

Each paper submitted by prospective authors are peer reviewed by internationally recognized experts, who are guided in their decisions about the publication by criteria of originality and overall quality, including its content and form, and of potential interest to readers (esp. professionals).

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl., followed by a hard copy addressed to
Prof. Wlodzimierz Okrasa,
GUS / Central Statistical Office
Al. Niepodległości 208, R. 287, 00-925 Warsaw, Poland

It is assumed, that the submitted manuscript has not been published previously and that it is not under review elsewhere. It should include an abstract (of not more than 1600 characters, including spaces). Inquiries concerning the submitted manuscript, its current status etc., should be directed to the Editor by email, address above, or w.okrasa@stat.gov.pl.

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: http://www.stat.gov.pl/pts/15_ENG_HTML.htm

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 359—370

NONPARAMETRIC BOOTSTRAP TEST FOR AUTOREGRESSIVE ADDITIVE MODELS

Luca Bagnato[1], Antonio Punzo[2]

ABSTRACT

Additive autoregressive models are commonly used to describe and simplify the behaviour of a nonlinear time series. When the additive structure is chosen, and the model estimated, it is important to evaluate if it is really suitable to describe the observed data since additivity represents a strong assumption. Although literature presents extensive developments on additive autoregressive models, few are the methods to test additivity which are generally applicable. In this paper a procedure for testing additivity in nonlinear time series analysis is provided. The method is based on: Generalized Likelihood Ratio, Volterra expansion and nonparametric conditional bootstrap (Jianqing and Qiwei, 2003). Investigation on performance (in terms of empirical size and power), and comparisons with other additivity tests proposed by Chen *et al.* (1995) are made recurring to Monte Carlo simulations.

Key words: Additive models; Generalized Likelihood Ratio; Volterra expansion; Bootstrap.

1. Introduction

Additive models face the trade-off between the misspecification problem and interpretability. It is not surprising if such tools are commonly used in the statistical applications to simplify data analysis. This wide family of models embodies a key simplifying assumption that, in some scale covariate, effects are separable. Furthermore, additive structures allow us to overcome the so-called problem of the "curse of dimensionality" (Bellman, 1961) by dimension reduction. In detail, additive autoregressive models, applied to time series analysis, assume that conditional expectation function of the dependent variable Y_t can be written as the sum of smooth terms in the lagged variables:

[1] Dipartimento di Metodi Quantitativi per le Scienze Economiche Aziendali Università degli Studi di Milano-Bicocca e-mail: luca.bagnato@unimib.it.
[2] Dipartimento di Impresa, Culture e Società Università di Catania e-mail: *antonio.punzo@unict.it*.

$$E\left[Y_t \mid \left(Y_{t-1},\ldots,Y_{t-p}\right)=\left(x_1,\ldots,x_p\right)\right]=m_1\left(x_1\right)+\cdots+m_p\left(x_p\right). \quad (1)$$

The advantages of using additive models for nonlinear autoregression are based on several reasons. First, they are easier to interpret because they do not involve interactions. Secondly, in many circumstances they can provide adequate approximations for many applications. Thirdly, under the additivity assumption, univariate smoothing techniques can be used directly in nonparametric estimation, resulting into a more comprehensive estimate. In addition, under additivity, the nonlinear contribution of each lagged variable to the response variable can be easily seen; it can be displayed graphically and in some cases can be interpreted. As regard the problem of computing the additive components, many algorithms like, for example, backfitting (Buja *et al.*, 1989) and marginal integration (Linton and Nielsen, 1995), have been provided and improved.

When the additive structure is chosen, and the model estimated, it is important to evaluate if it is really suitable to describe the observed data since, however, additivity represents a strong assumption. The question arises from the misspecification problem which leads to wrong conclusions and erroneous forecasting. The diagnostic checking stage is not merely to determine whether there is evidence of lack of fit but also to suggest ways in which the model may be modified when this is necessary. There are two basic methods for model validation: *overfitting* and *diagnostic checks* applied to the residuals. This paper focuses on the overfitting approach where the model is deliberately overparameterized in a way it is expected to be needed and in a manner such that the entertained model is obtained by setting certain parameters in the more general model at fixed values, usually zero (Box and Pierce, 1970). This traditional approach, mainly based on parametric assumptions, consists of using a large family of parametric models under the alternative hypothesis. The implicit assumption is that the large family of parametric models specifies the form of the true underlying dynamics correctly. However, this is not always warranted and leads naturally to a nonparametric alternative hypothesis. Naturally, the problems increase when also the null hypothesis is nonparametric (additive structure in this specific case).

Although extensive developments on nonparametric estimation techniques, there are few generally applicable methods for testing additivity (see Chen *et al.*, 1995). Our proposed procedure, that comes on top of the procedures proposed by Chen *et al*. (1995), is based on the Generalized Likelihood Ratio (GLR) which is a generally applicable tool for testing parametric hypotheses against nonparametric alternatives. An extension for using such a procedure to the nonparametric (additive) null hypothesis case will be made. Although the GLR method has been developed for independent data, the idea can be applied to time series data. In fact, it is expected that under mixing conditions, the results should also hold for the dependent data (Jianqing and Qiwei, 2003).

The paper is organized as follows: an introduction to the GLR method and nonparametric conditional bootstrap is presented in Section 2; the proposed

procedure for testing additivity is described in Section 3 and, in the last section, it is applied to additive and nonadditive models often used in the time series literature.

2. The generalized likelihood ratio

Before introducing GLR it is worth to remember the classic maximum likelihood ratio test which is generally applicable to most parametric hypothesis-testing procedure. The fundamental property that contributes to the success of the maximum likelihood ratio tests is that their asymptotic distributions under the null hypothesis are independent of nuisance parameters. This property was referred to as the "Wilks phenomenon" by Fan *et al*. (2001). Assuming such a property, one can determine the null distribution of the likelihood ratio statistic by using either the asymptotic distribution or the Monte Carlo simulation by setting nuisance parameters at some fitted values. The latter is also referred to as the parametric bootstrap.

The question arises naturally whether the maximum likelihood ratio test is still applicable to the problems with nonparametric models as alternative. First, nonparametric maximum likelihood estimators (MLE) usually do not exist. Even when they exist, they are hard to compute. To mitigate these difficulties, the maximum likelihood estimator under the alternative hypothesis can be replaced by any reasonable nonparametric estimator. This is the essence of the *generalized likelihood ratio*.

Let $\mathbf{f}$ be the vector of functions of main interest and $\boldsymbol{\eta}$ be the vector of nuisance parameters. Suppose that the logarithm of the likelihood of a given set of data is $\ell(\mathbf{f}, \boldsymbol{\eta})$. Given $\boldsymbol{\eta}$, a good nonparametric estimator $\hat{\mathbf{f}}_{\boldsymbol{\eta}}$ can be obtained. The nuisance parameters $\boldsymbol{\eta}$ can be estimated by the *profile likelihood* by maximizing $\ell(\hat{\mathbf{f}}_{\boldsymbol{\eta}}, \boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$, resulting in the profile likelihood estimator $\hat{\boldsymbol{\eta}}$. This gives the profile likelihood $\ell(\hat{\mathbf{f}}_{\hat{\boldsymbol{\eta}}}, \hat{\boldsymbol{\eta}})$, which is not the maximum likelihood since $\hat{\mathbf{f}}_{\hat{\boldsymbol{\eta}}}$ is not an MLE.

Now, suppose that we are interested in testing whether a parametric family $\mathbf{f}_\theta$ fits a given set of data. Formally, the null hypothesis is

$$H_0 : \mathbf{f} = \mathbf{f}_\theta, \qquad \theta \in \Theta, \tag{2}$$

and we use the nonparametric model $\mathbf{f}$ as alternative. Let $\hat{\theta}_0$ and $\hat{\boldsymbol{\eta}}_0$ be the maximum likelihood estimators under the null model (2) obtained by maximizing the function $\ell(\mathbf{f}_\theta, \boldsymbol{\eta})$. Then $\ell(\mathbf{f}_{\hat{\theta}_0}, \hat{\boldsymbol{\eta}}_0)$ is the maximum likelihood under the null

hypothesis. The GLR statistic simply compares the log-likelihood under the two competing classes of models:

$$T = \ell\left(\hat{\mathbf{f}}_{\hat{\boldsymbol{\eta}}}, \hat{\boldsymbol{\eta}}\right) - \ell\left(\mathbf{f}_{\hat{\theta}_0}, \hat{\boldsymbol{\eta}}_0\right). \tag{3}$$

Example 2.1 (Univariate nonparametric model) Let $\{(X_i, Y_i)\}_{i=1}^n$ be a sample from the nonparametric model:

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{4}$$

where $\{\varepsilon_i\}_{i=1}^n$ are a sequence of *i.i.d.* random variables from $N(0, \sigma^2)$. Consider testing the simple linear regression model:

$$H_0 : m(x) = \beta_0 + \beta_1 x \qquad H_1 : m(x) \neq \beta_0 + \beta_1 x, \tag{5}$$

with nonparametric alternative model (4). Then, the conditional log-likelihood function given $X_1, \ldots, X_n$ is

$$\ell(m, \sigma) = -n \ln\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [Y_i - m(X_i)]^2. \tag{6}$$

In this specific case results $\mathbf{f}=m$, $\boldsymbol{\eta}=\sigma$ and $\theta = (\beta_0, \beta_1)$. Ultimately θ identifies a particular model contained in the linear model class. For a given σ, let $\hat{m}(\cdot)$ be, for example, the local linear estimator based on the data $\{(X_i, Y_i)\}_{i=1}^n$, which is independent of σ. Substituting it into (6), the following profile likelihood is obtained:

$$\ell(\hat{m}, \sigma) = -n \ln\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2} \mathrm{RSS}_1, \tag{7}$$

where $\mathrm{RSS}_1 = \sum_{i=1}^n \left[Y_i - \hat{m}(X_i)\right]^2$. Maximizing (7) with respect to σ, it results that $\hat{\sigma}_1^2 = \mathrm{n}^{-1}\mathrm{RSS}_1$. Hence the profile likelihood is

$$\ell(\hat{m}, \hat{\sigma}) = -\frac{n}{2} \ln\left(\frac{\sqrt{2\pi}\mathrm{RSS}_1}{n}\right) - \frac{n}{2}. \tag{8}$$

Under H_0 the maximum likelihood estimator $\hat{\theta}_0 = \left(\hat{\beta}_0, \hat{\beta}_1\right)$ can be obtained. Then, the profile likelihood under the null hypothesis results

$$\ell\left(m_{\hat{\theta}_0}, \hat{\sigma}_0\right) = -\frac{n}{2} \ln\left(\frac{\sqrt{2\pi}\mathrm{RSS}_0}{n}\right) - \frac{n}{2}, \tag{9}$$

where $\mathrm{RSS}_0 = \sum_{i=1}^{n}\left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right]^2$. Using quantities RSS_0 and RSS_1 the statistic (3) can be obtained as follows:

$$T = \frac{n}{2}\ln\left(\frac{\mathrm{RSS}_0}{\mathrm{RSS}_1}\right). \tag{10}$$

This is a GLR test statistic.

As with parametric inference, the GLR test does not have to use the true likelihood. For example, the test statistic T in Example 2.1 applies to problem (5) whether ε_i is normally distributed or not. The normality assumption is simply used to motivate the procedure. Similarly, the GLR statistic does not have to require the MLE under the null hypothesis.

Such considerations suggested using the GLR method for testing additivity. In fact, neither MLE nor error distribution assumptions are made. In the next section it will be shown how to provide the alternative structure to compare with the additive one. In order to do this, the Volterra expansion will be used. Furthermore, to utilize the GLR statistic, the distribution under the null hypothesis needs to be provided. The question arises naturally whether the asymptotic null distribution depends on the nuisance parameter under the null hypothesis, namely, whether the Wilks phenomenon continues to hold for the GLR test. For a number of models and a number of hypotheses, studied by Fan *et al.* (2001), it has been shown that the Wilks type of results continue to hold. Such authors are not able to show that Wilks type of results hold for all problems, but their results indicate that such a phenomenon holds with sufficient generality.

3. A new test for autoregressive additivity

The Additive AutoRegressive (AAR) model is defined as follows:

$$Y_t = m\left(Y_{t-1},\ldots,Y_{t-p}\right) + \varepsilon_t, \tag{11}$$

with

$$m\left(Y_{t-1},\ldots,Y_{t-p}\right) = c_m + m_1\left(Y_{t-1}\right) + \cdots + m_p\left(Y_{t-p}\right), \tag{12}$$

where c_m is a constant, m_j, j=1,...,p, are univariate unknown functions, and the ε_t are independent and identically distributed (*i.i.d.*) with mean 0 and variance σ^2. Furthermore, ε_t is assumed independent of $\{Y_{t-k}\}_{k\geq 1}$ for any t. To ensure identifiability of the additive component functions m_j, it is assumed $E\left[m_j\left(Y_{t-j}\right)\right] = 0$ for all j=1,…,p. The intercept $c_m = E\left(Y_t\right)$ is typically estimated by

$\bar{Y} = \sum_{i=1}^{n} Y_i / n$. Technically, the *i.i.d.* assumption of the errors may be weakened when other theoretical explorations are made. However, as well known, a white noise process is no longer a pertinent building block for nonlinear models, as it is important to look for measures beyond the second moments to characterize the nonlinear dependence structure.

Once estimated the (11) under the additivity assumption (12), the obvious question is whether such a model is appropriate to describe the underlying structure. In order to deal with this model validation, the additive (null) hypothesis

$$H_0 : m\left(Y_{t-1}, \ldots, Y_{t-p}\right) = c_m + \sum_{j=1}^{p} m_j\left(Y_{t-j}\right), \tag{13}$$

will be compared with the (alternative) hypothesis that the conditional mean has one more general autoregressive structure, say

$$H_1 : m\left(Y_{t-1}, \ldots, Y_{t-p}\right). \tag{14}$$

The comparison will be made through the GLR statistic which is utilized here in the most general case, that is, when also the null hypothesis is nonparametric. Obviously, the model validation procedure (overfitting technique) needs to estimate a very general model under the alternative hypothesis. The main problem relates to the definition of a model more general than the additive one but not affected by the "curse of dimensionality". The Volterra expansion allows us to overcome such a difficulty (Chen *et al*., 1995). In particular, through the Volterra expansion, an autoregressive model can be rewritten in the following way

$$\begin{aligned} Y_t &= m\left(Y_{t-1}, \ldots, Y_{t-p}\right) + \varepsilon_t \\ &= \sum_{u=1}^{p} \phi_u Y_{t-u} + \sum_{u \le v}^{p} \phi_{uv} Y_{t-u} Y_{t-v} + \sum_{u \le v \le w}^{p} \phi_{uvw} Y_{t-u} Y_{t-v} Y_{t-w} + \cdots + \varepsilon_t \\ &= c_m + \sum_{u=1}^{p} m_u\left(Y_{t-u}\right) + \sum_{u \le v}^{p} \phi_{uv} Y_{t-u} Y_{t-v} + \sum_{u \le v \le w}^{p} \phi_{uvw} Y_{t-u} Y_{t-v} Y_{t-w} + \cdots + \varepsilon_t \end{aligned} \tag{15}$$

where

$$m_u\left(Y_{t-u}\right) = \phi_u Y_{t-u} + \phi_{uu} Y_{t-u}^2 + \phi_{uuu} Y_{t-u}^3 + \cdots.$$

Obviously, the case $u = v = w$ is excluded from the summation of the third-order term in (15). It is clear from expression (15) that if the model is additive, then all the coefficients of the higher-order terms in the equation should be zero. Furthermore, defining

$$m_{uv}\left(Y_{t-u} Y_{t-v}\right) = \phi_{uv} Y_{t-u} Y_{t-v} + \phi_{uuvv}\left(Y_{t-u} Y_{t-v}\right)^2 + \phi_{uuuvvv}\left(Y_{t-u} Y_{t-v}\right)^3 + \cdots,$$

the expression (15) can be rewritten as follows:

$$Y_t = c_m + \sum_{u=1}^{p} m_u\left(Y_{t-u}\right) + \sum_{u\le v}^{p} m_{uv}\left(Y_{t-u}Y_{t-v}\right) + \sum_{u\le v\le w}^{p} m_{uvw}\left(Y_{t-u}Y_{t-v}Y_{t-w}\right) + \cdots + \varepsilon_t. \quad (16)$$

Such a result suggests us a simple approximate form to apply for the alternative hypothesis (14). In particular if, for example, the (16) is truncated to the second summation, the hypothesis (14) can be assumed as:

$$H_1 : m\left(Y_{t-1},\ldots,Y_{t-p}\right) = c_m + \sum_{u=1}^{p} m_u\left(Y_{t-u}\right) + \sum_{u\le v}^{p} m_{uv}\left(Y_{t-u}Y_{t-v}\right). \quad (17)$$

This formulation allows to estimate a model that contains $p + \binom{p}{2}$ univariate additive functions. Although such a test is limited to the first-order cross-product terms, it should have acceptable power against a large class of nonadditive models. In the next Subsection the method for finding the distribution of the GLR statistic will be provided.

3.1. The conditional bootstrap test

The proposed step-procedure for testing additivity, that will be called from now on as Conditional Bootstrap Test, is described in what follows.

I. The two models, respectively under the null and the alternative hypotheses (13) and (14), are estimated and the GLR statistic T is calculated.

II. The *nonparametric conditional bootstrap* (Jianqing and Qiwei, 2003) is applied:

1. Generate the bootstrap residuals $\left\{\varepsilon_t^*\right\}$ of the empirical distribution of the centred residuals $\left\{\hat{\varepsilon}_t - \bar{\hat{\varepsilon}}\right\}$ from the alternative model, where $\bar{\hat{\varepsilon}}$ is the average of $\left\{\hat{\varepsilon}_t\right\}$. An assumption about the distribution error is made; thus, for example, a kernel density estimation can be applied. Construct the bootstrap sample: $Y_{t,1}^* = Y_{t-1},\ldots,Y_{t,p}^* = Y_{t-p}$ and

$$Y_t^* = c_m + \hat{m}_1\left(Y_{t-1}\right) + \cdots + \hat{m}_p\left(Y_{t-p}\right) + \hat{\varepsilon}_t^*$$

for t=p,…,n.

2. Estimate the additive and the alternative model based on the bootstrap sample:

$$\left\{\left(Y_{t,1}^*,\ldots,Y_{t,p}^*,Y_t^*\right)\right\}_{t=p}^{n}.$$

Calculate the GLR statistic

$$T = \frac{n-p+1}{2} \ln\left(\frac{\mathrm{RSS}_0^*}{\mathrm{RSS}_1^*}\right) \approx \frac{n-p+1}{2} \frac{\mathrm{RSS}_0^* - \mathrm{RSS}_1^*}{\mathrm{RSS}_1^*}.$$

3. Repeat the above two steps B times, and use the empirical distribution of $\{T^*\}$ as an approximation to the distribution of the GLR statistic T under H_0.

III. The estimated p-value of the test is the percentage of $\{T^*\}$ greater than the statistic T provided at point I.

4. Simulation study

In order to evaluate the performance of the proposed test, in terms of its empirical size and power, a simulation study is performed. Obtained results are also compared with size and power of three different additivity tests proposed by Chen *et al*. (1995): the *conditional mean test*, the *Lagrange multiplier test*, and the *permutation test*. The first uses the local conditional mean estimator of Truong (1993) and employs a procedure similar to the analysis of variance. The second applies the alternating conditional expectation (ACE) algorithm of Breiman and Friedman (1985) to fit an additive model to the data; additivity is then tested by means of a Lagrange multiplier type test. The third procedure uses the ACE algorithm as well, but it fits permuted residuals to some cross-product terms of the explanatory variables in order to obtain a reference distribution for the test statistic.

In order to make the above-mentioned comparison easier, two subsets of models considered in Chen *et al*. (1995) are used here – one for size considerations and the other for power analysis. The first set consists of the following two additive models:

$$Y_t = 0.8Y_{t-1} - 0.3Y_{t-2} + \varepsilon_t, \tag{18}$$

$$Y_t = 0.5Y_{t-1} - \sin(Y_{t-2}) + \varepsilon_t. \tag{19}$$

These models are used to study the behaviour of the conditional bootstrap test under the null hypothesis of additivity. They represent time series models commonly used in univariate analysis. The linear *AR* (2) model in (18) is chosen to ensure that the proposed test works well for this simple case, while the slightly more complicated model (19), containing a trigonometric sine function at lag 2, is often used in the time series literature to describe periodic series (Lewis and Ray, 1993).

The second set consists of the following two nonadditive models:

$$Y_t = 2\exp\left(-0.1Y_{t-1}^2\right)Y_{t-1} - \exp\left(-0.1Y_{t-1}^2\right)Y_{t-2} + \varepsilon_t, \tag{20}$$

$$Y_t = Y_{t-1}\sin\left(Y_{t-2}\right) + \varepsilon_t. \tag{21}$$

These models are used to study the power of the conditional bootstrap test. In particular, model (21) is a functional-coefficient *AR*(1) with a sine function of lag 2 (Chen and Tsay, 1993).

Like Chen *et al*. (1995), for each of the models in (18)-(21), we have applied the proposed test to 300 realizations, each with 300 observations. The sample size of 300 or larger is common in nonlinear time series analysis, especially when using nonparametric methods; indeed, it is often difficult to obtain a reliable estimate of the high-dimensional surface when the sample size is small. According to Chen *et al*. (1995), the innovations ε_t are independent $N(0,1)$. In applying the conditional bootstrap test, in order to make faster the procedure, we use a value B=100 and cross-product terms of degree one in the Volterra expansion. For details on the simulation factors used for the other three tests, see Chen *et al*. (1995).

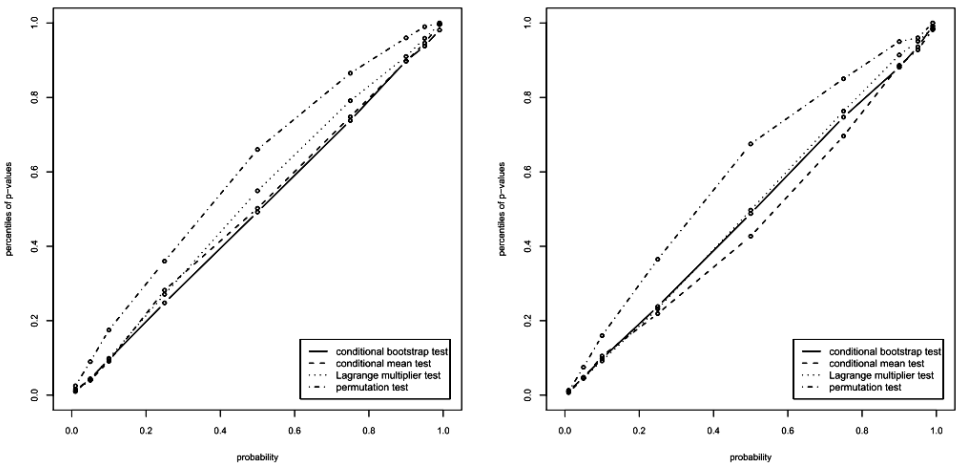
Table 1 shows the (simulated) empirical distribution function of the *p*-values for models (18) and (19), and for each of the four considered tests.

Table 1. Percentiles of *p*-values of the nonparametric bootstrap test, under the null hypothesis, in comparison with the tests proposed by Chen *et al*. (1995).

Probability	Conditional bootstrap test		Conditional mean test		Lagrange multiplier test		Permutation test	
	Model (18)	Model (19)	Model (18)	Model (19)	Model (18)	Model (19)	Model (18)	Model (19)
0.01	0.013	0.010	0.016	0.007	0.010	0.013	0.025	0.010
0.05	0.043	0.047	0.040	0.045	0.041	0.046	0.090	0.075
0.10	0.097	0.097	0.091	0.105	0.099	0.092	0.175	0.160
0.25	0.247	0.237	0.282	0.219	0.271	0.232	0.360	0.365
0.50	0.493	0.487	0.507	0.427	0.549	0.497	0.660	0.675
0.75	0.737	0.747	0.748	0.696	0.791	0.763	0.865	0.850
0.90	0.887	0.880	0.897	0.886	0.910	0.914	0.960	0.950
0.95	0.937	0.933	0.946	0.928	0.959	0.951	0.990	0.960
0.99	0.983	0.993	0.996	0.982	0.996	0.986	1.000	1.000

The graphical counterpart of Table 1 is also given in Figure 1; here, models (18) and (19) are separately considered in Figure 1(a) and Figure 1(b), respectively.

Figure 1. Percentiles of *p*-values of the nonparametric bootstrap test, under the null hypothesis, in comparison with the tests proposed by Chen *et al*. (1995).



(a) Model (18) (b) Model (19)

Figure 1 is useful because it simplifies the comparative analysis of performance; indeed, as expected, it is easy to see that the empirical distribution function of the *p*-values, for both models (18) and (19), is always close to the uniform distribution in the unit interval [0,1], which means that the nominal size equals the empirical one. This similarity is strikingly clear for the conditional bootstrap test. In these terms the conditional bootstrap test appears to be one of the best while the permutation test the worst.

Table 2 shows the percentages of rejection by the several tests under different significance levels for models (20) and (21).

Table 2. Percentages of rejection by the conditional bootstrap test, under the alternative hypothesis, in comparison with the tests proposed by Chen *et al*. (1995).

Probability	Conditional bootstrap test		Conditional mean test		Lagrange multiplier test		Permutation test	
	Model (20)	Model (21)	Model (20)	Model (21)	Model (20)	Model (21)	Model (20)	Model (21)
0.010	0.207	1.000	0.403	1.000	1.000	0.990	0.090	1.000
0.050	0.127	1.000	0.267	1.000	0.997	0.990	0.037	1.000
0.100	0.083	1.000	0.117	1.000	0.977	0.967	0.007	1.000

It can be easily noted how all the tests have a good power against the functional-coefficient autoregressive model (21). Unfortunately, above all the permutation test and the conditional bootstrap test do not have a good power against the exponential model in (20). This poor performance of the two tests is

understandable because the nonadditivity of the alternative model is in higher-order terms and only the simple cross-product term $Y_{t-1}Y_{t-2}$ has been used in the tests. In practice, it may be helpful to employ several cross-product terms in using these tests.

5. Concluding remarks

In this paper a new procedure for testing additivity, which we defined conditional bootstrap test, has been proposed by means of Generalized Likelihood Ratio, Volterra expansion, and nonparametric conditional bootstrap. This procedure does not require any strong assumption on innovations. A simulated analysis of performance, in terms of empirical size and power, suggests that the conditional bootstrap test is generally reliable under the null hypothesis, even if it may result in low power when the true alternative model is nonadditive in higher-order terms and the cross-product terms considered in the Volterra expansion are small in number. This drawback, also shared by the permutation test proposed in Chen *et al*. (1995), could be however overcome by employing several cross-product terms in the above said expansion (naturally, to the detriment of the computing time).

The paper hints at some further issues; for example, a sort of "rule of thumb" in order to select the suitable number of cross-product terms in the Volterra expansion could be interesting. In detail, future works will be directed to implementing information criterion techniques to select the "correct" model inside a wide family of models resulting from the Volterra expansion.

REFERENCES

BELLMAN, R., 1961. *Adaptive Control Process*. Princeton: Princeton University Press.

BOX, G. E. P. and PIERCE, D. A., 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332), pp. 1509–1526.

BREIMAN, L. and FRIEDMAN, J. H., 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391), pp. 580–619.

BUJA, A., HASTIE, T., and TIBSHIRANI, R., 1989. Linear smoothers and additive models. *The Annals of Statistics*, 17(2), pp. 453–510.

CHEN, R. and TSAY, R. S., 1993. Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88(421), pp. 298-308.

CHEN, R., LIU, J. S., and TSAY, R. S., 1995. Additivity tests for nonlinear autoregression. *Biometrika*, 82(2), pp. 369–383.

FAN, J., ZHANG, C., and ZHANG, J., 2001. Generalized likelihood ratio statistics and Wilks phenomenon. *The Annals of Statistics*, 29(1), pp. 153–193.

HASTIE, T. and TIBSHIRANI, R. J., 1990. *Generalised additive models*. London: Chapman and Hall.

JIANQING, F. and QIWEI, Y., 2003. *Nonlinear time series: nonparametric and parametric methods*. New York: Springer.

LEWIS, P. and RAY, B., 1993. Nonlinear modelling of multivariate and categorical time series using multivariate adaptive regression splines. In H. Tong, ed. *Dimension Estimation and Models*. Singapore: World Scientific, pp. 136-169.

LINTON, O. and NIELSEN, J. P., 1995. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, pp. 82(1), 93–100.

TRUONG, Y., 1993. A nonparametric framework for time series analysis. *New Directions in Time Series Analysis*, pp. 371–386. New York: Springer.

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 371—385

ON NON-NEGATIVE AND IMPROVED VARIANCE ESTIMATION FOR THE RATIO ESTIMATOR UNDER THE MIDZUNO-SEN SAMPLING SCHEME

Jigna Patel[1] and P. A. Patel[2]

ABSTRACT

Various studies on variance estimation showed that it is hard to single out a best and non-negative variance estimator in finite population. This paper attempts to find improved variance estimators for the ordinary ratio estimator under the Midzuno-Sen sampling scheme. A Monte Carlo comparison has been carried out. The suggested estimator has performed well and has taken non-negative values with probability 1.

Key words: Model-based estimation, Monte Carlo Simulation, Ratio estimator, Variance estimation

1. Introduction

A major problem in a large complex survey is the selection of a variance estimation procedure. Most of the basic theory developed in the standard sampling texts deals with variance estimation for linear estimators, and, therefore, is not applicable to complex survey involving ratio and composite estimation procedures. However, these variance estimators are not free from weaknesses. Also, these estimators have not incorporated the auxiliary information.

The regression and ratio estimators are widely used in survey practice. In the past more attention has been given to the ratio estimator because of its computational ease and applicability for general sampling designs. Many variance estimators for the ratio estimator have been proposed and compared. The most of them are design-based, see, e.g., Rao (1969), Rao and Beegle (1967), Rao (1968), Rao and Rao (1971), Rao and Kuzik (1974), Royall and Eberhardt (1975), Royall and Cumberland (1978, 1981), Krewski and Chakrabarty (1981) and Wu (1982)

[1] Department of Statistics, Sardar Patel University, Vallabh Vidhyanagar-388120, Gujarat, India, e-mail: jigna.stat@gmail.com.
[2] Department of Statistics, Sardar Patel University, Vallabh Vidhyanagar-388120, Gujarat, India, e-mail: patelpraful_a@indiatimes.com.

among others. The theoretical comparisons of various variance estimators have been made by assuming that the variables x and y, in population, satisfy some linear regression models. Several authors studied the estimation of model-variance of ratio predictor under various models, see, e.g., Mukhopadhyay (1996) and references cited there. The issue of variance estimation, for these widely used estimators, has not been finally resolved. Studies by Wu (1982), Wu & Deng (1983) and Deng & Wu (1987) show that it is hard to single out a 'best' variance estimator; optimality depends on the performance criterion use. This article deals with the estimation of variance of the ratio estimator under the Midzuno-Sen sampling scheme when auxiliary information is available.

In section 2, we suggest variance estimator for the ratio estimator. Section 3 presents the results of a Monte Carlo study that compares the suggested estimator with the standard and available estimators. Finally, our conclusions are given in Section 4.

Let $U = \{1,...,i,...,N\}$ be a finite population and let y_i and x_i be the values of the study variable y and an auxiliary variable x for the i th population unit, $i = 1,...,N$. If $A \subseteq U$, we write Σ_A for $\Sigma_{i\in A}$ and $\Sigma\Sigma_A$ for $\Sigma\Sigma_{i\neq j\in A}$. We seek to estimate the variance of the ratio estimator $\hat{Y}_R = \dfrac{\sum_s y_i}{\sum_s p_i}$ of the population total $Y = \sum_U y_i$, where $p_i = \dfrac{x_i}{X}$ with $X = \Sigma_U X_i$, ($i = 1,2,...,N$), on the basis of a sample s of fixed-size n drawn according to a sampling design $p(s)$ with positive inclusion probabilities $\pi_i = P(i \in s)$ and $\pi_{ij} = P(i, j \in s)$ for every i and j. $E_P(\cdot)$ and $V_P(\cdot)$ denote the design-expectation and design-variance of an estimator. The variance of $\hat{Y}_R$, suggested by Midzuno (1950), is given by

$$V_{P1}\left(\hat{Y}_R\right) = \sum_U \Lambda(s,ii) y_i^2 + \sum\sum_U \Lambda(s,ij) y_i y_j$$

where, for $i, j = 1,...,N,\ i \neq j$,

$$\Lambda(s,ij) = \left(\frac{1}{M_1}\sum_{s\ni i}\frac{1}{P_s} - 1\right) \quad \text{if } i = j$$

$$= \left(\frac{1}{M_1}\sum_{s\ni i,j}\frac{1}{P_s} - 1\right) \text{ if } i \neq j$$

and $P_s = \sum_s p_i$, $M_1 = \binom{N-1}{n-1}$.

Rao (1972) proposed

$$v_1(\hat{Y}_R) = \sum_s \Lambda(s,ii)\frac{y_i^2}{\pi_i} + \sum\sum_s \Lambda(s,ij)\frac{y_i y_j}{\pi_{ij}}$$

as an unbiased estimator of $V_{R1}(\hat{Y}_R)$. Further he stated that a sufficient condition for $v_1(\hat{Y}_R)$ to be positive is that $\Lambda(s,ij) \geq 0$ for all i, j. Chaudhuri (1975) assumes that the characteristic y can take negative values in which case the above sufficient condition is not valid. Therefore, he suggested alternative unbiased estimators by writing $V_{P1}\left(\hat{Y}_R\right)$ into different forms as

$$V_{P2}(\hat{Y}_R) = X^2\left[\sum_{i=1}^{N} t_i^2(nT_i - N) + \sum\sum_{i<j}^{N}(1-T_{ij})(t_i - t_j)^2\right],$$

$$V_{P3}(\hat{Y}_R) = \left(\frac{1}{2}\sum\sum_U Q_{ij}\right)X^2,$$

$$V_{P4}(\hat{Y}_R) = \left(\frac{1}{2}\sum\sum_U R_{ij}\right)X^2$$

where $T_i = \Lambda(s,ii)+1$, $T_{ij} = \Lambda(s,ij)+1$, $t_i = \frac{y_i}{X}$,

$$Q_{ij} = \left(\frac{T_i - 1}{N-1}t_i^2 + 2(T_{ij}-1)t_i t_j + \frac{T_j - 1}{N-1}t_j^2\right)$$

and

$$R_{ij} = \left(\frac{T_{ij}}{n-1} - \frac{1}{N-1}\right)t_i^2 + 2(T_{ij}-1)t_i t_j + \left(\frac{T_{ij}}{n-1} - \frac{1}{N-1}\right)t_j^2$$

His suggested variance estimators are then

$$v_2(\hat{Y}_R) = X^2\left[\sum_s \frac{t_i^2(nT_i - N)}{\pi_i} + \sum\sum_{i<j\in s}\frac{(1-T_{ij})(t_i - t_j)^2}{\pi_{ij}}\right]$$

$$v_3(\hat{Y}_R) = \left(\frac{1}{2}\sum\sum_s \frac{Q_{ij}}{\pi_{ij}}\right) X^2$$

$$v_4(\hat{Y}_R) = \left(\frac{1}{2}\sum\sum_s \frac{R_{ij}}{\pi_{ij}}\right) X^2$$

Chaudhuri (1976, 1981) and Chaudhuri and Arnab (1981) studied the problem of non-negative variance estimation and proposed sufficient conditions for non-negativity for their estimators

$$v_{11} = \sum \sum_{i<j\in s} C_{ij} x_i x_j \frac{(N-1)X}{(n-1)x_s}\left\{1 - \frac{1}{M_1}\sum_{s\ni i,j}\frac{X}{x_s}\right\}$$

$$v_{12} = v_{32} = \sum \sum_{i<j\in s} C_{ij} x_i x_j \left\{\frac{M_1 X}{M_2 x_s} - \frac{1}{\pi_{ij} M_1}\sum_{s\ni i,j}\frac{X}{x_s}\right\}$$

$$v_{13} = \sum \sum_{i<j\in s} C_{ij} x_i x_j \left\{\frac{X}{x_s}\left(\frac{N-1}{n-1} - \frac{1}{M_2}\sum_{s\ni i,j}\frac{X}{x_s}\right)\right\}$$

$$v_{23} = \sum \sum_{i<j\in s} C_{ij} x_i x_j \left\{\frac{1}{\pi_{ij}} - \frac{X}{M_2 x_s}\sum_{s\ni i,j}\frac{X}{x_s}\right\}$$

where $C_{ij} = \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2$, $P(s/i) = \frac{1}{M_1 p(s)}$, $M_i = \binom{N-i}{n-i}$ and $x_s = \sum_s x_i$

However, it should be noted that none of the estimators is of the form given by Rao (1979). Moreover, these sufficient conditions cannot be satisfied either at all or except in trivial case when size measures are equal.

Rao and Vijayan (1977) suggested

$$v_{10} = v_{30} = \sum \sum_{i<j\in s} C_{ij} x_i x_j \left\{\frac{1}{M_2 p(s)} - \frac{P(s/i)P(s/j)}{(p(s))^2}\right\}$$

and

$$v_{22} = \sum \sum_{i<j\in s} C_{ij} \frac{x_i x_j}{\Pi_{ij}}\left\{1 - \frac{1}{M_1}\sum_{s\ni i,j}\frac{X}{x_s}\right\},$$

which satisfy Rao's (1977) condition for the necessary form of non-negative quadratic unbiased estimators. Tracy and Mukhopadhyay (1994), also, addressed the same problem namely non-negative unbiased variance estimation for the Midzuno strategy. Vijayan et al. (1995) extended the results of Rao and Vijayan (1977) for non-negative unbiased variance estimation of quadratic forms and, in particular, discussed various estimators of variance of estimators of the population total.

2. The suggested estimator

Valliant et al. (2000) has mentioned that relatively little research has been directed toward deriving variance estimators that explicitly incorporate both design-based and model-based thinking. In this section model-design-based estimator is suggested.

Assume that $y_1, y_2, ..., y_N$ are exchangeable random variables having joint distribution ξ (See, Cassel et al., 1977, p.102) and that the relationship between y_i and x_i is

$$\left.\begin{aligned} & y_i = \beta x_i + \varepsilon_i \\ & \mathrm{E}_\xi(\varepsilon_i / x_i) = 0 \\ & V_\xi(\varepsilon_i / x_i) = \sigma^2 x_i^2 \\ & C_\xi(\varepsilon_i, \varepsilon_j / x_i, x_j) = \rho\sigma^2 x_i x_j \quad (i \neq j) \end{aligned}\right\} \tag{1}$$

where $\beta, \sigma^2 > 0$ and $\rho \in \left(-\frac{1}{N-1}, 1\right)$ are the parameters. Here $E_\xi(\cdot), \mathrm{V}_\xi(\cdot)$ and $\mathrm{C}_\xi(\cdot)$ denote ξ-expectation, ξ-variance and ξ-covariance, respectively. The parameter ρ in the model (1) was shown to be a quite generally redundant. (See, e.g., Brewer and Tam, 1990, and Patel and Shah, 1999). Henceforth, we assume $\rho = 0$.We try to make as efficient use of auxiliary information as possible through model. To find an optimal strategy (a combination of sampling design and estimator) direct minimization of design-variance, $V_p(v)$ is impossible, but given a model ξ, we can try to minimize the anticipated variance (See, Isaki and Fuller, 1982)

$$AV(v - V_{RY}) = E_\xi E_p[(v - V_{RY})^2] - [E_\xi E_p(v - V_{RY})]^2$$

Clearly, when v is p-unbiased, the $AV(\cdot)$ becomes

$$AV(v - V_{RY}) = E_\xi E_p[(v - V_{RY})^2]$$

Here, the optimality is interpreted in sense of minimizing $E_\xi E_p[(v-V_{RY})^2]$ subject to $E_p(v)=V_{RY}$, $\forall\, Y\in R_N$.

Under model (1), we have

$$E_\xi(V_{RY})=(\sigma^2+\beta^2)\sum_{i\in U}\Lambda(s,ii)x_i^2+\beta^2\sum\sum_{i\neq j\in U}\Lambda(s,ij)x_ix_j$$

We must find v to minimize

$$E_\xi E_p[(v-V_{RY})^2]=E_p\{V_\xi(v)\}+E_p\{B_\xi(v)\}^2$$

subject to $E_\xi E_p(v)=E_\xi(V_{RY})$, where $B_\xi(v)=E_\xi(v-V_{RY})$.

Proceeding on the same line as in lemma 4.5 of Cassel et al. (1977) (See, also, Patel and Shah, 1999), it is easy to show that

$$\sigma^2+\beta^2\hat{=}\sum_{i\in s}\frac{y_i^2}{nx_i^2}\text{ and }\beta^2\hat{=}\sum\sum_{i\neq j\in s}\frac{y_iy_j}{n(n-1)x_ix_j}\qquad(2)$$

are, respectively, unique $p\xi$-unbiased predictors of $\sigma^2+\beta^2$ and β^2. As p-unbiasedness implies $p\xi$-unbiasedness, the optimal p-unbiased predictor of V_{RY} is found to be

$$v_{OPT}=\sum_{i\in s}\Lambda(s,ii)\frac{y_i^2}{\pi_{i0}}+\sum\sum_{i\neq j\in s}\Lambda(s,ij)\frac{y_iy_j}{\pi_{ij0}}\qquad(3)$$

where,

$$\pi_{i0}=\frac{n\Lambda(s,ii)x_i^2}{\sum_{i\in s}\Lambda(s,ii)x_i^2}\quad\text{and}\quad\pi_{ij0}=\frac{n(n-1)\Lambda(s,ij)x_ix_j}{\sum\sum_{i\neq j\in s}\Lambda(s,ij)x_ix_j}$$

It is clear from (3) that v_{OPT} will reflect the true parameter value closely when the best linear fit between y_i and x_i goes through the origin and the residual from it are small.

Remark 1. The optimal inclusion probabilities are not consistent since $\Sigma\Sigma_U\pi_{ij0}\neq(n-1)\pi_{i0}$

Remark 2. It is unlikely that a design is chosen solely for the purpose of optimum estimation of a quadratic function.

3. Simulation

The preceding variance estimators are compared empirically on 22 natural populations listed in Table 4 of Appendix B. For comparison of the variance estimators, v_1, v_{12}, v_{13}, v_{22} and v_{OPT}, a sample of size $n = 4$ was drawn using Midzuno-Sen sampling scheme from each of the populations and these variance estimators were computed. These procedure is repeated $M = 10{,}000$ times.

For each variance estimate v, its relative percentage bias is calculated as

$$RB(v) = 100 * \frac{\bar{v} - V}{V} ,$$

the relative efficiency as

$$RE(v) = \frac{MSE(v_1)}{MSE(v)}$$

where $\bar{v} = \frac{1}{M}\sum_{j=1}^{M} v_{(j)}$, $MSE(v) = \frac{1}{M}\sum_{j=1}^{M}(v_{(j)} - V)^2$

Table 1 reports the values of RE of the estimators v_{12}, v_{13}, v_{22}, and v_{OPT}. The values of RB(%) and probability of taking negative values are reported in Table 2 and Table 3, respectively, given in Appendix A.

Remark 3. Among v_2, v_3 and v_4 (suggested by Chaudhuri, see Section 1), v_2 is better for most cases. Moreover, the performances of v_1 and v_2 are similar. For these reasons the simulated results on these estimators are not presented here.

Remark 4. Among the estimators v_{10}, v_{20} and v_{23}, v_{10} is better than v_{23} and v_{23} is better than v_{20}. For most of the populations their RBs are negative indicating that these estimators are under estimating the true variance. Also, these estimators have taken frequently the negative values for most of the populations. Moreover, v_{10} is consistently poorer than v_{22} in both criteria. In short, the estimators v_{10}, v_{20} and v_{23} are shatteringly bad and therefore the corresponding results are omitted from the respective tables.

Table 1. RE under Midzuno Sampling

Popl.	v_1	v_{12}	v_{13}	v_{22}	v_{OPT}
1	1	84473.59	76279.33	70338.32	124200.59
2	1	63.65	66.12	65.35	89.67
3	1	3205.06	3020.11	3015.52	287.83
4	1	156.62	214.59	213.11	238.16
5	1	286.67	271.66	274.04	538.19
6	1	953.42	658.26	518.02	1735.75
7	1	249.41	310.31	318.88	413.36
8	1	187.87	254.11	257.26	178.27
9	1	8.88	12.83	13.66	14.59
10	1	123.25	147.79	133.80	209.91
11	1	103.75	130.95	129.26	199.33
12	1	48.19	54.98	54.66	96.40
13	1	1298.57	1297.42	1288.07	1419.42
14	1	65.47	98.35	97.62	128.22
15	1	18.18	17.38	17.06	21.01
16	1	42.05	36.45	30.47	11.57
17	1	201.97	183.72	177.04	300.46
18	1	29.93	24.47	20.12	37.81
19	1	22.55	20.45	16.99	28.36
20	1	9.46	8.69	7.59	15.79
21	1	24.45	27.07	24.69	47.39
22	1	67.65	62.91	56.64	45.04

Tables 1–3 lead to the following comments

- The estimators v_{12}, v_{13} and v_{22} are comparable among each other from RE point of view. Among these estimators v_{12} has smaller absolute RBs and has taken negative values for a very few populations with negligible probabilities. Overall v_{12} is the best, v_{13} the middle and v_{22} the worst for most cases.
- v_{OPT} has smaller MSE but have bigger bias (in magnitude).
- Empirically, v_{OPT} is the only non-negative estimator for all the populations.
- The scatter plot of the populations 1-8, 10-14, 17, 19 and 21 reveals that a linear model $y_i = \beta x_i + \varepsilon_i$ might be appropriate and the relationship between y and x is strong. The populations 1-8, 10-14 have the variance structure $\mathcal{V}(y_i) \propto x_i$, whereas the populations 17, 19 and 21 have the

variance structure $\mathcal{V}(y_i) \propto x_i^2$. The relationship between y and x is curvilinear for the populations 18 and 20, whereas for the populations 9, 15, 16 and 20 no systematic pattern is found though the correlation between y and x is moderate to high. Clearly, the populations 1-8, 10-14, 17, 19 and 20 have fulfilled the requirements for the v_{OPT} estimator, given in (3). Obviously, for these populations v_{OPT} has performed better than the other (except one case).

4. Conclusions

Based on the empirical study in previous section we arrive at the following conclusions.

1) We can rank the performance of v_{12}, v_{13} and v_{22} as $v_{12} \succ\succ v_{13} \succ\succ v_{22}$, where '$\succ\succ$' means 'better than', with respect to all criteria considered here. Thus, the estimators v_{12} (with $\alpha = \beta = 0$) suggested by Chaudhuri (1981) performed very well for the population having moderate to high correlation between y and x.
2) It is clear that v_{OPT} will reflect the true variance clearly when the best linear fit goes through the origin and the residual from it are small. v_{OPT} will perform badly if the true model deviates from the assumed model.

Appendix A

Table 2. RB(%) under Midzuno Sampling.

Popl.	v_1	v_{12}	v_{13}	v_{22}	v_{OPT}
1	6985.55	-11.92	-20.48	-22.46	-9.53
2	178.62	-42.77	-46.84	-48.14	-55.91
3	363.68	-14.30	-23.95	-25.64	154.60
4	248.91	-15.00	-22.18	-24.03	-15.97
5	625.18	-25.90	-33.17	-35.54	-41.40
6	-535.44	2.27	10.27	13.64	8.68
7	552.97	7.44	0.64	-1.09	-2.88
8	165.00	-5.89	-9.26	-10.08	14.22
9	55.16	-8.77	-16.99	-19.46	17.47
10	12.04	-3.38	-4.82	-5.19	-6.27
11	155.22	-10.64	-15.49	-16.87	-13.51
12	117.84	-10.40	-15.48	-17.06	-19.75
13	113.12	-1.47	-1.93	-2.05	-0.56
14	113.78	-5.59	-11.26	-12.90	-5.03
15	59.47	-30.13	-32.41	-33.01	-35.97
16	-4.92	-0.64	1.00	1.47	69.32
17	368.93	-25.47	-30.86	-32.28	-38.85
18	-55.21	-0.18	5.35	7.27	19.05
19	45.16	1.56	-5.41	-7.58	8.52
20	-14.87	6.91	9.17	9.85	4.68
21	84.07	-3.23	-8.69	-10.63	-13.82
22	92.01	-11.99	-19.64	-21.34	0.99

Table 3. Probability of taking negative values

Popl.	p_1	p_{12}	p_{13}	p_{22}	p_{OPT}
1	0.2981	0	0.0003	0.0003	0
2	0.2579	0	0.0040	0.0047	0
3	0.0032	0.0004	0.0010	0.0011	0
4	0.2549	0	0	0.0008	0
5	0.2552	0	0.0004	0.0004	0
6	0.0044	0.0001	0.0004	0.0005	0
7	0.3729	0	0	0	0
8	0.0037	0	0	0	0
9	0.2600	0.0173	0.0373	0.0376	0
10	0.4071	0	0	0	0
11	0.3343	0	0	0	0
12	0.3114	0	0	0	0
13	0.0046	0	0	0	0
14	0.3296	0	0	0	0
15	0.2379	0	0	0	0
16	0.0031	0.0001	0.0004	0.0004	0
17	0.2860	0	0	0	0
18	0.0038	0.0001	0.0003	0.0004	0
19	0.0024	0	0.0003	0.0004	0
20	0.0030	0	0.0001	0.0001	0
21	0.2923	0.0078	0.0243	0.0267	0
22	0.2876	0	0.0018	0.0018	0

Appendix B

Table 4. Study Population

Popl.	N	CV(x)	CV(y)	ρ(x,y)	Source	x	y
1	15	0.388	0.376	0.996	D. Gujarati, p.173	Money Supply	GNP
2	15	0.425	0.571	0.995	D. Gujarati, p.224	Labor input (per thousand persons)	Real gross product millions of NT ($)
3	17	0.698	0.712	0.988	Murthy(1967), p.399	area under wheat (1963)	area under wheat (1964)
4	15	0.340	0.490	0.984	D. Gujarati, p.184	Money Supply	GNP
5	20	0.480	0.346	0.980	D. Gujarati, p.227	Aerospace industry sales	Defense budget outlays
6	17	0.732	0.760	0.977	Murthy(1967), p.399	area under wheat (1963)	area under wheat (1964)
7	14	0.334	0.565	0.975	D. Gujarati, p.352	GNP	Merchandise imports
8	24	0.235	0.098	0.968	Murthy(1967), p.228	Fixed Capital	Output for Factories
9	23	0.597	0.186	0.947	D. Gujarati, p.228	Real disposable income per capita ($)	Per capita consumption of chickens (lbs)
10	14	0.276	0.217	0.944	D. Gujarati, p.352	Wage income	Consumption
11	23	0.369	0.186	0.937	D. Gujarati, p.228	Composite real price of chicken substitutes per lb weighted avg.of x2 to x5	Per capita consumption of chickens (lbs)
12	23	0.414	0.186	0.935	D. Gujarati, p.228	Real retail price of beef per lb	Per capita consumption of chickens (lbs)
13	25	0.117	0.124	0.920	Murthy(1967), p.228	Fixed Capital	Output for Factories in a region
14	23	0.390	0.186	0.912	D. Gujarati, p.228	Real retail price of pork per lb	Per capita consumption of chickens (lbs)
15	17	0.136	0.287	0.887	D. Gujarati, p.230	Long-term interest rate (%)	Nominal money crores of rupees
16	23	0.642	0.702	0.881	Murthy(1967), p.128	number of persons (1951)	no. of cultivators
17	15	0.289	0.197	0.871	D. Gujarati, p.216	Real Capital input (millions of NT,$)	Real gross product millions of NT ($)
18	25	0.607	0.657	0.868	Murthy(1967), p.128	area in sq.miles	no. of cultivators
19	17	0.607	0.712	0.853	Murthy(1967), p.399	cultivated area (1961)	area under wheat (1964)
20	19	0.524	0.647	0.829	Murthy(1967), p.128	number of persons (1961)	no. of cultivators
21	22	0.518	0.375	0.800	D. Gujarati, p.279	Income	Savings
22	13	0.217	0.392	0.787	D. Gujarati, p.203	Expected or anticipated inflation rate (%) at time t	Actual rate of inflation (in %) at time t

Acknowledgment

We would like to thank anonymous referees and Prof. Parimal Mukhopadhyay for helpful comments and suggestions.

REFERENCES

BREWER, K.R.W. and TAM, S.M. (1990). Is the assumption of uniform intra-class correlation ever justified? *Australian Journal of Statistics*, 32(3), 411–423.

CASSEL, C. M., SÄRNDAL, C. E. and WRETMAN, J. H. (1977). *Foundation of Inference in Survey Sampling*, John Wiley, New York.

CHAUDHURI, A. (1975). On some inference problems with finite populations and related topics in survey sampling theory. Economic Statistics Papers. No.10, 1975, University of Sydney.

CHAUDHURI, A. (1976). A non-negativity criterion for a certain variance estimator. *Metrika*, 23, 201–205.

CHAUDHURI, A. (1981). Non- negative unbiased variance estimators. *Current topics in survey sampling*, Academic Press, Inc., 317–328.

CHAUDHURI, A. and ARNAB, R. (1981). On non- negative variance estimation. *Metrika*, 28, 1–12.

DENG, L.Y. and WU, C.F.J. (1987). Estimation of the regression estimator. *Journal of American Statistical Association*, 82, 568–576.

GUJARATI, D. N. (1995). *Basic Econometrics*, McGraw-Hill, Inc.

ISAKI, C. T. and FULLER (1982). Survey Design under the regression super-population model, *Journal of the American Statistical Association*, 7, 89–96

KREWSKI, D. and CHAKRABARTY, R. P. (1981). On the stability of the jackknife variance estimator in ratio estimation. *Journal of Statistical Planning and Inference*, 5, 71–78.

MIDZUNO, H. (1950). An outline of the theory of sampling systems. *Annals Institute of Statistics and Mathematics*, Tokyo, 1, 149–156.

MONTGOMERY, D. C.; PECK, E. A. and VINING, G. G. (2003). *Introduction to linear regression analysis*. Third Edition, John Wiley & Sons, Inc.

MUKHOPADHYAY, P. (1996). *Inferential Problems in Survey Sampling*, New Age International, New Delhi.

MURTHY, M. N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

PATEL, P. A. and SHAH, D. N. (1999). Model-based estimation for a finite population variance. *Journal of the Indian Statistical Association*, 37(1), 27–35.

RAO, J. N. K. (1968). Some small sample results in ratio and regression Estimation. *Journal of the Indian Statistical Association*, 6, 160–168.

RAO, J. N. K. (1969). Ratio and regression estimators, in new developments in Survey Sampling (N.L.John and H. Smith, eds.), Wiley, New York, 213–234.

RAO, J. N. K. (1979). On deriving mean square errors and their non-negative unbiased estimators in finite population sampling. Journal of the Indian Statistical Association, 17, 125–136.

RAO, J. N. K. and BEEGLE, L. D. (1967). A Monte Carlo study of some ratio Estimators. *Sankhya, Series B*, 29, 47–56.

RAO, J. N. K. and KUZIK, R. A. (1974). Sampling errors in ratio estimation. *Sankhya, Series C*, 36 (1974), 43–58.

RAO, J.N.K. and VIJAYAN, K. (1977). On estimating the variance in sampling with probability proportional to aggregate size. *Journal of the American Statistical Association*, 72, 579–584.

RAO, P.S.R.S. and RAO, J.N.K. (1971). Small sample results for ratio estimators. *Biometrika*, 58, 625–630.

RAO, T. J. (1972). On the variance of the ratio estimator for the Midzuno-Sen Sampling scheme. *Metrika*, 18, 209–215.

RAO, T. J. (1977). Estimating the variance of the ratio estimator for the Midzuno-Sen sampling scheme. *Metrika*, 24, 203–215.

ROYALL, R.M. and CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351–358.

ROYALL, R.M. and CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 355–359.

ROYALL, R.M. and EBERHADT, K.R. (1975). Variance estimates for ratio estimator. *Sankhya, Series C*, 37, 43–52.

TRACY, D. S. and MUKHOPADHYAY, P. (1994). On non-negative unbiased variance estimation for Midzuno strategy, Pak. J. Stat., 10 (3), 575–583.

VALLIANT R, DORFAM, A.H., ROYALL, R.M. (2000). *Finite population sampling and inference*, Wiley, New York.

VIJAYAN, K., MUKHOPADHYAY, P. and BHATTACHARYA, S. (1995). On non-negative unbiased estimation of quadratic forms in finite population, Australian Journal of Statistics, 37 (2), 168–178.

WU, C. F. J. (1982). Estimation of variance of the ratio estimator. *Biometrika*, 69, 183–189.

WU, C.F.J. and DENG, L.(1983). Estimation of variance of the ratio estimator: An empirical study. *In Scientific Inference Data Analysis and Robustness.*

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 387—395

ESTIMATION OF LIFE-TABLES UNDER RIGHT-CENSORING

Agnieszka Rossa[1]

ABSTRACT

The paper deals with a class of non-parametric estimators of conditional probabilities of failure prior $x+y$ given survival to x under the random and observable right-censorship model. The proposed estimators are based on a specific sequential sampling scheme. Application of the estimators in life-table analysis is presented.

Key words: Life-table analysis; non-parametric estimation; right-censored data; sequential sampling.

1. Introduction

Let T be a non-negative random variable representing a duration time between two well-defined events, i.e. an initial and a final event (the second event is usually called a failure). Let T be defined on a probability space (Ω, $\mathcal{A}$, **P**). The survival function of T is then defined as

$$\overline{F}(x) = \mathbf{P}(T > x), \quad x \in \mathbf{R},$$

with $\overline{F}(x) = 1$ for $x \leq 0$.

Consider a probability that the failure occurs in a time interval $(x,x+y]$ given $T>x$. Using the actuarial notation, such a probability is usually denoted by ${}_yq_x$. Thus, we have

$${}_yq_x = \mathbf{P}(T \leq x + y \mid T > x). \tag{1}$$

It can be also expressed in terms of $\overline{F}$ as

[1] For correspondence: Institute of Statistics and Demography, University of Łódź, ul. Rewolucji 1905, 41,90-214 Łódź, Poland; e-mail: agrossa@uni.lodz.pl.

$$ {}_y q_x = \frac{\overline{F}(x) - \overline{F}(x+y)}{\overline{F}(x)} = 1 - \frac{\overline{F}(x+y)}{\overline{F}(x)}. \tag{2} $$

Denoting

$$ {}_y p_x = \frac{\overline{F}(x+y)}{\overline{F}(x)} \tag{3} $$

we receive from (2) the obvious formula ${}_y q_x = 1 - {}_y p_x$.

Thus, ${}_y q_x$ is a probability of failure prior $x+y$ given survival to x, and ${}_y p_x$ is a probability of surviving beyond $x+y$ given survival to x (for $y \geq 0$).

In many applications, especially in demographic and actuarial studies, a table of numerical values of ${}_y q_x$ and ${}_y p_x$ for a certain selected values of x and y are considered, most commonly the integers. Typically, a complete table shows values of ${}_y q_x$ and ${}_y p_x$ for all integer values of $x = 0,1,\ldots,\omega-1$, and for y=1, where ω stands for a maximal possible lifetime. In such cases probabilities ${}_1 q_x$ and ${}_1 p_x$, denoted hereafter by q_x, p_x, are organized in a table of numbers (see Table 1).

Table 1. Tabular survival model

x	q_x	p_x
0	q_0	p_0
1	q_1	p_1
⋮	⋮	⋮
$\omega-1$	$q_{\omega-1}$	$p_{\omega-1}$

From probabilities ${}_y q_x$ and ${}_y p_x$ various life-table characteristics are derived, such as numbers of lives and failures in a cohort group, expected lifetimes, times of exposure etc.

2. Two censorship models

In many real-life situations duration times $T_1, T_2, \ldots, T_n$ for individuals cannot be fully observed. There can be one or more random events (other than the failure), occurring of which terminates the observation of some individuals. In such cases it is said that the observations are right-censored. For instance, very often it is impossible to keep a study open until all the individuals in the sample

have experienced the failure. Instead, a period of observation (follow-up) is chosen, and individuals are observed until the end of that period. If an *i*-th individual fails by the time of the analysis, then it yields a true duration time T_i, otherwise a censoring time Z_i, say, is observed, such that $Z_i < T_i$. If individuals enter the study at random times then the Z_i's are random variables, independent of the T_i's.

Under *the random censorship model* [see Efron (1967)] it is usually assumed, that the T_i's are iid, and independent of the Z_i's. It is also assumed that T_i is observed whenever $T_i \le Z_i$, and Z_i is observed whenever $Z_i < T_i$. In other words, a pair (X_i, δ_i) is observed, where

$$X_i = \min(T_i, Z_i), \tag{4}$$

$$\delta_i = \mathbf{1}(T_i \le Z_i), \tag{5}$$

and $\mathbf{1}(\cdot)$ denotes an indicator function, indicating whether X_i is right-censored ($\delta_i = 0$) or not ($\delta_i = 1$).

However, if censoring mechanism is only due to the termination of the observation period, then each Z_i represents the time elapsed from entering an *i*-th subject into the study to the end of the observation period. In such cases the Z_i's are fully observed and one observes (X_i, Z_i) for *i*=1,2,…. We will refer to this special type of the model (4)-(5) as *the random and observable censorship model*.

Throughout the rest of the paper we will assume that the duration times T_i, as well as the censoring times Z_i are mutually independent, continuous random variables in the probability space (Ω, $\mathcal{A}$, **P**) and that the T_i's have a common cumulative distribution function (cdf) *F*, whereas the Z_i's have a common cdf *G*. These assumptions imply that variables X_i defined in (4) have a common cdf expressed as

$$H = 1 - \overline{F} \cdot \overline{G},$$

where $\overline{F} = 1 - F$, $\overline{G} = 1 - G$.

3. Estimation of conditional probabilities under the random censorship model

Let us consider a random censored sample of a fixed size *n*

$$(X_1,\delta_1),(X_2,\delta_2),\ldots,(X_n,\delta_n). \tag{6}$$

Hereafter, observations $(X_i,\delta_i=1)$ will be called failures, whereas $(X_i,\delta_i=0)$ will be called losses. Let

$$0=x_0<x_1<\ldots<x_J<x_{J+1}=y_0 \tag{7}$$

be a partition of the range $[0,y_0]$ of observation into *J*+1 intervals $I_j=(x_j,x_{j+1}]$ for $j=0,1,\ldots,J$. We will consider the following statistics, defined for each $j=0,1,\ldots,J$.

$$D_j=\sum_{i=1}^{n}\mathbf{1}(X_i\in I_j,\delta_i=1), \tag{8}$$

$$L_j=\sum_{i=1}^{n}\mathbf{1}(X_i\in I_j,\delta_i=0), \tag{9}$$

$$M_j=\sum_{i=1}^{n}\mathbf{1}(X_i>x_j), \tag{10}$$

where D_j and L_j represent numbers of failures and losses in an interval $(x_j,x_{j+1}]$, respectively, whereas M_j represents number of items at risk at the beginning of the interval (i.e. numbers of individuals surviving beyond x_j). Note that $M_j=0$ implies $D_j=0$ and $L_j=0$.

Table 2. Life-Table Statistics Under Random Censorship Model

Intervals $I_j=(x_j,x_{j+1}]$	No. of survivors M_j	No. of failures D_j	No. of losses L_j
$(0,\ x_1]$	M_0	D_0	L_0
$(x_1,x_2]$	M_1	D_1	L_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(x_J,x_{J+1}]$	M_J	D_J	L_J

The well-known estimators of the conditional probabilities ${}_{y_j}q_{x_j}$ and ${}_{y_j}p_{x_j}$, where $y_j = x_{j+1} - x_j$, are usually defined by means of the statistics (8)-(10). For instance, the Standard Life-Table Estimators, encountered in the literature [e.g. Berkson and Gage (1950), Gehan (1965), Breslow and Crowley (1974), Daya (2005)], are defined as

$$ {}_{y_j}\hat{q}_{x_j} = \frac{D_j}{M_j - \frac{1}{2}L_j} \quad \text{and} \quad {}_{y_j}\hat{p}_{x_j} = 1 - {}_{y_j}\hat{q}_{x_j} \tag{11} $$

Both ${}_{y_j}\hat{q}_{x_j}$ and ${}_{y_j}\hat{p}_{x_j}$ are not defined if $M_j = 0$. In such cases it is usually assumed that ${}_{y_j}\hat{q}_{x_j} = 1$, ${}_{y_j}\hat{p}_{x_j} = 0$.

It is also well-known that the Standard Life-Table Estimators ${}_{y_j}\hat{q}_{x_j}$ are negatively biased, what indicates that ${}_{y_j}\hat{p}_{x_j}$ are biased positively. In general, both estimators are asymptotically biased and not consistent [see Breslow and Crowley (1974)].

Taking into account all these disadvantages mentioned above we will propose a new class of estimators of probabilities (2) and (3), which are unbiased and consistent. The proposed estimators are derived under the model of random and observable censorship and are based on a special type of sequential sampling.

4. Estimation of conditional probabilities under random and observable censoring

Consider the sequence of observations $(X_1, Z_1), (X_2, Z_2), \ldots$. Let k be a fixed integer $(k \geq 2)$ and y_0 be a fixed real value such that $0 < y_0 < \sup\{y : H(y) < 1\}$, where $H(\cdot)$ denotes a common cdf of the X_i's.

Suppose that individuals arrive at random into the study and the observation period terminates if for k individuals one observes $X_{i_j} > y_0$, $j = 1,2,\ldots,k$. Such a sequential sampling leads to a sample

$$ (X_1, Z_1), (X_2, Z_2), \ldots, (X_{N_k}, Z_{N_k}) \tag{12} $$

where the sample size N_k is a random variable distributed according to the negative binomial distribution with parameters k and $p = 1 - H(y_0)$.

Let us consider the partition (7) of the time interval $[0, y_0]$, and define the following statistics

$$R_{k,j} = \sum_{i=1}^{N_k} \mathbf{1}(X_i > x_{j-1}, Z_i > x_j), \quad j = 1,2,\ldots,J,J+1, \tag{13}$$

$$M_{k,j} = \sum_{i=1}^{N_k} \mathbf{1}(X_i > x_j), \quad j = 0,1,\ldots,J,J+1. \tag{14}$$

In the special case there is $M_{k,0} = N_k$ and $M_{k.J+1} = k.$

Definition.

Under the model of random and observable censorship the life-table estimators of the conditional probabilities ${}_{y_j}q_{x_j}$ and ${}_{y_j}p_{x_j}$ ($j = 0,1,\ldots,J$) take the form

$${}_{y_j}\widetilde{q}_{k,x_j} = \frac{R_{k,j+1} - M_{k,j+1}}{R_{k,j+1} - 1} \tag{15}$$

and

$${}_{y_j}\widetilde{p}_{k,x_j} = 1 - {}_{y_j}\widetilde{q}_{k,x_j}. \tag{16}$$

where $y_j = x_{j+1} - x_j$.

Proposition.

The estimators (15) and (16) are unbiased and consistent estimators of the conditional probabilities ${}_{y_j}q_{x_j}$ and ${}_{y_j}p_{x_j}$. Their variances satisfy the following equivalence

$$\mathbf{V}\left({}_{y_j}\widetilde{q}_{k,x_j}\right) = \mathbf{V}\left({}_{y_j}\widetilde{p}_{k,x_j}\right) = {}_{y_j}q_{x_j}\,\mathbf{E}\left(\frac{1}{R_{k,j+1}}\right) - {}_{y_j}q_{x_j}^2\,\mathbf{E}\left(\frac{1}{R_{k,j+1}+1`}\right) \tag{17}$$

where the expectations $\mathbf{E}\left(\frac{1}{R_{k,j+1}}\right)$ and $\mathbf{E}\left(\frac{1}{R_{k,j+1}+1}\right)$ can be expressed as

$$\mathbf{E}\left(\frac{1}{R_{k,j+1}}\right) = \left(\frac{p_j}{q_j}\right)^k \int_0^{q_j} u^{k-1}(1-u)^{-k}\,du,$$

$$\mathbf{E}\left(\frac{1}{R_{k,j+1}+1}\right) = \left(\frac{p_j}{q_j}\right)^k \frac{1}{q_j}\int_0^{q_j} u^k(1-u)^{-k}\,du,$$

and

$$ {}_{y_j}p_{x_j} = \frac{\overline{F}(x_{j+1})}{\overline{F}(x_j)}, \qquad {}_{y_j}q_{x_j} = 1 - {}_{y_j}p_{x_j}, $$

$$ p_j = \frac{\overline{H}(y_0)}{\overline{F}(x_j)\overline{G}(x_{j+1})}, \quad q_j = 1 - p_j. $$

The proof of the proposition was given by Rossa (2005), pp. 80-82.

It follows from (17) that variances of ${}_{y_j}\widetilde{q}_{k,x_j}$ and ${}_{y_j}\widetilde{p}_{k,x_j}$ can be estimated by means of the following expression

$$ \hat{\mathbf{V}}\left({}_{y_j}\widetilde{q}_{x_j} \right) = \hat{\mathbf{V}}\left({}_{y_j}\widetilde{p}_{x_j} \right) = {}_{y_j}\widetilde{q}_{x_j} \frac{1}{R_{k,j+1}} - {}_{y_j}\widetilde{q}^2_{x_j} \frac{1}{R_{k,j+1}+1}. \qquad (18) $$

5. A numerical example

To illustrate in detail the sampling scheme and the estimators proposed, a hypothetical study will be considered. Assume that the subject of observation is the time T_i elapsed from the issuance of a life-insurance policy up to the death of an *i*-th insured person (in years), and let Z_i denote the time elapsed from the issuance of his/her policy up to the termination of the follow-up study.

Due to right-censoring variables $T_1, T_2, \ldots$ are possibly unobserved. However, variables $X_i = \min(T_i, Z_i)$, as well as censoring variables Z_i for $i = 1,2,\ldots$ are fully observed.

We will assume that persons arrive at random into the study and the follow-up period terminates when for *k*=15 insured persons we observe $X_i > y_0$. Suppose here that $y_0 = 5$ (years).

Consider a partition of the range $[0, y_0]$ into subintervals $I_j = (x, x+1]$, where $x = 0,1,\ldots, y_0 - 1$. Table 3 (columns 2 and 3) contains exemplary values of statistics $R_{k,x+1}, M_{k,x+1}, x = 0,1,\ldots,4$ which can be observed under such a sampling scheme. The next three columns present estimates of the conditional probabilities (2) and (3) derived from the formulae (15) and (16) for $y_j = 1$, as well as estimates of their variances $\mathbf{V}(\widetilde{q}_{k,x})$ obtained from the formula (18).

Table 3. Estimates $\widetilde{q}_{15,x}$, $\widetilde{p}_{15,x}$ of probabilities q_x, p_x and $\mathbf{V}(\widetilde{q}_{15,x})$

x	$R_{15,x+1}$	$M_{15,x+1}$	$\widetilde{q}_{15,x}$	$\widetilde{p}_{15,x}$	$\hat{\mathbf{V}}(\widetilde{q}_{15,x})$
0	63	60	0,0484	0,9516	0,000805
1	48	45	0,0638	0,9362	0,001412
2	37	34	0,0833	0,9167	0,002434
3	28	24	0,1481	0,8519	0,006046
4	19	15	0,2222	0,7778	0,014163

6. Discussion

In the paper two classes of non-parametric estimators of conditional probabilities ${}_y q_x = \mathbf{P}(T \le x+y \mid T > x)$ and ${}_y p_x = \mathbf{P}(T > x+y \mid T > x)$ are proposed, derived under the so-called random and observable censorship model. The proposed estimators are unbiased and consistent, as opposed to the well-known Standard Life-Table Estimators.

Both classes of estimators are based on a special sequential sampling scheme. In this scheme an integer $k \ge 2$ and a positive value y_0 such that $y_0 < \sup\{y : H(y) < 1\}$ have to be fixed in advance, where H is a cdf of $X = \min(T, Z)$. Note that usually it is not difficult to choose a proper value of y_0, even if the function H is unknown. It is sufficient to know the maximal possible values t and z, say, of the duration and censoring times, respectively. Then, for any $y_0 \in (0, \min(t,z))$ and the condition $y_0 < \sup\{y : H(y) < 1\}$ is satisfied.

REFERENCES

BERKSON, J. & GAGE, R. 1950. Calculation of Survival Rates for Cancer. *Proc. of the Staff Meetings of the Mayo Clinic*, 25, pp. 270–286.

BRESLOW, N. E. & CROWLEY, J. J. 1974. Large Sample Study of the Life Table and Product Limit Estimates under Random Censorship. *Annals of Statistics,* 2, pp. 437–454.

DAYA, S. 2005. Life Table (Survival) Analysis to Generate Pregnancy Rates in Assisted Reproduction: Are We Overestimation Our Success Rates? *Human Reproduction*, 20, 1135–1143.

EFRON, B. 1967. The Two-Sample Problem with Censored Data. *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 4, pp. 831–853, University of California Press, Berkeley.

GEHAN, E. A. 1965. A Generalized Wilcoxon Test for Comparing Arbitrarily Single-Censored Samples. *Biometrika*, 52, pp. 203–234.

ROSSA, A. 2005. Estimation of Survival Distributions under Right-Censoring and Applications, University of Łódź (in Polish).

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 397—414

ESTIMATION OF MEAN UNDER IMPUTATION OF MISSING DATA USING FACTOR-TYPE ESTIMATOR IN TWO-PHASE SAMPLING

Diwakar Shukla, Narendra Singh Thakur, Sharad Pathak[1] **and Dilip Singh Rajput**[2]

ABSTRACT

In sample surveys, the problem of non-response is one of the most frequent and widely appearing, whose solution is required to obtain using relevant statistical techniques. The imputation is one such methodology, which uses available data as a source for replacement of missing observations. Two-phase sampling is useful when population parameter of auxiliary information is unknown. This paper presents the use of imputation for dealing with non-responding units in the setup of two-phase sampling. Two different two-phase sampling strategies (sub-sample and independent sample) are compared under imputed data setup. Factor-Type (F-T) estimators are used as tools of imputation and simulation study is performed over multiple samples showing the comparative strength of one over other. First imputation strategy is found better than second whereas second sampling design is better than first.

Key words: Estimation, Missing data, Imputation, Bias, Mean squared error (MSE), Factor Type (F-T) estimator, Two-phase sampling, Simple Random Sampling Without Replacement (SRSWOR), Compromised Imputation (C. I.).

1. Introduction

Let $\Omega = \{1, 2,..., N\}$ be a finite population with Y_i as a variable of main interest and X_i $(i = 1,2,..., N)$ an auxiliary variable. As usual, $\overline{Y} = N^{-1}\sum_{i=1}^{N} Y_i$,

[1] Diwakar Shukla, Narendra Singh Thakur, Sharad Pathak, Deptt. of Mathematics and Statistics, H.S. Gour University of Sagar, Sagar,(M.P.) INDIA, Pin-470003.
e-mails:diwakarshukla@rediffmail.com, nst_stats@yahoo.co.in, s.pathak_stats@yahoo.co.in.
[2] Dilip Singh Rajput, Govt. College, Rehli, Sagar (M.P.), INDIA.

$\overline{X} = N^{-1}\sum_{i=1}^{N} X_i$ are population means, $\overline{X}$ is assumed known and $\overline{Y}$ under investigation. Singh and Shukla (1987) proposed Factor Type (F-T) estimator to obtain the estimate of population mean under setup of SRSWOR. Some other contributions on Factor-Type estimator, in similar setup, are due to Singh and Shukla (1991) and Singh et al. (1993).

With $\overline{X}$ unknown, the two-phase sampling is used to obtain the estimate of population mean and Shukla (2002) suggested F-T estimator under this case. But when few of observations are missing in the sample, the F-T estimator fails to estimate. This paper undertakes the problem of Shukla (2002) with suggested imputation procedures for missing observations.

Rubin (1976) addressed three missing observation concepts: missing at random (MAR), observed at random (OAR) and parameter distribution (PD). Heitjan and Basu (1996) explained the concept of missing at random (MAR) and introduced the missing completely at random (MCAR). The present discussion is on MCAR wherever the non-response is quoted. Rao and Sitter (1995) discussed a new linearization variance estimator that makes more complete use of the sample data than a standard one. They have shown its application to 'mass' imputation under two-phase sampling and deterministic imputation for missing data. Singh and Horn (2000) suggested a Compromised Imputation (C-I) procedure in which the estimator of mean obtained through C-I remains better than obtained from ratio method of imputation and mean method of imputation. Ahmed et al. (2006) designed several generalized structure of imputation procedures and their corresponding estimators of the population mean. Motivation is derived from these and from Shukla (2002) to extend the content for the imputation setup.

Consider a preliminary large sample S' of size n' drawn from population Ω by SRSWOR and a secondary sample S of size n ($n < n'$) drawn in either of the following manners:

Case I: as a sub-sample from sample S' (denoted by design F_1) as in fig. 1(a),

Case II: independent to sample S' (denote by design F_2) as in fig. 1(b) without replacing S'.

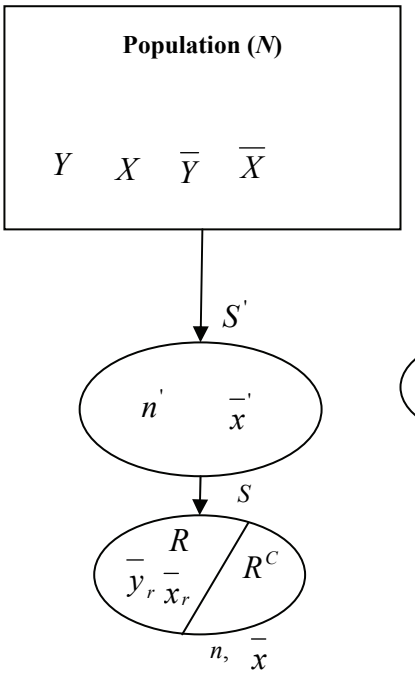


Fig. 1(a) [Case *I*, F_1]

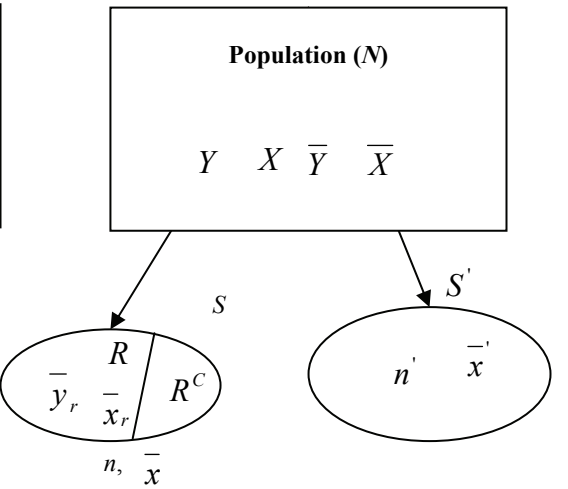


Fig. 1(b) [Case *II*, F_2]

Let sample *S* of *n* units contains *r* responding units ($r < n$) forming a subspace *R* and ($n - r$) non-responding with sub-space R^C in $S = R \bigcup R^C$. For every $i \in R$, the y_i is observed available. For $i \in R^C$, the y_i values are missing and imputed values are computed. The i^{th} value x_i of auxiliary variate is used as a source of imputation for missing data when $i \in R^C$. Assume for *S*, the data $x_s = \{x_i : i \in S\}$ and $\{x_{i'} : i' \in S'\}$ are known with mean $\bar{x} = (n)^{-1} \sum_{i=1}^{n} x_i$ and $\bar{x}' = (n')^{-1} \sum_{i'=1}^{n'} x_{i'}$ respectively.

2. *F-T* Imputation Strategies

Two proposed strategies d_1 and d_2 for missing data under both cases are :

$$\boldsymbol{d_1}: \ (y_{d1})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\bar{y}_r \left\{ \dfrac{(A+C)\bar{x}' + fB\bar{x}}{(A+fB)\bar{x}' + C\bar{x}} \right\} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \tag{2.1}$$

$$\boldsymbol{d_2}: \ (y_{d2})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\bar{y}_r \left\{ \dfrac{(A+C)\bar{x}' + fB\bar{x}_r}{(A+fB)\bar{x}' + C\bar{x}_r} \right\} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \tag{2.2}$$

Under (2.1) and (2.2) point estimators of $\bar{Y}$ are :

$$\left(\bar{y}_{d1}\right) = \bar{y}_r \left[\frac{(A+C)\bar{x}' + fB\bar{x}}{(A+fB)\bar{x}' + C\bar{x}} \right]; \tag{2.3}$$

$$\left(\bar{y}_{d2}\right) = \bar{y}_r \left[\frac{(A+C)\bar{x}' + fB\bar{x}_r}{(A+fB)\bar{x}' + C\bar{x}_r} \right]; \tag{2.4}$$

where $A = (k-1)(k-2)$; $B = (k-1)(k-4)$; $C = (k-2)(k-3)(k-4)$ and k ($0 < k < \infty$) is a constant.

2.1.Some Special Cases

(i) At $k = 1$; $A = 0$; $B = 0$; $C = -6$

$$\left(\bar{y}_{d1}\right) = \bar{y}_r \left(\frac{\bar{x}'}{\bar{x}} \right) \tag{2.5}$$

$$\left(\bar{y}_{d2}\right) = \bar{y}_r \left(\frac{\bar{x}'}{\bar{x}_r} \right) \tag{2.6}$$

(ii) At $k = 2$; $A = 0$; $B = -2$; $C = 0$

$$\left(\bar{y}_{d1}\right) = \bar{y}_r \left(\frac{\bar{x}}{\bar{x}'} \right) \tag{2.7}$$

$$\left(\bar{y}_{d2}\right)=\bar{y}_r\left(\frac{\bar{x}_r}{\bar{x}'}\right) \tag{2.8}$$

(iii) At $k = 3$; $A = 2$; $B = -2$; $C = 0$

$$\left(\bar{y}_{d1}\right)=\bar{y}_r\left(\frac{\bar{x}'+f\bar{x}}{(1-f)\bar{x}'}\right) \tag{2.9}$$

$$\left(\bar{y}_{d2}\right)=\bar{y}_r\left(\frac{\bar{x}'-f\bar{x}}{(1-f)\bar{x}'}\right) \tag{2.10}$$

(iv) At $k = 4$; $A = 6$; $B = 0$; $C = 0$

$$\left(\bar{y}_{d1}\right)=\bar{y}_r \tag{2.11}$$

$$\left(\bar{y}_{d2}\right)=\bar{y}_r \tag{2.12}$$

3. Properties of Imputation Strategies

Let $B(.)_t$ and $M(.)_t$ denote the bias and mean squared error (*M.S.E.*) of estimator under sampling design $t = I, II$ (or F_1, F_2). Large sample approximations are:

$$\bar{y}_r=\bar{Y}(1+e_1);\ \bar{x}_r=\bar{X}(1+e_2);\ \bar{x}=\bar{X}(1+e_3) \text{ and } \bar{x}'=\bar{X}(1+e_3')$$

Using two-phase sampling, following Rao and Sitter (1995) and the mechanism of MCAR, for given r, n and n', we write:

***(i)* Under design F_1 [Case *I*]:**

$$E(e_1)=E(e_2)=E(e_3)=E(e_3')=0;\ E(e_1^2)=\delta_1 C_Y^2;\ \ E(e_2^2)=\delta_1 C_X^2;$$
$$E(e_3^2)=\delta_2 C_X^2;\ \ E(e_3'^2)=\delta_3 C_x^2;\ \ E(e_1e_2)=\delta_1\rho\ C_Y\ C_X;$$
$$E(e_1e_3)=\delta_2\rho\ C_Y C_X;\ \ \ E(e_1e_3')=\delta_3\rho C_Y\ C_X;\ E(e_2e_3)=\delta_2\ C_X^2;$$
$$E(e_2e_3')=\delta_3\ C_X^2;\ \ E(e_3e_3')=\delta_3\ C_X^2;$$

***(ii)* Under design F_2 [Case *II*]:**

$$E(e_1)=E(e_2)=E(e_3)=E(e_3')=0;\ E(e_1^2)=\delta_4 C_Y^2;\ \ \ E(e_2^2)=\delta_4 C_X^2;$$
$$E(e_3^2)=\delta_5 C_X^2;\ \ E(e_3'^2)=\delta_3 C_x^2;\ \ E(e_1e_2)=\delta_4\rho\ C_Y\ C_X;$$

$$E(e_1e_3)=\delta_5\rho\ C_YC_X;\quad E(e_1e_3')=0;\ E(e_2e_3)=\delta_5\ C_X^2;\ E(e_2e_3')=0;$$
$$E(e_3e_3')=0$$

where

$$\delta_1=\left(\frac{1}{r}-\frac{1}{n'}\right);\ \delta_2=\left(\frac{1}{n}-\frac{1}{n'}\right);\ \delta_3=\left(\frac{1}{n'}-\frac{1}{N}\right);$$
$$\delta_4=\left(\frac{1}{r}-\frac{1}{N-n'}\right);\ \delta_5=\left(\frac{1}{n}-\frac{1}{N-n'}\right)$$

Remark 3.1: Let

$$\theta_1=\frac{A+C}{A+fB+C};\ \theta_2=\frac{fB}{A+fB+C};\ \theta_3=\frac{A+fB}{A+fB+C};\ \theta_4=\frac{C}{A+fB+C};$$
$$P=-(\theta_1-\theta_3)=(\theta_2-\theta_4);\ (\theta_1\theta_4+\theta_2\theta_3-2\theta_3\theta_4)=P(\theta_3-\theta_4);\ V=\rho\frac{C_Y}{C_X}.$$

Theorem 3.1: Estimators $(\bar{y}_{d1})$ and $(\bar{y}_{d2})$, in terms of $e_i;\ i=1,2,3$ and e_i', could be expressed :

(i) $\bar{y}_{d1}=\bar{Y}\left[1+e_1+P\left\{e_3-e_3'-\theta_4e_3^2+\theta_3e_3'^2+e_1e_3-e_1e_3'-(\theta_3-\theta_4)e_3e_3'\right\}\right]$ (3.1)

(ii) $\bar{y}_{d2}=\bar{Y}\left[1+e_1+P\left\{e_2-e_3'-\theta_4e_2^2+\theta_3e_3'^2+e_1e_2-e_1e_3'-(\theta_3-\theta_4)e_3e_3'\right\}\right]$ (3.2)

While ignoring terms $E\left[e_i^re_j^s\right]$, $E\left[e_i^r\left(e_j'\right)^s\right]$ for $r+s>2$, $r,s=0,1,2,...$ and $i,j=1,2,3,...$ which is first order of approximation [see Cochran (2005)].

Proof:

(i) $\bar{y}_{d1}=\bar{y}_r\left[\frac{(A+C)\bar{x}'+fB\bar{x}}{(A+fB)\bar{x}'+C\bar{x}}\right]$

$=\bar{Y}(1+e_1)(1+\theta_1e_3'+\theta_2e_3)(1+\theta_3e_3'+\theta_4e_3)^{-1}$

$=\bar{Y}\left[1+e_1+P\left\{e_3-e_3'-\theta_4e_3^2+\theta_3e_3'^2+e_1e_3-e_1e_3'-(\theta_3-\theta_4)e_3e_3'\right\}\right]$

(ii) $\bar{y}_{d2}=\bar{y}_r\left[\frac{(A+C)\bar{x}'+fB\bar{x}_r}{(A+fB)\bar{x}'+C\bar{x}_r}\right]$

$=\bar{Y}(1+e_1)(1+\theta_1e_3'+\theta_2e_2)(1+\theta_3e_3'+\theta_4e_2)^{-1}$

$=\bar{Y}\left[1+e_1+P\left\{e_2-e_3'-\theta_4e_2^2+\theta_3e_3'^2+e_1e_2-e_1e_3'-(\theta_3-\theta_4)e_2e_3'\right\}\right]$

Theorem 3.2: Biases of $\left(\bar{y}_{d1}\right)_t$ and $\left(\bar{y}_{d2}\right)_t$ under $t = I, II$ (or design F_1 and F_2), up to first order of approximation are:

(i) $B\left[\bar{y}_{d1}\right]_I = -\bar{Y}P(\delta_2 - \delta_3)\left(\theta_4 C_X^2 - \rho C_Y C_X\right)$ (3.3)

(ii) $B\left[\bar{y}_{d1}\right]_{II} = \bar{Y}P\left[(\theta_3\delta_3 - \theta_4\delta_5)C_X^2 + \delta_5\rho C_Y C_X\right]$ (3.4)

(iii) $B\left[\bar{y}_{d2}\right]_I = -\bar{Y}P(\delta_1 - \delta_3)\left(\theta_4 C_X^2 - \rho C_Y C_X\right)$ (3.5)

(iv) $B\left[\bar{y}_{d2}\right]_{II} = \bar{Y}P\left[(\theta_3\delta_3 - \theta_4\delta_4)C_X^2 + \delta_4\rho C_Y C_X\right]$ (3.6)

Proof:

(i) $B\left[\bar{y}_{d1}\right]_I = E\left[\bar{y}_{d1} - \bar{Y}\right]_I$

$= \bar{Y}E\left[1 + e_1 + P\left\{e_3 - e_3' - \theta_4 e_3^2 + \theta_3 e_3'^2 + e_1 e_3 - e_1 e_3' - (\theta_3 - \theta_4)e_3 e_3'\right\} - 1\right]$

$= -\bar{Y}P(\delta_2 - \delta_3)\left(\theta_4 C_X^2 - \rho C_Y C_X\right)$

(ii) $B\left[\bar{y}_{d1}\right]_{II} = E\left[\bar{y}_{d1} - \bar{Y}\right]_{II}$

$= \bar{Y}E\left[1 + e_1 + P\left\{e_3 - e_3' - \theta_4 e_3^2 + \theta_3 e_3'^2 + e_1 e_3 - e_1 e_3' - (\theta_3 - \theta_4)e_3 e_3'\right\} - 1\right]$

$= \bar{Y}P\left[(\theta_3\delta_3 - \theta_4\delta_5)C_X^2 + \delta_5\rho C_Y C_X\right]$

(iii) $B\left[\bar{y}_{d2}\right]_I = E\left[\bar{y}_{d2} - \bar{Y}\right]_I$

$= -\bar{Y}P(\delta_1 - \delta_3)\left(\theta_4 C_X^2 - \rho C_Y C_X\right)$

(iv) $B\left[\bar{y}_{d2}\right]_{II} = E\left[\bar{y}_{d2} - \bar{Y}\right]_{II}$

$= \bar{Y}P\left[(\theta_3\delta_3 - \theta_4\delta_4)C_X^2 + \delta_4\rho C_Y C_X\right]$

Theorem 3.3: Mean squared errors of $\left(\bar{y}_{d1}\right)_t$ and $\left(\bar{y}_{d2}\right)_t$ under design F_1 and F_2, up to first order of approximation are:

(i) $M\left[\bar{y}_{d1}\right]_I = \bar{Y}^2\left[\delta_1 C_Y^2 + (\delta_2 - \delta_3)\left(P^2 C_X^2 + 2P\rho C_Y C_X\right)\right]$ (3.7)

(ii) $M\left[\bar{y}_{d1}\right]_{II} = -\bar{Y}^2\left[\delta_4 C_Y^2 + (\delta_3 + \delta_5)P^2 C_X^2 + 2P\delta_5\rho C_Y C_X\right]$ (3.8)

(iii) $M\left[\bar{y}_{d2}\right]_I = \bar{Y}^2\left[\delta_1 C_Y^2 + (\delta_1 - \delta_3)\left(P^2 C_X^2 + 2P\rho C_Y C_X\right)\right]$ (3.9)

(iv) $M\left[\bar{y}_{d2}\right]_{II} = \bar{Y}^2\left[\delta_4 C_Y^2 + (\delta_3 + \delta_4)P^2 C_X^2 + 2P\delta_4\rho C_Y C_X\right]$ (3.10)

Proof:

(i) $M\left[\bar{y}_{d1}\right]_I = E\left[\bar{y}_{d1} - \bar{Y}\right]_I^2$

$$= \bar{Y}^2 E\left[e_1 + P\left(e_3 - e_3^{'}\right)\right]^2$$

$$= \bar{Y}^2 \left[\delta_1 C_Y^2 + (\delta_2 - \delta_3)\left(P^2 C_X^2 - 2P\rho C_Y C_X\right)\right]$$

(ii) $M\left[\bar{y}_{d1}\right]_{II} = E\left[\bar{y}_{d1} - \bar{Y}\right]_{II}^2$

$$= \bar{Y}^2 E\left[e_1 + P\left(e_3 - e_3^{'}\right)\right]^2$$

$$= \bar{Y}^2 \left[\delta_4 C_Y^2 + (\delta_3 + \delta_5) P^2 C_X^2 + 2P\delta_5 \rho C_Y C_X\right]$$

(iii) $M\left[\bar{y}_{d2}\right]_I = E\left[\bar{y}_{d2} - \bar{Y}\right]_I^2$

$$= \bar{Y}^2 E\left[e_1 + P\left(e_2 - e_3^{'}\right)\right]^2$$

$$= \bar{Y}^2 \left[\delta_1 C_Y^2 + (\delta_1 - \delta_3)\left(P^2 C_X^2 - 2P\rho C_Y C_X\right)\right]$$

(iv) $M\left[\bar{y}_{d2}\right]_{II} = E\left[\bar{y}_{d2} - \bar{Y}\right]_{II}^2$

$$= \bar{Y}^2 \left[\delta_4 C_Y^2 + (\delta_3 + \delta_4) P^2 C_X^2 + 2P\delta_4 \rho C_Y C_X\right]$$

Theorem 3.4: Minimum mean squared errors of $\left(\bar{y}_{d1}\right)_t$ and $\left(\bar{y}_{d2}\right)_t$ under design F_1 and F_2 are :

(i) $Min\left[M\left(\bar{y}\right)_{d1}\right]_I = \left[\delta_1 - (\delta_2 - \delta_3)\rho^2\right] S_Y^2$ when $P = -V$ (3.11)

(ii) $Min\left[M\left(\bar{y}_{d1}\right)\right]_{II} = \left[\delta_4 - (\delta_3 + \delta_5)^{-1} \delta_5^2 \rho^2\right] S_Y^2$

when $P = -\delta_5 V / (\delta_3 + \delta_5)$ (3.12)

(iii) $Min\left[M\left(\bar{y}_{d2}\right)\right]_I = \left[\delta_1 - (\delta_1 - \delta_3)\rho^2\right] S_Y^2$ when $P = -V$ (3.13)

(iv) $Min\left[M\left(\bar{y}_{d2}\right)\right]_{II} = \left[\delta_4 - (\delta_3 + \delta_4)^{-1} \delta_4^2 \rho^2\right] S_Y^2$

when $P = -\delta_4 V / (\delta_3 + \delta_4)$ (3.14)

Proof:

(i) $\frac{d}{dP}\left[M\left(\bar{y}_{d1}\right)_I\right] = 0 \Rightarrow P = -\rho \frac{C_Y}{C_X} = -V$ and using this in (3.7)

$$Min\left[M\left(\bar{y}_{d1}\right)\right]_I = \left[\delta_1 - (\delta_2 - \delta_3)\rho^2\right] S_Y^2$$

(ii) $\frac{d}{dP}\left[M\left(\bar{y}_{d1}\right)_{II}\right] = 0 \Rightarrow P = -\delta_5 V (\delta_3 + \delta_5)^{-1}$ and using this in (3.8)

$$Min\left[M\left(\bar{y}_{d1}\right)\right]_{II} = \left[\delta_4 - (\delta_3 + \delta_5)^{-1} \delta_5^2 \rho^2\right] S_Y^2$$

(iii) $\frac{d}{dP}\left[M\left(\bar{y}_{d2}\right)_I\right] = 0 \Rightarrow P = -\rho \frac{C_Y}{C_X} = -V$

$$Min\left[M\left(\bar{y}_{d2}\right)\right]_I = \left[\delta_1 - (\delta_1 - \delta_3)\rho^2\right]S_Y^2$$

(iv) $\dfrac{d}{dP}\left[M\left(\bar{y}_{d2}\right)_{II}\right] = 0 \Rightarrow P = -\delta_4 V\left(\delta_3 + \delta_4\right)^{-1}$

$$Min\left[M\left(\bar{y}_{d2}\right)\right]_{II} = \left[\delta_4 - (\delta_3 + \delta_4)^{-1}\delta_4^2\rho^2\right] S_Y^2$$

Lemma 3.0 [By Shukla (2002)]:
F-T estimator in two-phase sampling (without imputation) is

$$\left(\bar{y}_d\right)_w = \bar{y}\left[\frac{(A+C)\bar{x}' + fB\bar{x}}{(A+fB)\bar{x}' + C\bar{x}}\right] \tag{3.15}$$

With optimum MSE conditions

under design F_1 [Case I]: $P = -V$ (3.16)

under design F_2 [Case I]: $P = -V\left(1+\delta\right)^{-1}$ (3.17)

and optimum MSE expressions

$$opt\left[M\left(\bar{y}_d\right)_w\right]_I = \bar{Y}^2 V_{20}\left[1 - \rho^2\left(1-\delta\right)\right] \tag{3.18}$$

$$opt\left[M\left(\bar{y}_d\right)_w\right]_{II} = \bar{Y}^2 V_{20}\left[1 - \rho^2\left(1+\delta\right)^{-1}\right] \tag{3.19}$$

where $\delta = \left(\dfrac{1}{n'} - \dfrac{1}{N}\right)\left(\dfrac{1}{n} - \dfrac{1}{N}\right)^{-1}$;

$$V_{ij} = E\left[\left(\bar{y} - \bar{Y}\right)^i \left(\bar{x} - \bar{X}\right)^j\right] \Big/ \bar{Y}^i \bar{X}^j; \quad i = 0,1,2;\ j = 0,1,2$$

4. Appropriate Choice of *k* for Bias Reduction

(i) $B\left[\bar{y}_{d1}\right]_I = 0 \Rightarrow P\left(\theta_4 C_X^2 - \rho C_Y C_X\right) = 0$

If $P = 0 \Rightarrow (k-4)\left[k^2 - (5+f)k + (6+f)\right] = 0$ (4.1)

$$\left.\begin{aligned} &\text{and} \quad k = k_1 = 4 \\ &k = k_2 = \frac{1}{2}\left[(5+f) + \left(f^2 + 6f + 1\right)^{1/2}\right] \\ &k = k_3 = \frac{1}{2}\left[(5+f) - \left(f^2 + 6f + 1\right)^{1/2}\right] \end{aligned}\right\} \tag{4.2}$$

If $\theta_4 C_X^2 - \rho C_Y C_X = 0$

$$\Rightarrow AV + VfB + (V-1)C = 0 \tag{4.3}$$

$$\Rightarrow (V-1)k^3 - [(8-f)V - 9]k^2 - [(23+5f)V + 26]k - 2[(11-2f)V - 12] = 0 \tag{4.4}$$

(ii) $B\left[\bar{y}_{d1}\right]_{II} = 0 \Rightarrow P\left[(\theta_3\delta_3 - \theta_4\delta_5)C_X^2 + \delta_5\rho C_Y C_X\right] = 0$

If $P = 0$ we have solution as per (4.2) and

if $(\theta_3\delta_3 - \theta_4\delta_5)C_X + \delta_5\rho C_Y = 0$

$$\Rightarrow (V-1)k^3 + \left[\left(\frac{\delta_3}{\delta_5} + V\right)(1+f) - 9(V-1)\right]k^2 - \left[\left(\frac{\delta_3}{\delta_5} + V\right)(3+5f) - 26(V-1)\right]k$$

$$+ 2\left[\left(\frac{\delta_3}{\delta_5} + V\right)(1+2f) - 12(V-1)\right] = 0 \tag{4.5}$$

(iii) $B\left[\bar{y}_{d2}\right]_I = 0$ provides similar solution as in (i).

(iv) $B\left[\bar{y}_{d2}\right]_{II} = 0 \Rightarrow P\left[(\theta_3\delta_3 - \theta_4\delta_4)C_X^2 + \delta_4\rho C_Y C_X\right] = 0$

if $P = 0$ we have solution as per (4.2) and

if $(\theta_3\delta_3 - \theta_4\delta_4)C_X^2 + \delta_4\rho C_Y C_X = 0$

$$\Rightarrow [(A + fB)\delta_3 - C\delta_4] = -\delta_4 V(A + fB + C)$$

$$\Rightarrow (V-1)k^3 + \left[\left(\frac{\delta_3}{\delta_4} + V\right)(1+f) - 9(V-1)\right]k^2 - \left[\left(\frac{\delta_3}{\delta_4} + V\right)(3+5f) - 26(V-1)\right]k$$

$$+ 2\left[\left(\frac{\delta_3}{\delta_4} + V\right)(1+2f) - 12(V-1)\right] = 0 \tag{4.6}$$

5. Comparison of the Estimators

(i) $\Delta_1 = \min\left[M\left(\bar{y}_{d1}\right)_I\right] - \min\left[M\left(\bar{y}_{d2}\right)_I\right] = \left(\frac{1}{r} - \frac{1}{N}\right)\rho^2 S_Y^2$

$\left(\bar{y}_{d2}\right)_I$ is better than $\left(\bar{y}_{d1}\right)_I$ if $\Delta_1 > 0 \Rightarrow N > r$ which is always true.

(ii) $\Delta_2 = \min\left[M\left(\bar{y}_{d1}\right)_{II}\right] - \min\left[M\left(\bar{y}_{d2}\right)_{II}\right]$

$$= \left(\frac{\delta_4^2}{(\delta_3 + \delta_4)} - \frac{\delta_5^2}{(\delta_3 + \delta_5)}\right)\rho^2 S_Y^2$$

$\left(\bar{y}_{d2}\right)_{\text{II}}$ is better than $\left(\bar{y}_{d1}\right)_{\text{II}}$ if $\Delta_2 > 0$

$\Rightarrow (n-r)\left[N^3 - \left(n'n + n'r + nr\right)N + 2n'nr\right] > 0$

which generates two options as

(A) when $(n-r) > 0 \Rightarrow n > r$ and

(B) $\left[N^3 - \left(n'n + n'r + nr\right)N + 2n'nr\right] > 0$

if $n' \approx N$ $\left[i.e.\ n' \to N\right]$

then $N\left[N^2 - (n-r)N + nr\right] > 0$ (since $N > 0$ always)

$\Rightarrow (N-n)(N-r) > 0$

$\Rightarrow (N-n) > 0 \Rightarrow N > n$ and $N - r > 0 \Rightarrow N > r$

The ultimate is $N > n > r$, which is always true.

(iii) $\Delta_3 = \min\left[M\left(\bar{y}_{d2}\right)_I\right] - \min\left[M\left(\bar{y}_{d2}\right)_{II}\right]$

$$= \frac{(\delta_1 - \delta_4)(\delta_3 + \delta_4) + \left(\delta_4^2 + \delta_3^2 - \delta_1\delta_3 - \delta_1\delta_4 + \delta_3\delta_4\right)}{(\delta_3 + \delta_4)} \rho^2 S_Y^2$$

$\left(\bar{y}_{d2}\right)_{\text{II}}$ is better than $\left(\bar{y}_{d2}\right)_{\text{I}}$, if $\Delta_3 > 0$

$$\Rightarrow \rho^2 > \left[\frac{1+m}{1+2m}\right] \text{ where } m = \left[\frac{r\left(N-n'\right)}{n'\left(N-r\right)}\right]$$

$$\Rightarrow -1 < \rho < -\sqrt{\frac{1+m}{1+2m}} \quad \text{or} \quad \sqrt{\frac{1+m}{1+2m}} < \rho < 1$$

6. Empirical Study

The attached appendix A has an artificial population of size N = 200 [see Appendix A] containing values of main variable Y and auxiliary variable X. Parameters of this are given in table 6.1.

Table 6.1. Population Parameters

$\bar{Y}$	$\bar{X}$	S_Y^2	S_X^2	ρ	C_X	C_Y	$V = \rho\frac{C_Y}{C_X}$
42.485	18.515	199.0598	48.5375	0.8652	0.3763	0.3321	0.7635

Under design-I, we draw a preliminary random sample S' of size $n' = 110$ to compute $\bar{x}'$ and further draw a random sample S of size $n = 50$ such that

$S \subset S'$ by *SRSWOR*. The V is a stable quantity over time and assumed to be known [see Reddy (1978)].

Table 6.2.

Design	Optimum Condition for *MSE*	Three optimum values of k on one condition		
I	$P = -V$	$k_1 = 1.5206$	$k_2 = 2.4505$	$k_3 = 8.9456$
	$P = -\delta_5 V/(\delta_3 + \delta_5)$	$k_4 = 1.5880$	$k_5 = 2.8768$	$k_6 = 6.4279$
II	$P = -V$	$k_7 = 1.5206 = k_1$	$k_8 = 2.4505 = k_2$	$k_9 = 8.9456 = k_3$
	$P = -\delta_4 V/(\delta_3 + \delta_4)$	$k_{10} = 1.5645$	$k_{11} = 2.8572$	$k_{12} = 6.7221$

7. Simulation

The bias and optimum m.s.e. of proposed estimators under both designs are computed through 50,000 repeated samples n, n' as per design. Computations are in table 7.1 where efficiency loss measurement due to imputation is as $LI_t(\bar{y}_s) = \dfrac{Opt\left[M(\bar{y}_s)_t\right]}{Opt\left[M(\bar{y}_d)_w\right]}$ with $Opt\left[M(\bar{y}_s)_t\right]$ the optimum mean squared error of estimator $\bar{y}_s$,

$s = d, d_1, d_2; t = I, II, t = w =$ without imputation.

For design I and II the simulation procedure has following steps :

Step 1: Draw a random sample S' of size $n' = 110$ from the population of N = 200 by SRSWOR.

Step 2: Draw a random sub-sample of size $n = 50$ from S' for design *I* and independent random sample n= 50 from $(N - n')$ for design *II*.

Step 3: Drop down 5 units randomly from each second sample corresponding to Y in both *I* and *II*.

Step 4: Impute dropped units of Y by proposed methods and available methods and compute the relevant statistic.

Step 5: Repeat the above steps 50,000 times, which provides multiple sample based estimates $(\hat{y}_{1s})_t, (\hat{y}_{2s})_t, (\hat{y}_{3s})_t \ldots (\hat{y}_{50000s})_t$ for estimators $(\bar{y}_{d_1})_t$, $(\bar{y}_{d_2})_t$ and $(\bar{y}_d)_w$.

Step 6: Bias of $(\hat{y}_s)_t$ is $B(\hat{y}_s)_t = \dfrac{1}{50000}\sum_{i=1}^{50000}\left[(\hat{y}_{is})_t - \bar{Y}\right]$

Step 7: *M.S.E.* of $(\hat{y}_s)_t$ is $M(\hat{y}_s)_t = \frac{1}{50000}\sum_{i=1}^{50000}\left[(\hat{y}_{is})_t - \bar{Y}\right]^2$

Step 8: The efficiency comparisons are

Design efficiency $E_1 = \frac{M(\bar{y}_{d1})_I}{M(\bar{y}_{d1})_{II}} \times 100$;

Design efficiency $E_2 = \frac{M(\bar{y}_{d2})_I}{M(\bar{y}_{d2})_{II}} \times 100$

Estimator efficiency $E_3 = \frac{M(\bar{y}_{d1})_I}{M(\bar{y}_{d2})_I} \times 100$;

Estimator efficiency $E_4 = \frac{M(\bar{y}_{d1})_{II}}{M(\bar{y}_{d2})_{II}} \times 100$

Table 7.1. Bias and Mean Squared Error

Opt (k)	Design F_1				Design F_2			
	$(\bar{y}_{d_1})_{k_i}$		$(\bar{y}_{d_2})_{k_i}$		$(\bar{y}_{d_1})_{k_i}$		$(\bar{y}_{d_2})_{k_i}$	
	Bias	*MSE*	*Bias*	*MSE*	*Bias*	*MSE*	*Bias*	*MSE*
$k_1 = 1.3813$	-0.1314	2.5947	-0.2855	2.1727	0.3665	2.7997	0.3680	3.0111
$k_2 = 2.7576$	-0.1780	2.6282	-0.3339	2.2192	0.1485	2.7801	0.1287	2.9542
$k_3 = 9.9538$	-0.1350	2.5968	-0.2892	2.1758	0.3498	2.7911	0.3497	2.9982
$k_4 = 1.5880$	-0.0895	2.2549	-0.2892	2.1758	0.4131	3.7278	0.3680	3.0110
$k_5 = 2.8768$	-0.0855	2.2706	-0.2268	1.9645	0.2664	3.5517	0.2499	3.4296
$k_6 = 6.4279$	-0.0951	2.2580	-0.2114	1.9333	0.3832	3.7016	0.3778	3.5901
$k_7 = 1.3813$	-0.1314	2.5947	-0.2855	2.1727	0.3665	2.7997	0.3680	3.0111
$k_8 = 2.7576$	-0.1780	2.6282	-0.3339	2.2192	0.1485	2.7801	0.1287	2.9542
$k_9 = 9.9538$	-0.1350	2.5968	-0.2892	2.1758	0.3498	2.7911	0.3497	2.9982
$k_{10} = 1.5645$	-0.0953	2.2743	-0.2082	1.9476	0.4044	3.4592	0.4024	3.3839
$k_{11} = 2.8572$	-0.1287	2.2961	-0.2429	1.9705	0.2473	3.2895	0.2301	3.1970
$k_{12} = 6.7221$	-0.1015	2.7780	-0.2146	1.9515	0.3756	3.4273	0.3707	3.3472

Table 7.2. Estimator without Imputation [Lemma 3.0]

	Optimum k - values			
	k_1	k_2	k_3	Optimum *MSE*
Case - *I*	1.3813	2.7576	9.9538	1.9021
Case - *II*	1.5311	2.8337	7.1604	2.7702

Table 7.3. Loss due to Imputation $LI_t[\ ]$

Opt (k)	Design F_1		Design F_2	
	$LI_I(\bar{y}_{d1})_k$	$LI_I(\bar{y}_{d2})_k$	$LI_{II}(\bar{y}_{d1})_k$	$LI_{II}(\bar{y}_{d2})_k$
$k_1 = 1.3813$	1.364123863	1.142263814	1.01064905	1.08696123
$k_2 = 2.7576$	1.381735976	1.166710478	1.00357375	1.0664212
$k_3 = 9.9538$	1.365227906	1.143893591	1.00754458	1.08230453
$k_4 = 1.5880$	1.185479207	1.143893591	1.34567901	1.08692513
$k_5 = 2.8768$	1.193733242	1.032805846	1.28210959	1.23803335
$k_6 = 6.4279$	1.187108985	1.016402923	1.33622121	1.29597141
$k_7 = 1.3813$	1.364123863	1.142263814	1.01064905	1.08696123
$k_8 = 2.7576$	1.381735976	1.166710478	1.00357375	1.0664212
$k_9 = 9.9538$	1.365227906	1.143893591	1.00754458	1.08230453
$k_{10} = 1.5645$	1.195678461	1.023920929	1.2487185	1.22153635
$k_{11} = 2.8572$	1.207139477	1.035960254	1.18745939	1.1540683
$k_{12} = 6.7221$	1.460491036	1.025971295	1.23720309	1.20828821

Table 7.4. Efficiency Comparisons E_1, E_2, E_3, E_4

Opt (*k*)	E_1	E_2	E_3	E_4
$k_1 = 1.3813$	92.67%	72.15%	119.42%	92.97%
$k_2 = 2.7576$	94.53%	75.12%	118.43%	94.10%
$k_3 = 9.9538$	93.03%	72.57%	119.34%	93.09%
$k_4 = 1.5880$	60.48%	72.26%	103.63%	123.80%
$k_5 = 2.8768$	63.92%	57.28%	115.58%	103.56%
$k_6 = 6.4279$	61.00%	53.85%	116.79%	103.10%
$k_7 = 1.3813$	92.67%	72.15%	119.42%	92.97%
$k_8 = 2.7576$	94.53%	75.12%	118.43%	94.10%
$k_9 = 9.9538$	93.03%	72.57%	119.34%	93.09%
$k_{10} = 1.5645$	65.74%	54.24%	116.77%	102.22%
$k_{11} = 2.8572$	69.80%	61.63%	116.52%	102.89%
$k_{12} = 6.7221$	81.05%	58.30%	142.35%	102.39%

8. Almost Unbiased Imputation Methods

Using equations of section 4.0 we get

From (i) $k_1^{'} = 4$; $k_2^{'} = 3.4253$; $k_3^{'} = 1.8246$; $k_4^{'} = 0.1812$.

From (ii) $k_1^{'} = 4$; $k_2^{'} = 3.4253$; $k_3^{'} = 1.8246$; $k_4^{'} = 1.4236$; $k_5^{'} = 2.4469$; $k_6^{'} = 10.7864$. From (iii) similar to (i).

From (iv) $k_1^{'} = 4$; $k_2^{'} = 3.4253$; $k_3^{'} = 1.8246$; $k_4^{'} = 1.4339$; $k_5^{'} = 2.4488$; $k_6^{'} = 10.5426$

Using these *k*-values we can make proposed F-T imputation strategies almost unbiased. The best among them will be that having the lowest m.s.e. By this we have option to choose almost unbiased estimator with a control over mean squared error.

9. Discussion and Conclusion

The proposed estimators are found useful for situation when some observations are missing in the sample. As per table 7.3 for $\bar{y}_{d_1}$ and $\bar{y}_{d_2}$ under design F_1, the efficient performance of both is found when k = 1.5880, 1.5645, 6.4279 and 6.7221. On these specific choices the loss of efficiency with respect to without imputation is very low. Similarly, for $\bar{y}_{d_1}$ and $\bar{y}_{d_2}$ under design F_2, the efficient performance observed at k = 1.3813, 2.7576, 9.9538. It seems even by adopting imputation, the suggested estimators are loosing a little in terms of relative m.s.e. to the without imputation usual F-T estimator.

While mutual comparisons are in table 7.4, the design F_2 is uniformly efficient as F_1 at all the optimum *k*-values, over both suggested F-T strategies. Within F_1, the estimator $\bar{y}_{d_2}$ is more efficient than $\bar{y}_{d_1}$ whereas within F_2 it does not hold uniformly for all *k*-optimals. The $\bar{y}_{d_2}$ under F_2 found better when k = 1.2, 2.8, 6.4 and 6.7. One can get almost unbiased estimators also on choices k = 0.1812, 1.4236, 1.4339, 1.8246, 2.4469, 2.4488, 3.4253, 4, 10.5426, 10.7864. The most suitable will be that which has the lowest m.s.e. So these suggested strategies are almost unbiased with a reducing control over m.s.e. also.

Appendix A

Population (N = 200)

Y_i	45	50	39	60	42	38	28	42	38	35
X_i	15	20	23	35	18	12	8	15	17	13
Y_i	40	55	45	36	40	58	56	62	58	46
X_i	29	35	20	14	18	25	28	21	19	18
Y_i	36	43	68	70	50	56	45	32	30	38
X_i	15	20	38	42	23	25	18	11	09	17
Y_i	35	41	45	65	30	28	32	38	61	58
X_i	13	15	18	25	09	08	11	13	23	21
Y_i	65	62	68	85	40	32	60	57	47	55
X_i	27	25	30	45	15	12	22	19	17	21
Y_i	67	70	60	40	35	30	25	38	23	55
X_i	25	30	27	21	15	17	09	15	11	21
Y_i	50	69	53	55	71	74	55	39	43	45
X_i	15	23	29	30	33	31	17	14	17	19
Y_i	61	72	65	39	43	57	37	71	71	70
X_i	25	31	30	19	21	23	15	30	32	29
Y_i	73	63	67	47	53	51	54	57	59	39
X_i	28	23	23	17	19	17	18	21	23	20
Y_i	23	25	35	30	38	60	60	40	47	30
X_i	07	09	15	11	13	25	27	15	17	11
Y_i	57	54	60	51	26	32	30	45	55	54
X_i	31	23	25	17	09	11	13	19	25	27
Y_i	33	33	20	25	28	40	33	38	41	33
X_i	13	11	07	09	13	15	13	17	15	13
Y_i	30	35	20	18	20	27	23	42	37	45
X_i	11	15	08	07	09	13	12	25	21	22
Y_i	37	37	37	34	41	35	39	45	24	27
X_i	15	16	17	13	20	15	21	25	11	13
Y_i	23	20	26	26	40	56	41	47	43	33
X_i	09	08	11	12	15	25	15	25	21	15
Y_i	37	27	21	23	24	21	39	33	25	35
X_i	17	13	11	11	09	08	15	17	11	19
Y_i	45	40	31	20	40	50	45	35	30	35
X_i	21	23	15	11	20	25	23	17	16	18
Y_i	32	27	30	33	31	47	43	35	30	40
X_i	15	13	14	17	15	25	23	17	16	19
Y_i	35	35	46	39	35	30	31	53	63	41
X_i	19	19	23	15	17	13	19	25	35	21
Y_i	52	43	39	37	20	23	35	39	45	37
X_i	25	19	18	17	11	09	15	17	19	19

REFERENCES

AHMED, M. S., AL-TITI, O., AL-RAWI, Z. and ABU-DAYYEH, W. (2006): *Estimation of a population mean using different imputation methods,* Statistics in Transition, 7, 6, 1247–1264.

COCHRAN, W. G. (2005): *Sampling Techniques,* John Wiley and Sons, Fifth Edition New York.

HEITJAN, D. F. and BASU, S. (1996): *Distinguishing 'Missing at random' and 'missing completely at random'*, The American Statistician, 50, 207–213.

RAO, J. N. K. and SITTER, R. R. (1995): *Variance estimation under two-phase sampling with application to imputation for missing data*, Biometrica, 82, 453–460.

RUBIN, D. B. (1976): *Inference and missing data,* Biometrica, 63, 581–593.

SHUKLA, D. (2002): *F-T estimator under two-phase sampling*, Metron, 59, 1–2, 253–263.

SHUKLA, D., SINGH, V. K. and SINGH, G. N. (1991): *On the use of transformation in factor type estimator,* Metron, 49 (1–4), 359–361.

SINGH, S. and HORN, S. (2000): *Compromised imputation in survey sampling,* Metrika, 51, 266–276.

SINGH, V. K. and SHUKLA, D. (1987): *One parameter family of factor type ratio estimator,* Metron, 45, 1–2, 273–283.

SINGH, V. K. and SHUKLA, D. (1993): *An efficient one parameter family of factor - type estimator in sample survey*, Metron, 51, 1–2, 139–159.

SINGH, V. K. and SINGH, G. N. (1991): *Chain type estimator with two auxiliary variables under double sampling scheme*, Metron, 49, 279–289.

REDDY, V. N. (1978): *A study on the use of prior knowledge on certain population parameters in estimation*, Sankhya, C, 40, 29–37.

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 415—435

FULL INFORMATION EFFICIENT ESTIMATOR OF FINITE POPULATION VARIANCE

Manoj Kumar Srivastava[1], Namita Srivastava[2] , Housila P. Singh[3]

ABSTRACT

Second order or quadratic and finite population parametric functions may be expressed as total of variable-values on pairs of units in a derived population. Recently, Sitter and Wu (2002) utilized this approach for estimating variance under model calibration. In this paper an efficient design based full-information estimator of finite population variance has been suggested. The exact expression of its variance and its relative efficiency has also been derived. Finally, the proposed estimator has been shown to be superior to its competitors in an empirical investigation.

Key words: design based estimation; variance estimation; Rao-Blackwellization in survey sampling; estimation of polynomial finite population function.

1. Introduction

In finite population theory the problem of estimating quadratic or higher order finite population functions is an extensively explored area of research. Efficient estimators for finite population variances, covariance between two response variables or variance of linear estimators are highly desirable. Efficient estimation of these quadratic functions have been prohibitive (Sitter and Wu 2002), because of complex expressions of variances of these estimators. Liu (1974 a) suggested several design based estimators of finite population variance following which Chaudhuri (1978) noted that many of Liu's estimators sometimes take negative values. He then suggested alternatively few non-negative estimators and discussed their statistical properties. Later, Liu and Thompson (1983) showed nonexistence of the best unbiased estimator of population variance and asserted

[1] Associate Professor, Department of Statistics, Institute of Social Sciences, Dr. BRA University, Agra-282002.U.P. India, email : mks_iss@yahoo.co.in.
[2] Associate Professor, Department of Statistics, St. John's College, Agra-282002, U.P. India. email: drnamita.sjc@gmail.com.
[3] Professor, School of Studies in Statistics, Vikram University, Ujjain-456010, M.P. India, email: hpsujn@ rediffmail.com.

that the admissibility holds under certain restrictions. Swain and Mishra (1994) suggested more efficient estimator of population variance as compared to Liu's and Chaudhary's estimators and tried it on various natural and artificial populations. This estimator still suffers from a drawback that it takes negative values occasionally. Mukhopadhyay (1978) suggested model based predictors of population variance following the Royall's prediction approach. Mukhopadhyay and Bhattacharya (1989) suggested optimal estimators of population variance under regression superpopulation models. Sitter and Wu (2002) for the first time suggested very efficient model calibration estimators which were asymptotically design unbiased under linear and non-linear models. They suggested pseudoempirical likelihood estimators also which are free from taking negative values.

Bayesian estimation of finite population variance was first taken up by Ericson (1969) in his pioneering work on Bayesian estimation in survey sampling although the issue of estimation of population variance was briefly discussed. Subsequent references are Liu (1974 b), Chaudhuri (1978), Zacks and Soloman (1981), Ghosh and Meeden (1983, 1984),

Vardeman and Meeden (1983), Ghosh and Lahiri (1987), Lahiri and Tiwari (1991), Datta and Tiwari (1991), Datta and Ghosh (1993) among others.

Hanurav (1966) was first who attempted simplification of the problem of estimation of population variance by expressing it as a total of a function over a new population, whose elements are ordered pairs of units of the original population. Using this approach, Liu and Thompson (1983) showed nonexistence of best unbiased estimators of population variance and gave some admissibility results. More recently, Sitter and Wu (2002) named the population of pairs as "synthetic population" and under the Hanurav's approach they discussed model calibration and pseudoempirical likelihood estimators which were model assisted and asymptotically design unbiased under a general sampling design. We have suggested an efficient estimator named as full-information estimator of population variance for equal and unequal probability sampling design under the above set up by drawing samples from "synthetic population". We have used this estimator on a population of 124 countries (CO124, Sarn-dal, Swensson, Wretmen 1992) for estimating the variance of imports among 124 countries, and have shown that the proposed estimator is more efficient as compared to its competitors in terms of relative efficiency (RE).

1.1.Notations

We denote population by $U = \{U_1, U_2, \cdots, U_N\}$; and the population of ordered pairs by $U' = \{\mathbf{U}'_k : \mathbf{U}'_k = (U_{i1}, U_{i2}), U_{i1}, U_{i2}$ are units taken from U so that $i_1 < i_2\}$. The value of the study variable y on U_i be denoted by y_{i1} and on $\mathbf{U}'_\mathbf{k}$ be denoted

by (y_{i1}, y_{i2}). Define, for a real symmetric function h on (y_{i1}, y_{i2}), $t_k = h(y_{i1}, y_{i2})$ for some k in U', $\overline{T}_{U'} = \frac{1}{N'}\sum_{k=1}^{N'} t_k$, $N' = \binom{N}{2}$. $\overline{T}_{U'}$ for different h functions:

1. For $t_k(y) = \frac{1}{2}(y_i - y_j)^2, k = (i, j)$, we get $\overline{T}_{U'} = \frac{1}{N-1}\sum_U (y_i - \overline{y})^2 = S_y^2$: finite population variance.

2. For $t_k(y, z) = \frac{1}{2}(y_i - y_j)(z_i - z_j) = C_{yz}$, we get $\overline{T}_{U'} = \frac{1}{N-1}\sum_U (y_i - \overline{y})(z_i - \overline{z})$: population covariance.

3. For $t_k(y) = \frac{N(N-1)}{2}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$, we get $\overline{T}_{U'} = \sum_{i<j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 = V_{YG}$, which is Yates and Grundy-type variance of Horvitz Thompson (HT) estimator of a population total, $\hat{Y}_{HT} = \sum_s \frac{y_i}{\pi_i}$

4. For $t_k = \frac{|y_i - y_j|}{2\overline{y}}$, gives $\overline{T}_{U'} = \sum_{i<j\in U}\frac{|y_i - y_j|}{N(N-1)\overline{y}} = G$, which is population Gini-coefficient.

Note that in (1) the finite population variance S_y^2 has been expressed as mean of t_k values over the population U'.

2. Estimation under unequal probability design

Associated with U_i let x_i be the size measure so that $y_i \propto x_i$; $P_i = x_i/x$, $x = \sum_U x_k$. Note for the function h in (1) is such that $t_k(\mathbf{y}) \propto t_k(\mathbf{x})$; $t_k(\mathbf{x}) = \frac{1}{2}(x_i - x_j)^2$ whenever $y_i \propto x_i$. Let the size measures on U' be $t_k(\mathbf{x})$, so that $P'(U_k') = \frac{t_k(\mathbf{x})}{\sum_{U'} t_k(\mathbf{x})}$. We consider here a probability proportional to size design with replacement with sample size *n(s) = n.* Let the corresponding sample space be denoted by S'. Denote this design by $D'(U', S', P')$. Based on ppswr sample s_n of size n and the corresponding data $d = \{(\mathbf{U}_k', t_k) : \mathbf{U}_k' \in s_n\}$, the unbiased estimator of $\overline{T}_{U'}(S_y^2)$ is given by

$$\hat{\bar{T}}_{s_n} = \frac{1}{n}\sum_{s_n}\frac{t_k}{N'P'_k} \tag{2.1}$$

The variance of $\hat{\bar{T}}_{s_n}$ is given by

$$V\left(\hat{\bar{T}}_{s_n}\right) = \frac{1}{n}\sum_{U'}P'_kP'_l\left(\frac{t_k}{N'P'_k}-\frac{t_l}{N'P'_l}\right)^2 \tag{2.2}$$

Where $\sum_{U'}$ is summation over all different pairs of 2-tuples in U'. Furthermore the estimator of $V(\cdot)$ is given by

$$v\left(\hat{\bar{T}}_{s_n}\right) = \frac{1}{n^2(n-1)}\sum_{s_n}\left(\frac{t_k}{N'P'_k}-\frac{t_l}{N'P'_l}\right)^2 \tag{2.3}$$

Where the summation $\sum_{s_n}$ is over all pairs of 2-tuples in the sample s_n. Let us now consider a reducing transformation on s_n that ignores order and multiplicity of 2-tuples and results into a sample $s_{n'}$ and the corresponding data be denoted by $d^* = \{(\mathbf{U}'_k, t_k) : \mathbf{U}'_k \in s_{n'}\}$. Since the sample $s_{n'}$ is a sufficient statistic (Cassel, Sarndal, and Wretman 1976), therefore, by using Rao-Blackwellization, we get an improved unbiased estimator $\hat{\bar{T}}^*_{s_{n'}} = E^{S_n|s_{n'}}\left\{\hat{\bar{T}}_{s_n} \mid D^* = d^*\right\}$ since $E^{S_{n'}}\left\{\hat{\bar{T}}^*_{s_{n'}}\right\} = E^{S_{n'}}E^{S_n|S_{n'}}\{.\} = \bar{T}_{u'}$ with variance

$$V\left(\hat{\bar{T}}^*_{s_{n'}}\right) = E^{S_{n'}}\left\{E^{S_n|S_{n'}}\left\{\hat{\bar{T}}_{s_n} \mid D^* = d^*\right\}\right\}^2 - \bar{T}^2_{u'}$$

$$= \sum_{S_{n'}}\left[\sum_{\substack{s_n\in S' \\ \text{s.t. } s_n \text{ gives } s_{n'}}}\hat{\bar{T}}_{s_n}\cdot\frac{P'(s_n)}{P'(s_{n'})}\right]^2 P'(s_{n'}) - \bar{T}^2_{u'}$$

Where $P'(s_{n'}) = \sum_{\substack{s_n\in S' \\ \text{s.t. } s_n \text{ gives } s_{n'}}} P'(s_n)$ further $V\left\{\hat{\bar{T}}^*_{s_{n'}}\right\} \le V\left\{\hat{\bar{T}}_{s_n}\right\}$.

While utilizing information on already selected 2-tuples in a ppswr sample we get some additional information on few more 2-tuples. For example if $\mathbf{U}'_{\mathbf{i}_1} = (U_1, U_2)$ and $\mathbf{U}'_{\mathbf{i}_2} = (U_1, U_3)$ are selected in a sample, observing these 2-tuples also means observing U_1, U_2 and U_3. So we have effectively derived information on $\mathbf{U}'_{\mathbf{i}_3}$ in addition to information on $\mathbf{U}'_{\mathbf{i}_1}$ and $\mathbf{U}'_{\mathbf{i}_2}$ where, $\mathbf{U}'_{\mathbf{i}_3} = (U_2, U_3)$. We expect that this information on $\mathbf{U}'_{\mathbf{i}_3}$ could also be utilized to

get a better estimator though it was not directly selected. Let the sample containing information on $\mathbf{U}'_{\mathbf{i}_1}, \mathbf{U}'_{\mathbf{i}_2}, \mathbf{U}'_{\mathbf{i}_3}$ be called full-information sample denoted by $s\binom{n''}{2}$ and subset of n'' distinct units be denoted by $A_{n''} = \{U_{i_1}, U_{i_2}, \cdots, U_{n''}\} = \{U_1, U_2, U_3\}$ (as in the above example). $A_{n''}$ could be derived from ppswr sample s_n or by $s_{n'}$ (with their usual notations) by a reducing function on s_n, which gives only distinct units appeared on 2-tuples in $s_n\,(s_{n'})$. The probability distribution of such a random set $A_{n''}$ has been stated in the theorem 2.1. We have $P'(s_n) = P'\left(\mathbf{U}'_{\mathbf{i}_1}\right).P'\left(\mathbf{U}'_{\mathbf{i}_2}\right)....P'\left(\mathbf{U}'_{\mathbf{i}_n}\right)$. Define a random set

$$A_{n''} = \mathbf{U}'_{\mathbf{i}_1} \cup \mathbf{U}'_{\mathbf{i}_2} \cup \cdots \cup \mathbf{U}'_{\mathbf{i}_n}$$

with a meaning that $A_{n''}$ is a set of distinct units U_i's appearing in one or more $\mathbf{U}'_{\mathbf{i}_j}$. Clearly, $A_{n''}$ is a random set in the language of probability theory, taking its values from *U* as a subset. Let us denote the number of distinct units in $A_{n''}$ by $|\,A_{n''}|$, this number is also an integer-valued random variable. To discuss the statistical properties of the estimator based on $A_{n''}$ we would be interested in the probability distribution of the random set $A_{n''}$. Consider a set *B* of *M* units being fixed from *U* as a subset, namely, $U_{l_1} \cup U_{l_2} \cup \cdots \cup U_{l_M}$; generate a set of $\binom{M}{2} = M'$ 2-tuples purely made of units in *B,* denote it by $U'_B, |\,B| = M, 1 \le M \le N$, clearly $U'_B \subset U'$. If one draws probability proportional to size samples of size n from U'_B with replacement (ppswr), the corresponding sample space be given by S'_B. Note that S'_B is a collection of all 2-tuples made up entirely of the units (labels) in the subset *B* so that $S'_B \subseteq S'$. The probability of any such sample $S_n^{(B)} = \left\{\mathbf{U}'_{\mathbf{j}_1}, \mathbf{U}'_{\mathbf{j}_2} \cdots, \mathbf{U}'_{\mathbf{j}_n}\right\}$, each $\mathbf{U}'_{\mathbf{j}_i}$ is some 2-tuple made up of units in *B,* under the design D'

$$P'\left(s_n^{(B)}\right) = P'\left(\mathbf{U}_{\mathbf{j}_1}\right).P'\left(\mathbf{U}_{\mathbf{j}_2}\right).....P'\left(\mathbf{U}_{\mathbf{j}_n}\right)$$

where, $\mathbf{U}_{\mathbf{j}_i}$'s are not necessarily distinct. The probability distribution of the set $A_{n''}$ is given by :

Theorem 2.1. Consider a probability proportional to size sample s_n of size $n\,(s_n) = n$ drawn from U' with replacement under the ppswr design $D'(U', S', P')$. Let the constituent 2-tuples in s_n be denoted by $\mathbf{U}'_{\mathbf{i}_1}, \mathbf{U}'_{\mathbf{i}_2}, ..., \mathbf{U}'_{\mathbf{i}_n}$. Define a set valued random variable taking its values from U,

$$A_{n''} = \mathbf{U}'_{\mathbf{i}_1} \cup \mathbf{U}'_{\mathbf{i}_2} \cup \ldots\ldots \cup \mathbf{U}'_{\mathbf{i}_n}$$

as *the set of distinct units common to* $\mathbf{U}'_{\mathbf{i}_j}$*'s. The size of* $A_{n''}$*, i.e.*$|\ A_{n''}|$ *would range from* $\underline{k} = 2$ *to* $\overline{k} = \min\{2n, N\}$. *The probability distribution of* $A_{n''}$ *is given by*

$$P'(A_{n''})|_{\substack{A_{n''}=B, \\ |B|=\underline{k}}} = \sum_{S'_B} P'\left(s_n^{(B)}\right)$$

for every B of size $\underline{k}$. Next

$$P'(A_{n''})|_{\substack{A_{n''}=B, \\ |B|=\underline{k}+1}} = \sum_{S'_B} P'\left(s_n^{(B)}\right) - \sum_{\substack{\{A_{n''}:|\ A_n|=\underline{k}, \\ A_{n''}\subset B\}}} P'(A_{n''})$$

for every B of size $\underline{k}+1$, where, $P'(A_{n''})$ in the second term on the right hand side of the above expression comes from the previous step. Proceeding similarly, for $k = \underline{k} + j$

$$P'(A_{n''})|_{\substack{A_{n''}=B, \\ |B|=\underline{k}+j}} = \sum_{S'_B} P'\left(s_n^{(B)}\right) - \sum_{\substack{\{A_{n''}:|\ A_n|=\underline{k}+(j-1), \\ A_{n''}\subset B\}}} P'(A_{n''})$$

for every B of size $\underline{k}+j$. Proceeding similarly up to $j = \overline{k} - \underline{k}$, one gets complete distribution of $A_{n''}$ recursively. (proof of the theorem has been given in Appendix).

It can easily be seen that $\dfrac{P'(s_n)}{P'\left[s\binom{n''}{2}\right]}$ or $\dfrac{P'(s_{n'})}{P'\left[s\binom{n''}{2}\right]}$ independent of the parameter vector, $\mathbf{t} = (t_1, \ldots\ldots, t_{N'})'$, therefore, $s\binom{n''}{2}$ or $A_{n''}$ is a sufficient statistic.

Since, $\hat{\bar{T}}_{S_n}$ and $\hat{\bar{T}}^*_{s_{n'}}$ are both unbiased estimators of $\bar{T}_{U'}$ and $s\binom{n''}{2}$ or $A_{n''}$ is sufficient, therefore, Rao Blakwellization of both the estimators gives the same unbiased estimator $\tilde{\bar{T}}_{s\binom{n''}{2}|\,full-\inf}$ defined as

$$\tilde{\bar{T}}_{s\binom{n''}{2}|\,full-\inf} = \sum_{s_n \Rightarrow A_{n''}} \hat{\bar{T}}_{S_n} \cdot \frac{P'(s_n)}{P'(A_{n''})} = E^{S_n|\,A_{n''}}\left\{\hat{\bar{T}}_{S_n}\right\} \qquad (2.4)$$

where, $\sum$ is over all such ppswr samples of size s_n in S' that result into same $A_{n''}$ given $A_{n''}$ in U, $S'_{A_{n''}}$, where ppswr samples s_n reduces into a given $A_{n''}$.

S_n in the above expectation is a random variable taking its values from the subspace $S'_{A_{n''}}$ This estimator is unbiased, since, $E^{A_{n''}}\left\{\tilde{\tilde{T}}_{s\binom{n''}{2}} \mid full-\inf\right\} = E^{A_{n''}} E^{S_n \mid A_{n''}}\left\{\hat{\bar{T}}_{s_n}\right\} = \bar{T}_{U'}$ where $E^{A_{n''}}$ is over all the subsets of *U*. Further, the variance of the full-information estimator is given by

$$V\left\{\tilde{\tilde{T}}_{s\binom{n''}{2}} \mid full-\inf\right\} = E^{A_{n''}}\left[E^{S_n \mid A_{n''}}\left\{\hat{\bar{T}}_{s_n}\right\}\right]^2 - \bar{T}_{U'}^2,$$

$$= \sum_{A_{n''} \subset U}\left[\sum_{\substack{s_n \in S' \text{ so that} \\ s_n \text{ gives } A_{n''} \Rightarrow s\binom{n''}{2}}} \hat{\bar{T}}_{s_n} \cdot \frac{P'(s_n)}{P'(A_{n''})}\right]^2 P'(A_{n''}) - \bar{T}_{U'}^2,$$

where, $\sum_{s_n \in S'}$ is over all such ppswr samples in S' that gives the same $A_{n''}$ which is being fixed at the previous summation $\sum_{A_{n''} \subset U}$. Here, $P'(A_{n''})$ are obtained as the probability distribution of the random sets $A_{n''}$ derived in Theorem 2.1.

Finally, we get the following relationship

$$V\left\{\tilde{\tilde{T}}_{s\binom{n''}{2}} \mid full-inf\right\} \le V\left\{\tilde{\tilde{T}}^*_{s_{n'}}\right\} \le V\left\{\hat{\bar{T}}_{s_n}\right\} \tag{2.5}$$

3. Estimation under equal probability design

Let the design $D'(U', S', P')$ be srswr. We draw a simple random sample s_n of size *n* from U' with replacement (srswr) and denote the corresponding data by $d = \{(\mathrm{U}_k, t_k) : \mathrm{U}_k \in s_n\}$. Based on *d* we have an unbiased estimator of S_y^2

$$\hat{\bar{T}}_{s_n} = \frac{1}{n}\sum_{s_n} t_k = \frac{1}{n}\sum_{k=1}^{n'} f_k t_k \tag{3.1}$$

where f_k is the number of times the k^{th} 2-tuple is repeated in the sample s_n, and n' be the number of distinct 2-tuples in s_n. $V\left(\hat{\bar{T}}_{s_n} \mid srswr\right) = \frac{\binom{N}{2} - 1}{\binom{N}{2}} \cdot \frac{S_t^2}{n}$

where $S_t^2 = \frac{1}{\binom{N}{2}-1}\sum_{U'}\left(t_k - \bar{T}_{U'}\right)^2$. Using the result of Raj and Khamis (1958) and Des Raj (1968)-pp40, theorem 3.5, a better unbiased (Rao-Blackwellized) estimator of s_y^2 based on sample $s_{n'}$ of n' distinct 2-tuples in s_n is

$$\hat{\bar{T}}^*_{s_{n'}} = E^{S_n|s'_n}\left[\hat{\bar{T}}_{s_n}\right] = \frac{1}{n'}\sum_{s_{n'}} t_k \tag{3.2}$$

with $V\left[\hat{\bar{T}}^*_{s_{n'}}\right] = \frac{\binom{N}{2}-n'}{\binom{N}{2}n'}.S_t^2 \le V\left[\hat{\bar{T}}_{s_n}\right]$

We select n' 2-tuples from U', by srswor, consequently we get a set, $A_{n''}$, of n'' distinct units of U. These n'' distinct units allow us to know the information (t-values) on $\binom{n''}{2}$ distinct 2-tuples. For equal probability without replacement sampling the probability distribution of these n'' units is stated in the following corollary to theorem 2.1. In theorem 2.1 we noted that the probabilities $P'(A_{n''})$ were all different for different $A_{n''}$'s of the same size because these probabilities were depending on the units constituting $A_{n''}$ whereas for equal probability sampling the probabilities of sets $A_{n''}$ do not depend on the units constituting $A_{n''}$ rather on the size of $A_{n''}$, *i.e.,* $|A_{n''}|$.

Corollary 3.1. *Consider the design $D'(U', S', P')$ as simple random sampling without replacement. Consider a srswor sample $s_{n'}$ of size n' from U' under the design D', the probability distribution of distinct units n'' in $A_{n''}$ would be given by*

$$P'(n'' = \underline{k}) = \frac{\binom{\binom{\underline{k}}{2}}{n'}\binom{N}{\underline{k}}}{\binom{\binom{N}{2}}{n'}} \tag{3.3}$$

$$P'(n'' = \underline{k}+1) = \left[\frac{\binom{\binom{\underline{k}+1}{2}}{n'}}{\binom{\binom{N}{2}}{n'}} - \frac{(\underline{k}+1)P'(n'' = \underline{k})}{\binom{N}{\underline{k}}}\right]\binom{N}{\underline{k}+1} \tag{3.4}$$

$$P'\left(n'' = \underline{k} + j\right) = \left[\frac{\left(\!\!\binom{\underline{k}+j}{2}\!\!\right)}{\left(\!\!\binom{N}{2}\!\!\right)} - \sum_{k=\underline{k}}^{\underline{k}+j-1} \frac{\binom{k+j}{k} P'\left(n'' = k\right)}{\binom{N}{k}} \right] \binom{N}{k+j} \tag{3.5}$$

for $j = 1,.....,\bar{k} - \underline{k}$

where, $\underline{k} = \min\left\{k : \binom{k}{2} \geq n'\right\}$; and $\bar{k} = \min\{n'2, N\}$ (3.6)

We propose an estimator based on these $\binom{n''}{2}$ distinct 2-tuples and call it as full-information estimator, which has been defined as

$$\tilde{\bar{T}}_{s\left(\binom{n''}{2}\right)} = \frac{1}{\binom{n''}{2}} \sum_{k=1}^{\binom{n''}{2}} t_k \tag{3.7}$$

The above proposed estimator could be viewed as a Rao Blackwellization of the estimator $\hat{\bar{T}}^*_{s_{n'}}$ based on srswor of 2-tuples from U' given the sufficient statistics $s\left(n''_2\right)$, that is, $\tilde{\bar{T}}_{s\left(\binom{n''}{2}\right)} = E^{S_{n'} | s\left(\binom{n''}{2}\right)}\left(\hat{\bar{T}}^*_{S_{n'}}\right) = \frac{1}{\binom{n''}{2}} \sum_{k=1}^{\binom{n''}{2}} t_k$. Moreover, $V\left[\tilde{\bar{T}}_{s\left(\binom{n''}{2}\right)}\right] \leq V\left[\hat{\bar{T}}^*_{s_{n'}}\right]$.

We shall now state the expression for variance of the proposed estimator $\tilde{\bar{T}}_{s\left(\binom{n''}{2}\right)}$ and its relative efficiency in the following theorem.

Let us consider the three conditioning stages as follows:

Stage I: Decide about the number n'' of distinct units from *U*.

Stage II: Select a srswor of n'' units from *U*. Given these n'' units a total of $\binom{n''}{2}$ 2-tuples are generated.

Stage III: A srswor of size n' 2-tuples are selected from the $\binom{n''}{2}$ 2-tuples obtained at stage II.

Under these conditioning stages, we shall prove the following theorem :

Theorem 3.1. Under the above 3 stages of sample selection of $s_{n'}$, the following results hold :

1. The estimator $\bar{\bar{T}}_{s_{n'}} \mid n'', \mathrm{II} = \frac{1}{n'}\sum_{s_{n'}} t_k \mid n''$, *II and the proposed full-information estimator* $\bar{\bar{T}}_{s\left(n''_r\right)}$ *estimate* $\bar{T}_{U'}$ *unbiasedly for all values of* n''.

2. $V\left(\bar{\bar{T}}_{s_{n'}} \mid n'', \mathrm{II}\right) \geq V_2\left(\bar{\bar{T}}_{s\left(n''_2\right)} \mid n''\right) =$

$$A_1 \sum_{U'} t_k^2 + A_2 \left\{\sum_{U'}^{(1)} t_k t_{k'} + \sum_{U'}^{(2)} t_k t_{k'} + \sum_{U'}^{(3)} t_k t_{k'}\right\} + A_3 \sum_{U'}^{(4)} t_k t_{k'}$$

Where $A_1 = \frac{1}{\binom{n''}{2}\binom{N}{2}}\left(1 - \frac{n''(n''-1)}{N(N-1)}\right)$ $A_2 = \frac{2}{\binom{n''}{2}\binom{N}{2}}\left(\frac{(n''-2)}{(N-2)} - \frac{n''(n''-1)}{N(N-1)}\right)$;

and $A_3 = \frac{2}{\binom{n''}{2}\binom{N}{2}}\left(\frac{(n''-2)(n''-3)}{(N-2)(N-3)} - \frac{n''(n''-1)}{N(N-1)}\right)$

3. $V\left(\bar{\bar{T}}_{s_{n'}}\right) = E_1(A_1 + B_1)\sum_{U'} t_k^2 + E_1(A_2 + B_2)\left\{\sum_{U'}^{(1)} t_k t_{k'} + \sum_{U'}^{(2)} t_k t_{k'} + \sum_{U'}^{(3)} t_k t_{k'}\right\}$

$+ \; E_1(A_3 + B_3)\sum_{U'}^{(4)} t_k t_{k'} = C$ (say) *Where A's are as in (2), and B's are defined as* $B_1 = \frac{1}{\binom{N}{2}}\left[\frac{1}{n'} - \frac{1}{\binom{n''}{2}}\right]$; $B_2 = 2\left[\frac{1}{n'} - \frac{1}{\binom{n''}{2}}\right].\frac{1}{\binom{n''}{2}-1}\frac{\binom{n''}{3}}{\binom{n''}{2}\binom{N}{3}}$;

$B_3 = 2\left[\frac{1}{n'} - \frac{1}{\binom{n''}{2}}\right].\frac{1}{\binom{n''}{2}-1}\frac{\binom{n''}{4}}{\binom{n''}{2}\binom{N}{4}}$

And

4. $V\left[\bar{\bar{T}}_{s\left(n''_2\right)}\right] = E_1 V_2\left(\bar{\bar{T}}_{s\left(n''_2\right)}\right) = E_1(A_1)\sum_{U'} t_k^2 + E_1(A_2)\left\{\sum_{U'}^{(1)} t_k t_{k'} + \sum_{U'}^{(2)} t_k t_{k'} + \sum_{U'}^{(3)} t_k t_{k'}\right\}$

$+ \; E_1(A_3)\sum_{U'}^{(4)} t_k t_{k'} = D$ (say).

5. Relative efficiency (RE) of using $\bar{\bar{T}}_{s\left(n''_r\right)}$ *over* $\bar{\bar{T}}_{s_{n'}}$ *shall be given by*

$$\mathrm{RE} = \mathrm{C/D} \tag{3.8}$$

where, C and D are given in (3) and (4) and E_1 *corresponds to the probability distribution of distinct units* n'' *in the random set* $A_{n''}$ *obtained in corollary 2.1.* (proof of the theorem has been given in Appendix).

The calculations of RE could easily be performed by using the probability distribution of n'' obtained in corollary 2.1.

In case of simple random sampling without replacement design D', the full-information estimator $\tilde{\bar{T}}_{s\binom{n''}{2}}$ becomes $s_y^2 = \frac{1}{n''-1}\sum_1^{n''}(y_i - \bar{y})^2$ as an estimator of s_y^2 when $t_k(y) = \frac{1}{2}(y_i - y_j)^2$, $k = (i,j) = 1,2 \ldots \binom{N}{2}$;

$c_{yz} = \frac{1}{n''-1}\sum_1^{n''}(y_i - \bar{y})(z_i - \bar{z})$ of C_{yz}

when $v_{YG} = \sum_{i<j=1}^{n''} \frac{1}{\pi_{ij}}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$,

where $\pi_i = \pi_j = \frac{n''}{N} \cdot \pi_{ij} = \frac{n''(n''-1)}{N(N-1)}$ (since the present design D' is srswor) an estimator of V_{YG}.

A re-accounting and a note on the application of the above results is required at this stage. In the theorem 3.1 it has been shown that the full-information estimator $\tilde{\bar{T}}_{s\binom{n''}{2}}$ is superior to the one based on srswor sample $s_{n'}, \bar{\bar{T}}_{s_{n'}}$ Therefore, this suggests to draw n'' distinct units from U to generate $\binom{n''}{2}$ 2-tuples to calculate $\tilde{\bar{T}}_{s\binom{n''}{2}}$. One simplest consideration for deciding n'' could be the cost function $C = C_0 + c.n''$, where C_0 being the overhead cost, and c the cost of observing a unit in U. This cost function is reasonable when the major cost is born in observing the units and the cost of calculating the estimate is comparatively negligible. This consideration suggests us to select as large n'' distinct units from U as permitted by $n'' \le \frac{C - C_0}{c}$.

To sum up the above development, it is suggested that we shall initially select a sample of n'' units from U. Next, we consider all 2-tuples generated by it, $s\binom{n''}{2}$, and estimate $\bar{T}_{U'}$ by utilizing t-measures on generated sample, $s\binom{n''}{2}$, of 2-tuples. For the suggested full information estimator we need to know the partition of S', ppswr sample space, on which each sample reduces into a given $A_{n''}$ and also the conditional probabilities on this partition. Theorem 2.1 gives the required probability distribution of $A_{n''}$ in the present setting. This theorem also helps in identifying s_n's resulting into a fixed $A_{n''}$ starting with $n'' = 2$ and progressing recursively, instead of checking every sample s_n among N''^n samples in S' one by one that gives the same $A_{n''}$.

4. Empirical study

The proposed estimators were applied to estimate the population variance of imports on a real population C0124 of 124 countries consisting of data on 1983 import figures (in millions of U.S. dollars) and 1982 figures on gross national product (in tens of millions of U.S. dollars) (Sarndal, Swensson, Wretman (1991). The correlation between y(IMP) and x(GNP) and $t(\mathbf{y})$ and $t(\mathbf{x})$ are both high; 0.91 and 0.94 respectively. On the basis of computer simulations we have obtained the following results, for the efficiencies of discussed estimators as compared to the estimator $\hat{\bar{T}}_{s_n}$.

Table 1. Relative Efficiencies

Estimator of S_y^2	$\hat{\bar{T}}_{s_n}$	$\hat{\bar{T}}^*_{s'_n}$	$\tilde{\bar{T}}_{s\left(\frac{n''}{2}\right)}$	$s_{y\ conv.}^2$
Design: (ppswr)	1	2.11 2.09	4.59	1.96
Design: (srswr)	1		3.72	1.66

$RE = V\left(\hat{\bar{T}}_{s_n}\right) / V(.)$

In these calculations, the conventional estimator of s_y^2 has been taken as $s_{yconv}^2 = \frac{1}{n}\sum_{s_n} \frac{(y_i - \bar{y})^2}{(N-1)P_i}$ which is unbiased for S_y^2. Its variance has been given by $V\left(s_{yconv}^2\right) = \left\{\sum_U \frac{(y_i - \bar{y})^2}{(N-1)P_i} - S_y^4\right\}$. The results in table 1 show clearly that the proposed full information estimator claims overall superiority. Other estimators of variance available in literature have not been included in this study since these and proposed estimator are based on different designs, therefore, non comparable.

Appendix

PROOF OF THEOREM 2.1

Let the set B of M fixed units from U be given by $B = \left\{U_{l_1}, U_{l_2}, ..., U_{l_M}\right\}$, U_{l_j} 's being taken from U, so that $2 \le M \le N$. The ppswr sample, s_n, gives the random set $A_{n''} = \left\{U_{i_1}, U_{i_2}, ..., U_{i_n}\right\}$, $|A_{n''}| = n'' \le M$. Under the design D'

$$P'\left[A_{n''} \subset B\right] = \sum_{S'_B} P'\left(s_n^{(B)}\right) \tag{A.1}$$

The set $A_{n''}$ is a random set which is one set among the class of all subsets of *U*. The size of $A_{n''}$, *i.e.* $|\ A_{n''}|$ would range from $\underline{k} = 2$ to $\bar{k} = \min\{2n, N\}$. Further,

$$P'\left[A_{n''} \subset B\right] = \sum_{\{A_{n''}: A_{n''} \subset B\}} P'\left(A_{n''}\right) = \sum_{k=\underline{k}}^{\bar{k}} \sum_{\substack{\{A_{n''}: |\ A_{n''}| = k, \\ A_{n''} \subset B\}}} P'\left(A_{n''}\right) \tag{A.2}$$

On comparing (14) and (15), we get,

$$\sum_{k=\underline{k}}^{\bar{k}} \sum_{\substack{\{A_{n''}: |\ A_{n''}| = k, \\ A_{n''} \subset B\}}} P'\left(A_{n''}\right) = \sum_{S'_B} P'\left(s_n^{(B)}\right) \tag{A.3}$$

for every set *B,* so that $|\ B| = M \geq \underline{k}$.

If we solve (16) for $P'\left(A_{n''}\right)$ starting with every *B* of size $\underline{k}$, *i.e.,* $|\ B| = \underline{k}\,(M = 2)$, one gets probabilities of every sets $A_{n''}$ of size 2.

Next, for some fixed *B* of size $\underline{k} + 1(2 + 1)$, we have

$$P'\left(A_{n''}\right)|_{\substack{A_{n''} = B, \\ |\ B| = 2+1}} = \sum_{S'_B} P'\left(s_n^{(B)}\right) - \sum_{\substack{\{A_{n''}: |\ A_{n''}| = r, \\ A_{n''} \subset B\}}} P'\left(A_{n''}\right) \tag{A.4}$$

where the second term on right hand expression comes from probabilities $P'\left(A_{n''}\right)$ at the previous step. The probabilities of all such $A_{n''}$'s each of size (2 + 1) could be calculated simply by taking variation over all subsets *B* of size 2 + 1. Proceeding similarly for $j = 2$ to $\bar{k} = \min\{2n, N\}$, we get a complete probability distribution of $A_{n''}$. More generally, for $\underline{k} + j$ and for fixed *B,* $|\ B| = \underline{k} + j$

$$P'\left(A_{n''}\right)|_{\substack{A_{n''} = B, \\ |\ B| = \underline{k}+j}} = \sum_{S'_B} P'\left(s_n^{(B)}\right) - \sum_{\substack{\{A_{n''}: |\ A_{n''}| = \underline{k}+j-1, \\ A_{n''} \subset B\}}} P'\left(A_{n''}\right) \tag{A.5}$$

$$j = 1, \ldots, \bar{k} - \underline{k}$$

Varying *B,* each of size $(\underline{k} + j)$, one gets the probabilities $P'\left(A_{n''}\right)$ so that $A_{n''} = B$. Note that the second term on the right side of the above expression comes from probabilities $P'\left(A_{n''}\right)$ obtained at the previous step. The above recursive method enables one to get the complete probability distribution of $A_{n''}$. This completes the proof.

The advantage of the recursive procedure explained in Theorem 2.1 is that one needs only to calculate the probabilities of all such ppswr samples of size *n*, $s_n^{(M)}$'s drawn from U'_M, the space of all 2-tuples made of units from fixed set *B*, and proceed with $M = \underline{k}$. For variations in *B* of size $\underline{k}$, one gets $P'(A_{n''})$ with $B = A_{n''}$. These probabilities are then used to calculate *P'(A*$_n$*")*, $|A_{n''}| = \underline{k}+1$ for variations in *B*, $B = A_{n''}$ so that $|B| = \underline{k}+1$. Proceeding similarly, one gets the complete probability distribution of $A_{n''}$.

ILLUSTRATION OF THEOREM 2.1: Consider a ppswr design, $D(U,S,P)$, $U = \{U_1, U_2, U_3, U_4\}$; and *S* consists of $4^3 = 64$ ppswr samples of size *n* = 3; the sizes of the units U_1, U_2, U_3, U_4 are 1.00, 2.00, 3.00, 4.00 respectively, giving their selection probabilities 0.1, 0.2, 0.3 and 0.4 respectively. Let us now construct a design $D'(U', S', P')$ so that

$$U' = \{\mathbf{U}'_1 = (U_1, U_2), \mathbf{U}'_2 = (U_1, U_3), \mathbf{U}'_3 = (U_1, U_4), \mathbf{U}'_4 = (U_2, U_3),$$

$$\mathbf{U}'_5 = (U_2, U_4), \mathbf{U}'_6 = (U_3, U_4)\}; S'$$

is sample space of ppswr samples of size *n* = 3 having $(N')^n = 6^3 = 216$ samples, where, $N' = \binom{N}{2} = 6$, *n* = 3; and the induced probabilities, are $P'(\mathrm{U}'_1) = 0.0472 = P'_1$, $P'(\mathrm{U}'_2) = 0.0762 = P'_2$, $P'(\mathrm{U}'_3) = 0.111 = P'_3$, $P'(\mathrm{U}'_4) = 0.161 = P'_4$, $P'(\mathrm{U}'_5) = 0.233 = P'_5$, $P'(\mathrm{U}'_6) = 0.3714 = P'_6$. Note $\underline{k} = 2$ and $\bar{k} = \min\{nr, N\} = \min\{6, 4\} = 4$. Now starting with $M = \underline{k} = 2$ for the set $B = \{1, 2\}$, the theorem gives,

$$P(A_{n''} = \{1,2\}) = P'(s_n^{(B)}) = P'\{(1,2),(1,2),(1,2)\}$$

$$= \{P'(\mathbf{U}'_1)\}^3 = P_1'^3 = 1.052 \times 10^{-4}$$

similarly, we get $P(A_{n''} = \{1,3\}) = P_2'^3 = 4.425 \times 10^{-4}$;

$P(A_{n''} = \{1,4\}) = P_3'^3 = 1.3676 \times 10^{-3}$; $P(A_{n''} = \{2,3\}) = P_4'^3 = 4.173 \times 10^{-3}$; $P(A_{n''} = \{2,4\}) = P_5'^3 = 0.0126$; $P(A_{n''} = \{3,4\}) = P_6'^3 = 0.0512$.. Next, consider sets *B* of size $\underline{k}+1 = 3$, we get the probabilities of sets $A_{n''}$ of size 3. Let us fix *B* = {1, 2, 3}, we have U'_B = {(1, 2), (1, 3), (2, 3)}; sample space under ppswr from U'_B would be given by

$S'_B = \{\mathrm{U}'_1, \mathrm{U}'_1, \mathrm{U}'_1\}, \{\mathrm{U}'_2, \mathrm{U}'_2, \mathrm{U}'_2\}, \{\mathrm{U}'_4, \mathrm{U}'_4, \mathrm{U}'_4\}, \{\mathrm{U}'_1, \mathrm{U}'_1, \mathrm{U}'_2\}, \{\mathrm{U}'_1, \mathrm{U}'_1, \mathrm{U}'_4\}$ $\{\mathrm{U}'_2, \mathrm{U}'_2, \mathrm{U}'_1\}, \{\mathrm{U}'_2, \mathrm{U}'_2, \mathrm{U}'_4\}, \{\mathrm{U}'_4, \mathrm{U}'_4, \mathrm{U}'_1\}, \{\mathrm{U}'_4, \mathrm{U}'_4, \mathrm{U}'_2\}$ all with three

permutations, and $\{U_1', U_2', U_4'\}$ with 3 permutations. $\sum_{S_B'} P'\left(s_n^{(B)}\right)$ becomes $P_1'^3 + P_2'^3 + P_4'^3 + 3\left\{P_1'^2\left(P_2' + P_4'\right) + P_2'^2\left(P_1' + P_4'\right) + P_4'^2\left(P_1' + P_2'\right)\right\} + 3!\left(P_1'.P_2'.P_4'\right)$.

The result of the theorem gives,

$$P'\left(\{1,2,3\}\right) = \sum_{S_B'} P'\left(s_n^{(B)}\right) - \sum_{\substack{\{A_{n''}:|A_{n''}|=2,\\ A_{n''}\subset B\}}} P'\left(A_{n''}\right) \tag{A.6}$$

Note $P'\left(A_{n''}\right), |A_{n''}| = 2,$ have been calculated at the previous step. This gives,

$$P'\left(\{1,2,3\}\right) = 3\left\{P_1'^2\left(P_2' + P_4'\right) + P_2'^2\left(P_1' + P_4'\right) + P_4'^2\left(P_1' + P_2'\right)\right\} + 6\left(P_1'.P_2'.P_4'\right)$$

$$= 0.018282.$$

similarly $P'\left(\{1,2,4\}\right) = 0.04576$, $P'\left(\{1,3,4\}\right) = 0.12126$, $P'\left(\{2,3,4\}\right) = 0.38035$. Having obtained the probabilities of sets $A_{n''}$ of size 3, we would now obtain $A_{n''}$ = {1, 2, 3, 4}, fix B = {1, 2, 3, 4}. Note that $S_B' = S'$ since $A_{n''} = U$, therefore, $\sum_{S_B'} P'\left(s_n^{(B)}\right) = 1$. The result of the theorem gives,

$$P'\left(\{1,2,3,4\}\right) = \sum_{S_B'} P'\left(s_n^{(B)}\right) - \sum_{\substack{\{A_{n''}:|A_{n''}|=3,\\ A_{n''}\subset B\}}} P'\left(A_{n''}\right) - \sum_{\substack{\{A_{n''}:|A_{n''}|=2,\\ A_{n''}\subset B\}}} P'\left(A_{n''}\right) = 0.3645.$$

PROOF OF COROLLARY 3.1. Choose and fix M units from U and denote them by a subset B, $|B| = M$, $1 \le M \le N$ and draw a srswor sample $s_{n'}$ of size n' from U', $P'(s_{n'}) = \dfrac{1}{\binom{N}{n'}}$; the corresponding sample space is denoted by S' and the set of distinct units in $s_{n'}$ be denoted by $A_{n''} = \left\{\mathbf{U}_{\mathbf{i}_1}' \cup \mathbf{U}_{\mathbf{i}_2}' \cup \ldots\ldots \cup \mathbf{U}_{\mathbf{i}_{n'}}'\right\}$ with size $|A_{n''}| = n''$, where $\mathbf{U}_{\mathbf{i}_1}' \cup \mathbf{U}_{\mathbf{i}_2}' \cup \ldots\ldots \cup \mathbf{U}_{\mathbf{i}_{n'}}'$ gives distinct units which have appeared in $\mathbf{U}_{\mathbf{i}}'$'s in $s_{n'}$. In U' there are $\binom{M}{2}$, 2-tuples consisting entirely of the units in B, we denote them by $U_M', U_M' \subset U'$. Then

$$P'\left[A_{n''} \subset B\right] = \frac{\binom{M'}{n'}}{\binom{N'}{n'}} \tag{A.7}$$

since n' r-tuples in $s_{n'}$ could be chosen in $\binom{M'}{n'}$ ways from U'_M out of total ways $\binom{N'}{n'}$, where $M' = \binom{M}{2}$, $N' = \binom{N}{2}$, $|B| = M$ and $|A_{n''}| \le M$. Let the probabilities of $A_{n''}$ having k distinct units be denoted by $P'(n'' = k)$ for $k = \underline{k}, \underline{k}+1, ..., \bar{k}$. We shall observe that $P'(A_{n''})$ for $A_{n''} \subset U$, depends only on the size of $A_{n''}$, *i.e.*, $|A_{n''}|$, not on which labels constitute $A_{n''}$. In other words, the distribution of $A_{n''}$ is invariant under the permutations of U_i's. Thus for some $A_{n''} \subset U$ such that $|A_{n''}| = k$, we have

$$P'(A_{n''}) = \frac{P'(n'' = k)}{\binom{N}{k}} \tag{A.8}$$

But, $P'[A_{n''} \subset B] = \sum_{\{T : T \subset B\}} P'(T), \quad |B| = M$

$$= \sum_{k=\underline{k}}^{\bar{k}} \binom{M}{k} P'(A_{n''})|_{|A_{n''}| = k}$$

From (21) we get $P'[A_{n''} \subset B] = \sum_{k=\underline{k}}^{\bar{k}} \binom{M}{k} \frac{P'(n'' = k)}{\binom{N}{k}}$ (A.9)

On comparing (20) and (22) we get

$$\sum_{k=\underline{k}}^{\bar{k}} \frac{\binom{M}{k}}{\binom{N}{k}} P'(n'' = k) = \frac{\binom{M'}{n'}}{\binom{N'}{n'}} \qquad \forall M \ge \underline{k} \tag{A.10}$$

These equations can be recursively solved for $P'(n'' = k)$ starting with $M = \underline{k}$, we get

$$P'(n'' = \underline{k}) = \frac{\binom{\binom{\underline{k}}{2}}{n'} \binom{N}{\underline{k}}}{\binom{N'}{n'}}$$

$$P'(n'' = \underline{k}+1) = \left[\frac{\binom{\binom{\underline{k}+1}{2}}{n'}}{\binom{N}{2}{n'}} - \frac{(\underline{k}+1) P'(n'' = \underline{k})}{\binom{N}{\underline{k}}} \right] \binom{N}{\underline{k}+1}$$

$$P'(n'' = \underline{k} + j) = \left[\frac{\binom{\binom{\underline{k}+j}{r}}{n'}}{\binom{N}{r}{n'}} - \sum_{k=\underline{k}}^{\underline{k}+j-1} \frac{\binom{k+j}{k} P'(n'' = k)}{\binom{N}{k}} \right] \binom{N}{\underline{k}+j}$$

For $j = 1,....,\bar{k} - \underline{k}$.

This is the desired distribution of number of distinct units n'' in a srswor $s_{n'}$ taken from U' under the design $D'(U', S', P')$.

ILLUSTRATION OF COROLLARY 2.1

Let $N = 5$, $n' = 3$. Then $\underline{k} = 3$ and $\bar{k} = 5$ and the equation (7) and (8) reduces to

$$P'(n'' = 3) = \binom{5}{3} \left(\binom{\binom{3}{2}}{3} \right) \Big/ \left(\binom{\binom{5}{2}}{3} \right) = \frac{1}{12}; \qquad \frac{1}{\binom{5}{4}} P'(4) + \frac{4}{\binom{5}{3}} P'(3) = \left(\binom{\binom{4}{2}}{3} \right) \Big/ \left(\binom{\binom{5}{2}}{3} \right) = \frac{1}{6}$$

and $P'(5) + P'(4) + P'(3) = 1$.

Thus $P'(3) = \frac{1}{12}$; $P'(4) = \left(\frac{1}{6} - \frac{4}{10} . \frac{1}{12} \right) .5 = \frac{2}{3}$ and $P'(5) = \frac{1}{4}$

PROOF OF THEOREM 3.1

1. The conditional expectation of $\bar{\bar{T}}_{s_{n'}}$, given n'' and II is

$$E_3 \left(\bar{\bar{T}} \mid n'', \mathrm{II} \right) = \frac{1}{\binom{n''}{2}} \sum_{k=1}^{\binom{n''}{2}} t_k = \tilde{\bar{T}}_{s\binom{n''}{2}}$$

and $E_2 E_3 \left(\bar{\bar{T}}_{s_{n'}} \mid n'', \mathrm{II} \right) = E_2 \left(\tilde{\bar{T}}_{s\binom{n''}{2}} \right)$

$$= \frac{1}{\binom{n''}{2}} \sum_{k=1}^{\binom{n''}{2}} E_2 (t_k) = \sum_{U'} t_k P \left(\mathbf{U}'_{\mathbf{k}} = (U_i, U_j) \right)$$

$$= \sum_{U'} t_k . \frac{2}{N(N-1)} = \bar{T}_{U'}$$

This shows that $\bar{\bar{T}}_{s_{n'}}$, and $\tilde{\bar{T}}_{s\binom{n''}{2}}$ are unbiased estimators of $\bar{T}_{U'}$ for all values of n''.

2. Now $V\left(\bar{\bar{T}}_{s_{n'}} \mid n'', \mathrm{II}\right) = V_2 E_3\left(\cdot \mid n'', II\right) + E_2 V_3\left(\cdot \mid n'', II\right) \geq V_2 E_3\left(\cdot \mid n'', II\right)$

$= V_2\left(\tilde{\bar{T}}_{s\binom{n''}{2}}\right)$

Showing $\tilde{\bar{T}}_{s\binom{n''}{2}}$ is superior to $\bar{\bar{T}}_{s_{n'}} \mid n'', \mathrm{II}$. Let us now consider

$$V_2\left(\tilde{\bar{T}}_{s\binom{n''}{2}}\right) = \frac{1}{\left\{\binom{n''}{2}\right\}^2}\left[E_2\left(\sum_{k=1}^{\binom{n''}{2}} t_k\right)^2 - E_2^2\left(\sum_{k=1}^{\binom{n''}{2}} t_k\right)\right] \tag{A.11}$$

Here, $E_2\left(\sum_1^{\binom{n''}{2}} t_k\right)^2$ calculates to be

$$\frac{n''(n''-1)}{N(N-1)}\sum_{U'} t_k^2 + 2\frac{n''(n''-1)(n''-2)}{N(N-1)(N-2)}\left\{\sum_{U'}^{(1)} t_k t_{k'} + \sum_{U'}^{(2)} t_k t_{k'} + \sum_{U'}^{(3)} t_k t_{k'}\right\}$$

$$+2\frac{n''(n''-1)(n''-2)(n''-3)}{N(N-1)(N-2)(N-3)}\sum_{U'}^{(4)} t_k t_{k'}$$

where

$\sum_{U'}^{(1)} = \sum_{\substack{k\neq k' \\ k=(i,j),\ k'=(i,l) \\ i<j<l}}$; $\sum_{U'}^{(2)} = \sum_{\substack{k\neq k' \\ k=(i,j),\ k'=(j,l) \\ i<j<l}}$; $\sum_{U'}^{(3)} = \sum_{\substack{k\neq k' \\ k=(i,j),\ k'=(k,j) \\ i<k<j}}$;

$\sum_{U'}^{(4)} = \sum_{\substack{k\neq k' \\ k=(i,j),\ k'=(k,l) \\ i<j<k<l \\ i\neq j\neq k\neq l}}$; and $\left\{E_2\left(\sum_1^{\binom{n''}{2}} t_k\right)\right\}^2$ calculates to be $\left(\frac{n''(n''-1)}{N(N-1)}\right)^2$

$$\{\sum_{U'} t_k^2 + 2\sum_{U'}^{(1)} t_k t_{k'} + 2\sum_{U'}^{(2)} t_k t_{k'} + 2\sum_{U'}^{(3)} t_k t_{k'} + 2\sum_{U'}^{(4)} t_k t_{k'}\}.$$

On putting these values in (24), we get

$$V_2\left(\tilde{\bar{T}}_{s\binom{n''}{2}}\right) = A_1\sum_{U'} t_k^2 + A_2\left\{\sum_{U'}^{(1)} t_k t_{k'} + \sum_{U'}^{(2)} t_k t_{k'} + \sum_{U'}^{(3)} t_k t_{k'}\right\} + A_3\sum_{U'}^{(4)} t_k t_{k'}$$

Where $A_1 = \frac{1}{\binom{n''}{2}\binom{N}{2}}\left(1 - \frac{n''(n''-1)}{N(N-1)}\right)$; $A_2 = \frac{2}{\binom{n''}{2}\binom{N}{2}}\left(\frac{(n''-2)}{(N-2)} - \frac{n''(n''-1)}{N(N-1)}\right)$;

$A_3 = \frac{2}{\binom{n''}{2}\binom{N}{2}}\left(\frac{(n''-2)(n''-3)}{(N-2)(N-3)} - \frac{n''(n''-1)}{N(N-1)}\right)$;

3. Next, consider

$$V_3\left[\bar{\bar{T}}_{s_{n'}} \mid n''\right] = \left[\frac{1}{n'} - \frac{1}{\binom{n''}{2}}\right] \frac{1}{\binom{n''}{2} - 1} \left[\sum_1^{\binom{n''}{2}} t_k^2 - \frac{1}{\binom{n''}{2}} \left(\sum_1^{\binom{n''}{2}} t_k\right)^2\right]$$

Taking E_2 on $V_3\left[\bar{\bar{T}}_{s_{n'}}\right]$, some algebraic simplification gives

$$E_2 V_3\left[\bar{\bar{T}}_{s_{n'}}\right] = B_1 \sum_{U'} t_k^2 - B_2 \left\{\sum_{U'}^{(1)} t_k t_{k'} + \sum_{U'}^{(2)} t_k t_{k'} + \sum_{U'}^{(3)} t_k t_{k'}\right\} - B_3 \sum_{U'}^{(4)} t_k t_{k'} \quad \text{(A.12)}$$

Where $B_1 = \frac{1}{\binom{N}{2}}\left[\frac{1}{n'} - \frac{1}{\binom{n''}{2}}\right]$; $B_2 = \left[\frac{1}{n'} - \frac{1}{\binom{n''}{2}}\right] \cdot \frac{1}{\binom{n''}{2} - 1} \cdot \frac{\binom{n''}{3}}{\binom{n''}{2}\binom{N}{3}}$;

$$B_3 = 2 \cdot \left[\frac{1}{n'} - \frac{1}{\binom{n''}{2}}\right] \cdot \frac{1}{\binom{n''}{2} - 1} \cdot \frac{\binom{n''}{4}}{\binom{n''}{2}\binom{N}{4}}$$

So far we have been treating n'' as a quantity fixed in advance, but in actual practice we fix n' and select n' 2-tuples from $\binom{N}{2}$ 2-tuples by srswor. For any given sample $s_{n'}$, we get n'' distinct units of U. Thus n'' is actually a random variable.

The distribution of n'' has already been calculated in corollary 2.1. by recursive method. We now have,

$$V\left(\bar{\bar{T}}_{s_{n'}}\right) = E_1 E_2 V_3(\cdot) + E_1 V_2 E_3(\cdot) + V_1 E_2 E_3(\cdot)$$

$$= E_1(A_1 + B_1) \sum_{U'} t_k^2 + E_1(A_2 + B_2)\left\{\sum_{U'}^{(1)} t_k t_{k'} + \sum_{U'}^{(2)} t_k t_{k'} + \sum_{U'}^{(3)} t_k t_{k'}\right\}$$

$$+ E_1(A_3 + B_3) \sum_{U'}^{(4)} t_k t_{k'} = C \text{ (say)} \quad \text{(A.13)}$$

4. From (25), we get

$$V\left(\bar{\bar{T}}_{s\binom{n''}{2}}\right) = E_1 V_2\left(\bar{\bar{T}}_{s\binom{n''}{2}}\right)$$

$$= E_1(A_1) \sum_{U'} t_k^2 + E_1(A_2)\left\{\sum_{U'}^{(1)} t_k t_{k'} + \sum_{U'}^{(2)} t_k t_{k'} + \sum_{U'}^{(3)} t_k t_{k'}\right\}$$

$$+ E_1(A_3) \sum_{U'}^{(4)} t_k t_{k'} = D \text{ (say)} \quad \text{(A.14)}$$

5. The relative efficiency (RE) of the estimator $\tilde{\bar{T}}_{s\left(\frac{n''}{2}\right)}$ has been given by

$$RE = \frac{V\left(\bar{\bar{T}}_{s_{n'}}\right)}{V\left(\tilde{\bar{T}}_{s\left(\frac{n''}{2}\right)}\right)} = C/D$$

REFERENCES

[1] CASSEL, C.M., SARNDAL, C.E., and WRETMAN, J.H. (1977). *Foundations of Inference in Survey Sampling.* New York: Wiley.

[2] CHAUDHRI, A.(1978). On Estimating the Variance of a Finite Population. *Metrika* **25**, 65–76.

[3] DATTA, G.S. and GHOSE, M.(1993).Bayesian Estimation of Finite Population Variances with Auxiliary Information. *Sankhya, Ser.* B,**55**, 156–170.

[4] DATTA and TIWARI (1991).Bayesian Estimation of Finite Population Variances with Auxiliary Information. *Sankhya, Ser.* B,**55**, 156–170.

[5] GHOSH, M. and LAHIRI, P.(1987).Robust Empirical Bayes Estimation of Variance From Stratified Samples. *.J. Amer. Statist, Assoc.,***82**, 1153–1162.

[6] GHOSE, M. and MEEDEN, G.(1983).Estimation of the Variance in Finite Population Sampling. *Sankhya, Ser. B,**45*, 362–375.

[7] GHOSE, M. and MEEDEN, G.(1984).A new Bayesian analysis of a random effect model. *J.R. Statist. Soc.,B***46**, 474–482.

[8] ERICSON, W.A. (1969).Subjective Bayesian Models in Sampling Finite Populations(With discussion). *Journal of Royal Statistical Society, Ser.* B**31**, 195–233.

[9] GODAMBE, V.P.(1955). A unified theory of sampling from finite populations. *J. R. Statist. Soc. B,* **17**, 208–278.

[10] HANURAV, T.V.(1966). Some aspect of unified sampling theory. *Sankhya, Ser. A,* **28**, 175–204.

[11] LAHIRI, P. and TIWARI, R.C.(1991). NonParametric Bayes and Empirical Bayes Estimation of Variances from Stratified Sampling. *Sankhya, Ser.* S,**52**, 105–118.

[12] LIU, T.P. (1974a). Bayes Estimation for the Variance of a finite population. *Metrika* **36**, part I, 23–32.

[13] LIU, T.P. (1974b). A General Unbiased Estimator for the Variance of a Finite Population. *Sankhya, Ser. C,* **36**, part I, 23–32.

[14] LIU, T. P. and THOMPSON, M.E. (1983). Properties of Estimators of Quadratic Finite Population Function the Batch Approach. *Annals of Statistics* **11***,* 275–285.

[15] MUKHOPADHYAY, P.(1978). Estimating the Variance of a Finite Population Under a Superpopulation Model. *Metrika* **25***,* 115–122.

[16] MUKHOPADHYAY, P. and BHATTACHARYA, S. (1989). On Estimating the Variance of a Finite Population Under a Superpopulation Model. *Journal of the Indian Statistical Assoc.**27,*** 37–46.

[17] RAJ, DES (1968). *Sampling theory.* Tata McGraw-Hill Publishing Company Ltd. New Delhi.

[18] SWAIN, A.K.P.C. and MISHRA, G.(1994). Estimation of Finite Population Variance Under Unequal Probability Sampling. *Sankhya, Ser. B,***56***,* 374–388.

[19] SARNDAL.C.E., SWENSSON, B. and WRETMAN, J. (1992).*Model Assisted Survey Sampling,* Springer-Verlag,Newyork.

[20] VARDEMAN, S. and MEEDEN, G. (1983). Admissible Estimators in Finite Population Sampling Employing Various Types of Prior Information. *Journal of Statistical Planning Inference, 7,* 329–341.

[21] WU,C. and SITTER, R.R. (2002). Efficient Estimation of Quadratic Finite Population Functions in the Presence of Auxiliary Information. *J. Amer. Statist. Assoc.,**97**,,* 535–543.

[22] ZACKS, S. and SOLOMAN, H. (l981). Bayes and Equivariant Estimators of the Variance of a finite population:Part I, Simple random Sampling. *Commun. Statist-Theory Meth.,* **A10**, 407–426.

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 437—444

A NONPARAMETRIC CONFIDENCE INTERVAL FOR AT-RISK-OF-POVERTY-RATE

Wojciech Zieliński[1]

ABSTRACT

In the European Commission Eurostat document Doc. IPSE/65/04/EN page 11, the ``*at-risk-of-poverty rate*'' (ARPR) is defined as a percent of population with income smaller than 60% of population median. Zieliński (2008) proposed a distribution-free confidence interval for ARPR. In the paper, an example of application of the constructed confidence interval is shown.

Key words: binomial distribution, confidence interval, ARPR.

1. Introduction

In the European Commission Eurostat document Doc. IPSE/65/04/EN page 11, the ``*at-risk-of-poverty rate*'' (ARPR) is defined as follows. Let $EQ\_INC_i$ denote the equivalised disposable income of the $i$-th person and let $weight_i$ denote the weight of person $i$. The ``*at-risk-of-poverty threshold*'' (ARPT) is calculated as 60% of calculated median value, i.e.

$$\text{ARPT} = \text{At risk of poverty threshold} = 60\%\text{EQ\_INC}_{\text{MEDIAN}},$$

where

$$EQ\_INC_{MEDIAN} = \begin{cases} \frac{1}{2}(EQ\_INC_j + EQ\_INC_{j+1}), & if \quad \sum_{i=1}^{j} weight_i = \frac{W}{2} \\ EQ\_INC_{j+1}, & if \quad \sum_{i=1}^{j} weight_i < \frac{W}{2} < \sum_{i=1}^{j+1} weight_i \end{cases},$$

and

$$W = \sum_{All\ persons} weight_i.$$

[1] Department of Econometrics and Statistics Warsaw University of Life Sciences, Nowoursynowska 159, 02-776 Warszawa, e-mail: wojtek.zielinski@statystyka.info.

Then the ``*at-risk-of-poverty rate*'' is calculated as the percentage of persons (over the total population) with an equivalised disposable income below the *at-risk-of-poverty threshold* (i.e. the equivalised disposable income of each person is compared with *at-risk-of-poverty threshold*). The cumulated weights of persons whose equivalised disposable income is below the *at-risk-of-poverty threshold*, is divided by the cumulated weights of the total population (i.e. sum of all the personal weights):

$$ARPR = \frac{\sum_{All\ persons\ with\ EQ\_INC<at-risk-of-poverty\ threshold} weight_i}{W} \times 100\%.$$

The natural estimator of ARPR is as follows. Let $X_1,\ldots,X_n$ be a sample of disposable incomes of randomly drawn *n* persons and *Med* denotes the sample median. The estimator $\hat{A}RPR$ is defined:

$$\hat{A}RPR = \frac{1}{n}\#\{X_i \le 0.6 \cdot Med\},$$

where #*S* denotes the cardinality of the set *S*.

The properties of the above estimator (bias, variance etc.) depends strongly on the distribution *F* of population income. Zieliński (2006) showed that the estimator is almost unbiased, i.e.

$$E_F\,\hat{A}RPR \approx ARPR = F(0.6 \cdot Q(0.5))$$

for all continuous *F* (*Q(0.5)* stands for the median of the distribution *F*). He also calculated its variance.

However, the problem is in interval estimation. Zieliński (2007) proposed a nonparametric confidence interval for *ARPR*. It appeared that his confidence interval is too conservative, i.e. the true confidence level is significantly larger that the nominal one. In what follows, we construct a confidence interval for *ARPR,* the confidence level of which is near nominal one.

2. Confidence interval

Let *F* denotes the cdf of a distribution of population income. It is assumed that *F* is continuous. We are interested in estimation of the parameter

$$\theta = F(\alpha Q(q)),$$

for given $\alpha, q \in (0,1)$, where $Q(\cdot)$ denotes the quantile function $(Q(x) = F^{-1}(x))$. For $\alpha = 0.6$ and $q = 0.5$ parameter θ is *ARPR*. We are interested in constructing a confidence interval for θ.

Let $X_1, X_2, \ldots, X_n$ be a sample from F and let $X_{1:n} < \cdots < X_{n:n}$ be order statistics. As an estimator of θ we take

$$\hat{\theta} = \frac{1}{n} \# \{ X_i \le \alpha \cdot X_{M:n} \},$$

where $M = \lfloor qn \rfloor + 1$ ($\lfloor a \rfloor$ is the greatest integer not greater than *a*). Here, $X_{M:n}$ is an estimator of *q*-quantile *Q(q)* of the *F* distribution.

Let ξ be the number of observations not greater than $\alpha \cdot X_{M:n}$:

$$\xi = \# \{ X_i \le \alpha \cdot X_{M:n} \}.$$

The distribution of ξ is

$$P_F\{\xi = k\} = P_F\{\xi \ge k\} - P_F\{\xi \ge k+1\} = P_F\{X_{k:n} \le \alpha \cdot X_{M:n}\}$$

$$- P_F\{X_{k+1:n} \le \alpha \cdot X_{M:n}\}, \ k = 0, \ldots, M-1.$$

where (David and Nagaraja 2003)

$$P_F\{X_{k:n} \le \alpha \cdot X_{M:n}\} = \frac{n!}{(k-1)!(M-k-1)!(n-M)!} \int_{-\infty}^{\infty} (1 - F(v))^{n-M} \cdot$$

$$f(v) \int_{-\infty}^{\infty} F(u)^{k-1} [F(v) - F(u)]^{M-k-1} f(u) du dv$$

$$= \frac{n!}{(k-1)!(M-k-1)!(n-M)!} \int_0^1 (1-v)^{n-M} \int_0^1 u^{k-1} [v-u]^{M-k-1} du dv$$

$$= \int_0^1 B_{k,M-k} \left(\frac{F(\alpha Q(v))}{v} \right) b_{M,n-M+1}(v) dv$$

Here, $B_{a,b}(\cdot)$ and $b_{a,b}(\cdot)$ denotes cdf and pdf of beta distribution with parameters *(a,b)*, respectively. That distribution may be written in the form

$$P_F\{X_{k:n} \le \alpha \cdot X_{M:n}\} = B_{k,M-k} \left(\frac{F(\alpha Q(q))}{q} \right)$$

$$+ \int_0^1 \left[B_{k,M-k} \left(\frac{F(\alpha Q(v))}{v} \right) - B_{k,M-k} \left(\frac{F(\alpha Q(q))}{q} \right) \right] b_{M,n-M+1}(v) dv.$$

It is well known, that if S_n is a random variable distributed as binomial with parameters n and p, then

$$P_{n,p}\{S_n \le k\} = \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j} = B_{n-k,k+1}(1-p).$$

Hence, the distribution of ξ is almost binomial with parameters *M-1* and $\frac{F(\alpha Q(q))}{q}$.

If F is power distribution with shape parameter b, i.e.

$$F(x) = x^b, \quad 0 < x < 1,$$

then

$$\frac{F(\alpha Q(v))}{v} \equiv \alpha^b,$$

and

$$\int_0^1 \left[B_{k,M-k}\left(\frac{F(\alpha Q(v))}{v} \right) - B_{k,M-k}\left(\frac{F(\alpha Q(q))}{q} \right) \right] b_{M,n-M+1}(v)dv = 0.$$

Hence, in case of power function distribution, ξ is binomially distributed with parameters *M-1* and α^b. For this distribution $\theta = \alpha^b q$.

Let $\gamma \in (0,1)$. Consider an interval (see Appendix)

$$\left(qB^{-1}\left(\xi, M-\xi+1; \frac{1-\gamma}{2} \right); qB^{-1}\left(\xi+1, M-\xi; \frac{1+\gamma}{2} \right) \right), \qquad (*)$$

where $B^{-1}(a,b;\delta)$ is the δ quantile of beta distribution with parameters (a,b). If F is power function distribution, then

$$P_F\left\{ \theta \in \left(qB^{-1}\left(\xi, M-\xi+1; \frac{1-\gamma}{2} \right); qB^{-1}\left(\xi+1, M-\xi; \frac{1+\gamma}{2} \right) \right) \right\} \ge \gamma.$$

Hence, *(*)* is a confidence interval for θ. The question is: what is the value of

$$P_F\left\{ \theta \in \left(qB^{-1}\left(\xi, M-\xi+1; \frac{1-\gamma}{2} \right); qB^{-1}\left(\xi+1, M-\xi; \frac{1+\gamma}{2} \right) \right) \right\} \qquad (**)$$

for distributions F other the power function one? In general, it is impossible to calculate *(**)* because it strongly depends on F. In tables 1-3 there are calculated

probabilities *(**)* (denoted by $\hat{\gamma}$ and the mean length ($\hat{\Delta}$) of confidence interval *(*)* for:

Pareto distribution with pdf $b\left(\frac{1}{x+1}\right)^{b-1}$ for $x \in (0,\infty)$;

Gamma distribution with pdf $\frac{1}{\Gamma(b)} x^{b-1} e^{-x}$ for $x \in (0,\infty)$;

Lognormal distribution with pdf $\frac{1}{xb\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln x}{b}\right)^2\right]$ for $x \in (0,\infty)$.

Calculations were made for α=0.6, q=0.5, n=20, 200, 2000, $M = \mathrm{d}n/2\mathrm{t}+1$ and γ=0.95.

3. Conclusions

The confidence interval for $\theta=F(\alpha \cdot Q(q))$ may be constructed as a confidence interval in binomial distribution based on the number of observations not greater than $\alpha \cdot X_{M:n}$. Its confidence level for power function distribution is exactly the same as the confidence level of confidence interval for binomial proportion. For Pareto, Gamma and Lognormal distributions confidence level does not differ much from the nominal one. It may be expected that for other continuous distributions the confidence level will behave similarly.

Appendix: confidence interval for binomial proportion

Let η be a binomial random variable with parameters n and unknown p. It is well known that

$$P_p\{\eta \le k\} = B_{n-k,k+1}(1-p) \quad \text{and} \quad P_p\{\xi \ge k\} = B_{k,n-k+1}(p).$$

Let $\delta \in (0,1)$ be a given number. Confidence interval for p at the confidence level δ is defined as

$$P_p\{p_L(\eta) \le p \le p_U(\eta)\} = \delta, \quad \text{for all } p \in (0,1).$$

For given n and k let $p_L(k)$ be the solution of

$$B_{n-k+1,k}(1-p_L(k)) = \frac{1+\delta}{2} \quad \text{or equivalently} \quad B_{k,n-k+1}(p_L(k)) = \frac{1-\delta}{2}.$$

We obtain $p_L(k) = B^{-1}\left(k, n-k+1; \frac{1-\delta}{2}\right)$.

Similarly, we obtain $p_U(k) = B^{-1}\left(k+1, n-k; \frac{1+\delta}{2}\right)$.

Hence, the confidence interval for p at the confidence level δ is of the form

$$P_p\left\{B^{-1}\left(\eta, n-\eta+1; \frac{1-\delta}{2}\right) \le p \le B^{-1}\left(\eta+1, n-\eta; \frac{1+\delta}{2}\right)\right\} \ge \delta, \quad \text{for all } p \in (0,1).$$

The actual confidence level is higher than the nominal one because of discreteness of binomial distribution (see for example Brown et al. 2001).

Table 1. Pareto distribution

b	ARPR	n=20		n=200		n=2000	
		$\hat{\gamma}$	$\hat{\Delta}$	$\hat{\gamma}$	$\hat{\Delta}$	$\hat{\gamma}$	$\hat{\Delta}$
0.100	0.4738	0.9815	0.1804	0.9567	0.0473	0.9495	0.0136
0.111	0.4709	0.9761	0.1820	0.9476	0.0495	0.9448	0.0142
0.125	0.4672	0.9685	0.1835	0.9474	0.0521	0.9373	0.0148
0.143	0.4625	0.9917	0.1948	0.9259	0.0540	0.9433	0.0158
0.167	0.4565	0.9872	0.1991	0.9262	0.0580	0.9387	0.0168
0.200	0.4485	0.9795	0.2039	0.9093	0.0612	0.9311	0.0180
0.250	0.4377	0.9925	0.2170	0.9701	0.0656	0.9342	0.0196
0.333	0.4228	0.9849	0.2263	0.9442	0.0704	0.9358	0.0214
0.500	0.4024	0.9928	0.2435	0.9550	0.0776	0.9368	0.0234
1	0.3750	0.9800	0.2572	0.9531	0.0842	0.9453	0.0258
2	0.3585	0.9926	0.2705	0.9486	0.0870	0.9482	0.0269
3	0.3526	0.9908	0.2730	0.9578	0.0889	0.9465	0.0272
4	0.3496	0.9897	0.2742	0.9584	0.0893	0.9458	0.0273
5	0.3477	0.9890	0.2748	0.9543	0.0891	0.9494	0.0275
6	0.3465	0.9885	0.2753	0.9565	0.0896	0.9490	0.0276
7	0.3456	0.9576	0.2709	0.9573	0.0898	0.9527	0.0277
8	0.3450	0.9579	0.2712	0.9574	0.0899	0.9486	0.0276
9	0.3444	0.9582	0.2715	0.9574	0.0900	0.9485	0.0276
10	0.3440	0.9584	0.2717	0.9572	0.0900	0.9524	0.0278

Table 2. Gamma distribution

b	ARPR	n=20		n=200		n=2000	
		$\hat{\gamma}$	$\hat{\Delta}$	$\hat{\gamma}$	$\hat{\Delta}$	$\hat{\gamma}$	$\hat{\Delta}$
0.100	0.4751	0.9887	0.1818	0.9828	0.0462	0.9581	0.0134
0.111	0.4724	0.9851	0.1837	0.9544	0.0470	0.9627	0.0141
0.125	0.4691	0.9799	0.1858	0.9666	0.0497	0.9587	0.0148
0.143	0.4649	0.9965	0.1951	0.9714	0.0525	0.9527	0.0155
0.167	0.4595	0.9941	0.2000	0.9749	0.0562	0.9513	0.0166
0.200	0.4521	0.9894	0.2058	0.9750	0.0602	0.9533	0.0178
0.250	0.4416	0.9791	0.2124	0.9590	0.0642	0.9566	0.0195
0.333	0.4257	0.9911	0.2294	0.9665	0.0710	0.9544	0.0215
0.500	0.3986	0.9937	0.2498	0.9537	0.0786	0.9544	0.0242
1	0.3402	0.9600	0.2734	0.9563	0.0906	0.9514	0.0279
2	0.2668	0.9692	0.2916	0.9492	0.0959	0.9477	0.0297
3	0.2178	0.9649	0.2890	0.9484	0.0953	0.9433	0.0294
4	0.1813	0.9781	0.2866	0.9545	0.0932	0.9416	0.0285
5	0.1527	0.9802	0.2764	0.9518	0.0893	0.9432	0.0274
6	0.1297	0.9627	0.2609	0.9502	0.0851	0.9420	0.0260
7	0.1109	0.9807	0.2562	0.9623	0.0819	0.9461	0.0248
8	0.0952	0.9900	0.2493	0.9491	0.0767	0.9458	0.0235
9	0.0820	0.9748	0.2352	0.9473	0.0726	0.9452	0.0221
10	0.0709	0.9850	0.2297	0.9467	0.0685	0.9439	0.0209

Table 3. Lognormal distribution

b	ARPR	n=20		n=200		n=2000	
		$\hat{\gamma}$	$\hat{\Delta}$	$\hat{\gamma}$	$\hat{\Delta}$	$\hat{\gamma}$	$\hat{\Delta}$
0.100	$1.63\ 10^{-7}$	1.0000	0.1542	1.0000	0.0181	0.9997	0.0018
0.111	$2.14\ 10^{-6}$	0.9999	0.1542	0.9995	0.0181	0.9957	0.0018
0.125	$2.19\ 10^{-5}$	0.9992	0.1541	0.9953	0.0180	0.9990	0.0019
0.143	0.0002	0.9942	0.1534	0.9993	0.0184	0.9944	0.0021
0.167	0.0011	0.9687	0.1494	0.9773	0.0193	0.9752	0.0033
0.200	0.0053	0.9895	0.1606	0.9728	0.0256	0.9534	0.0065
0.250	0.0205	0.9839	0.1797	0.9532	0.0415	0.9536	0.0122
0.333	0.0627	0.9866	0.2243	0.9513	0.0654	0.9382	0.0197
0.500	0.1535	0.9727	0.2743	0.9413	0.0884	0.9313	0.0271
1	0.3047	0.9748	0.2853	0.9435	0.0933	0.9390	0.0288
2	0.3992	0.9924	0.2472	0.9450	0.0777	0.9460	0.0240
3	0.4324	0.9931	0.2229	0.9682	0.0685	0.9518	0.0206
4	0.4492	0.9861	0.2060	0.9503	0.0600	0.9541	0.0183
5	0.4593	0.9936	0.1985	0.9720	0.0560	0.9530	0.0166
6	0.4661	0.9966	0.1926	0.9728	0.0518	0.9523	0.0153
7	0.4709	0.9829	0.1835	0.9663	0.0486	0.9548	0.0143
8	0.4745	0.9880	0.1810	0.9804	0.0463	0.9535	0.0135
9	0.4774	0.9913	0.1789	0.9736	0.0439	0.9577	0.0128
10	0.4796	0.9935	0.1770	0.9617	0.0415	0.9530	0.0121

REFERENCES

BROWN L. D, CAI T. T, DASGUPTA A. (2001) Interval estimation for Binomial proportion, Statistical Science, 16, 101–133.

DAVID H. A., NAGARAJA H. N. (2003) Order Statistics, Third Edition, Wiley.

ZIELIŃSKI R. (2006) Exact distribution of the natural ARPR estimator in small samples from infinite populations, Statistics In Transition, 7, 881–888.

ZIELIŃSKI R. (2007) A confidence interval for ARPR – "at-risk-of-poverty-rate", Statistics In Transition, 8, 217–222.

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 445—456

KURTOSIS OF A RANDOM VECTOR SPECIAL TYPES OF DISTIBUTIONS

Katarzyna Budny[1], Jan Tatar[2]

ABSTRACT

In presented paper the authors attempt to generalize definition of kurtosis for the case of multidimensional and prove its essential properties. The generalized characteristic applied in the single-dimension case has the same properties as kurtosis, that is known in the literature on single-dimensional random variables. The basis of conducted considerations is the definition of *the power of a vector in space with the scalar product.*

Key words: Kurtosis, Moments of a random vector, Multidimensional distribution, Power of a vector.

1. Introduction

This paper is a continuation of previous studies by Tatar (1996,1999, 2000a, 2000b,2002, 2003), Osiewalski and Tatar (1999). Let us recoll:

Let there be any vector space $(R^n, R, +, \cdot)$ with the classical (Euclidean) scalar product form

$$\forall v = (v_1, v_2, ..., v_n), w = (w_1, w_2, ..., w_n) \in R^n : \ \langle v, w \rangle = \sum_{i=1}^{n} v_i w_i \ .$$

For any $v \in R^n$ and any number $k \in N_0 = N \cup \{0\}$ a following definition of *k*-th power of the vector *v* has been proposed.

Definition 1

$$v^o = 1 \in R$$

and

[1] Cracow University of Economics, Department of Mathematics budnyk@uek.krakow.pl.
[2] Cracow University of Economics, Department of Mathematics tatarj@uek.krakow.pl.

$$v^k = \begin{cases} v^{k-1} \cdot v & , \quad for \;\; k-odd \\ \left\langle v^{k-1}, v \right\rangle & , \quad for \; k-even \end{cases}$$

From the above-mentioned definition immediately result two following important properties

$$\forall v \in R^n \;\; , \;\; k \in N_0 \;\; : \;\; k-even \quad \Rightarrow \;\; v^k \in R$$
$$\forall v \in R^n \;\; , \;\; k \in N \;\; : \;\; k-odd \;\; \Rightarrow \;\; v^k \in R^n \; .$$

Definition 2

For any number $k \in N_0$ an ordinary moment of the order k of a random vector ξ we call the expected value of a random variables $g(\xi) = \xi^k$, it means $m_k = E(\xi^k)$, if $E(|\xi^k|) < +\infty$.

Particularly, for $n = 2$ and for any $k \in N$ we have (1)

$$m_k = \begin{cases} \sum_{i=0}^{k/2} m_{k-2i,2i} & , \; for \;\; k \text{ - even}, \\ \left(\sum_{i=0}^{(k-1)/2} \binom{k-1}{2}{i} m_{k-2i,2i} \; , \; \sum_{i=0}^{(k-1)/2} \binom{k-1}{2}{i} m_{k-2i,2i+1} \right) & , \; for \;\; k-odd \end{cases}$$

The symbols like $m_{k-2i,2i}$ used in the formula (1) means the classical mixed ordinary moments of the rank "$k-2i,2i$ ".

Similarly to the ordinary moments, we define central moments of multidimensional distribution. We only need to assume in the definition 2, that $g(\xi) = (\xi - E(\xi))^k$.

From the definition 1 and 2 results inter alia the following general conclusion:

$$\forall k \in N_0 : k \text{ - even} \Rightarrow m_k \in R$$

$$\forall k \in N : k \text{ - odd} \Rightarrow m_k \in R^n \quad .$$

In previously cited papers, on the basis of definition of the power of a vector and the ordinary moment of a random vector, important results for characterization of multidimensional probability distributions were obtained. Inter alia *Chebyshev's Inequality* known for single-dimension case was generalized, multidimensional versions of the *law of large numbers* were proposed, there were defined and applied tools which help to analyze and measure the *asymmetry of*

multidimensional probability distributions, the definition of *covariance and correlation coefficient* was generalized as well. We would like to emphasize that proposed and developed conception essentially and directly provides us with characteristics of analyzed random vector, not – properly constructed (single-dimensional) – functions of its terms.

To exercise our previous declaration, let's move to the proposition of measurement of concentration (which means flattening as well) probability distribution of a random vector.

2. Propositions

At the beginning let's recollect, that in the theory of single-dimensional random variables in capacity of concentration distribution measure around expected value it is assumed, inter alia, its fourth central moment. Usually – in order to obtain relative measure – the fourth order central moment is furthermore divided by square variance. The high values of indicator constructed in this way (*kurtosis*) indicate that there is the significant tendency to focus the distribution around its expected value and vice versa: indicators' low level shows its flattening.

The definition of random vector central moment let us to definite the kurtosis in a similar way for the multidimensional distribution.

Definition 3

The kurtosis of a random vector $X=(X_1,...,X_n):\Omega\to R^n$ the following value will be called:

$$\mathrm{Kurt}X=\frac{E\left[(X-EX)^4\right]}{\left(D^2X\right)^2}\quad . \tag{2}$$

In the proofs of theorems 1 and 2, which we'll form in the further part of this paper, we'll use the following lemma.

Lemma 1

Let $X=(X_1,...,X_n):\Omega\to R^n$ be a random vector fulfilling conditions:

a) $X_i\sim N\left(m_i,\sigma_i^2\right)$, for any $i\in\{1,2,...,n\}$

b) for all $i,j\in\{1,2,...,n\}$: if $i\neq j$, than X_i and X_j are stochastically independent (symbolically: $X_i\perp X_j$).

Then we have the following equalities:

$$1.\ E\left(\langle X,X\rangle^2\right)=\sum_{\substack{i,j=1\\ i\neq j}}^{n}\left(m_i^2+\sigma_i^2\right)\left(m_j^2+\sigma_j^2\right)+\sum_{i=1}^{n}\left(3\sigma_i^4+6m_i^2\sigma_i^2+m_i^4\right),$$

2. $E\left(\langle X,X\rangle\langle X,EX\rangle\right)=\sum\limits_{\substack{i,j=1\\ i\neq j}}^{n} m_j^2\left(m_i^2+\sigma_i^2\right)+\sum\limits_{i=1}^{n}\left(3m_i^2\sigma_i^2+m_i^4\right)$,

3. $E\left(\langle X,EX\rangle^2\right)=\sum\limits_{\substack{i,j=1\\ i\neq j}}^{n} m_i^2 m_j^2+\sum\limits_{i=1}^{n}\left(m_i^4+m_i^2\sigma_i^2\right)$,

4. $\langle EX,EX\rangle E\left(\langle X,X\rangle\right)=\sum\limits_{i,j=1}^{n}\left(m_i^2m_j^2+m_i^2\sigma_j^2\right)$,

5. $\langle EX,EX\rangle^2=\sum\limits_{i,j=1}^{n} m_i^2m_j^2$.

Proof:

Let's recollect, that if a random variable $\xi:\Omega\to R$ has a normal distribution with a mean of *m* and a variance of σ^2, that is $\xi\sim N\left(m,\sigma^2\right)$, then

$$E\left(\xi^2\right)=m^2+\sigma^2\ ,\quad E\left(\xi^3\right)=3m\sigma^2+m^3$$

$$\text{and}\quad E\left(\xi^4\right)=3\sigma^4+6m^2\sigma^2+m^4. \tag{3}$$

While using the properties of integral (alternatively sum) and the independence of random variables X_i^2 and X_j^2 for all $i,j\in\{1,2,...,n\}$, such that $i\neq j$ we get

$$E\left(\langle X,X\rangle^2\right)=E\left[\left(\sum_{i=1}^{n}X_i^2\right)\left(\sum_{j=1}^{n}X_j^2\right)\right]=\sum_{i,j=1}^{n}E\left(X_i^2X_j^2\right)= \tag{4}$$

$$=\sum_{\substack{i,j=1\\ i\neq j}}^{n}E\left(X_i^2\right)E\left(X_j^2\right)+\sum_{i=1}^{n}E\left(X_i^4\right).$$

Thus using (3) the equality (4) takes form:

$$E\left(\langle X,X\rangle^2\right)=\sum_{\substack{i,j=1\\ i\neq j}}^{n}\left(m_i^2+\sigma_i^2\right)\left(m_j^2+\sigma_j^2\right)+\sum_{i=1}^{n}\left(3\sigma_i^4+6m_i^2\sigma_i^2+m_i^4\right),$$

And that ends the proof for thesis 1.

Successively from the independence of random variables X_i^2 and X_j and form (3) of the ordinary moments of the second and third order a random variable with normal distribution, we get thesis 2.

In fact:

$$E\left(\langle X,X\rangle\langle X,EX\rangle\right)=E\left[\left(\sum_{i=1}^{n}X_i^2\right)\left(\sum_{j=1}^{n}X_j m_j\right)\right]=\sum_{i,j=1}^{n}E\left(X_i^2X_j m_j\right)=$$

$$=\sum_{\substack{i,j=1\\ i\neq j}}^{n}m_j E\left(X_i^2\right)E\left(X_j\right)+\sum_{i=1}^{n}m_i E\left(X_i^3\right)=\sum_{\substack{i,j=1\\ i\neq j}}^{n}m_j^2\left(m_i^2+\sigma_i^2\right)+\sum_{i=1}^{n}m_i\left(3m_i\sigma_i^2+m_i^3\right)=$$

$$=\sum_{\substack{i,j=1\\ i\neq j}}^{n}m_j^2\left(m_i^2+\sigma_i^2\right)+\sum_{i=1}^{n}\left(3m_i^2\sigma_i^2+m_i^4\right).$$

In order to prove the thesis 3 we use the independence of random variables X_i and X_j as well. So there is:

$$E\left(\langle X,EX\rangle^2\right)=E\left[\left(\sum_{i=1}^{n}X_iEX_i\right)\left(\sum_{j=1}^{n}X_jEX_j\right)\right]=\sum_{i,j=1}^{n}E\left(X_iX_jE\left(X_i\right)E\left(X_j\right)\right)=$$

$$=\sum_{i,j=1}^{n}E\left(X_i\right)E\left(X_j\right)E\left(X_iX_j\right)=\sum_{\substack{i,j=1\\ i\neq j}}^{n}m_i^2m_j^2+\sum_{i=1}^{n}\left(m_i^4+m_i^2\sigma_i^2\right).$$

Finally, for thesis 4 we only need to conduct the following reasoning:

$$\langle EX,EX\rangle E\left(\langle X,X\rangle\right)=\left(\sum_{i=1}^{n}m_i^2\right)E\left(\sum_{j=1}^{n}X_j^2\right)=\left(\sum_{i=1}^{n}m_i^2\right)\left(\sum_{j=1}^{n}E\left(X_j^2\right)\right)=$$

$$=\left(\sum_{i=1}^{n}m_i^2\right)\left(\sum_{j=1}^{n}\left(m_j^2+\sigma_j^2\right)\right)=\sum_{i,j=1}^{n}\left(m_i^2\left(m_j^2+\sigma_j^2\right)\right)=\sum_{i,j=1}^{n}\left(m_i^2m_j^2+m_i^2\sigma_j^2\right).$$

Thesis 5 is in this way obviously fulfilled.

Proof of the lemma has been finished.

At the moment, the theorem concerning the kurtosis of a random vector with normal marginal distributions, will be defined and proved.

Theorem 1

If $X=\left(X_1,...,X_n\right):\Omega\to R^n$ is a random vector fulfilling lemma 1 assumptions, then

$$KurtX=1+\frac{2\sum_{i=1}^{n}\sigma_i^4}{\sum_{i,j=1}^{n}\sigma_i^2\sigma_j^2}=1+\frac{2\sum_{i=1}^{n}\left(D^2X_i\right)^2}{\sum_{i,j=1}^{n}D^2X_iD^2X_j}. \qquad (5)$$

Proof:

Using the definition of even power of the vector, properties of scalar product and integral (sum), we get:

$$E\left[\left(X-EX\right)^4\right]=$$

$$E\left[\left\langle X-EX,X-EX\right\rangle^2\right]=E\left[\left(\left\langle X,X\right\rangle-2\left\langle X,EX\right\rangle+\left\langle EX,EX\right\rangle\right)^2\right]=$$

$$E\left[\left\langle X,X\right\rangle^2-4\left\langle X,X\right\rangle\left\langle X,EX\right\rangle+4\left\langle X,EX\right\rangle^2+2\left\langle X,X\right\rangle\left\langle EX,EX\right\rangle\right]+$$

$$-E\left[4\left\langle X,EX\right\rangle\cdot\left\langle EX,EX\right\rangle+\left\langle EX,EX\right\rangle^2\right].$$

Let's notice, that we also have equality:

$$E\left(\left\langle X,EX\right\rangle\left\langle EX,EX\right\rangle\right)=\left(\left\langle EX,EX\right\rangle\right)^2 \tag{6}$$

Indeed:

$$E\left(\left\langle X,EX\right\rangle\left\langle EX,EX\right\rangle\right)=$$

$$=E\left[\left(\sum_{i=1}^{n}X_iEX_i\right)\left(\sum_{j=1}^{n}\left(EX_j\right)^2\right)\right]=E\left[\left(\sum_{i=1}^{n}X_im_i\right)\left(\sum_{j=1}^{n}m_j^2\right)\right]=$$

$$=\left(\sum_{j=1}^{n}m_j^2\right)\left(\sum_{i=1}^{n}m_iEX_i\right)=\left(\sum_{j=1}^{n}m_j^2\right)\left(\sum_{i=1}^{n}m_i^2\right)=\left(\sum_{k=1}^{n}m_k^2\right)^2=\left(\left\langle EX,EX\right\rangle\right)^2.$$

Taking into consideration the equality (6), the fourth central moment can be shown in the following way

$$E\left[\left(X-EX\right)^4\right]=E\left(\left\langle X,X\right\rangle^2\right)-4E\left(\left\langle X,X\right\rangle\left\langle X,EX\right\rangle\right)+4E\left(\left\langle X,EX\right\rangle^2\right)+ \tag{7}$$

$$+2\left\langle EX,EX\right\rangle E\left(\left\langle X,X\right\rangle\right)-3\left\langle EX,EX\right\rangle^2.$$

Successively, from the lemma 1, equality (7) takes form:

$$E\left[\left(X-EX\right)^4\right]=\sum_{i,j=1}^{n}\sigma_i^2\sigma_j^2+2\sum_{i=1}^{n}\sigma_i^4\ . \tag{8}$$

Indeed:

$$E\left[(X-EX)^4\right]=\sum_{\substack{i,j=1\\ i\neq j}}^{n}\left(m_i^2+\sigma_i^2\right)\left(m_j^2+\sigma_j^2\right)+\sum_{i=1}^{n}\left(3\sigma_i^4+6m_i^2\sigma_i^2+m_i^4\right)+$$

$$-4\left(\sum_{\substack{i,j=1\\ i\neq j}}^{n}m_j^2\left(m_i^2+\sigma_i^2\right)+\sum_{i=1}^{n}\left(3m_i^2\sigma_i^2+m_i^4\right)\right)+4\left(\sum_{\substack{i,j=1\\ i\neq j}}^{n}m_i^2m_j^2+\sum_{i=1}^{n}\left(m_i^4+m_i^2\sigma_i^2\right)\right)+$$

$$+2\left(\sum_{i,j=1}^{n}\left(m_i^2m_j^2+m_i^2\sigma_j^2\right)\right)-3\sum_{i,j=1}^{n}m_i^2m_j^2=$$

$$=2\sum_{\substack{i,j=1\\ i\neq j}}^{n}m_i^2\sigma_j^2+\sum_{\substack{i,j=1\\ i\neq j}}^{n}m_i^2m_j^2+\sum_{\substack{i,j=1\\ i\neq j}}^{n}\sigma_i^2\sigma_j^2+\sum_{i=1}^{n}\sigma_i^4+2\sum_{i=1}^{n}\sigma_i^4+2\sum_{i=1}^{n}m_i^2\sigma_i^2+$$

$$+4\sum_{i=1}^{n}m_i^2\sigma_i^2+\sum_{i=1}^{n}m_i^4-4\left(\sum_{\substack{i,j=1\\ i\neq j}}^{n}m_j^2\left(m_i^2+\sigma_i^2\right)+\sum_{i=1}^{n}\left(3m_i^2\sigma_i^2+m_i^4\right)\right)+$$

$$+4\left(\sum_{\substack{i,j=1\\ i\neq j}}^{n}m_i^2m_j^2+\sum_{i=1}^{n}\left(m_i^4+m_i^2\sigma_i^2\right)\right)+2\left(\sum_{i,j=1}^{n}\left(m_i^2m_j^2+m_i^2\sigma_j^2\right)\right)-3\sum_{i,j=1}^{n}m_i^2m_j^2=$$

$$=2\sum_{\substack{i,j=1\\ i\neq j}}^{n}m_i^2\sigma_j^2+2\sum_{i=1}^{n}m_i^2\sigma_i^2+\sum_{\substack{i,j=1\\ i\neq j}}^{n}m_i^2m_j^2+\sum_{i=1}^{n}m_i^4+\sum_{\substack{i,j=1\\ i\neq j}}^{n}\sigma_i^2\sigma_j^2+\sum_{i=1}^{n}\sigma_i^4+$$

$$+2\sum_{i=1}^{n}\sigma_i^4+4\sum_{i=1}^{n}m_i^2\sigma_i^2-4\left(\sum_{i,j=1}^{n}m_i^2m_j^2+\sum_{i,j=1}^{n}m_i^2\sigma_j^2+2\sum_{i=1}^{n}m_i^2\sigma_i^2\right)+$$

$$+4\left(\sum_{i,j=1}^{n}m_i^2m_j^2+\sum_{i=1}^{n}m_i^2\sigma_i^2\right)+2\sum_{i,j=1}^{n}m_i^2\sigma_j^2+2\sum_{i,j=1}^{n}m_i^2m_j^2-3\sum_{i,j=1}^{n}m_i^2m_j^2=$$

$$=2\sum_{i,j=1}^{n}m_i^2\sigma_j^2+\sum_{i,j=1}^{n}m_i^2m_j^2+\sum_{i,j=1}^{n}\sigma_i^2\sigma_j^2+2\sum_{i=1}^{n}\sigma_i^4+4\sum_{i=1}^{n}m_i^2\sigma_i^2+$$

$$-4\sum_{i,j=1}^{n}m_j^2\sigma_i^2-8\sum_{i=1}^{n}m_i^2\sigma_i^2+4\sum_{i=1}^{n}m_i^2\sigma_i^2+2\sum_{i,j=1}^{n}m_i^2\sigma_j^2+2\sum_{i,j=1}^{n}m_i^2m_j^2-3\sum_{i,j=1}^{n}m_i^2m_j^2=$$

$$=4\sum_{i,j=1}^{n}m_i^2\sigma_j^2-4\sum_{i,j=1}^{n}m_j^2\sigma_i^2+4\sum_{i=1}^{n}m_i^2\sigma_i^2-8\sum_{i=1}^{n}m_i^2\sigma_i^2+4\sum_{i=1}^{n}m_i^2\sigma_i^2+$$

$$+\sum_{i,j=1}^{n} m_i^2 m_j^2 + 2\sum_{i,j=1}^{n} m_i^2 m_j^2 - 3\sum_{i,j=1}^{n} m_i^2 m_j^2 + \sum_{i,j=1}^{n} \sigma_i^2 \sigma_j^2 + 2\sum_{i=1}^{n} \sigma_i^4 = \sum_{i,j=1}^{n} \sigma_i^2 \sigma_j^2 + 2\sum_{i=1}^{n} \sigma_i^4 .$$

Furthermore, the following sequence of equalities is true:

$$\left(D^2 X\right)^2 = \left(\sum_{i=1}^{n} D^2 X_i\right)^2 = \left(\sum_{i=1}^{n} \sigma_i^2\right)^2 = \sum_{i,j=1}^{n} \sigma_i^2 \sigma_j^2 . \qquad (9)$$

In the view of (8), (9) and from definition 3, kurtosis of a random vector, fulfilling lemma 1 assumptions, can be expressed as:

$$KurtX = 1 + \frac{2\sum_{i=1}^{n} \sigma_i^4}{\sum_{i,j=1}^{n} \sigma_i^2 \sigma_j^2} .$$

And that finishes the proof of theorem 1.

It's easy to notice that simple and immediate consequence of theorem 1 is the following conclusion.

Conclusion 1

If $X = (X_1,...,X_n): \Omega \to R^n$ is a random vector fulfilling assumptions of lemma 1, and if there is also condition $\sigma_1 = \sigma_2 = ... = \sigma_n = \sigma$, then

$$KurtX = 1 + \frac{2}{n} .$$

From the conclusion above also comes out – quite obvious, but it let us to call the submitted proposition as "generalization" – a remark that kurtosis of a single-dimensional random variable, normally distributed equals 3.

In order to realize the purpose formulated at the beginning of this paper, in the next theorem we'll get the form of kurtosis of a random vector, which marginal variables have a Student's t-distributions.

Theorem 2

Let $X = (X_1,...,X_n): \Omega \to R^n$ be a random vector fulfilling following conditions:

a) $X_i \sim t_{\nu_i}$ (that is X_i has a Student's t-distribution with ν_i degrees of freedom) where $\nu_i > 4$, for each $i \in \{1,2,...,n\}$,

b) for all $i, j \in \{1,2,...,n\}$: if $i \neq j$, then X_i and X_j are stochastically independent (symbolically: $X_i \perp X_j$).

Then

$$KurtX = 1 + \frac{\sum_{i=1}^{n}\left(2+\frac{6}{\nu_i-4}\right)\left(\frac{\nu_i}{\nu_i-2}\right)^2}{\sum_{i,j=1}^{n}\left(\frac{\nu_i}{\nu_i-2}\right)\left(\frac{\nu_j}{\nu_j-2}\right)} \quad . \tag{10}$$

Proof:

We'll conduct the reasoning similar to the one in the proof for theorem 1. But first, let's recollect, that if random variable $\xi : \Omega \to R$ has a Student's t-distribution with *v* degrees of freedom (else: $\xi \sim t_\nu$), where $\nu > 4$, then:

$$E(\xi) = 0 \ , \quad E(\xi^2) = D^2\xi = \frac{\nu}{\nu-2} \ , \quad E(\xi^3) = 0 \ ,$$

$$E(\xi^4) = \left(3+\frac{6}{\nu-4}\right)\left(\frac{\nu}{\nu-2}\right)^2$$

and $Kurt\,\xi = 3 + \frac{6}{\nu-4}$, $Excess\ \xi = \frac{6}{\nu-4}$.

We also need to notice, that for vector *X*, which was discussed in the assumption of theorem 2, we have the equality $EX = (EX_1, EX_2, ..., EX_n) = (0,0,...,0)$. So here emerges conclusion that the thesis of lemma is equivalent for lemma 1 (with the difference that now we assume that the marginal variables of a random vector *X* have a Student's t-distribution) takes form:

1. $E(\langle X, X\rangle^2) = \sum_{\substack{i,j=1 \\ i\neq j}}^{n}\left(\frac{\nu_i}{\nu_i-2}\right)\left(\frac{\nu_j}{\nu_j-2}\right) + \sum_{i=1}^{n}\left(3+\frac{6}{\nu_i-4}\right)\left(\frac{\nu_i}{\nu_i-2}\right)^2$,
2. $E(\langle X, X\rangle \cdot \langle X, EX\rangle) = 0$,
3. $E(\langle X, EX\rangle^2) = 0$,
4. $\langle EX, EX\rangle \cdot E(\langle X, X\rangle) = 0$,
5. $\langle EX, EX\rangle^2 = 0$.

Then:

$$E\left[\left(X-EX\right)^4\right]=E\left(\left\langle X,X\right\rangle^2\right)= \tag{11}$$

$$=\sum_{\substack{i,j=1\\ i\neq j}}^{n}\left(\frac{\nu_i}{\nu_i-2}\right)\left(\frac{\nu_j}{\nu_j-2}\right)+\sum_{i=1}^{n}\left(3+\frac{6}{\nu_i-4}\right)\left(\frac{\nu_i}{\nu_i-2}\right)^2 .$$

and

$$\left(D^2X\right)^2=\left(\sum_{i=1}^{n}\frac{\nu_i}{\nu_i-2}\right)^2=\sum_{i,j=1}^{n}\left(\frac{\nu_i}{\nu_i-2}\right)\left(\frac{\nu_j}{\nu_j-2}\right). \tag{12}$$

So from (11) and (12) and from definition 3 we get:

$$KurtX=\frac{\sum_{\substack{i,j=1\\ i\neq j}}^{n}\left(\frac{\nu_i}{\nu_i-2}\right)\left(\frac{\nu_j}{\nu_j-2}\right)+\sum_{i=1}^{n}\left(3+\frac{6}{\nu_i-4}\right)\left(\frac{\nu_i}{\nu_i-2}\right)^2}{\sum_{i,j=1}^{n}\left(\frac{\nu_i}{\nu_i-2}\right)\left(\frac{\nu_j}{\nu_j-2}\right)}=$$

$$=\frac{\sum_{i,j=1}^{n}\left(\frac{\nu_i}{\nu_i-2}\right)\left(\frac{\nu_j}{\nu_j-2}\right)-\sum_{i=1}^{n}\left(\frac{\nu_i}{\nu_i-2}\right)^2+\sum_{i=1}^{n}\left(3+\frac{6}{\nu_i-4}\right)\left(\frac{\nu_i}{\nu_i-2}\right)^2}{\sum_{i,j=1}^{n}\left(\frac{\nu_i}{\nu_i-2}\right)\left(\frac{\nu_j}{\nu_j-2}\right)}=$$

$$=\frac{\sum_{i,j=1}^{n}\left(\frac{\nu_i}{\nu_i-2}\right)\left(\frac{\nu_j}{\nu_j-2}\right)+\sum_{i=1}^{n}\left(2+\frac{6}{\nu_i-4}\right)\left(\frac{\nu_i}{\nu_i-2}\right)^2}{\sum_{i,j=1}^{n}\left(\frac{\nu_i}{\nu_i-2}\right)\left(\frac{\nu_j}{\nu_j-2}\right)}=$$

$$=1+\frac{\sum_{i=1}^{n}\left(2+\frac{6}{\nu_i-4}\right)\left(\frac{\nu_i}{\nu_i-2}\right)^2}{\sum_{i,j=1}^{n}\left(\frac{\nu_i}{\nu_i-2}\right)\left(\frac{\nu_j}{\nu_j-2}\right)},$$

and that is the required thesis (10).

Also from theorem 2 results the following conclusion:

Conclusion 2

If $X = (X_1,...,X_n) : \Omega \to R^n$ is a random vector fulfilling assumptions of theorem 2 and if we have condition $\nu_1 = \nu_2 = ... = \nu_n = \nu > 4$ as well, then

$$KurtX = 1 + \frac{2 + \frac{6}{\nu - 4}}{n} .$$

Proof:

We only need to notice that in the view of equal number of degrees of freedom of all marginal variables the following equalities - from theorem 2 – are true:

$$KurtX = 1 + \frac{n\left(2 + \frac{6}{\nu - 4}\right)\left(\frac{\nu}{\nu - 2}\right)^2}{\left(n\frac{\nu}{\nu - 2}\right)^2} = 1 + \frac{2 + \frac{6}{\nu - 4}}{n} .$$

Obviously for n=1 the kurtosis of a random variable with a Student's t-distribution depends only on a number of its degrees of freedom and equals

$$KurtX = 3 + \frac{6}{\nu - 4} .$$

Let's notice that from conclusion 1 and 2 we obviously get:

Conclusion 3

If $X = (X_1,...,X_n) : \Omega \to R^n$ is a random vector fulfilling assumptions of lemma 1, but $\sigma_1 = \sigma_2 = ... = \sigma_n = \sigma$, and if $Y = (Y_1,...,Y_n) : \Omega \to R^n$ is a random vector fulfilling assumptions of theorem 2 and we have condition $\nu_1 = \nu_2 = ... = \nu_n = \nu > 4$, then $KurtY > KurtX$.

REFERENCES

FELLER, W., 1969. *Wstęp do rachunku prawdopodobieństwa, Volume 1 and 2*. Warszawa: PWN.

FISZ, M., 1969. *Rachunek prawdopodobieństwa i statystyka matematyczna*. Warszawa: PWN.

JAKUBOWSKI, J., SZTENCEL, R., 2004. *Wstęp do rachunku prawdopodobieństwa*. 3rd ed. Warszawa: Script.

OSIEWALSKI, J., TATAR, J., 1999. Przegląd Statystyczny. *Multivariate Chebyshev inequality based on a new definition of moments of a random vector*, 2.

PLUCIŃSKA, A., PLUCIŃSKI, E., 2006. *Probabilistyka. Rachunek prawdopodobieństwa, statystyka matematyczna, procesy stochastyczne*. Warszawa: WNT.

TATAR, J., 1996. Przegląd Statystyczny. *O niektórych miarach rozproszenia rozkładów prawdopodobieństwa*, 3/4 .

TATAR, J., 1999. Przegląd Statystyczny. *Moments of a random variable in a Hilbert space*, 2.

TATAR, J., (2000a), *Nowa charakteryzacja wielowymiarowych rozkładów prawdopodobieństwa.* Sprawozdanie z badań statutowych; um. nr: 92/KM/1/99/S; AE Kraków.

TATAR, J., 2000b. Momenty absolutne wielowymiarowych rozkładów prawdopodobieństwa. In: Polska Akademia Nauk, *Posiedzenie Komisji Statystyczno-Demograficznej PAN, O/Kraków* . Cracow, 17 November 2000.

TATAR, J., 2002. Nowe miary zależności wektorów losowych. In: Polska Akademia Nauk, *Posiedzenie Komisji Statystyczno-Demograficznej PAN, O/Kraków.* Cracow 22 May 2002.

TATAR, J., 2003. Prace naukowe AE we Wrocławiu. *Prawa wielkich liczb dla wielowymiarowych wektorów losowych*, 1006.

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 457—464

COMBINED EFFECT OF FAULT DETECTION AND FAULT INTRODUCTION RATE ON SOFTWARE RELIABILITY MODELLING

S. Chatterjee[1], L.N. Upadhyaya[2], J.B. Singh [3]and S. Nigam [4]

ABSTRACT

This paper proposes a software reliability growth model to study the combined effect of increasing error detection and decreasing error introduction rate under imperfect debugging. The model is developed based on non homogeneous Poisson process (NHPP) and can be used to estimate and predict the reliability as well as the cost of a software product. Some real life data has been used to validate the proposed model and to show its usefulness. Comparison of this model with other has been carried out.

Key words: Software Reliability, Imperfect Debugging, Multiple Failures, Non Homogeneous Poisson Process, Increasing Fault Detection Rate, Decreasing Fault Removal Rate.

Nomenclature

N (t) – counting process representing the cumulative number of errors
Poim (m(t)) – Poisson process with mean m(t)
A – mean initial error content in the software
b_i – fault detection rate per type i fault, i=1,2,3,…,k
p_i – content proportion of type i fault, i=1,2,3…,k
$\lambda_i(t)$ – fault detection rate per unit time of type i error, i=1,2,3,…,k
$\lambda(t)$ – fault detection rate per unit time
n(t) – number of errors to be eventually detected
β – error introduction rate
$m_i(t)$ – expected number of type i errors by time t, i=1,2,3,….,k
T – total test time i.e., release time
s_j – cumulative time

[1] Assistant Professor, Dept. of Applied Mathematics, ISMU, Dhanbad-826004, Corresponding author. Email: chatterjee_subhashis@rediffmail.com.
[2] Professor, Dept. of Applied Mathematics, ISMU, Dhanbad-826004.
[3] Research Scholar, Dept. of Applied Mathematics, ISMU, Dhanbad.
[4] Project Fellow, Department of Applied Mathematics, ISMU, Dhanbad.

R(x/t) – conditional reliability of software
mle – maximum likelihood estimate

Introduction

Computers are widely used in various fields of life including business and safety critical systems. Application of computer means the application of software. Therefore, there exists an increasing demand of highly reliable software. Since assessment of software reliability is one of the major concern in present-day software industries, one need a good mathematical model to estimate the reliability of software. Research in the area of software reliability has been going on since last three decades. Detail studies related to software reliability has been presented in [1,2,3]. Initially, various software reliability models were developed using the concept of perfect debugging process. Some of the important software reliability models among them are [4,5,6,7,8,9,10,11,12,13]. Latter stage various software reliability models were developed using the concept of imperfect debugging. Some of them also throws light to other aspects of software reliability studies like: estimation of cost, release time etc.[14,15,16,17,18,19,20,21,22].

Present-day software development process has became very complex. Therefore, there is a need to develop an efficient mathematical model which can give better prediction of software reliability and take care of different aspects of software development process as well. Among these aspects of software development process, error detection rate, i.e., FDR and error introduction rate, i.e., FIR during software debugging plays very crucial role in the growth of software reliability.

Pham[19] developed a software reliability model considering imperfect debugging and presence of multiple failures. He considered FDR and FIR both as constant and different for different types of errors. Also, he considered presence of three types of errors: type 1 errors, i.e., critical errors – which are very difficult to detect, type 2 errors, i.e., major errors – which are difficult to detect, type 3 errors, i.e., minor error – which are easy to detect. Software testing as well as debugging are done by human being. As time progress test personnel learns more about the software. As a result, FDR increases and FIR decreases with respect to time. In this paper the FIR β and FDR b are being considered as a function of cumulative time and a more realistic software reliability model has been developed to study the combined effect of increasing FDR and decreasing FIR on the growth of software reliability. As error introduction and detection depends on the knowledge of the test personnel, these factors cannot be different for different types of errors. Due to this reason, FDR has been considered the same for all types of errors and the same is the case for FIR. Also, in this paper, presence of k types of errors (for generalization) is considered. The proposed model is based on non homogeneous Poisson process and some real life data are used to validate the model.

Model development

A non homogeneous Poisson process based model considering imperfect debugging and presence of k types of error has been proposed in this section.

Model Assumptions

The following assumptions are made for the development of the model.

(i) During removal of detected errors, it is possible to introduce new errors.

(ii) The probability of finding an error in software is proportional to the number of remaining errors in the software.

(iii) There exist k types of errors in software.

(iv) The error detection process follows a non homogeneous Poisson process.

(v) Fault detection rate (FDR) b is same for k types of errors and increases with respect to cumulative time. Here b is considered as a logistic function, i.e., $b = \dfrac{1}{1+(\dfrac{1}{\alpha_0}-1)e^{-rs_i}}$, where, α_0 and r are constant, s_j is the cumulative time.

(vi) Fault introduction rate (FIR) β is same for k types of errors and decreases with respect to cumulative time, where $\beta = \dfrac{1}{1+s_j^2}$

For mathematical simplification the cumulative time s_j is considered here. Otherwise, analysis will be more complicated. According to the assumption (iv), the model has been formulated as $\Pr\{N(t)=n\} = poim(n, m(t)), n = 0,1,2.....$ To obtain the reliability of software the mean value function $m(t)$, i.e., the expected number of software failures to be determined. It is obtained by solving the following differential equations.

$$\frac{dm_i(t)}{dt} = \lambda_i(t) \quad , \quad \frac{dm_i(t)}{dt} = b[n_i(t) - m_i(t)]$$

$$\frac{dn_i(t)}{dt} = \beta \ \frac{dm_i(t)}{dt}, \ m(t) = \sum_{i=1}^{k} m_i(t) \ \text{ where } n_i(0) = ap_i, \ m_i(0) = 0$$

Solutions of the above differential equations give the expression for $m(t), \lambda(t)$ and $R(x/t)$ as follows.

$$m(t) = \sum_{i=1}^{k} \frac{ap_i[1-e^{-(1-\beta)bt}]}{(1-\beta)} \quad , \quad \lambda(t) = \sum_{i=1}^{k} ap_i be^{-(1-\beta)bt} \text{ and}$$

$$R(x/t) = e^{-[m(t+x)-m(t)]} = e^{-\sum_{i=1}^{k} m_i(x)e^{-(1-\beta)bt}}$$

Results & Discussion

To demonstrate the usefulness of the proposed model and for comparison with Pham model [19], presence of three types of errors, i.e., k=3 in a software is considered. Three types of errors are: type 1 errors, i.e., critical errors – which are very difficult to detect, type 2 errors, i.e., major errors – which are difficult to detect, type 3 errors, i.e., minor error – which are easy to detect. For illustration purpose the failure data of Misra [23] given in Table-I are used. Here $p_1 = 0.0173,\ p_2 = 0.342,\ p_3 = 0.6407$. The values of the constant r, α_0 are 1.5 and 0.05 respectively.

Table-I. Original failure data

Test Week	Test Hours	Critical errors	Major errors	Minor errors	Test Week	Test Hours	Critical errors	Major errors	Minor errors
1	62.5	0	6	9	20	25	0	2	3
2	44	0	2	4	21	12	0	1	1
3	40	0	1	7	22	55	0	3	2
4	68	1	1	6	23	49	0	2	4
5	62	0	3	5	24	64	0	4	5
6	66	0	1	3	25	26	0	1	0
7	73	0	2	2	26	66	0	2	2
8	73.5	0	3	5	27	49	0	2	0
9	92	0	2	4	28	52	0	2	2
10	71.4	0	0	2	29	70	0	1	3
11	64.5	0	3	4	30	84.5	1	2	6
12	64.7	0	1	7	31	83	1	2	3
13	36	0	3	0	32	60	0	0	1
14	54	0	0	5	33	72.5	0	2	1
15	39.5	0	2	3	34	90	0	2	4
16	68	0	5	3	35	58	0	3	3
17	61	0	5	3	36	60	0	1	2
18	62.6	0	2	4	37	168	1	2	11
19	98.7	0	2	10	38	111.5	0	1	9

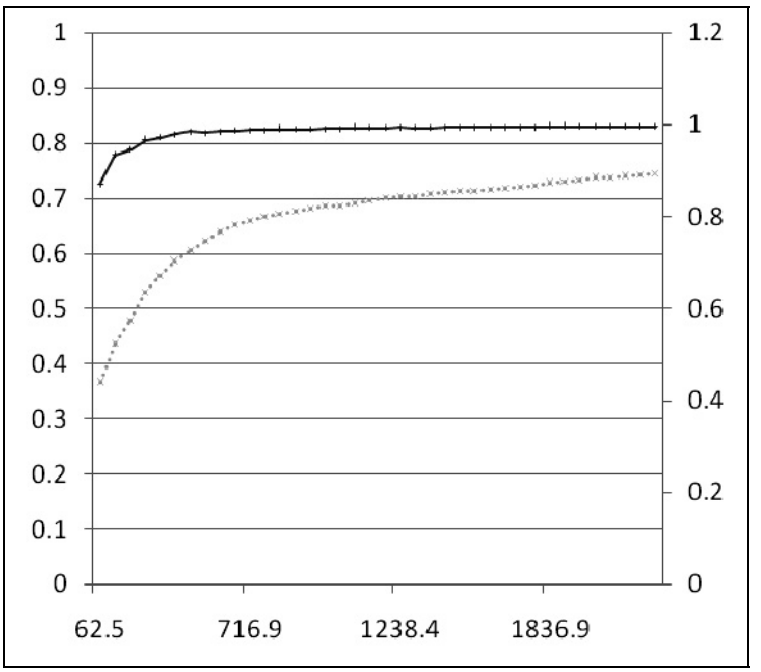
To estimate the error a, mle is used. Solving the mle equation, the estimated value of errors obtained is, $\hat{a} = 42$. While Pham [19] estimated the total number of errors using his model as $\hat{a} = 0.428$. Though the estimated total number of

errors $\hat{a} = 42$ using the proposed model is less than the actual errors present in the software, still it is better than the result obtained in [19]. Better results can be obtained using appropriate function for b and β. Conditional reliability $R(x/T)$ has been computed using the equation $R(x/t) = e^{-[m(t+x)-m(t)]}$ considering x=0.2. $R(x/T)$ and FDR b at each cumulative time are given in Table 2. The corresponding graph is given in Fig. 1.

Table 2. FDR & Conditional reliability corresponds to cumulative time

Failure Time	FDR (b)	R(x/T)	Failure Time	FDR (b)	R(x/T)
62.5	0.438	0.871	1226.4	0.842	0.9931
106.5	0.524	0.9345	1238.4	0.844	0.9938
146.4	0.575	0.9462	1293.4	0.845	0.9934
214.5	0.635	0.966	1342.4	0.849	0.9936
276.4	0.672	0.9727	1406.4	0.852	0.9939
342.4	0.702	0.9795	1432.4	0.855	0.994
415.5	0.727	0.9862	1496.4	0.857	0.9943
489	0.748	0.9828	1547.4	0.860	0.9945
581	0.769	0.9856	1599.4	0.863	0.9946
652.4	0.782	0.9876	1669.4	0.865	0.9949
716.9	0.792	0.9883	1753.9	0.869	0.9951
781.6	0.801	0.9892	1836.9	0.872	0.9953
817.6	0.806	0.9897	1896.9	0.876	0.9955
871.6	0.812	0.9903	1969.4	0.880	0.9956
911.1	0.817	0.9907	2059.4	0.884	0.9958
979.1	0.823	0.9914	2117.4	0.885	0.9959
1040.1	0.824	0.9918	2177.4	0.887	0.9961
1102.7	0.829	0.9923	2345.4	0.892	0.9963
1201.4	0.835	0.9929	2456.9	0.895	0.9965

Figure 1. Fault Detection Rate and Conditional Reliability w.r.t. Cumulative time



Conclusion

In this paper an attempt has been made to study the combined effect of increasing FDR and decreasing FIR on error estimation as well as reliability growth of software when the debugging process is imperfect. It has been observed that the assumptions made about the FDR and FIR are logical. The model can be used in other software failure data by assuming proper FDR & FIR. The proposed model can be used for estimating release time and cost of software.

Acknowledgement

Authors acknowledges University Grants Communication (UGC), New Delhi, India, for financial help in the project number F.No.33-115/2007(SR). Also, authors acknowledge ISM, Dhanbad, for providing facilities to carry out the work.

REFERENCES

[1]. MUSA, J.D.; IANNINO, A. and OKUMOTO, K.; "Software Reliability Measurement, Prediction, Application", *McGraw-Hill Int. Ed*., 1987

[2]. XIE, M.; "Software Reliability Modelling", *World Scientific Press*, 1991.

[3]. LYU; M.R.; "Handbook of Software Reliability Engineering", McGraw-Hill: NY, 1996.

[4] JELINSKI, Z. and MORANDA, P.B.; "Software Reliability Research", Statistical Computer Performance Evaluation**"**, *W. Freiberger. Ed. Academic, N.Y*., 1972, p. 465–484.

[5] SHOOMAN, M.L.; "Probabilistic Models for Software Reliability Prediction", *Statistical Computer Performance Evaluation, W. Freiberger, Ed., Academic, N.Y*., 1972, p. 485–502.

[6] WAGNOR, W.L.; "The Final Report on Software Reliability Measurement Study", *Report TOR-0074 (4112)-1, The Aerospace Corporation, El Segundo, C.A*., 1973.

[7] SCHICK, G.J. and WOLVERTON, R.W.; "An Analysis of Competing Software Reliability Model*", IEEE Trans. On Software Eng.*, vol. SE-4, 1978, p. 104–120.

[8] MUSA, J.D.; "A Theory of Software Reliability and Its Application", *IEEE Trans. On Software Eng*., vol. SE-1, 1975, p. 312–327.

[9] LITTLEWOOD, B. and VERRALL, J.L.; "A Bayesian Reliability Growth Model for Computer Software", Appl. *Statist*., vol. 22, 1973, p. 332–346.

[10] SINGPURWALLA, N.D. and SOYER. R.; "Assessing (Software) Reliability Growth Using A Random Co-efficient Autoregressive Process and Its Ramifications", *IEEE Trans. On Software Eng*., vol. SE-11, 1985, p. 1456–1464.

[11] XIE, M.; "A Shock Model for Software Reliability", Microelectron. *Rel*., vol. 27, 1987, p. 717–724.

[12] GOEL, A.L. and OKUMOTO, K.; "A Time-Dependent Error Detection Rate Model for Software Reliability and Other Performance Measure", IEEE *Trans. On Re*l., vol. R-28, 1979, p.206–211.

[13] CHATTERJEE, S; MISRA, R.B. and ALAM, S.,S.; "Joint Effect of Test Effort and Learning Factor on Software Reliability and Optimal Release Policy"; *International Journal of System Science*; Vol. 28; No. 4; 1997; p. 391–396.

[14] SUMITA, U. and SHANTIKUMAR J.G.; "A Software Reliability Model With Multiple-Error Introduction & Removal", *IEEE Trans. On Rel*., vol. R-35, 1986, p. 459–462.

[15] FAKHRE - ZAKERI, I. and SLUD, E.; "Mixture Models for Reliability of Software With Imperfect Debugging: Identifiably of Parameters" *IEEE Trans. On Rel*., vol. 44, 1995, p. 104–113.

[16] ZEEPHONGSEKUL, P.; XIA, G. and KUMAR, S.; "Software-Reliability Growth Model: Primary Failures Generate Secondary-Faults Under Imperfect Debugging", *IEEE Trans. On Rel*., vol. 43}, 1994, p. 408–413.

[17] KAREER, N.; KAPUR, P.K. and GROVER, P.S.; "A S-Shaped Software Reliability Growth Model With Two Types of Error", *Microelectron. Re*l., vol. 3, 1985, p.1085–1090.

[18] XIA, G.; ZEEPHONGSEKUL. P. and KUMAR, S.; "Optimal Software Release Policy With a Learning Factor for Imperfect Debugging", *Microelectron. Rel*., vol. 33, 1993, p. 81–86.

[19] PHAM, H.; "A Software Cost Model With Imperfect Debugging Random Life Cycle and Penalty Cost", Int*. J. of Sys. Sci*., vol. 27, 1996, p. 455–463.

[20] KAPUR, P.K.; SHARMA, K.D. and GARG. R.B.; "Transient Solutions of a Software Reliability Model With Imperfect Debugging and Error Generation", *Microelectron. Rel*., vol. 32, 1992, pp. 475–478.

[21] CHATTERJEE, S; MISRA, R.B. and ALAM, S.,S.; "A Generalized Shock Model for Software Reliability"; C*omputer and Electrical Engineering-An International Journal,* Vol. 24; 1998, p.no.: 363–368.

[22] ZHANG, X; TENG, X; and PHAM, H.; "Considering Fault Removal Efficiency in Software Reliability Assessment", *IEEE Trans. On Systems Man and Cybernetics-Part A: Systems and Humans*, Vol. 33, No. 1, Jan., 2003, p. 114–120.

[23] MISRA, P.N.; "Software Reliability Analysis", *IBM Syst. J*., Vol. 22, 1983, p. 262–272.

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 465—478

MONITORING WORKERS' REMITTANCES AND BENEFITS IN UGANDA: THE STATISTICAL ISSUES

E.S.K. Muwanga-Zake (PhD)[1]

ABSTRACT

Due to the increasing importance of remittances to Uganda, efforts are underway by the Central Bank and Central Statistics Office in the country to improve the regulatory and monitoring of environment. A multifaceted approach is used. This includes the enactment of a new law and regulations, improving administrative reporting and carrying out surveys in the major remitting countries and in Uganda. However, these have issues of collecting complete, accurate and timely data.

1. Background

Workers' remittances are current transfers by migrants employed in another country and have lived in those countries for at least one year. The players involved are mainly related persons (IMF, 1993). In many of the least developing countries (LDCs) international remittances now constitute the second largest capital flow after Foreign Direct Investment (FDI). Remittances constitute the fastest growing and most stable capital flow to developing countries. The exact amounts of these flows are, however, uncertain and the statistical compilation of remittances needs improvement especially in Sub-Saharan Africa. Official remittance statistics reported in the balance of payments (BOP) typically underestimate actual levels, and in several countries unrecorded remittances are significant. (Terry, F.D. et. al. 2005).

[1] Institute of Statistics and Applied Economics (ISAE), Makerere University, P.O. Box 7026 KAMPALA, Uganda: (Formerly, Director, Trade and External Debt Department (TEDD), Bank of Uganda, P.O. Box 7120 KAMPALA) E-Mail: emuwangazake@isae.mak.ac.ug; muwangazake@hotmail.com.
The views expressed are those of the author and none of the two institutions.

2. Why is the Bank of Uganda interested in workers' remittances?

In Uganda, Balance of Payments (BOP) estimates of remittances for the period 1996 to 2000, averaged 32% of exports. The figure increased to 65% of exports during the period 2001 to 2005 while for 2006 the figure is estimated at US$ 652million equivalent to about 6.5 percent of GDP (US$ 10 billion). Workers' remittances are, therefore, the second largest contributor to foreign exchange inflows after exports of goods and services, thereby contributing significantly to the BOP, exchange rate stability and also affecting monetary policies (Tumusiime Mutebile 2006).

More specifically, workers' remittances are important because:

a) They are a regular source of income of recipient households and spending power for their families.
b) They impact on poverty reduction and welfare improvement through their macro-economic effects.
c) They increase investments of households on education, capital for business ventures and health.
d) They are less volatile sources of foreign exchange as they tend to be counter cyclical increases in times of economic depression, political turmoil and natural disasters unlike other forms of inflows like Foreign Direct investment (FDI) and External Debt.
e) They are a steady stream of Foreign exchange earnings that can improve the countries' creditworthiness.

The Uganda Government, through the Bank of Uganda (the Central Bank) in conjunction with the Uganda Bureau of Statistics (the Central Statistics Office), has embarked on a programme to monitor workers' remittances to and from the country.

The major characteristics of remittances need to be clearly established in order to improve the understanding of remittance impacts. Studies and surveys are required in both the sending and receiving countries.

3. Conceptual and Measurement Issues of Remittances

The economic significance of remittances is not fully captured in the official BOP statistics in either sending or receiving countries. Monetary transfers increase the supply of foreign exchange in the country of origin of the migrant, while "in-kind" remittances in the form of goods and services save scarce foreign exchange in the recipient country. The measurement of both cash and in-kind remittances has proven to be very difficult, imprecise, and incomplete. Only some of these transactions are recorded.

Where remittances are sent through formal channels, they are recorded by the receiving country's official statistics as an inflow in the current account of the BOP. Conversely, cash remittances sent informally through couriers are usually

unrecorded in official statistics. In-kind remittances, or goods and services sent to households in the home country or brought on their return home, may be only partially captured as imports in official data. Very little data exists on the size of remittances in kind. Other transfers in the form of charitable donations or payments and deposits for relatives and friends (such as insurance premiums, tuition, and travel costs) function as an economic form of remittances, but are rarely recorded as such.

Choosing between official or unofficial transfer channels is an important decision when sending remittances and depends on several factors. These include the availability and adequacy of banking services, transaction costs, "know-your-customer" requirements, and the potential for additional earnings through unofficial channels. When official means of remittance transfer are unattractive, the private sector or groups of emigrants themselves often set up parallel systems. Dealers in informal transfer systems sometimes provide competitive exchange rates, as well as other assistance and services of a personal nature.

All these mean that the data in Uganda is incomplete and inaccurate. It is not possible to accurately disaggregate official estimates of remittances by source due to the residual method of estimation, which is used in making the estimates.

4. What is Being Done in Uganda?

The Bank of Uganda (BOU) is using a multifaceted approach of measures, outlined below, aimed at improving the recording, management and integrity of workers' remittances, while at the same time trying to maximize the benefits in terms contributing to economic development.

4.1. Publicity and awareness campaigns

A sensitization campaign amongst Ugandans in the Diaspora, and the recipients and service providers in the private sector is ongoing. The campaign emphasizes transparency, competition, and customer information, and is designed to highlight policy and operational issues. The campaign also works through the Uganda embassies and ethnic group based associations. The Bank is also considering creating links with web sites of the two Uganda newspapers, which are widely read by the Ugandans in the Diaspora.

4.2. Legal and Institutional Framework for the Money Transfer Business

The Ugandan Government has revised the law on foreign exchange transactions to provide for more money remittance services, thus making it easier for Ugandans abroad to send money home (Uganda Government, 2005). Apart from commercial banks and other financial institutions, Remittance Licenses may

now, under certain conditions, be obtained by International Money Transfer Agencies and Foreign Exchange Bureau. BOU has also facilitated notable improvements in the payment system to make it more efficient and wide-spread.

These efforts include: installation of electronic banking with the national electronic switch; 230 Automatic Teller Machines (ATMs) spread countrywide; and the introduction of credit cards and debit cards. Some financial institutions are already considering introducing Internet banking. Mobile Banking Services are expected to reach an estimated 20,000 clients over a period of five years.

All these developments should facilitate a smoother flow of remittances and attract remittances into the formal system, and thereby move some of the hitherto informal remittances into the formal (official) channels and hence make them easier to record. This will improve data collection.

To cap all these, BOU created a Trade Statistics and Remittances Division in the then Trade and External Debt Department (TEDD) to concentrate all efforts in the estimation of remittances.

4.3. New Methods of Reporting/ Reporting Requirements for Money Remittance Business

The new law strengthens the capacity of BOU to monitor and regulate the transactions. Under this law, all licensed remitters are expected to make weekly and monthly returns to BOU and therefore provide information on remittances on a continuous basis. Unfortunately, details for remittances through commercial banks and other financial institutions are difficult to identify. Similarly, returns from the newly-licensed remitters are still faulty and incomplete.

As at 30th June, 2007, there were twenty eight (28) Non Banking Financial Institutions (NBFIs) licensed to engage in money remittance business comprising: 23 forex bureaus, four (4) Micro Finance Deposit-Taking Institutions and two (2) Credit Institutions. Financial Institutions (Banks) licensed under the Financial Institutions Act 2004 are exempted from complying with the provisions of the Foreign Exchange Act 2004.

All Money Remitters (MR) licensed under Section 5 of the Foreign Exchange Act 2004 are required to submit to Bank of Uganda the following returns: Weekly (Send and Receive), Monthly (Send and Receive) and a summary Monthly transactions as specified in Section 27 (4) and Schedule 8 of the Foreign exchange (Forex Bureau and Money Remittance) Regulations 2006.

All the 28 NBFIs are submitting these returns as specified by the regulations. Compilation of the returns into meaningful reports commenced in the second quarter of 2007 and as at 30th June, 2007, total inflows amounted to US$17.301Million while outflows captured were US$0.585Million. This is a clear indication that either the formal means of remitting money is not in use as expected or the mode of capturing the data is not ideal/ accurate. It may therefore be preferable to capture such information from the source countries and through

commercial banks since the eventual settlement of such cross-border transactions is through a commercial bank.

4.4.Surveys on Remittances

Although balance of payment data on remittances are commonly used for estimating the volume of remittances, they could be misleading since informal remittance flows are not accounted for in these calculations. The Bank of Uganda, in collaboration with the Uganda Bureau of Statistics (UBOS), have institutionalized surveys on remittances to Uganda. The objective is to estimate the specifics of the remittances in terms of source and size of remittances, frequency and channels of remitting, use of remittances, seasonal pattern, the demographic characteristics of the recipients and the location of the remitters, the socio-economic conditions and intentions of the senders, etc.

Surveys have been carried out in the major sending countries, beginning with the UK, the USA and South Africa. The remittances to Uganda are mainly from the UK, the USA, Japan, South Africa and Sweden. Informal service providers reveal that the UK, the USA, and the Scandinavian countries account for about 60%, 20% and 10% of transfers, respectively.

In order to enhance the existing knowledge on remittances to developing countries, the Financial Market Integrity Unit (FPDFI) has launched a series of studies called Bilateral Remittance Corridor Analysis (BRCA). The intention is to use the knowledge gained through these studies to develop best practices to protect the integrity of remittance markets and to improve efficiency and transparency of transfer channels for remittance flows.

This study was initiated at the request of the Bank of Uganda (BoU) as a World Bank-BoU joint study. The purpose of the joint study is to share knowledge and expertise of World Bank's Bilateral Remittance Corridor Analysis (BRCA) with the Bank of Uganda. For this purpose, a mission was jointly conducted in the United Kingdom and the United States. For these missions, the German Gesellschaft für Technische Zusammenarbeit (GTZ) provided financial support. Being the first BRCA study to be conducted with the partnership of a local authority adds to the significance of this study.

The study evolved from an original United Kingdom–Uganda remittance corridor to three corridors including the US-Uganda and South Africa-Uganda. These two corridors were included in consequence of a supplementary request made by the Bank of Uganda. Initial remittance data from a pilot survey indicated that a large volume of remittances seems to originate in the United States and anecdotal evidence of remittance flows from South Africa. The South Africa-Uganda corridor added value to the study since it allows the study to compare North-South corridor (UK-Uganda and US-Uganda) and South-South corridor.

The main objective was to describe how characteristics of the three remittance corridors affect remittance flows to Uganda and implications of such characteristics

on the desired shift from informal to formal remittance flows. In order to achieve this objective, analysis was made for the remittance senders, remittance flows and market players, regulations, risks of money laundering and the existing Anti-Money Laundering and Combating the Financing of Terrorism (AML/CFT) framework, and challenges to expanding access to financial services. Based on the analysis, the report provides a set of focused policy recommendations to regulators and market players in order to foster a competitive and transparent formal market. This report is the culmination of the efforts and contributions of FPDFI, Central Bank of Uganda and The German Gesellschaft für Technische Zusammenarbeit, and many others.

The fieldwork was conducted in Uganda, the United Kingdom, the United States, and South Africa. They included interviews with the authorities, international organizations, private sector entities (banks, money transfer operators, and research organizations), Non-governmental organizations, and Ugandan migrants.

4.5. The Uganda National Household Survey (UNHS) 2005/06

During the UNHS 2005/06 carried out by UBOS, information was collected on receipt and use of both domestic and international remittances at household level. As shown in Table 1, overall proportion of recipients of remittances from local sources (41%) was much higher than that for remittances from abroad (2%). The capital City, Kampala, had the highest proportion of households receiving remittances from abroad (7%). The national mean monthly value of remittances from abroad was about US$40.

Table 1. Households that received a Remittance during the last 12 months by Residence (%)

Residence	Percentage of households		Mean monthly Value* of amount received (USh)	
	from domestic sources	from abroad	from domestic sources	from abroad
Rural/Urban				
Urban	39.6	5.0	38,000	119,300
Rural	41.8	1.7	14,500	39,700
Region				
Kampala	35.2	7.3	50,700	130,500
Central	47.0	2.4	20,500	98,500
Eastern	41.7	3.3	15,600	29,500
Northern	56.4	0.9	13,700	27,700
Western	26.9	0.7	14,700	28,300
Uganda	**41.4**	**2.3**	**18,500**	**70,500**

**Note: The Value of remittances includes both cash and in- kind.*
Source: UNHS 2005/06

As Table 2 shows, most recipients used the remittances to purchase consumption goods and services irrespective of their source. This was followed by payment for education expenses. It reinforces the findings by Bank of Uganda that the shilling tends to appreciate during the time of paying school fees and Christmas holidays when the immigrant workers send funds to support their families.

Table 2. Recipients by Purpose and Source of Remittances (%)

Main Purpose of Remittances	Source of Remittances	
	domestic	abroad
Purchase consumption goods and services	63.4	51.7
Pay for education expenses	13.6	26.1
Pay for health expenses	6.6	2.9
Working capital for non-farm enterprises	0.9	5.7
Purchase building materials	0.5	3.9
Buy land	0.1	1.5
Buy farm inputs, tools and implements	0.6	0.3
Pay for ceremonial expenses	1.1	2.0
Other	13.2	5.9
Total	**100.0**	**100.0**

Source: UNHS 2005/06

4.6. Pilot Survey 2006

A pilot survey was carried out in January 2007 to test the questionnaire for the subsequent countrywide survey on remittances. The preliminary findings revealed that almost 50% of the beneficiaries received under US$100 per annum. The same percentage of the respondents indicated preference to use informal channels for sending money to Uganda. The origins of remittances are diverse but the UK, and the USA dominate.

The results also showed that remittances are used mainly for consumption, health care, childcare and education. There is also a noticeable use of remittance funds for business support and real estate. For example, 16% of the recipients indicated that remittances are spent on building works and land purchase.

4.7. Survey on Remittances to Uganda 2007

The main objective of the nationwide survey, jointly conducted by BOU and UBOS, was to collect high quality data on remittances received in 2006 at household level so as to improve the accuracy of the BOP and therefore strengthen the formulation of monetary and exchange rate policies. There was

also a community questionnaire to determine the contribution of remittances at community level.

The survey covered over 4,000 randomly selected households. The survey sought to establish an accurate estimate of the value and volume of remittances received during 2006, the origin of such remittances and the characteristics, e.g. amounts, frequency, use of received remittances received. The results are outlined below.

The up-rated estimated remittance value from the survey of US$406.5 million was below the official estimate (US$665m) for the year 2006. However, the official estimate was derived as a residual from BOP computations. This was the country's first attempt in empirical assessment of the size of inward remittances, hence this estimate was thought to be encouraging and considered to be indicative of the magnitude of such flows. There were also identifiable methodological gaps which could be addressed through refinement of sampling procedure, limiting the geographical scope of the survey, revision of the survey instruments, enhanced training of field staff, a strategic sensitization programme to create awareness and build confidence in the respondents, among others.

More than half of the recipients received remittances once or twice a year. This finding is supported by the fact that December and January returned the highest proportion of remittances during 2006 confirming the practice of migrants sending money home during the festive season and school related expenditure. However, with one round of fieldwork it may not be possible to conclude on the regularity of inwards remittances.

In addition to monetary remittances, Ugandans also received remittances in kind mostly once a year and these were mainly clothes and domestic appliances. Such goods supplement household incomes. However, estimation of values for in-kind remittances was rather complicated due to omission of values at the onset. Future surveys should provide for this kind of estimation.

The sources of remittances to Uganda were diverse, with Europe and Africa being distinct in terms of share of total volume. This finding may be explained by the existence of historical ties in addition to proximity and ease in communication between Uganda and these regions. The results also revealed presence of south-to-south remittances to Uganda with 37 percent of the respondents indicating that they received remittances from Africa.

The findings on education and age revealed that the majority of remitters are young educated Ugandans. At the same time, more than half of the remitters had lived abroad for between 5-10 years. These findings are an indicator of the continued movement of Ugandans in search of employment abroad.

Over 65.5 percent of the respondents indicated that they would rather use formal channels for receiving remittances, a good indicator of how the population perceives and appreciates the formal financial services. The main reason for this preference is safety of the remittances. This inclination towards formal financial services should be leveraged in the financial deepening strategy and furtherance of the payments systems in Uganda. However, many Ugandans still use informal

methods, that is, friends and acquaintances for remittance transmission, which may be attributed to avoidance of transaction costs associated with formal channels. There is therefore need to sensitise users on the various service providers available, the type of services that they are authorized to engage in and the risks associated with the use of informal methods of remitting.

Remittances benefit the remitters and recipients. However, there is a distinct difference in the respective use, with consumption for recipient and investment for remitter. While Remitters remit to maintain the immediate and extended families, they personally seem to have an interest in long-term development projects.

Though the survey did not directly ask a question on beneficiary sectors, from the expenditure pattern one can easily link expenditure with three major sectors, namely wholesale and retail trade-for consumption goods, education and health, for the benefit of both recipients and remitters. In addition to these three, investment stands out in cases where remittances benefit the remitter.

The report identified the under-listed lessons learnt and areas of focus for future surveys;

- The need to revisit the methodology to facilitate better estimation of the magnitude of remittances:
  - The sample frame: Remittances seem to originate in urban areas and distribution takes place thereafter. It would therefore be of benefit to conduct an urban survey on remittances;
  - Assessment of regional distribution and "pockets" of remittances if any; and
  - Assessment of the three-tier distribution system in terms of role of service providers along the chain and market share of each.
- In order to build on the survey results, there is need to conduct similar studies on annual basis. This may be followed by censuses after every 5 years.
- Subsequent surveys are also necessary to confirm the revealed seasonal pattern of remittance inflows.
- Assessment of the impact of remittances calls for drawing comparisons between households that receive and those that do not receive remittances. This may be achieved through the census.
- An independent survey on outward remittances with appropriate methodology is necessary to complete the country's position on workers' remittances for Balance of Payments.
- Incorporate continuous sensitization in the remittance monitoring strategy, not only for the survey activities.
- Remittances in kind are a reasonable proportion of total inflows. It is important to address valuation issues for better estimation of the value of remittances in kind.

4.8. Findings in the United States of America (USA)

There is an annual meeting of the Uganda Northern America Association (UNAA) during the Labour Day Week-end. These meetings are attended by over 2000 Ugandans leaving in the USA, Canada, the Caribbean, the UK and, sometimes South America to share ideas, challenges, and also fellowship. BOU has participated in these meetings since 2004 to discuss remittances. A survey was carried out during the convention in New York in 2006 as part of the ongoing joint World Bank/Bank of Uganda study on the USA-Uganda remittance corridor.

A total of 124 questionnaires were returned. More than half of the remitters were graduates with a first degree, the majority having lived in the USA for periods ranging from 10 to 20 years. Western Union was the most preferred remittance channel with 39%, followed by MoneyGram (28%) and the banks (27%). Major reasons or considerations for choice of medium (in order of importance) included cost/rates, convenience, speed and safety/reliability of the given medium. Only 6% use informal channels. Costs are within 1-12% of the total remitted; remittance through banks was noted to be more costly as it involves payment of fees at both the sending and receiving points. Remitters prefer to use the money transfer companies for lower and banks for the bigger amounts. Family support was the major use (intended) of remittance as indicated by 59% of remitters while almost 37% were equally divided between education and investment/business projects. The mean remittance value was about US$700.

The 19th UNAA convention was held from 30th August to 3rd September 2007 in San Francisco, California. This and future UNAA conventions should provide a platform for feedback on the Bank's findings and more elaborate surveys on remittances.

4.9. Findings in the United Kingdom

Remittances service providers in the UK include banks, money transfer companies, and informal providers. The Financial Services Authority regulates the banks while the money transfer companies are registered, regulated and supervised by Her Majesty's Revenue and Customs (HMRC). There are 21 ethnic group based money transfer service providers to Uganda registered with HMRC under the Money Laundering Regulations (2003). These remitters have formed an association which could be used to collect data on the amounts of remittances. However, a few unregistered individuals are known to offer remittance services. The individuals operate outside the defined regulatory regime characterized by easy entry and operational requirements.

A preliminary study has been carried out on the registered remitters in the UK. The study gave some data on the volumes remitted and the costs of remitting. It confirmed that some of the remittances were very low. Remittance values of between 10 and 16 pounds (US$19-30) are not uncommon for remittances

channelled through ethnic group based service providers. They are preferred because of the relatively low rates charged, additional personal service offered, less paper work and higher speed of delivering services.

Transaction charges range between 7 to 15 percent of the remittance value, through banks and international Money Transfer Organizations while ethnic group based ones charge 2–3 percent of the value.

4.10. Findings from South Africa

In South Africa, MoneyGram is the only international MTO providing remittance services by collaborating with banks. Quantifying the remittance flows is a challenging task, resulting from lack of proper methodologies, accuracy of recorded data, and extensive use of informal mechanisms. South Africa sends and receives cross-border remittances with multiple remittance corridors within and outside Africa. As a sending country, South Africa is the largest source of remittances in Sub-Saharan Africa. The volume of remittance flows from South Africa to Uganda is unknown. Unlike the information of Ugandans in the United States and the United Kingdom, there is limited data available on the Ugandan population in South Africa. The lack of information with regard to the size of per capita remittance, the size of population, and income level makes it difficult to develop an estimate of the remittance volumes.

4.11. Summary of Findings in the Three Bilateral Remittance Corridors

- **Uganda is not a major remittance destination in the three countries** although these remittances are important to the Ugandan economy. Remittance data are not available from South Africa, which is doubly important as a South-South corridor as well as a regional remittance corridor.
- **Different migration generations of both documented and undocumented Ugandans have entered the three corridors over the past 5 decades.** Identifying the size of the undocumented Ugandan migrant population is difficult, making estimating remittances flows complicated.
- **Immigration policies of the three countries have made it difficult for low-skilled migrants** to seek temporary employment in these countries; hence many low-skilled workers do not have proper immigration documentation, which leads to limited or no access to formal financial services.
- **Access to remittance services varies** although documented migrants have full access to formal financial services while undocumented workers in all three countries have limited or no access to formal financial services. The undocumented Ugandans in the United Kingdom have

access to formal remittance facilities if they have any government-issued ID, including Ugandan national identification.

- **Investing in businesses and preparing for retirement in Uganda, in addition to supporting their families, is high on the agenda** of the Ugandan Diasporas in all three countries. Hence, most Ugandans show preferences to keep closer ties with their home country.
- **Registered money service businesses are allowed to provide remittance services** in the United Kingdom and the United States; on the other hand, in South Africa only authorized dealers, namely banks and foreign exchange bureaus (most of these entities are part of banks) are allowed to conduct remittances. In South Africa, only MoneyGram, among international MTOs, provides remittance services by collaborating with banks.
- **Remittance costs in the United Kingdom and South Africa are substantially higher than the costs in the United States.** The formal remittance market in South Africa lacks competition with only one international MTO, commercial banks and their subsidiary foreign exchange bureaus. In the UK–Uganda corridor, international MTOs have an advantage in overall pricing while ethnic MTOs are more competitive. In the US–Uganda market, international MTOs provide competitive services since they consider Africa as a single market and are competitive in all corridors to Africa.
- **AML/CFT laws and regulations exist in all three countries**, however, they are not similar. In the area of remittances, the FATF requires countries to ensure that money transfer businesses are registered or licensed. The United Kingdom and the United States require money service businesses to be registered; South Africa requires them to be licensed. The United Kingdom and South Africa have a centralized regulatory framework for the remittance market, while US regulations are fragmented among federal and state levels.

4.12. The African Development Bank – Workshop in May 2009

It was noted that data on remittances was initially collected as a residual of capital inflows after subtracting explainable inflows such as exports, and current transfers. More direct methods of estimating remittances have since been adopted, particularly the use of surveys in institutions through which remittances flow, i.e. banking institutions and companies registered to do the business of money transfer. It is believed that the new approaches produce more accurate data on remittances. The challenge Uganda is facing is that such surveys have been piecemeal. The key recommendation was to improve data collection on migration and remittances through a comprehensive survey to get a correct picture of remittances in Uganda.

5. Conclusion: Challenges and the Way Forward

From the above findings on remittances to Uganda, the following issues on the estimation of the remittances to Uganda can be identified:

(i) The preference for use of informal channels, for various reasons, and the attendant difficulties in obtaining data strongly suggest that available data on remittances could be understated which calls for improvement in methodology estimating the volume of remittances. Informality is explained by access issues – shallow banking system; poor technology; and low labour mobility explain the persistence of the informal system in remittances.

(ii) Issues of cost, less cumbersome paper work and personalized service seem to attract users to the informal system. The formal channels need to focus on the improving the delivery of these preferred characteristics of migrant remittances. While the "Know-Your-Customer" requirements are necessary for protecting the integrity of the financial sector, these clearly come out as constraints in the access to and use of formal channels. Harmonization of the more restrictive supervision and regulation of the financial sector at the receiving end with the more accommodative regime at the sending points is critical for increasing the use of formal channels of money transfer.

(iii) The use of cash as the major delivery instrument highlights the inefficiencies in, and the need for further development of the payments system.

(iv) The annual UNAA Meetings in USA and the Remitters' Association in the UK present unique opportunities to collect data on remittances to Uganda from the two major source countries. Obviously, it will be necessary to cover remitters who do not attend the UNAA convention or who do not remit through the Remitters' Association. Ugandans in Japan have also formed an Association and are planning a UNAA-like conventions.

(v) More work needs to be done to get complete and timely data from the formal remitters, including commercial banks and other financial institutions. It is also necessary to link remittances to poverty reduction and establish the seasonal pattern for purposes of exchange rate management, and conduct of monetary policy.

(vi) The diverse origin, small values and use to which remittances are channelled presents the challenge of identifying the respondents in any survey. However, the concentration of receiving points in the capital city, Kampala may mean that the surveys could be concentrated there to begin with.

(vii) The collection of data on remittances is a further illustration of the collaboration efforts in the Ugandan statistical system. Other examples include the collection of data on Private Capital Flows (PCF), Informal Cross Border Trade (ICBT), etc. Uganda has drawn up a Plan for National Statistical Development (PNSD).

REFERENCES

ABUKA, CHARLES (2008): The Government and Remittances: Uganda's Policy. Presentation at the Workshop on Remittances in Uganda: African Development Bank Group; 28 May, 2008.

African Development Bank (ADB) – Remittances in Uganda: The Way Forward, Serena Hotel, Kampala, Uganda 28 May 2009; WORKSHOP REPORT.

Bank of Uganda, Uganda Bureau of Statistics & GTZ; (2007); Inward Remittances 2006; Survey on Remittances to Uganda, 2007 Workers' Remittances Report.

International Monetary Fund (1993); Balance of Payments Manual; Washington, D.C.

MUWANGA-ZAKE, E.S.K (2004): Transferring Funds to Uganda: The Current Legal Ways and Issues, Paper Presented at the UNAA Convention, Minnesota September 2004.

OROZCO, MANUEL (2008): Remittance Transfers, its marketplace and Financial Intermediation in Uganda: Preliminary Findings, Lessons and Recommendations. Report Commissioned by the Inter-American Development Bank in co-operation with the African Development Bank.

TERRY F. DONALD; WILSON R. STEVEN (Editors) (2005); Beyond Small Change: Making Migrant Remittances Count; Inter-American Development Bank, Washington, D.C.

TUMUSIIME MUTEBILE, EMMANUEL (2006), Governor Bank of Uganda; Speech at the Second International Conference on Migrant Remittances: Remittance and Access to Finance; County Hall; London, United Kingdom; November 13–14, 2006.

Uganda Government, (2005); The Foreign Exchange Act 2004, and the Foreign Exchange (Forex Bureaux and Money Remittance) Regulations, 2006.

Uganda Bureau of Statistics (2006); Uganda National Household Survey, 2005/06.

World Bank, (2008) Bilateral Remittance Corridor Analysis (BRCA) Remittance Corridors to Uganda United Kingdom, United States, and South Africa: Challenges to linking remittances and use of formal financial services. Isaku Endo and Jane Namaaji; Editors.

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 479—503

ANALYSIS OF TEMPORARY ASPECTS OF POVERTY IN POLAND BETWEEN 1997–2000 BY HAZARD MODELS

Natalia Nehrebecka[1]

ABSTRACT

This paper examines the extent and characteristics of poverty in Poland analyzed on the basis of panel data from CHER (Consortium of Household Panels for European Socio-Economic Research) for the years 1997–2000. The analysis presented shows a low households' dynamic of income in this period. The total number of years spent in poverty as well as different sequences of entry to and exit from poverty suggest the tendency to a persistent form of this phenomenon in the population. During the period studied, the basis for the calculation of the number of years spent in poverty was the rate of exit from and entry to poverty. The calculations have been made according to the method of analyzing poverty dynamics by hazard models, considering observed and unobserved heterogeneity of individuals to "explain" a chance of exit and return to the sphere of poverty.

Key words: Permanent and transient poverty, Hazard models, Multi-spell poverty episodes, Unobserved heterogeneity

Introduction

Poverty has affected society since the beginning of mankind. It still is experienced not only by citizens of undeveloped countries but also by those living in highly developed economies. The existence of poverty on a large scale leads to a wide interest in this topic, although research on poverty is primarily conducted by countries in which the scope of this phenomenon is marginal. A different situation persisted in centrally-planned economies, where poverty was not officially recognized as the aim of socialist regimes was to eliminate poverty.

In recent years, income dynamics and duration of poverty have been more often discussed during public and academic discussions as socially important

[1] Department of Statistics National Bank of Poland, Faculty of Economic Sciences University of Warsaw, e-mail: Natalia.Nehrebecka@nbp.pl, nnehrebecka@wne.uw.edu.pl.
I would like to say special thank to prof. B. Górecki, prof. M. Wiśniewski for excellent suggestions, from which I benefited a lot.

factors completing information on income distribution. It has been agreed that the best way to measure and understand the problem is to supplement traditional studies based on cross-series data with long-term analysis based on panel data.

When a household's gross income declines below a minimum level and there are not enough financial resources to satisfy basic human needs, its members are considered to be poor. When this situation persists for an extended period of their lives then an entire society is affected by effects of poverty. A long-term poverty of the same group of people is difficult to accept not only because of social fairness but also due to high external costs of permanent marginalization and social exclusion. Political stability is also an important reason why the topic of poverty arouses so much interest both in highly developed and developing countries.

The aim of this paper is to analyze duration of poverty in Poland over the period of 1997–2000, using panel data from the CHER[1] database. An exact specification of the population permanently living in poverty and its differentiation from the category of people who experience this state temporarily may aid in preparing programs of social assistance targeted at these groups that are not able to exit poverty by themselves. Studies of this type are necessary due to a large scale of this phenomenon as well as limited amount of resources which, for the sake of a permanent budget deficit problem in Poland, may be of use to overcome the problem. If I succeed in defining the category of those chronically living in poverty, this analysis may prove helpful in targeting the financial assistance for the poor more effectively while cutting the budgetary resources earmarked for this purpose at the same time.

The main questions put forth in this paper are as follows:

- When does a household find itself in poverty and how long does the poverty last?
- Does "recurrence of poverty" exist i.e. whether repeated spells of poverty could be considered an adequate indicator of permanent poverty?
- What socio-economic characteristics of households allows for the best identification of poverty in Poland between 1997–2000?

This article uses econometric methods of poverty analysis by model hazard taking into account observable and unobservable heterogeneity of individuals. It carries out estimation of two complimentary log-log type models in discrete time. They differ in method by which they describe duration of so-called base hazard function. In the first model, it has a non-parameter shape while in the second it is a polynomial time function.

The article consists of five parts. The first part includes the methodology used in constructing the econometric model. The second part describes the data used in the analysis, definition of poverty as well as independent variables of the model. The third part provides a short description of poverty dynamics and its duration in Poland. To introduce a distribution of time spent in poverty an exit rate from and

[1] Consortium of Household Panels for European Socio-Economic Research.

a return rate to poverty have been used. The parametric method of analyzing poverty is presented in the fourth part. The last part is a summary of the results and conclusions.

1. **Duration theory**

To study the exit from and return to poverty rates of households, this article applies the Jenkins (1995, 2004) and Devicienti (2000, 2002) methodology. The hazard function, $h(t)$[1] is used for calculating exit and return probabilities.

Let $\langle \tau_1, \tau_n \rangle$ be an observation period of the sample, while $\langle t_1, t_k \rangle$ - period of poverty. In the study, one can have censored observations - that is, those that contain information how long the person "was affected" by poverty, but the exact moment of exit from and entry to poverty are unknown.

Let T_i be the length of the poverty episode of *i*-the individual. Let the length of this time be a realization of the continuous random variable T_i with the distribution function $F(\tau)$ and the density function $f(\tau)$, then the probability that a state of poverty will last shorter than τ years under a condition that it occurred is:

$$F(\tau) = \Pr(T_i < \tau) \tag{1}$$

However, the probability that occurrence of poverty will last a minimum of τ years under condition that it occurred is[2]:

$$S(\tau) = 1 - F(\tau) = \Pr(T_i \geq \tau) \tag{2}$$

where: $S(\tau)$ – survival function.

The continuous time hazard rate is:

$$\theta(\tau) = \lim_{\Delta\tau \to 0} \frac{\Pr(\tau \leq T_i < \tau + \Delta\tau \mid T_i \geq \tau)}{\Delta\tau} = \lim_{\Delta\tau \to 0} \frac{F(\tau + \Delta\tau) - F(\tau)}{\Delta S(\tau)} = \frac{f(\tau)}{1 - F(\tau)} = \frac{f(\tau)}{S(\tau)} \tag{3}$$

This rate can be interpreted as a probability of ending the episode of poverty in the range $[\tau, \tau + \Delta\tau]$ for small $\Delta\tau$ under condition of "survival" to time τ.

It is important to notice that hazard rate in continuous time does not satisfy all properties of probability and especially the hazard rate may be greater than 1.

The survival function as well as hazard rate are connected by simple direct relation, that is:

$$\theta(\tau) = -\frac{S'(\theta)}{S(\theta)} = -\frac{d}{d\tau} \ln S(\tau) \tag{4}$$

---

[1] h(t) – hazard rate in discrete time, θ(τ) – hazard rate in continuous time.

[2] The survival function provides a probability that the person will live longer than some given time τ or in other words – that they live until time τ (Stanisz, 2003).

and

$$S(\tau) = \exp\left[-\int_0^{\tau} \theta(z)dz \right] \tag{5}$$

This research had begun from non-parametric[1] estimation of survival and hazard functions utilizing the Kaplan-Meier estimator.

1.1. Kaplan-Meier Estimator

The Kaplan-Meier estimator is calculated in accordance with the following formula[2]:

$$\hat{S}(\tau_j) = \prod_{j|\tau_j<\tau_n} \frac{n_j - h_j}{n_j} = \prod_{j|\tau_j<\tau_n} \left(1 - \hat{\lambda}(\tau_j)\right) \tag{6}$$

where:

$\{\tau_j : j = 1,2,...,k\}$ - is a set of all moments of the events which occurred;

$\tau_1 < \tau_2 < ... < \tau_k < \infty$ - in the order of duration of the episodes;

n – a number of observations, $k < n$, because a part of the observations is censored;

h_j – a number of completed cases with a duration of τ_j;

n_j – a count of the risk set, that is a number of episodes exposed to completion in a time τ_j, $n_j = (m_j + h_j) + (m_{j+1} + h_{j+1}) + ... + (m_k + h_k)$; where: m_j - a number of censored observations with a censored duration in the interval $[\tau_j, \tau_{j+1})$;

$\lambda(\tau_j)$ – a probability of ending the occurrence in the near (right-sided) τ_j point, under a condition that the episode has lasted until point τ_j. $\lambda(\tau_j)$ is a

[1] A non-parametric analysis allows us to analyze data without making assumptions regarding distribution. This has certain benefits as well as pitfalls. On one hand, a possibility to analyze data without making assumptions about the real distribution of "life" lets one overcome potentially great errors. On the other hand, confidence levels related to non-parametric analysis are generally much wider than those based on parametric calculations; moreover in the latter case forecasts beyond the sample are possible, http://www.weibull.com/LifeDataWeb/nonparametric_analysis.htm.

[2] Kiefer (1988).

theoretical hazard function, and $\hat{\lambda}(\tau_j) = \frac{h_j}{n_j}$ is an empirical "estimator" of hazard function at point τ_j.

The Kaplan-Meier estimator can also be set using the life duration table methodology[1]. In this case, the estimator has a form of:

$$\hat{S}(j) = \prod_{k \le j} \left(1 - \frac{d_k}{r_k} \right) \tag{7}$$

where:

d_k – a number of occurrences which has finished in *k* range;

$r_k = r_{k-1} - \frac{1}{2} m_k$;

where:

m_k – a number of censored occurrences in a range (a_{k-1}, a_k);

r_{k-1} – a risk set - a set of units that lived to the uppermost part of the range (a_{k-2}, a_{k-1}) and so have a chance to enter (a_{k-1}, a_k) range.

Therefore, one can see that here a survival function is estimated and its only argument is time. Based on survival function, one can easily transform to hazard function of the following form:

$$\theta(\tau) = -\frac{d}{d\tau} \ln \hat{S}(\tau) \tag{8}$$

1.2. The parametric method: discrete time model, including observed and unobserved heterogeneity of individuals

When estimating the hazard rate with the use of the Kaplan-Meier method, one does not consider the heterogeneity of individuals, which depends on observed and unobserved variables. This problem is resolved by the parametric method.

Although exit from poverty can occur at any moment of time (stochastic process for continuous time), usually the duration of poverty episodes is observed in discrete, not continuous time[2]. The models with a discrete time, however, have some advantages. One of these comes from the fact that discrete time models

[1] This method is the oldest technique of estimating survival and hazard functions.

[2] Jenkins (2004).

combine variability in time with elastic specification of duration interdependencies[1].

For the needs of this analysis, discrete models presented by Jenkins (1995) have been used:

1) Prentice-Gloeckler model (1978);
2) an extended model by Meyer (1990), the Prentice-Gloeckler model (1978) which contains the gamma distribution that includes unobserved individual heterogeneity.

An exit hazard rate from a given state (poverty or wealth) in a discrete time for *i*-th individual in period *t* is specified by Prentice and Gloeckler (1978) as:

$$h_i(t) = 1 - \exp[-\exp(\theta_0(t) + \beta' X_i(t))] \tag{9}$$

where:

X_i(t) – independent variables (variable with time or fixed);

β – a vector of unknown parameters;

$\theta_0(t)$ – a base hazard rate, that is a hazard for a given individual when all independent variables' values are equal to zero.

The model also known as the *„complementary log-log"* can be interpreted as a model with discrete time which is directly related to hazard in continuous time[2]. The assumption of a base hazard form $\theta_0(t)$ may unnecessarily limit the scope of hazard and bring potential bias of the β estimator. That is why, it is especially important to include general nonparametric specifications.

2. Data and variable description and definition of poverty margin

The data used in this empirical research come from the CHER database. It is a harmonized and standardized microeconomic database created from already existing panels pertaining to living conditions of individuals and households in the European Union before its expansion in 2004, as well as for Poland and Hungary. The database contains detailed data on income and professional activities of individuals, their education, employment, employment history, and others. Variables describing social relations and the sentiments of the members of the households are also included there. In the CHER database two Polish panels are available: the first one referring to years 1994-1996, the second to 1997–2000. Okrasa (1999) analyzes household welfare trajectories during the period 1993-1996 to identify long-term poverty and determine the relevance of household asset endowments as determinants of household poverty and vulnerability over time. Overall income mobility in Poland, at that period, was high. In other words,

[1] Duration data analyses benefit from the use of discrete-time models. However, available econometric software is usually unable to account for the sampling method used, thereby raising the probability of sample selection bias (Jenkins, 1995).

[2] Jenkins (1995).

the percentage of households that had changed their original position was substantial. The trends in repeated poverty during the growth period were similar so that during the pre-transition era: the fraction of households experiencing two year poverty oscillated around 10 percent after it had worsened significantly during the recession itself (Okrasa, 1999).

This paper is based on the second panel referring to 1997–2000. In this panel 12208 households are incorporated.

Defining the category of poverty is a key element in assessing its range and depth[1]. The kind of definition accepted will determine which groups of society may be acknowledged as the most vulnerable to poverty risk. The choice of a method setting the poverty margin depends on whether poverty is treated as a relative or absolute category. In the first case, poverty is understood as a relative deprivation, the level of which is depicted in a relation to wealth of other, better situated members of the society. In the second case, poverty is understood as a lack of fixed sources of income independent of the level of resources available to an overwhelming part of society which the impoverished individual is a member of[2].

[1] Kot (2000, p. 182).
[2] Panek, Podgórski, Szulc (1999).

Table 1. A trend of real average equivalent income as well as characteristics of impoverished households

Characteristics of impoverished households				1997	1998	1999	2000
Real average equivalent income (for all households)			(in PLN)	9360	9130	8940	9200
			(in %)	100	97,52	95,54	98,30
Poverty margin	50% median income	Impoverished households (%)		7,16	7,85	7,59	7,55
		Real average equivalent income of poor families (in PLN)		2850	2600	2780	2860
		Real average equivalent income of wealthy families (in PLN)		10520	10370	10160	10390
		Average poverty gap[1] (in PLN)		1050	1215	1055	1015
		Income gap index[2]		0,27	0,31	0,27	0,26
	75% median income	Impoverished households (%)		23,05	22,82	21,97	21,54
		Real average equivalent income of poor families (in PLN)		4333	4132	4132	4130
		Real average equivalent income of wealthy families (in PLN)		11650	11420	11130	11370
		Average poverty gap (in PLN)		1610	1706	1630	1685
		Income gap index		0,27	0,29	0,28	0,29

Source: CHER database: Poland (1997–2000); own calculations.

Literature on the subject discusses the following lines of poverty: absolute, relative, subjective and official[3]. Nevertheless, most of the research[4] shows the poverty's boundary as a proportion of central tendency measures (average or median) of a distribution of a given population. In the poverty analysis it is reasonable to use various poverty boundaries[5]. In this article I consider a household (and thus all its members) as impoverished[6] if its real equivalent

[1] A poverty gap shows how far poor persons are from the poverty margin, that is, how much on average each family, living in the poverty zone, should receive in order to find itself exactly on the poverty margin.

[2] An index used to synthetically assess a depth of poverty is an income gap index defined as: $I = \frac{1}{N \cdot z} \sum_{i=1}^{N} \left(z - \frac{y_i}{m_i} \right)$, where: N – a number of households, z – a poverty margin, y_i – an income of household i, m_i – equivalence scale of i-th household (Panek, Podgórski, Szulc, 1999).

[3] Golinowska (1997, p. 20-24]. The lines of absolute poverty were used in research by Denisczuk, Sajkiewicz (1995), Kurowski (2008), relative lines were used by GUS in Poland (Kordos, Ochocki, 1993), subjective lines of poverty in the following studies (Kot, 1995, 1998, Kasprzyk, 2000).

[4] The relative lines of poverty were used in studies by Eurostat and by GUS in Poland. In the European Union countries as an official boundary of poverty - half of an average income - is used.

[5] Steward, Swaffied (1999), Cappellari (2000).

[6] It is important to notice that the level of poverty boundary and its proportion to delineating the equivalent income are a result of statistical agreement.

income[1] is lower than the poverty boundary line of 50% or 75% of real equivalent median income measured according to the OECD relative scale (100/70/50)[2]. However, it is important to notice that according to Hagenaars et al. (1994) the OECD scale is too much focused on large families and therefore , the so-called OECD modified scale[3] has been proposed. The choice of using the OECD original scale in research is justified for at least two reasons. Firstly, the original scale is used by Eurostat, and secondly – in the 1990s – it showed a relatively high conformity with scales estimated on the basis of real expenses of households taking part in family's budget research (Szulc, 1995).

Table 1 presents the primary characteristics of poverty in the analyzed sample for Poland between 1997–2000. Over this period the average income declined by approximately 2%. The average real income of the population – after a dramatic decrease at the beginning of the transformation period – was continuously increasing until 1998. From this moment, there was a vivid decline in high, over the last few years, economic dynamic followed by a later reduction of the rate of growth in the real income of the population. The number of individuals having income below a relative boundary of poverty of 50% median income, after an increase in 1998 began to drop. Yet, at the relative boundary of poverty of 75% median income, the number of impoverished households shows a declining trend since the beginning of the studied period. The amount needed to combat poverty with the use of the poverty boundary of 50% median income is on average 1100 PLN (about 1700 PLN when a poverty boundary of 75% median income is used). The average wealth of an impoverished household group is lower by about 27% than the poverty line (in 1999 even by 31%) with the use of a poverty boundary of 50% median income (on average 28% when a poverty boundary of 75% median income is applied).

2.1.Characteristics of independent variables used in the study

The choice of independent variables in the panel model was based on previous studies which pertained to poverty analysis. Variables that have been used in the two types of the models are as follows: poverty exit rates and poverty return rates are presented below. They include:

---

[1] In the study a net real equivalent income was considered, expressed in year 2000 currency (indices for the following years:1 [2000], 1·1,073 [1999], 1·1,073·1,118 [1998], 1·1,073·1,118·1,149 [1997]) (Inflation in the years 1997–1999 equalled: 14,9%, 11,8%, 7,3%, „Indices for prices of consumer goods and services for 1950–2002" can be found at, http://www.stat.gov.pl).

[2] Net equivalent income was set using a equivalent scale utilized by GUS, which gives a weight of 1 per head of the household and 0,7 for every other adult person (age>15 years) as well as 0,5 for every child in the household (age =<15 years).

[3] Net equivalent income was set using a modified equivalent scale, which gives a weight of 1 per head of the household and 0,5 for every other adult person (age>15 years) as well as 0,3 for every child in the household (age =<15 years).

- Duration of poverty (in years) - a variable used in exit from poverty model: **duration_poverty$_{it}$** (composed of three levels: 1 – exit from poverty after 1 year, 2 – exit from poverty after 2 years, 3 – exit from poverty after 3 years). A variable used in return to poverty model: duration of wealth (in years) – **duration_wealth$_{it}$** (composed of three levels: 1 – return to poverty after 1 year, 2 – return to poverty after 2 years, 3 – return to poverty after 3 years).
- Gender of the head of the household: **gender$_{it}$** (composed of two types: 1 – male, 2 – female).
- Professional status of the head of the household: **status$_{it}$** (composed of four levels: 1 – employed and self-employed, 2 – retired, 3 – unemployed, 4 – not active professionally for other reasons).
- Education status of the head of the household: **education$_{it}$** (composed of three levels: 1 – incomplete primary and primary, 2 – secondary and vocational, 3 – higher).
- Place of living: city**, countryside$_{it}$** (composed of two types: 1 – city, 2 – countryside).
- Geographical region[1]: **region$_{it}$** (composed of four types: 1 – Eastern Poland and Warmia and Mazury, 2 – Southern Poland, 3 – Western Poland and Pomeranian, 4 – Centre).

[1] The following regions form the voivodships: Eastern Poland and Warmia and Mazury: lubelskie, podkarpackie, podlaskie, warmińsko-mazurskie; Southern Poland: małopolskie, opolskie, śląskie, świętokrzyskie; Western Poland and Pomeranian: dolnośląskie, lubuskie, pomorskie, zachodnio-pomorskie; Centre: kujawsko-pomorskie, łódzkie, mazowieckie, wielkopolskie. This classification is different from the division of the country into regions by GUS (GUS, 2008), however, for the purpose of this paper such an assumption was the most convenient.

Table 2. Characteristics of independent variables are used in the model

Independent variables	The rate of exit from poverty		The rate of return to poverty	
	Impoverished families	Wealthy families	Impoverished families	Wealthy families
Variables pertaining to the head of the family				
Gender (%):				
– male	62	66	58	65
– female	38	34	42	35
Age (average in years)	45	45	46	45
Education (%):				
– incomplete primary and primary	38	32	32	32
– secondary and vocational	61	65	67	64
– higher	0,8	3	1	3
Professional status (%):				
– employed and self-employed	60	70	63	69
– retired	10	10	9	12
– unemployed	5	2	5	0,2
– not active for other reasons	25	18	23	18
Extent of the balance in the family budget (%):				
– surplus	4	24	77	97
– deficit	96	76	23	3
Variables pertaining to the household				
Number of children below six years of age (average)	1,4	0,4	–	–
Number of children of the age between of 6 and 16 (average)	1,3	0,4	–	–
Place of employment (%):				
– city	–	–	36	42
– countryside	–	–	64	58
Region (%):				
– Eastern Poland and Warmia and Mazury	28	28	32	28
– Southern Poland	21	21	19	21
– Western Poland and Pomeranian	20	17	15	18
– Centre	31	33	33	33
Macroeconomic variables				
Unemployment rate (average in %)	14	13	–	–
Year of entry into poverty (%):			–	–
– 1997	26	-	–	–
– 1998	26	33	–	–
– 1999	25	38	–	–
– 2000	24	29	–	–
Number of households	2722	775	192	1238

Source: CHER database: Poland (1997–2000); own calculations.

- Extent of the balance in the family's budget: **budget$_{it}$** (composed of two types: 1 – surplus, 2 – deficit).
- Year of entry into poverty: **year$_{it}$** (composed of four levels: 1 – 1997, 2 – 1998, 3 – 1999, 4 – 2000).
- Age of the head of the household: **age$_{it}$** and age squared: age_2$_{it}$ (literature on the subject emphasizes a nonlinear relation between age and exit from and return to poverty).
- Number of children below 6 years of age: **number_children6$_{it}$**.
- Number of children between the ages of 6 and 16: **number_children16$_{it}$**.
- Unemployment rate according to voivodship: **rate_unemployment$_{it}$**.

For the correctness of conducted analysis, some variables have been transformed into 0–1 variables[1]. Table 2 includes characteristics of these independent variables used in the model[2].

3. Nonparametric method

This part presents the results of the study using the nonparametric method[3] on the topic of duration of poverty. It also provides distributions of the years spent in poverty in one-spell or multi-times episodes.

3.1.Statistics describing the dynamic of poverty

Table 3 shows that when using a poverty margin of 50% median income, 17% of the population studied was affected by poverty (40% at a poverty margin of 75% median income), in this case, 1,2% (8,6%) of the population was included in a prolonged poverty. The expected amount of time spent in poverty at a poverty margin of 50% median income, for those entering the panel and studied during

[1] Some variables were transformed into 0–1 variables: duration_poverty_02$_{it}$, duration_poverty_03$_{it}$ (1if a family exits the poverty after 2 years, 3 years, respectively), duration_wealth_02$_{it}$, duration_wealth_03$_{it}$ (1 if a household returns to poverty after 2 years in wealth, after 1 year in wealth, respectively), gender_02$_{it}$ (1 if the head of the household is a female), status_02$_{it}$, status_03$_{it}$, status_04$_{it}$ (1 if the head of the household is: retired, not employed, non active for other reasons), education_02$_{it}$, education_03$_{it}$ (1 if the head of the household has one of the following education types: vocational and secondary, higher), city_countryside_02$_{it}$ (1 if the household is located in the countryside), region_02$_{it}$, region_03$_{it}$, region_04$_{it}$ (1 if the household is located as follows: Southern Poland, Western Poland, the Centre, respectively), budget_02$_{it}$ (1 if the family had a negative balance of the family budget), year_02$_{it}$, year_03$_{it}$, year_04$_{it}$ (1 if the household entered poverty in one of the following years 1998, 1999, 2000, respectively).

[2] In the parametric method property margin of 75% median income has been used (100/70/50 scale).

[3] The nonparametric method provides information about a change in individual's behavior dependent on time under assumption of nonexistence of a particular form of an event distribution (Frątczak, Babiker, Gach-Ciepiela, 2005).

the 4 year period, equals to 0,3 part of the year; for the poverty margin of 75% median income, it equals to 0,89 part of the year.

Table 3. Household distribution by the number of years spent in poverty

Number of years in poverty	Percentage share of impoverished households with the use of a relative margins of poverty	
	50% median	75% median
0	82,89	60,56
1	8,90	13,79
2	4,56	9,98
3	2,46	7,03
4	1,18	8,64

Source: CHER database: Poland (1997–2000); own calculations.

Table 4 presents sequences of income in years 1997–2000 for the studied sample. If a household was in poverty in a given year, it is marked as "U", if not, then as "N". The first column shows how long the poverty lasted, whether the studied individual exited poverty and whether they returned to it.

A long-term perspective presented in Table 3 and Table 4 might be compared to a short-term cross-section view proposed in Table 1. A rate of poverty at a given point of time equals, on average, approximately 23% at a poverty margin of 75% median income and about 8% at a poverty margin of 50% (Table 1). Table 3 shows though that 40% of the studied population at the poverty margin of 75% median income (17% at a poverty margin of 50%) had been affected by poverty at least once. However, the data in Table 4 demonstrates that in many cases, individuals experienced poverty in the subsequent years (multi-time episodes). One can thus reason that households entering (or exiting) poverty might be beginning a long period of remaining in that state and, what is more important, they are non-differentiable from those that are located below the poverty margin in only one or two observed years. As a result, the extent of the phenomenon of poverty is underestimated. In each of these four years, one can note that this applies to about 8% of impoverished households at a poverty margin of 50% of the median income (23% at a poverty margin of 75%), while in the entire four-year period, shorter or longer episode of poverty was overcome by as much as 17% of households at the poverty margin of 50% median income (40% at the poverty margin of 75%).

Table 4. Household distribution in a 4-year sequence of states (U – poverty, N – not in poverty)

State sequences	Percentage share of impoverished households using a relative margins of poverty	
	50% median	75% median
NNNN	82,89	60,56
NNNU	2,43	3,15
NNUN	1,81	2,66
NNUU	1,18	2,36
NUNN	2,36	3,28
NUNU	0,56	1,18
NUUN	0,66	1,58
NUUU	0,95	2,17
UNNN	2,30	4,70
UNNU	0,53	1,25
UNUN	0,69	1,08
UNUU	0,33	1,41
UUNN	0,95	2,53
UUNU	0,39	1,38
UUUN	0,79	2,07
UUUU	1,18	8,64
Total	100,00	100,00

Source: CHER database: Poland (1997–2000); own calculations.

3.2. Rates of exit from and re-entry to poverty (Kaplan-Meier estimator)

In this study of exit from and return to poverty rates a nonparametric method has been used. According to the Devicienti's [1] definition(2002), the exit rates that are relevant in this context are the ones that refer to a cohort of individuals just falling into poverty, hence at risk of remaining in poverty thereafter. The re-entry rates refer instead to a cohort of individuals just starting a spell out of poverty, and so at risk of re-entering. Exit rates are calculated by dividing the number of individuals ending a spell after *d* years in poverty by the total number with low income for at least *d* years. The re-entry rates were calculated analogously. To estimate the exit rate, the author used the following sequences: *NUxx* and *xNUx*, where $x = N,$ U (NNUN, NUNN, NUNU, UNUN, NNUU, NUUN, UNUU, NUUU), however the return rate: *UNxx* and *xUNx*, where $x = N, U$ (UNUU, UNUN, UUNU, NUNU, UUNN, NUNN, UNNU, UNNN). Unlike the simple calculation of the number of years in poverty, such a periodical approach can include the right-side censored observations. For the purposes of this analysis, the

[1] Devicienti (2002, p. 8–9).

left-side censored observations have been excluded, on account of which the research study starts from 1998 or later.

Table 5. Survival function and rates of exit from poverty (Kaplan-Meier estimator)

Years	Relative poverty margin			
	50% median income		75% median income	
	Survival function	Return rate	Survival function	Return rate
1	1 (-)	0,6346 (0,0494)	1 (-)	0,5219 (0,0330)
2	0,3654 (0,0299)	0,4082 (0,0913)	0,4781 (0,0228)	0,4211 (0,0608)
3	0,2162 (0,0312)	-	0,2768 (0,0258)	-

**) in brackets standard errors are provided.*
Source: CHER database: Poland (1997– 2000); own calculations.

Table 5 shows exit rates from poverty using a relative poverty margin. An estimated hazard rate confirms negative effects of poverty duration: the longer the individual lives in poverty, the lesser probability that this state will change in the next period. For the cohort of individuals that begin a period of poverty, the probability of exit after the first year is equal to about 64%, and after two years to approximately 41% (a relative poverty margin of 50% median income), however, for the same cohort the probability of exit from poverty after the first year is equal to about 52%, and after two years to approximately 42% (a relative poverty margin of 75% median income).

Table 6. Survival function and rates of re-entry to poverty (Kaplan-Meier estimator)

Years	Relative poverty margin			
	50% median income		75% median income	
	Survival function	Return rate	Survival function	Return rate
1	1 (-)	0,2429 (0,0314)	1 (-)	0,2008 (0,0203)
2	0,7571 (0,0273)	0,1860 (0,0465)	0,6992 (0,0203)	0,2099 (0,0341)
3	0,6162 (0,0388)	-	0,5524 (0,0265)	-

**) in brackets standard errors are provided.*
Source: CHER database: Poland (1997–2000); own calculations.

Generalizing results from Table 5 and Table 6 implies the thesis that low income includes a wide spectrum of households. Yet, it is not a fixed group. Although there are individuals present in it who are permanently poor, many families exit from and enter poverty.

3.3. Poverty duration in one-spell and multi-time episodes

The estimation of exit from and re-entry to poverty allows us to introduce "distribution of time spent in poverty". Such distribution is a basic measure of duration of poverty.

The distribution of years spent in one-spell episodes of poverty has been calculated using only exit rates presented in Table 5 (e.g. two years spent in poverty depicted as (*NUUN*), where *N* - a period of wealth and *U* - a period of poverty). It was assumed that $e(d)$ and $r(d)$ are respectively exit rate from and re-entry to poverty after d years. With exception of the left-side censored periods, the probability of this sequence is calculated as: $(1-e(1))*e(2)$.

The distribution of years spent in poverty in the case of multi-time episodes was calculated using the exit and re-entry rates presented in Table 5 and Table 6 (e.g. two years in poverty depicted as (*NUNU*)). With exception of the left-side censored periods, the probability of this sequence is calculated as: $(UNU)=e(1)*r(1)$.

To calculate the probability of observing two years of living in poverty within these four years, one must calculate the probability of occurrence of all possible combinations which generate a sum of two years in poverty and adding them up. In comparison to the forecasts for one-spell and multi-time periods, also a distribution of "time spent in poverty" was calculated using the following sequence: ($N\ U\ x\ x$), where $x = N, U$.

Comparing Columns 2 and 3 in Table 7 one can notice that in one-spell episodes the distribution of "time spent in poverty" was overestimated for one year spent in poverty. For two-year duration of poverty the use of "one-spell episode" approach in each case results in underestimation of distribution of the time spent in poverty.

Table 7. Distribution of years spent in poverty for the cohort of individuals beginning a period in poverty in 1998

Years in poverty	Distribution in one-spell episodes		Distribution of years spent in poverty in the following three years			
			expected[1]		actual[2]	
	50% median income	75% median income	50% median income	75% median income	50% median income	75% median income
1	0,6346	0,5219	0,4805	0,3649	0,5217	0,4000
2	0,1492	0,2013	0,3033	0,3583	0,2681	0,3360
3	0,2162	0,2768	0,2162	0,2768	0,2101	0,2640

Source: CHER database: Poland (1997–2000); own calculations.

[1] Expected distribution of years spent in poverty was delineated using exit and entry rates from/to poverty taken from the Kaplan-Maier model (Table 5 and Table 6).

[2] Actual distribution of years spent in poverty was delineated using the following sequence (*NUxx*), where x=*N*,*U*.

It is important to emphasize that the above mentioned analysis assumes that all observed episodes pertain to a homogeneous population. However, it is more probable that different families having special characteristics (observed and unobserved) meet various rates of exit and re-entry to poverty, which explains why permanently poor households exist. As a result one must change from unilateral to multilateral approach which allows for dependence of an exit rates and re-entry rates on important socio-economic correlations. In the parametric method the poverty margin of 75% median income is used because it gives more possibilities to study the processes of exit from and re-entry to poverty in a short-run.

4. Parametric method: model with discrete time taking into consideration observed and unobserved heterogeneity of individuals

This part presents the results of studies on duration of poverty obtained by the use of parametric method[1].The estimated model contains variables pertaining to characteristics of the head of a household as well as the household itself and also conditions on the labor market, which have impact on the probability of exit and re-entry into poverty later on.. The estimation of the two complementary log-log models with discrete time has been conducted. The first one (Model I) includes elastic non-parametric specification for the base hazard function. The second one (Model II) describes the base hazard, using the multi-dimensional duration function, that is: $\theta_0(t) = at+bt^2+ct^3$. The use of these two models was justified by the observations made by Meyer (1990) that: "parameters of the base hazard function depict an important characteristic of the data, which would be omitted, if the model would have been estimated by a simple parametric base hazard function". Models Ia and IIa that include an unobserved heterogeneity have been also estimated.

4.1. Who exits from poverty*?*

Table 8 presents the analysis of duration of episodes of poverty that end by the household gaining a wealth status. Model 1 confirms a negative relation between duration of poverty and the probability of leaving these conditions, which was earlier observed on the basis of the results from Table 5: the longer a family remains in poverty, the harder it is for them to re-enter the state of wealth. The probability of exit rate from poverty also depends on the characteristics of the head of a family (e.g. gender, education) and a household (e.g. number of children, family's annual budget). Basing on collected evidence, households

[1] About parametric models it is said that the analytical form of probability distribution of density is known.

headed by women remain in poverty longer, which is shown by Model II. The probability of exit from poverty decreases along with the increase in the number of children in a household (aged up to 6 years) as bearing children makes full-time employment more difficult (especially in the case of women).

Moreover, the results of the estimation show that the higher educational level of the head of a family, the lower the chance of experiencing a relatively long period of poverty by their family and the easier fight with poverty. Having a higher education gives a many-times greater chance of exit from poverty than in the case where the head of a family has a secondary or vocational education. An important aspect of the possibility to exit poverty is also the status of the head of a family on the labor market. Households where the family is headed by an unemployed person are in a decisively worse situation. Furthermore, 1% increase in an unemployment rate in a voivodship, decreases the probability of exit from poverty by approximately 2,5 % according to Model I. A family budget deficit has also a negative impact on exit from poverty, another factor in delineating duration of poverty being a macroeconomic situation of a country as well as a period on which the beginning of impoverishment fell (for household which became poor in 1998, 1999 and 2000, a chance to exit poverty was much higher that for those in the base group which entered poverty in 1997). This phenomenon could have taken place due to economic growth in Poland[1].

[1] Results of international research (Barro, 1999) and interdependencies between unequal general income and economic growth in Poland during the transformation period confirm that one cannot expect, in the conditions of Polish economy, a significant limitation of income disproportions. An observations of annual fluctuation of GDP increase rates and Gini factor over the years 1989–1999 confirm that in Poland a higher speed of economic growth was accompanied by a lower build up of inequalities in terms of social division of income(Jabłoński, 2002).

Table 8. Analysis of poverty duration in Poland – exit from poverty

Independent variables	Without unobserved heterogeneity		With unobserved heterogeneity	
	Model I	Model II	Model Ia	Model IIa
	Parametr [standard error]			
Duration of poverty:				
– two years	- 0,910* [0,095]	-	- 0,815* [0,121]	-
– three years	- 1,989* [0,097]	-	- 2,528* [0,161]	-
Year of entry into poverty:				
1998	2,586* [0,215]	-	2,191* [0,293]	-
1999	2,793* [0,223]	-	2,564* [0,326]	-
2000	2,679* [0,232]	-	3,177* [0,370]	-
Variables pertaining to the head of the family				
Gender (female)	-	- 0,2445* [0,0817]	-	- 0,422* [0,153]
Age	- 0,106* [0,010]	-	- 0,113* [0,012]	-
Age_2/100	0,091* [0,011]	-	0,098* [0,013]	-
Professional status:				
– retired	- 0,140 [0,169]	- 0,379* [0,131]	- 0,018 [0,203]	- 0,572** [0,249]
– unemployed	- 0,856* [0,336]	- 1,370* [0,338]	- 0,983** [0,456]	- 2,005* [0,542]
– not active professionally	- 0,269* [0,109]	- 0,557* [0,104]	- 0,383* [0,135]	- 1,192* [0,227]
Education:				
– vocational or secondary	-	0,032 [0,083]	-	0,121 [0,149]
– higher	-	1,165* [0,272]	-	2,098* [0,460]
Variables pertaining to the household				
Number of children up to 6 years	- 0,363* [0,060]	- 0,395* [0,055]	- 0,317* [0,075]	- 0,627* [0,133]
Number of children between 6 and 16 years of age	- 0,120* [0,040]	- 0,305* [0,039]	- 0,152* [0,050]	- 0,594* [0,107]
Budget deficit of the family	- 0,797* [0,084]	- 0,712* [0,078]	- 0,670* [0,101]	- 0,857* [0,129]
Unemployment rate	- 0,029** [0,011]	- 0,049* [0,010]	-0,0844* [0,028]	- 0,175 [0,029]
t	-	- 2,601* [0,295]	-	- 1,153* [0,427]
t^2	-	1,947* [0,155]	-	1,191* [0,231]
t^3	-	- 0,334* [0,024]	-	- 0,200* [0,036]
Number of families	3497			
Logarithm likelihood	- 1294,438	- 1557,785	- 1056,284	- 1269,667

*Stars signify a significant of parameters of the following levels: * 1%, ** 5%,*** 10%.*
Source: CHER database: Poland (1997–2000); own calculations.

The results of the estimations of Models Ia and IIa, taking into consideration unobserved heterogeneity of the studied individuals are included in Table 8. These differ from the results obtained on the basis of Models I and II. In some cases, an absolute value of coefficients is greater, which strengthens the impact of regressors on a chance of exit from poverty. Models Ia and IIa are also characterized by a greater value of the logarithm likelihood.

4.2. Who returns to poverty?

The results of the estimated models of the chances of the families to return to poverty are included in Table 9. It is important to notice that estimated coefficients are characterized by a greater variability than in the case of exit rates. Model I shows a positive interdependence between duration in wealth after exit from poverty and re-entry to it, which weakens with an increase in years spent in wealth. However, on the basis of the model including heterogeneity of the studied individuals, one can discern that even after two years spent in wealth there is a negative relationship between wealth and re-entering poverty. Therefore, the longer a person stays unaffected by poverty, the lower the chance of re-entering it. Moreover, the probability of returning to poverty decreases if: families live in the city, the head of a family is a man (an interesting fact is that this variable, in Model I for estimating the rate of exit from poverty, was insignificant) and when the education level of the head of a family increases.

Model II also includes variables describing a geographical region where the household is situated. The results of estimations do not show significant differences between voivodships (in each case there was a negative relation). Households in voivodships of Western Poland are, however, in a better situation in comparison to households in voivodships of the Eastern Poland as well as Warmia and Mazury.

Table 9. Analysis of poverty duration in Poland – return to poverty

Independent variables	Without unobserved heterogeneity		With unobserved heterogeneity
	Model I	Model II	Model Ia
	Parametr [standard error]		
Duration of wealth:			
– two years	0,456** [0,192]	-	- 0,131 [0,540]
– one year	1,461* [0,187]	-	1,625* [0,219]
Variables pertaining to the head of the family			
Gender (male)	- 0,496* [0,146]	- 0,687* [0,140]	- 0,553** [0,278]
Age_2/100	-0,041* [0,005]	- 0,035* [0,006]	- 0,042* [0,007]
Professional status: unemployed	1,191* [0,411]	1,297* [0,364]	0,803*** [0,463]
Education:			
– vocational or secondary	-0,672* [0,145]	- 0,545* [0,154]	- 0,776** [0,333]
– higher	-1,719** [0,717]	- 2,002* [0,720]	- 1,527** [0,725]
Variables pertaining to the household			
Budget deficit of the family	1,101* [0,089]	-	1,118* [0,098]
Region:			
– South	-	- 0,593* [0,201]	-
– Eastern	-	- 0,770* [0,221]	-
– Centre	-	- 0,552* [0,166]	-
Place of living: countryside	- 0,287** [0,146]	-	- 0,343** [0,155]
t	-	- 0,545** [0,255]	
t^2	-	0,286* [0,072]	
Number of families	1430		
Logarithm likelihood	- 487,939	- 552,704	- 412,127

*Stars signify a significant of parameters of the following levels: * 1%, ** 5%,*** 10%.*
Source: CHER database: Poland (1997–2000); own calculations.

Conclusions

Poverty has always been one of the most important problems of the contemporary Poland. The research based on panel data representative of the whole country has shown a low income dynamic of families living in Poland between 1997–2000. The total number of years spent in poverty and various sequences of entry to and exit from poverty have indicated the permanence of this phenomenon in the population.

As follows from the conducted analysis, a small number of families (less than 1,5% of the population at a poverty margin of 50% of median income and less than 9% at a poverty margin equal to 75%) was impoverished for the entire period of the study, yet a much higher percentage of the entire population experienced

poverty at least once (17% at the poverty margin of 50% of median income and 40% at the poverty margin equal to 75%). A relatively wide array of households live on low income . However, it is not a fixed group. Even though there are some individuals who are permanently impoverished poverty transitions occur. Approximately 65% of the studied families at a poverty margin of 50% (52% at the poverty margin equal to 75%) come out of poverty after the first year of being impoverished. Nevertheless, only 25% at a poverty margin of 50% (20% at the poverty margin of 75%) become poor once again. After two years of living in poverty, 40% of families exit it, but 18% at a poverty margin equal to 50% (20% at a poverty margin equal to 75%) re-enter a group of impoverished families. Because of a short time period of the sample the study could not have incorporated a wider analysis of this type. One can assume, however. that the longer the family lives in poverty, the more difficult it is for them to change their situation , and even if their income increases above the poverty margin, the probability of these households becoming poor once again is still quite high.

Another important issue to analyze is the distribution of the number of years spent in poverty, considering exits and re-entries of families to this situation. Taking into account the sequences of episodes of being impoverished and wealthy, where the families found themselves in particular years, one can forecast that as much as 50% of the studied families at a poverty margin equal to 50% (65% at a poverty margin equal 75%) will spent at least 2 following years in poverty. However, when including only one-spell episodes of becoming impoverished, these prognoses come to a level of ca.35% of the studied families at a poverty margin equal to 50% (48% at a poverty margin equal to 75%). The actual values amount to 48% at a poverty margin of 50% (60% at a poverty margin equal to75%) therefore, including the sequences of the episodes improves the results. The observed results also show that the length of time of poverty may depend on a type of the definition of exit from and re-entry to poverty rates that is accepted. The analysis using the parametric method illustrates that there are groups of population more vulnerable to dropping below the poverty margin with higher probability of remaining impoverished over a longer period of time. These are households composed of a larger number of not only children but also adults, where the head of a family is an older person (mainly a woman) with a low education level. The life in voivodships with a high rate of unemployment is a factor increasing poverty, and especially endangered are households where the head of a family is unemployed.

Okrasa (1999) used four-year panel data (1993–1999) from Poland's Household Budget Survey to explore the distinction between transitory and long-term poverty and examine poverty mobility. The section of population that could minimize or avoid chronic poverty in Poland included those living in urban areas, headed by older and better educated, with few children and unemployed members and possessing financial or physical assets. Households with a larger kinship network faced significantly less danger of falling into chronic poverty or vulnerability.

The results of the estimation of models including unobserved heterogeneity of studied individuals confirm a negative interdependence between the exit from poverty rate and its duration. For parents staying a relatively long periods of time in poverty it is much more difficult to exit it on their own. Though, the longer the household that exited poverty remains outside that situation, the lower the chance that they will return (a negative relation exists after two years of being above the poverty margin).

The above conclusions may be useful for preparing plans to deal with long-term poverty in Poland. They will allow for a better understanding of the poverty phenomenon in Poland and factors that cause it. The results obtained might be helpful for an effective policy formulation: adding to an income of employed household members, early identification of families often entering the sphere of poverty and permanently impoverished, modeling of conditions on the labor market so as to lower the incidence of poverty in Poland.

REFERENCES

ALLISON, P. D. (1982), *Discrete-time methods for the analysis of event histories*, in Sociological Methodology 1982 (S. Leinhardt, ed.), Jossey-Bass Publishers, San Francisco 1982.

BANE, M., ELLWOOD, D. (1986), *Slipping Into and Out of Poverty: The Dynamics of Spells, "Journal of Human Resources***"**, Vol. 21.

BARRO, R.. J. (1999), *Inequality, growth and investment*, Working Paper 7038, "National Bureau of Economic Research".

BIEWEN, M. (2003), *Who are the chronic poor? Evidence on the Extent and the Composition of Chronic Poverty in Germany*, "IZA Discussion Paper", No. 779.

CIURA, G. (2002), *Poverty and impoverished area*, The Bureau of Research, No. 884, (In Polish).

CAPPELLARI, L., JENKINS, S. (2002), *Modeling low income transition,* "Discussion Papers of DIW Berlin from DIW Berlin", German Institute for Economic Research, No. 288.

CIECIELĄG, J., TOMASZEWSKI, A. (2003), *Econometric analysis of panel data*, (In Polish), Warsaw 2003.

DEVICIENTI, F. (2002), *Poverty persistence in Britain: a multivariate analysis using the BHPS, 1991–1997*, "Journal Of Economics", No. 9.

DEVICIENTI, F. (2002), *Estimating poverty persistence in Britain,* "LABORatorio R. Revelli Working Papers Series".

DOLTON, P., VAN DER KLAAUW, W. (1995), *Leaving teaching in the UK: a duration analysis*, "Economic Journal", Vol. 105, No. 429.

DUNCAN, G. (1999), *The PSID and Me*, "IPR working papers".

FRĄTCZAK, E., BABIKER, H., GACH-CIEPIELA, U. (2005), *History occurances analysis – Elements, theory, chosen practical examples*, (In Polish), Warsaw 2005.

Central Statistical Office, *Existence situation of households in 2003*, (In Polish),

www.stat.gov.pl/dane_spol-gosp/warunki_zycia/syt_byt_gosp_2003/warunki_zycia03.doc

GOLINOWSKA, S. (1996), *Polish poverty. Criteria. Assessment. Counteraction.*, The Institute of Labour and Social Studies, (In Polish), Warsaw 1996.

GOLINOWSKA, S. (1997), *Polish poverty II. Criteria. Assessment. Counteraction.*, The Institute of Labour and Social Studies, (In Polish), Warsaw 1997.

HANSEN, J., WALHBERG, R. (2004), *Poverty persistence in Sweden*, „IZA Discussion Paper", No. 1209.

JABŁOŃSKI, Ł. (2002), *Economic growth or limiting income inequality in Poland?*, No 4, (In Polish), Rzeszów 2002.

JARVIS, S., JENKINS, S. (1997), *Low income dynamics in 1990s Britain,* "Fiscal Studies" No. 18.

JENKINS, S. (1995), *Easy estimation method for discrete-time duration models,* "Oxford Bulletin of Economics and Statistics", No. 57.

JENKINS, S. (1997), *Estimation of discrete time proportional hazards models*, "Stata Technical Bulletin Reprints".

JENKINS, S. (2004), *Survival Analysis,* unpublished manuscript, Institute for Social and Economic.

KASPRZYK, B. (2000), *Methodological aspects of assessing the level of wealth*, Cracow University of Economics, (In Polish), Cracow 2000.

KIEFER, N. (1988), *Economic Duration Data and Hazard Functions*, "Journal Of Economic Literature", Vol. XXVI.

KORDOS, J., OCHOCKI, A. (1993), *Problems with measuring poverty in EWG countries and Poland*, " Statistical News", (In Polish), Warsaw 1993.

KOT, S. M. (1995), *Modelling level of wealth. Theory and application*, Ossolineum, (In Polish), Wroclaw 1995.

KOT, S. M. (1998), *The Cracow Poverty Line*, Cracow University of Economics, (In Polish), Cracow 1998.

KUMOR, P., SZTAUDYNGER, J. (2007), *Optimal compensation differentiation in Poland – econometric analysis*, „Economist", No. 1, (In Polish), Warsaw 2007.

KUROWSKI, P. (2007), *Study of the modified minimum existance level in 2006*, The Institute of Labour and Social Studies, (In Polish), http://www.solidar.uni.lodz.pl/min.egzystencji.pdf.

KUROWSKI, P. (2008), *Study of the level and structure of the modified minimimum existance level in 2007*, The Institute of Labour and Social Studies, (In Polish), http://www.ipiss.com.pl/Wo-2007.pdf.

LILLARD, L. A., WILLIS, R. J. (1978), *Dynamics aspects of earning mobility,* "Econometrica", Vol. 46.

MEYER, B. D. (1990), *Unemployment insurance and unemployment spell*, "Econometrica", Vol. 58.

OKRASA, W. (1999), *Who Avoids and Who Escapes from Poverty during the Transition? Evidence from Polish Panel Data, 1993–96*, The World Bank, Policy Research Working Paper 2218, Washington.

PANEK, T., PODGÓRSKI, J., SZULC, A. (1999), *Poverty: theory and practical assessment,* Warsaw School of Economics, (In Polish), Warsaw 1999.

PANEK, T. (2007) *Poverty and inequality*, Social Statistics, Polish Economics Publishers, (In Polish), Warsaw 2007.

PRENTICE, R., GLOECKER, L. (1978), *Regression analysis of grouped survival data with application to breast cancer data*, "Biometrics", Vol. 34.

„*Vovoidship annual statistical yearbook*", Central Statistical Office, (In Polish), Warsaw, 1998, 1999, 2000, 2001.

RADZIUKIEWICZ, M. (2007), *Poverty reach in Poland*, (In Polish), Polish Economics Publishers 2007.

RUSNAK, Z., KOŚNY, M. (2004), *Impact of changes in equivalence scales on the income and equivalent expense distributions*, in: Applications of Statistics and Mathematics in Economics, Stasiewicz, W. (Eds.), Wrocław University of Economics, (In Polish) ,Wroclaw 2004.

SHORROCKS, A. K. (1978), *The Measurement of Mobility*. „Econometrica", Vol. 46, No. 5, p. 1013-1024.

STANISZ, A. (2003) *Practical medicine 2002/03*, Practical medicine, (In Polish).

STEVENS, A. (1999), *Climbing out of poverty, falling back in. Measuring the persistence of poverty over multiple spells*, "Journal of Human Resources", Vol. 34, No. 3.

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 505—524

APPLYING THE HADAMARD PRODUCT TO DECOMPOSE GINI, CONCENTRATION, REDISTRIBUTION AND RE-RANKING INDEXES

Achille Vernizzi * [1]

ABSTRACT

Gini and concentration indexes are well known useful tools in analysing redistribution and re-ranking effects of taxes with respect to a population of income earners. There are several attempts in the literature to decompose Gini and re-ranking indices to analyse potential redistribution effects and the unfairness of a tax systems, including ones that consider contiguous income groups being created by dividing the pre-tax income parade according to the same bandwidth. However, earners may be very often split into groups characterized by social and demographic aspects or by other characteristics: in these circumstances groups can easily overlap. In this paper we consider a more general situation that takes into account overlapping among groups; we obtain matrix compact forms for Gini and concentration indexes, and consequently, for redistribution and re-ranking indexes. In deriving formulae the so called matrix Hadamard product is extensively used. Matrix algebra allows to write indexes aligning incomes in a non decreasing order either with respect to post-tax income or to pre-tax incomes. Moreover, matrix compact formulae allow an original discussion for the signs of the *within group*, *across group*, *between* and *transvariation* components into which the Atkinson-Plotnick-Kakwany re-ranking index can split.

Key words: Gini and concentration indexes decompositions, Tax redistributive effects, Tax re-ranking effects, Hadamard product.

* Università degli Studi di Milano, DEAS, achille.vernizzi@unimi.it.

[1] This paper is part of a research project joint with Maria Giovanna Monti (Università degli Studi di Milano and Università degli Studi di Milano-Bicocca) and Mauro Mussini (Università degli Studi di Milano-Bicocca). The author thanks his research fellows for general discussions, observations and revisions. Moreover, the author expresses special gratitude to Mario Faliva, for having encouraged the matrix approach adopted in the present paper and for all his precious suggestions, and to Giorgio Pederzoli for his precise and helpful comments. Remaining deficiencies and mistakes are exclusively due to author's responsibility.

Introduction

It is known that, dealing with a transferable phenomenon where units are classifiable into groups, Gini index fails to decompose additively into a between and a within component if the group ranges overlap. Following Bahattacharya and Mahalanobis (1967), a number of Gini decompositions was proposed (Rao (1969), Pyatt (1976), Mookherjee and Shorrocks (1982), Silber (1989), Yitzhaki and Lerman (1991), Lambert and Aronson (1993), Yitzhaki (1994), Dagum (1997)) and after Lambert and Aronson (1993), the third component of the conventional Gini index decomposition is denoted by overlapping term.

Monti (2007) shows that the conventional and the Dagum (1997) decomposition are identical, so that an alternative way to calculate the overlapping term can be derived from the decomposition suggested by this author.

Aronson, Johnson and Lambert (1994), Urban and Lambert (2008), use Gini and concentration index decomposition to identify and evaluate potential distributive effects and unfairness in a tax system. These authors consider contiguous income groups created by dividing the pre-tax income parade according to an identical bandwidth, so that the pre-tax income parade excludes overlapping by construction.

In the present paper we consider incomes gathered into groups characterized by social, demographic or income sources characteristics, so that overlapping among groups need not to be excluded. Our results are obtained using the Gini index decomposition derived from Dagum decomposition (Monti and Santoro 2007, Monti 2008).

Making use of the Hadamard product, in the first section we present Gini and concentration indexes in compact matrix forms. In the second section we introduce groups, present Gini and concentration indexes and show how within groups, across, between groups and transvariation components can be written in matrix compact forms. Links from matrix compact forms and scalar forms are reported: some scalar expressions are well known in literature, while others appears as modifications of already well known forms.

Section 3 presents matrix forms for redistribution and re-ranking indexes, together with their within, across, between groups and transvariation components.

In the fourth section we show how the signs of Atkinson-Plotnick-Kakwani (Plotnick 1981) re-ranking index components can be analysed, thanks to the algebraic tools presented in the paper.

1. Matrix forms for concentration and Gini indexes

Let X and Y be two real non negative statistical variables that describe a transferable phenomenon for a population of K units, $K \in \mathrm{N}$. In this paper we suppose that X represents income before taxation and Y after-tax income; not

infrequently the pair (x_i , y_i) has associated a weight p_i (i=1,……,K), $\sum_{i=1}^{K} p_i = N$. Furthermore, in measuring concentration we generally need to rank either x_i or y_i in a non-decreasing order: when the X elements are ranked in a non-decreasing order, the sequence of (x_i, y_i, p_i) triplets will be indicated as $\{(x_i, y_i, p_i)\}_X$; analogously, $\{(x_i, y_i, p_i)\}_Y$ will denote the sequence of (x_i, y_i, p_i), when the Y elements are ranked in a non-decreasing order.

The concentration index [1] for Y, in the ordering $\{(x_i, y_i, p_i)\}_X$, is defined as [2]

$$C_{Y|X} = 1 - \sum_{i=1}^{K}\left(\sum_{j=1}^{i}\frac{y_j p_j}{\mu_Y N} + \sum_{j=1}^{i-1}\frac{y_j p_j}{\mu_Y N}\right)\frac{p_i}{N} = \frac{1}{\mu_Y N^2}\sum_{i=1}^{K}\sum_{j=1}^{i-1}(y_i - y_j)\,p_i p_j =$$

$$= \frac{1}{2\mu_Y N^2}\sum_{i=1}^{K}\sum_{j=1}^{K}(y_i - y_j)\,p_i p_j I_{i-j} \qquad (1)$$

$$I_{i-j} = \begin{cases} 1: & i-j>0 \\ 0: & i-j=0 \\ -1: & i-j<0 \end{cases}$$

where μ_Y is the weighed mean of the observations on Y. Obviously, in the ordering $\{(x_i, y_i, p_i)\}_Y$, the concentration index $C_{Y|Y}$ coincides with the Gini index G_Y and, analogously in the ordering $\{(x_i, y_i, p_i)\}_X$, $C_{X|X} \equiv G_X$ [3]. Generally, when tax effects are analyzed, one considers the Gini index for the pre-tax distribution G_X, the Gini index for the post-tax distribution G_Y, and the concentration index for the post tax distributions, $C_{Y|X}$, with incomes ranked according to the $\{(x_i, y_i, p_i)\}_X$ ordering.

---

[1] The author is in debt with Maria Monti for the suggestion to express the concentration index by differences between incomes: this suggestion is at the basis of this paper. It can be shown that in expressions (1) the first formula is equal to the second one: the proof can be easily obtained following the demonstration that Landenna (1994, Ch. 4, § 4.4.) gives for the Gini index. In the right hand side of (1), the first component calculates the normalized concentration. In the case where the y's are in a non decreasing order, the second one is the normalized mean absolute difference, that is $G_Y = (1/2\mu_Y N^2)\sum_{i=1}^{K}\sum_{j=1}^{K}|y_i - y_j|p_i p_j = \Delta/2\mu_Y$.

[2] The indicator function I_{i-j} is a particular case of generalized functions considered in Faliva (2000): this article can be consulted for I_{i-j} properties.

[3] For definitions concerning concentration indexes and their relations with Gini indexes, see e.g. Kakwani (1980), in particular Ch. 5 and 8.

In order to pass to a matrix representation, we stack the K observations on X, Y and the weights P into $K\times 1$ vectors: when referring to the ordering $\{(x_i, y_i, p_i)\}_X$, the vectors will be indicated as $\mathbf{x}$, $\mathbf{y}_X$ and $\mathbf{p}_X$, while, referring to the ordering $\{(x_i, y_i, p_i)\}_Y$, the vectors will be labelled as $\mathbf{x}_Y$, $\mathbf{y}$ and $\mathbf{p}_Y$, that is, when elements in a vector are ranked in a non-decreasing order no label will be added, conversely, when they are ordered according to a non-decreasing order for another variable, this variable will be explicitly indicated.

We also introduce the following definitions:

$\mathbf{S} = \left[s_{i,j} \right]$ will denote a $K\times K$ semi-symmetric matrix with diagonal elements equal to zero, super-diagonal elements equal to 1 and sub-diagonal elements equal to -1;

$\mathbf{j}$ for a $K\times 1$ vector that has entries equal to 1;

$\mathbf{D}_X$ and $\mathbf{D}_Y$ will denote the $K\times K$ matrices $\mathbf{D}_X = (\mathbf{jx'} - \mathbf{xj'})$, $\mathbf{D}_Y = (\mathbf{jy'} - \mathbf{yj'})$.

Then, by making use of the Hadamard product $\circ$, we can express the indexes G_Y and G_X as follows [1]:

$$G_Y = \frac{1}{2\mu_Y N^2} \mathbf{p}_Y' (\mathbf{S} \circ \mathbf{D}_Y) \mathbf{p}_Y \qquad G_X = \frac{1}{2\mu_X N^2} \mathbf{p}_X' (\mathbf{S} \circ \mathbf{D}_X) \mathbf{p}_X \tag{2}$$

where μ_Y and μ_X are the weighed mean of the observations on Y and on X, respectively.

In addition, by introducing the $K\times K$ matrix $\mathbf{D}_{Y|X} = (\mathbf{j}\mathbf{y}_X' - \mathbf{y}_X\mathbf{j}')$, we can write the concentration index in compact form as

$$C_{Y|X} = \frac{1}{2\mu_Y N^2} \mathbf{p}_X' (\mathbf{S} \circ \mathbf{D}_{Y|X}) \mathbf{p}_X \tag{3}$$

The transformation from vectors $\mathbf{y}$ and $\mathbf{p}_Y$ to vectors $\mathbf{y}_X$ and $\mathbf{p}_X$ can be performed by a proper $K\times K$ permutation matrix $\mathbf{E}$. The reverse transformation from $\mathbf{y}_X$ and $\mathbf{p}_X$ to $\mathbf{y}$ and $\mathbf{p}_Y$ can be obtained through the matrix $\mathbf{E}^{-1}$ which is equal to $\mathbf{E'}$. Formally

[1] The Hadamard product for two matrices **A** and **B** is defined if both of them have the same number of rows and the same number of columns: $\left[a_{i,j} \right] \circ \left[b_{i,j} \right] = \left[a_{i,j} \cdot b_{i,j} \right]$. For the definition and properties of the Hadamard product see, e.g., Faliva (1983, Appendix) and (1987, Ch. 3), Schott (2005, Ch. 5).

$$\begin{cases} \mathbf{y}_X = \mathbf{E}\mathbf{y}, & \mathbf{y} = \mathbf{E}'\mathbf{y}_X \\ \mathbf{x} = \mathbf{E}\mathbf{x}_Y, & \mathbf{x}_Y = \mathbf{E}'\mathbf{x} \\ \mathbf{p}_X = \mathbf{E}\mathbf{p}_Y, & \mathbf{p}_Y = \mathbf{E}'\mathbf{p}_X \end{cases} \tag{4}$$

We shall show that, with some suitable algebraic permutations of the elements of **S**, it is possible to reformulate both the matrices **D** and the vectors **p** in (2) and (3) according either to the $\{(x_i, y_i, p_i)\}_X$ or to the $\{(x_i, y_i, p_i)\}_Y$ ordering, maintaining both Gini and concentration indexes unchanged. This leads to rewrite the expressions of formula (2) as

$$G_Y = \frac{1}{2\mu_Y N^2}\mathbf{p}_X'\left(\mathbf{ESE}' \square \mathbf{D}_{Y|X}\right)\mathbf{p}_X \quad \text{and} \quad G_X = \frac{1}{2\mu_X N^2}\mathbf{p}_X'\left(\mathbf{S} \square \mathbf{D}_X\right)\mathbf{p}_X \tag{5}$$

or as

$$G_Y = \frac{1}{2\mu_Y N^2}\mathbf{p}_Y'\left(\mathbf{S} \square \mathbf{D}_Y\right)\mathbf{p}_Y \quad \text{and} \quad G_X = \frac{1}{2\mu_X N^2}\mathbf{p}_Y'\left(\mathbf{E}'\mathbf{SE} \square \mathbf{D}_{X|Y}\right)\mathbf{p}_Y \tag{6}$$

where $\mathbf{D}_{Y|X} = \left(\mathbf{j}\mathbf{y}_X' - \mathbf{y}_X\mathbf{j}'\right)$ and $\mathbf{D}_{X|Y} = \left(\mathbf{j}\mathbf{x}_Y' - \mathbf{x}_Y\mathbf{j}'\right)$, respectively.

Moreover, $C_{Y|X}$ can be given in the following alternative form:

$$C_{Y|X} = \frac{1}{2\mu_Y N^2}\mathbf{p}_Y'\left(\mathbf{E}'\mathbf{SE} \square \mathbf{D}_Y\right)\mathbf{p}_Y \tag{7}$$

Proof

Consider G_X as specified in (2) and (6). As $\mathbf{EE}' = \mathbf{E}'\mathbf{E} = \mathbf{I}$, the following holds:

$$\mathbf{p}_X'\left(\mathbf{S} \square \mathbf{D}_X\right)\mathbf{p}_X = \mathbf{p}_X'\mathbf{EE}'\left(\mathbf{S} \square \mathbf{D}_X\right)\mathbf{EE}'\mathbf{p}_X = \mathbf{p}_Y'\left(\mathbf{E}'\mathbf{SE} - \mathbf{E}'\mathbf{D}_X\mathbf{E}\right)\mathbf{p}_Y$$

by keeping in mind the noteworthy property of the Hadamard product, $\mathbf{E}'\left(\mathbf{S} \square \mathbf{D}_X\right)\mathbf{E} = \left(\mathbf{E}'\mathbf{SE}\right) \square \left(\mathbf{E}'\mathbf{D}_X\mathbf{E}\right)$ (Faliva 1996, property *vii,* page. 157).

Noticing that

$\mathbf{E}'\mathbf{D}_X\mathbf{E} = \mathbf{E}'\left(\mathbf{j}\mathbf{x}' - \mathbf{x}\mathbf{j}'\right)\mathbf{E} = \left(\mathbf{j}\mathbf{x}'\mathbf{E} - \mathbf{E}'\mathbf{x}\mathbf{j}'\right) = \left(\mathbf{j}\mathbf{x}_Y' - \mathbf{x}_Y\mathbf{j}'\right) = \mathbf{D}_{X|Y}$, as $\mathbf{E}'\mathbf{j} = \mathbf{j}$ and $\mathbf{j}'\mathbf{E} = \mathbf{j}'$, the equivalence of expression (2) and expression (6) for G_X is proved.

The equivalence of expressions (2) and (5) for G_Y can be likewise proved. Indeed, the following holds:

$$\mathbf{p}_Y'\left(\mathbf{S} \square \mathbf{D}_Y\right)\mathbf{p}_Y = \mathbf{p}_Y'\mathbf{E}'\mathbf{E}\left(\mathbf{S} \square \mathbf{D}_Y\right)\mathbf{E}'\mathbf{E}\mathbf{p}_Y = \mathbf{p}_X'\left(\mathbf{ESE}' - \mathbf{D}_{Y|X}\right)\mathbf{p}_X$$

upon noticing that $\mathbf{E}\mathbf{D}_Y\mathbf{E}' = \mathbf{j}\mathbf{y}'\mathbf{E}' - \mathbf{E}\mathbf{y}\mathbf{j}' = \mathbf{j}\mathbf{y}_X' - \mathbf{y}_X\mathbf{j}' = \mathbf{D}_{Y|X}$.

As far as $C_{Y|X}$ is concerned, expression (3) turns out to be equivalent to expression (7), upon noticing that

$$\mathbf{E'D}_{Y|X}\mathbf{E} = \mathbf{jy}_X\mathbf{'E} - \mathbf{E'y}_X\mathbf{j'} = \mathbf{jy'} - \mathbf{yj'} = \mathbf{D}_Y .$$

2. Introducing groups

A population of income earners can be partitioned into H groups, $H \in Nn$, which can be characterized by income sources or by social and demographic aspects: typical group characterizations are family composition, dependent/non-dependent worker, men/women, geographic area, etc.

Dagum (1997) decomposes the Gini coefficient into *within groups* (henceforth *W*) and an *across groups* (henceforth *AG*) component. Dagum calls this latter component *gross between*).

Hence $G_Y = G_Y^W + G_Y^{AG}$. In addition, Dagum splits the *AG* component into a *between* and a *transvariation* component: $G_Y^{AG} = G_Y^B + G_Y^T$. The between component G_Y^B is the Gini (weighed) index which results when all values within the same group are replaced by their (weighed) average; the transvariation component G_Y^T measures the overlapping among groups: it is zero when no overlapping exists and it is equal to G_Y^{AG} when all group averages are equal [1]. Extending Dagum's decompositions to concentration indexes, we can split $C_{Y|X}$ into the two components *W* and *AG*, and write $C_{Y|X} = C_{Y|X}^W + C_{Y|X}^{AG}$, accordingly with

$$C_{Y|X}^W = \frac{1}{2\mu_Y N^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(y_i - y_j\right)p_i p_j \cdot I_{i,j\in h} \cdot I_{i-j} \tag{8}$$

$$C_{Y|X}^{AG} = \frac{1}{2\mu_Y N^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(y_i - y_j\right)p_i p_j \cdot \left(1 - I_{i,j\in h}\right) \cdot I_{i-j} \tag{9}$$

In (8) and (9) I_{i-j} is as defined in (1) above, and $I_{i,j\in h}$ is an indicator function: $I_{i,j\in h} = 1$ if both y_i and y_j belong to the same group *h* (*h*=1,2,...,*H*), $I_{i,j\in h} = 0$ if y_i and y_j do not.

[1] For more details on the expression of the Gini components in the Dagum decomposition, see e.g. Monti (2008).

Similar expressions hold for $C_{Y|Y}^{W} = G_{Y}^{W}$, $C_{Y|Y}^{AG} = G_{Y}^{AG}$ and $C_{X|X}^{W} = G_{X}^{W}$, $C_{X|X}^{AG} = G_{X}^{AG}$. In particular, for what concerns G^{W} and G^{AG}, the product $\left(y_i - y_j\right) \cdot I_{i-j}$ can be replaced by the absolute difference $\left|y_i - y_j\right|$.

In order to formalize compact matrix forms for $C_{Y|X}^{W}$ and $C_{Y|X}^{AG}$, it is worth to introduce a proper notation. More precisely, **J** will denote a $K \times K$ matrix with all elements equal to one, $\mathbf{W}_X = \sum_{h=1}^{H} \mathbf{w}_{X,h} \mathbf{w}_{X,h}'$ a $K \times K$ matrix in the $\left\{\left(x_i, y_i, p_i\right)\right\}_X$ ordering, where $\mathbf{w}_{X,h}$ stands for a $K \times 1$ vector with the i-th entry equal to one if the income in the i-th position belongs to group h (h=1,2,..,H), whereas it is zero otherwise. The matrix $\mathbf{W}_X$, when applied to $\mathbf{S} \square \mathbf{D}_{Y|X}$ in expression (3) allows to detect the $\sum_{h=1}^{H} K_h^2$ differences belonging to the same group from the whole K^2 $\left[s_{i,j} \cdot \left(y_i - y_j\right)\right]$ income differences. Conversely, the matrix $\left(\mathbf{J} - \mathbf{W}_X\right)$, when applied to $\mathbf{S} \square \mathbf{D}_{Y|X}$, allows to detect the $\left(K^2 - \sum_{h=1}^{H} K_h^2\right)$ differences between incomes belonging to different groups. Consider now the following expressions for the W and AG components of $C_{Y|X}$:

$$C_{Y|X}^{W} = \frac{1}{2\mu_Y N^2} \mathbf{p}_X' \left(\mathbf{W}_X \square \mathbf{S} \square \mathbf{D}_{Y|X}\right) \mathbf{p}_X \tag{10}$$

$$C_{Y|X}^{AG} = \frac{1}{2\mu_Y N^2} \mathbf{p}_X' \left[\left(\mathbf{J} - \mathbf{W}_X\right) \square \mathbf{S} \square \mathbf{D}_{Y|X}\right] \mathbf{p}_X \tag{11}$$

It is immediate to verify that $C_{Y|X} = C_{Y|X}^{W} + C_{Y|X}^{AG}$. Similar expressions for $G_Y^{W} = C_{Y|Y}^{W}$ and for $G_Y^{AG} = C_{Y|Y}^{AG}$ can be obtained by substituting $\mathbf{p}_X$ with $\mathbf{p}_Y$, $\mathbf{W}_X$ with $\mathbf{W}_Y$ and $\mathbf{D}_{Y|X}$ with $\mathbf{D}_Y$. Likewise, the corresponding expressions for $G_X^{W} = C_{X|X}^{W}$ and $G_X^{AG} = C_{X|X}^{AG}$ are obtained by replacing μ_Y with μ_X, and $\mathbf{D}_X$ with $\mathbf{D}_{Y|X}$. Observe also [1] that $\mathbf{W}_Y = \mathbf{E}'\mathbf{W}_X\mathbf{E}$ and $\mathbf{W}_X = \mathbf{E}\mathbf{W}_Y\mathbf{E}'$.

Moreover, Dagum (1997) splits G_Y^{AG} into the components G_Y^{B} and G_Y^{T}, bringing subdivision to the fore. Let uss now label each subject triplet of observations on X, Y and P by a pair of indexes (h,i), instead of one as before: h refers to the group (h=1,2,….,H), whereas i (i=1,2,….,K_h) refers to the position

[1] $\mathbf{w}_{X,h} = \mathbf{E}\mathbf{w}_{Y,h}$ and $\mathbf{w}_{Y,h} = \mathbf{E}'\mathbf{w}_{X,h}$.

that the subject occupies within the *h*-th group; note that $\sum_{i=1}^{K_h} p_{h,i} = N_h$ and $\sum_{h=1}^{H}\sum_{i=1}^{K_h} p_{h,i} = \sum_{h=1}^{H} N_h = N$.

Dagum's representations are:

$$G_Y = \frac{1}{2\mu_Y N^2}\sum_{h=1}^{H}\sum_{g=1}^{H}\left(\sum_{i=1}^{K_h}\sum_{j=1}^{K_g}\left|y_{h,i} - y_{g,j}\right| p_{h,i}p_{g,j}\right) \tag{12}$$

$$G_Y^W = \frac{1}{2\mu N^2}\sum_{h=1}^{H}\sum_{i=1}^{K_h}\sum_{j=1}^{K_h}\left|y_{h,i} - y_{h,j}\right| p_{h,i}p_{h,j} \tag{13}$$

$$G_Y^{AG} = \frac{1}{2\mu_Y N^2}\sum_{h=1}^{H}\sum_{g\neq h}^{H}\left(\sum_{i=1}^{K_h}\sum_{j=1}^{K_g}\left|y_{h,i} - y_{g,j}\right| p_{h,i}p_{g,j}\right) \tag{14}$$

$$G_Y^B = \frac{1}{2\mu_Y N^2}\sum_{h=1}^{H}\sum_{g=1}^{H}\left(\sum_{i=1}^{K_h}\sum_{j=1}^{K_g}\left|\mu_{Yh} - \mu_{Yg}\right| p_{h,i}p_{g,j}\right) = \frac{1}{2\mu_Y N^2}\sum_{h=1}^{H}\sum_{g=1}^{H}\left|\mu_{Yh} - \mu_{Yg}\right| \overline{p}_h\overline{p}_g \tag{15a}$$

where μ_{Yh} represents the income average of the *h*-th group (*h*=1, 2, ...,*H*).

$$G_Y^B = \frac{1}{\mu_Y N^2}\sum_{h=2}^{H}\sum_{g=1}^{h-1}\left(\sum_{i=1}^{K_h}\sum_{j=1}^{K_g}\left(y_{h,i} - y_{g,j}\right) p_{h,i}p_{g,j}\right) \tag{15b}$$

$$G_Y^T = \frac{2}{\mu_Y N^2}\sum_{h=2}^{H}\sum_{g=1}^{h-1}\left(\sum_{i}^{K_g}\sum_{j}^{K_h}{}_{\{y_{h,i}<y_{g,j}\}}\left|y_{h,i} - y_{g,j}\right| p_{h,i}p_{g,j}\right) \tag{16}$$

where $\overline{p}_h = \sum_{i=1}^{K_h} p_{h,i}$ and $\overline{p}_g = \sum_{j=1}^{K_g} p_{g,j}$.

We refer to Monti and Santoro (2007), formula (6) in particular, for the derivation of expression (15b). Expressions (12) (13), (14) and the first term on the right hand side in (15a) do not need ranking *Y* values; whereas (15b) and (16) need groups to be ranked according to their averages.

Let us now order the *Y* values (and the related *P* and, possibly, *X* values) so that

(i) within each group they are ranked in a non-decreasing order;

(ii) groups are aligned in a non-decreasing order with respect to the their averages.

Then the *Y* values parade becomes

$$\mathbf{y}_A' = \left[\left(y_{1,1}, y_{1,2}, ...y_{1,K_1}\right), ..., \left(y_{h,1}, y_{h,2}, ...y_{h,K_h}\right), ..., \left(y_{H,1}, y_{H,2}, ...y_{H,K_H}\right)\right] \tag{17}$$

$y_{h,i} \le y_{h,i+1}$ (i=1,2,…,K_h) and $\mu_{Yh} \le \mu_{Yh+1}$ (h=1,2,….H) [1].

We shall denote the ordering given by (17) as the $\{(x_i, y_i, p_i)\}_{AY}$ ordering.

The $\{(x_i, y_i, p_i)\}_{AX}$ ordering can be introduced likewise: according to this ordering the X values, together with the related Y and P values, are distributed into the H groups such that

(i) within each group the x's are ranked in a non-decreasing order;

(ii) groups are in a non-decreasing order with respect to their X averages.

Thus, for what concerns the X values, the $\{(x_i, y_i, p_i)\}_{AX}$ ordering will appear as

$$\mathbf{x}_A' = \left[\left(x_{1,1}, x_{1,2}, ...x_{1,K_1}\right), ..., \left(x_{h,1}, x_{h,2}, ...x_{h,K_h}\right), ..., \left(x_{H,1}, x_{H,2}, ...x_{H,K_H}\right) \right] \qquad (18)$$

$x_{h,i} \le x_{h,i+1}$ (i=1,2,…,K_h) and $\mu_{Xh} \le \mu_{Xh+1}$ (h=1,2,….H) [2].

The vectors $\mathbf{y}_A$ in (17) and $\mathbf{x}_A$ in (18) can be expressed as functions of $\mathbf{y}$ and $\mathbf{x}$ respectively, by introducing proper $K \times K$ permutation matrices $\mathbf{A}_Y$ and $\mathbf{A}_X$, such that $\mathbf{y}_A = \mathbf{A}_Y \mathbf{y}$ and $\mathbf{x}_A = \mathbf{A}_X \mathbf{x}$. Since $\mathbf{A}_Y$ and $\mathbf{A}_X$ are permutation matrices, the following holds: $\mathbf{A}_Y^{-1} = \mathbf{A}_Y'$ and $\mathbf{A}_X^{-1} = \mathbf{A}_X'$.

The Y vector corresponding to the $\{(x_i, y_i, p_i)\}_{AX}$ ordering can be obtained as $\mathbf{y}_{AX} = \mathbf{A}_X \mathbf{y}_X$, and likewise $\mathbf{p}_{AX} = \mathbf{A}_X \mathbf{p}_X$.

Also $\mathbf{x}_{AY} = \mathbf{A}_Y \mathbf{x}_X$ and $\mathbf{p}_{AY} = \mathbf{A}_Y \mathbf{p}_X$ contain the Y and the P elements, respectively, aligned according to the $\{(x_i, y_i, p_i)\}_{AY}$ ordering.

If we work out (3), (10) and (11), by making use of the property $\mathbf{A}_X' \mathbf{A}_X = \mathbf{I}$, we get

$$C_{Y|X} = \frac{1}{2\mu_Y N^2} \mathbf{p}_{AX}' \left(\mathbf{A}_X \mathbf{S} \mathbf{A}_X' \square \mathbf{D}_{Y|AX} \right) \mathbf{p}_{AX} \qquad (19)$$

$$C_{Y|X}^{W} = \frac{1}{2\mu_Y N^2} \mathbf{p}_{AX}' \left(\mathbf{W}_{AX} \square \mathbf{S} \square \mathbf{D}_{Y|AX} \right) \mathbf{p}_{AX} \qquad (20)$$

$$C_{Y|X}^{AG} = \frac{1}{2\mu_Y N^2} \mathbf{p}_{AX}' \left[\left(\mathbf{J} - \mathbf{W}_{AX} \right) \square \mathbf{A}_X \mathbf{S} \mathbf{A}_X' \square \mathbf{D}_{Y|AX} \right] \mathbf{p}_{AX} \qquad (21)$$

where $\mathbf{W}_{AX} = \mathbf{A}_X \mathbf{W}_X \mathbf{A}_X'$ and $\mathbf{D}_{Y|AX} = \left(\mathbf{j} \mathbf{y}_X' \mathbf{A}_X' - \mathbf{A}_X \mathbf{y}_X \mathbf{j}' \right) = \left(\mathbf{j} \mathbf{y}_{AX}' - \mathbf{y}_{AX} \mathbf{j}' \right)$.

For what concerns $C_{Y|X}^{W}$ in (21), it is shown in Appendix A2 that

---

[1] It is not excluded that $y_{h,i} > y_{g,j}$, $g > h$.

[2] Here also it is not excluded that $x_{h,i} > x_{g,j}$, $g > h$.

$$\mathbf{W}_{AX} \square \mathbf{A}_X \mathbf{S} \mathbf{A}_X' = \mathbf{W}_{AX} \square \mathbf{S} .$$

Focusing on $C_{Y|X}^{AG}$ decomposition, notice that:

$$C_{Y|X}^{B} = \frac{1}{2\mu_Y N^2} \mathbf{p}_{AX}' \left[(\mathbf{J} - \mathbf{W}_{AX}) \square \mathbf{S} \square \mathbf{D}_{Y|AX} \right] \mathbf{p}_{AX} \tag{22}$$

$$C_{Y|X}^{T} = \frac{1}{2\mu_Y N^2} \mathbf{p}_{AX}' \left[(\mathbf{J} - \mathbf{W}_{AX}) \square (\mathbf{A}_X \mathbf{S} \mathbf{A}_X' - \mathbf{S}) \square \mathbf{D}_{Y|AX} \right] \mathbf{p}_{AX} \tag{23}$$

Summing (22) and (23) yields (21).

Should $C_{Y|Y} \equiv G_Y$, $C_{Y|Y}^{W} \equiv G_Y^{W}$, $C_{Y|Y}^{AG} \equiv G_Y^{AG}$, $C_{Y|Y}^{B} \equiv G_Y^{B}$ and $C_{Y|Y}^{T} \equiv G_Y^{T}$, then (19), (20), (21), (22) and (23) would take the following forms:

$$G_Y = \frac{1}{2\mu_Y N^2} \mathbf{p}_{AY}' (\mathbf{A}_Y \mathbf{S} \mathbf{A}_Y' \square \mathbf{D}_{AY}) \mathbf{p}_{AY} \tag{24}$$

$$G_Y^{W} = \frac{1}{2\mu_Y N^2} \mathbf{p}_{AY}' (\mathbf{W}_{AY} \square \mathbf{S} \square \mathbf{D}_{AY}) \mathbf{p}_{AY} \tag{25}$$

$$G_Y^{AG} = \frac{1}{2\mu_Y N^2} \mathbf{p}_{AY}' \left[(\mathbf{J} - \mathbf{W}_{AY}) \square \mathbf{A}_Y \mathbf{S} \mathbf{A}_Y' \square \mathbf{D}_{AY} \right] \mathbf{p}_{AY} \tag{26}$$

$$G_Y^{B} = \frac{1}{2\mu_Y N^2} \mathbf{p}_{AY}' \left[(\mathbf{J} - \mathbf{W}_{AY}) \square \mathbf{S} \square \mathbf{D}_{AY} \right] \mathbf{p}_{AY} \tag{27}$$

$$G_Y^{T} = \frac{1}{2\mu_Y N^2} \mathbf{p}_{AY}' \left[(\mathbf{J} - \mathbf{W}_{AY}) \square (\mathbf{A}_Y \mathbf{S} \mathbf{A}_Y' - \mathbf{S}) \square \mathbf{D}_{AY} \right] \mathbf{p}_{AY} \tag{28}$$

where $\mathbf{D}_{AY} = (\mathbf{j}\mathbf{y}'\mathbf{A}_Y' - \mathbf{A}_Y \mathbf{y}\mathbf{j}') = (\mathbf{j}\mathbf{y}_A' - \mathbf{y}_A \mathbf{j}')$ and $\mathbf{W}_{AY} = \mathbf{A}_Y \mathbf{W}_Y \mathbf{A}_Y'$.

The matrix compact forms (24), (25), (26) (27) and (28) correspond to the scalar expressions (19), (20), (21), (22) and (23), respectively.

We conclude this section by providing closed-form expressions for $C_{Y|X}^{B}$ and $C_{Y|X}^{T}$, by bearing in mind $C_{Y|X}^{AG}$, as specified in (11), under the $\{(x_i, y_i, p_i)\}_X$ ordering:

$$C_{Y|X}^{B} = \frac{1}{2\mu_Y N^2} \mathbf{p}_X' \left[(\mathbf{J} - \mathbf{W}_X) \square \mathbf{A}_X' \mathbf{S} \mathbf{A}_X \square \mathbf{D}_{Y|X} \right] \mathbf{p}_X \tag{29}$$

$$C_{Y|X}^{T} = \frac{1}{2\mu_Y N^2} \mathbf{p}_X' \left[(\mathbf{J} - \mathbf{W}_X) \square (\mathbf{S} - \mathbf{A}_X' \mathbf{S} \mathbf{A}_X) \square \mathbf{D}_{Y|X} \right] \mathbf{p}_X \tag{30}$$

3. Redistribution and re-ranking indexes

The redistributive effect of a tax system can be measured by the difference between the Gini index for the pre-tax income distribution X and the Gini index for the post- tax income distribution Y [1]: following e.g. Urban and Lambert (2008), we shall denote difference by the acronym *RE*.

The Atkinson-Plotnick-Kakwani index is generally applied to measure the re-ranking effect generated by a tax system; it is defined as the difference between the Gini index for the post-tax income distribution and the concentration index for net incomes Y in the $\{(x_i, y_i, p_i)\}_X$ ordering [2]. The Atkinson, Plotnick; Kakwani index is usually denoted by the acronym *R*.

In considering the effects of a tax, it may be interesting to evaluate how *RE* and *R* act within and across groups and, eventually, also how they modify both group average positions and group intersections. This can be attained by splitting either *RE* or *R* into the within groups, across groups, between groups and transvariation components, introduced in the previous section.

One of the advantages of the compact expressions introduced in the previous sections is that all indexes can be calculated either aligning incomes according to the pre-tax or according to the post-tax ranking. We will present the *RE* and the *R* indexes by writing **D** matrices and **p** vectors either according to the $\{(x_i, y_i, p_i)\}_X$ or the $\{(x_i, y_i, p_i)\}_Y$ orderings, when individual income units are considered. Here, for the sake of shortness, the decompositions of *RE* will be reported only according to the $\{(x_i, y_i, p_i)\}_X$ ordering, and, conversely, *R* decompositions will be written according to the $\{(x_i, y_i, p_i)\}_Y$ ordering. All indexes could be also represented either according to the $\{(x_i, y_i, p_i)\}_{AX}$ or to the $\{(x_i, y_i, p_i)\}_{AY}$ orderings [3].

3.1. The RE index

From the definition of *RE* we can write

$$RE = G_X - G_Y = \left(G_X^W + G_X^{AG}\right) - \left(G_Y^W + G_Y^{AG}\right) = \left(G_X^W + G_X^B + G_X^T\right) - \left(G_Y^W + G_Y^B + G_Y^T\right)$$

Rearranging terms we get

$$RE = \left(G_X^W - G_Y^W\right) + \left(G_X^{AG} - G_Y^{AG}\right) = RE^W + RE^{AG} \tag{31}$$

[1] See e.g. Lambert (2001, Ch. 2, Section 2.5).
[2] Plotnick (1981), Lambert (2001, Ch. 2, Section 2.5).
[3] The formulae that are not reported in this article, will be provided to anyone on request.

Here, in what concerns RE^{AG}, bearing in mind that $G^{AG} = G^B + G^T$, we get

$$RE^{AG} = \left(G_X^B - G_Y^B\right) + \left(G_X^T - G_Y^T\right) = RE^B + RE^T \tag{32}$$

From (5) ad (6) it follows that $RE = G_X - G_Y =$

$$= \frac{1}{2\mu_X \mu_Y N^2} \mathbf{p}_X' \left[\mu_Y \left(\mathbf{S} \hat{} \mathbf{D}_X \right) - \mu_X \left(\mathbf{ESE}' \hat{} \mathbf{D}_{Y|X} \right) \right] \mathbf{p}_X \tag{33a}$$

$$= \frac{1}{2\mu_X \mu_Y N^2} \mathbf{p}_Y' \left[\mu_Y \left(\mathbf{E}'\mathbf{SE} \hat{} \mathbf{D}_{X|Y} \right) - \mu_X \left(\mathbf{S} \hat{} \mathbf{D}_Y \right) \right] \mathbf{p}_Y \tag{33b}$$

The RE^W components can be written, according to (32) and bearing in mind (10) as $RE^W = G_X^W - G_Y^W =$

$$= \frac{1}{2\mu_X \mu_Y N^2} \mathbf{p}_X' \left\{ \mathbf{W}_X \hat{} \left[\mu_Y \left(\mathbf{S} \hat{} \mathbf{D}_X \right) - \mu_X \left(\mathbf{ESE}' \hat{} \mathbf{D}_{Y|X} \right) \right] \right\} \mathbf{p}_X \tag{34}$$

Likewise the RE^{AG} components can be written as $RE^{AG} = G_X^{AG} - G_Y^{AG} =$

$$= \frac{1}{2\mu_X \mu_Y N^2} \mathbf{p}_X' \left\{ \left(\mathbf{J} - \mathbf{W}_X \right) \hat{} \left[\mu_Y \left(\mathbf{S} \hat{} \mathbf{D}_X \right) - \mu_X \left(\mathbf{ESE}' \hat{} \mathbf{D}_{Y|X} \right) \right] \right\} \mathbf{p}_X \tag{35}$$

Resorting to (29) and (30), RE^B and RE^T can be rewritten as

$$RE^B = G_X^B - G_Y^B =$$

$$= \frac{1}{2\mu_X \mu_Y N^2} \mathbf{p}_X' \left\{ \left(\mathbf{J} - \mathbf{W}_X \right) \hat{} \left[\mu_Y \left(\mathbf{A}_X'\mathbf{SA}_X \hat{} \mathbf{D}_X \right) \right.\right.$$

$$\left.\left. - \mu_X \left(\mathbf{EA}_Y'\mathbf{SA}_Y\mathbf{E}' \hat{} \mathbf{D}_{Y|X} \right) \right] \right\} \mathbf{p}_X \tag{36}$$

$$RE^T = G_X^T - G_Y^T =$$

$$= \frac{1}{2\mu_X \mu_Y N^2} \mathbf{p}_X' \left\{ \left(\mathbf{J} - \mathbf{W}_X \right) \hat{} \left[\mu_Y \left(\mathbf{S} - \mathbf{A}_X'\mathbf{SA}_X \right) \hat{} \right.\right.$$

$$\left.\left. \mathbf{D}_X - \mu_X \mathbf{E}\left(\mathbf{S} - \mathbf{A}_Y'\mathbf{SA}_Y \right)\mathbf{E}' \hat{} \mathbf{D}_{Y|X} \right] \right\} \mathbf{p}_X \tag{37}$$

3.2. The R (Atkinson-Plotnick-Kakwani) index

From the definition of R we can write

$$R = G_Y - C_{Y|X} = \left(G_Y^W + G_Y^{AG}\right) - \left(C_{Y|X}^W + C_{Y|X}^{AG}\right) = \left(G_Y^W + G_Y^B + G_Y^T\right) - \left(C_{Y|X}^W + C_{Y|X}^B + C_{Y|X}^T\right)$$

Rearranging the terms we get

$$R = \left(G_Y^W - C_{Y|X}^W\right) + \left(G_Y^{AG} - C_{Y|X}^{AG}\right) = R^W + R^{AG} \tag{38}$$

and in particular, for what concerns R^{AG}, we have

$$R^{AG} = \left(G_Y^B - C_{Y|X}^B\right) + \left(G_Y^T - C_{Y|X}^T\right) = R^B + R^T \tag{39}$$

When considering income units individually, from (2), (3), (5) and (7) the index R, and its components, can be written as follows

$$R = G_Y - C_{Y|X} = \frac{1}{2\mu_Y N^2}\mathbf{p}_Y'\left[(\mathbf{S} - \mathbf{E'SE}) \square\ \mathbf{D}_Y\right]\mathbf{p}_Y \tag{40a}$$

$$= \frac{1}{2\mu_Y N^2}\mathbf{p}_X'\left[(\mathbf{ESE'} - \mathbf{S}) \square\ \mathbf{D}_{Y|X}\right]\mathbf{p}_X \tag{40b}$$

From (10) and (40a) it follows that

$$R^W = G_Y^W - C_{Y|X}^W = \frac{1}{2\mu_Y N^2}\mathbf{p}_Y'\left[\mathbf{W}_Y \square\ (\mathbf{S} - \mathbf{E'SE}) \square\ \mathbf{D}_Y\right]\mathbf{p}_Y \tag{41}$$

From (11) and (40a) it follows that

$$R^{AG} = G_Y^{AG} - C_{Y|X}^{AG} = \frac{1}{2\mu_Y N^2}\mathbf{p}_Y'\left[(\mathbf{J} - \mathbf{W}_Y) \square\ (\mathbf{S} - \mathbf{E'SE}) \square\ \mathbf{D}_Y\right]\mathbf{p}_Y \tag{42}$$

From (29) the component R^B of R can be expressed as

$$R^B = G_Y^B - C_{Y|X}^B =$$

$$\frac{1}{2\mu_Y N^2}\mathbf{p}_Y'\left[(\mathbf{J} - \mathbf{W}_Y) \square\ (\mathbf{A}_Y'\mathbf{SA}_Y - \mathbf{E'A}_X'\mathbf{SA}_X\mathbf{E}) \square\ \mathbf{D}_Y\right]\mathbf{p}_Y \tag{43}$$

From (30) the component R^T can be expressed as

$$R^T = G_Y^T - C_{Y|X}^T =$$

$$\frac{1}{2\mu_Y N^2}\mathbf{p}_Y'\left\{(\mathbf{J} - \mathbf{W}_Y) \square\ \left[(\mathbf{S} - \mathbf{A}_Y'\mathbf{SA}_Y) - \mathbf{E}'(\mathbf{S} - \mathbf{A}_X'\mathbf{SA}_X)\mathbf{E}\right] \square\ \mathbf{D}_Y\right\}\mathbf{p}_Y \tag{44}$$

Either from the definitions of R^{AG} and R^{B} or by rearranging the terms in (44), R^{T} can be given the following representations:

$$R^{T} = \left(R^{AG} - R^{B}\right) = \frac{1}{2\mu_{Y}N^{2}}\mathbf{p}_{Y}'\left\{\left(\mathbf{J} - \mathbf{W}_{Y}\right)\square\right.$$

$$\left.\left[\left(\mathbf{S} - \mathbf{E}'\mathbf{S}\mathbf{E}\right) - \left(\mathbf{A}_{Y}'\mathbf{S}\mathbf{A}_{Y} - \mathbf{E}'\mathbf{A}_{X}'\mathbf{S}\mathbf{A}_{X}\mathbf{E}\right)\right]\square\ \mathbf{D}_{Y}\right\}\mathbf{p}_{Y} \qquad (45)$$

4. The issue of the signs of *R* and its components

We will now analyse the signs of *R* and of its decompositions by making use of the matrix tools introduced in the previous sections. Although most of the results presented in this section are available in the specialized literature[1], we think that our reappraisal of the issue through a tailor-made matrix toolkit provides some additional insights on the matter. Demonstrations will be carried out by inspecting the quadratic form which the *R* index and its decompositions are proportional to.

R

It is well known that for the concentration *C* index the property $-G \leq C \leq +G$ holds[2], from which it follows that $R = G_{Y} - C_{Y|X} \geq 0$. This result will be proved considering expression (40a).

Statement 1

The quadratic form $\mathbf{p}_{Y}'\left[\left(\mathbf{S} - \mathbf{E}'\mathbf{S}\mathbf{E}\right)\square\ \mathbf{D}_{Y}\right]\mathbf{p}$ is non-negative definite.

Proof

Recall that (*i*) matrix $\mathbf{S} = \left[s_{i,j}\right]$ has all super-diagonal elements equal to +1 and sub-diagonal ones equal to −1; (*ii*) the elements of $\mathbf{E}'\mathbf{S}\mathbf{E} = \left[s_{i,j}^{e}\right]$ may not necessarily respect the same repartition as in $\mathbf{S}$, due to permutations performed by $\mathbf{E}$. Thus, for all entries of $\mathbf{S}$ and $\mathbf{E}'\mathbf{S}\mathbf{E}$ which present the same values, $s_{i,j} - s_{i,j}^{e} = 0$, otherwise for $i<j$ we would have $s_{i,j} - s_{i,j}^{e} = 2$ and, for $i>j$, $s_{i,j} - s_{i,j}^{e} = -2$. Bearing in mind that for $i<j$, the matrix $\mathbf{D}_{Y} = \left[d_{i,j}^{Y}\right]$ has super-diagonal elements non-negative and sub-diagonal ones non-positive, the product

[1] Mussini (2008, Ch. 6, § 6.1, page 92) discusses the signs of *R* and its components R^{W}, R^{B} and R^{T}. The author observes also that R^{T} can be positive, null or negative in the framework of non contiguous pre-tax income groups: the proofs reported here complete the author's statements, especially in what concerns R^{T}. See also Vernizzi (2007) for considerations on *G* and *C* components especially for pre-tax non overlapping groups.

[2] Kakwani (1980, Corollary 8.7, page 175).

$\left(s_{i,j}-s_{i,j}^{e}\right)\cdot d_{i,j}^{Y}$ will in any case result to be non-negative, which proves the Statement.

R^W and R^{AG}

We will prove that $R^W = G_Y^W - C_{Y|X}^W \geq 0$ and $R^{AG} = G_Y^{AG} - C_{Y|X}^{AG} \geq 0$, by considering expressions (41) and (42) respectively.

Statement 2

The quadratic forms

$\mathbf{p}_Y'\left[\mathbf{W}_Y \square \left(\mathbf{S}-\mathbf{E'SE}\right) \square \mathbf{D}_Y\right]\mathbf{p}_Y$ and $\mathbf{p}_Y'\left[\left(\mathbf{J}-\mathbf{W}_Y\right) \square \left(\mathbf{S}-\mathbf{E'SE}\right) \square \mathbf{D}_Y\right]\mathbf{p}_Y$

are non-negative definite. Statement 2 is just a corollary of Statement 1.

R^B

We now prove that $R^B = G_Y^B - C_{Y|X}^B \geq 0$. In order to carry out the proof as for the previous Statements, it is convenient to consider a matrix compact form that corresponds in a straightforward manner to the second term in the right hand side of (15a). Let us define the H×1 vector $\boldsymbol{\mu}_Y = \left[\mu_{Y1}, \mu_{Y2}, \ldots \mu_{YH}\right]'$ of group averages, $\mu_{Yh} \leq \mu_{Yh+1}$ (h=1,2,….H), the H×1 vector $\overline{\mathbf{p}}_Y = \left[\overline{p}_1, \overline{p}_2, \ldots \overline{p}_H\right]'$ of group weights $\overline{p}_h = \sum_{i=1}^{K_h} p_{h,i}$ and the H×H matrix $\overline{\mathbf{D}}_Y = \left(\mathbf{1}\boldsymbol{\mu}_Y' - \boldsymbol{\mu}_Y\mathbf{1}'\right)$ of group average differences. Then

$$G_Y^B = \frac{1}{2\mu_Y N^2}\sum_{h=1}^{H}\sum_{g=1}^{H}\left|\mu_{Y.h} - \mu_{Y.g}\right|\overline{p}_h\overline{p}_g = \frac{1}{2\mu_Y N^2}\overline{\mathbf{p}}_Y'\left[\mathbf{S} \square \overline{\mathbf{D}}_Y\right]\overline{\mathbf{p}}_Y \qquad (46)$$

where $\mathbf{S}$ is now an H×H matrix.

After having defined $\boldsymbol{\mu}_{Y|X}$ and $\overline{\mathbf{p}}_X$, respectively, as the H×1 vector of μ_{Yh} and the H×1 vector of $\overline{p}_h$, aligned according to the $\left\{\left(x_i, y_i, p_i\right)\right\}_{AX}$ order, and the H×H matrix $\overline{\mathbf{D}}_{Y|X} = \left(\mathbf{1}\boldsymbol{\mu}_{Y|X}' - \boldsymbol{\mu}_{Y|X}\mathbf{1}'\right)$, (22) can be rewritten in this way:

$$C_Y^B = \frac{1}{2\mu_Y N^2}\overline{\mathbf{p}}_X'\left[\mathbf{S} \square \overline{\mathbf{D}}_{Y|X}\right]\overline{\mathbf{p}}_X \qquad (47)$$

Finally, by denoting by $\overline{\mathbf{E}}$ the H×H full rank permutation matrix such that $\boldsymbol{\mu}_{Y|X} = \overline{\mathbf{E}}\boldsymbol{\mu}_Y$, $\boldsymbol{\mu}_Y = \overline{\mathbf{E}}'\boldsymbol{\mu}_{Y|X}$, $\overline{\mathbf{p}}_X = \overline{\mathbf{E}}\overline{\mathbf{p}}_Y$ and $\overline{\mathbf{p}}_Y = \overline{\mathbf{E}}'\overline{\mathbf{p}}_X$, R^B can be rewritten as

$$R^B = G_Y^B - C_Y^B = \frac{1}{2\mu_Y N^2}\overline{\mathbf{p}}_Y'\left[\left(\mathbf{S}-\overline{\mathbf{E}}'\mathbf{S}\overline{\mathbf{E}}\right) \square \overline{\mathbf{D}}_Y\right]\overline{\mathbf{p}}_Y \qquad (48)$$

Statement 4

The quadratic form

$\bar{\mathbf{p}}_Y'\left[\left(\mathbf{S}-\bar{\mathbf{E}}'\mathbf{S}\bar{\mathbf{E}}\right)\circ\ \bar{\mathbf{D}}_Y\right]\bar{\mathbf{p}}_Y$ is n. n. definite.

Proof

Considerations analogous to those reported above hold for $\left(\mathbf{S}-\bar{\mathbf{E}}'\mathbf{S}\bar{\mathbf{E}}\right)\circ\ \bar{\mathbf{D}}_Y$. In $\bar{\mathbf{D}}_Y$ the super-diagonal entries are non-negative, the sub-diagonal entries are non-positive: while the former are multiplied either by 0 or by +2 entries which are in the super-diagonal part of $\left(\mathbf{S}-\bar{\mathbf{E}}'\mathbf{S}\bar{\mathbf{E}}\right)$, the latter by 0 or by −2 entries which are in the sub-diagonal part of $\left(\mathbf{S}-\bar{\mathbf{E}}'\mathbf{S}\bar{\mathbf{E}}\right)$, and hence it is proved that $R^B \geq 0$.

R^T

Differently from R, R^W, G^T and R^B, that are all non-negative, R^T can be either positive or negative, and, obviously, equal to zero.

Statement 5

In expression (44) the quadratic form

$$\mathbf{p}_Y'\left\{\left(\mathbf{J}-\mathbf{W}_Y\right)\circ\ \left[\left(\mathbf{S}-\mathbf{A}_Y'\mathbf{S}\mathbf{A}_Y\right)-\mathbf{E}'\left(\mathbf{S}-\mathbf{A}_X'\mathbf{S}\mathbf{A}_X\right)\mathbf{E}\right]\circ\ \mathbf{D}_Y\right\}\mathbf{p}_Y$$

can be zero, positive or negative.

Proof

Both in matrix $\left(\mathbf{S}-\mathbf{A}_Y'\mathbf{S}\mathbf{A}_Y\right)=\left[{}_Y\omega_{i,j}\right]$ and in matrix $\left(\mathbf{S}-\mathbf{A}_X'\mathbf{S}\mathbf{A}_X\right)=\left[{}_X\omega_{i,j}\right]$ non zero super-diagonal entries are +2, non zero sub-diagonal are −2. Due to permutation performed by $\mathbf{E}'$ and $\mathbf{E}$, $\mathbf{E}'\left(\mathbf{S}-\mathbf{A}_X'\mathbf{S}\mathbf{A}_X\right)\mathbf{E}=\left[{}_X\omega^e_{i,j}\right]$ can present some −2 as super-diagonal entries and, symmetrically, some +2 as sub-diagonal entries: hence, not considering the cases when both ${}_Y\omega_{i,j}$ and ${}_X\omega^e_{i,j}$ are zero, the super-diagonal differences in $\left[\left(\mathbf{S}-\mathbf{A}_Y'\mathbf{S}\mathbf{A}_Y\right)-\mathbf{E}'\left(\mathbf{S}-\mathbf{A}_X'\mathbf{S}\mathbf{A}_X\right)\mathbf{E}\right]=\left\{\left[{}_Y\omega_{i,j}\right]-\left[{}_X\omega^e_{i,j}\right]\right\}$ may assume values $[2]-[2]=0$, $[2]-[0]=2$, $[2]-[-2]=4$, $[0]-[-2]=2$, $[0]-[2]=-2$. It follows that non-negative super-diagonal entries of $\mathbf{D}_Y$ can be multiplied by a negative value. Symmetrically, sub-diagonal entries of $\left[\left(\mathbf{S}-\mathbf{A}_Y'\mathbf{S}\mathbf{A}_Y\right)-\mathbf{E}'\left(\mathbf{S}-\mathbf{A}_X'\mathbf{S}\mathbf{A}_X\right)\mathbf{E}\right]$ can now be equal not only to $[-2]-[0]=-2$, $[-2]-[2]=-4$, and to $[0]-[2]=-2$, but also to $[0]-[-2]=2$, so that non-positive sub-diagonal entries of $\mathbf{D}_Y$ can be multiplied by a positive value, which proves the Statement.

Conclusions

By use of the Hadamard product, an elegant compact representation in matrix notation has been obtained not only for Gini, concentration indexes and for their decompositions, but for redistribution and re-ranking indexes and their decompositions as well. The matrix toolkit introduced in this paper paves the way to obtain informative expressions for both the said indexes and their components, with incomes aligned either according to the pre-tax non-decreasing order or to the post-tax non-decreasing order.

Moreover, the compact representation introduced in this paper leads to establish in a straightforward manner the signs of the Atkinson-Plotnick-Kakwani index and of its components. We prove that R, R^W, R^{AG} and R^B are non-negative quantities, both when pre-tax income groups do overlap and when do not. In the latter case $R^T \equiv G^T$ ($R^T \equiv R^{AJL}$, following Urban and Lambert, 2008, notation) is non-negative, whereas in the former case we show R^T can be either positive or negative. Even if it is well known that R and $G^T \equiv R^{AJL}$ are non-negative, the proofs presented in this paper are new.

Appendix

On simplifying C_W

We will prove the simplification used in formula (20), that is

$$\mathbf{W}_{AX} \square \mathbf{A}_X \mathbf{S} \mathbf{A}_X' = \mathbf{W}_{AX} \square \mathbf{S} \tag{A1}$$

Proof

The elements $w_{i,j}$ of matrix $\mathbf{W}_X$ and the elements $w^a_{l,m} = \mathbf{a}_l' \mathbf{W}_X \mathbf{a}_m$ of matrix $\mathbf{W}_{AX}$ are equal to 1 if the associated pair of incomes, x_i and x_j, belong to the same group, they are zero otherwise. As all super-diagonal elements in matrix $\mathbf{S}$ are plus 1 and sub-diagonal elements are -1, we have to prove that all super-diagonal elements of matrix $\mathbf{A}_X \mathbf{S} \mathbf{A}_X'$, that are selected by $\mathbf{W}_{AX}$, are 1, and all sub-diagonal elements of $\mathbf{A}_X \mathbf{S} \mathbf{A}_X'$ selected by $\mathbf{W}_{AX}$ are -1.

Observe that incomes belonging to the same group remain ranked in a non decreasing order within each group, also according to the $\{(x_i, y_i, p_i)\}_{AX}$ ordering: therefore

(i) if in the $\{(x_i, y_i, p_i)\}_X$ ordering x_i occupies the *i*-th position and x_j the *j-th* one, with $i<j$, in the $\{(x_i, y_i, p_i)\}_{AX}$ ordering, x_i will occupy the *l*-th position and x_j the *m*-th one with $l<m$;

(ii) symmetrically, in the $\{(x_i, y_i, p_i)\}_X$ ordering, all pairs of incomes $x_i > x_j$, belonging to the same group, will respectively be in positions i and j, $i>j$, and in the $\{(x_i, y_i, p_i)\}_{AX}$ ordering, in positions l and m, $l>m$, respectively.

This implies that the entry $s_{i,j}$ of $\mathbf{S}$ will be shifted to the entry $s^a_{l,m}$ of $\mathbf{A}_X\mathbf{S}\mathbf{A}_X'$, with $l<m$ if $i<j$, and $l>m$ if $i>j$, so that in the super-diagonal part of $\mathbf{W}_{AX} \square \mathbf{A}_X\mathbf{S}\mathbf{A}_X'$ all elements will be equal to 1, and in the sub-diagonal part, all elements will be equal to −1, which proves (A3).

REFERENCES

ARONSON R. J., P. J. LAMBERT (1993), "Inequality decomposition analysis and the Gini coefficient revisited", *The Economic Journal*, 103, pp. 1221–1227.

ARONSON R. J., P. J. LAMBERT (1994), "Decomposing the Gini coefficient to reveal the vertical, horizontal and re-ranking effects of income taxation", *National Tax Journal*, 47, pp. 273–294.

ARONSON R. J., P. J. JOHNSON, P. J. LAMBERT (1994), "Redistributive effect and unequal income tax treatment", *The Economic Journal*, 104, pp. 262–270.

BHATTACHARYA, N., B. MAHALANOBIS (1967). "Regional disparities in household consumption in India", *Journal of the American Statistical Association,* 62, pp. 143–161.

DAGUM C. (1997), "A new approach to the decomposition of Gini income inequality ratio", *Empirical Economics*, 22, pp. 515–531.

FALIVA M. (1983), *Identificazione e Stima nel Modello Lineare ad Equazioni Simultanee*, Vita e Pensiero, Milano.

FALIVA M. (1987), *Econometria, Principi e Metodi*, UTET, Torino.

FALIVA M. (1996), "Hadamard matrix product, graph and system theories: motivations and role in Econometrics", *Matrices and Graphs, Theory and Applications to Economics*, S. Camiz, S. Stefani eds., World Scientific, London.

FALIVA M. (2000), "Su alcune funzioni generalizzate utilizzate nell'analisi dei processi armonizzabili", *Statistica*, 60, 4, pp.655–667.

KAKWANI N. C. (1980), *Income Inequality and Poverty: Methods of Estimation and Policy Applications*, Oxford University Press.

LAMBERT P. (2001), *The Distribution and Redistribution of Income*, Manchester University Press.

LANDENNA G. (1994), *Fondamenti di Statistica Descrittiva*, il Mulino, Bologna.

MONTI M., A. SANTORO (2007), "The Gini decomposition: an alternative formulation with an application to tax reform", *ECINEQ*, 2nd Conference, Berlin, available at http://www.diw.de/documents/dokumentenarchiv/17/60106/papermonti_santoro.pdf.

MONTI, M. (2007), "On the Dagum decomposition of the Gini inequality index", *DEAS Università degli studi di Milano,* W.P. 2007–16.

MONTI M. (2008), "A note on the residual term R in the decomposition of the Gini Index", *Argumenta Oeconomica*, 20, 1, pp. 107–138.

MOOKHERJEE, D., A. SHORROCK (1982). "A Decomposition Analysis of the Trend in UK Income Inequality", *The Economic Journal,* 92, pp. 886–902.

MUSSINI M. (2008), *La misurazione del riordinamento: aspetti metodologici ed un'applicazione con riguardo ai redditi delle famiglie milanesi*, PhD thesis, Università degli Studi di MIlano-Bicocca.

PLOTNICK R. (1981), "A measure of horizontal inequity", *Review of economics and Statistics*, 63, 283–288.

PYATT, G. (1976). "On the Interpretation and desegregations of Gini Coefficient", *Economic Journal,* v86, 243–255.

RAO, V. (1969). "The Decomposition of the Concentration ratio", *Journal of the Royal Statistical Society,* 132, 418–25.

SCHOTT J. F. (2005), *Matrix Analysis for Statistics*, 2nd edition, Wiley, Hoboken, New Jersey.

SILBER, J.(1989). "Factor Components,Population Subgroups and the Computation of the Gini Index of Inequality", *The Review of Economics and Statistics,* 71, 107–115.

URBAN I., P. J. LAMBERT, (2008), "Redistribution, horizontal inequity and re-ranking: how to measure them properly", *Public Finance Review*, 20, n. 10, pp.1–24.

VERNIZZI A. (2007), "Una precisazione sulla scomposizione dell'indice di redistribuzione *RE* di Aronson-Johnson-Lambert e una proposta di estensione

dell'indice di Plotnick", *Economia Pubblica*, 37, n. 1–2, pp. 145–153, and *DEAS*, *Università degli Studi di Milano*, WP 2006–28.

YITZAHAKI, S., R. LERMAN, (1991)."Income stratification and income inequality", *Review of income and wealth,* 37, 313–329.

YITZAHAKI, S. (1994). "Economic distance and overlapping of distributions", *Journal of Econometrics,*. 61, pp. 147–159.

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 525—528

BOOK REVIEW

Task force on the quality of the labour force survey Final report, Eurostat, methodologies and working papers, 2009 edition, 69 pages

The Eurostat publication on quality of the labour force survey will be useful to official statisticians and academic statisticians, practitioners, researchers, analytics and student who are interested in data quality assessment and improvement.

The present document reports the views of the Task Force on the Quality of the Labour Force Survey (LFS) as a result of its six meetings between June 2007 and April 2009. It also takes into account the views expressed by some of the main European institutional users, namely the European Commission's Directorate General "Employment, Social Affairs and Equal Opportunities" (DG EMPL) and the European Central Bank (ECB), and the feedback from the Labour Market Statistics (LAMAS) Working Group in September 2007, April and September 2008.

The Task Force was set up by the LAMAS Working Group at its March 2007 meeting. It was coordinated by Eurostat and composed of national delegates with substantial expertise on the LFS from nine Member States: France, Germany, Greece, Italy, the Netherlands, Poland, Portugal, Spain and the United Kingdom.

This initiative is in line with the continuous work to improve the quality of the LFS. It was conceived to consolidate the gains achieved in recent years and reinforce the status of the LFS – due to its history, sample size and richness of characteristics – as the main statistical source on the labour market.

The goal

The goal of the Task Force was to review the quality of the LFS along with the dimensions of the quality framework for statistical output of the European Statistical System (ESS), detect weaknesses and recommend improvements. The focus of the review was on the estimates of employment and unemployment, as these are the most relevant and largely used indicators produced by the LFS.

Following this review, the Task Force formulated forty-three recommendations on:

- sampling design and sampling errors,

- weighting schemes,
- non-response,
- interviewers and fieldwork organization,
- survey modes and questionnaire,
- information for users,
- coherence,
- comparability of employment and unemployment statistics,
- relevance of the ILO concept of employment and unemployment,
- timeliness and punctuality.

Sampling design and sampling errors

As concerns sampling design and sampling errors, the sample should be balanced over geographical areas and reference weeks. This would both improve the national quarterly and yearly estimates and increase the relevance of the LFS by enabling the production of good monthly estimates.

Target population, sampling frame and population estimates

Moreover, target population, sampling frame and population estimates should be consistent and up to date in order to avoid overcoverage and undercoverage**.**

The importance of ***harmonized rotation patterns*** which allow comparable longitudinal analysis at European level was also highlighted.

Finally, the need was recognized for a clarification of the wording of the precision requirements in Council Regulation 577/98 and for an agreed method to assess compliance with the Regulation.

Non-response

Non-response in the EU, EFTA and candidate countries is rather high (about 20% on average). It is usually selective with respect to employment and unemployment, thus affecting the accuracy of their estimation. Recommendations cover studying, preventing and correcting for non-response. Information on the characteristics of non-respondents should be regularly collected to assess and adjust for non-response bias and to improve fieldwork strategies.

Suitable tools to reassure respondents

Suitable tools to reassure respondents (such as free-toll numbers or presentation letters) should be introduced, with a special view to increase the participation of non-nationals. The use of the wave approach and of dependent interviewing should be considered to reduce response burden.

Weighting schemes

Finally, weighting schemes should take into account specific characteristics of non-respondents to correct for ***non-response bias***.

Role of interviewers

The role of interviewers is crucial for the accuracy of the survey results. Several recommendations and good practices concerning interviewers' contractual features, training, monitoring, and, in general, on the field-work organization were identified with a view to common guidance, as national arrangements concerning these features tend to vary. In particular, in order to boost motivation and minimize turnover, permanent professional interviewers should be used and their remuneration should be adequate to their crucial role for the quality of the survey. Interviewers' training should cover not only the survey content but also how to conduct the interview and to prevent non-response. Periodic debriefing and focus groups should be organized to review and tackle issues. Interviews should be carried out as close as possible to the reference period, to avoid recall problems and support timely production of results.

Computer-assisted questionnaires

The LFS should always be carried out by computer-assisted questionnaires, given that the traditional paper-interviewing mode is no longer suitable to cope with the complexity of the survey. However, the impact of self-administered electronic data collection, including web-based modes, on the measurement of ILO labour status should be carefully investigated. The use of mixed modes should be considered in the light of possible gains relating to response rates, burden and costs, along with likely the mode effects. In any case, any changes to modes, questionnaires and other explanatory survey material should be carefully tested and their impact assessed before introduction.

Lack of coherence

Lack of coherence between LFS and national accounts employment estimates is a major concern, as it may harm the credibility of statistics. In this regard, distinguishing between differences in coverage, scope and definitions from inconsistencies that can be ascribed to the accuracy of the different statistics is of the utmost importance. For this purpose the Task Force recommended the use of reconciliation tables between LFS and National Accounts estimates. The value of appropriate communication to users on the nature of incoherence and the need to provide guidance on which source fits which purpose were also recognised.

Higher input harmonization

The idea of moving towards higher input harmonization is considered too difficult for the moment because of national specificities and needs. Council Regulation no. 577/1998 together with the 12 principles for the formulation of the questions on labour status laid down in Commission Regulation 1897/2000 remain therefore the basis at European level for comparable statistics on employment and unemployment. However, the principles should be reviewed in order to clarify particular ambiguous points. Such clarifications should not necessarily imply changes in the regulation (necessarily via a new legal act), but

should instead be provided as much as possible in working documents such as the explanatory notes.

Introducing innovations

Care should be taken when introducing innovations, as these can negatively impact on comparability of statistics over time. National statistical institutes should always adequately plan and monitor all changes initiated either by Eurostat or by countries in order to assess the statistical effect on time series. Consistent time-series should be produced and disseminated, at least for the headline indicators. For its part, Eurostat should group together innovations it proposes in order to limit the number of potential breaks in time series.

The relevance

The relevance of the ILO labour force concept was confirmed, although the need for supplementary indicators for the ILO unemployment rate, both capturing a wider extent of the labour reserve and allowing longitudinal analysis, was recognized. The variable "Main Status as perceived by respondents", which offers a complementary view to the ILO economic activity status, should be mandatory in the EU-LFS.

The timeliness

The timeline*ss* of the EU-LFS can be significantly improved. This would further enhance its relevance for short-term economic analysis. Establishing a release calendar would be similarly helpful. For this purpose it is essential that the twelve-week deadline in the Regulation as the one for final, not first, data transmission is respected.

Recommended practices

All recommended practices are effective for improving the quality of the LFS and are feasible, as they are already in use in at least one country. Most of the recommendations apply to national statistical institutes, whereas several apply to Eurostat and a few to both.

The full list of recommendations, grouped by subject, is provided at the end of the report. Page numbers in brackets at the end of each recommendation refer to the point in the text where they are discussed.

I hope that similar report will be prepared for other sample surveys, such the Household Budget Survey and the EU-SILC (Statistics of Income and Living Conditions).

Prepared by Jan Kordos,
Warsaw School of Economics

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 529—532

REPORT

The Demographic Future of Poland – a scientific conference Łódź, 17–18 September 2009

Between 17 and 18 November 2009, the Chair of Demography and Social Gerontology, University of Łódź, organized in Łódź a conference entitled „The Demographic Future of Poland". The conference focused on future changes in both demographic structures and the processes that shape them. Changing population structures and processes have short-term as well as long-term consequences. The increasing awareness of impacts induced by population changes makes the latter more interesting not only to demographers, but also to researchers active in other fields of science. Hence, in addition to demographers, the 50 conference participants represented also sociology, social policy, statistics, and gerontology.

The conference was conceived to provide a platform for presenting some selected population projections and exchanging opinions on the pace, directions and consequences of the aforementioned changes. The research areas addressed in the delivered papers and during the discussions concentrated around the below detailed topics:

- Outcomes of the demographic projections for Poland worked out by various forecasting institutions, including the presentation of Poland against the demographic map of Europe.
- Impacts of the changes in population structure, affecting the labour market, social security systems, health situation.
- Changes in the demographic structure with respect to migration processes.
- The influence of demographic changes on the evolution of the consumer goods and services market and predicting future tendencies in this area.
- The methods for forecasting demographic processes.

The conference was opened by Prof. Jolanta Grotowska-Leder, Associate Dean at the Faculty of Economics and Sociology, University of Łódź, Prof. Czesław Domański, Director of the Institute of Statistics and Demography, and Prof. Agnieszka Rossa, Head of the Chair of Demography and Social Gerontology.

The conference consisted of five thematic sessions. The first session, "The Future of the Family and Procreative Behaviour", was chaired by Prof. Zofia Zarzycka. The discussed topics concerned not only the changes in procreative

behaviour and their effect on the future condition and structure of population, but also family policy and its challenges. The relevant consequences were presented by Dr. Milena Lange in the context of the results of demographic forecasts. The presentation delivered by Prof. Andrzej Ochocki and Marta Kawińska, M.Sc., dealt with family policy problems in a broader, European context. Dr. Piotr Szukalski's deliberations concentrated on factors that could enable the Polish population to return to the strict replacement rate. In his opinion, the fertility rate will not exceed 1.7 in the near future, although, as the author stressed, a skilfully pursued pronatalist policy could help reverse this unfavourable trend in the next decades. Another paper presented during the session concerned the influence of religiousness on the family planning decisions. Anna Majdzińska, M.Sc., and Witold Śmigielski, M.Sc., conducted a questionnaire survey among students of the 4th and 5th grades of the University of Łódź to explore how they envisage their family future.

The leading theme of the second session, chaired by Prof. Andrzej Ochocki, was the influence of changing structure and age of population on the labour market. Most speakers during the session, entitled „The Demographic Determinants of Changing Situation of Households and Labour Resources", focused on the problem of aging labour force and its consequences for both employers and employees. One aspect addressed during this part of the conference was the economic activity of Poles in the pre-retirement period, which is the lowest in the EU. It was discussed by Prof. Halina Worach-Kardas and Szymon Kostrzewski, M.Sc. The authors believe that the high rate of unemployment among persons aged 50+ is caused by objective factors, such as workers' lack of relevant occupational skills and health condition, and subjective determinants, including prejudice against older workers. Katarzyna Bałandynowicz-Panfil, M.Sc., was of the opinion that *age management* could be a response to the aging of labour resources. This term encapsulates actions aimed at improving older persons' situation in the workplace, e.g., involving the introduction of flexible working time, training and workplace adaptation to the needs of older workers. It should be remembered, though, as stressed by the author, that these actions must be taken not only by the employers, but also by the governmental institutions. Employment flexibility that Dr. Zofia Szymanek referred to is one of the most important instruments capable of preventing further reduction in the activity of persons at pre-retirement age. An increased proportion of flexible work offers is likely to help activate persons who have prematurely withdrawn from the labour market.

The third session, „The Transforming Demand for Health and Social Benefits", was chaired by Prof. Ireneusz Kuropka. This was another session where the aging of Poland's population and the process' consequences for healthcare and long-term aid were the dominant themes. Health condition and morbidity among persons aged 65 years and older were addressed in the presentation prepared by Olga Gajewska, M.Sc., Prof. Irena Maniecka-Bryła, and Dr. Marek Bryła. Their analysis provided an insight into the situation of persons

living in the county of Płock and using the services of the primary health care physicians. The other speakers in the session concentrated on giving care to seniors. Factors determining the take-up of social aid services by persons aged 60 years and older were presented and analysed by Monika Szlawska, M.Sc., Prof. Irena Maniecka-Bryła and Dr. Marek Bryła. The authors studied the population living in the town and in the rural commune of Zgierz. The next paper dealt with financing care services for the elderly and investigated the long-term care insurance, a subject that is frequently raised during discussions on old age. Dr. Barbara Dembowska pointed out that the growing population of older persons, particularly of seniors in the fourth age, would consequently increase the demand for care that the in-family support will not be able to meet. The growing number of persons in need of assistance will entail the problem of earmarking larger funds for aid services extended to older persons. Zofia Szweda-Lewandowska, M.Sc., attempted in her talk to estimate the future demand for beds in nursing homes among persons aged 75 years and older.

The second day of the conference was broken down into two sessions. The first of them, „The Socio-Economic Challenges in an Aging Society", was chaired by Dr. Alina Potrykowska. The session continued to explore the aging of Poland's population that was already discussed on the first day of the conference. The presentation opening the session was delivered by Prof. Barbara Szatur-Jaworska and its subject was social policy towards old people, exemplified by the case of Warsaw. The author emphasized that one of the key concerns should be a correct diagnosis of the situation of senior citizens, one providing a starting point for formulating programmes for the subpopulation of old persons. The situation of old persons and the quality of their lives were also evaluated by Dr. Dorota Jachimowicz-Wołoszynek. The speaker had studied some attributes of seniors, such as their gender, age, and educational attainment, in the context of their influence on the assessment of life quality. Further, Prof. Bernard Rzeczyński introduced issues related to „gerontotechnology" and discussed how urban layout interferes with the functioning of the elderly in urban space. Dr. Dorota Kozieł and Dr. Małgorzata Kaczmarczyk presented the challenges that ageing society faces in the area of education and preventing the exclusion of seniors. According to the authors, a strategy promoting life-long learning needs to be implemented, so that the socio-economic activity of persons in older age groups can be supported. Karolina Jaskólska, M.Sc., was another speaker considering the situation and role of old persons in society. She underlined that in the near future the role of seniors in a fast-changing knowledge-based society will appear more and more often in social debates. Another presentation delivered by Prof. Agnieszka Rossa dealt with mortality forecasting. The author proposed to use the dynamic life tables based on the Lee-Carter model in computing the amounts of pensions to be drawn by future old-age pensioners. The last presentation delivered during the session discussed the role of the market for consumer goods and services used by the elderly. Patrycja Woszczyk, M.Sc., discussed both the present situation in the area and the relevant forecasts.

The closing session of the conference, chaired by Prof. Jerzy Kowaleski, concerned the international dimension of demographic changes in Poland. In their presentation, Prof. Eugeniusz Zdrojewski and Małgorzata Guzińskia, M.Sc., discussed the size and destinations of permanent migration in Poland. Among other things, the authors pointed to the problem of emigrants with high occupational qualifications. The central theme of the talk presented by Dr. Alina Potrykowska was the future development of demographic processes in Poland in the context of migration.

Prof. Jerzy Kowaleski, the chairperson of the last session, concluded the conference by stressing the importance of the topics addressed and pointing to the diversity of issues that were raised in the presentations. Although ageing and its consequences dominated amongst the topics, in fact all the main areas of interest to demographers were presented.

Speaking on behalf of the conference organizers, Prof. Agnieszka Rossa thanked the participants for their attendance, their presentations and active participation in the discussions. She also invited them to take part in another conference that the Chair of Demography and Social Gerontology, UŁ, is organizing in Łódź – "Life Quality of the Elderly – the Presence and the Future" (22-23 June, 2010). All information relevant to the forthcoming conference can be sought on the website: www.demografia.uni.lodz.pl and www.gerontologia.uni.lodz.pl.

Zofia Szweda-Lewandowska
Agnieszka Rossa

STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 533—537

REPORT

XXVIII Conference on Multivariate Statistical Analysis, MSA 2009, Łódź, Poland, 16–18 November 2009

The XXVIII Conference on **Multivariate Statistical Analysis** was held from 16th to 18th November 2009 in Łódź. Organization of the conference was charged to the Chair of Statistical Methods, University of Łódź and Polish Statistical Association. The conference presented the latest theoretical achievements in the field of the multivariate statistical analysis and its applications. This is a continuation of the issues undertaken on the conferences organized in the past years. The scientific programme of MSA 2009 covered a range of statistical problems, such as: multivariate distributions, statistical tests, nonparametric inference, discrimination analysis, Monte Carlo analysis, Bayesian inference, application of statistical methods finance, insurance, capital markets and risk management.

This year the open meeting of Committee of Statistics and Econometrics PAN was held as part of the Conference. On the meeting, Prof. Agnieszka Rossa delivered a lecture "*Stochastic models of population dynamics"*.

Altogether, there were 70 participants from various academic and research centres in Poland. Concerning the papers, 43 papers were presented in 13 sessions.

The conferences were opened by the Chairman of the Organizing Committee: **Prof. Czesław Domański**. The opening speech was also given by the Pro-vice Chancellor of the University of Łódź **Prof. Antoni Różalski** and the Dean of the Faculty of Economics and Sociology of the University of Łódź **Prof. Jan Gajda.**

The first **plenary session** (chair: Prof. Janusz Wywiał) was devoted to famous Polish statisticians: ***Stanisław Marcin Ulam (1909–1984)*** and ***Ludwik Krzywicki (1859–1941).***

Prof. Mirosław Krzyśko (Adam Mickiewicz University in Poznań) reported a paper titled ***"Stanislaw Marcin Ulam".*** Stanislaw Marcin Ulam (born April 13, 1909, Lwów, died May 13, 1984, Santa Fe, New Mexico, U.S.) was a Polish and American Mathematician (in 1943 he became a U.S. citizen) who helped to develop the Teller-Ulam design which powers the hydrogen bomb, as well as a number of other important mathematical tools (branching processes, Monte Carlo method).

Prof. Czesław Domański (University of Łódź) delivered a lecture about ***"Ludwik Krzywicki"***. Ludwik Krzywicki (born August 21, 1859, Płock, died,

June 10, 1941, Warsaw) was a Polish anthropologist, economist and sociologist. One of the early champions of sociology in Poland, he approached historical materialism from sociological viewpoint. Krzywicki studied mathematics at the University of Warsaw in Poland. Later he began studying anthropology, archaeology and ethnology among others in Paris. From 1919 to 1936 he was a professor at the University of Warsaw.

The titles of the papers of the next sessions of the conference MSA, with the names of the authors, are presented below:

16 November 2009:

Plenary Session II:
Chair: Prof. Mirosław Krzyśko

- *Simulation analysis of accuracy estimation of population mean on strategy dependent on order statistic of auxiliary variable* (Janusz L. Wywiał, Katowice);
- *Comparison between principal component analysis and factor analysis: an informational perspective* (Thierry Dhorne, Francja- Vannes);
- *Analytical interpretation of nonresponse error* (Mirosław Szreder, Gdańsk).

Session III A:
Chair: Prof. Mirosław Szreder

- *The comparison of model based clustering with the heuristic clustering methods* (Ewa Witek, Katowice);
- *The influence of irrelevant variables on classification error in rules induction* (Mariusz Kubus, Opole);
- *Comparison of stability of algorithms in classical and ensemble approach in taxonomy* (Dorota Rozmus, Katowice);
- *Shapley value regression for nonparametric multiple imputation* (Ewa Nowakowska, Krzysztof Puszczak, GFK Polonia);
- *Multiple classification analysis in demography* (Aleksander Suseł, Nowy Sącz).

Session III B:
Chair: Prof. Wiesław Wagner

- *Transformation of economic statistics* (Elżbieta Gołata, Grażyna Dehnel, Poznań);
- *Determinants of supply chains of blood in Poland* (Przemysław Jeziorski, Sebastian Twaróg, Katowice);
- *Innovations and usage of new technologies in Polish small and medium-sized enterprises results of survey* (Tomasz Jurkiewicz, Damian Gajda, Gdańsk);

- *The use of bank services by small and medium-sized enterprises before and after the financial crisis* (Tomasz Jurkiewicz, Damian Gajda, Gdańsk);
- *The estimation of the corruption perception index* (Aleksandra Baszczyńska, Dorota Pekasiewicz, Łódź).

17 November 2009:

Plenary Session I:
Chair: Prof. Janusz Wywiał
- *Some tests for quantile regression models* (Grażyna Trzpiot, Katowice);
- *Comparison of estimators of a probability of success in two models* (Wojciech Zieliński, Warszawa);
- *On effectiveness of Hellwig's variable choice method in linear regression model* (Tadeusz Bednarski, Filip Borowicz, Wrocław).

Session II A:
Chair: Prof. Grażyna Trzpiot
- *Model selection criteria for reduced rank multivariate time series with application in identification of periodic components* (Marcin Hławka, Maciej Kawecki, Wrocław)
- *Depth based strategies to robust estimation of ARIMA parameters* (Daniel Kosiorowski, Kraków);
- *Comparison of selected methods for variable selection in support vector machines* (Michał Trzęsiok, Katowice).

Session II B:
Chair: Prof. Agnieszka Rossa
- *Comparison of selected methods for variable selection in support vector machines* (Wiesław Wagner, Andrzej Mantaj, Rzeszów);
- *On some issues in the practice of domain fraction prediction* (Tomasz Żądło, Katowice);
- *Sustainable development in regional dimension – soft model* (Dorota Perło, Białystok).

Session III A:
Chair: Prof. Wojciech Zieliński
- *Multidimensional smoothing in tables of fertility rates* (Tomasz Jurkiewicz, Gdańsk);
- *Two-sample tests for crossing survival curves at small samples* (Tomasz Jurkiewicz, Ewa Wycinka, Gdańsk);
- *Socioeconomic well-being – soft model* (Dorota Mierzyńska, Białystok);

- *An analysis of job seniority among unemployed. Application of model* (Beata Jackowska, Ewa Wycinka, Gdańsk).

Session III B:
Chair: Prof. Krystyna Katulska
- *Optimum chemical balance weighing design under certain condition* (Bronisław Ceranka, Małgorzata Graczyk, Poznań);
- *Note on the optimum chemical balance weighing design for p=v+1 objects* (Bronisław Ceranka, Małgorzata Graczyk, Poznań);
- *Characteristics of two-dimensional binominal distribution* (Wiesław Wagner, Rzeszów);
- *On the monitoring of the process mean based on the sequence of permutation tests* (Grzegorz Kończak, Katowice).

18 November 2009:

Plenary Session I
Chair: Prof. Mirosław Krzyśko
- *Testing for tail independence in extreme value models – application on Polish stock exchange* (Grażyna Trzpiot, Justyna Majewska, Katowice);
- *Extreme value index of left and right tails for financial time series* (Wiesław Dziubdziela, Michał Stachura, Barbara Wodecka, Kielce);
- *Choosing variables in cluster analysis by means of entropy and detecting unimodal distributions* (Jerzy Korzeniewski, Łódź).

Session II A:
Chair: Prof. Krystyna Pruska
- *The influence of the sample size on the net premium rates in car liability insurance CR* (Anna Szymańska, Łódź);
- *Construction of the aggregative index of work efficiency* (Jacek Białek, Łódź);
- *Statistical analysis of the efficiency of pension systems of the EU and EFTA in 2005–2007* (Artur Mikulec, Łódź).

Session II B:
Chair: Prof. Krystyna Katulska
- *Comparison of methods determination loss distribution function from credit portfolio in CreditRisk+ model* (Agnieszka Pietrzak, Łódź);
- *Non-sampling errors in opinion polls* (Aleksandra Fijałkowska, Łódź).

Plenary Session III:

Chair: Prof. Bronisław Ceranka

- *The role of non-financial enterprises in Poland* (Czesław Domański, Magdalena Motyl, Łódź);
- *The estimation of the Stieltjes transform of spectral functions of covariance matrix* (Anna Witaszczyk, Łódź);
- *The role of probability theory and statistics in methodology of the exact sciences* (Janusz Kupczun, Łódź).

The next Conference on Multivariate Statistical Analysis will be held on November 8–10, 2010 in Lodz. Scientists interested in attending the Conference are kindly requested to send their application to the Scientific Secretary of MSA'2010 to the following address:

Anna Witaszczyk
29th Conference MSA'2010
Chair of Statistical Methods, University of Łódź
90-214 Łódź, Rewolucji 1905 r. nr 41, Poland
phone: (4842) 635 51 78; fax: (4842) 635 53 07
e-mail: msa@uni.lodz.pl
www.msa.uni.lodz.pl

Katarzyna Bolonek-Lasoń
Monika Zielińska-Sitkiewicz